Application of Machine Learning To Decision Support Systems

A Dissertation Presented

by

Han Yu

to

The Graduate School

in Partial Fulfillment of the

Requirements

For the Degree of

Doctor of Philosophy

in

Applied Mathematics and Statistics

Stony Brook University

December 2013

Stony Brook University

The Graduate School

Han Yu

We, the dissertation committee for the above candidate for the

Doctor of Philosophy degree, hereby recommend

acceptance of this dissertation.

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

Dr. Wei Zhu - Chairperson of Defense Committee
Professor, Applied Mathematics and Statistics

Dr. Song Wu – Defense Committee Member
Assistant Professor, Applied Mathematics and Statistics

Dr. Sangjin Hong – Outside Member
Professor, Department of Electrical and Computer Engineering

This dissertation is accepted by the Graduate School

Charles Taber
Interim Dean of the Graduate School

Abstract of the Dissertation

Application of Machine Learning To Decision Support Systems

by

Han Yu

Doctor of Philosophy

in

Applied Mathematics and Statistics

Stony Brook University

2013

This dissertation presents applications of machine learning methods to clinical data sets and development of a decision support system. Our goal is to develop machine learning methods to predict potential healthcare problems before the onset of the actual diseases. Our research involves two examples with high-dimensional data. The independent variables are selected depending on the quality of prediction, and the models will be trained on the subspaces of the training data set. We also employed feature extraction technique to the original feature space such as PCA, FCA, data transformation, etc. This projection of the original feature space to lower the dimension is proven to be efficient in reducing the dimension of the data set. Recently, a non-probabilistic classifier called support vector machines (SVM) has been developed. The main idea of SVM is about mapping the input vectors into a high-dimensional feature space, and then a linear decision surface is constructed. Thus, the prediction will be based on the relative position of the data point with the decision surface. A tree-based ensemble method called Random Forest also received attention. Using a random selection of features to split each node yields error rates that make this method compare favorably to Adaboost. In this dissertation, we applied several machine learning methods and techniques to develop a reliable decision support system.

Table of Contents

List of Algorithms and Tables

List of Figures

Chapter 1 Introduction

1.1 Machine learning and the history of classification

In statistics and computer science, machine learning theory is a newly proposed direction about the learning and the study of models which can uncover the hidden patterns in the data, and furthermore, to predict the unknowns based on the properties learned from training data sets. In recent years, a great number of new regression and classification methods such as support vector networks and artificial neural networks (ANN) have been proposed by researchers. On the other hand, classifier combination methods have also become an active research area in machine learning theory. Many studies have been published to illustrate the advantages and disadvantages of these combination techniques.

There are a number of well-known classification tools in machine learning theory, for example, Support Vector Machines (SVM), proposed by Cortes and Vapnik (1995), which was originally used for binary classification problems. Its base idea is to mapping the data points to a high-dimensional space, and then constructs a hyper plane or set of hyper planes in this space, which can split the data points into two distinguishable classes, and thus can be used for classification, regression, or other tasks. Good separation is achieved by maximizing the marginal distance to the nearest training data points of any classes. In addition to linearly separable problems, SVM can also be extended to learn linearly non-separable problems by first projecting the input data onto a high-dimensional feature space using kernel functions and formulating a linear classification problem in that feature space. Lee and Lin (2001) have explained the multi-category SVM models, which extend the binary SVM to multi-category classification problem. Based on these techniques, many research and applications using SVM have been carried out. Lu, Plataniotis and Ventesanopoulos (2001), for example, argue that SVM is superior to traditional empirical risk minimization principle employed by most of neural networks. Based on these studies, it has been proven by many researchers that SVM is becoming a state-of-the-art model in machine learning theory, and has been successfully applied to a number of applications ranging from classification and regression, face detection and verification, hand writing recognition, and speech verification.

Another direction of machine learning models is followed by the recent introduction of ensemble-voting approaches. Random Forest (RF) proposed by Breiman (2001), is one of such ensemble method, it combines the results of multiple decision tree classifiers by majority voting. In modern statistics, it is well-known that models based on data sets which contain a huge number of predictors and relatively small sample sizes are unstable. Breiman (2001) has genuinely avoided such difficulties by dividing the original feature space into a number of small partitions, and then assigning each partition to a single individual base classifier. Thus, RF predicts the unknown class category of a test data by combining the votes of multiple base classifiers. Many variations of the classical RF model have been proposed, which demonstrate the advantages of the combination paradigm over the individual classifier models. In this dissertation we also propose a new model by combining the optimal kernel selection algorithm with the RF ensemble method, and thus to further improve the diversity property of the classifier ensemble technique.

In this dissertation, we first introduce multiple classification models and their variations in chapter 2, and then we introduce two examples, to develop and compare the performance of these classification models. In chapter 3 and 4, we explain the details and research backgrounds for these examples, along with model validations, variable selection and predictions. Discussion and conclusion are given in chapter 5, with potential improvements and future studies in chapter 6.

1.2 Introduction to Detection of Gastrointestinal Bleeding (GIB) and pathology background

It is well known that prophetic diagnostic techniques are becoming an essential part of modern healthcare controls. Preventive measures are progressively introduced by healthcare specialists and play a vital part for the treatment of patients with possible healthcare problems. However, identifying underlying healthcare problems with little presence of symptoms has been proven to be a challenging issue for modern healthcare system. A decision support system is needed for predicting potential risks for individuals with inconspicuous symptoms. Our goal was to develop statistical methods to formulate a decision support system by using clinical and laboratory information, and thus to facilitate the detection of latent healthcare problems.

In the prediction of potential bleeding source and cohort identification among patients with acute gastrointestinal bleeding (GIB) study, we have developed a model to help gastroenterologists to determine whether the patients need urgent intervention. Patients with acute GIB are usually evaluated by home primary doctors. Patients with acute GIB may require urgent endoscopy to avoid further damage to the digestive system, with utilizing limited healthcare resources for those who need it the most. In our study, we utilize the ICD-9 codes to measure the patients with acute GIB. Three data sets have been introduced from hospital medical records database, with all variables required to develop and test models. Machine learning models including SVM and RF are trained and tested. And in the model validation stage, we compared the performance of these models using sensitivity (SN), specificity (SP), Negative Predictive Value (NPV), Positive Predictive Value (PPV) and overall accuracy (ACC).

1.3 Introduction to Schizophrenia detection and pathology background

In the prediction of risk for developing schizophrenia study, we have collected 165 cases in our dataset, among them 51 cases were healthy control, 19 cases were schizoaffective, 60 cases were schizophrenia patients, 20 cases were bipolar patients and 14 cases were major depressive disorder (MDD) patients. RF and SVM have been used to classify these data. We first use cross validation to evaluate the performance of these classification methods by comparing sensitivity, specificity, PPV, NPV and prediction accuracy on retained data sets. Then we applied the tuned models to the control group to predict the potential risks of developing schizophrenia before symptoms solidify into psychosis. The model with best predictive power for this data set was proposed and summarized in Chapter 4 and 5.

Chapter 2 Methodology

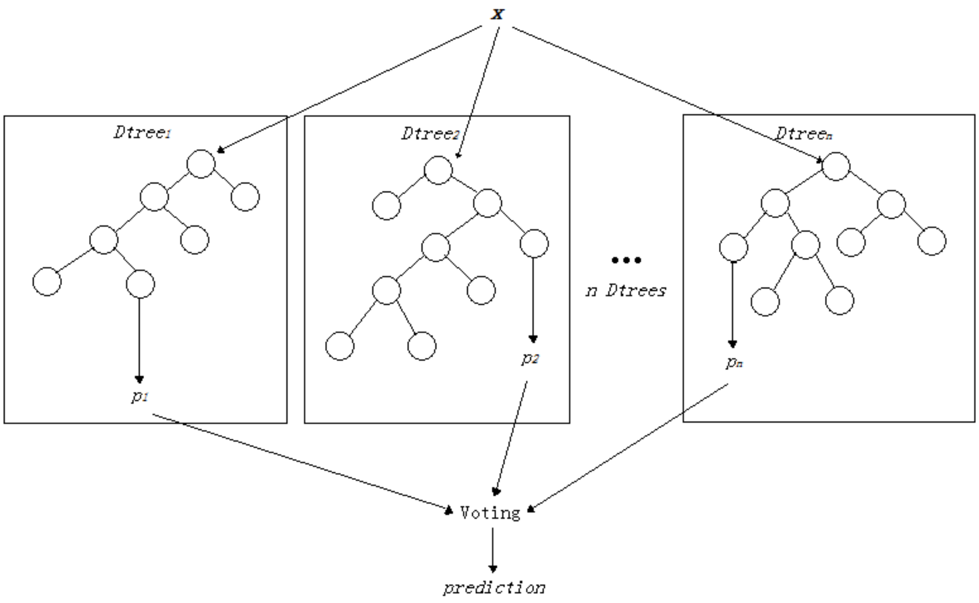
In our research, we applied multiple machine learning models to analyze the data set, each model was trained and their performances were compared. The key idea of these models was to use selected explanatory variables to predict the response variables using various statistical techniques. Methods we applied were, but not limited to, data interpolation, feature extraction, discriminant analysis and kernel learning.

Model training was performed on a randomly selected subset of patients, and tested on the remaining data set in model validation part. Data interpolation technique and robust methods were applied so that patients with missing data will not be discarded; categorical variables were changed into indicator variables. For each missing value in the data set, we used median to fit in the empty entry if the variable is numerical, and we used mode to fit in if the variable is categorical. This interpolation method may not provide any additional information, but in this way we can retain as much as information the original data set provided us. If we simply delete patients with missing data, all the cases with other entries that are not missing will be lost.

We applied 10-fold cross validation (CV) method to estimate how accurately a predictive model would perform in practice. The following statistics were calculated, if applicable: SN, SP, PPV, NPV and ACC.

2.1 Introduction to RF and its variations

RF is an ensemble classifier that combines multiple decision trees using the bagging algorithm. Bagging algorithm is a widely used ensemble based algorithm (Breiman et al., 1996), and the random selection of features, introduced by Ho et al. (1995), Amit and Geman (1997), is to construct an ensemble of decision trees with controlled variation with maximum feature diversity. In this algorithm, different training data sets are randomly drawn with replacement from the original training data set. Each training data set is used to train one individual classifier. The result is then given as a combination of individual classifiers by taking a majority vote of their decisions.



In 2009, the Apache Software Foundation initiated an open resource project to provide free implementations of machine learning algorithms on the Hadoop platform (https://cwiki.apache.org/MAHOUT/mahout-wiki.html, https://cwiki.apache.org/confluence/display/MAHOUT/Random+Forests). For a 2-way classification problem, let *X* be a matrix of *N* rows and *M* columns, where x_{ij} is the value of the j_{th} attribute in the i_{th} input data point. Let ***Y*** be a vector of *n* elements,$y_i \in \{+1, -1\}$ where y_{i} is the response class of the i_{th} data point. The ***Y*** values are categorical. Each decision tree is grown to its full potential and the simplified pseudo code is explained in Algorithm 2.1.1.

Algorithm 2.1.1. Construction of a decision tree

```
        Define function: Decision_Tree($\mathbf{X},\mathbf{Y}$)
        Select m variable at random with replacement out of M variables
        For j=1…m
        If $j_{th}$ attribute is categorical then
        $IG_j = H(Y) - H(Y|X_j) = -\sum_{l=1}^{m} p_l log_2 p_l - \sum_l p_l H(Y|X_j = v_j)$ (here $p_l$ is defined
as the entropy of $\mathbf{Y}$, i.e. $P(Y = y_l) = p_l$ for $l = 1, ..., N$, $H(Y|X = v)$ is the entropy of $\mathbf{Y}$ among
only those records in which $\mathbf{X}$ has value v)
        Else if $j_{th}$ attribute is real-valued
        $IG_j = max_t\{H(Y) - (H(Y|X_j < t)P(X_j < t) + H(Y|X_j \geq t)P(X_j \geq t))\}$
        Let $j^* = \mathrm{argmax}_j IG_j$
        If j* is categorical then
        For each value v of the $j_{th}$ attribute
        Let $\mathbf{X}^v$ = subset of rows of $\mathbf{X}$ in which $X_{ij} = v$. Let $Y^v$ = corresponding subset of $\mathbf{Y}$
        Let $Child^v$ = Decision_Tree ($\mathbf{X}^v,\mathbf{Y}^v$)
Return a decision tree node, splitting on $j_{th}$ attribute. The number of children equals the number
of values of the $j_{th}$ attribute, and the $v_{th}$ child is $Child^v$
        Else j* is real-valued and let t be the best split threshold
        Let $\mathbf{X}^{LO}$ = subset of rows of $\mathbf{X}$ in which $X_{ij} <= t$. Let $\mathbf{Y}^{LO}$ = corresponding subset of $\mathbf{Y}$
```

```
    Let $Child^{LO}$ = Decision_Tree ($\mathbf{X}^{LO}$,$\mathbf{Y}^{LO}$)
    Let $\mathbf{X}^{HI}$ = subset of rows of $\mathbf{X}$ in which $X_{ij} > t$. Let $\mathbf{Y}^{HI}$ = corresponding subset of $\mathbf{Y}$
    Let $Child^{HI}$ = Decision_Tree ($\mathbf{X}^{HI}$,$\mathbf{Y}^{HI}$)
Return a decision tree node, splitting on $j_{th}$ attribute. It has two children corresponding to whether the $j_{th}$ attribute is above or below the given threshold.
```

RF consists of a collection of tree-structured classifiers $\{h(\mathbf{x}, \theta_k), k = 1, ...\}$ where $\{\theta_k\}$ are independent and identically distributed random vectors and each tree casts a unit vote for the most popular class at input $\boldsymbol{X}$. Each tree is grown 1) sample as many cases from the original data set, but with replacement, to form the training data set. 2) sample m ($m<<M$) columns at random for each different node, where M is the number of input columns. Run the decision tree algorithm to fully grow a tree structured model using greedy algorithm. The value of m will be held constant during the entire training step. 3) Each decision tree model is grown to its full potential. There is no pruning. For an individual decision tree, the prediction is based on the trained nodes of the original tree, a vector $\boldsymbol{V}$ of M columns where V_j = the value of the j_{th} attribute. The prediction algorithm is illustrated as in Algorithm 2.1.2.

Algorithm 2.1.2. Prediction of the label of a single case

```
Define function Classify(node,$\mathbf{V}$)
if node.attribute = j then the split is done on the $j_{th}$ attribute
If node is a Leaf then
  Return the value predicted by node
Else
```

```
Let j = node.attribute
If j is categorical then
  Let $v = V_j$
  Let $child^{v}$ = child node corresponding to the attribute's value v
  Return Classify($child^{v}$,V)
Else j is real-valued
  Let t = node.threshold
  If $V_j < t$ then
    Let $child^{LO}$ = child node corresponding to (<t)
    Return Classify($child^{LO}$,V)
  Else
    Let $child^{HI}$ = child node corresponding to (>=t)
    Return Classify($child^{HI}$,V)
```

Breiman (2001) suggested that we pick a large number of base classifiers, and let $m = \sqrt{M}$ may optimize RF model's performance. For prediction, a new sample is plugged in for each tree. The leaf class that this sample ended up with will become its prediction. The sample point will be plugged in all the trees in the forest, and the ensemble will use majority voting to determine the final prediction.

Much literature has suggested that RF will constantly deliver a lower generalization error than many other ensemble methods. For example, Dietterich et al. (1998) have reported that it performs better than bagging. Breiman suggested that to improve accuracy, we need to minimize the correlation between individual base classifiers, at the same time maintaining strength.

As the classification combination technique is more popular, people are actively search for the reason. Banfield et al. (2004) reported a comparison of ensemble methods, Bauer et al. (1999) and Shi et al. (2007) also carried out an empirical comparison of voting classification algorithms. From the results of these empirical researches, Adaboost turned out to be a very comparative ensemble technique. However random selection of subspaces and bagging may deliver better results in certain cases. Base classifier's diversity is a very important property of an ensemble, thus explained (Banfield et al., 2004) the advantage of the classical AdaBoost. Based on the idea proposed by Breiman (2001), Rodríguez et al. (2006) proposed a new ensemble construction method, which aims at building accurate and diverse classifiers. Its key idea is to apply feature extraction methods to the partitions divided from the original feature space, and reconstruct a new feature set for each base classifier in the ensemble. In the application part, Rodríguez et al. (2006) has proposed to apply principal component analysis (PCA) as the feature extraction method.

As Breiman (2001) proposed, Rodríguez et al. (2006) also chose the decision tree as the base classifier for their advantage of sensitivity of change of axis. Aside from PCA, other alternative feature extraction methods may be used as well (Heijden et al., 2004; Webb, 1999; Fern and Brodley 2003).

Aside from feature extractions in sub-dimensional space, there are also other proposals to improve the performance of RF. From the structure of the construction of RF, it is natural to consider the quality of base classifiers. In the following paragraph, we briefly introduce the method that we applied in our research.

Instead of majority voting, Tsymbal et al. (2006) proposed that by replacing the combination function for RF, the performance can be improved further. In classical RF, Breiman proposed a majority voting technique to combine the results of base classifiers, which is one of the most popular and simplest techniques used to combine the results of base classifiers in a classification ensemble. Another commonly employed technique is weighted voting, where the combination function will gather each vote according to the weight proportional to the estimated performance of the corresponding classifier. By applying this technique the ensemble model will usually achieve a better predictive performance than simple majority voting (Bauer, 1999). As a selective voting method, Tsymbal et al. (2006) proposed a new method which is called dynamic integration information selection. In contrast to the static methods introduced previously, this method considers each new instance to be processed into the model.

In order to combine the dynamic methods to RF, Tsymbal et al. (2006) studied the internal structure of RF. As a state-of-the-art machine learning method, it has an appealing property that each tree is built on a bootstrap sample of the original training set. Thus RF can use the remaining data (out-of-bag samples) to evaluate the base classifier's performance including correlation feature importance and marginal accuracy. The bagging algorithm is described as following: Given a training data set T of size n, bagging algorithm will generates m new training sets T_i, each of size $n' < n$, by sampling from T uniformly and with replacement. By sampling with replacement, some observations may be repeated in each T_i. Then m models are fitted using the m bootstrap samples, the ensemble will combine the results of all the base classifiers by averaging the output (for regression) or voting (for classification). Breiman (2001) suggested that by applying this technique, RF can reach an improved stability and better accuracy. It can also reduce variance and helps to avoid over fitting. By taking advantage of this property, the distance function to determine the neighborhood of the current test instance in dynamic integration can be defined. Thus a neighborhood for local performance estimates can be calculated and transformed into the voting weights. In our research we followed their steps to use the heterogeneous Euclidean distance which is defined as:

$$d_{heom}(x_1, x_2) = \sqrt{\sum\nolimits_{a=1}^{m} heom_a^2(x_1, x_2)}$$

$$heom_a(\mathbf{x}_1, \mathbf{x}_2) = \begin{cases} \text{if } a \text{ is categorical, } \begin{cases} 0, \text{ if } x_{1a} = x_{2a} \\ 1, \text{otherwise} \end{cases} \\ \text{else, } \dfrac{|x_{1a} - x_{2a}|}{range_a} \end{cases}$$

where ***x₁*** and ***x₂*** are two instances, *a* is a numeric feature, and *m* is the number of features. The Euclidean distance for numeric features and the overlap distance for categorical features were proved to be robust in many applications (Wilson et al., 1997; Shi et al. 2012). In addition, from the original construction of RF, it has an inbuilt instance similarity metric. Breiman (2001) suggested that the proportion of base classifier, i.e., the decision tree, where similarity between two cases can be measured if they fall into the same leaf. Therefore, based on their definitions and derivations, Tsymbal et al. (2006) concluded that the definition of the dynamic weight for the i_{th} model for a new instance *x* would be

$$w_i(\mathbf{x}) = \frac{\sum_{j=1}^{k} I(\mathbf{x}_j \in OOB_i) \cdot \sigma(\mathbf{x}, \mathbf{x}_j) \cdot mr_i(\mathbf{x}_j)}{\sum_{j=1}^{k} I(\mathbf{x}_j \in OOB_i) \cdot \sigma(\mathbf{x}, \mathbf{x}_j)}$$

where *k* is the size of the neighborhood, OOB_i is the set of out-of-bag instances for model *i*, *I(.)* is an indicator function, $\sigma(x, x_j)$ is a distance-based relevance coefficient and

$$mr_i(\mathbf{x}) = \begin{cases} 1, & h_i(\mathbf{x}) = y(\mathbf{x}) \\ -1, & h_i(\mathbf{x}) \neq y(\mathbf{x}) \end{cases}$$

Aside from the modification of voting procedure, recent studies put more importance on the construction of the ensemble methods, and the prediction quality of the base classifiers is also an important factor to consider. Bernard et al. (2012) proposed a new induction algorithm for RF by constructing an adaptive tree. The main idea is to guide the tree induction so that each tree will complement the existing trees in the ensemble as much as possible. As stated in Breiman's original RF paper, a classical RF will build each base classifier independently from each other. This means each new tree is arbitrarily added to the forest, and the attribute selection is purely random. Thus, it is reasonable to suspect the quality of each base classifier within the ensemble. Bernard et al. (2009) suggested that sometimes, blindly introducing such base classifiers can worsen the ensemble performance, and if we carefully select the base classifiers, and create criteria for the entrance of the base classifier, the final ensemble may outperform the classical RF. Bernard et al. (2012) further illustrates this concept by comparing the OOB prediction accuracy using the sequential forward search, which proves that there exists at least one sub-forest which outperforms the classical RF. In order to avoid this drawback, they proposed the Dynamic Random Forest, by making the tree induction dependent on the ensemble under construction. More specifically, the DRF algorithm is actually taking a weight of each randomly selected training subsample according to the predictions given by all the trees already added to the forest. The proposed contribution of each training subsample for the induction of the next tree is as:

$$c(\boldsymbol{x}, y) = \frac{1}{\mid h_{oob} \mid} \times \sum_{h_i \in h_{oob}} I(h_i(\boldsymbol{x}) = y)$$

where ***x*** is an input data point and *y* is its true class, $h_i(\mathbf{x})$ is the i_{th} classifier ouput, h_{oob} is the set of out of bag trees of ***x***.

Bernard et al. (2012) has explicitly illustrated the algorithm for constructing the DRF, which is referenced here in Algorithm 2.1.5 for comparison with previously illustrated algorithms for different variations of RF, including the original RF by Breiman (2001).

Algorithm 2.1.5. Construction of DRF

```
Let: T the training set $(\mathbf{x}_i, y_i)$
Let: N the number of training instances in T
Let: M the number of features
Let: L the number of trees in the forest to be built
Let: $W(c(\mathbf{x}, y))$ a weighting function inversely proportional to $c(\mathbf{x}, y)$
Ensure: forest the ensemble of trees that compose the forest
for all $\mathbf{x}_i$ in T do
$D_1(\mathbf{x}_i)=1/N$
```

```
    end for
    for l from 1 to L do
        $T_l$ =a bootstrap sample, made with randomly sampled (with replacement)
training        instances from T, according to a uniform distribution
    $T_l$=$T_l$ weighted with $D_l$
    tree=RandomTree($T_l$)
    fores=forest ∪ tree
    Z=0
    for all $x_i$ in T do
        if ooBTrees($\mathbf{x}_i$) is not empty then
            $D_{l+1}(\mathbf{x}_i)=W(c(\mathbf{x}_\mathbf{i}, yi))$
        Else
            $D_{l+1}(\mathbf{x}_i)= D_l(\mathbf{x}_i)$
        end if
        $Z=Z + D_{l+1}(\mathbf{x}_i)$
    end for
    for all xi in T do
        $D_{l+1}(\mathbf{x}_i)= D_{l+1}(\mathbf{x}_i)/Z$
    end for
    end for
    return forest
```

2.2 RF with Optimal Kernel Selection (RF-OK),

In addition to modify the voting methods for RF, it is natural to consider the improvement of the performance of base classifiers in an ensemble. Chen et al. (2013) proposed a new model using Fisher’s linear discriminant analysis (LDA) as the base classifier in RF. In this paper, the splitting feature of RF to construct subspaces of sample points is kept, and then they applied the Canonical LDA to each subset to serve as a base classifier of the ensemble method. A majority voting is carried out to summarize all the predictions of the base classifiers. Chen et al. (2013) applied this method to 27 real and simulated data sets and claims that Canonical Forest is significantly higher in accuracy than other ensemble methods in the majority of theses 27 data sets. According to his paper, Canonical Forest performs well in reducing variance compared to other ensemble methods.

It is well known that LDA is an efficient linear classifier and dimension reduction method for linearly separable data sets, and for linearly non-separable data sets, people proposed the kernel LDA method to solve this problem. The main idea is to first map the data points from its original space into a feature space, then apply LDA in the feature space to classify the points. Therefore picking a good kernel is the key to achieve a good classification performance. Ensemble methods such as RF may involving the splitting of the original sample space and then construct new subsample spaces for each base classifier, therefore it is natural to believe that linearly non-separable data sets may exist in such subsamples, and due to its structural nature of kernels and a huge number of subspaces in ensemble, it is not possible to manually pick kernels for each base classifier in such scenarios. Kim et al. (2006) proposed a method to numerically find the optimal kernel for Kernel Fisher’s Discriminant Analysis (KFDA) over a given convex set of kernels. The key idea is to reformulate this problem as a tractable convex optimization problem which interior-point methods can solve. Therefore, by taking the advantage of the optimal kernel selection technique for KFDA, it is possible to construct an ensemble method with KFDA as its base classifier and each of them may own its specific optimal kernel automatically selected by an optimization algorithm. This method is proposed by Kim et al. (2006) and their paper is summarized here to explain our approach. Based on their previous work, we propose a new ensemble method based on the concept explained above.

KFDA has been well studied by many researchers over recent years as an efficient classifier and dimension reduction method. Its main goal is to find a direction in the feature space $\boldsymbol{H_K}$, onto which the projections of sets $\{x_i\}^+$ and $\{x_i\}^-$ are well separated. As usual, the seperation of these two sets are measured by the ratio of the variance between class, $(w^T\mu^+ - w^T\mu^-)^2$, and the variance within the class, $w^T(\sum_+ s + \sum_- s)w$. Therefore, the KFDA will try to find an optimal direction which maximize the ratio

$$F(w,k) = \frac{\{w^T\mu_k^+ - w^T\mu_k^-\}^2}{w^T(\sum_k^- + \sum_k^- + \delta I)w}$$

where λ is a positive regularization parameter, I is the identity matrix in $\boldsymbol{H_K}$. Kim et al. (2006) shows that the weight factor

$$\boldsymbol{w}^* = \left(\sum\nolimits_k^- + \sum\nolimits_k^- + \delta I\right)^{\mathbf{-1}} (\mu_k^+ - \mu_k^-)$$

maximizes $F(w,k)$ by using Cauchy-Schwartz inequality.

Mika et al. (2003) showed that the optimal weight vector can be found within the span of the image of the training inputs through the feature mapping. That is, there exists $\alpha^* \in R^m$ such that

$$w^* = \sum\nolimits_{i=1}^{m} \alpha_i^* \, \varphi_k(x_i) = U_k \alpha^*$$

where $U_k = [\varphi_k(x_1), \dots, \varphi_k(x_m)]$. And a closed form expressions for α^* can be found (Kim et al. 2006) as:

$$\alpha^* = \frac{1}{\delta}[I - J(\delta I + J G_k J)^{-1} J G_k] a$$

Where

$$a = a_+ - a_-$$

$$a_+ = \begin{bmatrix} \left(\frac{1}{m_+}\right) 1_{m_+} \\ 0 \end{bmatrix}, a_- = \begin{bmatrix} 0 \\ \left(\frac{1}{m_-}\right) 1_{m_-} \end{bmatrix}$$

$$J = \begin{bmatrix} J_+ & 0 \\ 0 & J_- \end{bmatrix}$$

$$J_+ = \frac{1}{\sqrt{m_+}}\left(I - \frac{1}{m_+} 1_{m_+} 1_{m_+}^T\right), J_- = \frac{1}{\sqrt{m_-}}(I - \frac{1}{m_-} 1_{m_-} 1_{m_-}^T),$$

$G_k = \{G_{ij}\}$ and $G_{ij} = K(x_i, x_j), K(x_i, x_j) =< \varphi(x_i), \varphi(x_j) >_{H_k}$

Here 1_n is the vector of all ones in R^n.

It is impractical and unnecessary to consider all possible kernels, thus we can consider a special case of kernels, that is, the set of all convex kernels ***K*** consists of convex combinations of user given kernels, i.e.

$$\left\{K \middle| K = \sum_{i=1}^{n} w_i K_i\,, \sum_{i} w_i = 1, w_i > 0\right\}$$

For a given set of kernels $K_1, \ldots, K_n$,. To numerically implement this technique, Kim et al. (2006) proposed that by using the Schur complement technique (Boyd and Vandenberghe, 2004), we can rewrite this problem into an Semi definite programming problem as

$$\begin{array}{ll} \text{minimize} & (1/\lambda)\left(t - \sum_{i=1}^{p} \theta_i a^T G_i a\right) \\ \text{subject to} & H(t,\theta) \succeq 0, \\ & \theta \succeq 0, \\ & \mathbf{1}^T \theta = 1, \end{array}$$

where $t \in R$, $\theta \in R^m$ and

$$H(t,\theta) = \left[\begin{array}{cc} \lambda I + \sum_{i=1}^{p} \theta_i J G_i J & \sum_{i=1}^{p} \theta_i J G_i a \\ \sum_{i=1}^{p} \theta_i a^T G_i J & t \end{array}\right]$$

Predefining a set of kernels and finding an appropriate semi definite numerical solver (such as CSDP or SeDuMi) to solve this optimization problem will give us the optimal kernel for each individual base KFDA classifier. Therefore, the ensemble which splits the original sample space into subspaces and applies KFDA as its base classifier can be viewed as a variation of RF. The algorithm of constructing RF-OK can be expressed as in Algorithm 2.2.1.

Algorithm 2.2.1. Construction of RF-OK.

```
Training phase

Let Tn be the training data set
Let M be the number of variables
Let N be the number of rows
Let y be the response variable
Let n be the rows of bootstrap samples, n<N
Let p be the number of base classifiers
Let $m = \sqrt{M}$
For i=1 to p
    Randomly pick m variables and the response variable y, to construct the $i_{th}$ framework $T_i$
    For j=1 to n
        Randomly pick one row from the training data set Tn (taking a bootstrap sample) and then fill up the empty entries with the m corresponding variables and the response variable y and store this row into $T_i$, jth row
End for
    Train the $i_{th}$ KFDA base classifier based on sub training set $T_i$, and store it in the R
```

```
S4 object $K_i$
End for
Classification phase
Let $\mathbf{x}$ be the given prediction sample vector
For i=1 to p
    Split the vector as we did in training phase, subset $T_i$, store it into vector $\mathbf{x}_i$
    Plug $\mathbf{x}_i$ into base classifier $K_i$, calculate the classification prediction $pred_i$
End for
Summarize the results based on the prediction vector $\mathbf{pred}$ using majority voting
```

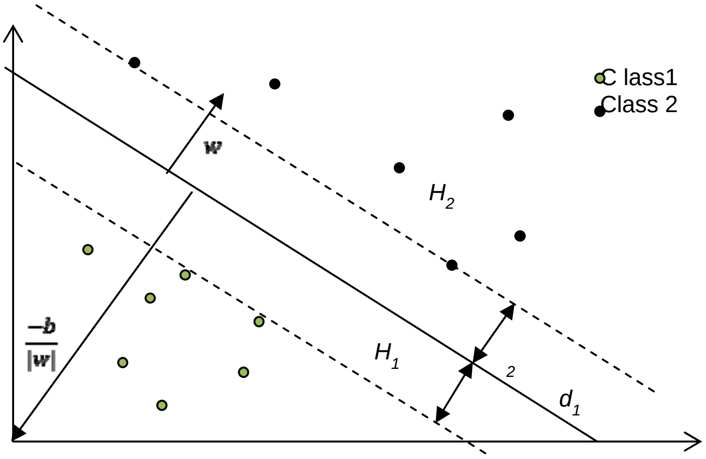
2.3 SVM and its application

The Support Vector Network is a machine learning classification model that first developed for 2-way classification problem. The key idea of SVM is by multi-dimensional splitting. Assuming the input data matrix is a combination of multidimensional data points, which can be mapped into a high dimensional feature space, SVM tries to construct a super linear decision surface based on several support vectors. This decision surface splits the training data set into two decision subspace so that a classification decision can be made.

SVM belong to the supervised learning class, which can be used for multiple application types, including classification, regression etc. The method was initially proposed by Cortes and Vapnik (1995), and was first used to solve the linearly separable data sets. They introduced the concept of classification margin and the essence of SVM - margin maximization. This methodology is then extended to data which are not fully linearly separable by introducing the kernel tricks and the soft margin concept. By implementing the idea of slack variables and the trade-off between maximizing the margin and minimizing the number of misclassified labels, SVM can be further extended to classify linearly non-separable problems and regression problems.

SVM maps the input vectors into high dimensional feature space ***Z*** through some non-linear mapping. The algorithm will construct a hyper-linear decision surface according to an optimization function which will ensure a high generalization ability of the classification machine.

For example (Cortes and Vapnik, 1995), to obtain a decision surface to a



polynomial of degree 2, we can create a feature space **Z** such that there are $N = \frac{n(n+3)}{2}$ coordinates:

$z_1 = x_1, ..., z_n = x_n$	n coordinates,
$z_{n+1} = x_1^2, ..., z_{2n} = x_n^2$	n coordinates,
$z_{2n+1} = x_1 x_2, ..., z_N = x_n x_{n-1}$	$\frac{n(n-1)}{2}$ coordinates,

where $\mathbf{x} = (x_1, ..., x_n)$. Then the hyper plane is constructed. For an input pattern, it was first transformed into some high-dimensional feature space, and then based on the optimal hyper plane constructed, the output classification result is determined.

In order to optimize the performance of SVM, the construction of a good separating hyper plane is a necessity. Fletcher (2009) briefly explained the background theory and application of constructing such planes. The algorithm can be implemented using multiple programming languages. The basic theory behind SVM is as follows: Assume we have N data points as the training data set. For each data point x_i , there are H attributes available. The response variable is of two classes y_i = -1 or +1. That is:

$$\{x_i, y_i\}\text{where } i = 1, \dots, N \text{ , } y_i \in \{-1, +1\}, x \in R^H$$

Here we assume that the data are linearly separable. Under these assumptions this hyper plane can be described by $w \cdot x + b = 0$, where w is the normal to the hyper plane. The closest data points to the hyper plane are defined as the support vectors.

The aim of SVM is to construct a plane such that this plane can be as far as possible from the closest members of both classes. In order to position the hyper plane to be as far from the support vectors as possible, we need to maximize this margin. Geometrically we can find that the margin is equal to $\frac{1}{||w||}$ (Cortes and Vapnik, 1995, Fletcher, 2009), thus we can convert this problem into an optimization problem formulated as

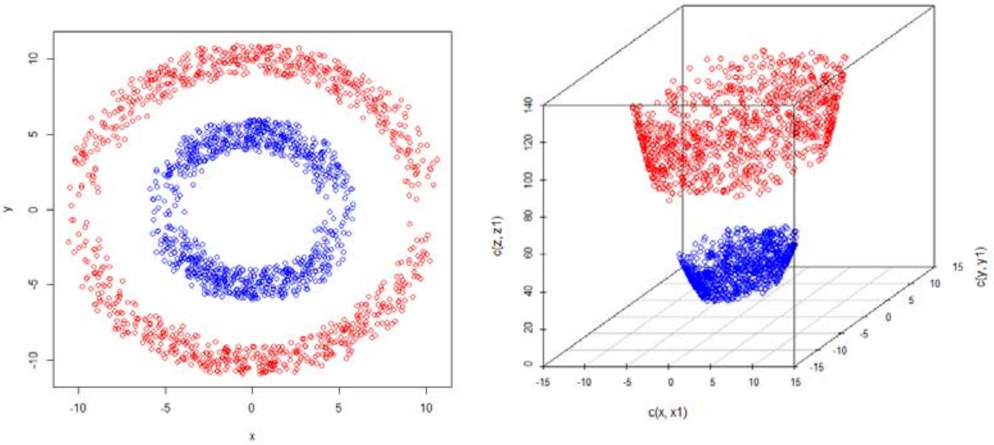
$$min||w||$$

$$\text{s.t.} \quad y_i(x_i \cdot w + b) - 1 \geq 0 \ \text{ for any } i$$

By applying quadratic programming optimization and Lagrange multipliers this problem can be rewritten into a convex quadratic optimization problem (QP), in which a QP solver will suffice to solve.

In addition to performing linear classification, SVM can also perform non-linear classification by applying the kernel trick, which implicitly mapping the data points into high-dimensional feature spaces and then perform classification analysis. In practice, it is possible that the data points are linearly non-separable. Therefore, to treat this kind of data we must start with maps and transformations of the original data. But due to insufficient memory and possible computational limitation, it is not possible to explicitly map great amounts of data points into feature space, and then perform classification analysis in the new feature space. Thus we can construct the target function in optimization problem discussed above, by applying kernel tricks to avoid such difficulty. Figure 2.3.1 is an illustration of kernel trick by converting linearly non-separable problem into linearly separable problem.

Figure 2.3.1. Illustration of kernel trick.



Dot product or inner product is an algebraic operation defined in high-dimensional feature spaces. Kernel function is used to implicitly calculate this product from its original space. In other words, it is capable of performing the multiplication in feature space without actually mapping the vectors into the feature space. Only inner products of the mapped inputs in the feature space need to be determined without the necessity to explicitly calculate the mapping. Therefore, by the help of kernel tricks and non-linearly mapping technique, we can construct the SVM in feature space for data sets that are originally linearly non-separable.

Common kernels applied in SVM are Linear, Radial basis, Polynomial, sigmoidal kernels given as:

Linear: $u^T v$

Polynomial: $(\gamma u^T v + coef0)^d$

Radial basis: $\epsilon(-\gamma|u - v|^2)$

Sigmoid: $tanh(\gamma u^T v + coef0)$

Figure 2.3.2 illustrates the difference between hard margin and soft margin in SVM,

Figure 2.3.3 illustrates the difference between large penalties and small penalties in SVM.

Figure 2.3.2. Hard margin of SVM (linearly separable) and soft margin of SVM (training error)

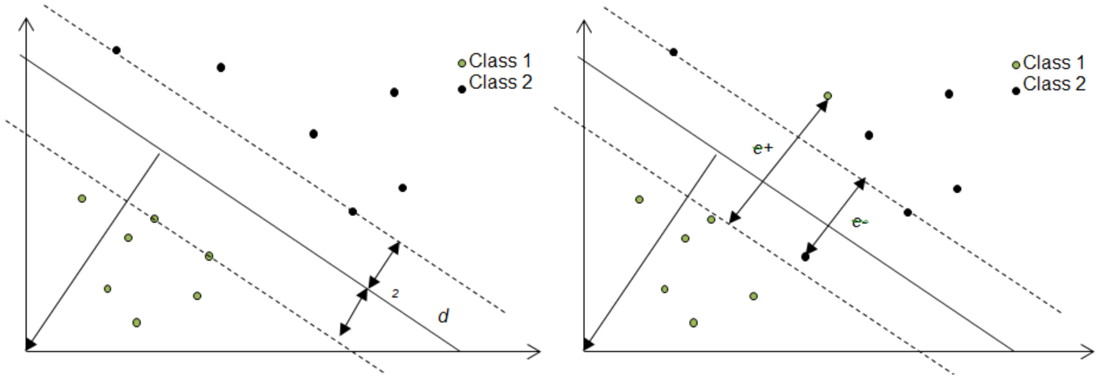
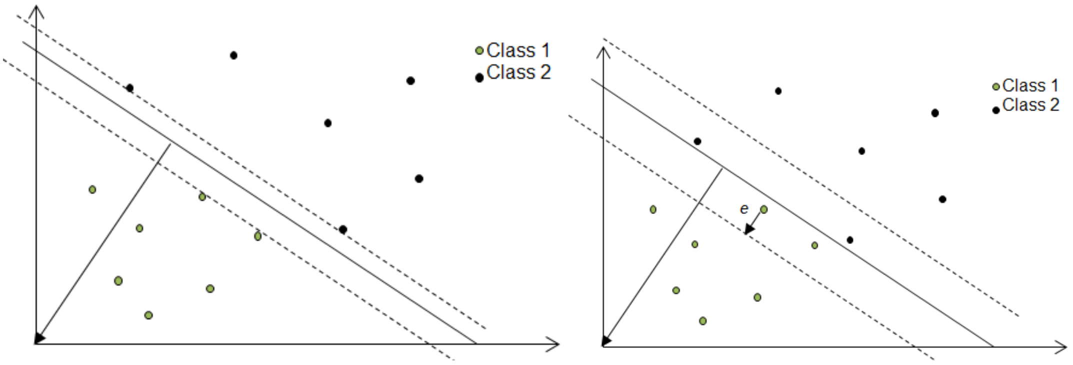


Figure 2.3.3, different penalty effect for training error, large *C* and small *C* respectively.



SVM can also be extended to multi-class classification problems. Literature suggests that there are mainly two ways to solve this problem. First is the one to others method, i.e., for each different class, take the sample with same label as one class, and all the others as the other class. For the n (n>2) class problem, n SVM classifiers will be trained and denote the i_{th} classifier as C_i. Then we pick the class which has the largest margin as the prediction.

Another method is pair wise distinguishing. Different class labels will be arranged as a tree-like structure, and multiple SVMs will be constructed for each splitting node to distinguish the classes. A bottom-up elimination tree was proposed by Pontil et al. (1998), for recognition of 3D objects and was applied to face recognition in (Guodong et al., 2000). Figure 2.3.4 (b) shows a binary tree structure for 4 classes. For a coming test data point, two pairs are compared, the winner will be tested in an upper level until the top of the tree is reached. For these two methods, usually the one to others method are favored by researchers for better time complexity. But empirical study showed that the classification performance of these two methods are much similar to each other (Nakajima et al., 2000).

Like RF, SVM can also provide variable importance analysis based on the evaluation of variable prediction power. Variable selection algorithms require a ranking criterion to distinguish important variables from less important variables. The ranking starts with a calculation of the criteria *Ct*, which resembles to ANN variable selection methods (Leray and Gallinari et al., 1999). Two approaches can be used to rank the criteria as below.

Zero order method: The algorithm repeatedly removes the variable that produces the smallest value of *Ct*. The ranking criterion will be $Rc^{(i)}=Ct^{(i)}$ when the i_{th} variable is removed.

First order method: Variables are ranked according to their influence on the criterion, which is measured by the absolute value of the derivative $Rc^{(i)}=|\bigtriangledown Ct|$

The most commonly used *Ct* is the weight vector $||w||^2$ defined as

$$R_e(i) = ||w^{(i)}||^2 = \sum_{k,j} \alpha_k^{*(i)} \alpha_j^{*(i)} y_k y_j K^{(i)}(x_k, x_j)$$

where $K^{(i)}$ is the Gram matrix of the training dataset, when the variable *i* has been removed.

SVM has been applied to multiple scenarios after its proposition, and is shown to be accurate and efficient. Byun et al. (2002) conducted a survey on variations of SVM and its applications, and found that the most commonly applied area is the pattern recognition part. Byun et al. (2002) classified the pattern recognitions into seven categories according to their aims, and the survey (partial) is summarized as follows:

Face detection

Frontal face detection

orthogonal Fourier-Mellin Moments as an input, Terrillon et al. (2000)

Combination of multiple methods

Eigenface for a coarse face detection followed by an SVM for fine detection, Li et al. (2000)

Face Verification

Reformulated Fisher's linear discriminant ratio to quadratic problem to apply SVM, Tefas et al. (2001)

Object Recognition

Input feature for SVM was extracted by PCA, Guo et al. (2001) Guodong et al. (2000)

3D range data for 3D shape features and 2D textures are projected onto PCA subspace and PC.s are input to SVMs, Wang et al. (2002)

Handwritten Character/ Digit Recognition

SVM, global view model for recognition, Choisy et al. (2001)

Speaker/ Speech Recognition

SVMs are used to accept keyword or reject non-keyword for speech recognition, Ma et al. (2001)

Combined Gaussian Mixture Model in SVM outputs text independent speaker verification, Dong et al. (2001)

Image Retrieval

Boundaries between classes were obtained by SVM, Guo et al. (2001)

SVMs were used to separate two classes of relevant and irrelevant images, Druker et al. (2002) Tian et al. (2000) Zhang et al. (2001)

Analytical Models for Understanding Misbehavior and MAC Friendliness in CSMA networks, Shi et al. (2009)

Prediction

C-ascending SVMs were suggested based on the assumption that it was better to give more weights on recent data than distant data, Tay et al. (2001)

Other Classifications

Misbehavior and MAC Friendliness in CSMA Networks Shi et al. (2007)
Competition, cooperation, and optimization in Multi-Hop CSMA networks, Shi et al. (2011)
Hyperspectral data classification, Zhang et al. (2001)
storm cell classification, Ramirez et al. (2001)
Image classification, Zhang et al. (2001)

It is a well known problem that due to different angle, lights, weather and wearing glasses or not, people's face can be much different from time to time, thus render the face recognition a very difficult problem for machines. Illuminations and other environmental parameters can also affect the machine's accuracy. Due to SVM's structural simplicity and high level degree of freedom, researchers are actively using SVM with different input features and different kernels to achieve a better performance.

Chapter 3 Prediction of potential bleeding source and cohort identification among patients with GIB data set

3.1 Explanation of the data set

Due to an aging population, acute GIB is drawing more and more attention in modern healthcare system. Further reductions in mortality will most likely require introduction of novel strategies to aid identification of the cohort requiring aggressive resuscitation and endoscopic intervention to prevent complications and death from ongoing bleeding. Delays in intervention usually result from failure to adequately recognize the source and severity of the bleed. The goal of this study is to utilize mathematical models to formulate a decision support system utilizing clinical and laboratory information to predict the source.

Chu et al. (2007) suggested a statistical method to predict the bleeding source and identify the cohort amongst patients with acute gastrointestinal bleeding (GIB) who require urgent intervention, including endoscopy. They applied RF, SVM, shrunken centroid (SC), linear discriminant analysis (LDA), k-nearest neighbor (kNN), logistic regression (logistic), boosting and ANN to their data set, and compared the performance using out of sample prediction accuracy and ROC curves. From the test results and Figure 3.1.1 to 3.1.4, Chu et al. (2007) concluded that, RF consistently provided a best prediction accuracy given the collected data set. Therefore, generalization of mathematical methods to other data sets and build a decision support system for evaluation and management of patients with acute GIB may lead to a promising future.

In our analysis, we generalized Chu et al. (2007) method to two other clinical data sets, i.e. Mayo clinic's data set (Mayo) and Blatchford's data set (Blatchford), along with Chu et al. (2007) data set (Chu). A survey was carried out and the data was collected accordingly.

In Chu's data set, there are 123 patients, 6273 entries with 852 missing values. Variables with too many missing values were deleted. 82 patients suffer from upper bowl bleeding, 30 patients suffer from lower bowl bleeding, and 11 patients suffer from middle bowl bleeding. 78 patients have received blood resuscitation and 79 patients have received urgent endoscopy. 5 patients were receiving care in home, 77 patients were receiving care in ICU, and the rest patients were receiving care in monitor floor.

In Mayo Clinic's data set, there are 400 patients, 11200 entries with 605 missing values. Variables with too many missing values were deleted. 134 patients suffer from upper bowl bleeding, 74 patients suffer from lower bowl bleeding, and 192 patients suffer from middle bowl bleeding. 29 patients have received blood resuscitation. 294 patients were receiving care in home, 89 patients were receiving care in ICU, and the rest patients were receiving care in monitor floor.

In Blatchford's data set, there are 1895 patients, 115595 entries with 87126 missing values. Variables with too many missing values were deleted. 369 patients have received blood resuscitation and 380 patients have received urgent endoscopy. 1132 patients were receiving care in home. The rest 763 patients were receiving care in ICU or in monitor floor.

Four response variables were considered to predict the risk of GIB, i.e.

- Bleeding source (Source): The irrefutable identification of a bleeding source at upper endoscopy, colonoscopy, small bowel enteroscopy or capsule endoscopy.
- Blood resuscitation (Resuscitation): Urgent blood resuscitation referred specifically to the administration of blood and blood products to correct loss of intravascular volume and presence of coagulopathy.
- Urgent endoscopy (Endoscopy): Technology used to identify patients with acute upper-GI bleeding, provide additional information about patient’s digestive system.
- Disposition: a tendency, either physical or mental, toward a given disease.

And we used patient demographics, presenting symptoms, comorbidities, clinical exam/blood tests to predict the response variables. A complete list of independent variable explanations is as the following :

- Prior GI Bleed: patient GI bleeding history.
- Hematochezia: the passage of red blood through the rectum. The cause is usually bleeding in the colon or rectum, but it may result from the loss of blood higher in the digestive tract although blood passed from the stomach or small intestine generally loses its red coloration through enzymatic activity on the erythrocytes. Cancer, colitis, and ulcers are among causes of hematochezia.

- Hematemesis: vomiting of bright red blood, indicating rapid upper GI bleeding, commonly associated with esophageal varices or peptic ulcer. The rate and the source of bleeding are determined by endoscopic examination. Any blood found in the stomach is removed by nasogastric suction.
- Melena: abnormal black tarry stool that has a distinctive odor and contains digested blood. It usually results from bleeding in the upper GI tract and is often a sign of peptic ulcer or small bowel disease.
- Duration: Duration of the symptoms.
- Syncope/Presyncope: A brief loss of consciousness caused by a sudden fall of blood pressure or failure of the cardiac systole, resulting in cerebral anemia.
- Unstable CAD: atherosclerosis of the coronary arteries, which may cause angina pectoris, myocardial infarction, and sudden death.
- COPD: Chronic obstructive pulmonary disease. A term used to describe chronic lung diseases, like chronic bronchitis, emphysema, and asthma.
- CRF: chronic renal failure (CRF), gradual loss of kidney function, with progressively more severe renal insufficiency until the stage called chronic irreversible kidney failure or end-stage renal disease.
- Risk for stress ulcer: indicator variable for whether patient exposed to the risk of developing a certain type of ulcer.
- Cirrhosis: Cirrhosis is a chronic degenerative disease in which normal liver cells are damaged and are then replaced by scar tissue.
- ASA/NSAIDS: Abbreviation for nonsteroidal antiinflammatory drugs, under drug; for example, aspirin, ibuprofen.

- PPI: Abbreviation for inorganic pyrophosphate (diphosphate).
- Systolic BP: The phase of blood circulation in which the heart's pumping chambers (ventricles) are actively pumping blood. The ventricles are squeezing (contracting) forcefully, and the pressure against the walls of the arteries is at its highest.
- Diastolic BP: The phase of blood circulation in which the heart's pumping chambers (ventricles) are being filled with blood. During this phase, the ventricles are at their most relaxed, and the pressure against the walls of the arteries is at its lowest.
- Heart Rate: heart pumping speed or frequency.
- Orthostatis: maintenance of an upright standing posture. In some medical tests a patient may need to maintain orthostasis for a long period to stimulate a rise in aldosterone concentration.
- NG lavage: the irrigation or washing out of an organ, as of the stomach or bowel.
- Rectal: Of, relating to, or situated near the rectum.
- Hct: Hematocrit, measures how much space in the blood is occupied by red blood cells. It is useful when evaluating a person for anemia.
- Cr: Conditioned reflex.
- BUN: Blood urea nitrogen, a waste product that is formed in the liver and collects in the bloodstream; patients with kidney failure have high BUN levels.

- INR: international normalized ratio index of blood coagulability (normal INR = 1); anticoagulant therapy (e.g. warfarin) adjusts INR to 2-4 (i.e. anticoagulated blood takes twice to four times as long to clot)
- Source: source of bleeding.
- Severity: indicator variable of measure of severity.
- Ulcer: defined as mucosal erosions equal to or greater than 0.5 cm of an area of the gastrointestinal tract that is usually acidic and thus extremely painful.
- Varix: refers to distended veins.
- MW tear: A Mallory-Weiss tear occurs in the mucus membrane of the lower part of the esophagus or upper part of the stomach, near where they join. The tear may bleed.
- Diverticula: an outpouching of a hollow (or a fluid filled) structure in the body. Usually implies that the structure is not normally present.
- AVM: a congenital disorder of the veins and arteries that make up the vascular system.
- Dieulafoy: a medical condition characterized by a large tortuous arteriole in the stomach wall that erodes and bleeds.

3.2 Variable selection

In RF, the rationale behind variable importance ranking is by permutation (Breiman 2001). The key idea is the following: If the variable contains crucial information regarding the response variable, then if we randomly permute it, the link between this predictor variable and response variable might be weaken, thus if we still use this predictor variable to predict the response variable then the prediction accuracy should be substantially decreased. On the other hand if this variable is not closely related to the response variable then the prediction accuracy should not be changed significantly. Several trials have been performed to evaluate the variable importance in the GIB data. The result is summarized in Table 3.2.1.

Table 3.2.1. Variable importance ranking.

Endoscopy		Disposition		Resuscitation	
	RF		RF		RF
Syncope	5.6	HR	6	Syncope	8.4
HR	5.1	Hct	5.1	DBP	6.7
DBP	4.3	SBP	4.6	HR	4
Hct	4	DBP	4.2	Hct	3.5
BUN	4	BUN	2.5	SBP	3.1
Hematemesis	3.4	Cr	2.3	Hematemesis	2
SBP	3.2	Syncope	2.1	BUN	1.6
Cr	1.9	Melena	0.9	Cr	0.9
PPI	0.8	Hematemesis	0.8	Melena	0.5
Hx_of_GIB	0.7	Hx_of_GIB	0.6	Hx_of_GIB	0.5
Cirrhosis	0.7	Unstable_CAD	0.5	PPI	0.4
Sex	0.6	Cirrhosis	0.5	ASA_NSAID	0.3
ASA_NSAID	0.4	PPI	0.4	Cirrhosis	0.3
Unstable_CAD	0.36	Sex	0.3	Sex	0.3
Melena	0.3	ASA_NSAID	0.3	Unstable_CAD	0.2

Figure 3.2.2. Variable importance ranking for Endoscopy.

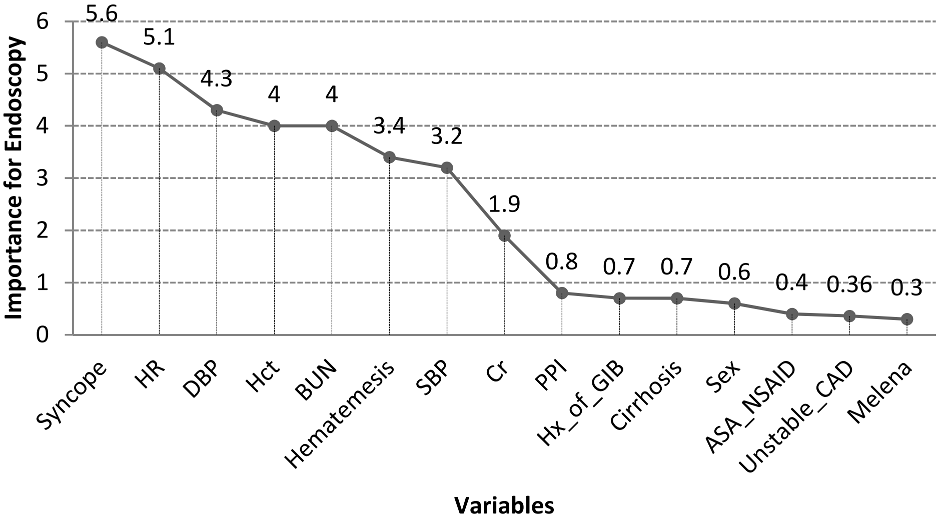


Figure 3.2.3. Variable importance ranking for Disposition.

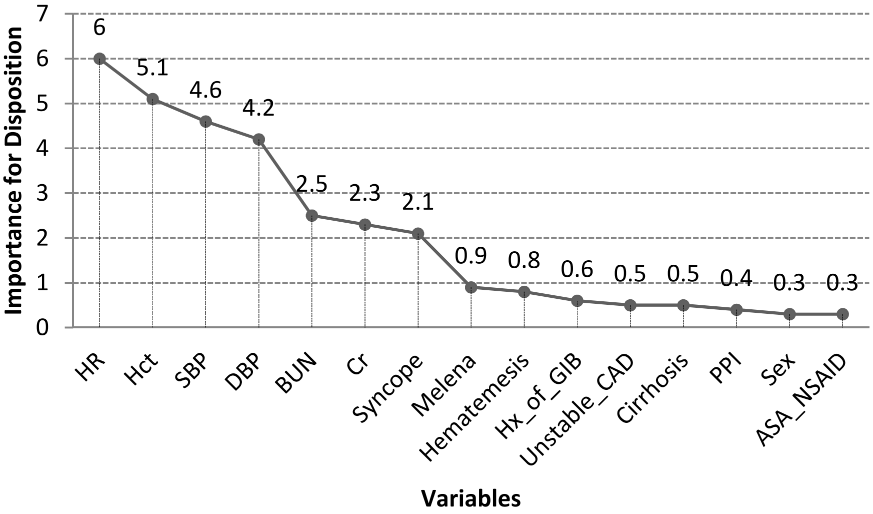
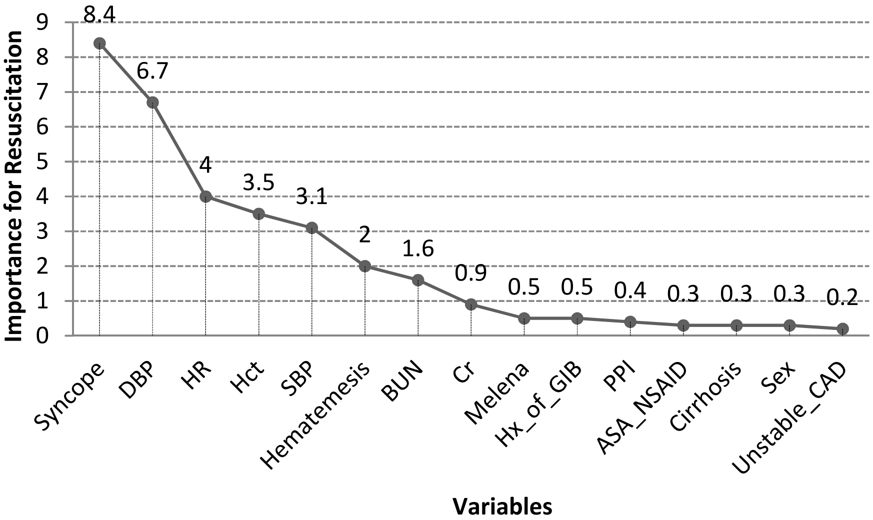


Figure 3.2.4. Variable importance ranking for Resuscitation.



As we can see from Table 3.2.1 and Figure 3.2.2 to 3.2.4, the variable importance ranking by RF agrees with the result in Chu et al. (2007), but several important variables for the prediction are missing. This may negatively affect the prediction results.

3.3 Statistical analysis and modeling

We applied RF and SVM to analyze the data set using 10 fold cross validation. Model training was performed on a randomly selected subset of patients, and tested on the remaining data set. Categorical variables were changed to indicator/dummy variables. For variable with missing data, we used median to fit in the empty entry if the variable is numerical, and we used mode if the variable is categorical. Multiple runs of 10-fold cross validation (CV) were performed. For every 10-fold CV, the following statistics were calculated: SN, SP and ACC.

For SVM, the R-software package e1071 is used. Variables were scaled and the tolerance level was set to .001. Radial, Sigmoid and linear kernels were considered. The program can dynamically search for optimal kernel parameters for different data sets. The results are summarized in Table 3.3.1 to Table 3.3.4.

Table 3.3.1. Prediction statistics for Resuscitation, train on Mayo and Chu merged dataset, 10-fold cross validation 50 repetitions.

Weighted_SVM	Mean	Std	Min	Max
Overall	0.823	0.041	0.743	0.895
RF	Mean	Std	Min	Max
Overall	0.808	0.041	0.695	0.876

Table 3.3.2. Prediction statistics for Endoscopy, train on Mayo and Chu merged dataset, 10-fold cross validation 50 repetitions.

Weighted_SVM	Mean	Std	Min	Max
Overall	0.888	0.027	0.81	0.943
Sensitivity	0.874	0.038	0.791	0.953
Specificity	0.937	0.05	0.824	1
PPV	0.66	0.08	0.486	0.818
NPV	0.983	0.013	0.957	1
RF	Mean	Std	Min	Max
Overall	0.952	0.018	0.905	0.99
Sensitivity	0.969	0.013	0.94	0.989
Specificity	0.884	0.072	0.6	1
PPV	0.878	0.048	0.778	0.962
NPV	0.97	0.019	0.912	1

Table 3.3.3. Prediction statistics for Disposition, train on Mayo and Chu merged dataset, 10-fold cross validation 50 repetitions.

Weighted_SVM	Mean	Std	Min	Max
Overall	0.745	0.049	0.61	0.848
RF	Mean	Std	Min	Max
Overall	0.765	0.037	0.676	0.848

Table 3.3.4. Prediction statistics for Source of Bleeding, train on Mayo and Chu merged dataset, 10-fold cross validation 50 repetitions.

Weighted_SVM	Mean	Std	Min	Max
Overall	0.587	0.046	0.495	0.667
RF	Mean	Std	Min	Max
Overall	0.634	0.045	0.543	0.743

We have also applied RF to Blatchford's data set, Mayo clinic data set and Chu et al. (2007) data, the parameters are adjusted through several cross validation trials: 500 trees were grown, the number of variables randomly sampled at each node was $\sqrt{p}$, where p is the number of explanatory variables, the option of variable importance was set to True, proximity was set to be True so that the proximity measure among the rows will be calculated. The out of sample prediction results are summarized in Table 3.3.1 to 3.3.6.

Table 3.3.5. Prediction statistics for Resuscitation, Train on Adrienne, test on Adrienne, 10-fold cross validation 50 repetitions.

RF	Mean	Std	Min	Max
Overall	0.918	0.058	0.76	1
Sensitivity	0.926	0.086	0.615	1
Specificity	0.918	0.079	0.647	1
PPV	0.951	0.059	0.706	1
NPV	0.876	0.108	0.571	1

Table 3.3.6. Prediction statistics for Endoscopy, Train on Adrienne, test on Adrienne, 10-fold cross validation 50 repetitions.

RF	Mean	Std	Min	Max
Overall	0.81	0.072	0.6	0.96
Sensitivity	0.712	0.176	0.222	1
Specificity	0.869	0.086	0.556	1
PPV	0.851	0.097	0.55	1
NPV	0.741	0.146	0.429	1

One more difficulty we encountered is that in these datasets, the response variables are unbalanced. Most of the classification models tend to predict individuals with low prediction confidence to the majority class to improve the overall accuracy. In order to overcome this, we used different decision thresholds in RF’s prediction, that is, in RF, individual decision trees were generated and classification is done by majority voting within the forest of these trees. We can modify this voting procedure by setting different thresholds and pick the threshold that optimizes our prediction accuracy. The results with different threshold are summarized in Figures 3.3.7 to 3.3.12 and Tables 3.3.13 to 3.3.15.

Figure 3.3.7. Sensitivity for Blatchford's data set changing threshold RF.

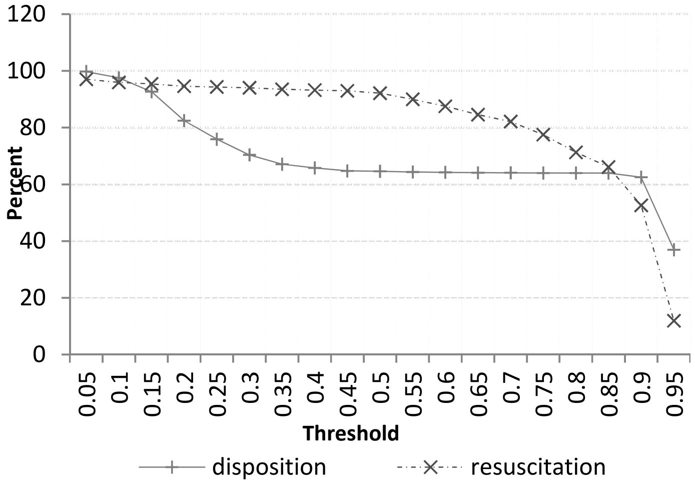


Figure 3.3.8. Specificity for Blatchford's data set changing threshold RF.

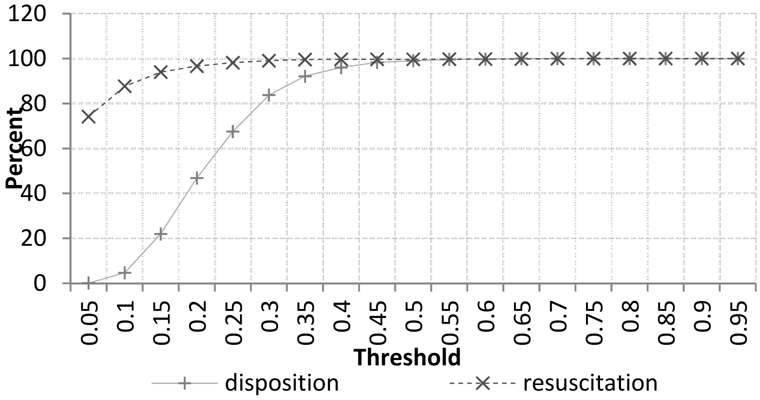


Figure 3.3.9. NPV for Blatchford's data set changing threshold of RF.

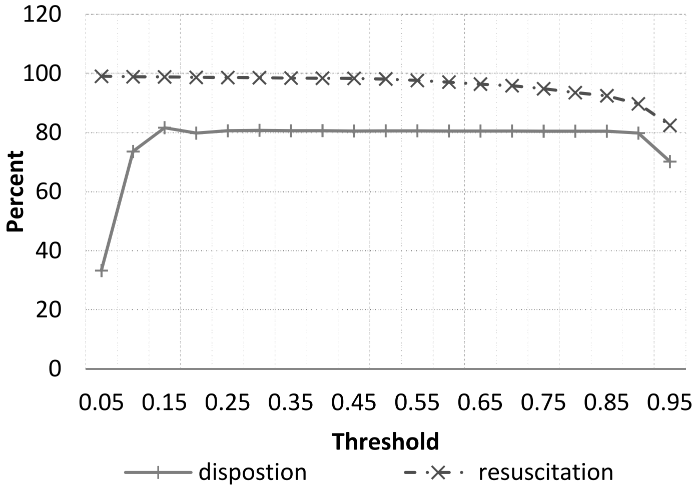


Figure 3.3.10. NPV for RF train on Chu’s data set and test on Blatchford's data set.

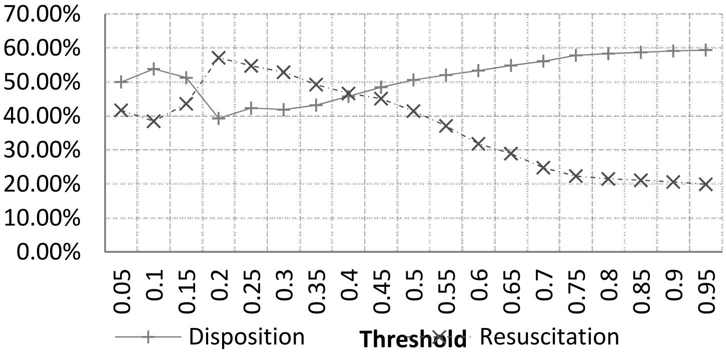


Figure 3.3.11. Sensitivity for RF train on Chu’s data set and test on Blatchford's data set.

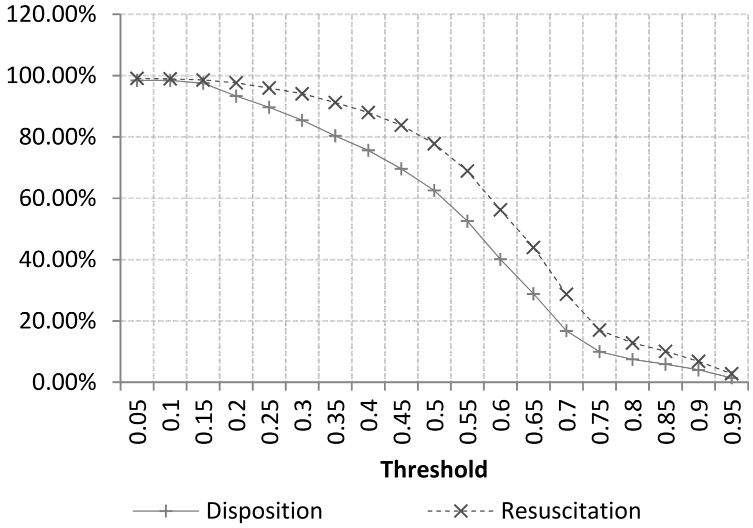


Figure 3.3.12. Specificity for RF train on Chu's data set and test on Blatchford's data set.

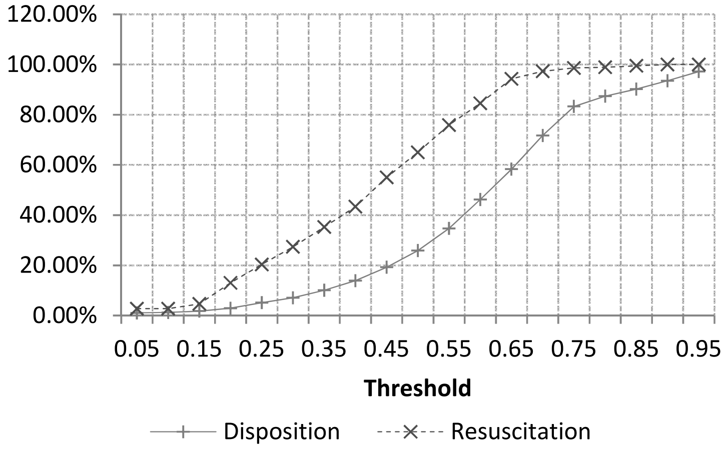


Table 3.3.13. Prediction statistics for 10-fold CV of RF for Disposition changing thresholds For Blatchford's dataset

Variable	threshold	sensitivity	specificity	NPV	accuracy
Disposition	0.05	99.74	0.09	33.33	40.21
Disposition	0.1	97.51	4.68	73.61	42.06
Disposition	0.15	92.66	22	81.64	50.45
Disposition	0.2	82.44	46.82	79.82	61.16
Disposition	0.25	75.88	67.58	80.61	70.92
Disposition	0.3	70.38	83.75	80.75	78.36
Disposition	0.35	67.1	92.14	80.6	82.06

Disposition	0.4	65.79	96.02	80.64	83.85
Disposition	0.45	64.74	98.32	80.54	84.8
Disposition	0.5	64.61	99.03	80.59	85.17
Disposition	0.55	64.35	99.65	80.57	85.44
Disposition	0.6	64.22	99.73	80.53	85.44
Disposition	0.65	64.09	100	80.51	85.54
Disposition	0.7	64.09	100	80.51	85.54
Disposition	0.75	63.96	100	80.45	85.49
Disposition	0.8	63.96	100	80.45	85.49
Disposition	0.85	63.96	100	80.45	85.49
Disposition	0.9	62.52	100	79.83	84.91
Disposition	0.95	36.96	100	70.18	74.62

Note: From the cut off analysis we can see that there is not much difference from threshold 0.5 to 0.9, the model performance may achieve an optimal prediction accuracy at 0.5 empirically.

Table 3.3.14. Prediction statistics for 10-fold CV of RF for Resuscitation changing thresholds For Blatchford’s dataset

Variable	threshold	sensitivity	specificity	NPV	accuracy
Resuscitation	0.05	97.02	74.18	99.04	93.07
Resuscitation	0.1	95.93	87.68	98.89	89.29
Resuscitation	0.15	95.39	93.97	98.83	94.25
Resuscitation	0.2	94.58	96.66	98.66	96.25
Resuscitation	0.25	94.31	98.17	98.62	97.41
Resuscitation	0.3	94.04	99.08	98.57	98.1
Resuscitation	0.35	93.5	99.54	98.44	98.36
Resuscitation	0.4	93.22	99.67	98.38	98.42
Resuscitation	0.45	92.95	99.67	98.32	98.36
Resuscitation	0.5	92.14	99.67	98.13	98.21
Resuscitation	0.55	89.97	99.74	97.63	97.84
Resuscitation	0.6	87.53	99.8	97.07	97.41
Resuscitation	0.65	84.55	99.8	96.39	96.83
Resuscitation	0.7	82.11	99.93	95.85	96.46
Resuscitation	0.75	77.51	100	94.84	95.62
Resuscitation	0.8	71.27	100	93.5	94.41
Resuscitation	0.85	66.12	100	92.43	93.4

Resuscitation	0.9	52.57	100	89.71	90.77
Resuscitation	0.95	11.92	100	82.44	82.85

Table 3.3.15. Prediction statistics for 10-fold CV of RF for Resuscitation changing thresholds

Variable	threshold	sensitivity	specificity	NPV	accuracy
Resuscitation	0.05	99.08	2.71	41.67	80.32
Resuscitation	0.1	98.95	2.71	38.46	80.21
Resuscitation	0.15	98.56	4.61	43.59	80.26
Resuscitation	0.2	97.64	13.01	57.14	81.16
Resuscitation	0.25	95.94	20.33	54.74	81.21
Resuscitation	0.3	94.1	27.37	52.88	81.11
Resuscitation	0.35	91.22	35.23	49.24	80.32
Resuscitation	0.4	88.01	43.36	46.65	79.31
Resuscitation	0.45	83.81	55.01	45.11	78.21
Resuscitation	0.5	77.79	65.04	41.45	75.3
Resuscitation	0.55	68.87	75.88	37.09	70.24
Resuscitation	0.6	56.23	84.55	31.84	61.74
Resuscitation	0.65	43.97	94.31	28.93	53.77
Resuscitation	0.7	28.7	97.29	24.81	42.06
Resuscitation	0.75	17.04	98.65	22.33	32.93
Resuscitation	0.8	12.84	98.92	21.53	29.6
Resuscitation	0.85	10.09	99.46	21.1	27.49
Resuscitation	0.9	6.82	100	20.6	24.96
Resuscitation	0.95	2.82	100	19.92	21.74

By choosing an optimal threshold for each response variable in Figure 3.3.7 to 3.3.12, we can remarkably increase the diagnostic statistics for a certain type of response variable with an acceptable cost of other test statistics and provide a suitable diagnostic in accordance with the priority. In clinical trials, for example, people are more concerned with life threatening emergencies, thus, a high NPV is more preferable than a high PPV. Doctors can make a more confident decision based on a high NPV to relieve a patient from ICU and rearrange the limited medical resource to more needing instances.

3.4 Model validation and prediction

Figure 3.4.1. Illustration of a 3-fold Cross Validation procedure.

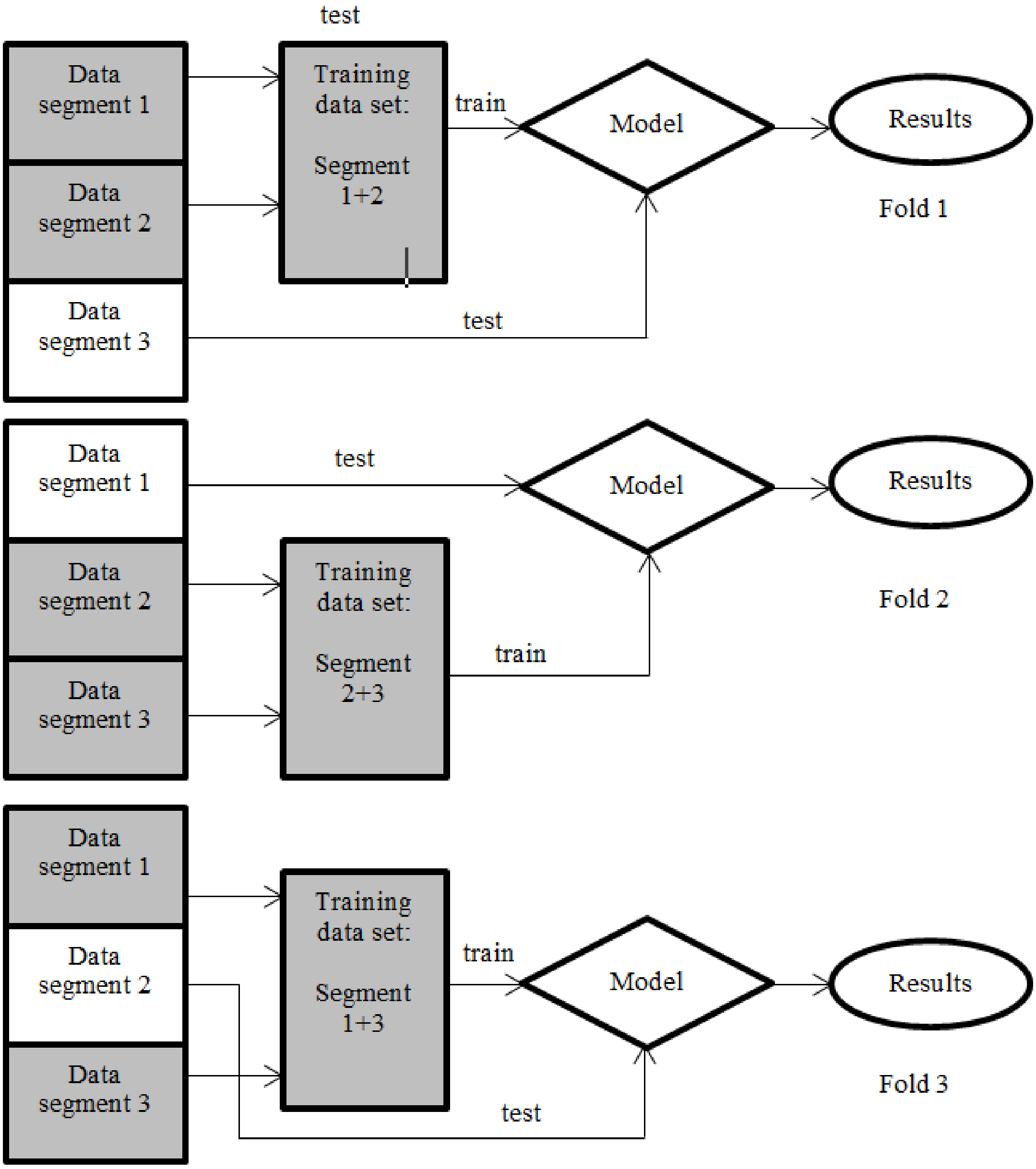


Table 3.3.1 through Table 3.3.6 summarizes the results for each response variable of each data set. Both models have achieved accuracy higher than 80% for Disposition and Resuscitation. In Table 3.3.3 and 3.3.4, the low performance may be because of different measure scales used in different data sets, or missing of important predictors.

From the tables we can see that the performance of RF in Chu's data set is very good for all of the test statistics, Resuscitation prediction accuracy is over 91%, and Endoscopy prediction accuracy is over 80%. However, the sensitivity for Endoscopy is around 70%. For models trained on Mayo clinic's data set and tested on Chu's data set, SVM and RF yield similar results for Resuscitation and Disposition. RF can achieve better prediction accuracy in Endoscopy than weighted SVM. The reason might because the importance ranking of predictors for these two models is different. For Source of bleeding, the prediction accuracy for both models is not good. One potential reason is that the predictors used may not provide enough information for this response variable. The absolute value of predictors in importance ranking is not as high as those calculated for other response variables.

Chapter 4 Prediction of the risk for developing schizophrenia dataset

4.1 Explanation of the data set

There were 165 cases in this dataset. Among them, 51 cases were healthy control, 19 cases were schizoaffective, 60 cases were schizophrenia patients, 20 cases were bipolar patients and 14 cases were MDD patients. We had one missing observation. Our data set contains demographic information, Neuropsychological Test Scores, PANSS scales and global assessment function of current and past scores. A summary of the data is given in Table 4.1.1.

Table 4.1.1, Summary of the base statistics of variables for Schizo dataset.

	AGE	ONSET	MOM AGE	DAD AGE	MNTHRES	UPSIT	FSIQ	PTOT	NTOT	GTOT	GAFCUR	GAFPAS
N	165	103	122	115	128	144	132	136	138	138	59	61
Mean	35.1	22.5	27.8	31.4	4.5	32	100.6	10.4	11.7	23.9	37.3	46
SD	11	7.3	6.4	6.9	1.5	4.4	17.6	5.6	5.5	8.8	16.1	12.5

We have included the following variables to explain the potential risks of developing schizophrenia:

- BRAGE_M: Mother’s age at birth
- BRAGE_F: Father’s age at birth
- FAMHXANY: Family history of schizophrenia
- MNTHRES: Mean OLF Threshold

- UPSIT: University of Pennsylvania Smell Identification Test
- VIQ: Verbal IQ
- PIQ: Performance IQ
- VIQP: Verbal ‐ performance differential score
- FSIQ: Full Scale Intelligence Quotient
- VCI, POI, WMI, PSI, WSINFS, WSPIXCS, WSDSS, WSPIXAS, WSVOCS, WSBDS, WSARITHS, WSOBASSS, WSCOMPS, WSDSYMS, WSSIMILS, WSMATRS, WSLETNMS, WSSYMSES, WSSYMBS: verbal subtests (arithmetic, digit span, information, vocabulary, comprehension, similarities) and the performance subtests (object assembly, picture arrangement, picture completion, digit symbol, block design)
- PTOT: Positive Syndrome Scale
- NTOT: Negative Syndrome Scale
- GTOT: General Syndrome Scale
- GAF ‐ CUR: Global Assessment Function, current
- GAF ‐ PAS: Global Assessment Function, past
- DIAGALL: control (0), schizo-affective (1), schizo (2), bipolar (3), MDD (4)

And the base statistics are summarized in Table 4.1.2

Table 4.1.2, base statistics for variables in Schizo dataset in different response classes.

		AGE	ONSET	MOM AGE	DAD AGE	MNTHRES	UPSIT	FSIQ	PTOT	NTOT	GTOT	GAFCUR	GAFPAS
0	N	51	0	40	37	41	44	37	40	41	41	0	0
	Mean	33.1		29.2	32.2	4.5	33	106	7	7.6	16.8		
	SD	11.6		6.3	6.7	1.2	4	13	0	1.6	1.7		
1	N	19	17	11	10	17	17	17	15	15	15	12	12
	Mean	36.5	22.1	24.3	27.1	4.3	31.9	95.6	12.4	13.7	27.4	32.3	39.8
	SD	9.3	7.7	5	7.2	1.3	4.1	19.7	4.4	4.9	8.1	9.6	8.7
2	N	60	58	46	46	46	54	50	57	57	57	32	34
	Mean	35.1	23.2	27.8	32.6	4.4	31	96.7	12.9	13.6	26.5	31	43.7
	SD	10.7	6.1	6.5	6.3	1.9	5.1	20.1	7	5.8	10	12	10.4
3	N	20	16	16	15	12	16	18	12	13	13	9	9
	Mean	34.5	21.2	27.1	30.7	4.6	32.9	100.4	8.4	11.9	25.8	55.2	56.3
	SD	8.9	8	7.6	8.1	0.9	3.2	13.9	1.7	5.8	4.2	14.3	12.8
4	N	14	12	9	7	12	12	10	12	12	12	6	6
	Mean	40.5	21.6	26.7	27.1	4.5	31.8	109	9.8	14.1	29.1	54.5	55.8
	SD	14	10.8	4.7	5.9	1	2.9	15.7	4.6	5.9	9	17.9	16.9

4.2 Variable selection

As we introduced in section 3.2, we first need to calculate the measure of importance for each independent variable, and see the amount of information provided by each predictor for each response variable.

Table 4.2.1, RF variable importance ranking based on GINI information, for case 1, Patient=0, Control=1

	MeanDecreaseAccuracy		MeanDecreaseAccuracy
GTOT	1.60249	wscomps	0.11387
NTOT	0.82428	wmi	0.06118
FAMHXANY	0.70125	vci	0.05366
wssymbs	0.56408	psi	0.05126
PTOT	0.50378	wsvocs	-0.0072
poi	0.41329	MNTHRES	-0.0259
wsariths	0.40423	upsittot	-0.0412
wsdsyms	0.33302	wsbds	-0.1069
viq	0.23051	BRAGE_M	-0.1353
wsletnms	0.22943	wsinfs	-0.1591
wssimils	0.22209	wspixas	-0.1735
fsiq	0.18969	SEX	-0.1996
wsmatrs	0.17318	wsobasss	-0.2998
piq	0.12694	wsdss	-0.3154
wspixcs	0.11738	BRAGE_F	-0.3171

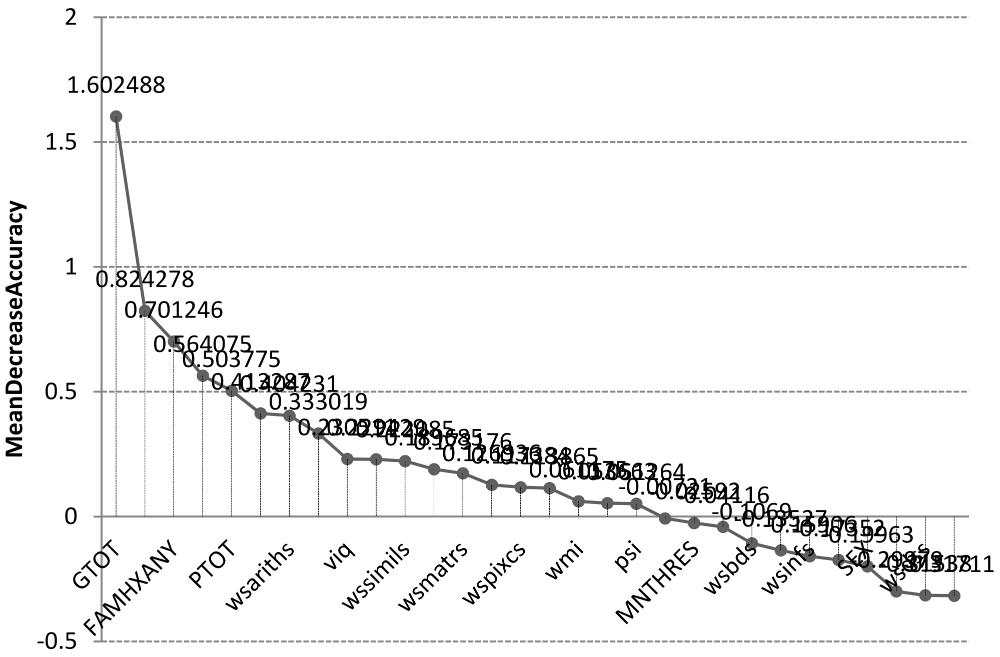


Table 4.2.2, RF variable importance ranking based on GINI information, for case 2, Schizophrenia=0, other patients=1 (Schizoaffective, Bipolar, MDD)

	MeanDecreaseAccuracy		MeanDecreaseAccuracy
FAMHXANY	2.06931	wspixas	0.4761
BRAGE_F	1.4703	wsvocs	0.45682
piq	1.02641	poi	0.44095
wssimils	0.98579	viq	0.43473
GTOT	0.97018	wsinfs	0.41911
GAF_CUR	0.96445	wmi	0.4011
GAF_PAS	0.83952	wsdss	0.30026
wsdsyms	0.81262	NTOT	0.26068
wsletnms	0.8096	wsbds	0.20949
wsariths	0.80477	BRAGE_M	0.09341
vci	0.69907	wscomps	0.0826
psi	0.63355	SEX	0.00286
wssymbs	0.61744	wsmatrs	-0.1066
PTOT	0.60839	wspixcs	-0.1236
MNTHRES	0.60402	wsobasss	-0.2416
fsiq	0.51841	upsittot	-0.4441

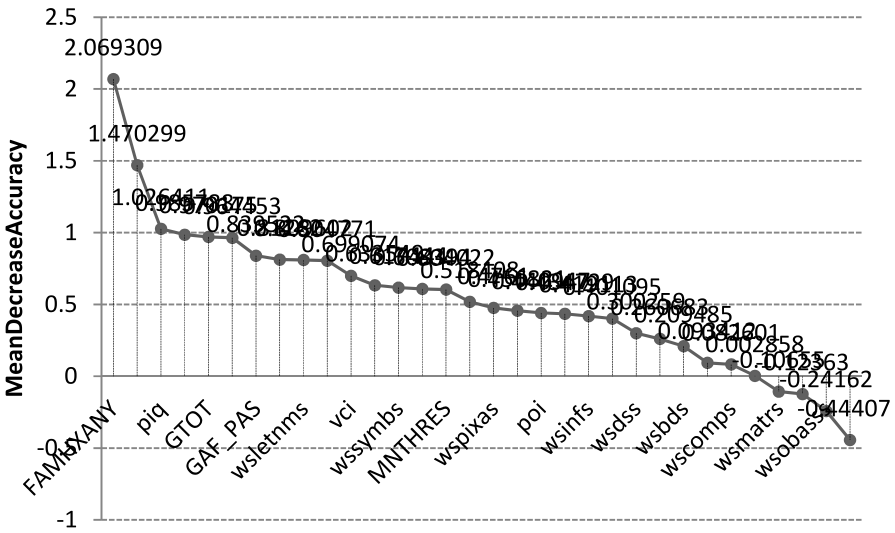
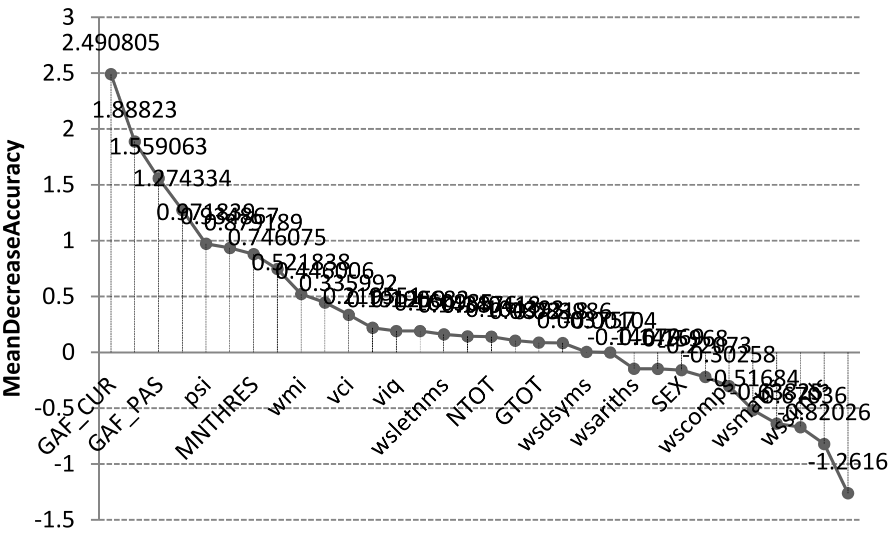


Table 4.2.3, RF variable importance ranking based on GINI information, for case 3, Schizophrenia & Schizoaffective=0, Bipolar & MDD=1

	MeanDecreaseAccuracy		MeanDecreaseAccuracy
GAF_CUR	2.49081	NTOT	0.14042
FAMHXANY	1.88823	piq	0.10439
GAF_PAS	1.55906	GTOT	0.08722
wsvocs	1.27433	fsiq	0.08399
psi	0.97184	wsdsyms	0.00376
PTOT	0.93487	wspixas	-0.001
MNTHRES	0.87919	wsariths	-0.1468
wssymbs	0.74608	wsinfs	-0.1477
wmi	0.52184	SEX	-0.1597
upsittot	0.44601	wsbds	-0.2207
vci	0.33599	wscomps	-0.3026
BRAGE_F	0.21955	BRAGE_M	-0.5168
viq	0.19126	wsmatrs	-0.6383
wspixcs	0.19068	wsdss	-0.6704
wsletnms	0.16099	wssimils	-0.8203
poi	0.14388	wsobasss	-1.2616



4.3 Statistical analysis and modeling

In this study, we applied RF model to predict the risk for developing schizophrenia. To evaluate the performance of RF, we ran multiple rounds of 10-fold cross validation and checked the out of sample prediction accuracy. Based on this evaluation, we picked the model with the best prediction performance, and then applied this model to the control group. In this way, we can assess the risk of developing schizophrenia among the participants who are currently in the healthy control group.

To achieve this goal, we first divide the participants into two groups: schizophrenia patients and the healthy control group based on doctor’s diagnoses. Then, we use the symptoms of schizophrenia patients to further split the patient group into several small groups. Thus, we can split this problem into 3 different scenarios:

1) Patient=0, Control=1
2) Schizophrenia=0, other patients=1 (Schizoaffective, Bipolar, MDD)

3) Schizophrenia & Schizoaffective=0, Bipolar & MDD=1

In each scenario, we trained and tested RF model using 10-fold CV with 50 repetitions, the test statistics were summarized. The analysis was performed using R package RandomForest (Liaw and Wiener, 2002) with all default parameters. Independent variables included in the model were: Paternal age, Maternal age, Family history, Mean OLF Threshold ,SEX, UPSIT, VIQ, PIQ, VCI, POI, WMI, PSI, WSINFS, WSPIXCS, WSDSS, WSPIXAS, WSVOCS, WSBDS, WSARITHS, WSOBASSS, WSCOMPS, WSDSYMS, WSSIMILS, WSMATRS, WSLETNMS, WSSYMSES, WSSYMBS, PTOT, NTOT and GTOT.

Another state-of-the-art machine learning model we applied to this data set was SVM. In SVM, weights are a biasing mechanism for specifying the relative importance of target values (classes). By default, SVM will automatically assign equal weights to all response classes. However, if the training data does not represent a realistic distribution, one can bias the model to compensate for class values that are under-represented. If we increase the weight for a class, the percent of correct predictions for that class will increase.

For 2-way classification, given a training set of instance-label pairs (x_i, y_i); $i = 1, \ldots, l,$ where x_i belongs to R^n and y_i belongs to {1,-1}, SVM (Cortes and Vapnik, 1995) require the solution of the following optimization problem:

$$\begin{aligned} \min_{\mathbf{w},b,\boldsymbol{\xi}} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l}\xi_i \\ \text{subject to} \quad & y_i(\mathbf{w}^T\phi(\mathbf{x}_i) + b) \geq 1 - \xi_i, \\ & \xi_i \geq 0. \end{aligned}$$

we can rewrite this target function into:

$$\min_{w,b,\xi} \frac{1}{2}||w||^2 + C_1 \sum_{\xi_i:y_i=-1}^{l} \xi_i + C_2 \sum_{\xi_i:y_i=+1}^{l} \xi_i$$

subject to $y_i(w^T\varphi(x_i) + b) \geq 1 - \varepsilon_i$

Therefore, we can choose constants C_1 and C_2 inversely proportional to the class sizes. That is, if we have l_1 training samples in class *1* and l_2 in class *2*, assign C_1 and C_2 such that $C_1/C_2 = l_2/l_1$.

In this study, we applied SVM, with and without class weight adjustment, to the same training and testing data sets. Their performance is summarized in Table 4.3.1.

Table 4.3.1, Out of sample prediction statistics of SVM in different scenarios.

	SVM with class weight performance						SVM without class weight performance				
Case 1		Mean	Std	Min	Max	Case 1		Mean	Std	Min	Max
	Overall	0.808	0.068	0.636	0.939		Overall	0.753	0.075	0.545	0.939
	Sensitivity	0.786	0.088	0.571	0.957		Sensitivity	0.886	0.072	0.68	1
	Specificity	0.851	0.117	0.429	1		Specificity	0.473	0.16	0	0.857
	PPV	0.646	0.123	0.375	0.9		PPV	0.657	0.2	0	1
	NPV	0.921	0.063	0.765	1		NPV	0.788	0.086	0.531	0.96
Case 2		Mean	Std	Min	Max	Case 2		Mean	Std	Min	Max
	Overall	0.741	0.068	0.609	0.87		Overall	0.694	0.079	0.478	0.87
	Sensitivity	0.771	0.126	0.5	1		Sensitivity	0.684	0.141	0.357	1
	Specificity	0.718	0.13	0.462	1		Specificity	0.723	0.118	0.429	1
	PPV	0.779	0.108	0.5	1		PPV	0.718	0.127	0.429	1
	NPV	0.723	0.114	0.467	1		NPV	0.686	0.133	0.385	1
Case3		Mean	Std	Min	Max	Case 3		Mean	Std	Min	Max
	Overall	0.704	0.069	0.522	0.87		Overall	0.698	0.072	0.565	0.826
	Sensitivity	0.646	0.198	0.2	1		Sensitivity	0.145	0.132	0	0.6
	Specificity	0.73	0.104	0.5	0.938		Specificity	0.944	0.057	0.765	1
	PPV	0.824	0.089	0.583	1		PPV	0.718	0.082	0.55	0.889
	NPV	0.524	0.139	0.2	0.833		NPV	NA	NA	NA	NA

Table 4.3.1 shows that SVM with class weight adjustment constantly report a higher prediction accuracy than SVM without weight adjustment. This phenomenon is result from the imbalance of the original data set. The ratio of patients that are schizo/schizo effective over Bip/MDD is 2.31/1, and as a result, SVM model have a great tendency to predict majority class, which leads to a very low sensitivity. In detecting potential patient stage, SVM without class weight adjustment will classify nearly every person as potential patient if we don’t assign appropriate class weights to the model.

4.4 Model validation

In our study, the performance of machine learning models is evaluated by ACC, SN, SP, PPV and NPV.

Overall accuracy (ACC) is obtained by the total number of correct predictions divided by total number of predictions.

Sensitivity (SN) measures the proportion of actual positives which are correctly identified as such. Specificity (SP) measures the proportion of actual negatives which are correctly identified as such.

Positive predictive value (PPV) is the proportion of true positives among the positive predictions and negative predictive value (NPV) is the proportion of true negatives among the negative predictions.

Table 4.4.1, Out of sample prediction statistics in Model validation part for case 1, patient=0, control=1

SVM	Mean	Std	Min	Max
Overall	0.808	0.068	0.636	0.939
Sensitivity	0.786	0.088	0.571	0.957
Specificity	0.851	0.117	0.429	1
PPV	0.646	0.123	0.375	0.9
NPV	0.921	0.063	0.765	1
RF	Mean	Std	Min	Max
Overall	0.845	0.011	0.829	0.866
Sensitivity	0.891	0.008	0.876	0.903
Specificity	0.744	0.029	0.706	0.804
PPV	0.755	0.017	0.725	0.784
NPV	0.885	0.012	0.87	0.91

Table 4.4.2, Out of sample prediction statistics in Model validation part for case 2, Schizophrenia=0, other patients=1 (Schizoaffective, Bipolar, MDD)

SVM	Mean	Std	Min	Max
Overall	0.741	0.068	0.609	0.87
Sensitivity	0.771	0.126	0.5	1
Specificity	0.718	0.13	0.462	1
PPV	0.779	0.108	0.5	1
NPV	0.723	0.114	0.467	1
RF	Mean	Std	Min	Max
Overall	0.701	0.013	0.673	0.717
Sensitivity	0.684	0.026	0.623	0.717
Specificity	0.716	0.025	0.683	0.767
PPV	0.72	0.014	0.695	0.741
NPV	0.681	0.017	0.648	0.708

Table 4.4.3, Out of sample prediction statistics in Model validation part for case 3, Schizophrenia & Schizoaffective=0, Bipolar & MDD=1

SVM	Mean	Std	Min	Max
Overall	0.704	0.069	0.522	0.87
Sensitivity	0.646	0.198	0.2	1
Specificity	0.73	0.104	0.5	0.938
PPV	0.824	0.089	0.583	1
NPV	0.524	0.139	0.2	0.833
RF	Mean	Std	Min	Max
Overall	0.73	0.011	0.708	0.752
Sensitivity	0.365	0.026	0.294	0.412
Specificity	0.887	0.013	0.861	0.911
PPV	0.765	0.007	0.747	0.778
NPV	0.583	0.029	0.522	0.65

Table 4.4.1, 4.4.2 and 4.4.3 suggests that the prediction accuracy for distinguishing patients and healthy control is over 80% for both models. In case 2, SVM achieved a higher prediction accuracy and sensitivity than RF, but also a little unstable than RF. In case 3, RF reported a higher overall accuracy, but sensitivity of RF is significantly lower than SVM due to imbalanced data set.

4.5 Potential patients identification

In order to detect the risk for developing schizophrenia, we first apply machine learning models to distinguish potential patients from healthy people. In this study, we randomly divided the control group into 3 sub-groups: G_1, G_2 and G_3. Based on these 3 sub-groups, we proposed a test protocol to identify potential patients who are currently in healthy control group, i.e.

1) Use G_1 and G_2 plus 113 patients to form a training data set, use G_3 as the test data set to train and test our model.
2) Use G_1 and G_3 plus 113 patients to form a training data set, use G_2 serve as the test set.
3) Use G_2 and G_3 plus 113 patients to form a training data set, use G_1 serve as the test data set.

Therefore, each person in control group was evaluated once, and we summarized the results in Table 4.5.1.

Table 4.5.1 indicates that, among the 51 healthy controls, 18 of them might be at risk for schizophrenia. PANSS scores and family history of schizophrenia were identified as very important variables in variable importance ranking analysis.

Secondly, we divide the patient group into 2 sub-groups: Schizophrenia versus Schizoaffective/Bipolar/MDD patients, and applied same test protocol to healthy control group to distinguish Schizophrenia from other mental disorders. Table 4.5.2 indicates that 11 people may potentially affected by mental disorders and 5 of them might be schizophrenic.

Thirdly, we further divide the patient group into 2 sub-groups based on their symptoms, and applied same test protocol. The prediction results are summarized in Table 4.5.3. Participants with ID number 11051 and 11503, who were identified as “other diseases” in earlier analysis were classified as Bipolar/MDD patient in this analysis as well.

Table 4.5.1, Prediction of potential patients using RF with all default parameters, case 1, patient=0, control=1

Patient ID	Prediction	Patient ID	Prediction	Patient ID	Prediction
3	potential patients	11053	potential patients	11020	control
6	potential patients	5	control	11022	control
7	potential patients	9	control	11023	control
19	potential patients	11	control	11024	control
11003	potential patients	11001	control	11025	control
11008	potential patients	11002	control	11026	control
11017	potential patients	11004	control	11027	control
11021	potential patients	11006	control	11029	control
11028	potential patients	11007	control	11030	control
11033	potential patients	11009	control	11032	control
11040	potential patients	11010	control	11034	control
11043	potential patients	11011	control	11035	control
11044	potential patients	11012	control	11036	control
11046	potential patients	11013	control	11037	control
11047	potential patients	11016	control	11042	control
11048	potential patients	11018	control	11045	control
11051	potential patients	11019	control	11050	control

Table 4.5.2, Prediction of potential patients using RF with all default parameters, case 2, Schizophrenia=0, other patients=1 (Schizoaffective, Bipolar, MDD)

Patient ID	Prediction	Patient ID	Prediction	Patient ID	Prediction
11003	others	11002	control	11025	control
11017	others	11004	control	11026	control
11043	others	11006	control	11027	control
11044	others	11007	control	11029	control
11051	others	11008	control	11030	control
11053	others	11009	control	11032	control
7	Scz	11010	control	11033	control
19	Scz	11011	control	11034	control
11021	Scz	11012	control	11035	control
11028	Scz	11013	control	11036	control
11046	Scz	11016	control	11037	control
3	control	11018	control	11040	control
5	control	11019	control	11042	control
6	control	11020	control	11045	control
9	control	11022	control	11047	control
11	control	11023	control	11048	control
11001	control	11024	control	11050	control

Table 4.5.3, Prediction of potential patients using RF with all default parameters, case 3, Schizophrenia & Schizoaffective=0, Bipolar & MDD=1

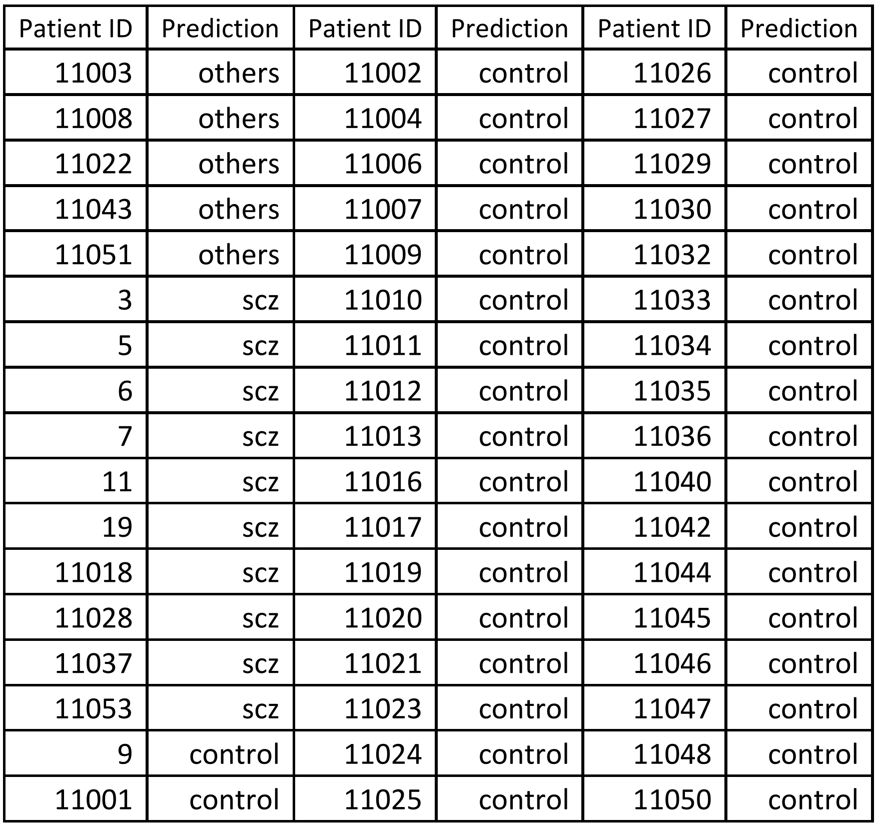
Patient ID	Prediction	Patient ID	Prediction	Patient ID	Prediction
11047	bip	11004	control	11025	control
11051	bip	11006	control	11026	control
11053	bip	11007	control	11027	control
3	scz	11008	control	11029	control
7	scz	11009	control	11030	control
19	scz	11010	control	11032	control
11021	scz	11011	control	11034	control
11028	scz	11012	control	11035	control
11033	scz	11013	control	11036	control
11046	scz	11016	control	11037	control
5	control	11017	control	11040	control
6	control	11018	control	11042	control
9	control	11019	control	11043	control
11	control	11020	control	11044	control
11001	control	11022	control	11045	control
11002	control	11023	control	11048	control
11003	control	11024	control	11050	control

In model validation part, RF is reporting a higher overall accuracy in distinguishing patients among healthy control (case 1), but in distinguishing schizophrenia from other mental disorders (case 2) and distinguishing schizophrenia/schizoaffective from bipolar/MDD analysis (case 3), the performance of RF is exceeded by SVM with class weight adjustment. Therefore, it is reasonable to assume that SVM with class weight adjustment may deliver a higher prediction confidence in such cases. Thus we re-ran the entire analysis using SVM with class weight adjustment and summarized the results in Table 4.5.4 to 4.5.6.

Table 4.5.4, Prediction of potential patients using SVM, class proportion as the class weights, case 1, patient=0, control=1

Patient ID	Prediction	Patient ID	Prediction	Patient ID	Prediction
3	potential patient	11001	control	11025	control
5	potential patient	11002	control	11026	control
6	potential patient	11004	control	11027	control
7	potential patient	11006	control	11029	control
11	potential patient	11007	control	11030	control
19	potential patient	11009	control	11032	control
11003	potential patient	11010	control	11033	control
11008	potential patient	11011	control	11035	control
11018	potential patient	11012	control	11036	control
11022	potential patient	11013	control	11037	control
11028	potential patient	11016	control	11040	control
11034	potential patient	11017	control	11042	control
11045	potential patient	11019	control	11043	control
11046	potential patient	11020	control	11044	control
11051	potential patient	11021	control	11047	control
11053	potential patient	11023	control	11048	control
9	control	11024	control	11050	control

Table 4.5.5, Prediction of potential patients using SVM, using class proportion as the class weights, case 2, Schizophrenia=0, other patients=1 (Schizoaffective, Bipolar, MDD)



Patient ID	Prediction	Patient ID	Prediction	Patient ID	Prediction
11003	others	11002	control	11026	control
11008	others	11004	control	11027	control
11022	others	11006	control	11029	control
11043	others	11007	control	11030	control
11051	others	11009	control	11032	control
3	scz	11010	control	11033	control
5	scz	11011	control	11034	control
6	scz	11012	control	11035	control
7	scz	11013	control	11036	control
11	scz	11016	control	11040	control
19	scz	11017	control	11042	control
11018	scz	11019	control	11044	control
11028	scz	11020	control	11045	control
11037	scz	11021	control	11046	control
11053	scz	11023	control	11047	control
9	control	11024	control	11048	control
11001	control	11025	control	11050	control

Table 4.5.6, Prediction of potential patients using SVM, using class proportion as the class weights, case 3, Schizophrenia & Schizoaffective=0, Bipolar & MDD=1

Patient ID	Prediction	Patient ID	Prediction	Patient ID	Prediction
11003	bip	11002	control	11026	control
11008	bip	11004	control	11027	control
11022	bip	11006	control	11029	control
11043	bip	11007	control	11030	control
11051	bip	11009	control	11032	control
3	scz	11010	control	11033	control
5	scz	11011	control	11034	control
6	scz	11012	control	11035	control
7	scz	11013	control	11036	control
11	scz	11016	control	11040	control
19	scz	11017	control	11042	control
11018	scz	11019	control	11044	control
11028	scz	11020	control	11045	control
11037	scz	11021	control	11046	control
11053	scz	11023	control	11047	control
9	control	11024	control	11048	control
11001	control	11025	control	11050	control

In Table 4.5.4, SVM have identified 16 participants as potential patients, compare to RF results in Table 4.5.1, participants with ID 3, 6, 7, 19, 11003, 11008, 11028, 11046, 11051, 11053 are classified as potential patients by both models. Table 4.5.5 indicates that 15 people may suffer from mental disorder and 10 of them are schizophrenia. Compare to RF, 4 more people are identified as patients, this could result from a higher sensitivity/PPV of SVM model, and a low NPV of RF indicates a higher probability of false "green light" among healthy control group. In case 3, sensitivity for RF is only 0.365, thus it is reasonable to expect more potential patients from SVM model. Participants with ID 3, 6, 7, 19, 11003, 11028, 11051 are classified as potential patients from all three cases and from both models, and participant with ID 11053 is classified as potential patient in all the cases by both models, but in case 2 and 3, this person is classified as schizophrenia by SVM model, and Bip/MDD by RF model.

Chapter 5 Results and Conclusion

From the out of sample prediction results we can see that the prediction performances are very promising for a computer based decision support system. The goal of this development is to provide an appropriate method to alleviate the pressure of limited medical resources, and to provide a preventive solution for patients with health concerns. By implementing this technology, professionals can get fast and stable results from this system and make judgments with more confidence and efficiency. Patients can get a self-diagnose by using this system and make suitable decision based on the prediction results.

In the study of GIB detection, it is impractical and economically unjustifiable to subject every patient with acute GIB to an urgent endoscopy, as only 20% of patients with acute GIB require urgent intervention. In our study, we have developed multiple machine learning predictive models, and successfully applied them to clinical data sets to predict clinical outcomes in a variety of conditions. But due to the inaccuracy of the records of our retrospective data set, and randomness of the decisions for Endoscopy, the prediction performance might be unstable for some data points. However, computer-based methods still prove to be a valuable decision support system, and as such, to optimize the cares of patients with acute gastrointestinal bleeding. Our models successfully achieved over 80-90% accuracies for Disposition and Resuscitation for the Blatchford data set, and the accuracies for Chu’s data set are over 81% and 91% for Endoscopy and Resuscitation respectively. Tuned threshold RF may provide a more flexible prediction and better performance in specified area. RF and SVM are designed for high-dimensional data with a large feature space (i.e. large number of predictor variables), and their numerical application algorithms are designed accordingly as well, so that they may reach a good computational efficiency.

In the study of Schizophrenia detection, Schizophrenia is characterized by a breakdown of thought processes and by a deficit of typical emotional responses. Many studies in modern neurology proven that this deficit is biologically measurable. Like other mental disorders, Schizophrenia is relatively inconspicuous and hard to notice. It may share common features with other mental diseases. Examining every potential patient using hospital standards is impractical and hard to implement. Therefore, developing computer based application to predict potential risks for such mental disorder can be highly efficient and economically practical. In our study, we developed multiple machine learning methods to uncover the possible relationship between patients' independent variable and response variable, to distinguish between Schizophrenia, MDD, Bipolar and Schizoaffective. Furthermore, by applying such models, we can identify potential patients among healthy control group and predict the risk of developing mental disorder, such as Schizophrenia, in the near future. In this study, we split the process into 3 major cases. In the first case, RF achieved a better overall accuracy, sensitivity and PPV, which are 84.5%, 89.1% and 75.5% respectively. In the second case, SVM had a better performance in all aspects. And in the third case, although RF achieved a better overall accuracy, this was at the cost of extremely low sensitivity due to a highly imbalanced data set, the performance of weight adjusted SVM is much more stable compared to RF in this case.

Chapter 6, Future Study

In our study, we have introduced several machine learning models and their variations to provide a statistical inference on clinical data sets. The primary goal is to build a suitable decision support system for both professionals and patients. To achieve this goal, we can split this big project into two parts. Firstly, we need to develop appropriate statistical methods to uncover the potential connections between predictors and response variables. Secondly, we need to apply numerical methods to implement these methods. Currently, we are still in the stage of developing appropriate methods that can deliver accurate predictions. Back testing results indicates that models introduced here have a promising future in computer-based online diagnostic systems. However, several issues still remain unanswered, e.g., the imbalanced response classes, sensitivity of Endoscopy in GIB study, inefficient optimal weight selection algorithm, ineffectual accuracy maximization algorithm.

Literature suggest that cost sensitive modeling, or proportional bootstrapping resampling technique may have a good impact in dealing with imbalanced data set. Applying approximation methods to rewrite the weight assigning algorithm into a quadratic programming problem and applying numerical solver, like Gradient Descent or Newton-like methods, may prove to be both efficient and accurate. In software development, the model performance might be affected by computer hardware and its operating systems. Appropriate programming skills are a necessity to find a fitting remedy for such problems. The computational efficiency of numerical methods is also being a concern.

References

[1] N. Ancona, G. Cicirelli, A. Branca, and A. Distante. Goal detection in football by using support vector machines for classification. In Proceedings of Int. Joint Conference on Neural Networks, vol.1, 611-616, 2001.

[2] M.R. Azimi-Sadjadi and S.A. Zekavat. Cloud classification using support vector machines. In Proceedings of IEEE Geoscience and Remote Sensing Symposium, vol. 2, 669-671, 2000.

[3] Y. Amit, D. Geman. Shape quantization and recognition with randomized trees. Neural Computation, vol. 9, 1545–1588, 1997.

[4] E. Bauer and R. Kohavi. An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants. Machine Learning, vol. 36, 105-139, 1999.

[5] E. Bauer, R. Kohavi. An empirical comparison of voting classification algorithms: bagging, boosting, and variants. Machine Learning, vol. 36, 105-139, 1999.

[6] L. Breiman. Random forests. Machine Learning, 45:5-32, 2001.

[7] L. Breiman. Bagging predictors. Machine Learning, 24:123-140, 1996.

[8] R.E. Banfield, L.O. Hall, K.W. Bowyer, D. Bhadoria, W.P.Kegelmeyer, and S. Eschrich. A Comparison of Ensemble Creation Techniques. Proc Fifth International Workshop Multiple Classifier Systems, 2004.

[9] S. Boyd , L. Vandenberghe. Convex optimization. Cambridge University Press, 2004.

[10] H. Byun and S.W. Lee. Applications of Support Vector Machines for Pattern Recognition: A Survey. Springer-Verlag Berlin Heidelberg, LNCS 2388, 213-236, 2002.

[11] S. Bernard, S.E. Adam, L. Heutte. Dynamic Random Forests. Pattern Recognition Letters 33, 1580-1586, 2012.

[12] S. Bernard, L. Heutte, S. Adam. On the selection of decision trees in random forests. International Joint Conference on Neural Network 302-307, 2009.

[13] S. Bernard, L. Heutte, S. Adam. Forest-RK: A new random forest induction method. Proceedings of the International Conference on Intelligent Computing (ICIC'08), vol. 5227 of Lecture Notes in Computer Science, Springer, 430-437, 2008.

[14] N. Bassiou, C. Kotropoulos, T. Kosmidis, and I. Pitas. Frontal face detection using support vector machines and back-propagation neural networks. In Proceedings of Int. Conference on Image Processing, 1026-1029, 2001.

[15] Y.C. Chen, H. Han, H. Kim and H. Ahn. Canonical Forest. Submitted, 2013.

[16] C. Choisy and A. Belaid. Handwriting recognition using local methods for normalization and global methods for recognition. In Proceedings of Sixth Int. Conference on Document Analysis and Recognition, 23-27, 2001.

[17] C. Cortes, V. Vapnik. Support Vector Networks. Machine Learning, 20, 1995.

[18] A. Chu, H. Ahn, B. Halwan, B. Kalmin, E. L. Artifon, A. Barkun, M. G. Lagoudakis, A. Kumar. A decision support system to facilitate management of patients with acute gastrointestinal bleeding. Artificial Intelligence in Medicine. vol. 42, 247-59, 2007.

[19] X. Dong and W. Zhaohui. Speaker recognition using continuous density support vector machines. Electronics Letters, vol. 37, 1099-1101, 2001.

[20] H. Druker, B. Shahrary, and D.C. Gibbon. Support vector machines: relevance feedback and information retrieval. Information Processing & Management, vol.38, Issue 3, 305-323, 2002.

[21] H. Drucker, D. Wu, and V. Vapnik. Support vector machines for spam categorization. IEEE Transactions on Neural Networks, vol. 10, 1048-1054 1999.

[22] T. G. Dietterich. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. Neural Computation, vol. 10, 1895-1924, 1997.

[23] X.Z. Fern and C.E. Brodley. Random Projection for High Dimensional Data Clustering: A Cluster Ensemble Approach. Proc. 20th International conference on Machine Learning, 2003.

[24] A. Fan and M. Palaniswami. Selecting bankruptcy predictors using a support vector machine approach. In Proceeding of IEEE-INNS-ENNS Int. joint Conference, vol. 6, 354-359, 2000.

[25] T. Fletcher. Support Vector Machines Explained. 2009.

[26] G. Guodong, S. Li, and C. Kapluk. Face recognition by support vector machines. In Proceedings of IEEE Int. Conference on Automatic Face and Gesture Recognition, 196-201, 2000.

[27] G. Guodong, S. Li, and C. Kapluk. Face recognition by support vector machines. In Proceedings of IEEE Int. Conference on Automatic Face and Gesture Recognition, 196-201, 2000.

[28] G. Guo, H.J. Zhang, and S.Z. Li. Distance-from-boundary as a metric for texture image retrieval. In Proceedings of IEEE Int. Conference on Acoustics, Speech, and Signal Processing, vol. 3, 1629-1632, 2001.

[29] S. Gutta, J.R.J. Huang, P. Jonathon, and H. Wechsler. Mixture of experts for classification of gender, ethnic origin, and pose of human. IEEE Trans. on Neural Networks, vol. 11, 948-960, 2000.

[30] F. van der Heijden, R.P.W. Duin, D. de Ridder, and D.M.J. Tax. Classification, Parameter Estimation and State Estimation. Wiley, 2004.

[31] B. Heisele, P. Ho, and T. Poggio. Face Recognition with support vector machines: global versus component-based approach. In Proceedings of Eighth IEEE Int. Conference on Computer Vision, vol. 2, 688-694, 2001.

[32] B. Heisele, P. Ho, and T. Poggio. Face Recognition with support vector machines: global versus component-based approach. In Proceedings of Eighth IEEE Int. Conference on Computer Vision, vol. 2, 688-694, 2001.

[33] T. K. Ho. Random decision forests. Proceedings of the Third International Conference, Document Analysis and Recognition, vol. 1, 278 – 282, 1995.

[34] K.I. Kim, J. Kim, and K Jung. Recognition of facial images using support vector machines. In Proceedings of 11th IEEE Workshop on Statistical Signal Processing, 468-471, 2001.

[35] S.J Kim, A. Magnani, S. Boyd. Optimal Kernel Selection in Kernel Fisher Discriminant Analysis. In Proceedings of the 23rd International Conference on Machine Learning, 2006.

[36] J. Lu, K.N. Plataniotis, and A.N. Ventesanopoulos. Face recognition using feature optimization and v-support vector machine. IEEE Neural Networks for Signal Processing XI, 373-382, 2001.

[37] Y. Li, S. Gong, and H. Liddell. Support vector regression and classification based multi-view face detection and recognition. In Proceedings of Face and Gesture Recognition, 300-305, 2000.

[38] Y. Li, S. Gong, J. Sherrah, and H. Liddell. Multi-view Face Detection Using Support Vector Machines and Eigen space Modeling. In Proceedings of Fourth Int. Conference on Knowledge-Based Intelligent Engineering Systems & Allied Technologies, 241-244, 2000.

[39] Y. Lee, Y. Lin, G. Wahba. Multicategory Support Vector Machines. Computing Science and Statistics 33, 2001.

[40] P. Leray, P. Gallinari. Feature selection with neural networks. Behaviormetrika (special issue on Analysis of Knowledge Representation in Neural Network Models), vol. 26, 145-166, 1999

[41] A. Liaw, M. Wiener. Classification and Regression by Random Forest. R News, vol. 2, 18-22, 2002.

[42] P. Mitra, C.A. Murthy, and S.K. Pal. Data condensation in large database by incremental learning with support vector machines. In Proceedings of 15th Int. Conference on Pattern Recognition, vol. 2, 708-711, 2000.

[43] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, A. Smola, and K. R. Muller. Constructing descriptive and discriminative non-linear features: Rayleigh coefficients in kernel feature spaces. IEEE Transactions on Pattern Analysis and Machine Intelligence, 25, 623-628. 2003.

[44] C. Nakajima, M. Pontil, and T. Poggio. People recognition and pose estimation in image sequences. In Proceedings of IEEE Int. Joint Conference on Neural Net-works, vol. 4, 189-194, 2000.

[45] J. Ng and S. Gong. Performing multi-view face detection and pose estimation using a composite support vector machine across the view sphere. In Proceedings of IEEE Int. Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems, 1999.

[46] J. Ng and S. Gong. Composite support vector machines for detection of faces across views and pose estimation. Image and Vision Computing, vol. 20, Issue 56, 359-368, 2002.

[47] C. Nakajima, M. Pontil, and T. Poggio. People recognition and pose estimation in image sequences. In Proceedings of IEEE Int. Joint Conference on Neural Networks, vol. 4, 189-194, 2000.

[48] C. Ongun, U. Halici, K. Leblebicioglu, V. Atalay, M. Beksac, and S. Beksac. Feature extraction and classification of blood cells for an

automated differential blood count system. In Proceedings of Int. Joint Conference on Neural Networks, 2461-2466, 2001.

[49] M. Pontil and A. Verri. Support vector machines for 3-D object recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence, 637-646, 1998.

[50] J. Platt, N. Christianini, and J. S. Taylor. Large margin DAGs for multiclass classification. Advances in Neural Information Processing Systems, 2000.

[51] M. Pontil and A. Verri. Support vector machines for 3-D object recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence, 637-646, 1998.

[52] D. Roobaert and M.M. Van Hulle. View-based 3D object recognition with support vector machines. In Proceedings of IX IEEE Workshop on Neural Networks for Signal Processing, 77-84, 1999.

[53] L. Ramirez, W. Pedrycz, and N. Pizzi. Severe storm cell classification using support vector machines and radial basis approaches. In Proceedings of Canadian Conference on Electrical and Computer Engineering, vol. 1, 87-91, 2001.

[54] J. J. Rodríguez, L. I. Kuncheva, C. J. Alonso. Rotation forest: A new classifier ensemble method. IEEE Trans. Pattern analysis and machine intelligence, vol. 28, 1619-1630, 2006.

[55] Z. Shi, C. Beard, K. Mitchell. Analytical Models for Understanding Misbehavior and MAC Friendliness in CSMA Networks. Performance Evaluation archive, vol. 66, 469-487, 2009.

[56] Z. Shi, C. Beard, K. Mitchell. Competition, Cooperation, and Optimization in Multi-Hop CSMA Networks with Correlated Traffic. International Journal of Next-Generation Computing. Vol. 3, 2012.

[57] Z. Shi, C. Beard, K. Mitchell. Misbehavior and MAC Friendliness in CSMA Networks. Wireless Communications and Networking Conference, 2007.

[58] Z. Shi, C. Beard, K. Mitchell. Analytical Models for Understanding Space, Backoff and Flow Correlation in CSMA Wireless Networks. WIRELESS NETWORKS, Springer, 2012.

[59] T. J. Terrillon, M.N. Shirazi, M. Sadek, H. Fukamachi, and S. Akamatsu. Invariant face detection with support vector machines. In Proceedings of 15th International Conference, vol. 4, 210-217, 2000.

[60] A. Tefas, C. Kotropoulos, and I. Pitas. Using support vector machines to enhance the performance of elastic graph matching for frontal face authentication. IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 23. No. 7, 735-746, 2001.

[61] L.N. Teow and K.F. Loe. Robust vision-based features and classification schemes for off-line handwritten digit recognition. Pattern Recognition, 2002.

[62] Q. Tian, P. Hong, and T.S. Huang. Update relevant image weights for content based image retrieval using support vector machines. In Proceedings of IEEE Int. Conference on Multimedia and Expo, vol.2, 1199-1202, 2000.

[63] F. Tay and L.J. Cao. Modified support vector machines in financial time series forecasting. Neurocomputing, 2001.

[64] L.N. Teow and K.F. Loe. Robust vision-based features and classification schemes for off-line handwritten digit recognition. Pattern Recognition, 2002.

[65] A. Tsymbal, M. Pechenizkiy, P. Cunningham. Dynamic Integration with Random Forests. In proceeding of: Machine Learning 17th ECML, 18-22, 2006.

[66] D.R. Wilson, T.R. Martinez. Improved heterogeneous distance functions. Journal of Artificial Intelligence Research, 6(1), 1-34, 1997.

[67] Y. Wang, C.S. Chua, and Y.K, Ho. Facial feature detection and face recognition from 2D and 3D images. Pattern Recognition Letters, 2002.

[68] V. Wan and W.M. Campbell. Support vector machines for speaker verification and identification. In Proceedings of IEEE Workshop on Neural Networks for Signal Processing X, vol. 2, 2000.

[69] W.F. Xie, D.J. Hou, and Q. Song. Bullet-hole image classification with support vector machines. In Proceedings of IEEE Signal Processing Workshop on Neural Networks for Signal Processing, vol.1, 318-327, 2000.

[70] M. H. Yang and B. Moghaddam. Gender classification using support vector machines. In Proceedings of IEEE Int. Conference on Image Processing, vol. 2, 471-474, 2000.

[71] Y. Yao, G. L. Marcialis, M. Pontil, P. Frasconi, and F. Roli. Combining flat and structured representations for fingerprint classification with recursive neural networks and support vector machines. Pattern Recognition, 1-10, 2002.

[72] B. Zhao, Y. Liu, and S.W. Xia. Support vector machines and its application in handwritten numerical recognition. In Proceedings of 15th Int. Conference on Pattern Recognition, vol. 2, 720-723, 2000.

[73] L. Zhang, F. Lin, and B. Zhang. Support vector machine learning for image retrieval. In Proceedings of Int. Conference on Image Processing, 721-724, 2001.

[74] Y. Zhang, R. Zhao, and Y. Leung. Image Classification by support vector machines. In Proceedings of Int. Conference on Intelligent Multimedia, Video and Speech Processing, 360-363, 2001.