Stony Brook University

Classification of Gastrointestinal Bleeding Data

A Dissertation Presented

by

Adrienne Michelle Chu

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Applied Mathematics and Statistics

(Statistics)

Stony Brook University

May 2009

Stony Brook University

The Graduate School

Adrienne Michelle Chu

We, the dissertation committee for the above candidate for the

Doctor of Philosophy degree, hereby recommend

acceptance of this dissertation.

Dr. Hongshik Ahn – Dissertation Advisor
Professor, Applied Mathematics and Statistics

Dr. Stephen Finch – Chairperson of Defense
Professor, Applied Mathematics and Statistics

Dr. Wei Zhu – Member
Professor, Applied Mathematics and Statistics

Dr. Atul Kumar – Outside Member
Gastroenterologist, Veterans Affairs Medical Center

This dissertation is accepted by the Graduate School

Lawrence Martin
Dean of the Graduate School

Abstract of the Dissertation

Classification of Gastrointestinal Bleeding Data

by

Adrienne Michelle Chu

Doctor of Philosophy

in

Applied Mathematics and Statistics

(Statistics)

Stony Brook University

2009

Acute gastrointestinal bleeding (GIB) is an increasing healthcare problem due to rising NSAID (non-steroidal anti-inflammatory drugs) use in an aging population. In the emergency room (ER), the ER physician can misdiagnose a GIB patient at least 50% of the time. While it is best for a gastroenterologist to diagnose GIB patients, it is not feasible due to time and cost constraints. Classification models can be used to assist the ER physician to diagnose GIB patients more efficiently and effectively, targeting scarce healthcare resources to those who need it the most.

Currently, there have not been models developed which can predict all three sources of bleeding simultaneously (upper, middle, and lower bleeding). Eight classification models were trained and tested by performing ten repetitions of ten-fold cross validation on a 192 patient dataset. The classification models considered were: artificial neural network, boosting, k-nearest neighbor, linear discriminant analysis, logistic regression, random forest, shrunken centroid, and support vector machine. The four response variables classified were: source of bleeding, need for urgent resuscitation, need for urgent endoscopy, and disposition. Performance was assessed by accuracy and balance of sensitivity and specificity. The top three models (random forest, support vector machine, and artificial neural network) were externally validated. It was determined that random forest performed the best overall.

The Rockall and Blatchford scores have been previously developed for upper GIB patients. The random forest model was found to be comparable to these scores for upper GIB patients. In addition, simulation studies were done to compare the eight classification models and to compare to the results obtained with the actual GIB data.

Simulated GIB data that was unbalanced versus balanced and correlated versus independent was considered, with accuracy and balance of sensitivity and specificity being the performance measures of the models. Random forest was again seen to be the best performing model. An online tool was developed for a user-friendly interface that physicians and nurses can utilize. This online tool will be utilized in future studies in the hope this tool or something similar can be adopted for routine use in caring for GIB patients.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would like to give a huge thanks to my advisor, Dr. Hongshik Ahn, for his patience and guidance not only on the dissertation but also with career decisions and for imparting his knowledge of teaching to me. I have learned so much because of Dr. Ahn and feel I have grown intellectually tremendously.

I would like to give a big thanks to Dr. Atul Kumar – it has been a pleasure to work with him and I am grateful for all the opportunities he has given me. Without Dr. Kumar, I would not have the statistical experience I do now.

I would like to thank Dr. Stephen Finch and Dr. Wei Zhu for serving on my preliminary exam committee and dissertation committee as well as teaching me so much that I know about statistics. Thanks also to Dr. Nancy Mendell, for sharing all her statistical knowledge with me. I have learned many valuable lessons from all of them both inside and outside of the classroom.

I also wish to thank Dr. Alan Tucker – his countless emails and chats with him have helped me grow as a teacher and I have learned many things from him. I am grateful he has given me the opportunity to teach here at Stony Brook University for the past 3 ½ years and provided me with tools I can use in my later teaching years.

I would especially like to thank my family and friends for all of their support, in particular, my parents (Tony and Elizabeth), sister (Jennifer), fiancé (Wan Cheong), and best friend (Jen). They have always been there for me, no matter what, and have always believed in me.

The text of this dissertation in part is a reprint of the materials as it appears in the journal article entitled “A decision support system to facilitate management of patients with acute gastrointestinal bleeding,” published in the Artificial Intelligence in Medicine journal. Permission has been granted by the Elsevier publisher, under license number 2154320444575. I wish to acknowledge all co-authors for their consent to use material from this article in my dissertation. The co-authors are: Dr. Hongshik Ahn, Dr. Bhawna Halwan, Dr. Bruce Kalmin, Dr. Everson L.A. Artifon, Dr. Alan Barkun, Dr. Michail G. Lagoudakis, and Dr. Atul Kumar.

Background and Introduction

Acute gastrointestinal bleeding (GIB) is an increasing healthcare problem due to rising non-steroidal anti-inflammatory drug (NSAID) use in an aging population (1). NSAIDs, such as aspirin and ibuprofen, are used to reduce pain, fever, or inflammation. Often times acute GIB occurs in the emergency room (ER) and it is a frontline physician (non-gastroenterologist) who is diagnosing the patient. Delays in intervention usually result from failure to adequately recognize the source and severity of the bleed. Using symptoms alone, physicians predict the location of a gastrointestinal lesion with only up to 40% accuracy as compared to endoscopy on some studies (2). Aside from lack of resources, it would not be feasible for a gastroenterologist to diagnose every single case of acute GIB due to time and cost constraints. In order to reduce further complications and mortality of patients, new strategies need to be developed to help identify those patients in need of urgent resuscitation and endoscopic intervention (3,4).

Although several scoring systems have been developed to risk stratify patients, there is no single model which is popular that helps identify patients with both upper and lower GIB that require urgent intervention (5,6,7,8). The Rockall score, which utilizes clinical data as well as data from endoscopic findings, is used to identify only upper GIB patients who are at high risk. The Blatchford score, only applicable for upper GIB patients as well, uses solely clinical data. This score identifies those patients who are in need of urgent treatment – i.e. who need urgent resuscitation or endoscopy. Other models have been developed, but not one that can be used for both upper and lower GIB and that is straightforward enough for a non-gastroenterologist to use. There are several computational models that are potentially useful and that can identify patients with both upper and lower GIB and determine their need for treatment as well as disposition. Using clinical and laboratory information available within a few hours of patient presentation, the models can be used to predict the source of bleeding, need for intervention (resuscitation and endoscopy) and disposition in patients with acute upper, mid, and lower GIB. The hope is that these classification models can assist the ER physician in diagnosing the patients more efficiently and effectively.

The models considered in classifying responses for the GIB data were artificial neural networks (ANN), boosting, *k*-nearest neighbor (kNN), linear discriminant analysis (LDA), logistic regression (logistic), random forests (RF), shrunken centroid (SC), and support vector machines (SVM). Boosting and random forests are ensemble-based voting methods. SVM predicts new cases by seeing where the new case lies with respect to class boundaries. kNN is an instance-based learning method, while SC and LDA classify new cases by computing the distance away from class centroids. More specific information about each model will be briefly described in the following paragraphs.

1. Model Information

Artificial neural networks (ANNs) are thought to be models for the human brain (9). An ANN consists of "artificial neurons" or nodes connected together by different weights, the connections representing the synapses of the brain. Functions such as sigmoid functions are applied to the nodes and then combined, and the output is obtained (by applying for example the softmax function). If there is no hidden layer in the network, the ANN is simply a linear regression model. If there are one or more hidden layers in the network, then the ANN becomes a non-linear generalization of the linear regression model. The idea of neural networks started in the 1940s, with the first practical ANN being implemented in the 1950s by Frank Rosenblatt (9). The first ANN, a perceptron, was a simple feed-forward model. Later on in the 1970s and 1980s, other more complex models were conceived – multi-layer perceptron networks, Hopfield networks, and Boltzmann machines. An advantage of using ANNs is that they can be very complex models. However, this may lead to overfitting. There are several practical applications for ANNs such as classification of data (medical diagnosis), pattern recognition (identification of faces or object recognition), and sequence recognition (handwritten text recognition).

In 1990, Robert E. Schapire came up with an ensemble voting method called boosting, where several weak classifiers are combined by weighted majority voting (10). By combining the weaker classifiers, one single more powerful and accurate classifier is produced. One well-known boosting algorithm is AdaBoost, which was developed by Schapire and Yoav Freund (11). AdaBoost is an iterative process – for a classifier being trained on a given iteration, when a wrong prediction is made for a particular case, this case gets more heavily weighted on the next iteration. At the end, a sequence of classifiers is obtained, with each new classifier "learning from its mistakes." The final decision is made by weighted majority voting among all the classifiers. Boosting relies heavily on the data that it is given. AdaBoost has been extended and modified to incorporate other methods and ideas, such as using bagging (BagBoost) (12) or using the binomial log-likelihood function in place of the exponential function as the loss function (LogitBoost) (13).

The method of *k*-nearest neighbor classifies a data point by considering the closest *k* neighbors to it. It is thought that the neighbors will be similar to each other (9). "Closest" is defined by either the Euclidean distance or the Mahalanobis distance. The difference between the two types of distances is that the Mahalanobis distance considers the correlations of the dataset and is scale-invariant. This model is simple to implement except it does not perform well with high-dimensional data because neighbors may not be "nearby" (9). Evelyn Fix and J.L. Hodges came up with the idea of nearest neighbors in 1951 (9). In 1986, Craig Stanfill and David Waltz applied the nearest neighbor concept

to the artificial intelligence area (9). The value for *k* is found by performing cross-validation and it is best to use an odd value for *k*, to break ties.

Linear discriminant analysis is a model where the decision boundaries are linear (14). If there are two variables, the decision boundary will be a line. The decision boundary will be a plane for three variables and a hyperplane for more than three variables. For diagonal linear discriminant analysis (DLDA), an equal diagonal covariance matrix is assumed, simplifying the problem. A new case is classified based on how far it is from each class group using the Mahalanobis distance function. The distance is calculated between the point where the new case lies and the "average" point for each group.

Logistic regression is a regression model that fits the log odds of the response to a linear combination of the explanatory variables. It is used mainly for binary responses, although there are extensions for multi-way responses as well. Coefficients are determined by maximizing the likelihood function. Numerical methods, such as the Newton-Raphson algorithm are applied iteratively to find the coefficients. An advantage for using logistic regression is that a model can be clearly and succinctly represented but on the flip side, it might not be able to produce complex models, leading to underfitting. Logistic regression is widely used in areas such as medical and social sciences.

Leo Breiman and Adele Cutler (15) developed the random forest method in 2001. It is an ensemble-based method where a forest of classification trees is grown. A subset of the original data samples is randomly selected with replacement for the training set to grow each tree. At each node on a tree, a random sample from all the variables is selected to determine the best split for that node. The number of variables selected at the first node is the number of variables selected for every node thereafter. All trees are grown to their fullest, with no pruning done. Once the trees are grown, majority voting among the trees classifies a test case. The random forest classifier works well for high dimensional data like microarray data or DNA data.

Nearest shrunken centroid classifiers compare a new case to all possible class centroids (16). These centroids are calculated by taking the difference between the average value for that class and the overall centroid, and then dividing by the standard deviation for that class. The centroids are shrunk towards zero by a threshold value. If a component of the centroid passes zero, then this component is set to zero. Whichever class centroid the new case is closest to becomes its predicted class. The method of shrunken centroid classifiers is an extension of the centroid classifier method – the extension was done by Tibshirani et al, in 2002 (16). Shrunken centroid classification has been applied to gene expression and cancer data.

Support vector machines (SVM) first introduced by Vladimir Vapnik in 1995, finds a linearly separable boundary between the classes (17). If the data is nonseparable in the original feature space, the data is transformed to a higher dimensional space, where the data becomes linearly separable. In classifying a new case, whichever side of the

boundary the case falls on is the predicted class. Solving for the boundaries is essentially a convex optimization problem.

These eight different classification models were evaluated and validated on a 192 patient database. Further, the best performing models were externally validated using an external database, RUGBE (Registry in patients with Upper Gastrointestinal Bleeding under an Endoscopy). To assess whether it would be beneficial to use classification models in practice, the top performing model was compared to existing GIB scoring methods (Rockall and Blatchford scores). In order to use the classification model in practice, it would need to be accessible to anyone. A user-friendly web interface was developed in order to compare how well the classification model performs versus a physician at the hospital. The ultimate goal is for a classification model to be put into routine use at a hospital to help diagnose patients more efficiently and effectively (either through a website or integrated directly into the hospital system).

2. Comparison of Classification Models Applied to GIB Data

2.1 Methods

Eight predictive models including artificial neural networks (ANN), boosting, *k*-nearest neighbor (kNN), linear discriminant analysis (LDA), logistic regression (logistic), random forest (RF), shrunken centroid (SC), and support vector machines (SVM) were trained and their performances compared. All models were run in R (versions 2.3.0 and 2.4.1, downloadable from http://cran.r-project.org/) except for ANN, which was run in STATISTICA (version 7.1, Statsoft, Inc, Tulsa, OK). Model training (evaluation step) was performed on one set of patients and testing (validation step) was done on the remaining patients. The total number of patients was 192 (122 used for the evaluation step and the remaining 70 used for the validation step).

The primary approach was to use selected explanatory variables to predict the response variable, discarding any patients with missing data (there was only at most n=3 patients dropped). In addition, for predicting resuscitation and endoscopy, an alternative selection of input variables was tried. The predicted value of the "source of bleeding" response and other selected input variables (to be detailed later) were utilized to predict the need for resuscitation and endoscopy. Only the evaluation step was done for the alternative approach. Any categorical variables were changed accordingly to indicator variables. In addition, distributions of the variables as well as correlation between variables were examined. The specific explanatory variables chosen to predict each response will be given in Section 2.1.6.

2.1.1 Model Evaluation Step

Ten runs of 10-fold cross validation (CV) were performed for each iteration to obtain a reliable result with low mean square error (MSE) and bias (18). *k*-fold cross validation divides the data into *k* groups, and one group is used as the testing set while the remaining *k*-1 groups are used as the training set. This is repeated until each group is used as a testing set once. Thus for 10-fold cross validation, 90% of the data is used for the training set with the remaining 10% being used for the testing set. For every 10-fold CV, the following statistics were calculated: accuracy (ACC: the sum of correct predictions divided by total predictions), sensitivity (SN: probability that the patient was predicted as positive given patient is truly positive), specificity (SP: probability that the patient was predicted as negative given patient is truly negative), positive predictive value (PPV: probability that the patient is truly positive given a positive prediction), and negative predictive value (NPV: probability that a patient is truly negative given a

negative prediction). For each classification model, the results from all 10 repetitions were then averaged together to give a single result for each statistic calculated.

Besides using accuracy to compare the models' performances, ROC curves were created and areas under the curve (AUC) were compared to assess models' ability to balance sensitivity and specificity. For creating the ROC curves, cutoff points were values between 0 and 1, in increments of 0.1 (first cutoff was 0, second cutoff was 0.1, third cutoff was 0.2, etc.). The Mann-Whitney statistic was calculated, which is equivalent to the area under an ROC curve (19). Additionally, accuracies between the model with the highest accuracy and the other models were compared using McNemar's test. Using the Bonferroni correction to account for multiple comparisons of models, an appropriate alpha value was used for each test to control the error rate. For example, there were 8 models for the resuscitation response, so there were 7 pair-wise comparisons. Thus for an original alpha value of 0.05, the new alpha used for a two-sided test was 0.05/7=0.0071. Models that had good accuracy and high values for area under the ROC curve were considered the best.

2.1.2 Model Validation Step

After models had been trained using a 122 patient database, a separate 70 patient database was utilized as a test set to validate the model. Accuracies from each model were compared for the validation step.

2.2 Individual Model Parameters

Artificial neural network (Neural Networks package): The network used was multilayer perceptrons with back propagation. The error function used was the cross entropy function. A linear synaptic function was used and a combination of the following four activation functions were used: linear, hyperbolic, softmax, and logistic. The number of epochs to train the model was set to 100 although the network always converged in fewer epochs. The learning rate was set to 0.01 and there was one hidden layer in the network. For the source of bleeding response, there were 30 input neurons, 11 hidden neurons (neurons in the hidden layer), and 3 output neurons. For the resuscitation response, there were 28 input neurons, 10 hidden neurons, and 1 output neuron. For the endoscopy response, there were 31 input neurons, 12 hidden neurons, and 1 output neuron. For the disposition response, there were 34 input neurons, 34 hidden neurons, and 1 output neuron. These neurons were chosen automatically by the Neural Networks package.

Boosting (package boost): No features were pre-selected (presel=0) and boosting ran for 10 iterations (mfinal=10). The optimal settings were found by trying different combinations of presel and mfinal. The R Boosting package contains four different

variations of boosting: AdaBoost, LogitBoost, L2Boost and BagBoost. AdaBoost was excluded, as often it did not classify patients correctly and this is the oldest, original boosting method. Performance of the other three boosting methods were similar. Among these three approaches, LogitBoost is included in the comparison.

***k*-nearest neighbor (package class):** Depending on what response variable was being classified, a different *k* was chosen accordingly. By performing CV on several different *k* values, the *k* that yielded the highest accuracy was chosen. The final values chosen were k=7 for source of bleeding and resuscitation, k=11 for endoscopy, and k=3 for disposition.

Linear discriminant analysis (package sma) and logistic regression (package stats) models: All default settings were used.

Random forest (package randomForest): All default settings were used – 500 trees were grown (mtry=500), the number of cases to select in the bootstrap sample for each tree was equal to the number of patients in the dataset (parameter sampsize), and the number of variables randomly sampled at each node of a classification tree was $\text{floor}\left(\sqrt{p}\right)$ where p is the number of explanatory variables (mtry= $\text{floor}\left(\sqrt{p}\right)$). The cutoff parameter used was the default, with each class having equal weight. The variable option of importance was set to true. To predict the outcome of a certain patient, voting was done without normalizing.

Shrunken centroid (package pamr): The best threshold value for each response was found by cross validation. The best threshold value was one that gave the highest accuracies – a threshold of 0.2 was used for all responses.

Support vector machine (package e1071): All default settings were used; variables were scaled; tolerance value to terminate algorithm was set to 0.001; epsilon set to 1 (for insensitive-loss function). Both the radial basis (default) and linear kernels were considered.

2.3 Importance of Each Variable

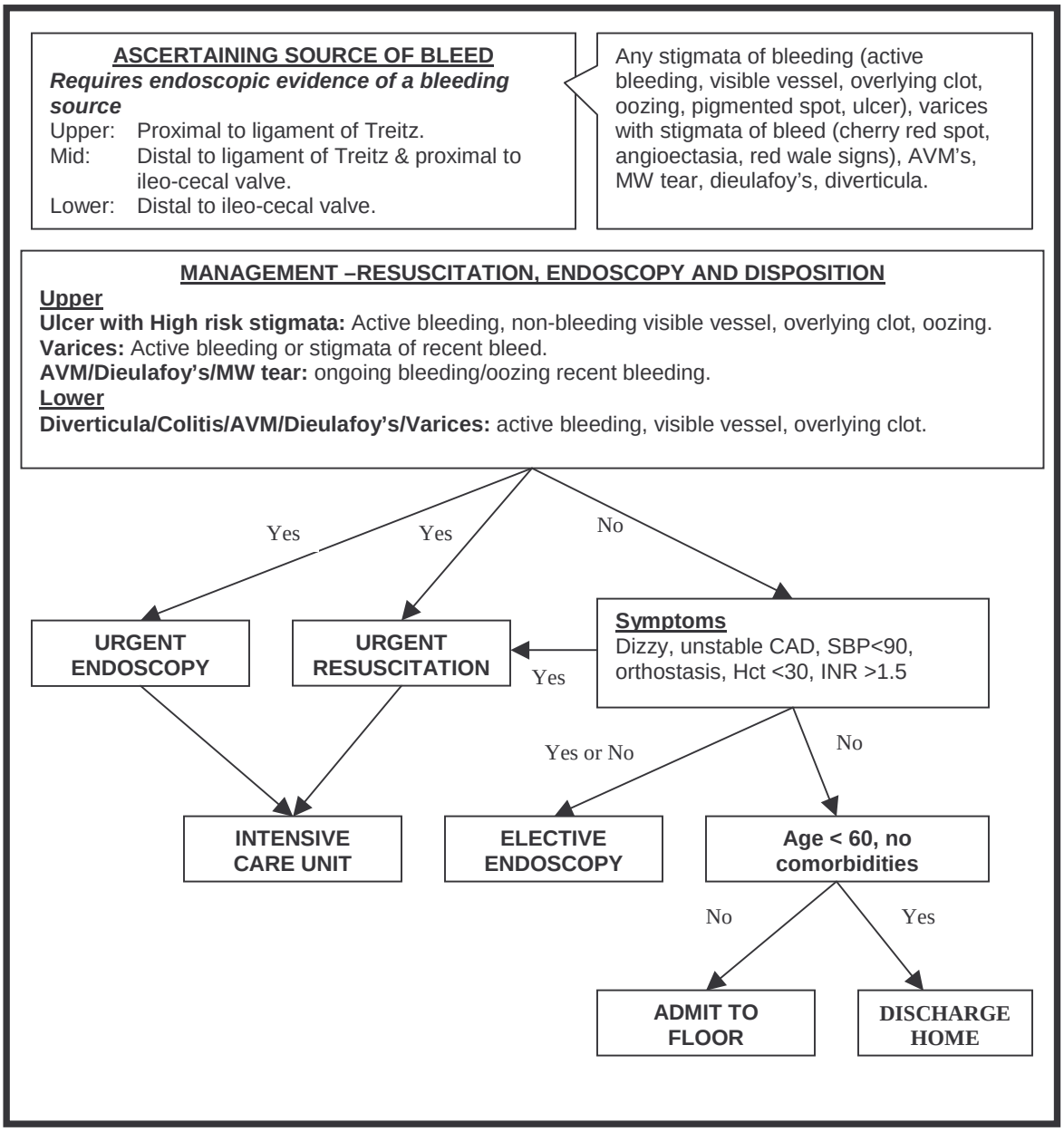
For RF and ANN, the variable importance option was set to true to see if these pre-selected variables were significant for predicting outcomes by the models. Variable importance for RF is given in terms of the mean decrease in accuracy. Hence the higher the number, the greater the importance of the variable. ANN provides information about the importance of a variable without any ranking. The variable importance feature for ANN was therefore repeated ten times in order to obtain rankings for the variables. The number of times a variable was shown to be important was counted – the closer the number was to 10, the more important the variable.

2.4 Patients

Patients with acute GIB were identified from the hospital medical records database using ICD-9 codes for GIB. The study was carried out in compliance with all institutional human investigations committee guidelines. Eligible patients were those presenting with clinical manifestations of acute upper or lower gastrointestinal bleeding who had undergone endoscopy within 24 hours for suspected upper GIB and within 48 hours for suspected lower GIB or if the upper endoscopy was negative. If no obvious source of GIB was identified at either upper or lower endoscopy the patient was expected to have undergone small bowel enteroscopy or capsule endoscopy within 1 week of the initial episode of acute GIB. Records of patients for whom a definite source of bleeding could not be identified or those with missing clinical variables required for model building and testing were discarded. See Figure 1 for the algorithm for the actual diagnosis of patients.

A database of 122 patients was identified from retrospective chart analysis. Clinical data on each patient was entered into a scannable data entry form that was then scanned into an SQL database and manually reviewed for errors. Variables to ascertain clinical outcomes corresponding to patient data included clinical and endoscopic data as listed in Table 1. These variables were identified from review of literature for their ability to predict outcomes amongst patients with acute GIB (20,21). Initially only 70 patients were available when first building the models and having a total of 122 patients resulted in an insignificant improvement in performance of the models (data not shown). A sample size of 122 patients was therefore deemed to be adequate. The additional 70 patients used in the validation step of the model were collected at a later time in a similar study, in compliance with all guidelines.

Figure 1. Schematic of Ascertaining Outcomes for GIB Patients



Note: Refer to Table 1 for abbreviations

Table 1. Clinical and Endoscopic Variables Used to Determine Patient Outcomes

Presentation: Hematemesis, Hematochezia, Melena, Duration of Symptoms, Syncope/Presyncope
Demographics: Age, Gender
Past history: Prior History of GIB, Unstable CAD (coronary artery disease), COPD (chronic obstructive pulmonary disease) Exacerbation, CRF (chronic renal failure), Risk of Stress Ulcer, Cirrhosis, ASA/NSAID (aspirin /non steroidal anti-inflammatory drug) Use, PPI (proton pump inhibitor)
Clinical Exam: SBP/DBP (systolic blood pressure/diastolic blood pressure), HR (heart rate), Orthostasis, NG (nasogastric) Lavage, Rectal Exam
Laboratory Data: Hct (Hematocrit), Drop in Hemotocrit, Plt (Platelet) Count, Cr (Creatinine), BUN (blood urea nitrogen), PT/INR (Prothrombin Time / International Normalized Ratio)
Endoscopic Data: Ulcer, Varix, MW (Mallory-Weiss) Tear, Diverticula, AVM (arterio-venous malformation), Diuelafoy, Other

2.5 Explanatory and Response Variables Used

Several studies in the past have evaluated risk factors of adverse outcomes and clinical predictors of source, severity and outcomes in patients with acute upper and lower GIB. Clinical correlates of source, severity and outcomes amongst patients with acute gastrointestinal bleeding were reviewed and are listed below for each corresponding response variable (6,7,8,20,22,23,24,25,26,27,28,29,30). The response variables predicted are source of bleeding (upper, middle, or lower intestine), need for urgent resuscitation (yes or no), need for urgent endoscopy (yes or no), and disposition (should patient be placed in the intensive care unit (ICU) or not the ICU). See Table 2 for a summary of the response variables.

Table 2. Response Variables

Source: Upper, Mid, or Lower
Resuscitation: Yes or No
Endoscopy: Yes or No
Disposition: ICU or not ICU

Bleeding source: The definitive source of bleeding was the irrefutable identification of a bleeding source at upper endoscopy, colonoscopy, small bowel enteroscopy, or capsule endoscopy. Input variables utilized to predict the source of GIB included: prior history of GIB, hematochezia, hematemesis, melena, syncope/presyncope, risk for stress ulceration, cirrhosis, ASA/NSAID use, blood pressure – systolic and diastolic (SBP/DBP), heart rate (HR), orthostasis, NG lavage, rectal exam, platelet count (Plt.), creatinine (Cr.), BUN, and INR.

Need for urgent blood resuscitation: Urgent blood resuscitation refers specifically to the administration of blood and blood products to correct loss of intravascular volume, and coagulopathy. Variables to predict this outcome included hematochezia, hematemesis, melena, duration of symptoms, syncope/presyncope, unstable CAD, blood pressure, heart rate, orthostasis, NG lavage, rectal exam, hematocrit (Hct.), drop in hematocrit, creatinine, BUN, and INR. Variables used to predict resuscitation using the second approach as described in Section 2.1 included the predicted value of the source of bleeding response, hematochezia, hematemesis, syncope/presyncope, blood pressure, heart rate, orthostasis, NG lavage, and INR.

Need for urgent endoscopy: Variables to predict need for urgent endoscopy included hematochezia, hematemesis, melena, duration of symptoms, syncope/presyncope, cirrhosis, ASA/NSAID use, blood pressure, heart rate, orthostasis, NG lavage, rectal exam, hematocrit, hematocrit drop, platelet count, creatinine, BUN, and INR. Variables used to predict endoscopy using the second approach (described in Section 2.1) are the same variables used to predict resuscitation using the second approach.

Disposition: Variables to predict disposition of the patient included age, hematochezia, hematemesis, melena, duration of symptoms, syncope/presyncope, unstable CAD, COPD, CRF, risk for stress ulcer, cirrhosis, blood pressure, heart rate, orthostasis, NG lavage, rectal exam, hematocrit, drop in hematocrit, platelet count, creatinine, BUN, and INR.

2.6 Results

2.6.1 Comparison of Classification Models Results

Tables 3 through 6 summarize the results for each outcome prediction variable for the evaluation step using the primary approach. Only six of the eight models were utilized to predict source of bleeding, since logistic regression and boosting can only be used for 2 way classification problems in R (source of bleeding response included three outcomes). The results for only the linear kernel for SVM are shown because for classifying the source of bleeding response, the linear kernel had slightly higher values. For the rest of the responses, the linear and radial kernel for SVM gave similar results.

Table 3. Evaluation Step – Model Performance for Source of Bleeding (standard error)

	ACC	SN	SP	PPV	NPV	AUC
ANN	0.917 (0.008)	0.972 (0.005)	0.936 (0.007)	0.968 (0.005)	0.944 (0.007)	0.999
kNN	0.697 (0.013)	0.901 (0.009)	0.287 (0.013)	0.717 (0.013)	0.591 (0.014)	0.658
LDA	0.931 (0.007)	0.965 (0.005)	1.000 (0.000)	1.000 (0.000)	0.935 (0.007)	0.987
RF	0.943 (0.007)	0.980 (0.004)	0.932 (0.007)	0.967 (0.005)	0.959 (0.006)	0.998
SC	0.914 (0.008)	0.965 (0.005)	0.890 (0.009)	0.946 (0.007)	0.927 (0.008)	0.978
SVM	0.930 (0.007)	0.965 (0.005)	0.945 (0.007)	0.973 (0.005)	0.932 (0.007)	0.979

Note: ACC – accuracy; SN – sensitivity; SP – specificity; PPV – positive predictive value; NPV – negative predictive value; AUC – area under ROC curve

Table 4. Evaluation Step – Model Performance for Resuscitation (standard error)

	ACC	SN	SP	PPV	NPV	AUC
ANN	0.921 (0.008)	0.927 (0.008)	0.910 (0.008)	0.946 (0.007)	0.880 (0.009)	0.993
kNN	0.884 (0.009)	0.903 (0.009)	0.852 (0.010)	0.914 (0.008)	0.835 (0.011)	0.890
LDA	0.922 (0.008)	0.904 (0.009)	0.955 (0.006)	0.972 (0.005)	0.852 (0.010)	0.937
Logistic	0.923 (0.008)	0.939 (0.007)	0.895 (0.009)	0.940 (0.007)	0.897 (0.009)	0.985
LogitBoost	0.647 (0.014)	0.916 (0.008)	0.184 (0.011)	0.662 (0.014)	0.481 (0.014)	0.381
RF	0.932 (0.007)	0.937 (0.007)	0.923 (0.008)	0.954 (0.006)	0.894 (0.009)	0.982
SC	0.915 (0.008)	0.929 (0.007)	0.891 (0.009)	0.936 (0.007)	0.879 (0.009)	0.920
SVM	0.941 (0.007)	0.938 (0.007)	0.945 (0.007)	0.968 (0.005)	0.899 (0.009)	0.964

Table 5. Evaluation Step – Model Performance for Endoscopy (standard error)

	ACC	SN	SP	PPV	NPV	AUC
ANN	0.778 (0.012)	0.801 (0.012)	0.733 (0.013)	0.849 (0.010)	0.665 (0.014)	0.913
kNN	0.796 (0.012)	0.876 (0.010)	0.648 (0.014)	0.822 (0.011)	0.737 (0.013)	0.766
LDA	0.833 (0.011)	0.821 (0.011)	0.857 (0.010)	0.914 (0.008)	0.720 (0.013)	0.843
Logistic	0.787 (0.012)	0.871 (0.010)	0.831 (0.014)	0.815 (0.011)	0.726 (0.013)	0.853
LogitBoost	0.627 (0.014)	0.891 (0.009)	0.138 (0.010)	0.658 (0.014)	0.403 (0.014)	0.404
RF	0.790 (0.012)	0.854 (0.010)	0.671 (0.014)	0.828 (0.011)	0.712 (0.013)	0.871
SC	0.811 (0.011)	0.838 (0.011)	0.760 (0.012)	0.866 (0.010)	0.717 (0.013)	0.801
SVM	0.803 (0.011)	0.859 (0.010)	0.700 (0.013)	0.842 (0.011)	0.728 (0.013)	0.820

Table 6. Evaluation Step – Model Performance for Disposition (standard error)

	ACC	SN	SP	PPV	NPV	AUC
ANN	0.850 (0.010)	0.829 (0.011)	0.889 (0.009)	0.928 (0.007)	0.752 (0.013)	0.972
kNN	0.876 (0.010)	0.923 (0.008)	0.798 (0.012)	0.886 (0.009)	0.858 (0.010)	0.881
LDA	0.897 (0.009)	0.891 (0.009)	0.909 (0.008)	0.943 (0.007)	0.830 (0.011)	0.901
LogitBoost	0.584 (0.014)	0.819 (0.011)	0.184 (0.011)	0.629 (0.014)	0.377 (0.014)	0.324
RF	0.883 (0.009)	0.907 (0.008)	0.843 (0.011)	0.908 (0.008)	0.841 (0.011)	0.967
SC	0.897 (0.009)	0.916 (0.008)	0.866 (0.010)	0.921 (0.008)	0.858 (0.010)	0.891
SVM	0.887 (0.009)	0.929 (0.007)	0.816 (0.011)	0.896 (0.009)	0.872 (0.010)	0.922

Figures 2 through 5 depict the accuracies obtained for each individual model and each response variable. The points shown are the accuracies with whiskers extending which represent the standard errors. Table 7 shows summary results from doing the McNemar's

test. The model that had the smallest statistically significant difference to the highest accuracy model is shown. The model that had the smallest non-statistically significant difference to the highest accuracy model is also shown. Average computing time to run 10 repetitions of 10-fold CV for all models for one response was 30-40 minutes.

Figure 2. Accuracies for Source of Bleeding Response (evaluation step)

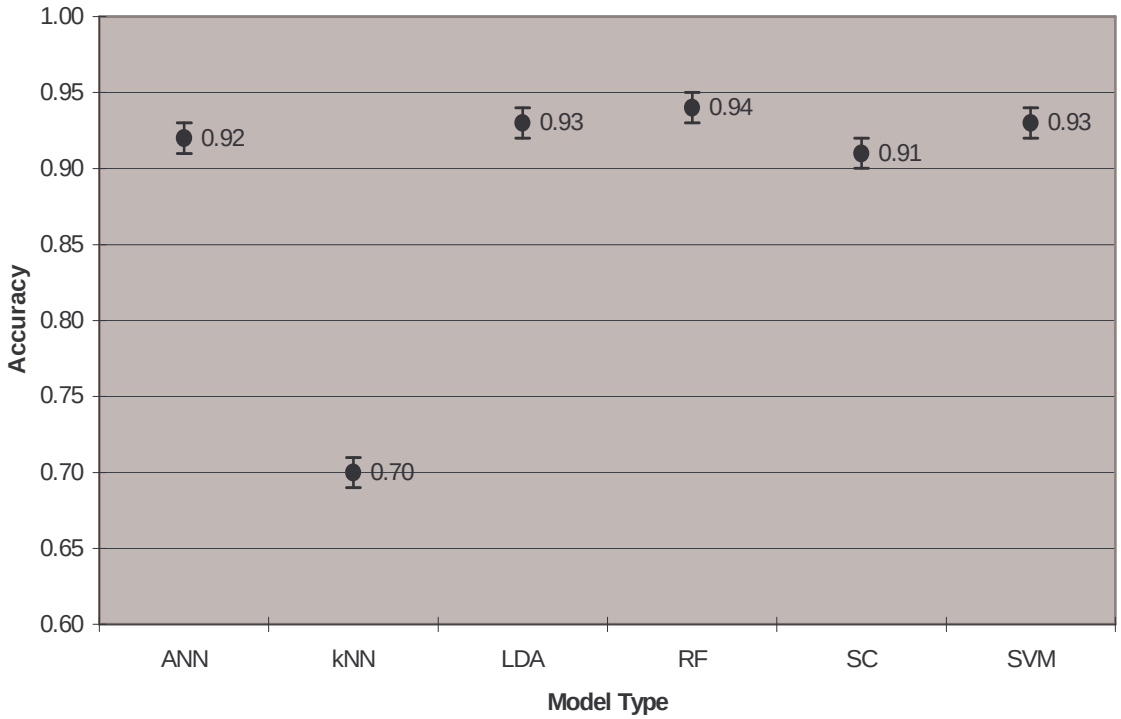


Figure 3. Accuracies for Resuscitation Response (evaluation step)

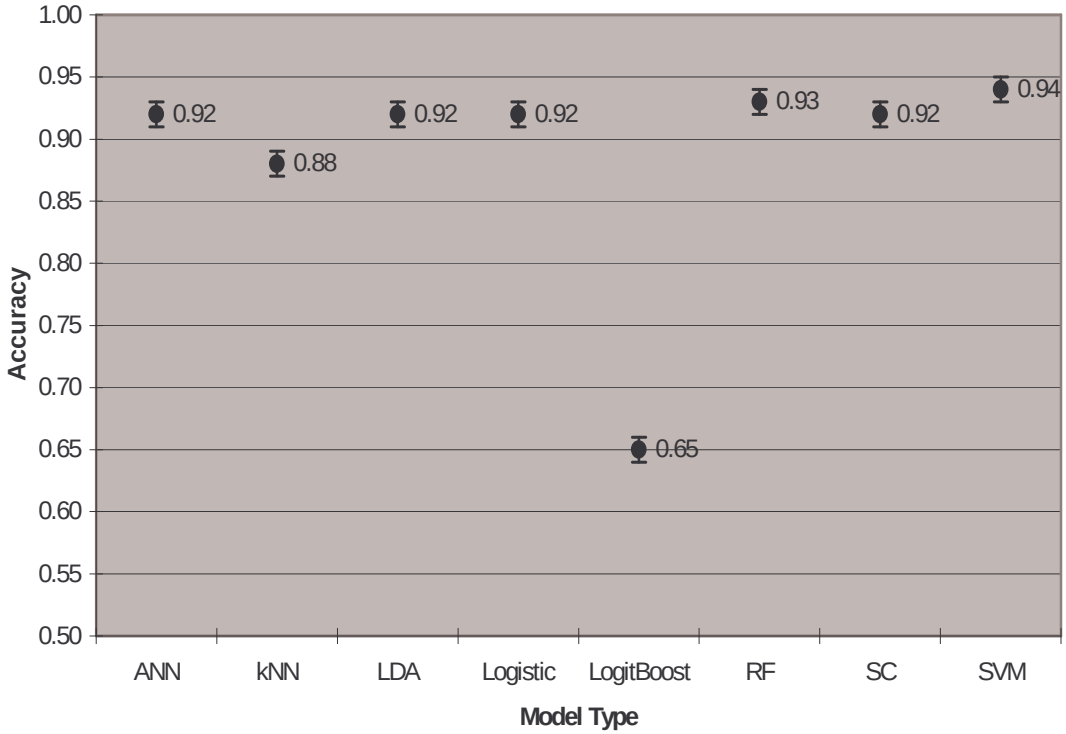


Figure 4. Accuracies for Endoscopy Response (evaluation step)

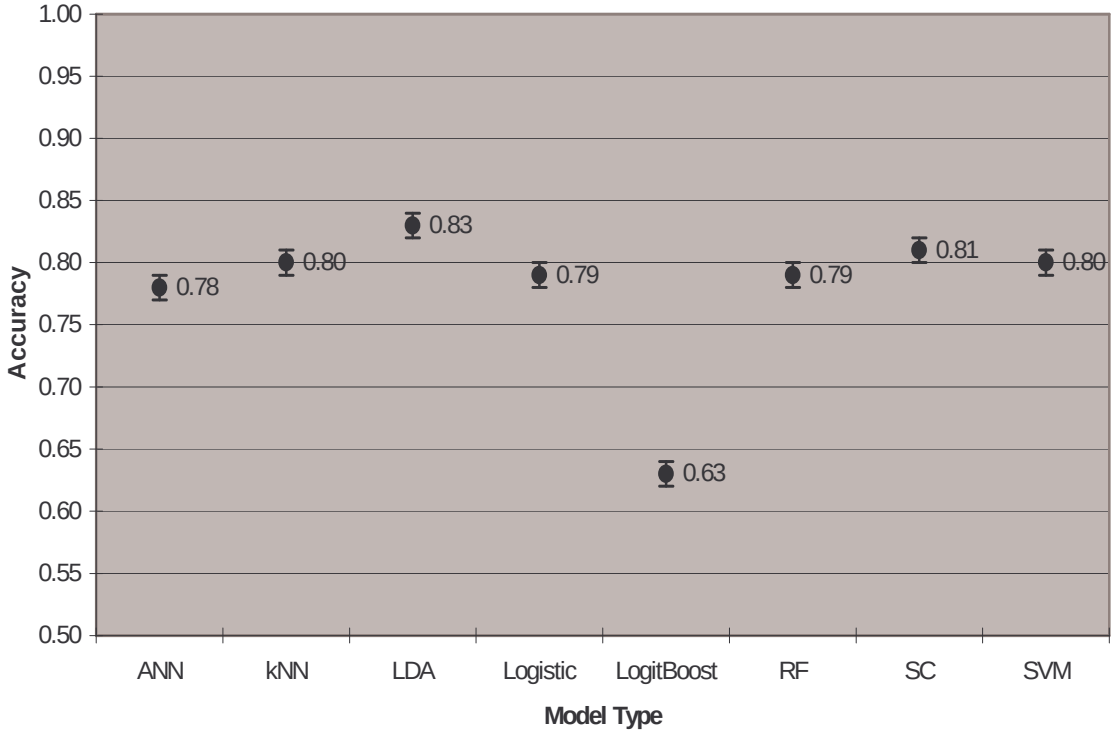


Figure 5. Accuracies for Disposition Response (evaluation step)

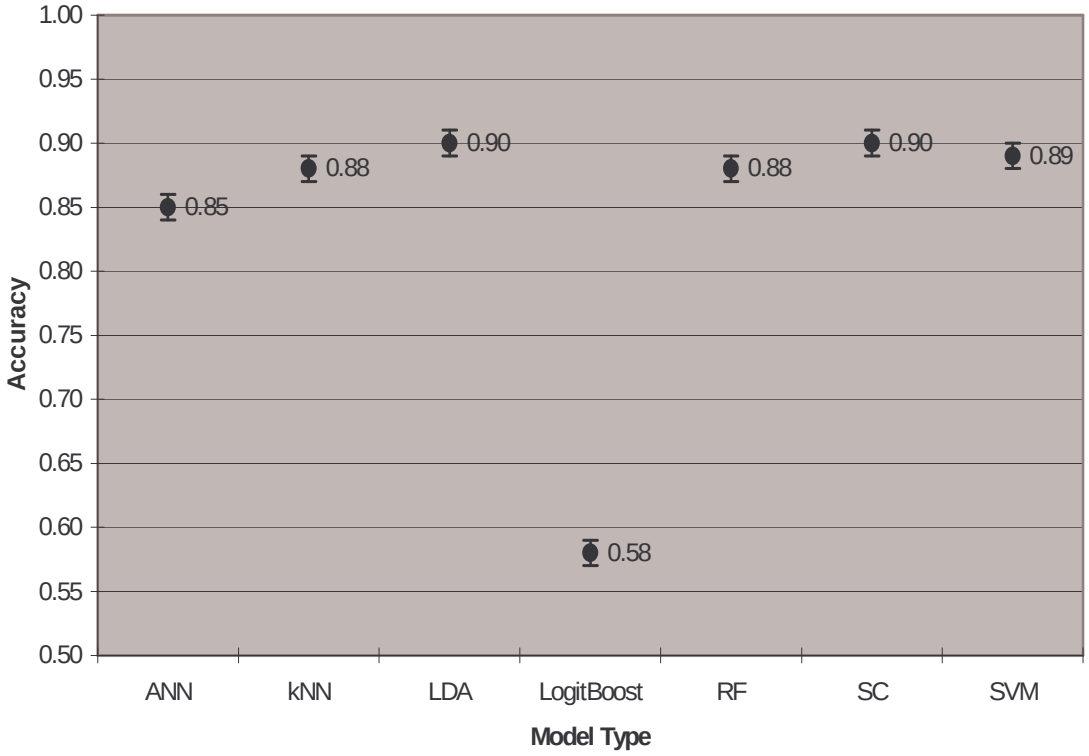


Table 7. Summary of McNemar's Test Results

	Model (accuracy)	Model (accuracy)	p-value
Source of bleeding			
Least significantly different	RF (0.943)	LDA (0.931)	0.0004
Least not significantly different[a]	–	–	–
Resuscitation			
Least significantly different	SVM (0.941)	RF (0.932)	0.0027
Least not significantly different[a]	–	–	–
Endoscopy			
Least significantly different	LDA (0.833)	SC (0.811)	<0.0001
Least not significantly different[a]	–	–	–
Disposition			
Least significantly different	LDA (0.897) or SC (0.897)	SVM (0.887)	0.0016
Least not significantly different	LDA (0.897)	SC (0.897)	1.0000

[a] All models were significantly different from highest accuracy model

Overall, accuracies obtained using SVM, RF, and LDA were generally higher than the accuracies for the other models. These models predicted the source of bleeding, need for resuscitation, and disposition correctly 88%-94% of the time. The need for endoscopy was correctly predicted about 80% of the time using kNN, SVM, SC, and LDA; RF's accuracy was just below 80% (79%). In terms of accuracy, logistic regression did well for predicting resuscitation and endoscopy. However it did not do well for disposition. This is because the algorithm for obtaining the model rarely converged. That is, it was unstable (results are not reported for the disposition response). The performance of boosting was worst among all models. kNN performed the worst for predicting source of bleeding. The linear discriminant analysis model appeared to demonstrate good overall performance with regards to all statistics (accuracy, sensitivity, specificity, PPV, NPV). Boosting, on the other hand, revealed an imbalance between sensitivity and specificity. ROC curves were constructed (Figures 6 through 9 and Tables 3 through 6), and overall RF and ANN have the highest AUC (area under the ROC curve), followed by SVM and LDA. For predicting resuscitation and endoscopy, the logistic model had excellent AUC. In the validation step, RF consistently performed well compared to the other models (Table 8).

Figure 6. ROC Curves for Predicting Source of Bleeding (evaluation step)

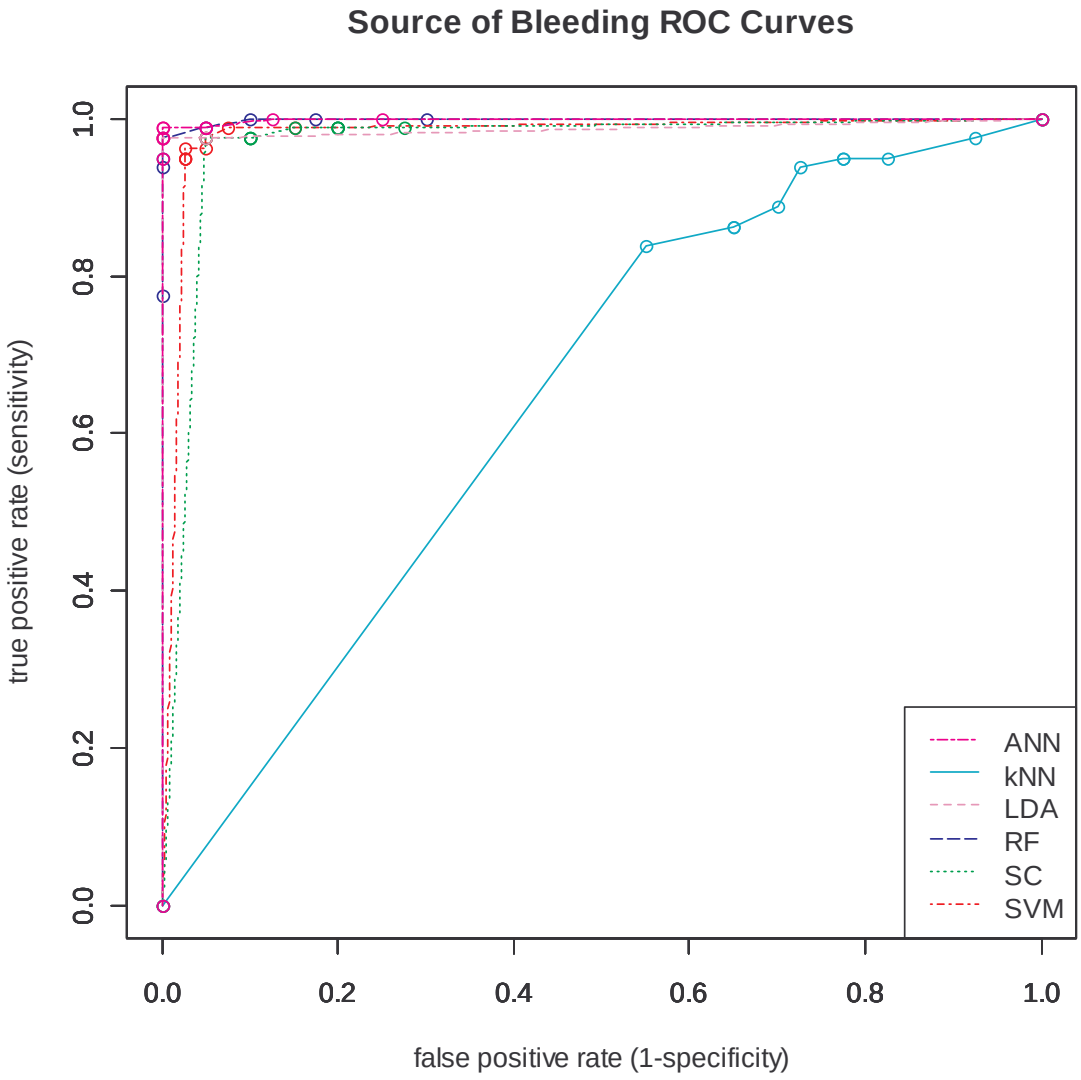


Figure 7. ROC Curves for Predicting Resuscitation (evaluation step)

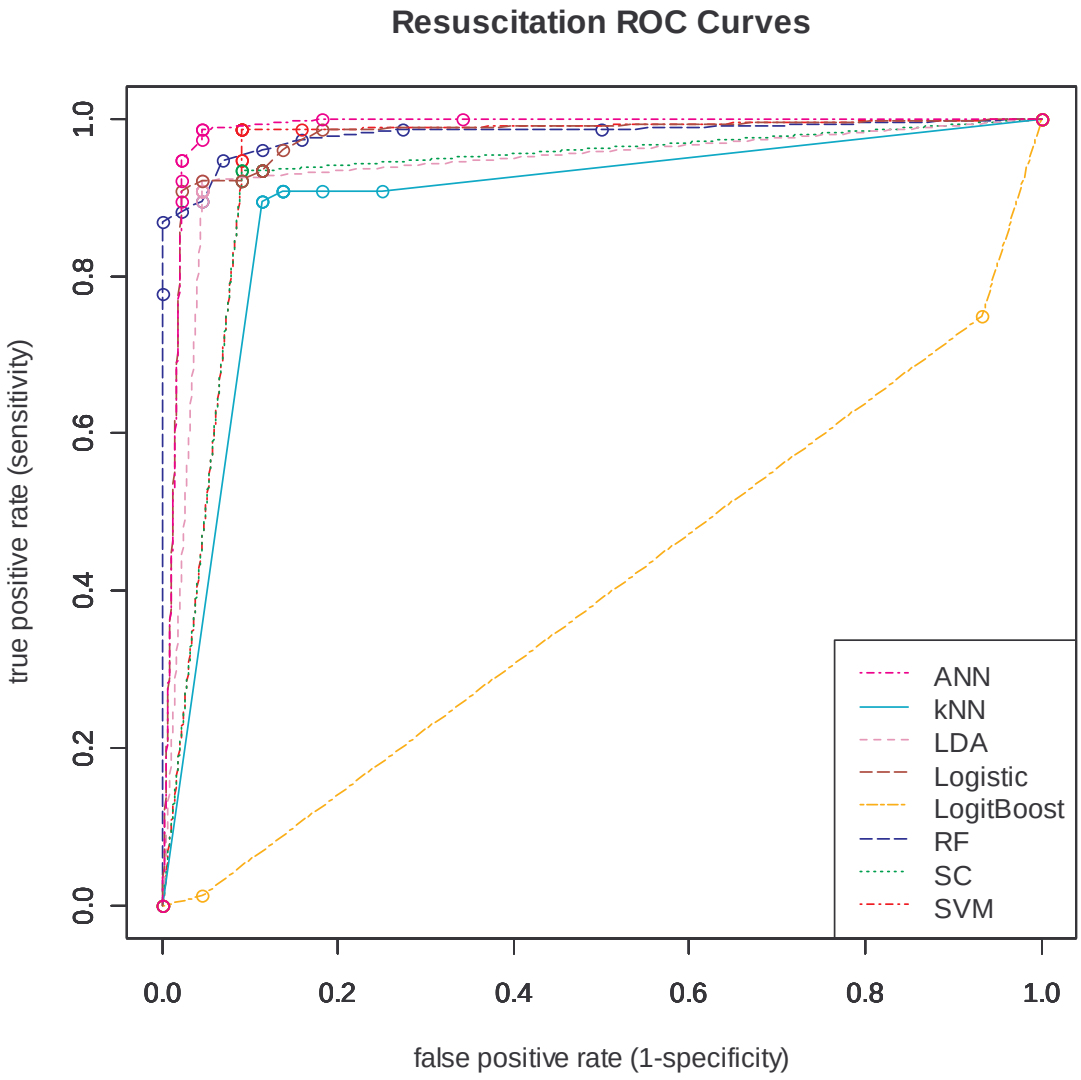


Figure 8. ROC Curves for Predicting Endoscopy (evaluation step)

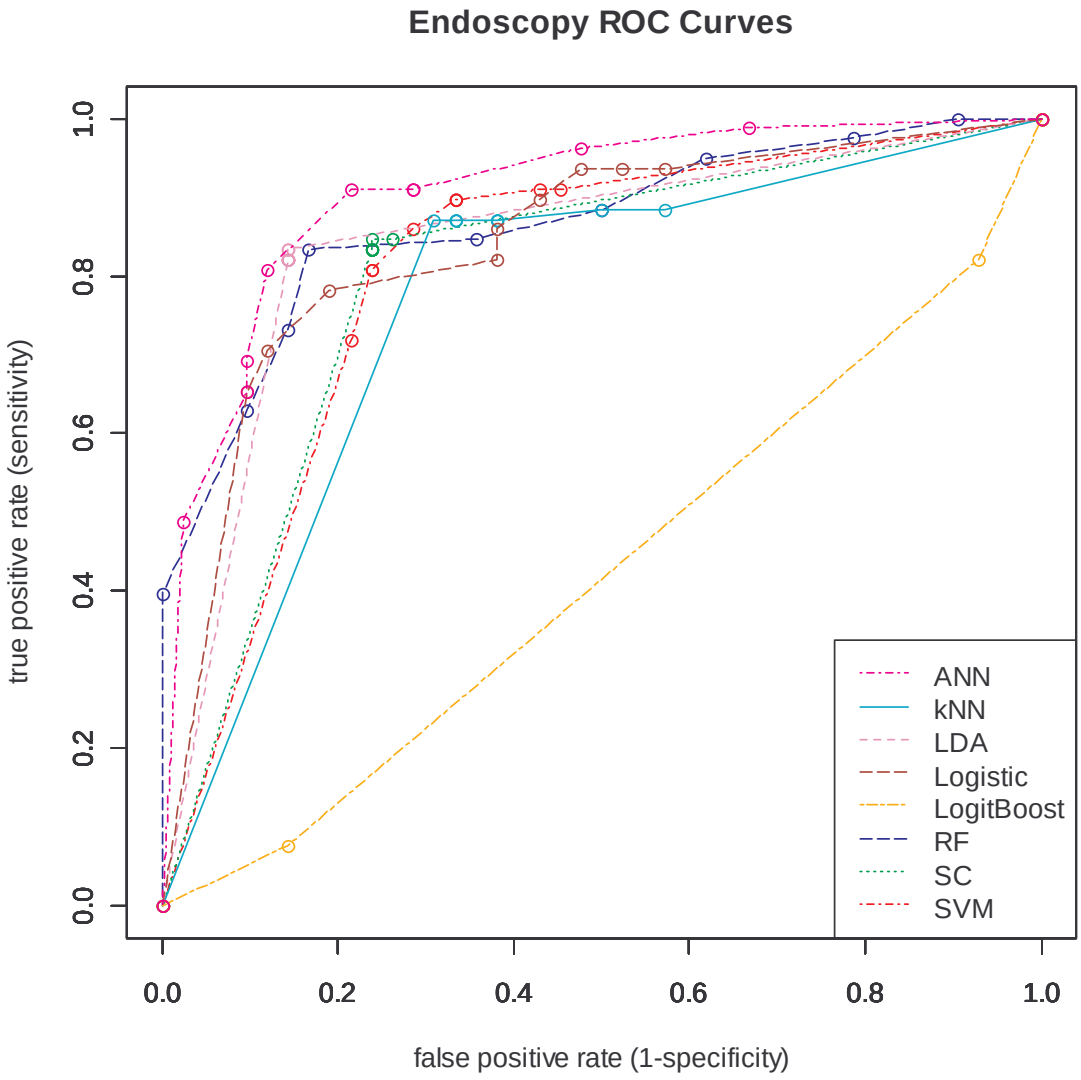


Figure 9. ROC Curves for Predicting Disposition (evaluation step)

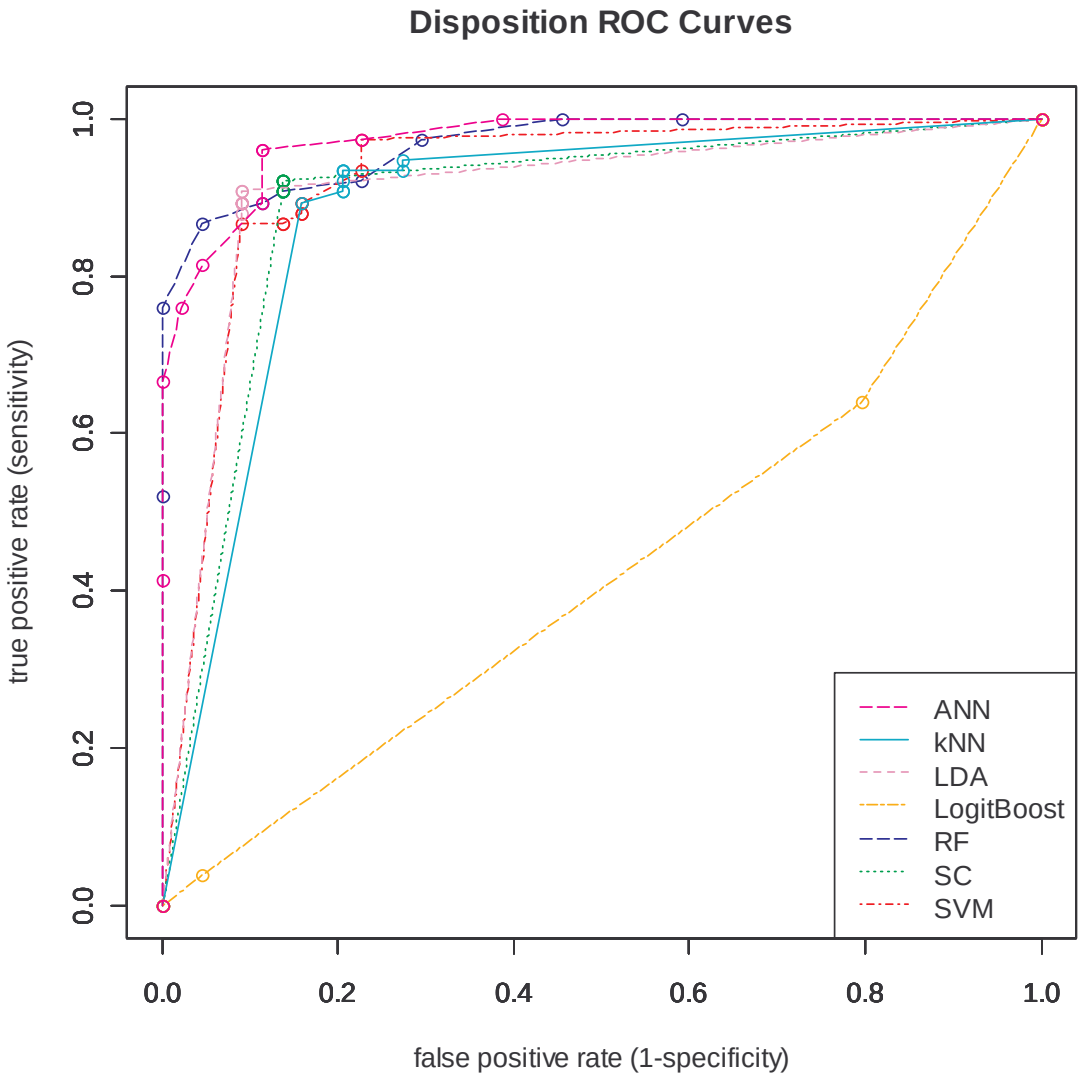


Table 8. Validation step – Predictive Accuracies Using a 70 Patient Database

	Source of bleeding	Resuscitation	Endoscopy	Disposition
ANN	0.884	0.821	0.638	0.754
kNN	0.783	0.821	0.667	0.783
LDA	0.768	0.821	0.681	0.797
Logistic	N/A	0.791	0.710	N/A
LogitBoost	N/A	0.567	0.551	0.362
RF	0.928	0.851	0.753	0.797
SC	0.855	0.866	0.696	0.768
SVM	0.826	0.791	0.681	0.783

Overall, data from the evaluation and validation steps suggest that the RF model consistently performs the best. For the secondary approach to predict resuscitation and endoscopy, there were no significant improvements seen for the majority of the results (Tables 9 and 10).

Table 9. Evaluation Step – Predicting Resuscitation Using Secondary Approach (standard error)

	ACC	SN	SP	PPV	NPV
ANN	0.915 (0.008)	0.911 (0.008)	0.919 (0.008)	0.951 (0.006)	0.857 (0.010)
kNN	0.883 (0.009)	0.896 (0.009)	0.861 (0.010)	0.919 (0.008)	0.826 (0.011)
LDA	0.922 (0.008)	0.909 (0.008)	0.945 (0.007)	0.967 (0.005)	0.856 (0.010)
Logistic	0.918 (0.008)	0.932 (0.007)	0.893 (0.009)	0.939 (0.007)	0.884 (0.009)
LogitBoost	0.631 (0.014)	0.883 (0.009)	0.191 (0.011)	0.660 (0.014)	0.442 (0.014)
RF	0.934 (0.007)	0.943 (0.007)	0.918 (0.008)	0.953 (0.006)	0.902 (0.009)
SC	0.909 (0.008)	0.909 (0.008)	0.909 (0.008)	0.946 (0.006)	0.851 (0.010)
SVM	0.935 (0.007)	0.952 (0.006)	0.905 (0.008)	0.946 (0.006)	0.915 (0.008)

Table 10. Evaluation Step – Predicting Endoscopy Using Secondary Approach (standard error)

	ACC	SN	SP	PPV	NPV
ANN	0.783 (0.012)	0.781 (0.012)	0.758 (0.012)	0.854 (0.010)	0.656 (0.014)
kNN	0.794 (0.012)	0.841 (0.011)	0.709 (0.013)	0.840 (0.011)	0.711 (0.013)
LDA	0.818 (0.011)	0.821 (0.011)	0.814 (0.011)	0.889 (0.009)	0.714 (0.013)
Logistic	0.792 (0.012)	0.840 (0.011)	0.705 (0.013)	0.838 (0.011)	0.708 (0.013)
LogitBoost	0.631 (0.014)	0.913 (0.008)	0.121 (0.009)	0.653 (0.014)	0.452 (0.014)
RF	0.783 (0.012)	0.838 (0.011)	0.684 (0.013)	0.828 (0.011)	0.700 (0.013)
SC	0.809 (0.011)	0.832 (0.011)	0.767 (0.012)	0.866 (0.010)	0.716 (0.013)
SVM	0.823 (0.011)	0.817 (0.011)	0.835 (0.011)	0.900 (0.009)	0.715 (0.013)

2.6.2 Variable Importance Results

The importance of each variable in predicting outcomes when using random forest and ANN is shown in Table 11. Variables were common to both RF and ANN about half the time. For predicting source, explanatory variables hematemesis through HR (heart rate) were considered significant variables. The remaining variables had mixed ratings except for the last variable, ASA/NSAID, which did not seem to have a great influence on the performance of a model. For predicting resuscitation, both models utilized variables syncope, orthostasis, hematocrit, hematocrit drop, blood pressure, heart rate, hematemesis and melena as important predictor variables. The rest of the variables have mixed importance except for duration and unstable CAD, which do not seem to have a large impact on how well the model does. Looking at the remaining two response variables, the results show there are also certain variables that are considered to be important for both models while the rest have varied importance. These results (Section 2.6) have been published in Chu et al (31).

Table 11. Variable Importance Using RF and ANN

SOURCE			**RESUSCITATION**		
	RF	**ANN**		**RF**	**ANN**
Hematemesis	0.1388	10	Syncope	0.1097	10
NG Lavage	0.0692	10	Orthostasis	0.0782	10
Hematochezia	0.0611	10	Hct. Drop	0.0495	10
BUN	0.0484	10	DBP	0.0276	10
Rectal	0.0374	10	Hct.	0.0232	10
Melena	0.0221	9	HR	0.0139	10
Orthostasis	0.0116	10	SBP	0.0114	10
Hx. of GIB	0.0088	8	Hematemesis	0.0052	10
HR	0.0066	10	Melena	0.0032	10
Cr.	0.0055	7	Cr.	0.0019	1
SBP	0.0051	4	NG Lavage	0.0019	9
DBP	0.0043	9	BUN	0.0018	6
Syncope	0.0032	9	Hematochezia	0.0015	10
INR	0.0031	5	Duration	0.0011	6
Plt.	0.0010	6	INR	0.0009	6
Risk for Stress Ulcer	0.0008	10	Rectal	0.0006	7
Cirrhosis	0.0004	10	Unstable CAD	-0.0016	5
ASA/NSAID	-0.0010	1			

Table 11. (cont'd) Variable Importance Using RF and ANN

ENDOSCOPY			DISPOSITION		
	RF	**ANN**		**RF**	**ANN**
Syncope	0.0507	10	Orthostasis	0.0585	10
Orthostasis	0.0259	10	HR	0.0431	10
Hct.	0.0223	10	Hct.	0.0340	10
HR	0.0213	9	SBP	0.0287	10
Hct. Drop	0.0188	10	Syncope	0.0271	10
DBP	0.0178	9	DBP	0.0217	10
Hematemesis	0.0133	10	Hct. Drop	0.0189	10
Rectal	0.0054	9	Rectal	0.0094	10
BUN	0.0051	5	Age	0.0070	10
SBP	0.0046	9	NG_Lavage	0.0055	8
INR	0.0043	2	BUN	0.0054	8
Cirrhosis	0.0039	8	INR	0.0051	4
Melena	0.0022	9	Risk for Stress Ulcer	0.0037	8
Plt.	0.0020	3	Plt.	0.0036	6
Risk for Stress Ulcer	0.0015	9	Melena	0.0026	9
Duration	0.0008	8	Hematochezia	0.0022	6
Cr.	0.0003	2	Hematemesis	0.0010	7
ASA/NSAID	0.00008	4	Cirrhosis	0.0007	8
Hematochezia	0.00001	9	COPD	0.0006	8
NG Lavage	-0.0004	8	CRF	0.00007	4
			Duration	0.00001	4
			Cr.	-0.00006	3
			Unstable CAD	-0.0002	5

2.7 Discussion of Findings

Although it would be best for patients with acute GIB to be cared for by gastroenterologists (32), it is too costly and logistically impossible that a gastroenterologist evaluate every patient. It is also impractical to justify intensive resuscitation and urgent endoscopy for every patient with acute GIB due to limited healthcare resources. "Expert systems" can immensely help non-gastroenterologists triage patients who may benefit most from urgent resuscitation and endoscopy. These classification models or predictive models have been successfully utilized to optimize treatment and predict clinical outcomes in a variety of other conditions (33,34,35,36),

such as computerized interpretation of the electrocardiogram (37) and to help streamline and optimize care of patients with acute myocardial infarction (38), especially in a busy practice or in the emergency room (39).

Our objective was to develop a model to provide diagnostic and specific treatment recommendations for patients presenting with acute GIB. The recommendations were designed to be in agreement with current evidence based guidelines for management of acute GIB. The models were able to provide patient specific recommendations with accuracies exceeding 70-80%. In the study, RF, SVM and LDA, all performed well in classification of the four response variables, in agreement with previous studies. RF in particular performed exceptionally well having both high accuracies and high AUC. RF and SVM are designed for high-dimensional data with a large feature space (large number of predictor variables) compared to the sample size and are likely to outperform other methods for high-dimensional data, which are unlike the current GIB data set (40). Logistic regression is a widely used standard regression model for binary data, and it can be expanded to data with more than two classes. However, it often shows computational instability, such as failure to converge or the predicted value being extremely close to 1 or 0 due to the nature of the model. Our results support the conclusion by Ahn et al. that boosting strategies in general provide poor accuracies (40). Furthermore, given the complexity, it is cumbersome and unwieldy compared to other methods. Although not relevant to our problem, LDA and kNN require a variable pre-selection for an optimal performance unlike RF or SVM for high-dimensional data. A statistical variable selection is often dependent on the criteria and is computer intensive.

With regards to the analysis of the importance of a variable, we show that both the RF and ANN models considered half of the variables to be important and the remaining half of varied importance. This appears to be consistent with prior knowledge in regards to importance of variables identified to predict source and severity of acute GIB. Every pre-selected variable was important for one model or the other and therefore consistent with their identification in prior multivariate analyses. Different models assigned varied importance to the variables due to the unique methods for evaluating variable importance. In RF, for every tree grown in the forest, test samples are used to count the number of votes cast for the correct class. RF randomly permutes the values of a selected variable in the test set and put these test cases down the tree a second time. It finds the number of votes for the correct class in the data with this permuted variable. It subtracts this number from the number of votes for the correct class in the original data without permutation. The average of this number of all trees in the forest is the raw importance score for the variable. In the Statistica software, a combination of different methods is used to quantify variable importance for ANN. These methods include searching algorithms, regularization (to avoid overfitting the model), connection weight approach, and sensitivity based approaches.

Since the output is based on training examples, bias can potentially be introduced through training examples. We have the ability to influence the recommendations by modulating outcomes associated with each training example. The model is limited by the

extent of the examples that is utilized to train the model. Despite these flaws, the availability of such systems may facilitate and standardize the care of patients presenting with acute GIB. Given that computer based tools are more likely to work if integrated with clinical care, validation of such a model could potentially facilitate care of patients with acute GIB. It is also possible to continue to train the model prospectively to adapt to changing guidelines and varied clinical scenarios, allowing these predictive models to be portable to a broad range of locales. Steps toward integrating a model such as random forest into hospital computer systems will be discussed in Section 8.

3. Comparison of Classification Models Based on Variable Importance Rankings

In Section 2, variables used were those thought to be important by literature research (mostly found by logistic regression) and those that are important to gastroenterologists. In this section, variable rankings were obtained by combining four different statistical methods. By using a subset of these ranked variables, the models were again evaluated. The methods used to obtain variable rankings were: RF (using the variable importance option), ANN (using the variable importance option), Support Vector Machine-Recursive Feature Elimination (SVM-RFE), and BW ratio. The same evaluation step as described in Section 2 was performed and accuracies as well as areas under the ROC curves were compared to the original analysis of models.

3.1 Description of the Different Variable Importance Ranking Methods

There are several ways to rank variables. Among them, are variable rankings done using classification models such as random forest, support vector machine, and artificial neural networks. Another way is by ranking variables according to their BW ratios. By computing the between-group sum of squares (BSS) and dividing it by the within-group sum of squares (WSS), we obtain the BW ratio. This is done for each variable, where the groups are the different classes of the response variable. The higher the BW ratio, the more important the variable is (41). For a particular variable j, the BW ratio is as follows:

$$\text{BW ratio} = \frac{BSS(j)}{WSS(j)} = \frac{\sum_i \sum_k I\left(y_i = k\right)\left(\overline{x}_{kj} - \overline{x}_{\bullet j}\right)^2}{\sum_i \sum_k I\left(y_i = k\right)\left(x_{ij} - \overline{x}_{kj}\right)^2}$$

where $\overline{x}_{\bullet j}$ represents the average for variable j across all cases, $\overline{x}_{kj}$ represents the average for variable j for a particular class k, and x_{ij} represents a particular case value for variable j. To obtain variable rankings using random forest, consider at first one single tree in the forest. As mentioned in the Discussion of Section 2, test cases are put down the tree and the predicted class is recorded. This is done for all the trees in the forest and the number of times the correct class is predicted is recorded (#VOTES_ORIGINAL). Then the data for one variable, X, for the test cases are randomly scrambled up, and the test cases are put down each tree again. We record the number of times the correct class is predicted again (#VOTES_PERMUTED). The number of times the correct class is predicted using the scrambled data is subtracted from the number of times the correct class is predicted

using the unscrambled data (#VOTES_ORIGINAL - #VOTES_PERMUTED). Taking the average of this result across all the trees in the forest is known as the raw importance score for variable X. This process is repeated for all variables. The higher the raw importance score is for a variable, the more important the variable (15). The raw importance score is also known as the mean decrease in accuracy, which is what is outputted in R.

The Intelligent Problem Solver in the Neural Networks package in Statistica selects the variables that are thought to be important in a model. The Intelligent Problem Solver uses a combination of several different techniques including searching algorithms, regularization (to avoid overfitting the model), and sensitivity based approaches. Support Vector Machine-Recursive Feature Elimination (SVM-RFE) removes features one by one in a backward elimination fashion (42). An SVM is trained with all the features and the features are assigned weights. A ranking score is assigned by squaring the weights and the feature with the smallest ranking score is eliminated. This is the feature that is least important. Another SVM is trained with the remaining features and a second feature is eliminated. This is repeated until only one feature remains (this would be considered the most important feature).

3.2 Methods for Choosing the Variables

The 4 different approaches of variable rankings were considered, with BW ratio and RF being done in R. Statistica was used to run ANN while Matlab (version R2007a) was used to run SVM-RFE. The number of times the variable was said to be important over 10 runs was recorded. All the variables were ranked in decreasing order of importance. The results from all 4 methods were combined by adding the standardized ranking values of each variable.

Since the artificial neural network method only indicates whether a variable is important or not in the model, for each repetition, the network was run an additional 10 different times in order to obtain some variable ranking. For ANN, the rankings were assigned by counting the number of times the variable was said to be important to the model (so the highest number is 10, the lowest is 0). Also since the history of GIB explanatory variable (Hx_of_GIB) wasn't broken down into indicator variables explicitly for ANN, the same number was assigned for each indicator variable of Hx_of_GIB so that we could combine the results with the other methods. For SVM-RFE, just a list was given of the decreasing importance of the variables. So the variables were renumbered, starting with 1 as the least important, 2 as the second least important, and 29 as the most important. The variable ranking values were standardized by subtracting each respective mean and dividing by each respective standard deviation. These standardized values were then added together to obtain an overall variable importance ranking. When a

higher drop-off in the values was seen, these variables were discarded from being used in the model.

3.3 Comparison Between Previous Results and Current Results

The same procedure was done for evaluating the models as described in Section 2.1 (evaluation step only). In particular, accuracies and area under the curves were compared for each response for each model using a significance level of α=0.05. The highest accuracies were compared to each other for each response using the McNemar's test. The same was done for the highest area under ROC curves. For situations where multiple models were being compared, Bonferroni's correction was used to control the error rate, with an overall α=0.05. The statistical program, R (version 2.4.1), was used to run all the models, except for ANN, which was run in Statistica (version 7.1).

3.4 Individual Model Parameters

All the same parameters were used as in the original analysis, except new values for the threshold were found for the shrunken centroid model and new values for *k* were found for the *k*-nearest neighbor model. Also a new value for the parameter mfinal for boosting was found. Using 10-fold cross-validation, the best threshold value was 0.4. For kNN, *k*=3 was used for the source of bleeding response, *k*=13 was used for the resuscitation response, and *k*=9 was used for the remaining two responses. The parameter mfinal in the boosting model was set to 15 for the resuscitation and endoscopy responses and 10 for the disposition response.

3.5 Results and Discussion/Conclusion from Evaluating Models Using Variable Importance Rankings to Select the Variables

Tables 12-15 show the combined variable importance rankings for each response.

Table 12. Variable Importance Rankings for Source of Bleeding

Variable	Rank	Variable	Rank
Hematemesis	9.527	DBP	-0.321
NG Lavage	4.230	CRF	-0.390
Hematochezia	2.998	Cirrhosis	-0.533
BUN	1.956	Hx. of GIB – upper	-0.611
Melena	1.688	Cr.	-1.129
Rectal	1.271	Syncope/Presyncope	-1.285
Hct. Drop	0.918	Plt.	-1.430
Hx. of GIB – mid	0.759	COPD	-1.468
Age	0.436	INR	-1.805
Orthostasis	0.373	Unstable CAD	-2.417
Hct.	-0.034	SBP	-2.565
Hx. of GIB – lower	-0.079	Sex	-2.857
Risk for Stress Ulcer	-0.161	PPI	-2.870
HR	-0.173	ASA/NSAID	-3.722
Duration	-0.307		

Table 13. Variable Importance Rankings for Resuscitation

Variable	Rank	Variable	Rank
Syncope/Presyncope	6.237	Hx. of GIB – lower	-0.343
Orthostasis	6.206	Hx. of GIB – upper	-0.348
Hct. Drop	3.106	Hx. of GIB – mid	-0.382
HR	2.599	NG Lavage	-1.029
Hct.	2.248	Cirrhosis	-1.427
DBP	2.107	Hematochezia	-1.478
Age	1.353	Unstable CAD	-1.464
BUN	1.341	Cr.	-1.549
Risk for Stress Ulcer	0.580	INR	-1.839
Hematemesis	0.435	CRF	-2.628
Plt.	0.430	COPD	-2.933
Rectal	0.291	Sex	-3.439
SBP	0.089	PPI	-3.660
Duration	0.077	ASA/NSAID	-4.471
Melena	-0.112		

Table 14. Variable Importance Rankings for Endoscopy

Variable	Rank	Variable	Rank
Syncope/Presyncope	7.599	Hx. of GIB – lower	-0.578
HR	4.609	Hx. of GIB – mid	-0.578
Orthostasis	4.359	Cirrhosis	-0.761
Hct.	4.240	Melena	-0.954
Hct. Drop	3.551	INR	-1.372
DBP	2.485	BUN	-1.432
SBP	1.588	PPI	-1.686
Hematemesis	1.224	NG Lavage	-1.739
Age	1.042	Hematochezia	-2.494
Rectal	0.527	COPD	-2.735
Duration	0.261	ASA/NSAID	-3.053
Unstable CAD	-0.078	Sex	-3.922
Hx. of GIB – upper	-0.446	Cr	-4.180
Risk for Stress Ulcer	-0.489	CRF	-4.490
Plt.	-0.499		

Table 15. Variable Importance Rankings for Disposition

Variable	Rank	Variable	Rank
Orthostasis	6.774	Unstable CAD	-0.628
HR	5.939	NG Lavage	-0.751
SBP	4.780	Cirrhosis	-1.474
Hct.	4.765	Hx. of GIB – lower	-1.928
Syncope/Presyncope	3.576	Hx. of GIB – upper	-2.016
DBP	2.608	Hx. of GIB – mid	-2.066
Hct. Drop	2.554	Hematochezia	-2.066
Age	1.753	Duration	-2.159
Rectal	0.636	COPD	-2.271
BUN	0.568	ASA/NSAID	-2.673
Risk for Stress Ulcer	0.151	Sex	-2.916
Plt.	0.099	CRF	-3.481
Hematemesis	-0.218	Cr.	-3.790
Melena	-0.413	PPI	-4.746
INR	-0.622		

Tables 16-19 and Figures 10-17 show the statistics calculated and the ROC curves respectively for evaluating the models using the variable importance rankings to select the variables to be in the models. Table 20 shows a summary of results from the McNemar's test.

Table 16. Source of Bleeding Results (based on variable importance rankings)

	ACC	SN	SP	PPV	NPV	AUC
ANN	0.912 (0.008)	0.969 (0.005)	0.940 (0.007)	0.969 (0.005)	0.941 (0.007)	0.999
kNN	0.762 (0.012)	0.871 (0.010)	0.602 (0.014)	0.811 (0.011)	0.706 (0.013)	0.789
LDA	0.923 (0.008)	0.975 (0.004)	1.000 (0.000)	1.000 (0.000)	0.953 (0.006)	0.988
RF	0.947 (0.006)	0.986 (0.003)	0.951 (0.006)	0.975 (0.004)	0.973 (0.005)	0.998
SC	0.923 (0.008)	0.976 (0.004)	0.907 (0.008)	0.954 (0.006)	0.952 (0.006)	0.968
SVM (linear)	0.922 (0.008)	0.976 (0.004)	0.932 (0.007)	0.965 (0.005)	0.953 (0.006)	0.974

Table 17. Resuscitation Results (based on variable importance rankings)

	ACC	SN	SP	PPV	NPV	AUC
ANN	0.915 (0.008)	0.906 (0.008)	0.929 (0.007)	0.956 (0.006)	0.858 (0.010)	0.990
kNN	0.874 (0.010)	0.926 (0.008)	0.784 (0.012)	0.879 (0.009)	0.864 (0.010)	0.864
LDA	0.921 (0.008)	0.899 (0.009)	0.958 (0.006)	0.973 (0.005)	0.848 (0.010)	0.938
Logistic	0.928 (0.007)	0.945 (0.007)	0.900 (0.009)	0.941 (0.007)	0.906 (0.008)	0.975
LogitBoost	0.600 (0.014)	0.822 (0.011)	0.224 (0.012)	0.641 (0.014)	0.423 (0.014)	0.496
RF	0.915 (0.008)	0.920 (0.008)	0.907 (0.008)	0.943 (0.007)	0.870 (0.010)	0.982
SC	0.911 (0.008)	0.924 (0.008)	0.889 (0.009)	0.934 (0.007)	0.873 (0.010)	0.921
SVM (linear)	0.964 (0.005)	0.974 (0.005)	0.949 (0.006)	0.970 (0.005)	0.955 (0.006)	0.980

Table 18. Endoscopy Results (based on variable importance rankings)

	ACC	SN	SP	PPV	NPV	AUC
ANN	0.800 (0.012)	0.808 (0.011)	0.785 (0.012)	0.871 (0.010)	0.697 (0.013)	0.933
kNN	0.761 (0.012)	0.859 (0.010)	0.584 (0.014)	0.789 (0.012)	0.695 (0.013)	0.761
LDA	0.823 (0.011)	0.803 (0.011)	0.860 (0.010)	0.913 (0.008)	0.706 (0.013)	0.848
Logistic	0.774 (0.012)	0.844 (0.010)	0.647 (0.014)	0.812 (0.011)	0.696 (0.013)	0.836
LogitBoost	0.614 (0.014)	0.856 (0.010)	0.174 (0.011)	0.653 (0.014)	0.393 (0.014)	0.528
RF	0.798 (0.012)	0.849 (0.010)	0.705 (0.013)	0.839 (0.011)	0.720 (0.013)	0.866
SC	0.820 (0.011)	0.846 (0.010)	0.772 (0.012)	0.871 (0.010)	0.734 (0.013)	0.815
SVM (linear)	0.765 (0.012)	0.823 (0.011)	0.660 (0.014)	0.815 (0.011)	0.673 (0.013)	0.790

Table 19. Disposition Results (based on variable importance rankings)

	ACC	SN	SP	PPV	NPV	AUC
ANN	0.840 (0.011)	0.806 (0.012)	0.896 (0.010)	0.929 (0.008)	0.740 (0.013)	0.976
kNN	0.875 (0.010)	0.936 (0.007)	0.770 (0.012)	0.874 (0.010)	0.876 (0.010)	0.907
LDA	0.896 (0.009)	0.888 (0.009)	0.909 (0.008)	0.943 (0.007)	0.827 (0.011)	0.901
LogitBoost	0.555 (0.014)	0.800 (0.012)	0.139 (0.010)	0.612 (0.014)	0.300 (0.013)	0.401
RF	0.884 (0.009)	0.913 (0.008)	0.834 (0.011)	0.904 (0.009)	0.849 (0.010)	0.968
SC	0.900 (0.009)	0.917 (0.008)	0.870 (0.010)	0.924 (0.008)	0.861 (0.010)	0.902
SVM (linear)	0.868 (0.010)	0.896 (0.009)	0.820 (0.011)	0.895 (0.009)	0.823 (0.011)	0.910

Figure 10. Accuracies for Source of Bleeding Response (based on variable importance rankings)

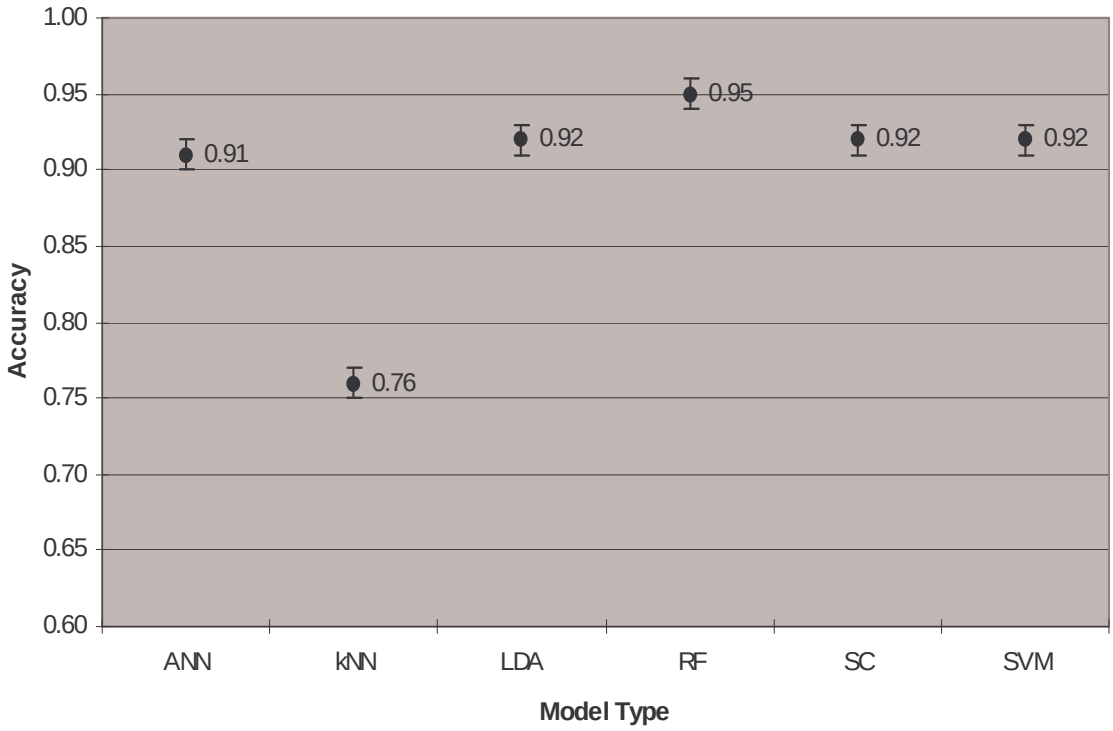


Figure 11. Accuracies for Resuscitation Response (based on variable importance rankings)

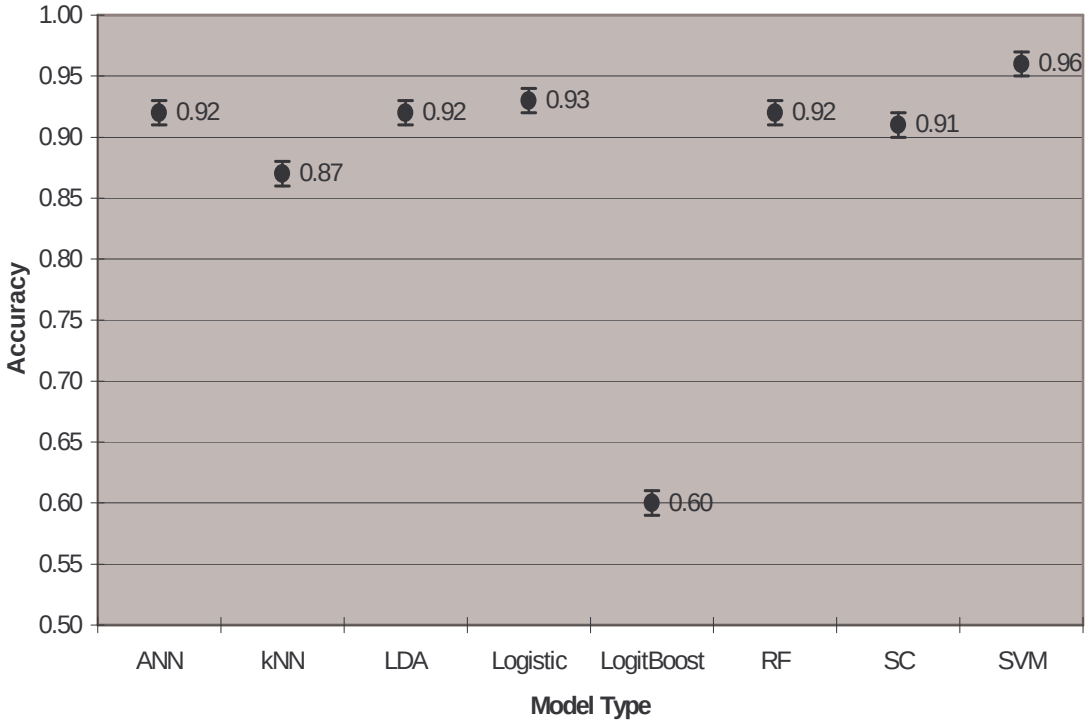


Figure 12. Accuracies for Endoscopy Response (based on variable importance rankings)

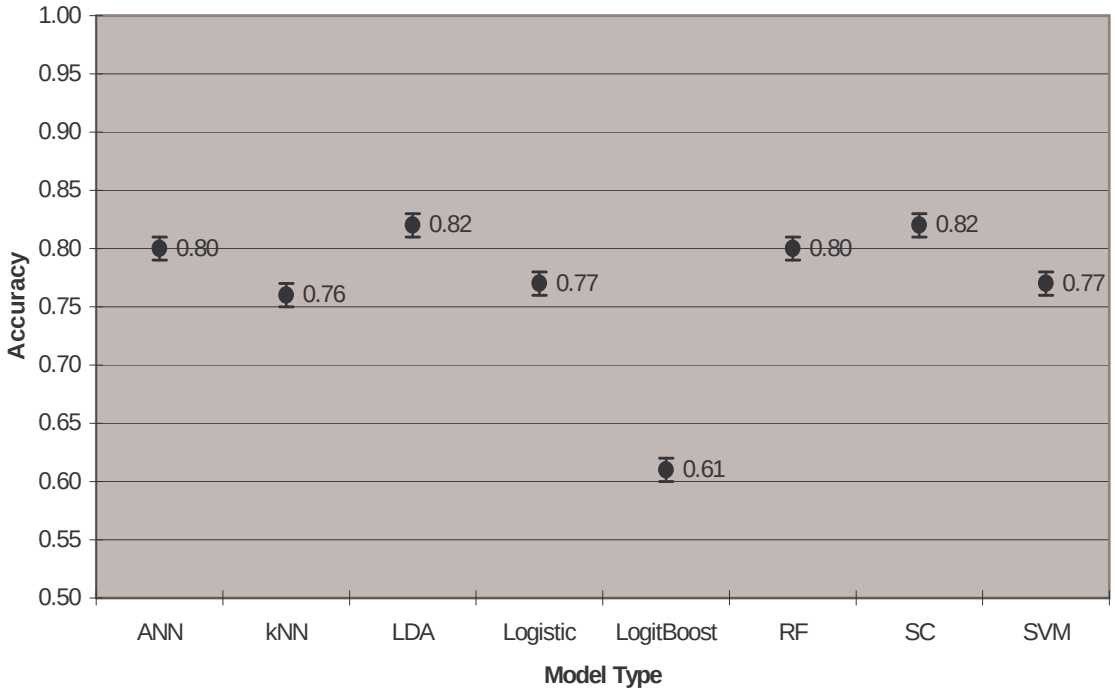


Figure 13. Accuracies for Disposition Response (based on variable importance rankings)

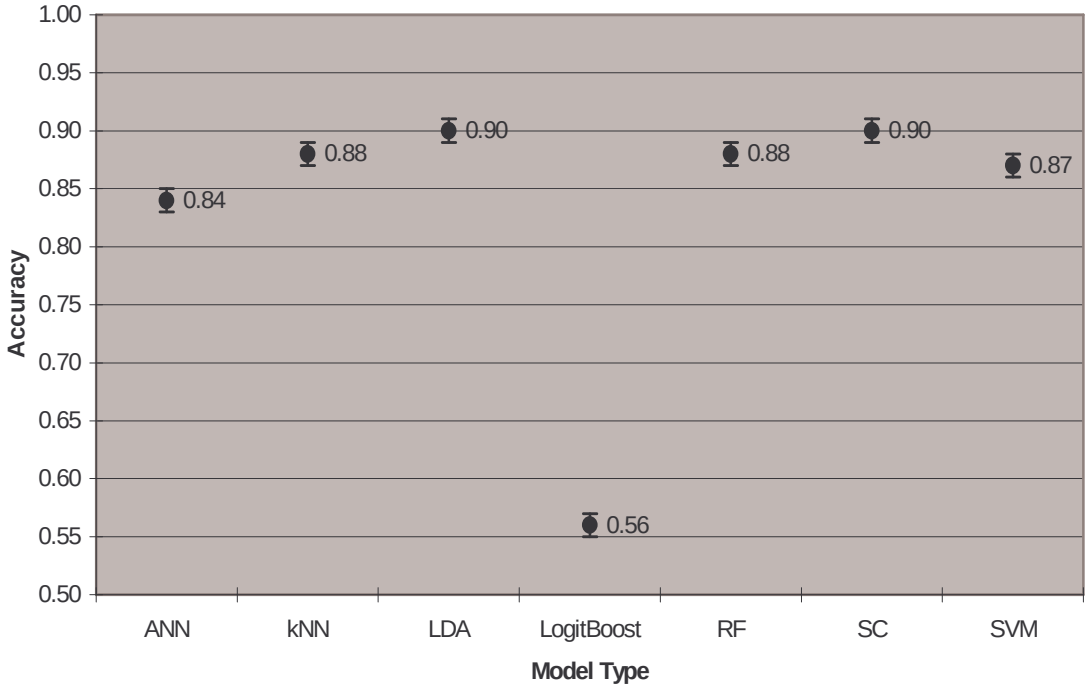


Table 20. Summary of McNemar's Test Results (based on variable importance rankings)

	Model (accuracy)	Model (accuracy)	p-value
Source of bleeding			
Least significantly different	RF (0.947)	LDA (0.923) or SC (0.923)	<0.0001
Least not significantly different[a]	–	–	–
Resuscitation			
Least significantly different	SVM (0.964)	Logistic (0.928)	<0.0001
Least not significantly different[a]	–	–	–
Endoscopy			
Least significantly different	LDA (0.823)	ANN (0.800)	<0.0001
Least not significantly different	LDA (0.823)	SC (0.820)	0.1675
Disposition			
Least significantly different	SC (0.900)	RF (0.884)	<0.0001
Least not significantly different	SC (0.900)	LDA (0.896)	0.0848

[a] All models were significantly different from highest accuracy model

Figure 14. ROC Curves for Predicting Source of Bleeding (based on variable importance rankings)

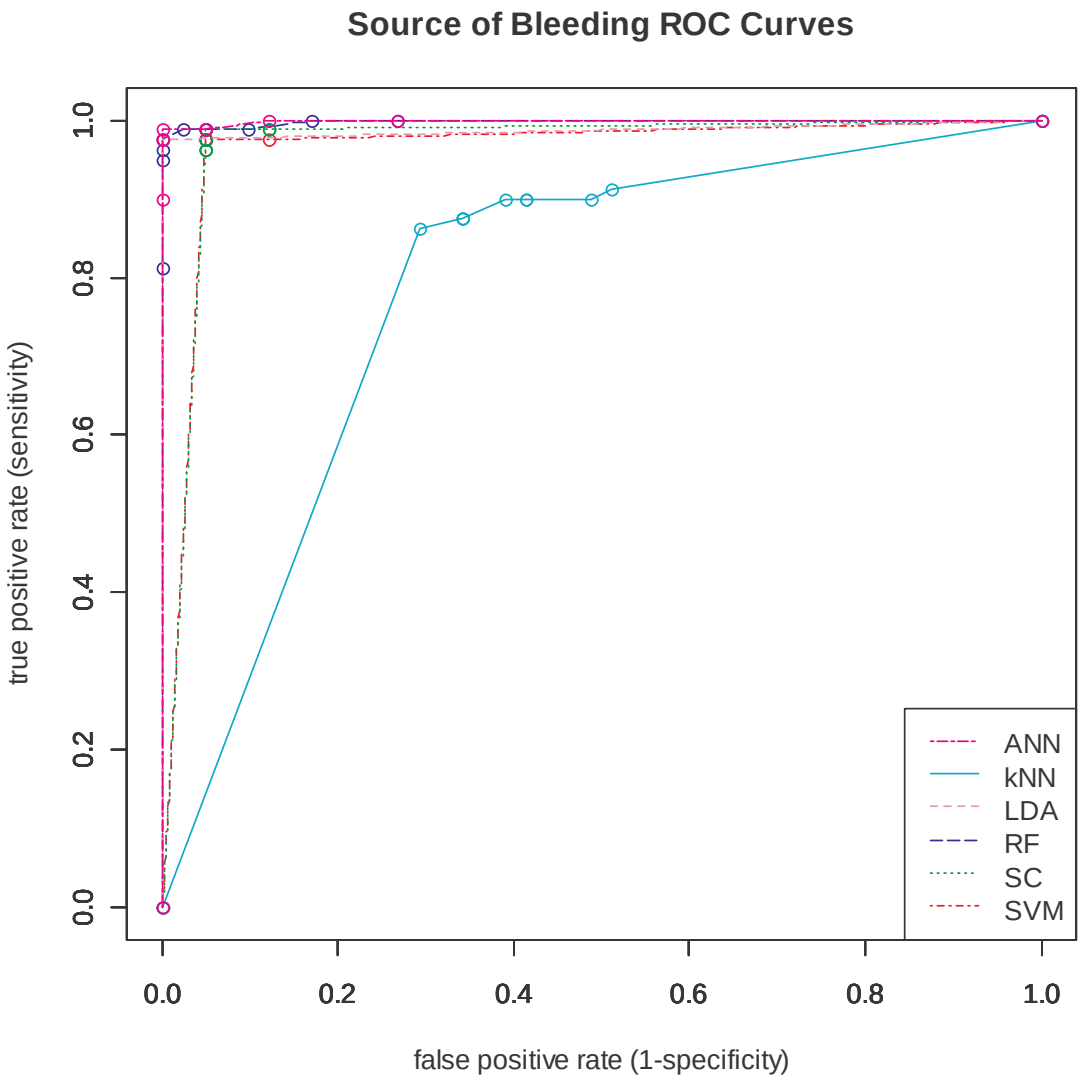


Figure 15. ROC Curves for Predicting Resuscitation (based on variable importance rankings)

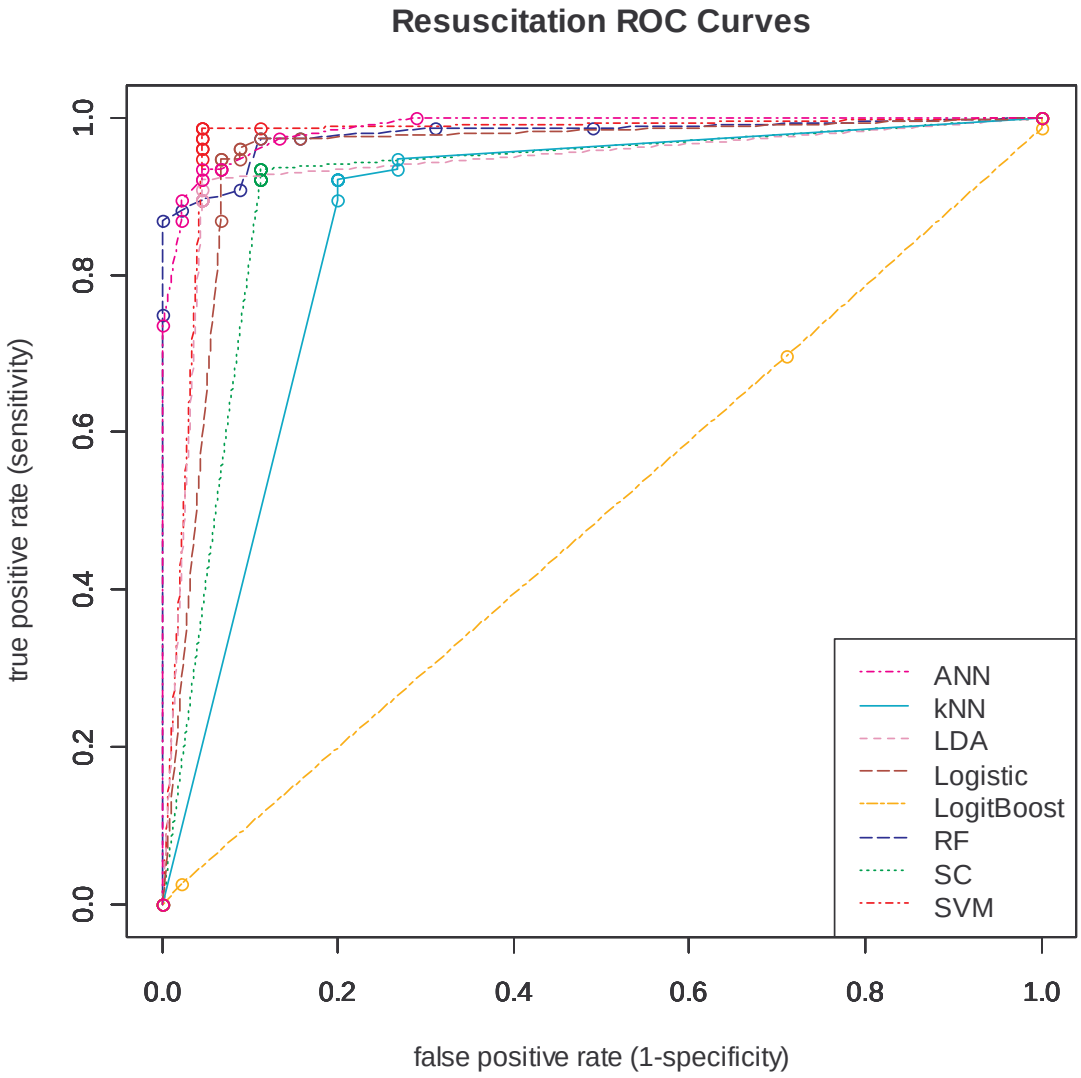


Figure 16. ROC Curves for Predicting Endoscopy (based on variable importance rankings)

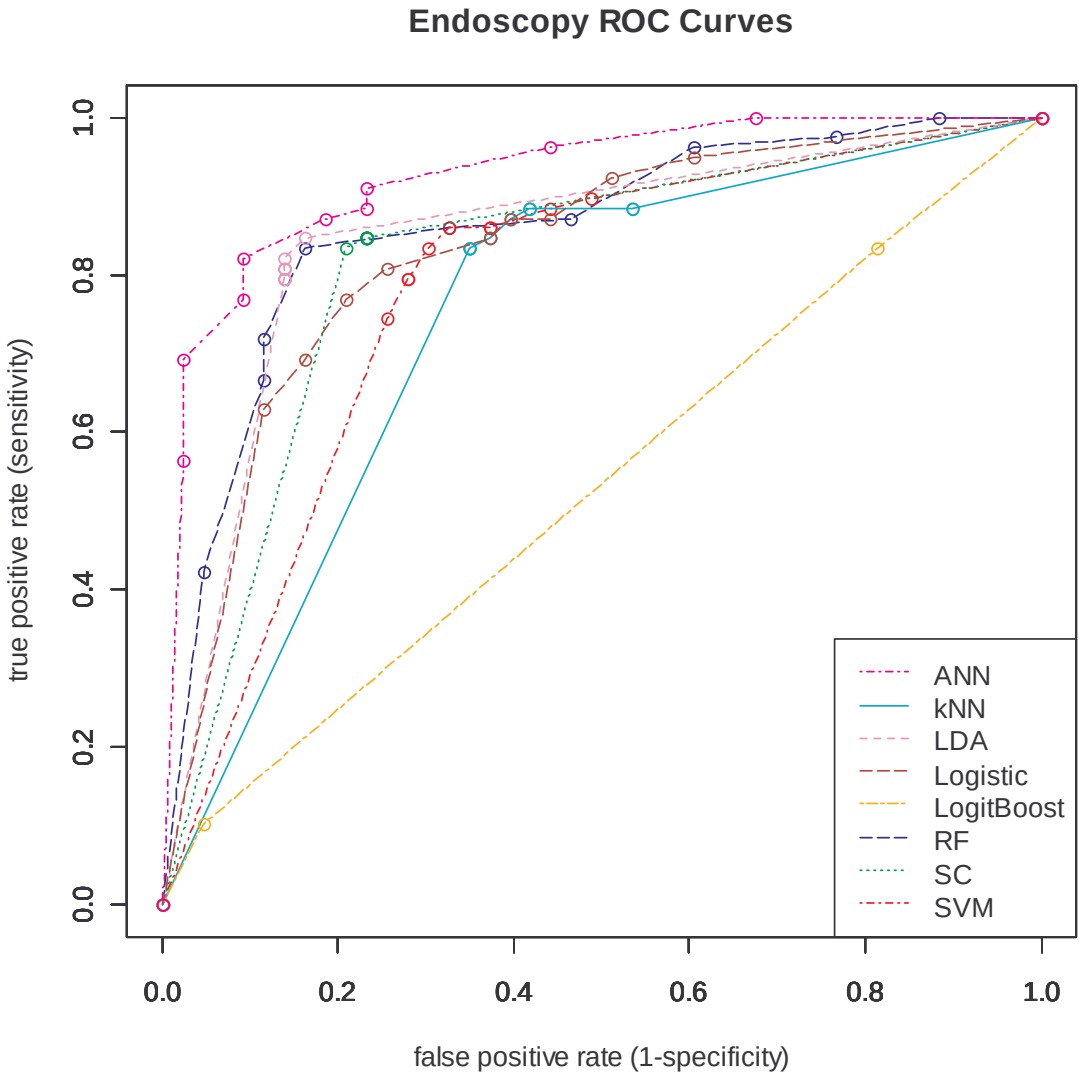
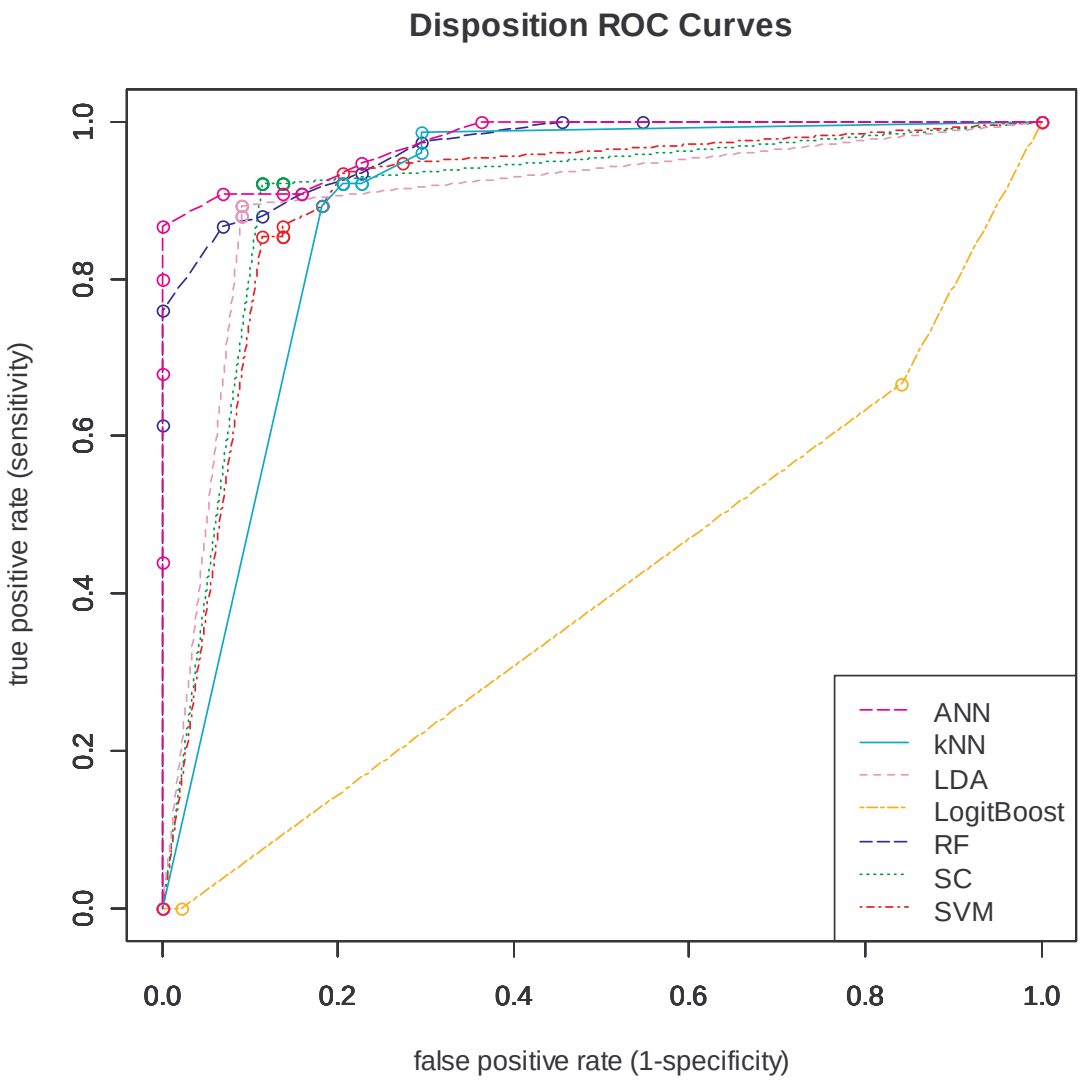


Figure 17. ROC Curves for Predicting Disposition (based on variable importance rankings)



For the response variable source of bleeding, the variables from Cr. and on were discarded (see Table 12). For resuscitation, the variables from NG_Lavage and on were discarded (see Table 13). For endoscopy, the variables from INR and on were discarded (Table 14). For disposition, the variables from Cirrhosis and on were discarded (Table 15). To obtain the variable importance rankings, each method took less than 5 minutes to run. To run the all models with the new chosen variables for one particular response doing 10 repetitions of 10-fold CV took about 30-40 minutes.

We observed that RF still performed best overall (in terms of accuracy and balance with sensitivity and specificity) with the newly selected subset of variables. Very few accuracies were statistically significantly higher and there were no big differences that

would be useful seen when comparing the AUC values. All the new accuracies were either about the same or lower for LDA. For the kNN model, at a significance level of 0.05, the new accuracy for source of bleeding was significantly higher while the remaining responses were about the same. For the disposition response, we observed that the logistic regression model performs poorly as before (results are not reported). For LogitBoost all the new accuracies were statistically significantly lower and the new AUCs were statistically significantly higher. SC tended to show slightly higher results or the same results. SVM's new accuracies were lower except for the resuscitation response – this accuracy was significantly higher. Hence there was not a consistent trend to be seen in comparing the accuracies and AUCs. We do see however that random forest still performed consistently well for all responses using the new subset of variables. Since there were no significant improvements with selecting variables using variable importance rankings that would be useful to us, all the models from now on will be those using the variables originally selected (as in Section 2).

4. External Validation Applied to RUGBE Database

The RUGBE (Registry in patients with Upper Gastrointestinal Bleeding undergoing an Endoscopy) database is a large Canadian database of patients with acute upper GIB from 18 participating sites across Canada. RUGBE is a network of 6 community and 12 tertiary care institutions from which source data is collected and entered into specialized electronic databases by specially trained research assistants. All patients presenting for medical attention for overt upper GI bleeding or with a history of hematemesis/coffee ground vomiting, melena, hematochezia, or a combination of any of the above within 24 hours preceding admission are considered for inclusion. Patients are entered in the registry only if an upper GI endoscopy is performed and patients with varices bleeding are excluded from the database. All data is reviewed at a single national location for internal logic of patient flow and biological plausibility, and ten percent of all records are audited on a quarterly basis by comparing them to the source data recorded in the hospital charts, thus further validating the abstracted information. All participating research staff and monitors used a glossary that included definitions of all variables entered in the registry to facilitate and standardize abstracted information.

4.1 Methods

The performance of these classification models on RUGBE data was analyzed to evaluate our approach of application of the model to clinical practice. The entire 122 patient dataset was used to train each model and the RUGBE dataset was the test set. Because the datasets were not completely alike, some alterations and recoding had to be done to both sets to assess the performance of the models. The set of predictor variables used included the following: history of prior GI bleeding, hematochezia, hematemesis, melena, ASA/NSAID use, blood pressure, heart rate, NG lavage, rectal exam, platelet count, and INR. Each model was run ten separate times and an average accuracy was obtained from the runs. The top three performing models were used to predict the RUGBE data.

4.2 Results from RUGBE Dataset

Patients with any missing data were deleted. However, this led to a large reduction of sample size because of many missing values for two variables: NG lavage and rectal exam. For this study, four separate strategies were undertaken (either deleting the rectal and/or NG lavage variable or keeping them both in). Keeping the NG lavage variable in resulted in a higher accuracy (94%~97%). See Table 21. Imputation of the missing

values for NG lavage and rectal were also tried. Since they were categorical variables, the mode was used to impute missing values. Imputation resulted in an increase of accuracy for ANN and RF.

Table 21. Results from RUGBE Data for Source of Bleeding Response

	# Patients used	ANN	RF	SVM
Using all 8 variables	94	0.885 (0.065)	0.884 (0.003)	0.926 (0.000)
Deleted NG Lavage	423	0.669 (0.207)	0.57 (0.033)	0.740 (0.006)
Deleted rectal	293	0.952 (0.024)	0.941 (0.002)	0.966 (0.000)
Deleted NG Lavage and rectal	1317	0.481 (0.115)	0.646 (0.021)	0.822 (0.005)
Imputation of NG Lavage and rectal	1317	0.967 (0.005)	0.969 (0.005)	0.903 (0.008)

Due to reduction of variables and recoding of data, LDA did not perform well on the RUGBE dataset. So ANN, a comparable model to RF and SVM, was used in its place. ANN was comparable in terms of accuracy and area under the ROC curve. RF and SVM were run in R (version 2.4.1) and ANN was run in Statistica (version 7.1). Running time for the three models was 10-15 minutes.

5. Comparison of Top Performing Model to Existing GIB Scores

5.1 Methods

In Section 2, we saw that the random forest model was our top performing model, having a high accuracy and good balance of sensitivity and specificity. We want to assess whether it would be beneficial to use this random forest model over existing GIB scores, such as the Rockall and Blatchford scores. Thus we compare the random forest model, the Rockall score, and Blatchford score. We assessed performance by comparing accuracies and area under ROC curves. Since the Rockall and Blatchford scores are applicable for upper GIB patients only, these were the patients considered. There were a total of 192 patients, with 126 of them being upper GIB patients. The areas under the correlated curves were compared for significant differences using the method described by DeLong and DeLong (43). Their method was implemented in R.

Although the Rockall score was originally designed to triage patients to high risk/low risk (admit to ICU/early discharge), it can also be used to determine those who are in need of urgent endoscopy. Only the initial Rockall score (without endoscopic data) was of true interest to us because the random forest model and Blatchford score both do not use endoscopic data. The full Rockall score was also examined merely to show that our random forest model is comparable to this even though our model does not include endoscopic data. The Blatchford score is used to identify those patients in need of urgent treatment. Thus we focused on the need for urgent endoscopy response.

Ten runs of 10-fold cross validation were performed. Parameters ntree (number of trees grown) and mtry (number of variables randomly sampled at each node split) were set to 200 and 1 respectively for the random forest model. These were the parameters that gave the best performance for random forest. The initial Rockall score (pre-endoscopy diagnosis data) ranges from 0 to 7 and the full Rockall score ranges from 0 to 11 while the Blatchford score ranges from 0 to 23. For creating the ROC curves, cutoff points for these scores were all possible integers the scores could take on. These cutoff points were then scaled to be between 0 and 1. Cutoff points for random forest were values between 0 and 1, in increments of 0.1. Areas under the ROC curves were calculated by finding the Mann-Whitney statistic and 95% confidence intervals were found to compare random forest to the two existing scores. In addition, accuracies for the scores and the random forest model were calculated. This analysis was run in R (version 2.4.1).

5.2 Results of Comparing Random Forest, Rockall Score, and Blatchford Score

Figure 18 shows the endoscopy ROC curves. Table 22 shows the area under the curve for each model/score. Without performing any tests, random forest seems to perform better than the Rockall score and slightly better than the Blatchford score in regards to the area under the ROC curve. Table 23 shows the 95% confidence intervals for testing the difference between random forest and the two scoring systems. The 95% confidence intervals have been adjusted for the multiple comparisons of scores and the random forest model (43). Comparing the initial Rockall score and the random forest model, there is a marginally significant difference between them, indicating the random forest model is slightly better than the initial Rockall score. There are no significant differences between the full Rockall score and random forest, or the Blatchford score and random forest. When the analysis was first performed with only 79 upper GIB patients (out of 122 patients total), the 95% confidence interval for the difference between the initial Rockall score and the random forest model was [-0.160,-0.008].

Table 22. Accuracies and Area Under ROC Curves for Comparison of Scores and RF Model

Score system / model	Accuracy	Area under ROC curve
Initial Rockall score	0.754	0.777
Full Rockall score	0.770	0.857
Blatchford score	0.817	0.863
Random forest	0.810	0.869

Table 23. 95% Confidence Intervals for the Difference Between a Scoring System and Random Forest

Comparison	95% confidence interval for the difference
Initial Rockall vs. random forest	[-0.182,-0.002]
Full Rockall vs. random forest	[-0.090,0.067]
Blatchford vs. random forest	[-0.069,0.057]

Figure 18. ROC Curves for Endoscopy (Comparing Rockall and Blatchford Scores and Random Forest)

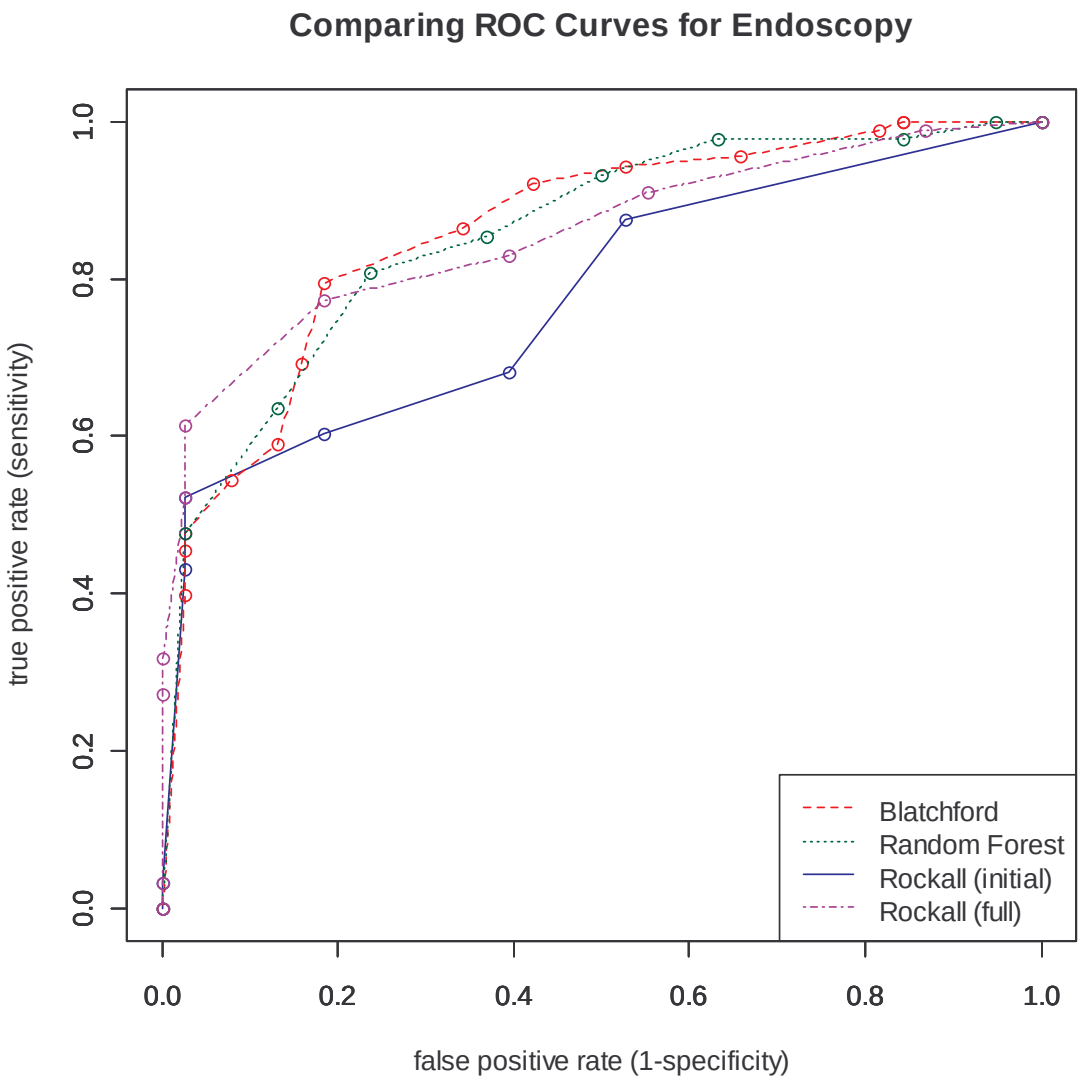


Table 22 shows the accuracies calculated for each score and the random forest model. The Blatchford score's accuracy and random forest's accuracy were virtually the same. Standard deviations could not be calculated for the Rockall and Blatchford scores because there was no way the score could ever vary for a given patient. The standard deviation for the random forest model however was 0.018. Given our results, we see that random forest performed statistically significantly better than the initial Rockall score. While random forest was not statistically significant to the full Rockall or Blatchford score, given the higher AUCs, there is indication that random forest is a good, reliable model to use in classifying GIB data. Random forest was shown to be comparable and not significantly worse than the full Rockall and Blatchford score. Further, random forest

is applicable to all sources of bleeding (i.e. upper, mid and lower GIB) not just one source of GIB bleeding, and is more versatile in that it can be used to predict different responses as well. Running time for this analysis took approximately 5 minutes.

6. Simulation Study

Aside from comparing the models using the GIB data, we want to compare the models using simulated GIB data to see in general how well they perform. All eight models were considered. For each response variable, simulated data was generated for four cases – a combination of unbalanced versus balanced data and correlated versus independent data. The sample size was 300 patients and the explanatory variable distributions were approximated by looking at the distributions from the actual GIB data.

For the source of bleeding response, the unbalanced data was divided as follows: 80% of the patients had upper bleeding, 15% had lower bleeding, and the remaining 5% had middle bleeding. For the resuscitation and endoscopy response, the unbalanced data was divided as 20% patients categorized as "Yes" with the rest (80%) as "No." Similarly, for the disposition response, 20% of the patients were categorized as placed in the ICU and the remaining 80% were classified as not being placed in the ICU. These resemble the actual proportions for GIB patients. For the balanced data, patients were classified evenly, i.e. 150 patients for each class (100 patients for source of bleeding response because there are 3 classes). Correlated data was simulated by generating a random number from the Uniform distribution, Uniform(0,0.3) for the correlation between two variables. Using values higher than 0.3 caused problems in obtaining the covariances and generating the multivariate normal distributions.

The data was initially generated from multivariate normal distributions and then changed accordingly to discrete distributions and right-skewed distributions where appropriate. For each class of a particular response variable, the means and standard deviations were taken from the real GIB data for each variable but their values exaggerated – means were spread further apart to indicate clearly the distinction between classes, and standard deviations used for the simulation data were two times the actual standard deviations. Variables were discretized by translating the original value to the area under the normal curve, making it a value between 0 and 1. Then, according to the proportions from the real GIB data, these values were assigned into a class. For example, if the input variable was 30% of the time "Yes" and 70% of the time "No", then values between [0,0.3] would be "Yes" and values between (0.3,1.0] would be "No." Proportions from the real GIB data were not followed exactly, but roughly followed to give a general idea. Variables were converted into skewed distributions by scaling the values down and exponentiating them, so that the variable would follow a log-normal distribution, which is right-skewed, and the peak of the distribution would start close to x=0. The values were then shifted over as needed to mimic the behavior of GIB variables.

Ten-fold cross validation was done, and 100 data files were generated for each case for each response, resulting in a total of 1600 files in total. The following statistics were calculated for each 10-fold cross validation, and their results were averaged together: accuracy, sensitivity, specificity, positive predictive value, and negative predictive value. As an alternative analysis, only a learning and test set were created, with 300 patients being in the learning set and 300 patients being in the test set. One hundred learning sets and 100 corresponding test sets were generated. The same statistics were calculated for this. We wanted to see whether there would be significant differences in 10-fold cross validation versus just a single learning and test set. McNemar's test was used to determine whether there were any significant differences in accuracies between the models. As with the actual GIB data, ROC curves were created and area under the ROC curves were found – this was done with the 10-fold cross validation results. Additionally, accuracies between the model with the highest accuracy and the other models were compared using McNemar's test. Using the Bonferroni correction to account for multiple comparisons of models, an appropriate alpha value was used for each test to control the error rate.

Since the analysis of the actual GIB data, there had been upgrades in the softwares used. The simulation study was run using R (version 2.7.2) for all models except ANN, which was run in Statistica (version 8.0). Due to changes in the Statistica software, the cross-validation option was no longer available. Random subsampling of the data was used. So, for each dataset, subsampling was done ten times to simulate 10-fold cross-validation as closely as possible. Further, due to computational constraints, only 5 files could be considered to create the ANN ROC curves instead of all 100 files. Using all 100 files would take 5 months (given a 40 hour work week) to simply prepare the data in a format that could be used to create the ROC curves. This is considering if nothing else was done but prepare the data (time does not include running the models and analyses themselves). For the learning/test set analysis, the learning and test sets had to be combined for ANN and then 50% of the dataset was randomly selected to be used as the test set. There were no options to specify particular cases to be included in either the training or test set. The only time particular cases could be selected was when they were to be included or excluded from the analysis.

6.1 Individual Model Parameters

All the same parameters were used as in the original analysis (Section 2) except new values were found for the shrunken centroid threshold, the *k* for *k*-nearest neighbor, and the mfinal parameter for boosting. A threshold of 2 was used for all responses for shrunken centroid. For kNN, k=3 for all the correlated data except for the endoscopy response, which was k=5. For the uncorrelated data, k=5 except for the disposition response, which was k=3. For boosting, mfinal was set to 5.

6.2 Simulation Study Results

Tables 24-43 give the results from the simulation study for each response and each combination (unbalanced versus balanced data and correlated versus independent data) as well as summary results from performing the McNemar's test. Figures 19-34 show the 16 ROC curves.

Table 24. Simulation Study Results for Source of Bleeding Response (unbalanced and correlated data) (standard error)

	ACC	SN	SP	PPV	NPV	AUC
ANN[a]	0.844 (0.002)	0.970 (0.001)	0.362 (0.003)	0.862 (0.002)	0.778 (0.002)	0.772
kNN	0.786 (0.002)	0.841 (0.002)	0.588 (0.003)	0.891 (0.002)	0.482 (0.003)	0.766
LDA	0.671 (0.003)	0.674 (0.003)	0.783 (0.002)	0.925 (0.002)	0.378 (0.003)	0.760
RF	0.954 (0.001)	0.999 (0.000)	0.777 (0.002)	0.947 (0.001)	0.994 (0.000)	0.992
SC	0.800 (0.002)	1.000 (0.000)	0.000 (0.000)	0.800 (0.002)	0.005 (0.000)	0.500
SVM – linear	0.816 (0.002)	0.963 (0.001)	0.232 (0.002)	0.834 (0.002)	0.606 (0.003)	0.651
SVM – radial	0.800 (0.002)	1.000 (0.000)	0.004 (0.000)	0.801 (0.002)	0.038 (0.001)	0.503

[a] Using only 5 files with subsampling

Table 25. Simulation Study Results for Source of Bleeding Response (unbalanced and not correlated data) (standard error)

	ACC	SN	SP	PPV	NPV	AUC
ANN[a]	0.837 (0.002)	0.959 (0.001)	0.378 (0.003)	0.861 (0.002)	0.745 (0.003)	0.775
kNN	0.791 (0.002)	0.859 (0.002)	0.549 (0.003)	0.885 (0.002)	0.493 (0.003)	0.754
LDA	0.694 (0.003)	0.708 (0.003)	0.763 (0.002)	0.923 (0.002)	0.399 (0.003)	0.769
RF	0.954 (0.001)	0.998 (0.000)	0.779 (0.002)	0.948 (0.001)	0.992 (0.001)	0.990
SC	0.800 (0.002)	1.000 (0.000)	0.000 (0.000)	0.800 (0.002)	0.000 (0.000)	0.500
SVM – linear	0.812 (0.002)	0.963 (0.001)	0.218 (0.002)	0.832 (0.002)	0.581 (0.003)	0.639
SVM – radial	0.801 (0.002)	0.999 (0.000)	0.012 (0.001)	0.802 (0.002)	0.077 (0.002)	0.510

[a] Using only 5 files with subsampling

Table 26. Simulation Study Results for Source of Bleeding Response (balanced and correlated data) (standard error)

	ACC	SN	SP	PPV	NPV	AUC
ANN[a]	0.770 (0.002)	0.686 (0.003)	0.869 (0.002)	0.740 (0.003)	0.848 (0.002)	0.856
kNN	0.787 (0.002)	0.470 (0.003)	0.961 (0.001)	0.857 (0.002)	0.784 (0.002)	0.734
LDA	0.706 (0.003)	0.530 (0.003)	0.873 (0.002)	0.678 (0.003)	0.788 (0.002)	0.731
RF	0.982 (0.001)	0.973 (0.001)	0.986 (0.001)	0.973 (0.001)	0.987 (0.001)	0.998
SC	0.714 (0.003)	0.344 (0.003)	0.926 (0.002)	0.714 (0.003)	0.739 (0.003)	0.659
SVM – linear	0.740 (0.003)	0.556 (0.003)	0.859 (0.002)	0.665 (0.003)	0.795 (0.002)	0.764
SVM – radial	0.481 (0.003)	0.487 (0.003)	0.578 (0.003)	0.366 (0.003)	0.694 (0.003)	0.513

[a] Using only 5 files with subsampling

Table 27. Simulation Study Results for Source of Bleeding Response (balanced and not correlated data) (standard error)

	ACC	SN	SP	PPV	NPV	AUC
ANN[a]	0.759 (0.002)	0.664 (0.003)	0.873 (0.002)	0.733 (0.003)	0.842 (0.002)	0.896
kNN	0.763 (0.002)	0.392 (0.003)	0.972 (0.001)	0.875 (0.002)	0.762 (0.002)	0.726
LDA	0.707 (0.003)	0.533 (0.003)	0.861 (0.002)	0.658 (0.003)	0.787 (0.002)	0.728
RF	0.981 (0.001)	0.973 (0.001)	0.985 (0.001)	0.970 (0.001)	0.987 (0.001)	0.998
SC	0.717 (0.003)	0.359 (0.003)	0.925 (0.002)	0.714 (0.003)	0.744 (0.003)	0.663
SVM – linear	0.741 (0.003)	0.557 (0.003)	0.859 (0.002)	0.665 (0.003)	0.796 (0.002)	0.760
SVM – radial	0.480 (0.003)	0.479 (0.003)	0.581 (0.003)	0.364 (0.003)	0.692 (0.003)	0.509

[a] Using only 5 files with subsampling

Table 28. Summary of McNemar's Test Results (Source of Bleeding Response)

	Model (accuracy)	Model (accuracy)	p-value
Unbalanced/correlated data			
Least significantly different	RF (0.954)	ANN (0.844)	<0.0001
Least not significantly different[a]	–	–	–
Unbalanced/not correlated data			
Least significantly different	RF (0.954)	ANN (0.837)	<0.0001
Least not significantly different[a]	–	–	–
Balanced/correlated data			
Least significantly different	RF (0.982)	kNN (0.787)	<0.0001
Least not significantly different[a]	–	–	–
Balanced/not correlated data			
Least significantly different	RF (0.981)	kNN (0.763)	<0.0001
Least not significantly different[a]	–	–	–

[a] All models were significantly different from highest accuracy model

Table 29. Simulation Study Results for Resuscitation Response (unbalanced and correlated data) (standard error)

	ACC	SN	SP	PPV	NPV	AUC
ANN[a]	0.949 (0.001)	0.785 (0.002)	0.990 (0.001)	0.950 (0.001)	0.949 (0.001)	0.933
kNN	0.833 (0.002)	0.201 (0.002)	0.991 (0.001)	0.852 (0.002)	0.832 (0.002)	0.624
LDA	0.938 (0.001)	0.741 (0.003)	0.987 (0.001)	0.937 (0.001)	0.938 (0.001)	0.883
Logistic	0.933 (0.001)	0.789 (0.002)	0.969 (0.001)	0.867 (0.002)	0.948 (0.001)	0.925
LogitBoost	0.800 (0.002)	0.000 (0.000)	1.000 (0.000)	0.000 (0.000)	0.800 (0.002)	0.518
RF	0.988 (0.001)	0.951 (0.001)	0.997 (0.000)	0.988 (0.001)	0.988 (0.001)	0.999
SC	0.904 (0.002)	0.518 (0.003)	1.000 (0.000)	1.000 (0.000)	0.893 (0.002)	0.777
SVM – linear	0.943 (0.001)	0.768 (0.002)	0.986 (0.001)	0.934 (0.001)	0.945 (0.001)	0.912
SVM – radial	0.821 (0.002)	0.112 (0.002)	0.998 (0.000)	0.140 (0.002)	0.821 (0.002)	0.562

[a] Using only 5 files with subsampling

Table 30. Simulation Study Results for Resuscitation Response (unbalanced and not correlated data) (standard error)

	ACC	SN	SP	PPV	NPV	AUC
ANN[a]	0.940 (0.001)	0.763 (0.002)	0.984 (0.001)	0.931 (0.001)	0.944 (0.001)	0.914
kNN	0.819 (0.002)	0.112 (0.002)	0.996 (0.000)	0.860 (0.002)	0.818 (0.002)	0.619
LDA	0.938 (0.001)	0.738 (0.003)	0.988 (0.001)	0.939 (0.001)	0.938 (0.001)	0.881
Logistic	0.928 (0.001)	0.764 (0.002)	0.970 (0.001)	0.863 (0.002)	0.943 (0.001)	0.912
LogitBoost	0.800 (0.002)	0.000 (0.000)	1.000 (0.000)	0.000 (0.000)	0.800 (0.002)	0.500
RF	0.988 (0.001)	0.949 (0.001)	0.998 (0.000)	0.990 (0.001)	0.987 (0.001)	0.999
SC	0.905 (0.002)	0.523 (0.003)	1.000 (0.000)	1.000 (0.000)	0.894 (0.002)	0.780
SVM – linear	0.938 (0.001)	0.744 (0.003)	0.986 (0.001)	0.930 (0.001)	0.939 (0.001)	0.899
SVM – radial	0.814 (0.002)	0.076 (0.002)	0.999 (0.000)	0.094 (0.002)	0.814 (0.002)	0.541

[a] Using only 5 files with subsampling

Table 31. Simulation Study Results for Resuscitation Response (balanced and correlated data) (standard error)

	ACC	SN	SP	PPV	NPV	AUC
ANN[a]	0.907 (0.002)	0.874 (0.002)	0.938 (0.001)	0.938 (0.001)	0.884 (0.002)	0.958
kNN	0.733 (0.003)	0.551 (0.003)	0.914 (0.002)	0.865 (0.002)	0.671 (0.003)	0.776
LDA	0.864 (0.002)	0.763 (0.002)	0.964 (0.001)	0.956 (0.001)	0.803 (0.002)	0.882
Logistic	0.892 (0.002)	0.857 (0.002)	0.926 (0.002)	0.921 (0.002)	0.867 (0.002)	0.932
LogitBoost	0.498 (0.003)	0.498 (0.003)	0.498 (0.003)	0.498 (0.003)	0.498 (0.003)	0.493
RF	0.985 (0.001)	0.987 (0.001)	0.983 (0.001)	0.983 (0.001)	0.987 (0.001)	0.999
SC	0.820 (0.002)	0.651 (0.003)	0.989 (0.001)	0.984 (0.001)	0.740 (0.003)	0.834
SVM – linear	0.896 (0.002)	0.840 (0.002)	0.951 (0.001)	0.946 (0.001)	0.856 (0.002)	0.920
SVM – radial	0.526 (0.003)	0.671 (0.003)	0.382 (0.003)	0.523 (0.003)	0.544 (0.003)	0.470

[a] Using only 5 files with subsampling

Table 32. Simulation Study Results for Resuscitation Response (balanced and not correlated data) (standard error)

	ACC	SN	SP	PPV	NPV	AUC
ANN[a]	0.886 (0.002)	0.848 (0.002)	0.924 (0.002)	0.923 (0.002)	0.861 (0.002)	0.926
kNN	0.688 (0.003)	0.465 (0.003)	0.910 (0.002)	0.839 (0.002)	0.630 (0.003)	0.753
LDA	0.852 (0.002)	0.745 (0.003)	0.959 (0.001)	0.948 (0.001)	0.790 (0.002)	0.870
Logistic	0.869 (0.002)	0.826 (0.002)	0.912 (0.002)	0.904 (0.002)	0.840 (0.002)	0.912
LogitBoost	0.492 (0.003)	0.489 (0.003)	0.495 (0.003)	0.492 (0.003)	0.492 (0.003)	0.490
RF	0.985 (0.001)	0.986 (0.001)	0.985 (0.001)	0.985 (0.001)	0.986 (0.001)	0.999
SC	0.819 (0.002)	0.651 (0.003)	0.987 (0.001)	0.981 (0.001)	0.739 (0.003)	0.834
SVM – linear	0.875 (0.002)	0.806 (0.002)	0.945 (0.001)	0.936 (0.001)	0.830 (0.002)	0.902
SVM – radial	0.528 (0.003)	0.647 (0.003)	0.408 (0.003)	0.525 (0.003)	0.540 (0.003)	0.477

[a] Using only 5 files with subsampling

Table 33. Summary of McNemar's Test Results (Resuscitation Response)

	Model (accuracy)	Model (accuracy)	p-value
Unbalanced/correlated data			
Least significantly different	RF (0.988)	ANN (0.949)	<0.0001
Least not significantly different[a]	–	–	–
Unbalanced/not correlated data			
Least significantly different	RF (0.988)	ANN (0.940)	<0.0001
Least not significantly different[a]	–	–	–
Balanced/correlated data			
Least significantly different	RF (0.985)	ANN (0.907)	<0.0001
Least not significantly different[a]	–	–	–
Balanced/not correlated data			
Least significantly different	RF (0.985)	ANN (0.886)	<0.0001
Least not significantly different [a]	–	–	–

[a] All models were significantly different from highest accuracy model

Table 34. Simulation Study Results for Endoscopy Response (unbalanced and correlated data) (standard error)

	ACC	SN	SP	PPV	NPV	AUC
ANN[a]	0.965 (0.001)	0.853 (0.002)	0.993 (0.001)	0.970 (0.001)	0.964 (0.001)	0.972
kNN	0.814 (0.002)	0.188 (0.002)	0.970 (0.001)	0.610 (0.003)	0.827 (0.002)	0.588
LDA	0.950 (0.001)	0.794 (0.002)	0.989 (0.001)	0.947 (0.001)	0.950 (0.001)	0.904
Logistic	0.935 (0.001)	0.828 (0.002)	0.962 (0.001)	0.846 (0.002)	0.957 (0.001)	0.947
LogitBoost	0.800 (0.002)	0.000 (0.000)	1.000 (0.000)	0.000 (0.000)	0.800 (0.002)	0.503
RF	0.968 (0.001)	0.870 (0.002)	0.992 (0.001)	0.966 (0.001)	0.968 (0.001)	0.993
SC	0.939 (0.001)	0.699 (0.003)	0.999 (0.000)	0.993 (0.000)	0.930 (0.001)	0.861
SVM – linear	0.952 (0.001)	0.842 (0.002)	0.979 (0.001)	0.911 (0.002)	0.961 (0.001)	0.940
SVM – radial	0.800 (0.002)	0.000 (0.000)	1.000 (0.000)	0.000 (0.000)	0.800 (0.002)	0.500

[a] Using only 5 files with subsampling

Table 35. Simulation Study Results for Endoscopy Response (unbalanced and not correlated data) (standard error)

	ACC	SN	SP	PPV	NPV	AUC
ANN[a]	0.958 (0.001)	0.819 (0.002)	0.994 (0.000)	0.972 (0.001)	0.956 (0.001)	0.960
kNN	0.810 (0.002)	0.107 (0.002)	0.986 (0.001)	0.655 (0.003)	0.815 (0.002)	0.574
LDA	0.957 (0.001)	0.810 (0.002)	0.994 (0.000)	0.968 (0.001)	0.954 (0.001)	0.913
Logistic	0.935 (0.001)	0.822 (0.002)	0.964 (0.001)	0.851 (0.002)	0.956 (0.001)	0.942
LogitBoost	0.800 (0.002)	0.000 (0.000)	1.000 (0.000)	0.010 (0.001)	0.800 (0.002)	0.492
RF	0.970 (0.001)	0.878 (0.002)	0.993 (0.000)	0.971 (0.001)	0.970 (0.001)	0.992
SC	0.942 (0.001)	0.711 (0.003)	1.000 (0.000)	0.997 (0.000)	0.933 (0.001)	0.868
SVM – linear	0.949 (0.001)	0.833 (0.002)	0.978 (0.001)	0.904 (0.002)	0.959 (0.001)	0.934
SVM – radial	0.800 (0.002)	0.000 (0.000)	1.000 (0.000)	0.000 (0.000)	0.800 (0.002)	0.500

[a] Using only 5 files with subsampling

Table 36. Simulation Study Results for Endoscopy Response (balanced and correlated data) (standard error)

	ACC	SN	SP	PPV	NPV	AUC
ANN[a]	0.926 (0.002)	0.905 (0.002)	0.949 (0.001)	0.948 (0.001)	0.910 (0.002)	0.963
kNN	0.688 (0.003)	0.574 (0.003)	0.802 (0.002)	0.745 (0.003)	0.654 (0.003)	0.730
LDA	0.888 (0.002)	0.808 (0.002)	0.967 (0.001)	0.961 (0.001)	0.835 (0.002)	0.901
Logistic	0.905 (0.002)	0.892 (0.002)	0.918 (0.002)	0.915 (0.002)	0.895 (0.002)	0.955
LogitBoost	0.502 (0.003)	0.503 (0.003)	0.501 (0.003)	0.502 (0.003)	0.502 (0.003)	0.502
RF	0.955 (0.001)	0.944 (0.001)	0.965 (0.001)	0.965 (0.001)	0.945 (0.001)	0.991
SC	0.884 (0.002)	0.783 (0.002)	0.984 (0.001)	0.980 (0.001)	0.820 (0.002)	0.894
SVM – linear	0.912 (0.002)	0.888 (0.002)	0.936 (0.001)	0.933 (0.001)	0.893 (0.002)	0.936
SVM – radial	0.448 (0.003)	0.525 (0.003)	0.371 (0.003)	0.452 (0.003)	0.438 (0.003)	0.320

[a] Using only 5 files with subsampling

Table 37. Simulation Study Results for Endoscopy Response (balanced and not correlated data) (standard error)

	ACC	SN	SP	PPV	NPV	AUC
ANN[a]	0.918 (0.002)	0.896 (0.002)	0.939 (0.001)	0.941 (0.001)	0.902 (0.002)	0.972
kNN	0.673 (0.003)	0.535 (0.003)	0.811 (0.002)	0.740 (0.003)	0.636 (0.003)	0.719
LDA	0.898 (0.002)	0.823 (0.002)	0.974 (0.001)	0.969 (0.001)	0.846 (0.002)	0.911
Logistic	0.898 (0.002)	0.881 (0.002)	0.914 (0.002)	0.912 (0.002)	0.885 (0.002)	0.948
LogitBoost	0.504 (0.003)	0.500 (0.003)	0.508 (0.003)	0.504 (0.003)	0.505 (0.003)	0.517
RF	0.956 (0.001)	0.946 (0.001)	0.965 (0.001)	0.965 (0.001)	0.947 (0.001)	0.991
SC	0.891 (0.002)	0.795 (0.002)	0.987 (0.001)	0.984 (0.001)	0.829 (0.002)	0.902
SVM – linear	0.905 (0.002)	0.877 (0.002)	0.933 (0.001)	0.930 (0.001)	0.884 (0.002)	0.930
SVM – radial	0.450 (0.003)	0.507 (0.003)	0.392 (0.003)	0.452 (0.003)	0.444 (0.003)	0.317

[a] Using only 5 files with subsampling

Table 38. Summary of McNemar's Test Results (Endoscopy Response)

	Model (accuracy)	Model (accuracy)	p-value
Unbalanced/correlated data			
Least significantly different	RF (0.968)	SVM – linear (0.952)	<0.0001
Least not significantly different	RF (0.968)	ANN (0.965)	0.5364
Unbalanced/not correlated data			
Least significantly different	RF (0.970)	LDA (0.957)	<0.0001
Least not significantly different	RF (0.970)	ANN (0.958)	0.6604
Balanced/correlated data			
Least significantly different	RF (0.955)	SVM – linear (0.912)	<0.0001
Least not significantly different	RF (0.955)	ANN (0.926)	0.8129
Balanced/not correlated data			
Least significantly different	RF (0.956)	SVM – linear (0.905)	<0.0001
Least not significantly different	RF (0.956)	ANN (0.918)	0.8499

Table 39. Simulation Study Results for Disposition Response (unbalanced and correlated data) (standard error)

	ACC	**SN**	**SP**	**PPV**	**NPV**	**AUC**
ANN[a]	0.979 (0.001)	0.910 (0.002)	0.996 (0.001)	0.985 (0.001)	0.978 (0.001)	0.967
kNN	0.912 (0.002)	0.576 (0.003)	0.997 (0.000)	0.977 (0.001)	0.904 (0.002)	0.800
LDA	0.967 (0.001)	0.907 (0.002)	0.982 (0.001)	0.930 (0.001)	0.977 (0.001)	0.958
LogitBoost	0.800 (0.002)	0.000 (0.000)	0.999 (0.000)	0.000 (0.000)	0.800 (0.002)	0.523
RF	0.991 (0.001)	0.955 (0.001)	1.000 (0.000)	0.999 (0.000)	0.989 (0.001)	1.000
SC	0.924 (0.002)	0.621 (0.003)	1.000 (0.000)	1.000 (0.000)	0.914 (0.002)	0.824
SVM – linear	0.968 (0.001)	0.893 (0.002)	0.987 (0.001)	0.945 (0.001)	0.974 (0.001)	0.964
SVM – radial	0.800 (0.002)	0.000 (0.000)	1.000 (0.000)	0.000 (0.000)	0.800 (0.002)	0.500

[a] Using only 5 files with subsampling

Table 40. Simulation Study Results for Disposition Response (unbalanced and not correlated data) (standard error)

	ACC	SN	SP	PPV	NPV	AUC
ANN[a]	0.980 (0.001)	0.917 (0.002)	0.996 (0.000)	0.983 (0.001)	0.980 (0.001)	0.968
kNN	0.909 (0.002)	0.557 (0.003)	0.997 (0.000)	0.976 (0.001)	0.900 (0.002)	0.792
LDA	0.969 (0.001)	0.925 (0.002)	0.979 (0.001)	0.921 (0.002)	0.981 (0.001)	0.967
LogitBoost	0.800 (0.002)	0.000 (0.000)	1.000 (0.000)	0.000 (0.000)	0.800 (0.002)	0.493
RF	0.992 (0.001)	0.962 (0.001)	1.000 (0.000)	0.999 (0.000)	0.991 (0.001)	1.000
SC	0.926 (0.002)	0.632 (0.003)	1.000 (0.000)	1.000 (0.000)	0.916 (0.002)	0.829
SVM – linear	0.972 (0.001)	0.903 (0.002)	0.989 (0.001)	0.955 (0.001)	0.976 (0.001)	0.969
SVM – radial	0.800 (0.002)	0.000 (0.000)	1.000 (0.000)	0.000 (0.000)	0.800 (0.002)	0.500

[a] Using only 5 files with subsampling

Table 41. Simulation Study Results for Disposition Response (balanced and correlated data) (standard error)

	ACC	SN	SP	PPV	NPV	AUC
ANN[a]	0.966 (0.001)	0.956 (0.001)	0.975 (0.001)	0.976 (0.001)	0.958 (0.001)	0.993
kNN	0.833 (0.002)	0.712 (0.003)	0.955 (0.001)	0.940 (0.001)	0.768 (0.002)	0.860
LDA	0.917 (0.002)	0.975 (0.001)	0.859 (0.002)	0.875 (0.002)	0.973 (0.001)	0.932
LogitBoost	0.509 (0.003)	0.511 (0.003)	0.508 (0.003)	0.509 (0.003)	0.510 (0.003)	0.521
RF	0.994 (0.000)	0.991 (0.001)	0.997 (0.000)	0.997 (0.000)	0.991 (0.001)	1.000
SC	0.952 (0.001)	0.928 (0.001)	0.977 (0.001)	0.976 (0.001)	0.932 (0.001)	0.962
SVM – linear	0.945 (0.001)	0.936 (0.001)	0.953 (0.001)	0.953 (0.001)	0.937 (0.001)	0.969
SVM – radial	0.435 (0.003)	0.515 (0.003)	0.355 (0.003)	0.441 (0.003)	0.421 (0.003)	0.274

[a] Using only 5 files with subsampling

Table 42. Simulation Study Results for Disposition Response (balanced and not correlated data) (standard error)

	ACC	SN	SP	PPV	NPV	AUC
ANN[a]	0.966 (0.001)	0.955 (0.001)	0.977 (0.001)	0.977 (0.001)	0.956 (0.001)	0.991
kNN	0.831 (0.002)	0.709 (0.003)	0.954 (0.001)	0.939 (0.001)	0.767 (0.002)	0.860
LDA	0.915 (0.002)	0.982 (0.001)	0.848 (0.002)	0.867 (0.002)	0.979 (0.001)	0.931
LogitBoost	0.491 (0.003)	0.489 (0.003)	0.493 (0.003)	0.491 (0.003)	0.491 (0.003)	0.502
RF	0.994 (0.000)	0.991 (0.001)	0.997 (0.000)	0.997 (0.000)	0.991 (0.001)	1.000
SC	0.960 (0.001)	0.944 (0.001)	0.977 (0.001)	0.976 (0.001)	0.946 (0.001)	0.971
SVM – linear	0.951 (0.001)	0.942 (0.001)	0.961 (0.001)	0.960 (0.001)	0.943 (0.001)	0.973
SVM – radial	0.434 (0.003)	0.501 (0.003)	0.367 (0.003)	0.439 (0.003)	0.422 (0.003)	0.275

[a] Using only 5 files with subsampling

Table 43. Summary of McNemar's Test Results (Disposition Response)

	Model (accuracy)	Model (accuracy)	p-value
Unbalanced/correlated data			
Least significantly different	RF (0.991)	ANN (0.979)	0.0013
Least not significantly different[a]	–	–	–
Unbalanced/not correlated data			
Least significantly different	RF (0.992)	ANN (0.980)	0.0010
Least not significantly different[a]	–	–	–
Balanced/correlated data			
Least significantly different	RF (0.994)	ANN (0.966)	<0.0001
Least not significantly different[a]	–	–	–
Balanced/not correlated data			
Least significantly different	RF (0.994)	ANN (0.966)	<0.0001
Least not significantly different[a]	–	–	–

[a] All models were significantly different from highest accuracy model

Note: On the following ROC curves, the ANN ROC curve was found using subsampling with only 5 files

Figure 19. Simulation Study ROC Curves for Source of Bleeding Response (unbalanced and correlated data)

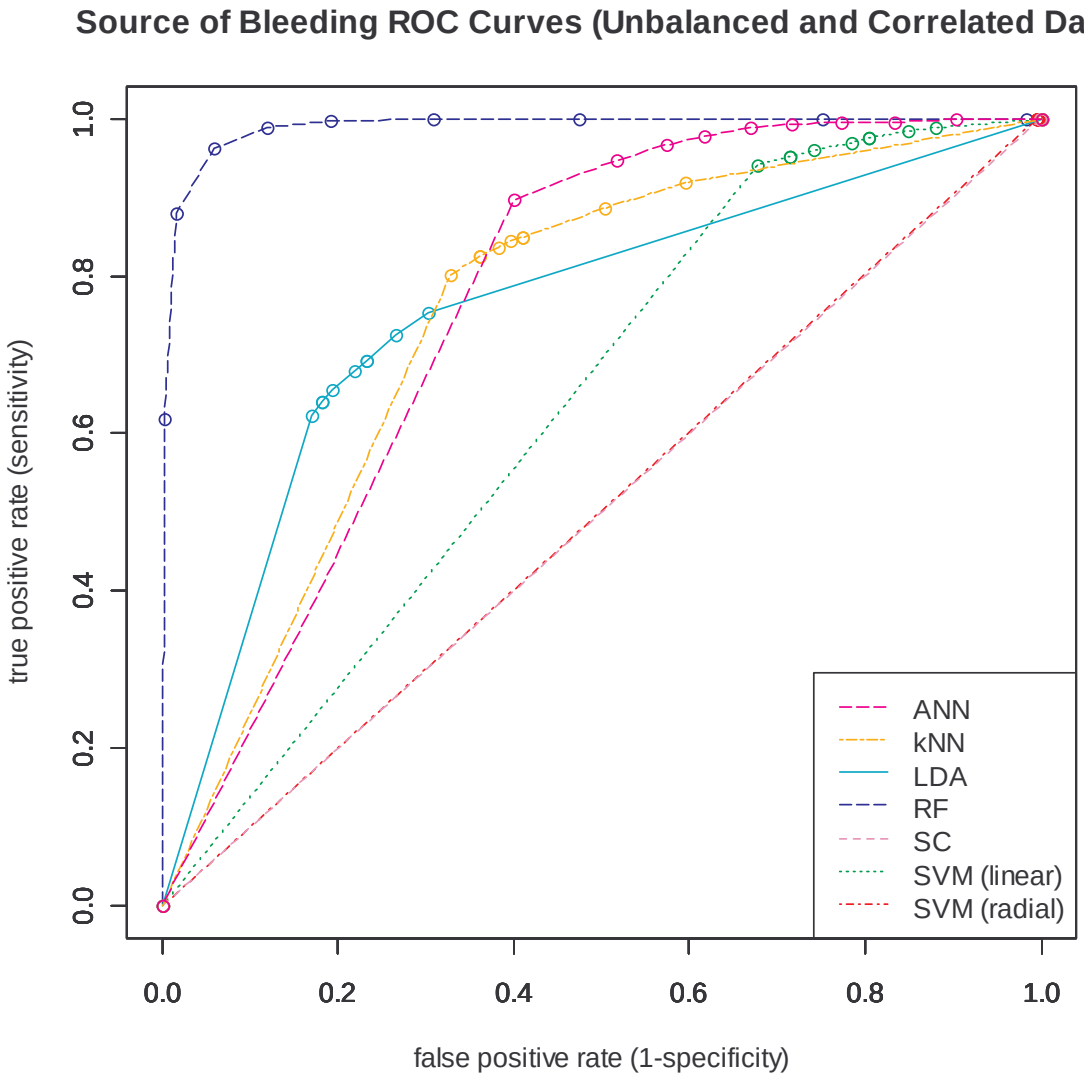


Figure 20. Simulation Study ROC Curves for Source of Bleeding Response (unbalanced and not correlated data)

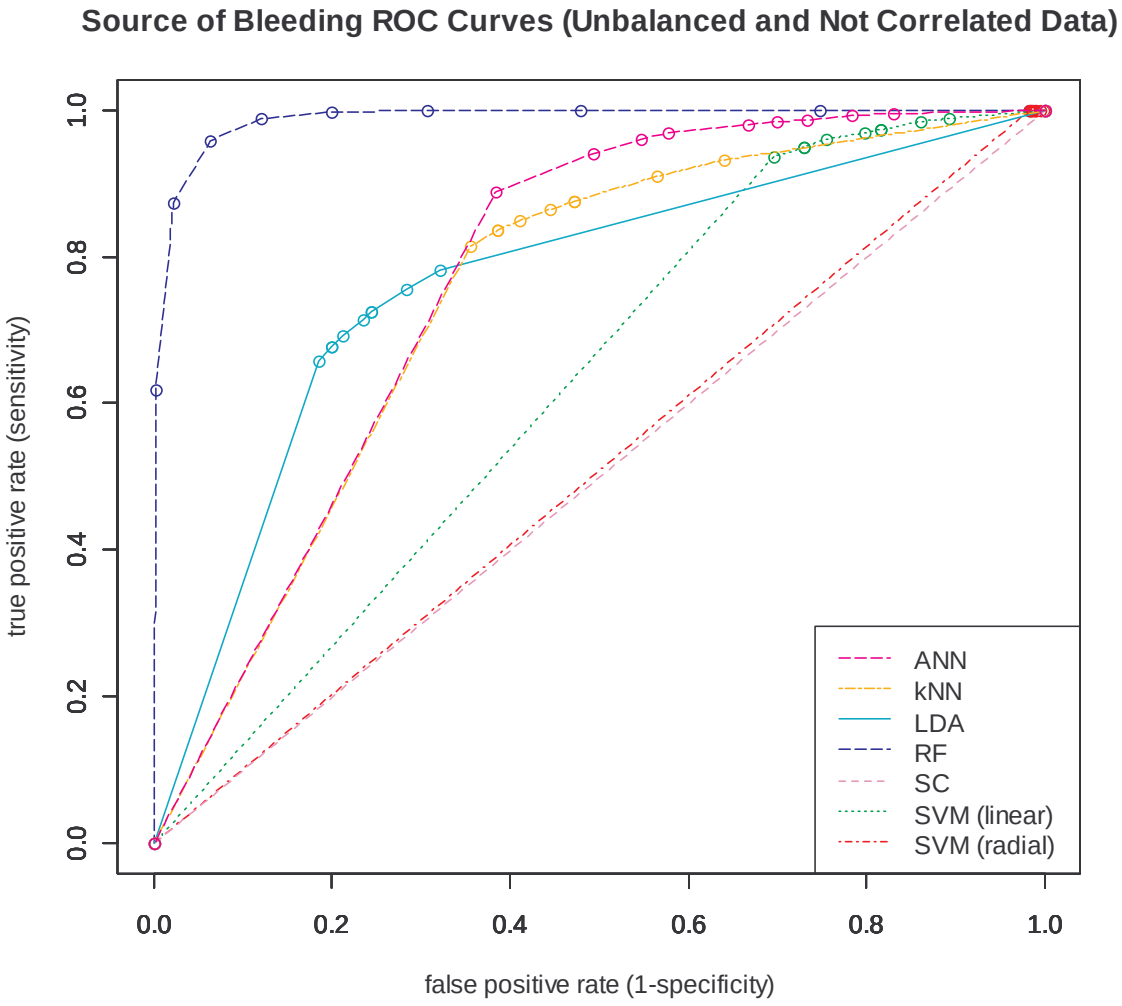


Figure 21. Simulation Study ROC Curves for Source of Bleeding Response (balanced and correlated data)

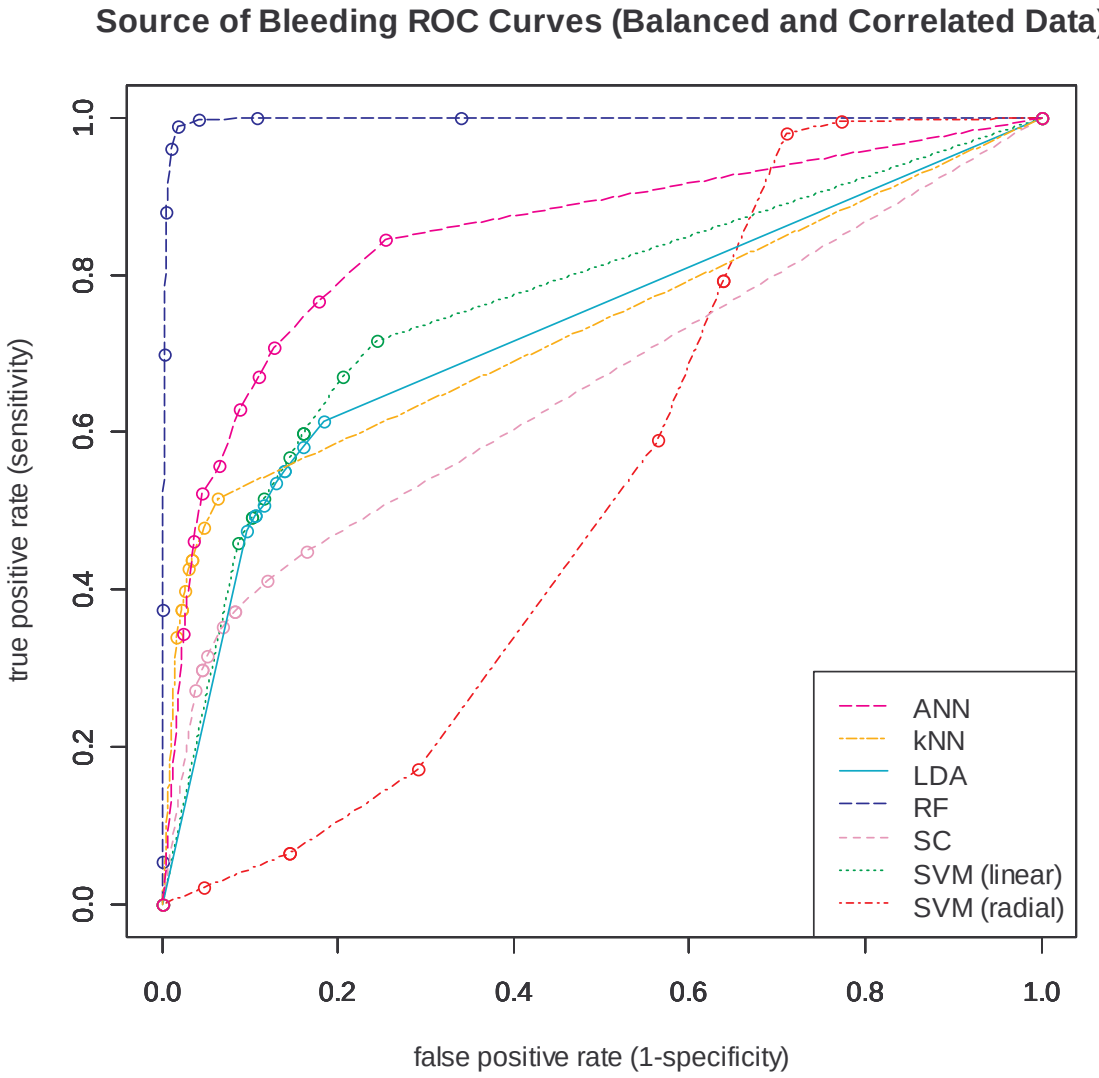


Figure 22. Simulation Study ROC Curves for Source of Bleeding Response (balanced and not correlated data)

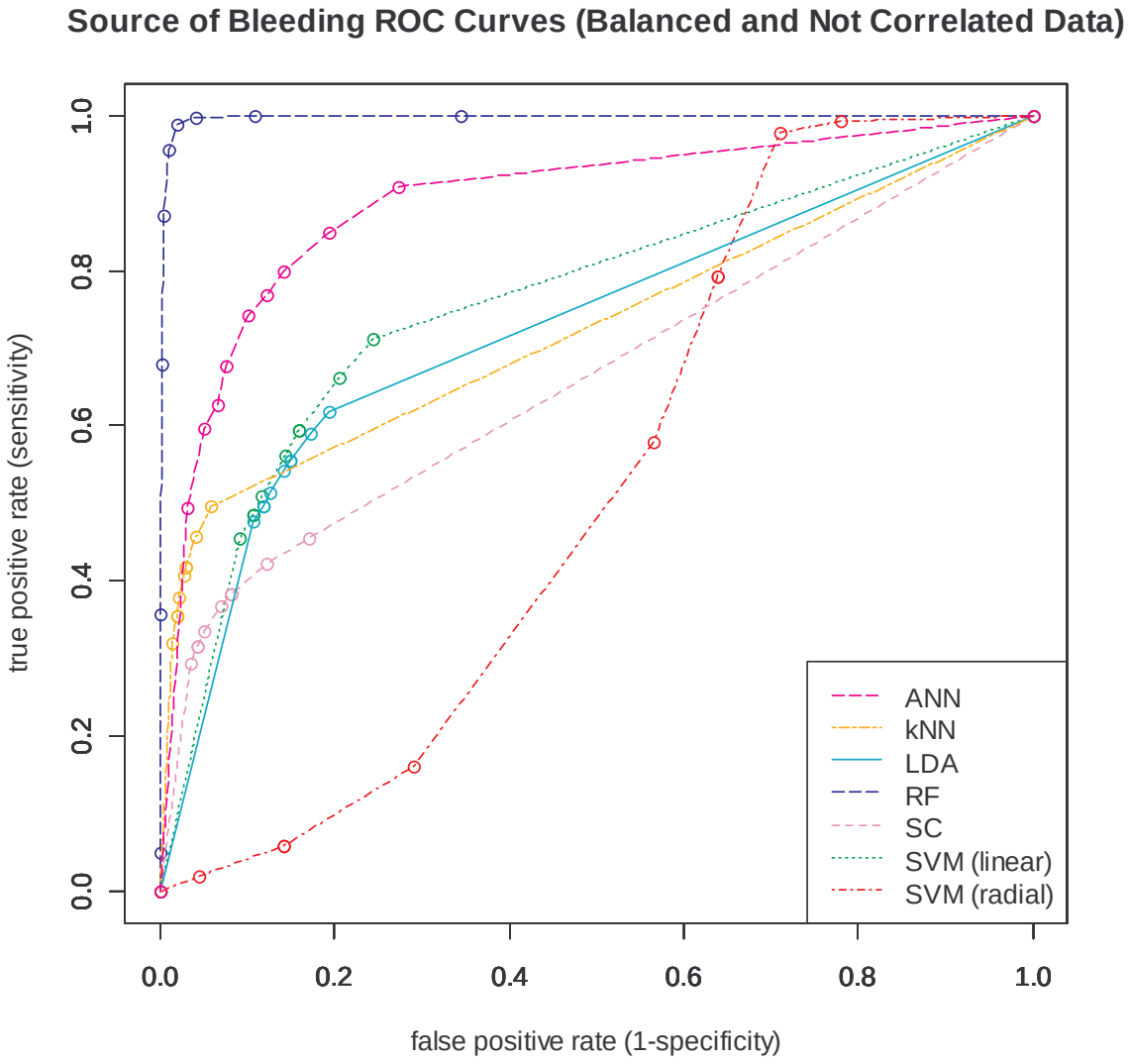


Figure 23. Simulation Study ROC Curves for Resuscitation Response (unbalanced and correlated data)

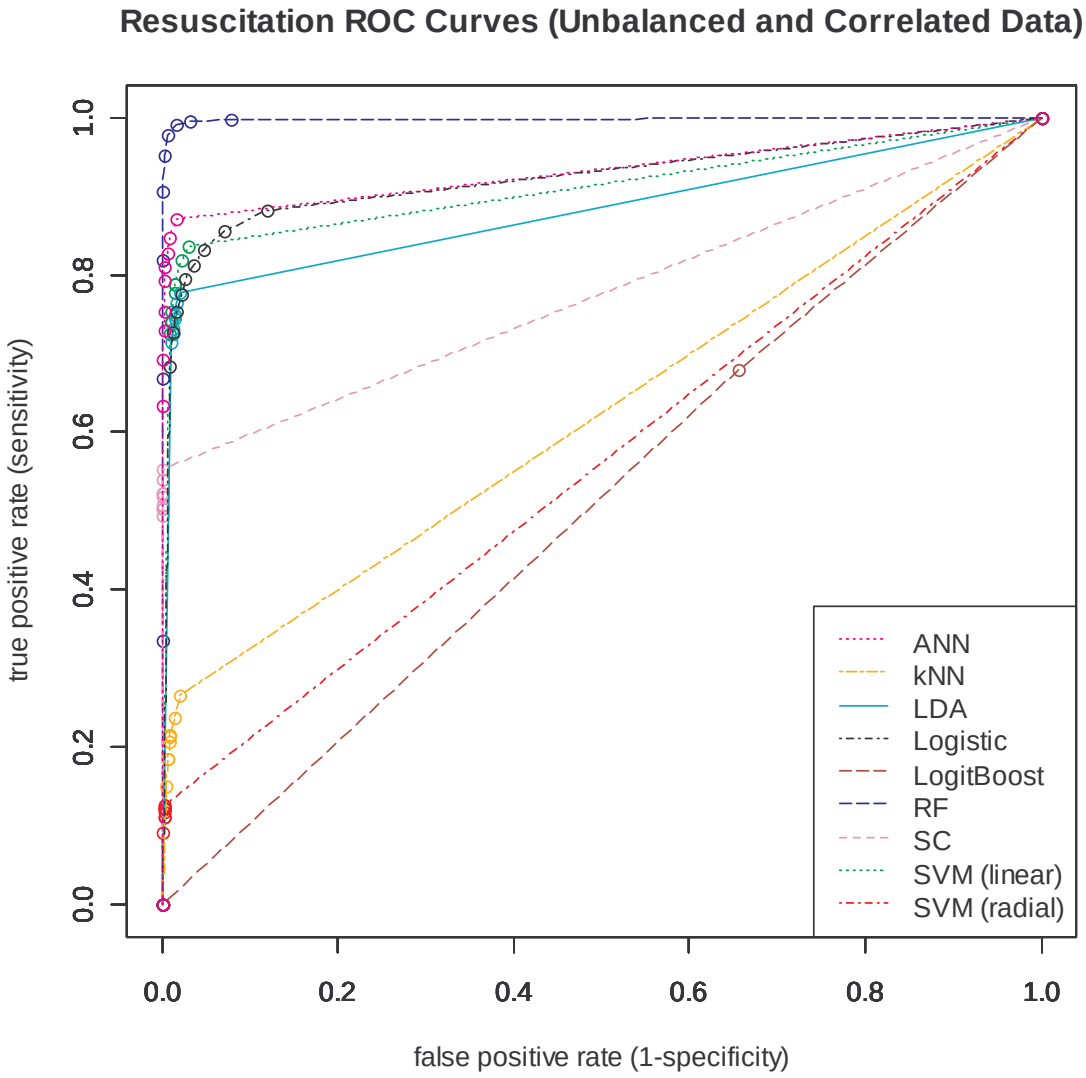


Figure 24. Simulation Study ROC Curves for Resuscitation Response (unbalanced and not correlated data)

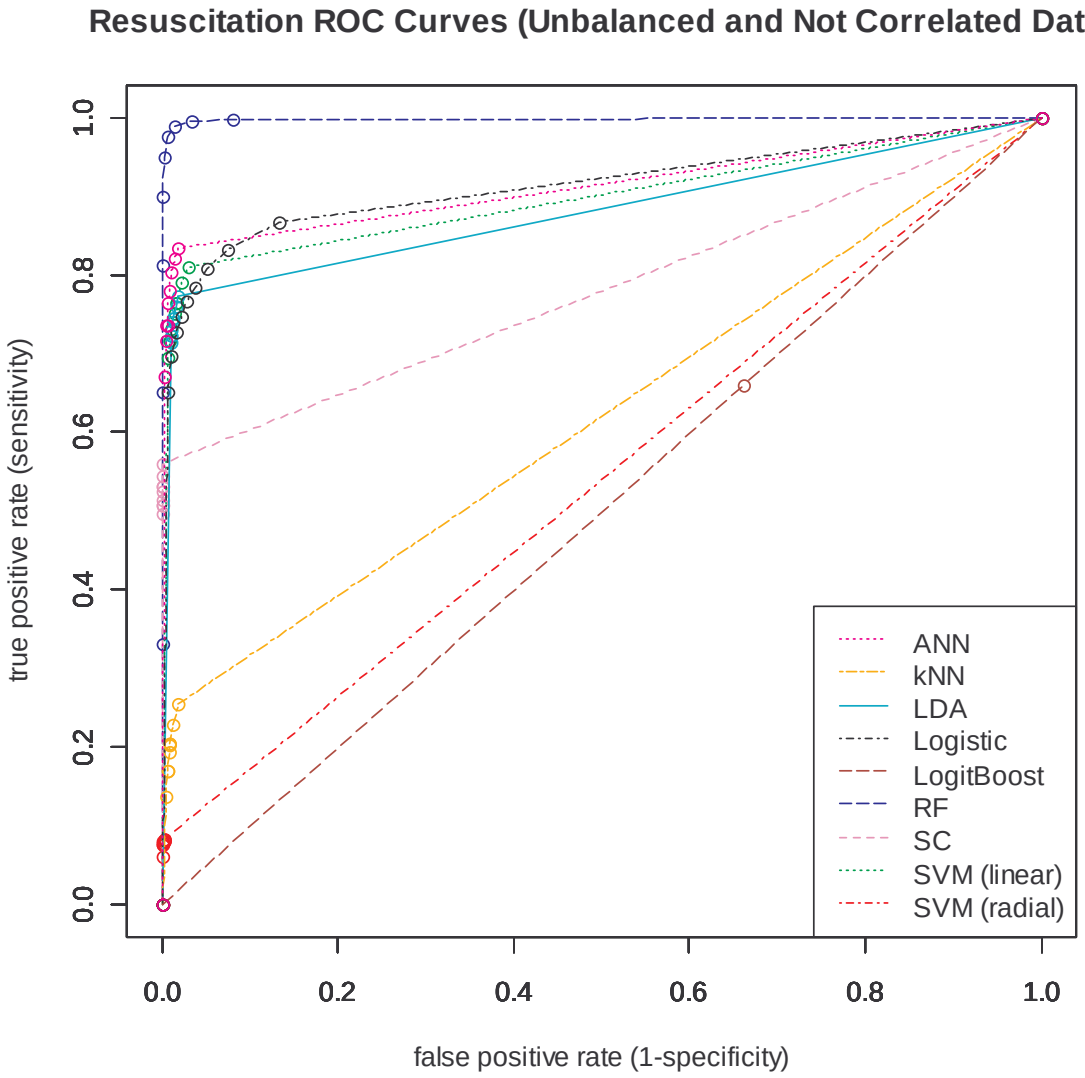


Figure 25. Simulation Study ROC Curves for Resuscitation Response (balanced and correlated data)

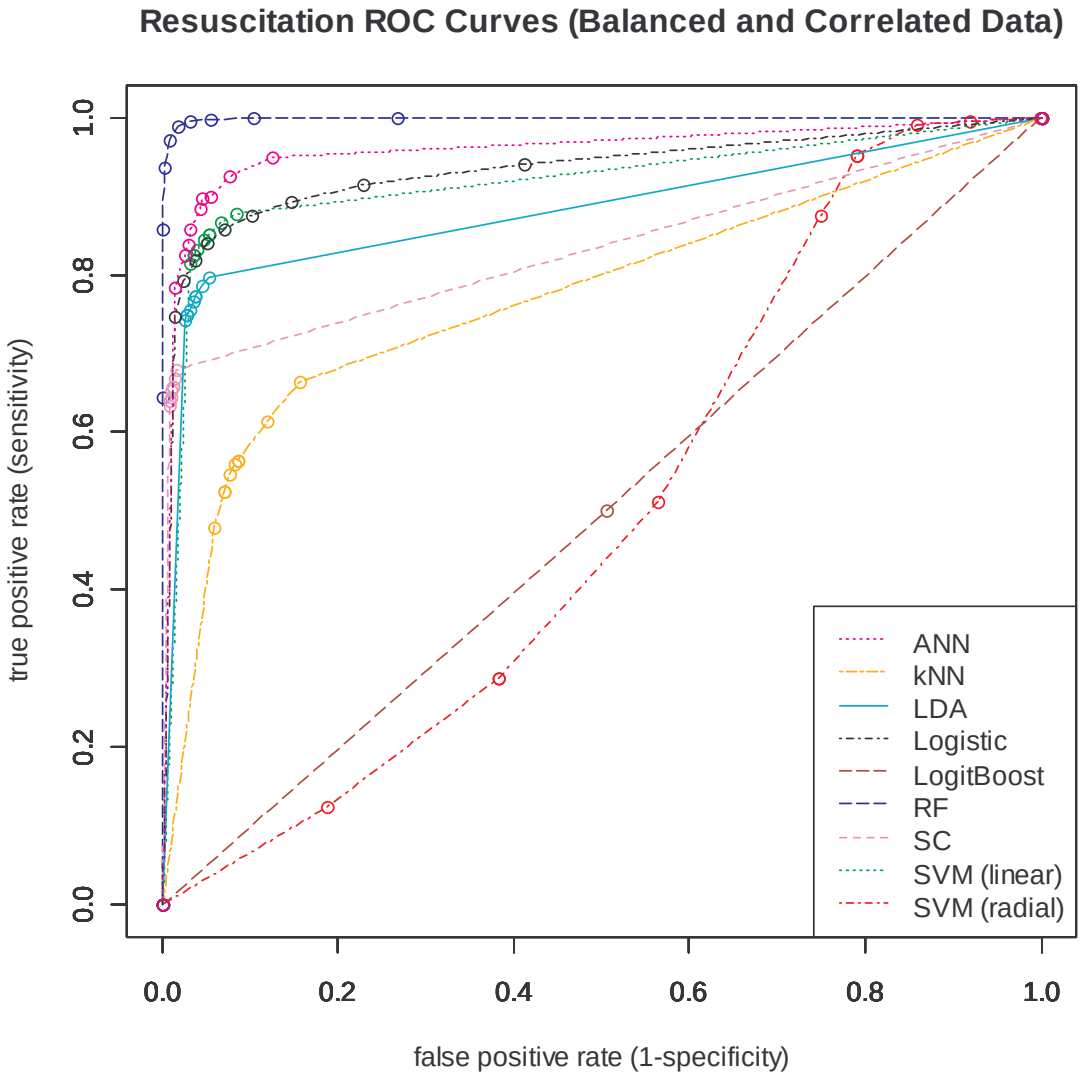


Figure 26. Simulation Study ROC Curves for Resuscitation Response (balanced and not correlated data)

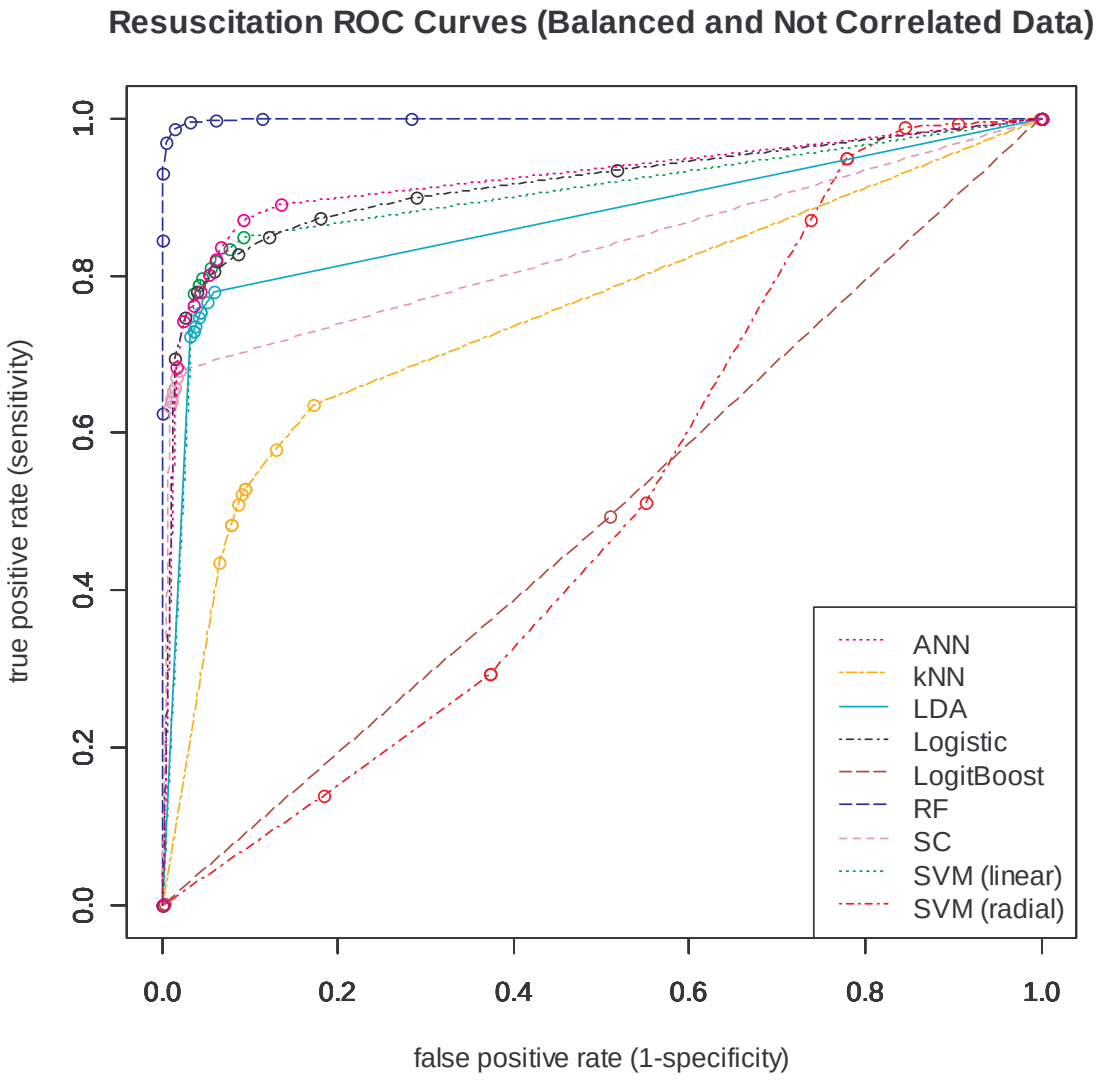


Figure 27. Simulation Study ROC Curves for Endoscopy Response (unbalanced and correlated data)

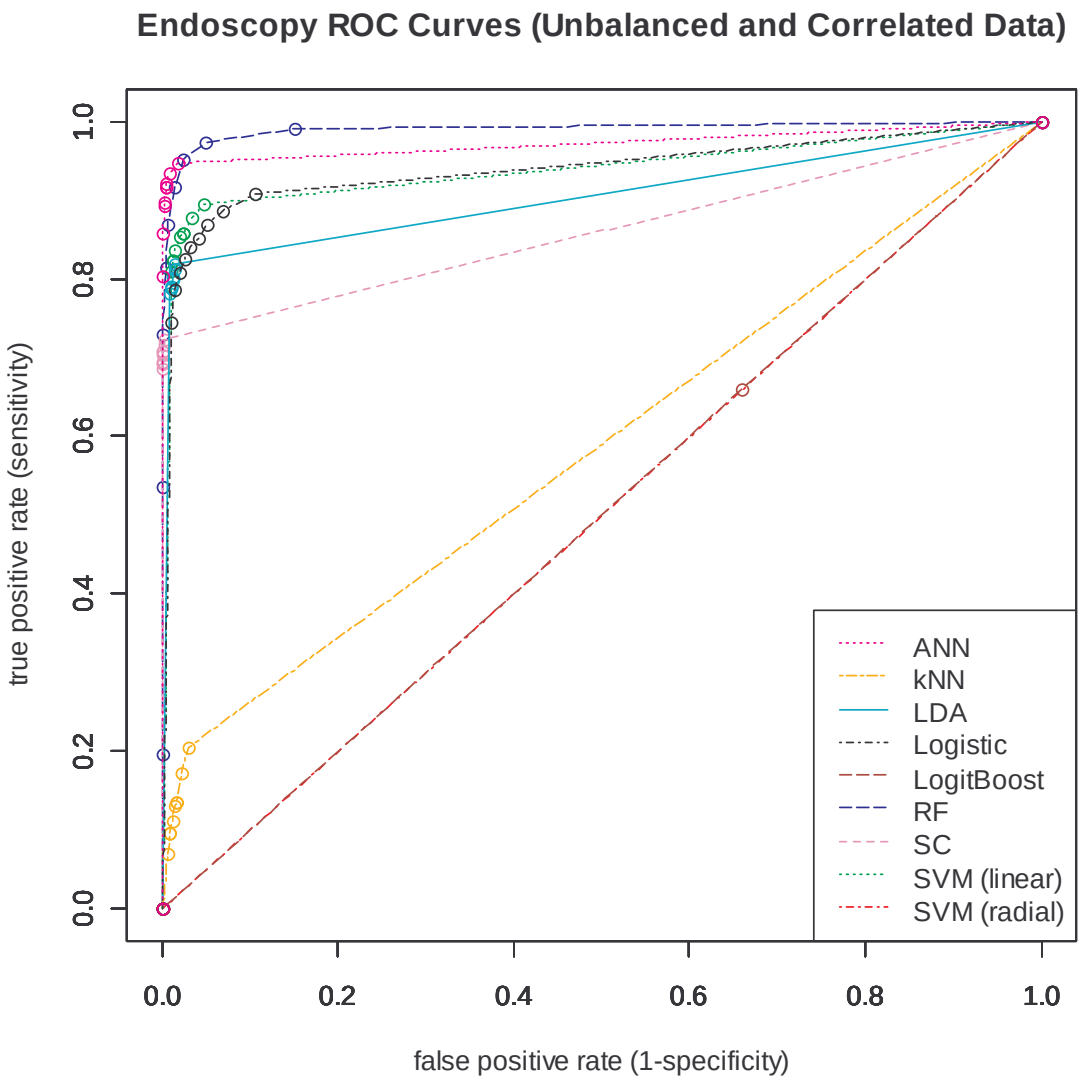


Figure 28. Simulation Study ROC Curves for Endoscopy Response (unbalanced and not correlated data)

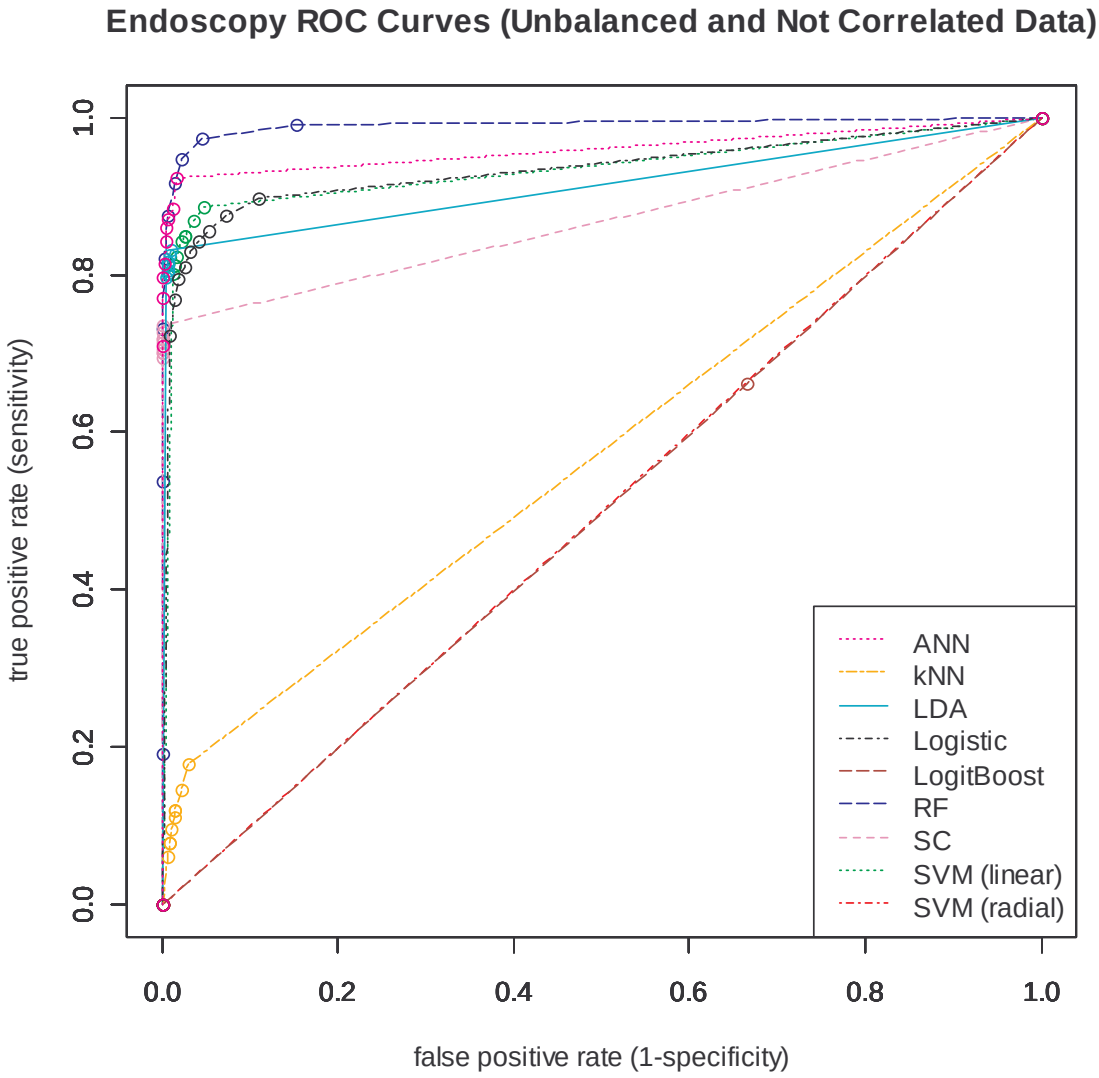


Figure 29. Simulation Study ROC Curves for Endoscopy Response (balanced and correlated data)

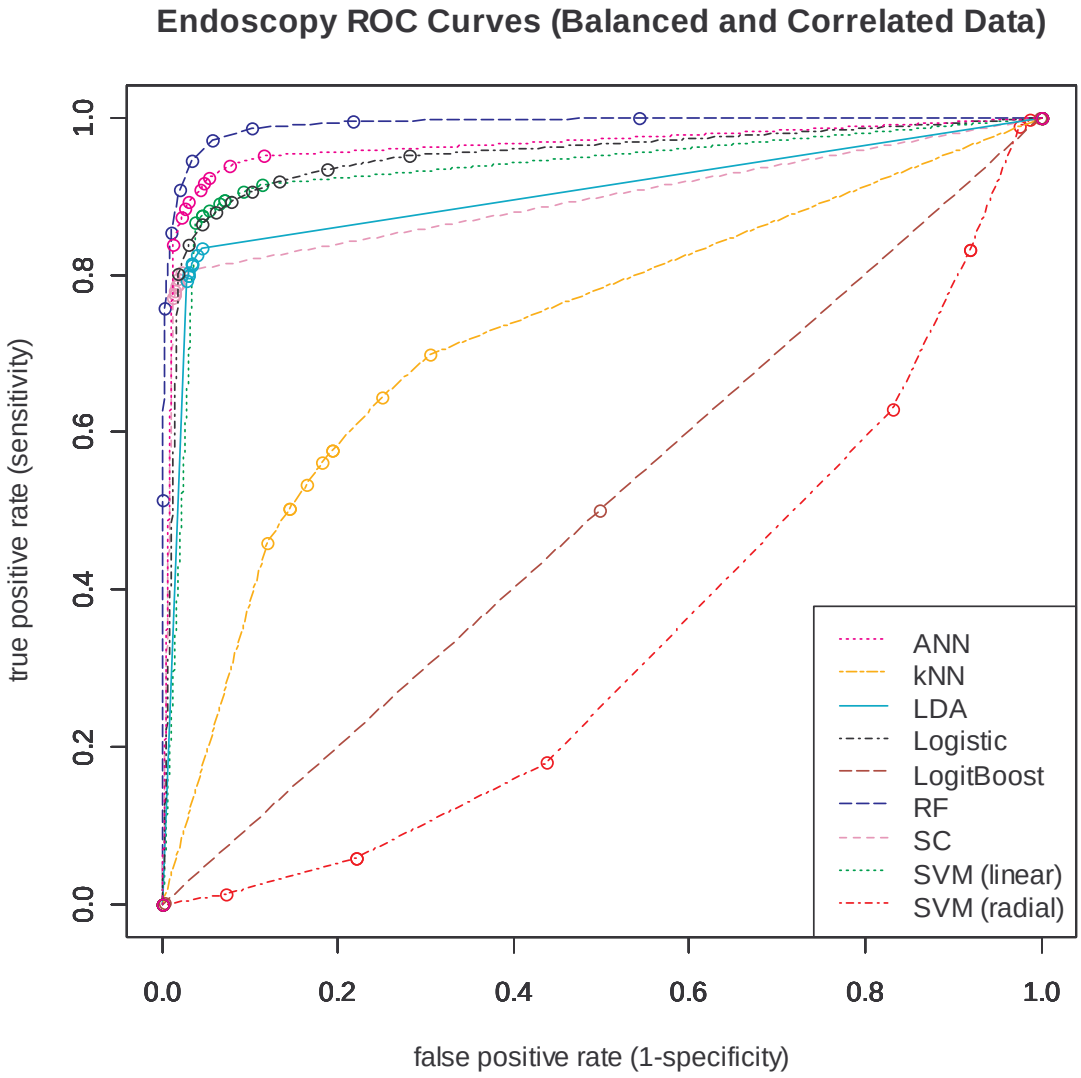


Figure 30. Simulation Study ROC Curves for Endoscopy Response (balanced and not correlated data)

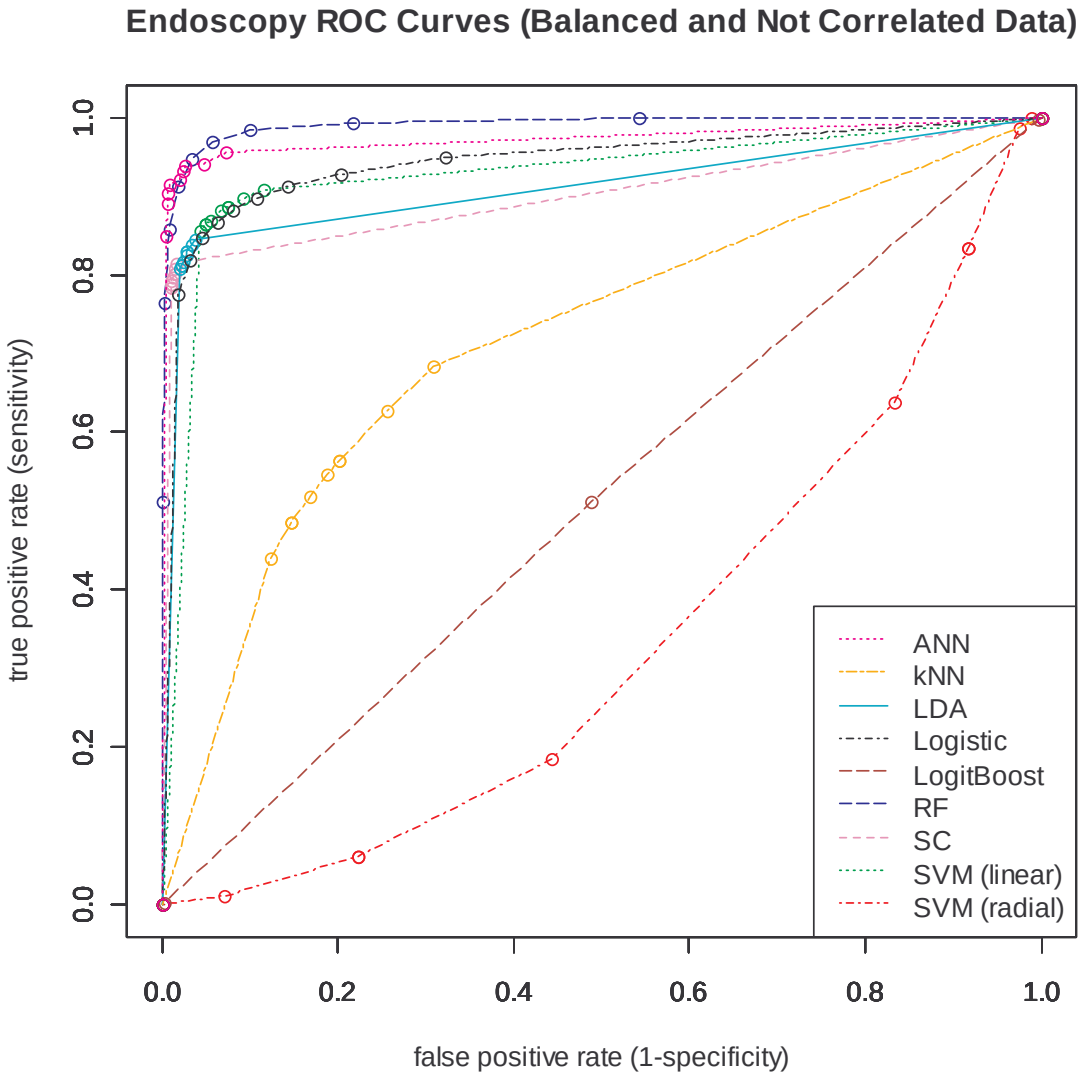


Figure 31. Simulation Study ROC Curves for Disposition Response (unbalanced and correlated data)

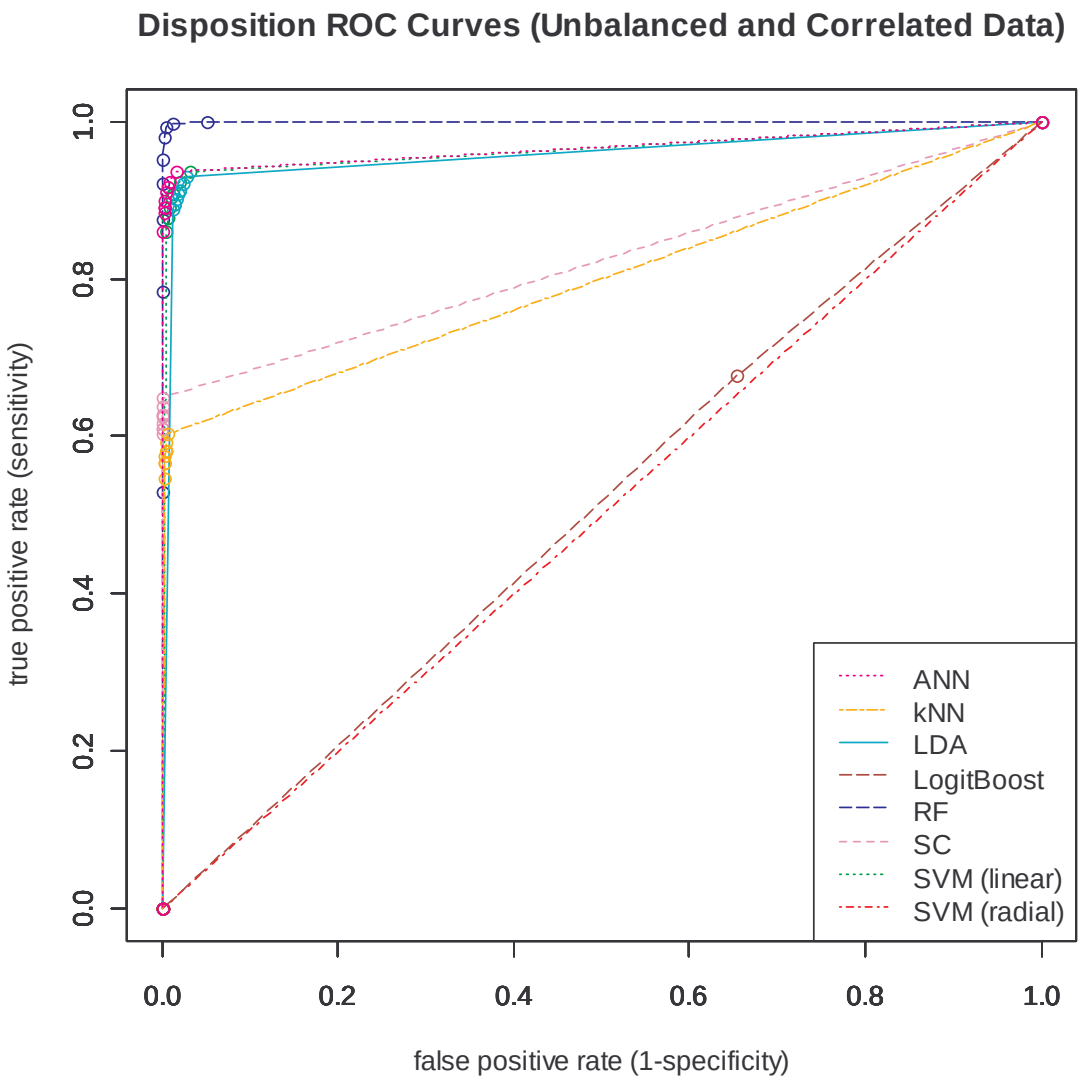


Figure 32. Simulation Study ROC Curves for Disposition Response (unbalanced and not correlated data)

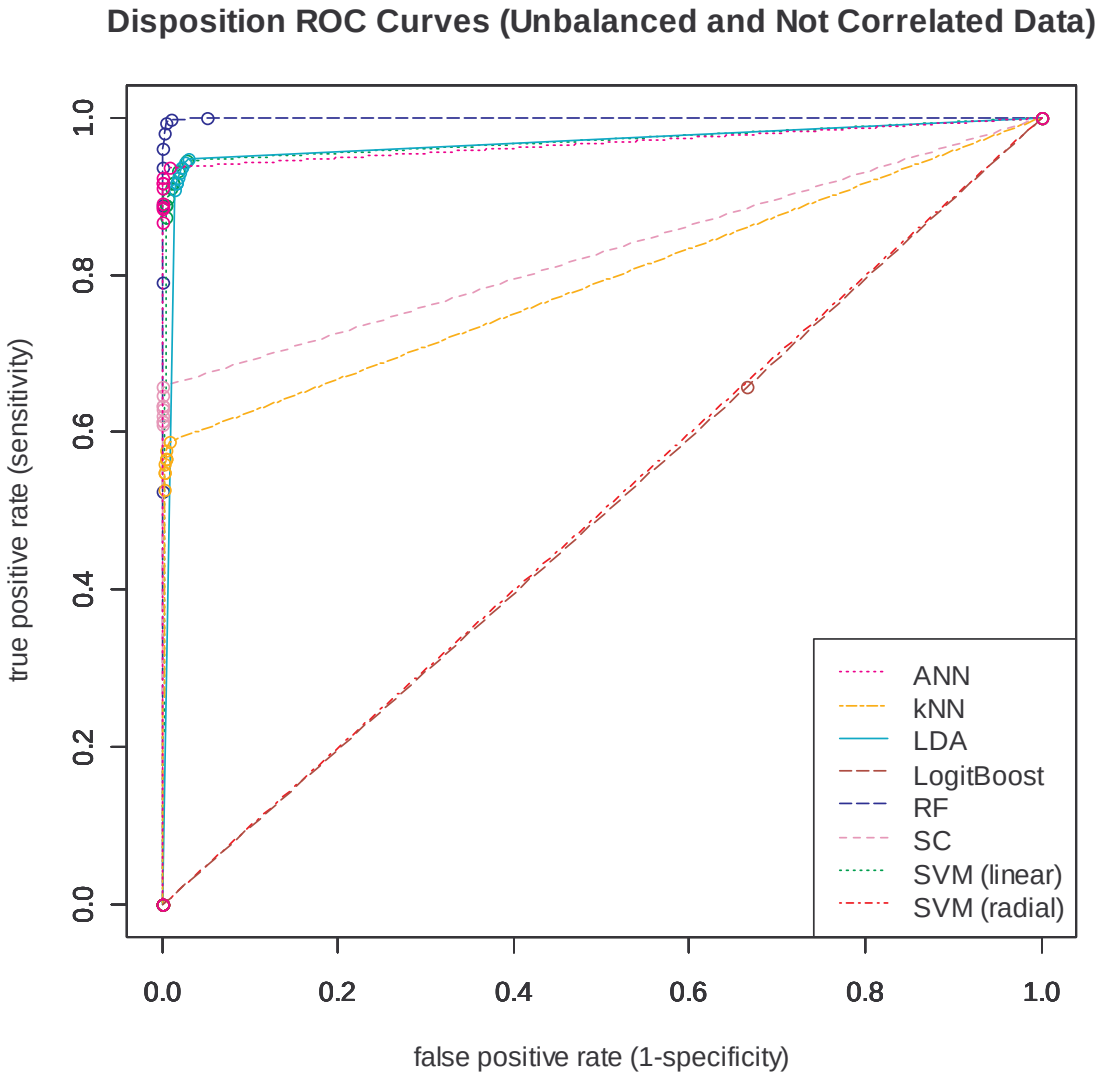


Figure 33. Simulation Study ROC Curves for Disposition Response (balanced and correlated data)

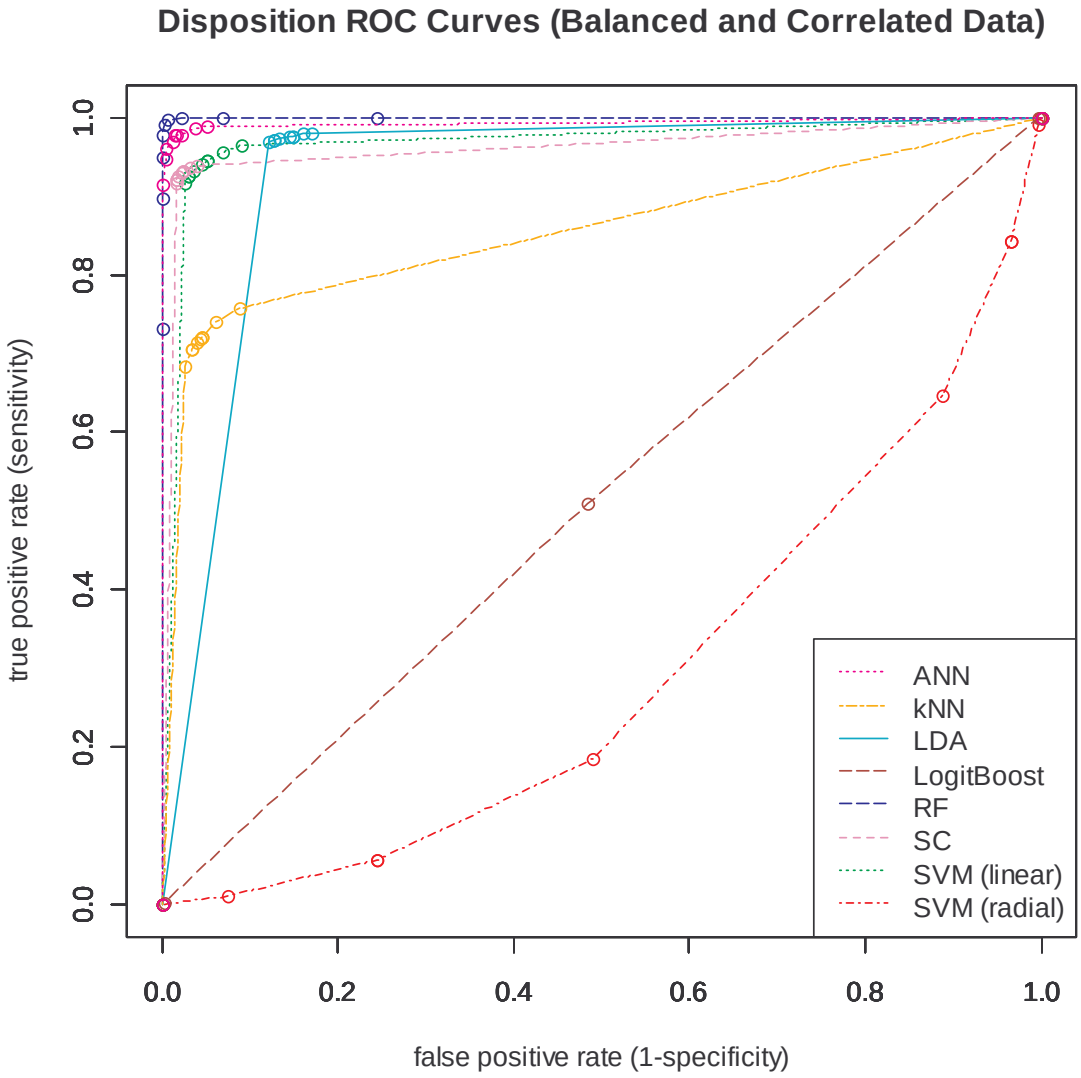
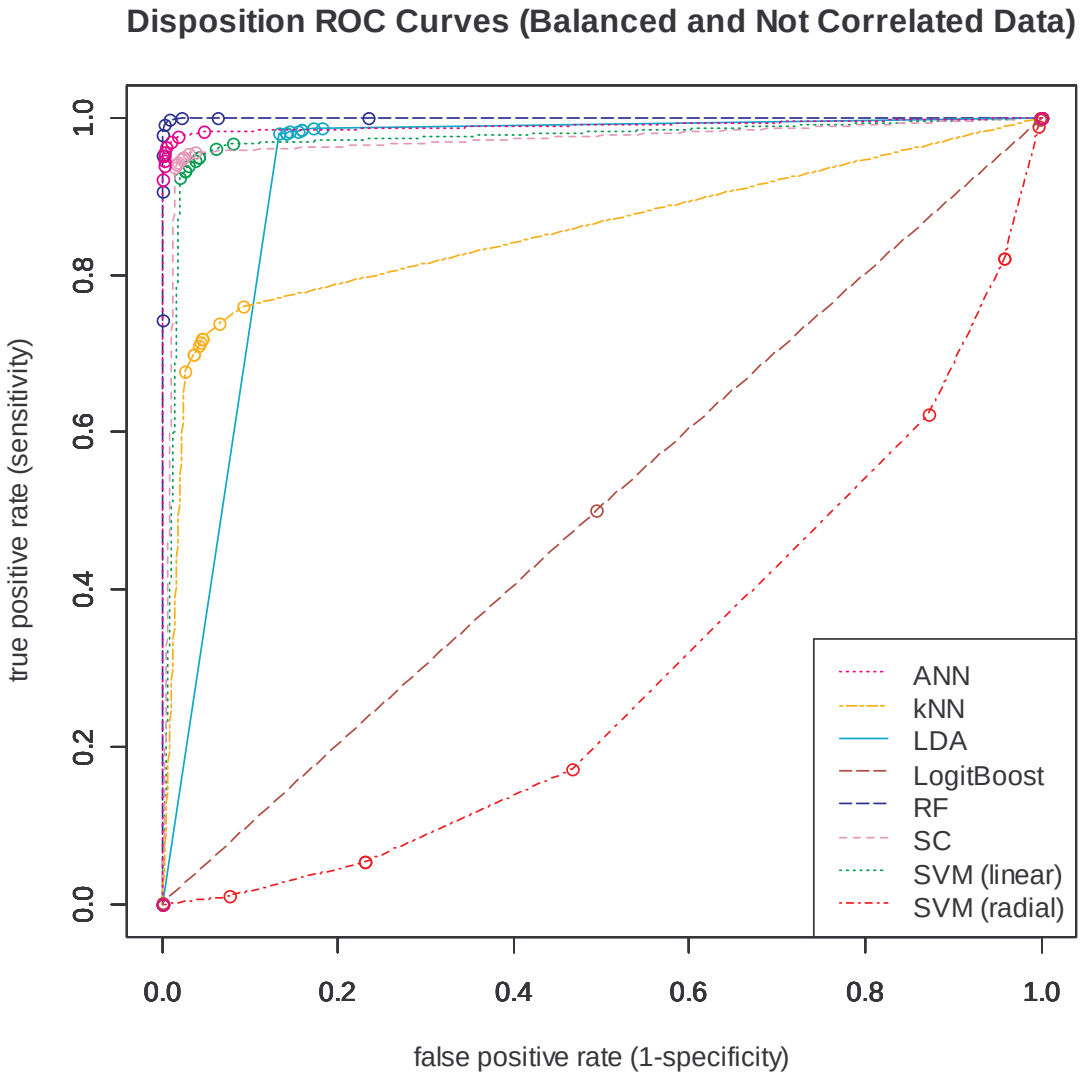


Figure 34. Simulation Study ROC curves for Disposition Response (balanced and not correlated data)



As seen in Section 2 with the real GIB data analysis, random forest performs the best overall out of all the models. With the simulated data, random forest significantly outperforms all the other models, having by far the highest accuracies and highest AUCs for all responses and all combinations of data. We see that most of the other models follow a similar trend with a few differences compared to that with the real GIB data analysis. For source of bleeding, kNN was one of the models that performed the worst. kNN may have performed poorly due to neighbors not being "nearby" the given test case. There have already been improvements on the kNN model to consider weighting the neighbors differently to take into account how far away they are from the test case (44). LDA, which had performed well prior, did not do well in predicting source of bleeding.

Boosting was seen to perform well with unbalanced data, but performed poorly with balanced data. Poor performance of the boosting model could be attributed to the training set being overfitted. Examples which were just noise could have been over emphasized (45). Logistic regression for disposition performed poorly as observed previously since the algorithm did not converge (results not reported for disposition response). A similar pattern was observed with SVM using the radial kernel – it performed well with unbalanced data but poorly with balanced data. Attempt to change the default values for the parameters epsilon and tolerance only showed significant improvement when the tolerance was set to 0.1 for the source of bleeding and resuscitation response (only balanced and correlated data were tried). However, some might consider this too large of a tolerance. See Tables 44-47.

Table 44. Accuracies Using Different Values for Epsilon and Tolerance for SVM – Radial (Source of Bleeding, balanced and correlated)

		Epsilon			
		0.1	**0.15**	**0.2**	**0.25**
Tolerance	**0.001**	0.41	0.44	0.46	0.44
	0.01	0.47	0.46	0.46	0.46
	0.1	0.49	0.46	0.42	0.44

Table 45. Accuracies Using Different Values for Epsilon and Tolerance for SVM – Radial (Resuscitation, balanced and correlated)

		Epsilon			
		0.1	**0.15**	**0.2**	**0.25**
Tolerance	**0.001**	0.53	0.57	0.55	0.49
	0.01	0.51	0.54	0.48	0.52
	0.1	0.50	0.61	0.49	0.50

Table 46. Accuracies Using Different Values for Epsilon and Tolerance for SVM – Radial (Endoscopy, balanced and correlated)

		Epsilon			
		0.1	**0.15**	**0.2**	**0.25**
Tolerance	**0.001**	0.46	0.45	0.50	0.47
	0.01	0.47	0.45	0.46	0.46
	0.1	0.48	0.45	0.41	0.45

Table 47. Accuracies Using Different Values for Epsilon and Tolerance for SVM – Radial (Disposition, balanced and correlated)

		Epsilon			
		0.1	**0.15**	**0.2**	**0.25**
Tolerance	**0.001**	0.44	0.44	0.44	0.40
	0.01	0.45	0.43	0.43	0.46
	0.1	0.47	0.46	0.43	0.43

Performance of SVMs can be data dependent, so for this particular balanced simulated GIB data, a line rather than a non-linear curve might separate the data better. Perhaps by using the radial kernel, the data becomes very spread out and sparse when transformed to a higher-dimensional space. Having sparse data would make it more difficult to classify new test cases (46). Both boosting and SVM (radial kernel) had a poor balance between sensitivity and specificity with unbalanced data. As noted with the actual GIB data analysis, logistic regression was not reliable in predicting the disposition response. With the simulated data, we see that it yielded very poor results. In agreement with the real GIB data analyses, the models with the best AUC were RF, LDA, SVM (linear kernel), and ANN.

The results for the data generated using a learning and test set (no cross-validation) are in Tables 48-63. Same or similar results were seen for the learning/test set data as the 10-fold cross validated data. There were no statistically significant differences except for the SVM – radial model for balanced data. This showed a significant improvement when using only the learning and test set. This is most likely attributed to random chance. For the remainder of the dissertation, the method used will be cross-validation.

Table 48. Simulation Study Results for Source of Bleeding Response (unbalanced and correlated data, learning/test set) (standard error)

	ACC	SN	SP	PPV	NPV
ANN[a]	0.815 (0.002)	0.963 (0.001)	0.268 (0.003)	0.841 (0.002)	NaN[b]
kNN	0.793 (0.002)	0.842 (0.002)	0.620 (0.003)	0.899 (0.002)	0.498 (0.003)
LDA	0.659 (0.003)	0.655 (0.003)	0.800 (0.002)	0.930 (0.001)	0.370 (0.003)
RF	0.960 (0.001)	0.998 (0.000)	0.808 (0.002)	0.954 (0.001)	0.993 (0.000)
SC	0.800 (0.002)	1.000 (0.000)	0.000 (0.000)	0.800 (0.002)	0.020 (0.000)
SVM – linear	0.809 (0.002)	0.954 (0.001)	0.240 (0.002)	0.835 (0.002)	0.593 (0.003)
SVM – radial	0.800 (0.002)	0.999 (0.000)	0.006 (0.000)	0.801 (0.002)	0.057 (0.001)

[a] Using only 5 files with subsampling
[b] Output produced NaN (Not a Number)

Table 49. Simulation Study Results for Source of Bleeding Response (unbalanced and not correlated data, learning/test set) (standard error)

	ACC	SN	SP	PPV	NPV
ANN[a]	0.821 (0.002)	0.968 (0.001)	0.244 (0.002)	0.841 (0.002)	NaN[b]
kNN	0.789 (0.002)	0.858 (0.002)	0.536 (0.003)	0.881 (0.002)	0.489 (0.003)
LDA	0.694 (0.003)	0.705 (0.003)	0.762 (0.002)	0.923 (0.002)	0.395 (0.003)
RF	0.959 (0.001)	0.999 (0.000)	0.802 (0.002)	0.953 (0.001)	0.994 (0.000)
SC	0.800 (0.002)	1.000 (0.000)	0.001 (0.000)	0.800 (0.002)	0.018 (0.001)
SVM – linear	0.812 (0.002)	0.961 (0.001)	0.225 (0.002)	0.833 (0.002)	0.606 (0.003)
SVM – radial	0.802 (0.002)	0.999 (0.000)	0.016 (0.000)	0.802 (0.002)	0.086 (0.002)

[a] Using only 5 files with subsampling
[b] Output produced NaN (Not a Number)

Table 50. Simulation Study Results for Source of Bleeding Response (balanced and correlated data, learning/test set) (standard error)

	ACC	SN	SP	PPV	NPV
ANN[a]	0.721 (0.003)	0.582 (0.003)	0.858 (0.002)	0.690 (0.003)	0.809 (0.002)
kNN	0.796 (0.002)	0.489 (0.003)	0.965 (0.001)	0.876 (0.002)	0.791 (0.002)
LDA	0.701 (0.003)	0.531 (0.003)	0.870 (0.002)	0.677 (0.003)	0.788 (0.002)
RF	0.984 (0.001)	0.978 (0.001)	0.987 (0.001)	0.974 (0.001)	0.989 (0.001)
SC	0.722 (0.003)	0.363 (0.003)	0.932 (0.001)	0.753 (0.002)	0.746 (0.003)
SVM – linear	0.734 (0.003)	0.542 (0.003)	0.860 (0.002)	0.667 (0.003)	0.790 (0.002)
SVM – radial	0.567 (0.003)	0.998 (0.000)	0.351 (0.003)	0.435 (0.003)	0.997 (0.000)

[a] Using only 5 files with subsampling

Table 51. Simulation Study Results for Source of Bleeding Response (balanced and not correlated data, learning/test set) (standard error)

	ACC	SN	SP	PPV	NPV
ANN[a]	0.719 (0.003)	0.613 (0.003)	0.843 (0.002)	0.671 (0.003)	0.814 (0.002)
kNN	0.769 (0.002)	0.404 (0.003)	0.973 (0.001)	0.885 (0.002)	0.766 (0.002)
LDA	0.709 (0.003)	0.534 (0.003)	0.868 (0.002)	0.672 (0.003)	0.789 (0.002)
RF	0.983 (0.001)	0.977 (0.001)	0.986 (0.001)	0.973 (0.001)	0.988 (0.001)
SC	0.726 (0.003)	0.346 (0.003)	0.945 (0.001)	0.784 (0.002)	0.744 (0.003)
SVM – linear	0.743 (0.003)	0.561 (0.003)	0.859 (0.002)	0.669 (0.003)	0.797 (0.002)
SVM – radial	0.574 (0.003)	0.989 (0.001)	0.368 (0.003)	0.445 (0.003)	0.993 (0.000)

[a] Using only 5 files with subsampling

Table 52. Simulation Study Results for Resuscitation Response (unbalanced and correlated data, learning/test set) (standard error)

	ACC	SN	SP	PPV	NPV
ANN[a]	0.933 (0.001)	0.749 (0.003)	0.981 (0.001)	0.912 (0.002)	0.939 (0.001)
kNN	0.833 (0.002)	0.202 (0.002)	0.991 (0.001)	0.849 (0.002)	0.832 (0.002)
LDA	0.938 (0.001)	0.744 (0.003)	0.987 (0.001)	0.935 (0.001)	0.939 (0.001)
Logistic	0.930 (0.001)	0.779 (0.002)	0.968 (0.001)	0.865 (0.002)	0.946 (0.001)
LogitBoost	0.800 (0.002)	0.000 (0.000)	1.000 (0.000)	0.000 (0.000)	0.800 (0.002)
RF	0.990 (0.001)	0.957 (0.001)	0.998 (0.000)	0.993 (0.000)	0.989 (0.001)
SC	0.906 (0.002)	0.528 (0.003)	1.000 (0.000)	1.000 (0.000)	0.895 (0.002)
SVM – linear	0.937 (0.001)	0.759 (0.002)	0.982 (0.001)	0.918 (0.002)	0.942 (0.001)
SVM – radial	0.822 (0.002)	0.115 (0.002)	0.998 (0.000)	0.133 (0.002)	0.822 (0.002)

[a] Using only 5 files with subsampling

Table 53. Simulation Study Results for Resuscitation Response (unbalanced and not correlated data, learning/test set) (standard error)

	ACC	SN	SP	PPV	NPV
ANN[a]	0.925 (0.002)	0.716 (0.003)	0.978 (0.001)	0.895 (0.002)	0.932 (0.001)
kNN	0.820 (0.002)	0.119 (0.002)	0.996 (0.000)	0.893 (0.002)	0.819 (0.002)
LDA	0.935 (0.001)	0.735 (0.003)	0.985 (0.001)	0.927 (0.002)	0.937 (0.001)
Logistic	0.926 (0.002)	0.768 (0.002)	0.965 (0.001)	0.854 (0.002)	0.943 (0.001)
LogitBoost	0.800 (0.002)	0.000 (0.000)	1.000 (0.000)	0.000 (0.000)	0.800 (0.002)
RF	0.989 (0.001)	0.956 (0.001)	0.998 (0.000)	0.992 (0.001)	0.989 (0.001)
SC	0.907 (0.002)	0.536 (0.003)	1.000 (0.000)	1.000 (0.000)	0.896 (0.002)
SVM – linear	0.935 (0.001)	0.750 (0.003)	0.981 (0.001)	0.912 (0.002)	0.940 (0.001)
SVM – radial	0.816 (0.002)	0.089 (0.002)	0.998 (0.000)	0.101 (0.002)	0.817 (0.002)

[a] Using only 5 files with subsampling

Table 54. Simulation Study Results for Resuscitation Response (balanced and correlated data, learning/test set) (standard error)

	ACC	SN	SP	PPV	NPV
ANN[a]	0.875 (0.002)	0.836 (0.002)	0.915 (0.002)	0.913 (0.002)	0.850 (0.002)
kNN	0.726 (0.003)	0.541 (0.003)	0.911 (0.002)	0.860 (0.002)	0.666 (0.003)
LDA	0.859 (0.002)	0.758 (0.002)	0.959 (0.001)	0.950 (0.001)	0.800 (0.002)
Logistic	0.877 (0.002)	0.839 (0.002)	0.915 (0.002)	0.909 (0.002)	0.851 (0.002)
LogitBoost	0.501 (0.003)	0.505 (0.003)	0.498 (0.003)	0.502 (0.003)	0.500 (0.003)
RF	0.983 (0.001)	0.986 (0.001)	0.981 (0.001)	0.981 (0.001)	0.986 (0.001)
SC	0.821 (0.002)	0.654 (0.003)	0.987 (0.001)	0.981 (0.001)	0.741 (0.003)
SVM – linear	0.881 (0.002)	0.820 (0.002)	0.942 (0.001)	0.935 (0.001)	0.840 (0.002)
SVM – radial	0.634 (0.003)	0.982 (0.001)	0.286 (0.003)	0.595 (0.003)	0.972 (0.001)

[a] Using only 5 files with subsampling

Table 55. Simulation Study Results for Resuscitation Response (balanced and not correlated data, learning/test set) (standard error)

	ACC	SN	SP	PPV	NPV
ANN[a]	0.863 (0.002)	0.814 (0.002)	0.913 (0.002)	0.908 (0.002)	0.833 (0.002)
kNN	0.701 (0.003)	0.480 (0.003)	0.923 (0.002)	0.862 (0.002)	0.640 (0.003)
LDA	0.851 (0.002)	0.745 (0.003)	0.958 (0.001)	0.948 (0.001)	0.790 (0.002)
Logistic	0.873 (0.002)	0.830 (0.002)	0.916 (0.002)	0.909 (0.002)	0.844 (0.002)
LogitBoost	0.500 (0.003)	0.488 (0.003)	0.512 (0.003)	0.499 (0.003)	0.501 (0.003)
RF	0.984 (0.001)	0.986 (0.001)	0.982 (0.001)	0.982 (0.001)	0.986 (0.001)
SC	0.820 (0.002)	0.651 (0.003)	0.990 (0.001)	0.985 (0.001)	0.740 (0.003)
SVM – linear	0.877 (0.002)	0.811 (0.002)	0.944 (0.001)	0.937 (0.001)	0.834 (0.002)
SVM – radial	0.630 (0.003)	0.985 (0.001)	0.275 (0.003)	0.589 (0.003)	0.969 (0.001)

[a] Using only 5 files with subsampling

Table 56. Simulation Study Results for Endoscopy Response (unbalanced and correlated data, learning/test set) (standard error)

	ACC	SN	SP	PPV	NPV
ANN[a]	0.946 (0.001)	0.820 (0.002)	0.978 (0.001)	0.923 (0.002)	0.956 (0.001)
kNN	0.813 (0.002)	0.128 (0.002)	0.984 (0.001)	0.678 (0.003)	0.819 (0.002)
LDA	0.949 (0.001)	0.798 (0.002)	0.987 (0.001)	0.942 (0.001)	0.951 (0.001)
Logistic	0.930 (0.001)	0.828 (0.002)	0.955 (0.001)	0.832 (0.002)	0.957 (0.001)
LogitBoost	0.800 (0.002)	0.000 (0.000)	1.000 (0.000)	0.000 (0.000)	0.800 (0.002)
RF	0.968 (0.001)	0.873 (0.002)	0.991 (0.001)	0.964 (0.001)	0.969 (0.001)
SC	0.940 (0.001)	0.706 (0.003)	0.999 (0.000)	0.992 (0.001)	0.932 (0.001)
SVM – linear	0.943 (0.001)	0.835 (0.002)	0.970 (0.001)	0.881 (0.002)	0.959 (0.001)
SVM – radial	0.800 (0.002)	0.000 (0.000)	1.000 (0.000)	0.000 (0.000)	0.800 (0.002)

[a] Using only 5 files with subsampling

Table 57. Simulation Study Results for Endoscopy Response (unbalanced and not correlated data, learning/test set) (standard error)

	ACC	SN	SP	PPV	NPV
ANN[a]	0.946 (0.001)	0.814 (0.002)	0.979 (0.001)	0.917 (0.002)	0.955 (0.001)
kNN	0.810 (0.002)	0.105 (0.002)	0.986 (0.001)	0.666 (0.003)	0.815 (0.002)
LDA	0.956 (0.001)	0.814 (0.002)	0.992 (0.001)	0.963 (0.001)	0.815 (0.002)
Logistic	0.939 (0.001)	0.817 (0.002)	0.970 (0.001)	0.878 (0.002)	0.955 (0.001)
LogitBoost	0.800 (0.002)	0.000 (0.000)	1.000 (0.000)	0.000 (0.000)	0.800 (0.002)
RF	0.970 (0.001)	0.876 (0.002)	0.993 (0.000)	0.971 (0.001)	0.970 (0.001)
SC	0.941 (0.001)	0.709 (0.003)	0.999 (0.000)	0.997 (0.000)	0.932 (0.000)
SVM – linear	0.948 (0.001)	0.824 (0.002)	0.979 (0.001)	0.911 (0.002)	0.957 (0.001)
SVM – radial	0.800 (0.002)	0.000 (0.000)	1.000 (0.000)	0.000 (0.000)	0.800 (0.002)

[a] Using only 5 files with subsampling

Table 58. Simulation Study Results for Endoscopy Response (balanced and correlated data, learning/test set) (standard error)

	ACC	SN	SP	PPV	NPV
ANN[a]	0.903 (0.002)	0.881 (0.002)	0.926 (0.002)	0.926 (0.002)	0.888 (0.002)
kNN	0.691 (0.003)	0.568 (0.003)	0.815 (0.002)	0.756 (0.002)	0.654 (0.003)
LDA	0.883 (0.002)	0.806 (0.002)	0.959 (0.001)	0.953 (0.001)	0.833 (0.002)
Logistic	0.888 (0.002)	0.879 (0.002)	0.898 (0.002)	0.898 (0.002)	0.882 (0.002)
LogitBoost	0.491 (0.003)	0.493 (0.003)	0.488 (0.003)	0.494 (0.003)	0.487 (0.003)
RF	0.952 (0.001)	0.949 (0.001)	0.956 (0.001)	0.956 (0.001)	0.949 (0.001)
SC	0.885 (0.002)	0.785 (0.002)	0.985 (0.001)	0.981 (0.001)	0.822 (0.002)
SVM – linear	0.893 (0.002)	0.874 (0.002)	0.912 (0.002)	0.910 (0.002)	0.879 (0.002)
SVM – radial	0.557 (0.003)	0.968 (0.001)	0.146 (0.002)	0.534 (0.003)	0.904 (0.002)

[a] Using only 5 files with subsampling

Table 59. Simulation Study Results for Endoscopy Response (balanced and not correlated data, learning/test set) (standard error)

	ACC	SN	SP	PPV	NPV
ANN[a]	0.892 (0.002)	0.872 (0.002)	0.913 (0.002)	0.913 (0.002)	0.881 (0.002)
kNN	0.676 (0.003)	0.544 (0.003)	0.808 (0.002)	0.741 (0.003)	0.640 (0.003)
LDA	0.896 (0.002)	0.823 (0.002)	0.969 (0.001)	0.964 (0.001)	0.847 (0.002)
Logistic	0.897 (0.002)	0.882 (0.002)	0.912 (0.002)	0.911 (0.002)	0.886 (0.002)
LogitBoost	0.520 (0.003)	0.515 (0.003)	0.526 (0.003)	0.522 (0.003)	0.519 (0.003)
RF	0.954 (0.001)	0.948 (0.001)	0.960 (0.001)	0.961 (0.001)	0.949 (0.001)
SC	0.892 (0.002)	0.797 (0.002)	0.987 (0.001)	0.984 (0.001)	0.830 (0.002)
SVM – linear	0.901 (0.002)	0.875 (0.002)	0.927 (0.002)	0.925 (0.002)	0.882 (0.002)
SVM – radial	0.555 (0.003)	0.985 (0.001)	0.126 (0.002)	0.530 (0.003)	0.895 (0.002)

[a] Using only 5 files with subsampling

Table 60. Simulation Study Results for Disposition Response (unbalanced and correlated data, learning/test set) (standard error)

	ACC	SN	SP	PPV	NPV
ANN[a]	0.972 (0.001)	0.887 (0.002)	0.994 (0.000)	0.975 (0.001)	0.972 (0.001)
kNN	0.914 (0.002)	0.580 (0.003)	0.997 (0.000)	0.980 (0.001)	0.905 (0.002)
LDA	0.966 (0.001)	0.915 (0.002)	0.978 (0.001)	0.919 (0.002)	0.979 (0.001)
LogitBoost	0.800 (0.002)	0.000 (0.000)	1.000 (0.000)	0.000 (0.000)	0.800 (0.002)
RF	0.991 (0.001)	0.957 (0.001)	1.000 (0.000)	1.000 (0.000)	0.989 (0.001)
SC	0.923 (0.002)	0.617 (0.003)	1.000 (0.000)	1.000 (0.000)	0.913 (0.002)
SVM – linear	0.965 (0.001)	0.889 (0.002)	0.983 (0.001)	0.934 (0.001)	0.973 (0.001)
SVM – radial	0.800 (0.002)	0.000 (0.000)	1.000 (0.000)	0.000 (0.000)	0.800 (0.002)

[a] Using only 5 files with subsampling

Table 61. Simulation Study Results for Disposition Response (unbalanced and not correlated data, learning/test set) (standard error)

	ACC	SN	SP	PPV	NPV
ANN[a]	0.977 (0.001)	0.901 (0.002)	0.996 (0.000)	0.981 (0.001)	0.976 (0.001)
kNN	0.904 (0.002)	0.523 (0.003)	1.000 (0.000)	0.997 (0.000)	0.894 (0.002)
LDA	0.967 (0.001)	0.935 (0.001)	0.974 (0.001)	0.905 (0.002)	0.984 (0.001)
LogitBoost	0.800 (0.002)	0.000 (0.000)	1.000 (0.000)	0.000 (0.000)	0.800 (0.002)
RF	0.992 (0.001)	0.960 (0.001)	1.000 (0.000)	1.000 (0.000)	0.990 (0.001)
SC	0.927 (0.002)	0.634 (0.003)	1.000 (0.000)	1.000 (0.000)	0.916 (0.002)
SVM – linear	0.973 (0.001)	0.900 (0.002)	0.991 (0.001)	0.962 (0.001)	0.976 (0.001)
SVM – radial	0.800 (0.002)	0.000 (0.000)	1.000 (0.000)	0.000 (0.000)	0.800 (0.002)

[a] Using only 5 files with subsampling

Table 62. Simulation Study Results for Disposition Response (balanced and correlated data, learning/test set) (standard error)

	ACC	SN	SP	PPV	NPV
ANN[a]	0.950 (0.001)	0.935 (0.001)	0.965 (0.001)	0.964 (0.001)	0.938 (0.001)
kNN	0.837 (0.002)	0.719 (0.003)	0.956 (0.001)	0.943 (0.001)	0.773 (0.002)
LDA	0.916 (0.002)	0.974 (0.001)	0.859 (0.002)	0.876 (0.002)	0.971 (0.001)
LogitBoost	0.519 (0.003)	0.519 (0.003)	0.519 (0.003)	0.521 (0.003)	0.517 (0.003)
RF	0.994 (0.000)	0.991 (0.001)	0.998 (0.000)	0.998 (0.000)	0.991 (0.001)
SC	0.957 (0.001)	0.931 (0.001)	0.983 (0.001)	0.982 (0.001)	0.936 (0.001)
SVM – linear	0.943 (0.001)	0.933 (0.001)	0.953 (0.001)	0.953 (0.001)	0.934 (0.001)
SVM – radial	0.517 (0.003)	0.998 (0.000)	0.036 (0.001)	0.509 (0.003)	0.892 (0.002)

[a] Using only 5 files with subsampling

Table 63. Simulation Study Results for Disposition Response (balanced and not correlated data, learning/test set) (standard error)

	ACC	SN	SP	PPV	NPV
ANN[a]	0.952 (0.001)	0.940 (0.001)	0.964 (0.001)	0.964 (0.001)	0.942 (0.001)
kNN	0.831 (0.002)	0.690 (0.003)	0.972 (0.001)	0.962 (0.001)	0.759 (0.002)
LDA	0.913 (0.002)	0.978 (0.001)	0.848 (0.002)	0.868 (0.002)	0.976 (0.001)
LogitBoost	0.503 (0.003)	0.509 (0.003)	0.498 (0.003)	0.504 (0.003)	0.502 (0.003)
RF	0.994 (0.000)	0.990 (0.001)	0.997 (0.000)	0.997 (0.000)	0.990 (0.001)
SC	0.960 (0.001)	0.945 (0.001)	0.974 (0.001)	0.974 (0.001)	0.947 (0.001)
SVM – linear	0.948 (0.001)	0.940 (0.001)	0.957 (0.001)	0.957 (0.001)	0.941 (0.001)
SVM – radial	0.514 (0.003)	0.979 (0.001)	0.049 (0.001)	0.497 (0.003)	0.888 (0.002)

[a] Using only 5 files with subsampling

All previous analyses were done on a 1.70 GHz Windows XP Professional laptop with 512 MB of RAM. To run the simulation study (without ANN) and obtain the statistics for 100 files for one particular combination (for example unbalanced and correlated data), required approximately 6 hours for source of bleeding and 2-3 hours for the other responses. To obtain the files needed to create the ROC curves took approximately 63-65 hours to get one combination for the source of bleeding response. The remaining responses and combinations were run on a 2 GHz Vista laptop with 2 GB of RAM. The running times were approximately 7-10 hours, 6-8 hours, and 10-13 hours for the resuscitation, endoscopy, and disposition responses respectively. These running times were for a single response, one combination type of data. Getting the statistics for ANN for any one response, one combination, took approximately 1 hour. To get the files needed to create the ROC curves for ANN for one response, one combination, took approximately 30-45 minutes. The learning/test set data, run on the Vista laptop, took a shorter amount of time to run the models and obtain the results.

7. Optimizing the Performance of Random Forest

Throughout our studies, we have seen that the random forest model consistently performs the best. We know random forest already performs well, but can it do even better? Random forest is known to perform more poorly with very unbalanced datasets. By optimizing the parameters used, can we improve random forest's performance? Three parameters of random forest that can be optimized are ntree, mtry, and cutoff.

7.1 Methods

The parameter ntree refers to the number of trees grown in the forest. The different values tried for ntree were: 100, 200, 500 (default value), 1000, and 2000. The mtry parameter refers to the number of variables that are sampled randomly at each node split in a given tree. Values tried for mtry started at 5, going up in increments of 5's (5, 10, 15, etc.), the last value tried was the largest value which was still less than the number of variables total. For example, if the number of variables was 23, then the largest value of mtry that is tried would be 20. The default value for mtry is $\text{floor}\left(\sqrt{p}\right)$, where p is the number of input variables there are altogether. The cutoff parameter (which is a vector) indicates what class to place a new test case in. The default is to weight each class equally (the vector containing the cutoffs would then have values 1/c, where c is the number of classes there are). However, if the data is imbalanced, then it might be better to weight one class more heavily than the other. Cutoff values that were tried for resuscitation, endoscopy, and disposition were varied by increasing the cutoff for the "No/not ICU" class by 0.05 on each iteration. Thus, the cutoff for the "Yes/ICU" class would be decreased by 0.05 on each iteration. For source of bleeding, the "upper" class would be increased by 0.05 and the "lower" and "middle" class would be decreased by 0.02 and 0.03 respectively on each iteration. The stopping point for cutoff values would be when the maximum cutoff value was reached for the "majority" class (i.e. the class that had the most data in it).

Ten runs of 10-fold cross-validation were done and the simulated data and actual GIB data were analyzed. Parameters were optimized on each fold of the cross-validation. Statistics calculated were: accuracy, sensitivity, specificity, positive predictive value, and negative predictive value. The optimal parameters on each fold of CV were stored and at the end, the average optimal parameter was found along with their respective standard errors. Due to computational constraints, only 10 different files were considered for the simulated data, not 100 files.

7.2 Optimizing Random Forest Results

Tables 64-73 show the results obtained when optimizing the random forest parameters. The statistics calculated are shown along with the average optimal parameters obtained and their corresponding standard errors. Because of the order of the parameter combinations tried, the lowest ones were tried first (for example for the simulated resuscitation data, the parameters first tried were ntree=100, mtry=5, and cutoff=(0.2,0.8)). We see that these lowest parameters were already the optimal ones in most cases. If the highest parameters were tried first, these would be the optimal ones. Essentially, it doesn't matter which parameters are chosen. Random forest will do well regardless. This analysis was done on the Vista laptop. Running times were roughly 4-7 hours for all responses for the simulated data. Running times for the actual data ranged from 1-3 hours.

Table 64. Optimizing RF Results for Source of Bleeding Response, Simulated Data (standard error)

	ACC	SN	SP	PPV	NPV
Unbalanced and correlated data	0.961 (0.004)	0.997 (0.001)	0.820 (0.007)	0.957 (0.004)	0.984 (0.002)
Unbalanced and not correlated data	0.967 (0.003)	0.999 (0.001)	0.840 (0.007)	0.962 (0.003)	0.994 (0.001)
Balanced and correlated data	0.989 (0.002)	0.992 (0.002)	0.988 (0.002)	0.976 (0.003)	0.996 (0.001)
Balanced and not correlated data	0.984 (0.002)	0.980 (0.003)	0.987 (0.002)	0.974 (0.003)	0.990 (0.002)

Table 65. Optimal RF Parameters for Source of Bleeding Response, Simulated Data (standard error)

	ntree	mtry	cutoff
Unbalanced and correlated data	100 (0.000)	6.55 (0.020)	0.393 (0.001)
Unbalanced and not correlated data	101 (0.058)	5.9 (0.013)	0.383 (0.001)
Balanced and correlated data	100 (0.000)	5.55 (0.011)	0.356 (0.000)
Balanced and not correlated data	105 (0.238)	5.35 (0.001)	0.379 (0.001)

Table 66. Optimizing RF Results for Resuscitation Response, Simulated Data (standard error)

	ACC	SN	SP	PPV	NPV
Unbalanced and correlated data	0.989 (0.002)	0.962 (0.004)	0.996 (0.001)	0.983 (0.002)	0.990 (0.002)
Unbalanced and not correlated data	0.988 (0.002)	0.955 (0.004)	0.996 (0.001)	0.983 (0.005)	0.989 (0.002)
Balanced and correlated data	0.987 (0.002)	0.992 (0.002)	0.983 (0.002)	0.983 (0.002)	0.992 (0.002)
Balanced and not correlated data	0.985 (0.002)	0.985 (0.002)	0.984 (0.002)	0.984 (0.002)	0.985 (0.002)

Table 67. Optimal RF Parameters for Resuscitation Response, Simulated Data (standard error)

	ntree	mtry	cutoff
Unbalanced and correlated data	111 (0.525)	5.25 (0.008)	0.212 (0.000)
Unbalanced and not correlated data	100 (0.000)	5.7 (0.014)	0.227 (0.000)
Balanced and correlated data	100 (0.000)	5.45 (0.009)	0.214 (0.000)
Balanced and not correlated data	100 (0.000)	5.45 (0.009)	0.216 (0.000)

Table 68. Optimizing RF Results for Endoscopy Response, Simulated Data (standard error)

	ACC	SN	SP	PPV	NPV
Unbalanced and correlated data	0.968 (0.003)	0.878 (0.006)	0.991 (0.002)	0.960 (0.004)	0.970 (0.003)
Unbalanced and not correlated data	0.977 (0.003)	0.912 (0.005)	0.993 (0.002)	0.970 (0.003)	0.978 (0.003)
Balanced and correlated data	0.959 (0.004)	0.953 (0.004)	0.964 (0.003)	0.964 (0.003)	0.954 (0.004)
Balanced and not correlated data	0.958 (0.004)	0.949 (0.004)	0.967 (0.003)	0.966 (0.003)	0.950 (0.004)

Table 69. Optimal RF Parameters for Endoscopy Response, Simulated Data (standard error)

	ntree	mtry	cutoff
Unbalanced and correlated data	103 (0.099)	7 (0.022)	0.249 (0.001)
Unbalanced and not correlated data	101 (0.058)	6.7 (0.020)	0.233 (0.000)
Balanced and correlated data	101 (0.058)	6.55 (0.025)	0.258 (0.001)
Balanced and not correlated data	100 (0.000)	6.4 (0.021)	0.243 (0.000)

Table 70. Optimizing RF Results for Disposition Response, Simulated Data (standard error)

	ACC	SN	SP	PPV	NPV
Unbalanced and correlated data	0.991 (0.002)	0.960 (0.004)	0.999 (0.001)	0.997 (0.001)	0.990 (0.002)
Unbalanced and not correlated data	0.992 (0.002)	0.962 (0.004)	0.999 (0.001)	0.997 (0.001)	0.991 (0.002)
Balanced and correlated data	0.994 (0.001)	0.991 (0.002)	0.997 (0.001)	0.997 (0.001)	0.991 (0.002)
Balanced and not correlated data	0.993 (0.002)	0.991 (0.002)	0.994 (0.001)	0.994 (0.001)	0.991 (0.002)

Table 71. Optimal RF Parameters for Disposition Response, Simulated Data (standard error)

	ntree	mtry	cutoff
Unbalanced and correlated data	100 (0.000)	6.15 (0.021)	0.215 (0.046)
Unbalanced and not correlated data	104 (0.231)	5.45 (0.012)	0.216 (0.000)
Balanced and correlated data	100 (0.000)	5.2 (0.006)	0.208 (0.000)
Balanced and not correlated data	100 (0.000)	5.3 (0.013)	0.208 (0.000)

Table 72. Optimizing RF Results for Actual GIB Data (standard error)

	ACC	SN	SP	PPV	NPV
Source of bleeding response	0.946 (0.007)	0.979 (0.004)	0.945 (0.007)	0.973 (0.003)	0.957 (0.004)
Resuscitation response	0.928 (0.008)	0.932 (0.007)	0.918 (0.008)	0.952 (0.004)	0.886 (0.006)
Endoscopy response	0.789 (0.012)	0.847 (0.010)	0.681 (0.014)	0.832 (0.007)	0.706 (0.008)
Disposition response	0.877 (0.010)	0.903 (0.009)	0.834 (0.011)	0.903 (0.005)	0.834 (0.007)

Table 73. Optimal RF Parameters for Actual GIB Data (standard error)

	ntree	mtry	cutoff
Source of bleeding response	109 (0.412)	5.85 (0.027)	0.367 (0.001)
Resuscitation response	111 (0.830)	5.2 (0.009)	0.513 (0.000)
Endoscopy response	113 (0.309)	6.9 (0.035)	0.525 (0.000)
Disposition response	113 (0.548)	6.6 (0.033)	0.521 (0.000)

Comparing these optimized results with the results for the corresponding simulated data and actual GIB data, there are not any significant improvements seen, although there are improvements seen for sensitivity and specificity values on some of the simulated data. Given that there were no significant improvements seen with regards to accuracy, it might not be practical to optimize the parameters, given the much increased computational time. Random forest already does so well, that attempting to optimize the parameters did not make a big difference. Although there were significant improvements seen sometimes with sensitivity and specificity values, it was not consistent for every case. Optimizing parameters may however be dependent on the dataset, so these findings might not carry over to other datasets.

8. Online (Web-Based) Tool for Classifying GIB Data

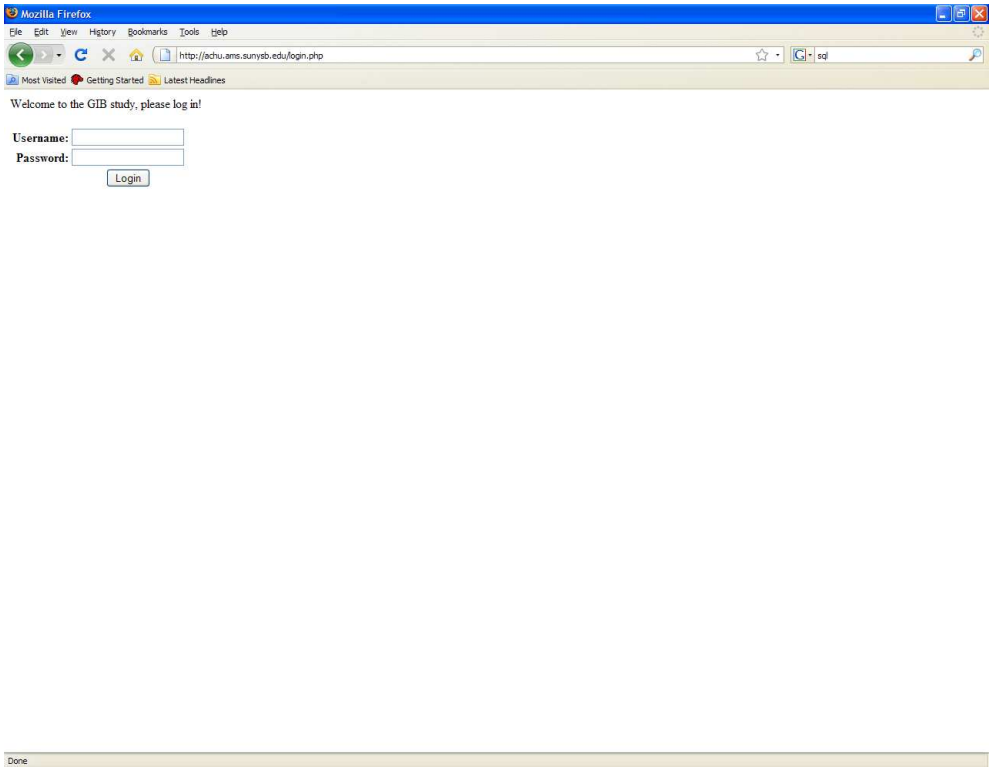
There would have to be a way to present the use of the random forest model that is easy to understand and simple to use for non-computer scientists and non-statisticians/mathematicians. In a nutshell, we can have a website where physicians enter in the patient input data and by clicking a button, the model predictions will be displayed on the following screen. This concept can be further extended so that instead of the physician going to a website, the same steps can be integrated directly into the hospital computer system, making it a seamless process. However, before this could be put into practice, we would have to determine whether this is something that would be accepted and what is the best way to present the material in the most user-friendly fashion.

This involves a two-step study. First, the website is developed and implemented. Second, a survey is conducted to assess whether the web-based tool is easy to use and what improvements can be made. Also, we want to find out if this is something physicians would consider using and what their thoughts are on the tool. The second step involves testing out the website in an actual hospital setting and seeing whether there can be an improvement in patient management and care. The data that physicians enter in will be stored in a database and the accuracies of the physician predictions, model predictions, and physician predictions after seeing the model predictions will all be compared to each other. In the second step, the web tool will be tested at 3 different sites, including the Stony Brook University Medical Center and Northport Veterans Affairs Medical Center. This study will be done as future research.

8.1 What the User Sees

On the website, the physician must first enter in a username and password before gaining access to the HTML form (to ensure only those we want to have access are accessing the page – we don't want any "garbage" data entries). See Figure 35.

Figure 35. Web-Based Tool – Log-In Page for Physicians



The physicians input onto an HTML form the patient clinical data and their predictions of the 4 responses (source of bleeding, need for urgent resuscitation, need for urgent endoscopy, and disposition). See Figure 36 and Figure 37.

Figure 36. Web-Based Tool – HTML Form Where Physicians Input Data

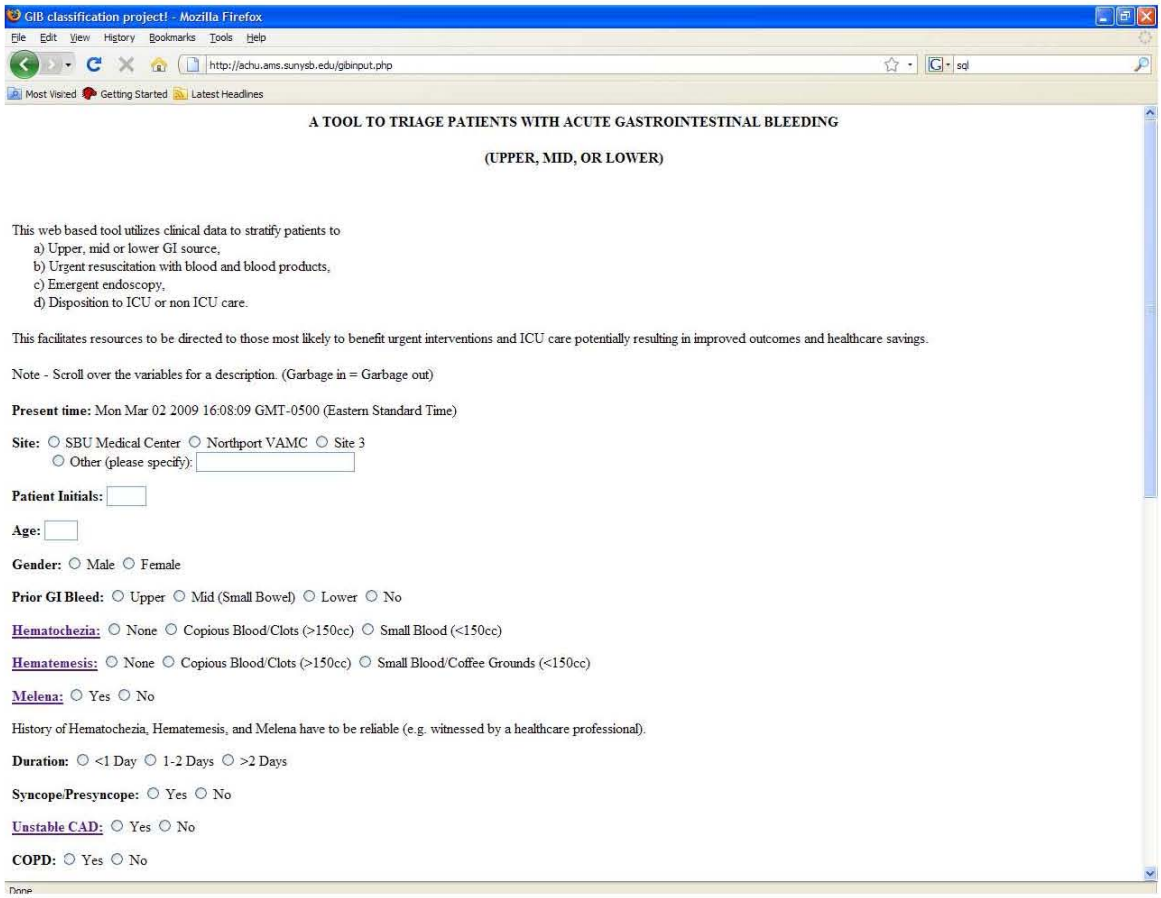
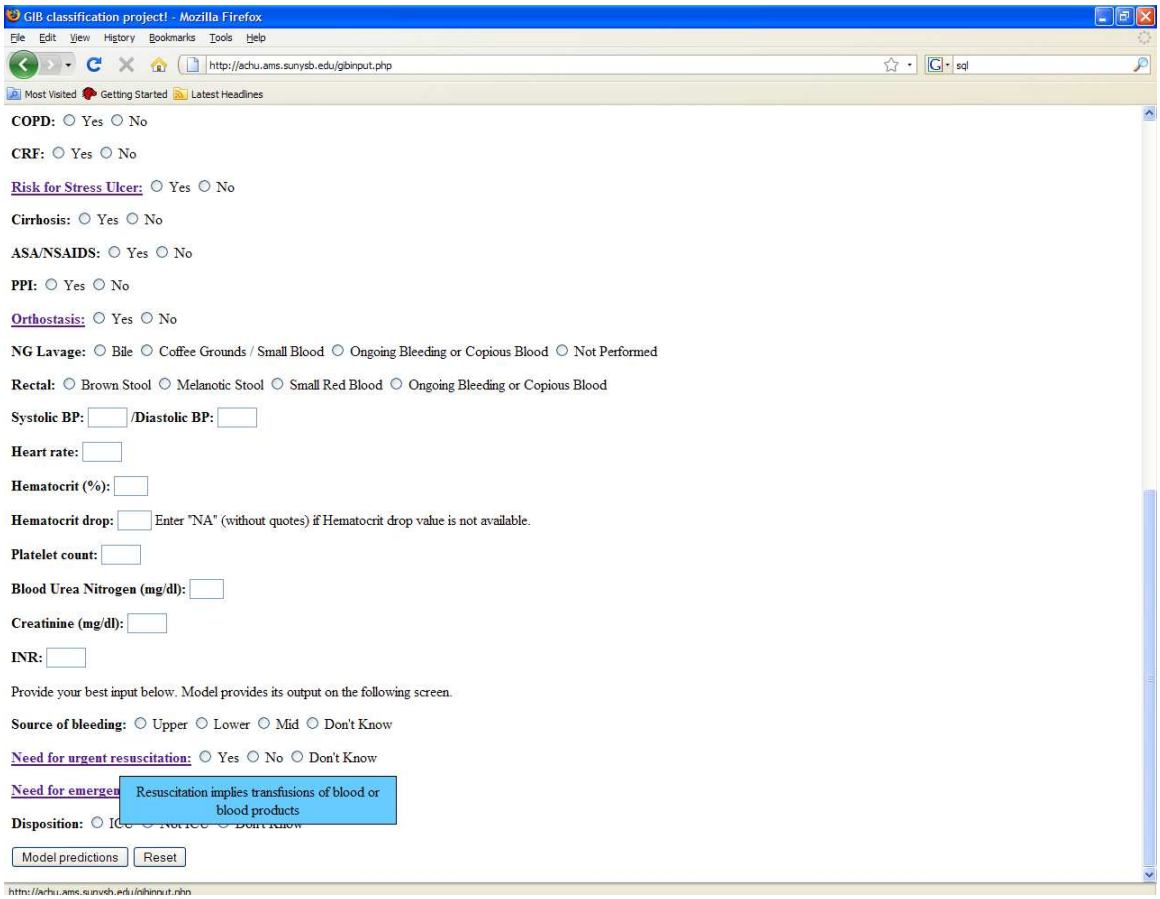
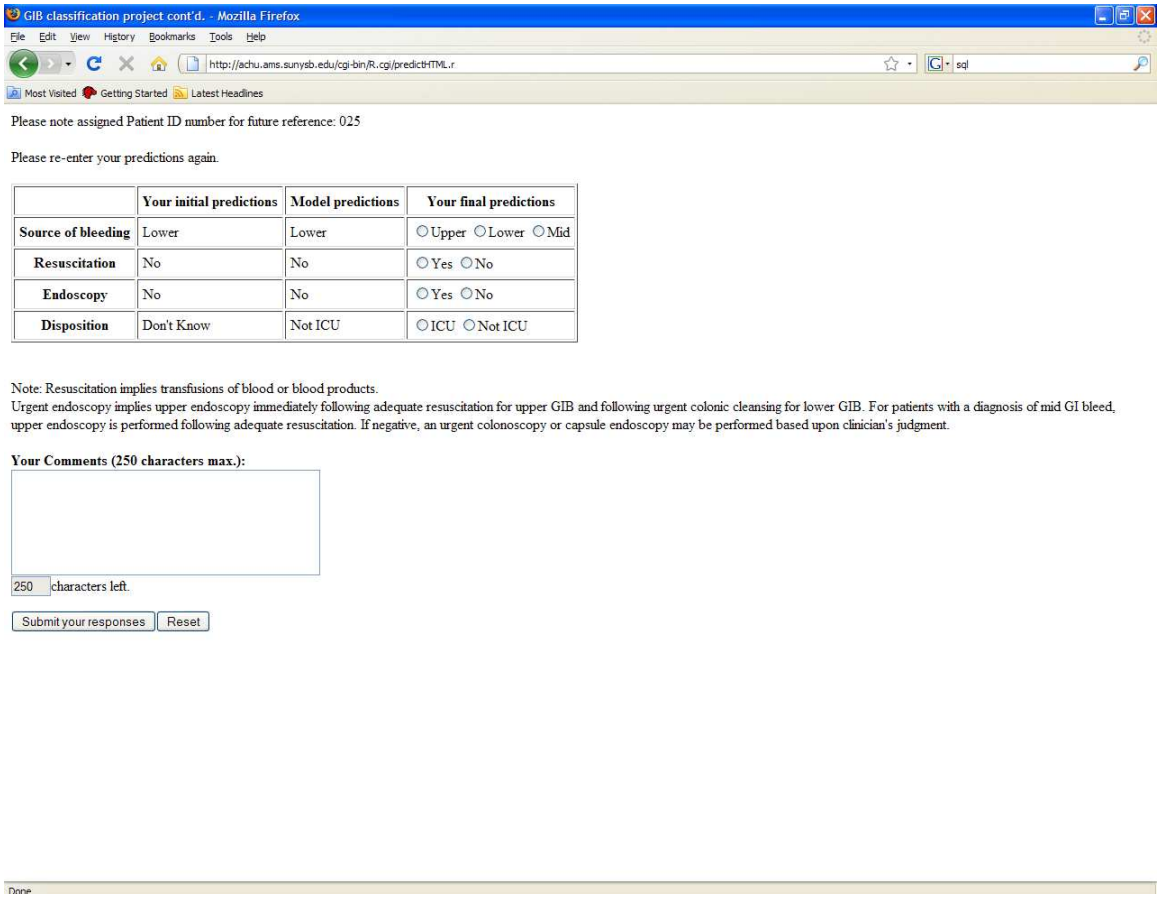


Figure 37. Web-Based Tool – HTML Form Where Physicians Input Data (At End of Page, Physician Presses the "Model Predictions" Button)



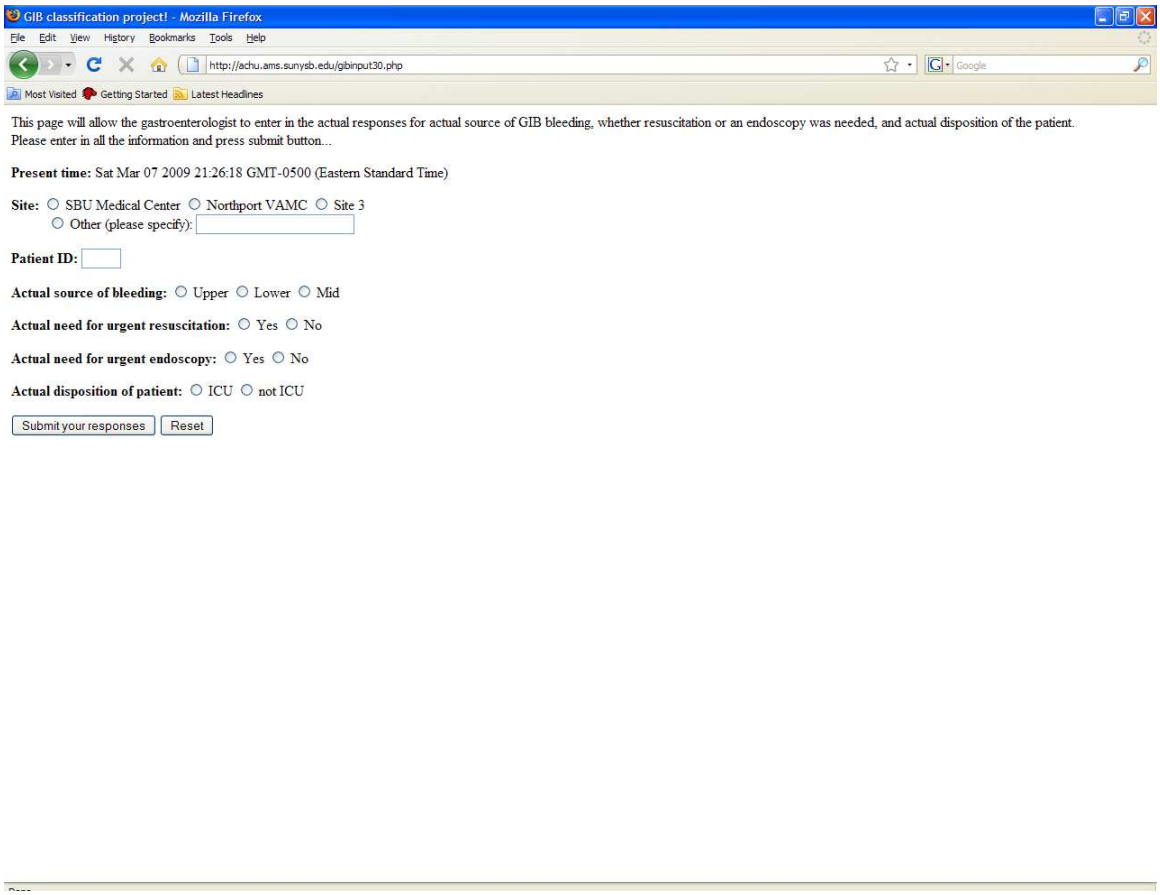
Upon submission of the form, the predictions from the model are displayed on the following page. Physicians will then be asked to make a second round of predictions once they have seen the model's predictions. See Figure 38. We refer to the second and final physician predictions as the model+physician's predictions.

Figure 38. Web-Based Tool – Model Predictions



For the first step of the study, a questionnaire will be displayed on the page as well, asking for information and getting feedback about how they like the web tool. On a separate webpage, following each patient after 30 days, the gastroenterologists will look at the clinical data and endoscopic data and enter in the actual diagnosis. That way we can compare the accuracies of the initial physician's predictions, model's predictions, and model+physician's predictions. See Figure 39.

Figure 39. Web-Based Tool – HTML Form Where Gastroenterologists Enter in Actual Diagnosis



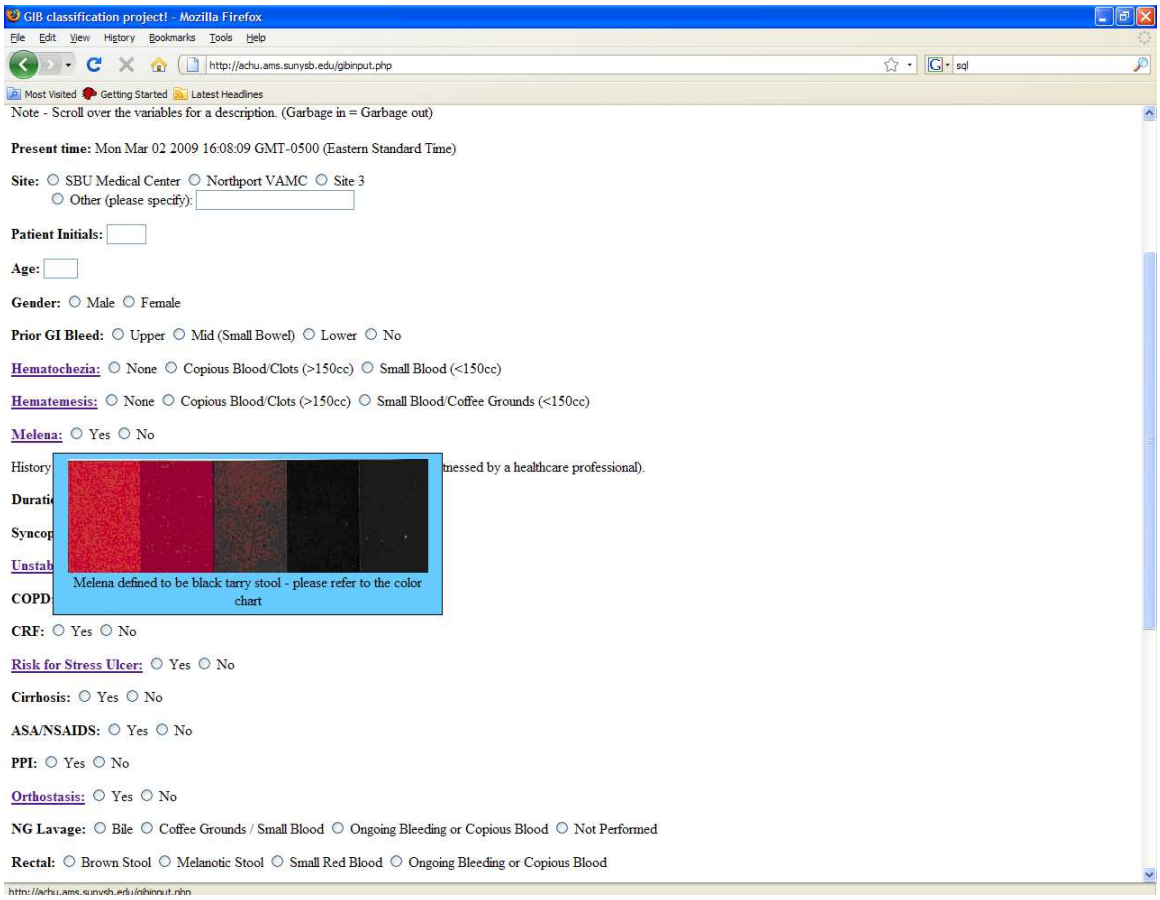
8.2 Behind-the-Scene: How the Website Works

8.2.1 Inner Workings of the Website

In this section, we describe the technical details of how the website was developed, implemented and how it is maintained. PHP (PHP: Hypertext Preprocessor), a scripting language that is used in conjunction with HTML, is used to validate the username and password. Cookies are used to ensure that nobody is bypassing the login page and attempting to access the main webpage directly. If they haven't entered in the correct username and password and try to access the main webpage directly, they are redirected back to the login page. JavaScript, a scripting language used on the client side, is used to create tool-tips (boxes that appear when the mouse hovers over a specific link) to give

further explanation to the input values (i.e. what units the values should be in or how the variable is defined). The client here is the web browser (for example Mozilla Firefox or Internet Explorer). See Figure 40.

Figure 40. Web-Based Tool – HTML Form Where Physicians Input Data (Demonstrating Tool-Tips)



By using the CGIwithR package in the R statistical program, our R program can communicate with the web server through a CGI (Common Gateway Interface) script. The web server here is a computer that accepts HTTP requests and processes them, sending back the appropriate reply, which can be for example a new web page. An HTML form is sent by the web server to a CGI script, which then passes the contents of the form to the R program. A CGI script can be thought of as a messenger for communicating information between a CGI program (the R program in this case) and the web server. The CGI script is a standard Internet protocol that translates information from the web server so that it is in a format that the R program can understand.

The data from the HTML form is validated through an R program to ensure there is no invalid data entered. Examples of invalid data would be if non-numerical characters (letters A-Z or a-z, symbols such as ?, !, @, #, $, etc.) were entered when an integer value was supposed to be entered, more than one decimal point was entered for decimal values, or if there were negative values entered when they were supposed to be non-negative. If there is invalid data, then the user is informed of what data is invalid and is asked to correct the invalid data before continuing on to obtain the model's predictions. The R program also makes sure that important information is filled in, such as what site the physician is at and what the physician's predictions are. Once the form's contents have been validated, the random forest model makes the predictions and the model's predictions are displayed on the webpage. The physician is asked to enter in their final predictions upon viewing the model's predictions. These predictions must be made in order to reach the final page.

8.2.2 MySQL and R

All of the data (input data, physician's predictions, model's predictions, model+physician's predictions, answers to the survey, and gastroenterologist's actual diagnoses) are written out to a MySQL database. MySQL is a very popular open source database system that uses the SQL language to retrieve information from the database and manage the database. Data in MySQL is stored in tables. By using the RMySQL package in R, R is able to connect to a specified MySQL database and extract information from the database or submit information to be stored into the database. To ensure each patient entry is unique, the database automatically assigns a different patient id to each patient for a given site. There is a separate table in the database for each website page, so by assigning an id to each patient, all the data for a particular patient can be matched up accordingly at the end of the study. The R programs also do error checking to ensure that no patient information is entered that shouldn't be. For example, if the physician for some reason presses the Back button on their browser and tries to submit their data again, they won't be allowed to submit again, otherwise it would cause a double entry for one patient. On the gastroenterologist page, if a gastroenterologist tries to enter in the actual diagnosis for patient 106, but there was not any initial information for patient 106 previously, then the gastroenterologist will not be allowed to enter in this information.

The database has been normalized – redundant data has been eliminated in tables and only data that is relevant is stored in each table. Patient IDs as well as patient initials have been encrypted so in the event the database is compromised, no links to patients will be revealed. Further, backups are done regularly to ensure minimal loss of data. Routine checks are also done to make sure the database has not been corrupted and is functioning properly. The web-based tool is located at: http://achu.ams.sunysb.edu. If you wish to be able to login, please send an email to achu.sunysb@gmail.com and information will be sent on how to log in.

9. Final Conclusions

Acute gastrointestinal bleeding has become an increasing healthcare concern due to rising NSAID use in an aging population. With non-gastroenterologists incorrectly diagnosing GIB patients at least 50% of the time, it would be very beneficial to find a way to assist them. We proposed that a classification model could be developed to help non-gastroenterologists and improve patient diagnosis and care. These classification models are part of a broader group, referred to as decision support systems, i.e. a computerized system that takes raw information, processes it, and makes informed decisions and comes up with solutions. We compared 8 different classification models, using actual GIB data as well as simulated data and externally validating the top performing models. The best performing model, random forest, was also compared to existing GIB scoring systems and was seen to be comparable. Most of the models performed very well with the actual GIB data and the simulated data, definitely improving on the accuracy of a non-gastroenterologist. The random forest model had excellent performance and stood out from all the other models. We also attempted to optimize random forest, but our results showed no significant improvement. It is the hope and our goal that classification models can be used in practice in the hospital setting to assist physicians in predicting GIB responses to better care for GIB patients.

10. Future Studies

In Section 8, we described the web-based tool that was implemented. In future studies, we will develop a questionnaire to get information about ways to improve the tool and to assess whether the tool would be useful. We also will perform a study in three hospitals, to test the tool out and see whether it is practical in the real setting. Four hundred fifty patient observations will be collected and the accuracies of the physician's predictions, model's predictions, and model+physician's predictions will be compared. If the study is successful and improvement is shown with using the model over physician's diagnosis alone, a similar study will be implemented at more sites (five or six hospitals or medical centers) and a larger number of patient observations will be collected and analyzed. The ultimate goal is to be able to use a classification model to assist the doctors in order to run the hospital more efficiently and for it to be more cost effective. The next step would be to integrate the model directly into a hospital computer system and potentially put it into practice all across Long Island.

References

1. Rockall TA, Logan RFA, Devlin HB, Northfield TC, on behalf of the Steering Committee and members of the National Audit of Acute Upper Gastrointestinal Haemorrhage. Incidence of and mortality from acute upper gastrointestinal haemorrhage in the United Kingdom. British Medical Journal 1995 July; 311: 222-26.
2. Kollef MH, O'Brien JD, Zuckerman GR, Shannon W. BLEED: A classification tool to predict outcomes in patients with acute upper and lower gastrointestinal hemorrhage. Critical Care Medicine; 1997 July. 25(7): 1125-32.
3. Baradarian R, Ramdhaney S, Chapalamadugu R, Skoczylas L, Wang K, Rivilis S, Remus K, Mayer I, Iswara K, Tenner S. Early Intensive Resuscitation of Patients with Upper Gastrointestinal Bleeding Decreases Mortality. American Journal of Gastroenterology. 2004 April; 99(4): 619-22.
4. Elta GH. Urgent colonoscopy for acute lower-GI bleeding. Gastrointestinal Endoscopy. 2004 March; 59(3): 402-8.
5. Das A, Wong RC, Prediction of Outcome of Acute GI Hemorrhage: A review of risk scores and predictive models. Gastrointestinal Endoscopy. 2004 July; 60(1): 85-93.
6. Rockall TA, Logan RFA, Devlin HB, Northfield TC. Risk assessment after acute upper gastrointestinal haemorrhage. Gut. 1996 March; 38(3) :316-21.
7. Rockall TA, Logan RFA, Devlin HB, Northfield TC. Selection of patients for early discharge or outpatient care after acute upper gastrointestinal haemorrhage. National Audit of Acute Upper Gastrointestinal Haemorrhage. Lancet. 1996 April; 347(9009): 1138-40.
8. Blatchford O, Murray WR, Blatchford M. A risk score to predict need for treatment for upper-gastrointestinal haemorrhage. Lancet. 2000 October; 356(9238): 1318-21.
9. Russell S, Norvig P. Artificial Intelligence – A Modern Approach. New Jersey: Prentice Hall. 2003.
10. Schapire R. The Strength of Weak Learnability. Machine Learning. 1990 June; 5(2): 197-227.
11. Freund Y, Schapire R. Experiments with a New Boosting Algorithm. Machine Learning: Proceddings of the Thirteenth International Conference, 1996.
12. Dettling M. BagBoosting for Tumor Classification with Gene Expression Data. Bioinformatics. 2004 October; 20(18): 3583-3593.
13. Friedman J, Hastie T, Tibshirani R. Additive logistic regression: a statistical view of boosting. Annals of Statistics. 2000 April; 28(2): 337-407.
14. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning – Data Mining, Inference, and Prediction. 2001.
15. Breiman L. Random Forest. Machine Learning. 2001 October; 45(1): 5-32.
16. Tibshirani R, Hastie T, Balasubramanian N, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proceedings of the National Academy of Sciences of the United States of America. 2002 May; 99(10): 6567-6572.

17. Vapnik, V. The nature of statistical learning theory. New York, NY: Spring Verlag. 1999.
18. Molinaro AM, Simon R, Pfeiffer RM. Prediction error estimation: a comparison of resampling methods. Bioinformatics. 2005. 21(15): 3301-3307.
19. Bamber D. The area above the ordinal dominance graph and the area below the receiver operating graph. Journal of Mathematical Psychology 1975; 12: 387-415.
20. Barkun A, Bardou M, Marshall JK. Nonvariceal Upper GI Bleeding Consensus Conference Group. Consensus Recommendations for Managing Patients with Nonvariceal Upper Gastrointestinal Bleeding. Annals of Internal Medicine. 2003 November; 139(10): 843-57.
21. Strate LL, Orav EJ, Syngal S. Early Predictors of Severity in Acute Lower Intestinal Tract Bleeding. Archives of Internal Medicine. 2003 April; 163(7): 838-43.
22. Klebl F, Bregenzer N, Schofer L, Tamme W, Langgartner J, Scholmerich J, Messmann H. Risk factors for mortality in severe upper gastrointestinal bleeding. International Journal of Colorectal Disease. 2005. 20(1): 49-58.
23. Velayos FS, Williamson A, Sousa KH, Lung E, Bostrom A, Weber EJ, Ostroff JW, Terdiman JP. Early predictors of severe lower gastrointestinal bleeding and adverse outcomes: a prospective study. Clinical Gastroenterology and Hepatology. 2004 June; 2(6): 485-90.
24. Kalula SZ, Swingler GH, Louw JA. Clinical predictors of outcome in acute upper gastrointestinal bleeding. South African Medical Journal. 2003 April; 93(4): 286-90.
25. Bordley DR, Mushlin AI, Dolan JG, Richardson WS, Barry M, Polio J, Griner PF. Early clinical signs identify low-risk patients with acute upper gastrointestinal hemorrhage. The Journal of the American Medical Association. 1985 June; 253(22): 3282-5.
26. Mortensen PB, Nohr M, Moller-Petersen JF, Balsley I. The diagnostic value of serum urea/creatinine ratio in distinguishing between upper and lower gastrointestinal bleeding. A prospective study. Danish Medical Bulletin. 1994 April; 41(2): 237-40.
27. Zimmerman J, Siguencia J, Tsvang E, Beeri R, Amon R. Predictors of mortality in patients admitted to hospital for acute upper gastrointestinal hemorrhage. Scandinavian Journal of Gastroenterology. 1995 April; 30(4): 327-31.
28. Terdiman JP, Ostroff JW. Risk of persistent or recurrent and intractable upper gastrointestinal bleeding in the era of therapeutic endoscopy. The American Journal of Gastroenterology. 1997 October; 92(10): 1805-11.
29. Corley DA, Stefan AM, Wolf M, Cook EF, Lee TH. Early indicators of prognosis in upper gastrointestinal hemorrhage. American Journal of Gastroenterology. 1998 March; 93(3): 336-40.
30. Zaragoza Marcet A, Tenias Burillo JM, Llorente Melero MJ. Pre-endoscopic prognostic factors in non-varicose upper gastrointestinal bleeding. Development of a predictive algorithm. Revista Espanola de Enfermedades Digestivas. 2002 March; 94(3): 139-48.
31. Chu A, Ahn H, Halwan B, Kalmin B, Artifon E, Barkun A, Lagoudakis M,

Kumar A. A decision support system to facilitate management of patients with acute gastrointestinal bleeding. Artificial Intelligence in Medicine. 42(3): 247-259. March 2008.
32. Quirk DM, Barry MJ, Aserkoff B, Podolsky DK. Physician specialty and variations in the cost of treating patients with acute upper gastrointestinal bleeding. Gastroenterology 1997 November; 113(5): 1443-8.
33. Timmerman D, Verrelst H, Bourne TH, De Moor B, Collins WP, Vergote I, Vandewalle J. Artificial neural network models for the preoperative discrimination between malignant and benign adnexal masses. Ultrasound in Obstetrics and Gynecology. 1999 January; 13(1): 17-25.
34. Rosenblatt KP, Bryant-Greenwood P, Killian JK, Mehta A, Geho D, Espina V, Petricoin III EF, Liotta LA. Serum Proteomics in Cancer Diagnosis and Management. Annual Review of Medicine. 2004 February; 55: 97-112.
35. Selaru FM, Xu Y, Yin J, Zou T, Liu TC, Mori Y, Abraham JM, Sato F, Wang S, Twigg C, Olaru A, Shustova V, Leytin A, Hytiroglou P, Shibata D, Harpaz N, Meltzer SJ. Artificial neural networks distinguish among subtypes of neoplastic colorectal lesions. Gastroenterology. 2002 March; 122(3): 606-13.
36. Chong CF, Li YC, Wang TL, Chang H. Stratification of Adverse Outcomes by Preoperative Risk Factors in Coronary Artery Bypass Graft Patients: An Artificial Neural Network Prediction Model. AMIA Annual Symposium Proceedings. 2003: 160-4.
37. Lund LH. Comment on: Computerized interpretation of the electrocardiogram. Archives of Internal Medicine. 2004 August; 164(15): 1698-9.
38. Kennedy RL, Harrison RF, Burton AM, Fraser HS, Hamer WG, MacArthur D, McAllum R, Steedman DJ. An artificial neural network system for diagnosis of acute myocardial infarction (AMI) in the accident and emergency department: evaluation and comparison with serum myoglobin measurements. Computer Methods and Programs in Biomedicine. 1997 February; 52(2): 93-103.
39. Lisboa PJ. A review of evidence of health benefit from artificial neural networks in medical intervention. Neural Networks. 2002 January; 15(1): 11-39.
40. Ahn H, Moon H, Fazzari MJ, Lim N, Chen JJ, Kodell RL. Classification by ensembles of random partitions of high-dimensional data. Computational Statistics and Data Analysis. 2007 August; 51(12): 6166-6179.
41. Duduoit S, Fridlyand J, Speed TP, Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data. Journal of the American Statistical Association. 2002 March; 97(457): 77-87.
42. Guyon I, Weston J, Barnhill S, Vapnik V. Gene Selection for Cancer Classification using Support Vector Machines. Machine Learning. 2002; 46(1-3): 389-422.
43. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. Biometrics. 1988 September; 44(3): 837-845.
44. Dudani SA. The distance-weighted k-nearest-neighbor rule. IEEE Transactions on Systems, Man, Cybernetics. 1976. SMC-6:325-327.

45. Freund Y, Schapire R. Experiments with a New Boosting Algorithm. Machine Learning: Proceedings of the Thirteenth International Conference. 1996. 148-156.
46. Ayat NE, Cheriet M, Remaki L, Suen CY. KMOD – a new support vector machine kernel with moderate decreasing for pattern recognition. Application to digit image recognition. Proceedings of the Sixth International Conference on Document Analysis and Recognition. 2001. 1215-19.