# MODELLING EARLY DETECTION OF PROSTATE CANCER

by

Zhengjun Liu

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

in

Biostatistics

Department of Mathematical and Statistical Sciences
University of Alberta

# Abstract

Prostate cancer is one of the most common cancers among men in the world (excluding non-melanoma skin cancers). According to the statistics from the Canadian Cancer Society [1], it is the third leading cause of death from cancer in men in Canada. In general, prostate cancer is treatable with 5-year survival rate of 99% for early-stage. However, if the cancer has spread to nearby organs, the 5-year survival rate drops to 28%. Unfortunately, 92% of patients are diagnosed at an advanced stage.

Early detection is an ongoing challenge for prostate cancer treatment. Existing statistical models (e.g. principal component analysis) are more likely to inform us the statistical relationship between each metabolite. However, they have a poor performance of predicting the early prostate cancer.

In order to improve predictions, in this thesis, we developed models using metabolite profiles to identify patients who are likely to suffer prostate cancer. Several predictive methods such as Support Vector Machine (SVM), K-Nearest neighbor (KNN), Random Forest, LASSO, and PLS-DA were used.

# Preface

The project completed in this thesis was developed as part of the MITACS Accelerate program led by Dr. Michael Li in the Department of Mathematical and Statistical Sciences at the University of Alberta, and the partner company, Metabolomics Technologies Inc. (MTI), Canada.

The metabolite data were provided by MTI. The information were collected based on serum samples of prostate cancer group and healthy control group.

I was responsible for the model formulation and analysis, parameter estimation, model validation and interpretation of results.

## Acknowledgements

I would like to express my great appreciation to my supervisor, Dr. Michael Li, for the support of my MSc. study and offering creative advice as well as helpful suggestions on my thesis.

I would also like to thank the members of the thesis committee for the time and help in the completion of this thesis.

Many thanks to Dr. Lu Deng, Senior Scientist at MTI, the partner company of MITAC program, for introducing me to the topic of metabolic studies that lead to the research project for this thesis. I would also like to thank MTI for providing data for the thesis.

I would like to offer my thanks to the Department of Mathematical and Statistical Sciences and the MITACS Accelerate program for providing funding for my thesis research.

Finally, my sincere thanks goes to my family and friends for their continuous support and encouragements through the process of my graduate studies and research.

# Table of Contents

**1 Introduction** **1**

1.1 Background information . . . . . 1

1.2 Introduction of metabolomic research . . . . . 2

1.3 Existing models for prostate cancer prediction . . . . . 3

1.4 Summary of research methodology . . . . . 4

**2 Formulation of predictive model for prostate cancer** **6**

2.1 Penalized Logistic Regression Method . . . . . 6

2.2 Support Vector Machine Method . . . . . 9

2.3 Random Forest Method . . . . . 11

2.4 K-nearest neighbors Method . . . . . 13

2.5 Neural Network Method . . . . . 14

2.6 Partial Least Squares Method . . . . . 15

**3 Model Fitting** **17**

3.1 Software . . . . . 17

3.2 Data . . . . . 17

3.3 Model Selection and Evaluation . . . . . 18

3.4 Modelling with All Detectable Metabolites . . . . . 19
3.5 Modelling with Selected Metabolites . . . . . 28
3.6 Discussion . . . . . 39

**4 Conclusion** **40**

**Bibliography** **42**

**Appendices** **46**

# List of Tables

3.1 Logistic Regression - LASSO with top 5 features . . . . . 19

3.2 AUC of ROC curve for models with all features . . . . . 26

3.3 AUC of ROC curve for models with selected features . . . . . 38

# List of Figures


3.1 ROC curve - LASSO model (Train) . . . . . 20
3.2 ROC curve - LASSO model (Test) . . . . . 20
3.3 ROC curve - SVM Linear model (Train) . . . . . 21
3.4 ROC curve - SVM Linear model (Test) . . . . . 22
3.5 ROC curve - SVM Nonlinear model (Train) . . . . . 22
3.6 ROC curve - SVM Nonlinear model (Test) . . . . . 23
3.7 ROC curve - Random Forest model (Train) . . . . . 24
3.8 ROC curve - Random Forest model (Test) . . . . . 24
3.9 ROC curve - Neural Network model (Train) . . . . . 25
3.10 ROC curve - Neural Network model (Test) . . . . . 25
3.11 ROC curve - PLS-DA model (Train) . . . . . 27
3.12 ROC curve - PLS-DA model (Test) . . . . . 27
3.13 Loading weight of comp.1 and comp.2 . . . . . 28
3.14 ROC curve - LASSO model with selected features (Train) . . . 29
3.15 ROC curve - LASSO model with selected features (Test) . . . 30
3.16 ROC curve - SVM Linear model with selected features (Train) 31
3.17 ROC curve - SVM Linear model with selected features (Test) . 31
3.18 ROC curve - SVM Nonlinear model with selected features (Train) 32

3.19 ROC curve - SVM Nonlinear model with selected features (Test) 32
3.20 ROC curve - Random Forest model with selected features (Train) 33
3.21 ROC curve - Random Forest model with selected features (Test) 34
3.22 ROC curve - PLS-DA model with selected features (Train) . . 35
3.23 ROC curve - PLS-DA model with selected features (Test) . . . 35
3.24 Loading weight of comp.1 and comp.2 . . . . . 36
3.25 ROC curve - Neural Network model with selected features (Train) 37
3.26 ROC curve - Neural Network model with selected features (Test) 37

# Chapter 1

# Introduction

## 1.1 Background information

Prostate cancer is one of the most common cancers among men in the world (excluding non-melanoma skin cancers). According to 2017 cancer statistics from the Canadian Cancer Society [1], in Canada, prostate cancer is the third leading cause of death from cancer in men. Prostate cancer is treatable if a patient is diagnosed at a early stage. It indicates that early detection is important for increasing survival rate. In Alberta, Alberta Health Services (AHS) recommends that men ages 50 to 74 have a prostate specific antigen (PSA) test and a digital rectal exam (DRE) annually. Doctors suggest that men with PSA value of 18% or less should get a prostate biopsy. However, the low accuracy of PSA predictions results in unnecessary biopsies and the side effects such as bleeding, infection, pain, etc. Based on the statistical result [2], the sensitivity of the combination of PSA and DRE is 38% which indicates that only 38% of patients are correctly identified as having prostate cancer.

In order to address these issues, we would like to construct models to improve the predictive accuracy.

The prostate cancer treatments include active surveillance, surgery, radiation therapy etc. The risk of prostate cancer could be determined by Gleason score which is in the range of 2 to 10. The American Society of Clinical Oncology (ASCO) guidelines suggest active surveillance for most patients with low risk and senior population [10]. A low score (usually lower than 6) means that the tumour is less likely to spread. On the contrary, a high score (larger than 7) means that the patient need advanced treatment immediately. The Gleason score is calculated by the pattern of cells under a microscope. After receiving samples from a prostate biopsy, the total score is produced by the combination of dominant cell pattern and non-dominant cell pattern.

## 1.2 Introduction of metabolomic research

Metabolomics is a new field of disease diagnostics. There are over 6,500 metabolites in human body. The sample of metabolites can be thought of as a metabolic "fingerprint" representative of an individual's current state of health [11].

Recent research shows that metabolism plays an important role in prostate cancer detection. It is because metabolic syndrome includes several cited risk factors. For instance, high concentrations of inflammation-related biomarkers will enhance tumour growth and some prostatic fluids (e.g. spermine, citrate) are not commonly existed in prostate urine samples of healthy group. Current

findings indicates that prostate cancer cells will lose the capacity to accumulate zinc. This situation can be captured by metabolites profiling. It illustrates the patients at high risk of prostate cancer could be distinguished from healthy group by analyzing the unique metabolism.

## 1.3 Existing models for prostate cancer prediction

In order to figure out the important features of prostate cancer, one of the common methods is principal component analysis (PCA) [9].

The basic idea behind principal component analysis is to convert observations of dependent variables into sets of linear combinations.

Suppose that we have $n$ observations with a set of $p$ features, $X_1, X_2, ..., X_p$. Before starting analysis, data visualization is a useful step. For 2 dimensional data, we could plot scatterplot to determine if there are statistical relationships between features. However, when $p$ is large, it is impossible to obtain the relationships among the features. Therefore, an alternative method would be finding a low dimensional space that includes information as much as possible. PCA provides a useful tool to analyze high dimensional data.

The first principal component of $p$ features, $X_1, X_2, ..., X_p$, is the normalized linear combination,

$$Z_1 = \alpha_{11}X_1 + \alpha_{21}X_2 + ... + \alpha_{p1}X_p, \text{ subject to } \sum_{j=1}^{p} \alpha_{j1}^2 = 1.$$

It represents the largest variance of the exploratory data set. The coefficient

elements $\alpha_{11}, \alpha_{21}, ..., \alpha_{p1}$ are the weights of the first principal component.

The approach used to compute first principal component is the following optimization problem:

$$\max_{\alpha_{11},\alpha_{21},...,\alpha_{p1}} \{\frac{1}{n}\sum_{i=1}^{n}\{\alpha_{11}x_{i1} + \alpha_{21}x_{i2} + ... + \alpha_{p1}x_{ip}\}^2\}, \text{ subject to } \sum_{j=1}^{p}\alpha_{j1}^2 = 1.$$

Once the first principal component of the features determined, we can generate the second principal component based on the following definition. The second principal component is the linear combination of $X_1, X_2, ..., X_p$ that has maximum variance among all linear combinations that are not correlated with the first principal component.

## 1.4 Summary of research methodology

**Data.** In this project, we had succeeded in collecting 188 NMR-detectable metabolites along with 2 clinical features: age and smoking history. Samples obtained from the APCaRI group biobank were broken down by 300 cancer serums and 300 normal serums (a total of 600). The analytical platform was Biocrates p180 kit (quantifies 188 metabolites) on a LC-MS/MS platform. There were two types of MS workflows. The first one used directed flow injection (DI) followed by MS analysis in order to identify the key metabolites. The second one used high performance liquid chromatography (LC) separation followed by MS analysis to help addressing some of the challenges associated with analyzing complex samples like serums samples.

**Statistical modelling.** Models for statistical analysis of data included logistic regression with LASSO, Support Vector Machine (linear and nonlinear), Random Forest, K-Nearest neigbhor (KNN), Neural Network, and PLS-DA. Metabolites profiles (including 188 qualified metabolites) from normal group and cancer group were analyzed to identify a classifier to predict the likelihood of a individual at risk of prostate cancer. For data pre-processing, we scaled data using log transformation and replaced the missing value of metabolites profiles with 0.5* LOD (limit of detection). Also we removed the metabolite that had more than 30% missing values.

**Model selection.** For model selection, we introduced the AUC (Area Under The Curve) - ROC (Receiver Operating Characteristics) curve which is one of the most common evaluation metrics for testing the classification's performance. The better model is with higher AUC value.

# Chapter 2

# Formulation of predictive model for prostate cancer

In this chapter, we discuss several predictive models for early detection of prostate cancer.

## 2.1 Penalized Logistic Regression Method

Logistic regression is a predictive analysis method that is used to model the probability of an event existing such as healthy/sick. It is one of generalized linear models (GLM) where the outcome $Y$ given $X$ is a Bernoulli variable [9]. The distribution can be represented as:

$$p(y|x) = \sigma(\langle \mathbf{w}, \mathbf{x} \rangle)^y (1 - \sigma(\langle \mathbf{w}, \mathbf{x} \rangle))^{1-y},$$

where $\sigma$ is a sigmoid function given by

$$\sigma(t) = \frac{1}{1 + e^{-t}}$$

and $\mathbf{w} = [w_1, w_2, \ldots, w_d]$ is a set of unknown coefficients that will be learned from data. Note that if $y = 1$, then $p(y = 1|x) = \sigma(\langle \mathbf{w}, \mathbf{x} \rangle)$.

In this thesis, we aimed to predict the probability that an event occurs (class =1). Thus, given this probability, we can indicate $p(y = 0|x) = 1 - \sigma(\langle \mathbf{w}, \mathbf{x} \rangle)$.

Maximum likelihood estimation (MLE) is a good method to estimate unknown coefficients in GLM. We assume that dataset $D = \{x_i, y_i\}$, $i$ from 1 to n is an i.i.d sample from probability distribution $p(x, y) = p(y|x) * p(x)$. Then the negative log-likelihood function could be written as $-ll(\mathbf{w}) = \sum_{i=1}^{n} -ll_i(\mathbf{w})$, where

$$\begin{aligned} ll_i(\mathbf{w}) &= \log\ p(y_i|\mathbf{x}) \\ &= y_i \log \sigma(\langle \mathbf{w}, x_i \rangle) + (1 - y_i) \log(1 - \sigma(\langle \mathbf{w}, x_i \rangle)). \end{aligned}$$

Next step is to take derivative of the negative log-likelihood function using the chain rule. Let $\theta_i = \langle \mathbf{w}, x_i \rangle$, the first part can be written

$$\begin{aligned} \frac{\partial y_i \log \sigma(\langle \mathbf{w}, x_i \rangle)}{\partial w_j} &= y_i \frac{\partial \log \sigma(\theta_i)}{\partial \theta_i} \frac{\partial \theta_i}{\partial w_j} \\ &= y_i (1 - \sigma(\theta_i)) x_{ij}. \end{aligned}$$

Similarly, we take the derivative of second component

$$\frac{\partial(1-y_i)\log(1-\sigma(\langle \mathbf{w}, x_i\rangle))}{\partial w_j} = y_i\sigma(\theta_i)(1-\sigma(\theta_i))x_{ij}.$$

Finally, the gradient of negative log-likelihood per sample is

$$-\frac{\partial ll_i(\mathbf{w})}{\partial w_j} = (\sigma(\langle \mathbf{w}, x_i\rangle) - y_i)x_{ij}.$$

We can estimate the unknown coefficients $\mathbf{w}$ by solving the gradient of negative log-likelihood function.

In order to improve the performance of regression model, we want to shrink the coefficients of the less contributive variables toward zero. There are several methods that are commonly used to remove the unimportant features. In this thesis, we focused on Least Absolute Shrinkage and Selection Operator (LASSO) method [9].

The goal of LASSO is to minimize

$$\min\left(\text{ log loss function } + \lambda\sum_{j=1}^{d}|w_j|.\right),$$

where

$$\text{log loss function} = -\frac{1}{n}\sum_{i=1}^{n}[y_i\log p(y_i) + (1-y_i)\log(1-p(y_i))].$$

Unlike other shrinkage methods, LASSO has no closed form. In terms of solving the function above, we used R package "glmnt".

## 2.2 Support Vector Machine Method

Support vector machine (SVM) is a machine learning tool for learning predictions in high dimensional space. The key idea of support vector machine (SVM) is to classify a test observation depending on which side of a hyperplane it lies [9].

First, we discuss about the SVM model with linear boundary. Let $S = (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$ be a training dataset, where $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i = \pm 1$. $y_i = 1$ labels as "normal group"; $y_i = -1$ represents "high risk group". Each $y_i$ can be written as:

$$y_i = \text{sign}(\langle \mathbf{w}, \mathbf{x}_i \rangle + b),$$

where $\mathbf{w} = [w_1, w_2, \ldots, w_d]$ is the vector of parameters.

Next, we want to determine the maximum margin that is used to find the boundary. Define the distance between $\mathbf{x}$ and the hyperplane by $|\langle \mathbf{w}, \mathbf{x}_i \rangle + b|$, where $\|\mathbf{w}\| = \sqrt{\sum_{j=1}^{d} \mathbf{w}_j^2} = 1$. Therefore, the closest observation in training set that used to separate hyperplane is $\min |\langle \mathbf{w}, \mathbf{x}_i \rangle + b|$ for each $i \in m$. Let $M = \min_{\forall i \in m} |\langle \mathbf{w}, \mathbf{x}_i \rangle + b|$ . Our objective is to find a solution for the following optimization problem,

$$\begin{aligned} &\max_{(\mathbf{w},b):\|\mathbf{w}\|=1} \quad M \\ &\text{subject to: } y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) > 0 \text{ for each } i. \end{aligned} \tag{2.1}$$

Another equivalent formula for Hard-SVM is :

$$\begin{aligned} (\mathbf{w}_0, b_0) &= \min_{(\mathbf{w},b)} \ \|\mathbf{w}\|^2 \\ &\text{subject to: } y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) > 1 \text{ for each } i, \end{aligned} \tag{2.2}$$

where $M = 1/\|\mathbf{w}\|$. Then the solution of (2.2) is given by $\hat{\mathbf{w}} = \frac{\mathbf{w}_0}{\|\mathbf{w}_0\|}, \hat{b} = \frac{b_0}{\|\mathbf{w}_0\|}$.

Then we estimate $(\hat{\mathbf{w}}, \hat{b})$ by using Lagrange function as follows,

$$L_P = \|\mathbf{w}\|^2 - \sum_{i=1}^{m} \alpha_i y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b - 1), \text{with Lagrange multiplier: } \alpha_i > 0.$$

The gradients are computed by [12]:

$$\begin{cases} \bigtriangledown_{\mathbf{w}} L_P = \mathbf{w} - \sum_{i=1}^{m} \alpha_i y_i x_i, \\ \dfrac{\partial L_P}{\partial b} = \sum_{i=1}^{m} \alpha_i y_i. \end{cases}$$

Set the respective derivative to zero:

$$\begin{cases} \bigtriangledown_{\mathbf{w}} L_P = 0 \Longrightarrow \mathbf{w} = \sum_{i=1}^{m} \alpha_i y_i x_i, \\ \dfrac{\partial L_P}{\partial b} = 0 \Longrightarrow \sum_{i=1}^{m} \alpha_i y_i = 0. \end{cases}$$

We obtain that

$$\hat{w} = \sum_{i=1}^{m} \hat{\alpha}_i y_i x_i,$$

with non-zero coefficient $\hat{\alpha}_i$ for those observations that are selected as margin points.

Non-linear SVM is another type of SVM method in which the boundary

conditions are nonlinear. We introduce the kernel method that is used as a bridge from linearity to non-linearity [9]. There are many nonlinear kernel functions (i.e. k-degree polynomial, radial basis kernel):

- k-degree polynomial kernel function: $K(\mathbf{x}, \mathbf{x}^{'}) = (1 + \langle \mathbf{x}, \mathbf{x}^{'} \rangle)^k$.
- Radial basis kernel function is

$$K(\mathbf{x}, \mathbf{x}^{'}) = \exp\left(-\frac{\|(\mathbf{x} - \mathbf{x}^{'}\|^2}{2\sigma}\right) \text{ with given } \sigma > 0.$$

Then, the function of hyperplane can be written as

$$f(x) = \sum_{i=1}^{m} \alpha_i y_i K(x, x_i) + b_0,$$

where $K$ is a kernel function.

To figure out the estimators of the above function, we can use R packages "svmpath" and "e1071".

## 2.3 Random Forest Method

Random forest is an ensemble learning method for classification that operates by generating a multitude of decision trees [9]. Before discussing about Random Forest method, we would get started with the introduction of decision tree. To generate a decision tree, each internal node is labeled with an input feature. We apply GINI index as our cost function which can be written as,

$$\text{GINI Index} = \sum_{k \neq k'} \hat{p}_{nk} \hat{p}_{nk'} = \sum_{k=1}^{K} \hat{p}_{nk}(1 - \hat{p}_{nk'}),$$

where $\hat{p}_{nk} = \frac{1}{N_n} \sum_{x_i \in R_n} I(y_i = K)$ represents the proportion of class $k$ in node $n$ with $N_n$ observations in the region $R_n$. For the binary classification (i.e. Yes/No), if $p$ represents the probability in positive event, then the GINI index can be re-written as $2p(1-p)$. The feature for each node can be selected by minimum value of GINI Index.

However, one of the biggest disadvantages for decision tree is overfitting. In other words, the decision tree is excessively dependent on features of the training dataset with the result that it has a poor performance for unseen instances. In order to reduce overfitting of decision tree, we introduce the random forest method. The algorithm of random forest includes the following steps:

**Step 1.** randomly select samples from original dataset to obtain a bootstrapped dataset that is the same size as the original one.

**Step 2.** create a decision tree using the bootstrapped dataset but use a random subset of features at each step.

**Step 3.** go back to step 1 and repeat the selection process.

This bootstrapping procedure results in a better model performance as it can decrease the variance of the model without increasing the bias. As we mentioned before, the predictions of a single decision tree are highly related to the noise in training dataset. In that case, if a model is trained on a single training set, it would give strongly correlated trees. To address the issue, bootstrap sampling is an option of decreasing the chance of correlating the trees by providing them different training sets. The final predictions can be made

by taking the majority vote.

## 2.4 K-nearest neighbors Method

K-nearest neighbors method (KNN) is a non-parametric method that is commonly used for classification and regression. The input consists of the k closest training observations in the feature space. Unlike previous approaches, KNN approach does not require any model to fit it. It means the predictions are based on "memory". The main idea behind is that a classifier can be identified by $K$ points which are closet to the observation $x_0$ [9]. The conditional probability of $Y$ belongs to class $j$ is

$$Pr(Y = j|X = x_0) = \frac{1}{K} \sum_{i \in N_0} I(y_i = j),$$

where $N_0$ is a set of the $K$ points. Then KNN classifies the observation $x_0$ to the class with the largest probability. The major advantage of KNN is simplicity, and KNN has good performance in a large number of classification task. To discuss Bayes error rate, we need to introduce some basic background of Bayes rule.

Let $p(c_i)$ denote a prior distribution of class $i, 1 \leq i \leq N$. $p(x|c_i)$ denotes the condition probability density of $x$ given that it belongs to class $i$. The posteriori probability $p(c_i|x)$ is given by Bayes rule,

$$p(c_i|x) = \frac{p(x|c_i)p(c_i)}{p(x)},$$

where $p(x) = \sum_{i=1}^{N} p(x|c_i)p(c_i)$. Then, the Bayes classifier is the classification

that assigns the observation $x$ to the class with the highest posteriori. The Bayes error can be expressed as,

$$Bayes\ Error = 1 - \sum_{i=1}^{N} \int_{C_i} p(x|c_i)p(c_i)dx = 1 - p_{c_i*}(x),$$

where $C_i$ is the region of class $i$ with highest posteriori.

A famous result of Cover and Halt (1967) shows that asymptotically the error rate of 1-nearest neighbor classifier is never more than twice the Bayes error rate. This result provides a brief idea that the Bayes error rate can be interpreted as the lowest possible prediction error that can be achieved.

## 2.5 Neural Network Method

Neural networks are a set of algorithms, that are computing systems inspired by biological neural networks in human brain. Such systems learn to perform tasks by interpreting data through a kind of artificial intelligence machine to label or cluster raw dataset. For neural network method, the goal is to learn a function of inputs to generate a prediction of the target function $h(x)$ [9].

First, we discuss a simple example by considering one input observation with one output with 2-dimensional hidden layer. The hidden layer is used to map from the input $x$ to a new representation. In general, we apply a sigmoid transfer $\sigma$ to get the hidden layer. Then, the new representations are

$$\mathbf{h} = [h_1, h_2] \quad \text{with} \quad h_1 = \sigma(xw_1^{(1)}) \quad \text{and} \quad h_2 = \sigma(xw_2^{(1)})$$

Next, we assume $h$ to be new input and learn a GLM on the late layer so

that the output $y$ equals $\mathbf{h}\mathbf{w}^{(2)}$ with the learned weights $\mathbf{w}^{(2)}$.

Now, we consider more general case with $d$ inputs, $m$ outputs and $k_1$-dimensional hidden layers. Following the prior definition, the general format of new representations is

$$\mathbf{h} = \sigma(\mathbf{w}^{(1)}\mathbf{x}) = \left[\sigma(\mathbf{x}\mathbf{w}_1^{(1)}), \sigma(\mathbf{x}\mathbf{w}_2^{(1)}), \ldots, \sigma(\mathbf{x}\mathbf{w}_{k_1}^{(1)})\right]',$$

where $\sigma$ is sigmoid function applied to each input. Therefore, the outputs are $\sigma(\sigma(\mathbf{x}\mathbf{w}^{(1)})\mathbf{w}^{(2)})$. In addition, we could apply this formula to the situation of any number of hidden layers. In that case, the outputs with H-1 hidden layers of $k_1, \ldots, k_{H-1}$ dimensions can be written as,

$$\sigma(\sigma(\ldots\sigma(\mathbf{x}\mathbf{w}^{(1)})\mathbf{w}^{(2)})\ldots)\mathbf{w}^{(H)}),$$

where $\mathbf{w}^{(1)} \in \mathbb{R}^{d*k_{H-1}}, \ldots, \mathbf{w}^{(H)} \in \mathbb{R}^{k_1*m}$. We solved the above coefficients by running R package.

## 2.6 Partial Least Squares Method

Partial Least Squares (PLS) is a statistical approach used to predict variables. Instead of finding the hyperplanes of maximum variance between outcomes and explanatory variables, it provides a linear combination model by projecting the response and the observable features to a new space. Unlike PCA, PLS is a supervised learning method with principal components of both exploratory data $X$ and the response $Y$ [9].

To generate principle direction of PLS, usually we need to standardize the

$p$ predictors before using PLS, the first PLS component $Z_1$ can be written as

$$Z_1 = \sum_{j=1}^{p} \varphi_{j1} X_j,$$

where $\varphi_{j1}$ is the coefficient from the linear regression of the response Y on the $X_j$. The highest weight on the variables means the strong related to the response.

Then, we create the second PLS direction using the residuals of the regression. These residuals can be interpreted as the remainder of the information that can not be explained by the first PLS direction. The algorithm can be repeated by $M$ times to identify PLS components $Z_1, \ldots, Z_M$.

The partial least square discriminant analysis (PLS-DA) is often used when the response is the categorical variable.

# Chapter 3

# Model Fitting

## 3.1 Software

All codes in this thesis were written in R version 3.4.3. For SVM, we used the R library: e1071(version 1.6). For PLS-DA, KNN, Random Forest and neural network methods, we used R library: caret (version 5.15). R library: glmnet(version 1.8) was applied to LASSO approach.

## 3.2 Data

We had succeeded in collecting 188 NMR-detectable urinary metabolites of 600 individuals along with 2 clinical features: age, smoking history. Samples obtaining from the APCaRI group biobank were broken down by 300 cancer serums and 300 normal serums (total of 600).

Data pre-processing was performed using code written in R version 3.4.3. Metabolites that were not detected in 30% of the samples were removed from

the initial list of 188 metabolites. If a metabolite concentration was lower than LOD (limit of detection) , it was replaced with half of the value of LOD. We scaled metabolites profile (except clinic features) and separated the dataset into training set and testing set by matching age and cancer event rate. 70% of the source data was used for training the model, and 30% of the source data, for testing the model. The average age of total observations and prostate cancer group was 63.5 and 65.5 respectively.

## 3.3 Model Selection and Evaluation

For model selection and evaluation, there were several methods such as confusion matrix, Gini coefficients etc. In this thesis, we were interested in predicting the likelihood of a patient at risk of prostate cancer. So, our objective was a classification problem. One of the most common evaluation metrics for testing the classification's performance is AUC (Area Under The Curve) - ROC (Receiver Operating Characteristics).

ROC is a curve of true positive rate (TPR) and false positive rate (FPR) by different thresholds. In our prostate cancer detection example, TPR measures the percentage of accurate predictions who are having prostate cancer in actual. FPR is the percentage of correctly classify an individual as disease - free. AUC represents the area under the curve. The range of AUC values is from 0 to 1. A prefect model has $AUC = 1$ and a random classifier has $AUC = 0.5$. Normally, a model will score somewhere in between 0.5 and 1. We selected our model with higher AUC value in both training dataset and testing dataset.

## 3.4 Modelling with All Detectable Metabolites

In this section, we constructed predictive models using all NMR-detectable metabolites. Then we selected the best performance model by maximum the AUC value of the ROC curve.

For logistic regression with LASSO penalty method, we split the data set into 5 folds. For each fold, the remaining data were used as training data. The final model was built based on the robust features. The top 5 metabolites was represented in Table 3.1. It included the metabolites such as lysoPC.a, Dopamine, PC.aa.C30.2, PC.aa.C42.1 and lysoPC.a.C18.2. Figure 3.1 showed the ROC curve of the model. It had an AUC value of 0.69 with a specificity of 55% and sensitivity of 83.4%. However, the AUC value of test dataset was down to 0.596.

| Metabolite | Coefficient |
|---|---|
| lysoPC.a | 0.287 |
| Dopamine | 0.182 |
| PC.aa.C42.1 | 0.098 |
| PC.aa.C30.2 | -0.083 |
| lysoPC.a.C18.2 | 0.015 |

**Table 3.1:** Logistic Regression - LASSO with top 5 features

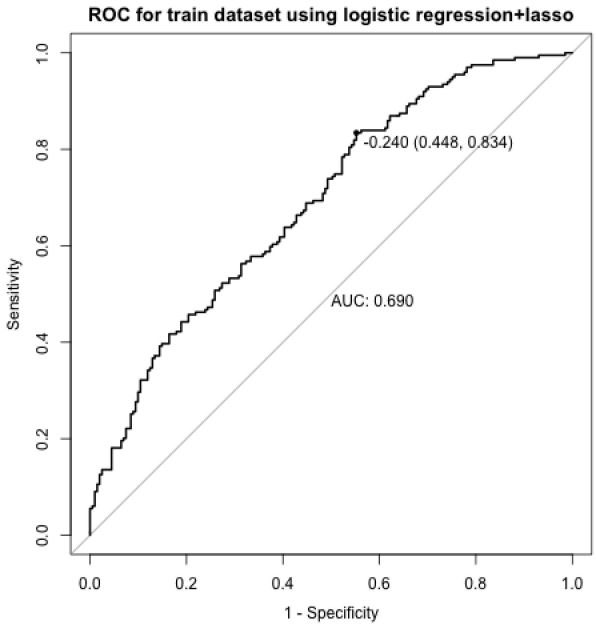


**Figure 3.1:** ROC curve - LASSO model (Train)

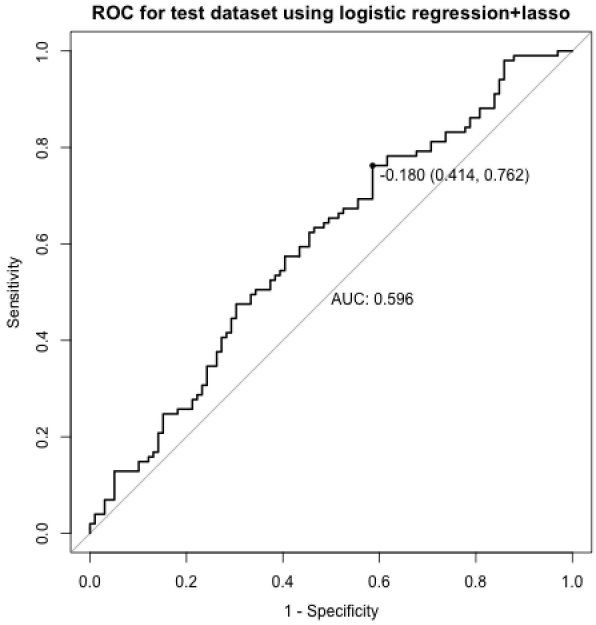


**Figure 3.2:** ROC curve - LASSO model (Test)

For SVM models, we observed that SVM with nonlinear boundary had a better performance than linear one. In terms of SVM - Nonlinear model, the polynomial kernel function with gramma value of 0.1 and degree of 2 had best performance according to 5-folder cross validation. The final model of SVM - Nonlinear had AUC value of 0.858 in the train set, however the model was a bit overfitting since the AUC value of test dataset was dropped to 0.574.

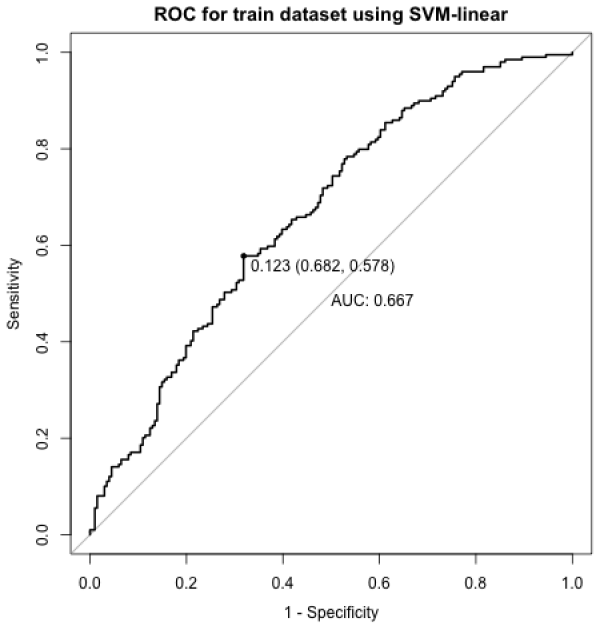


**Figure 3.3:** ROC curve - SVM Linear model (Train)

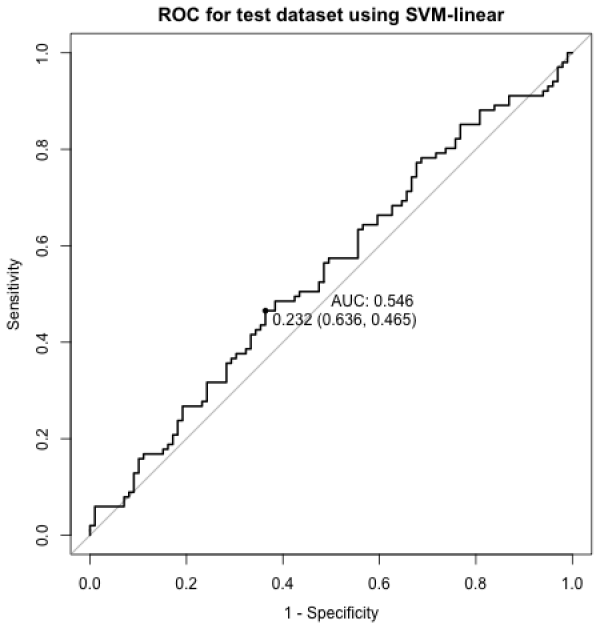


**Figure 3.4:** ROC curve - SVM Linear model (Test)

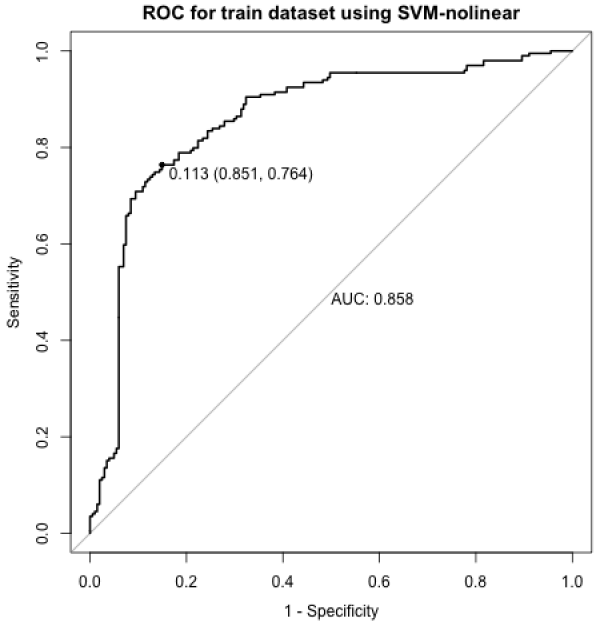


**Figure 3.5:** ROC curve - SVM Nonlinear model (Train)

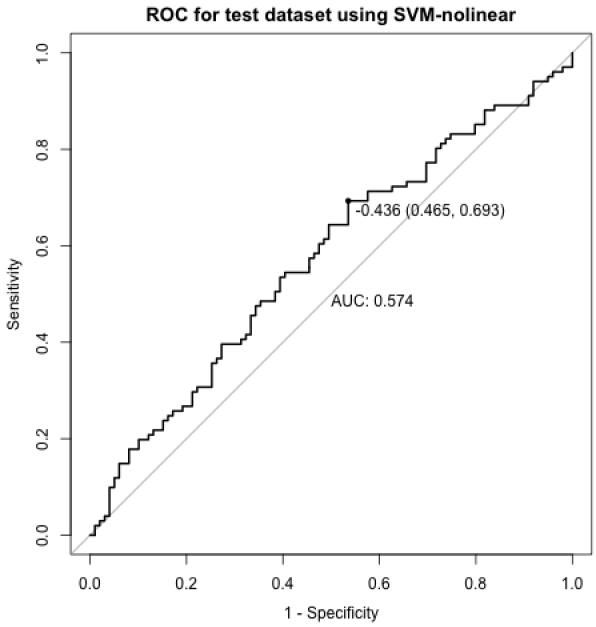


**Figure 3.6:** ROC curve - SVM Nonlinear model (Test)

From the ROC curves, random forest and neural network models were also overfitting in the training dataset. For random forest, the value of optimal number of selected fields was 13. Figure 3.7 showed that the AUC value of train set was 0.893. However, the AUC value of test dataset was 0.47 for random forest method.

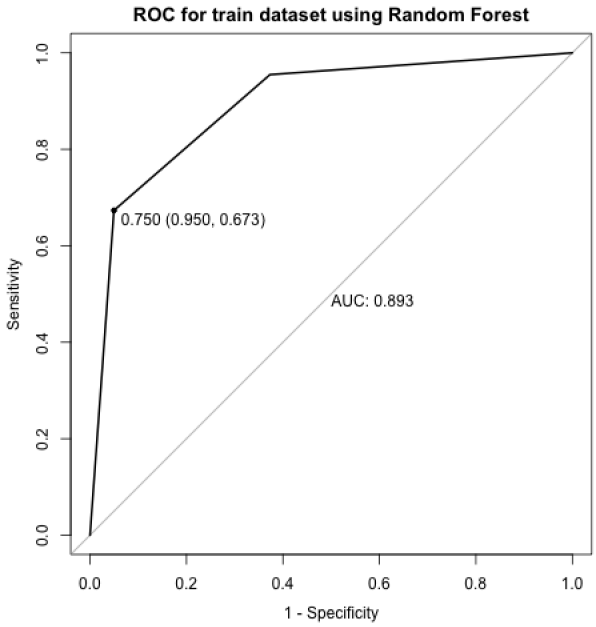


**Figure 3.7:** ROC curve - Random Forest model (Train)

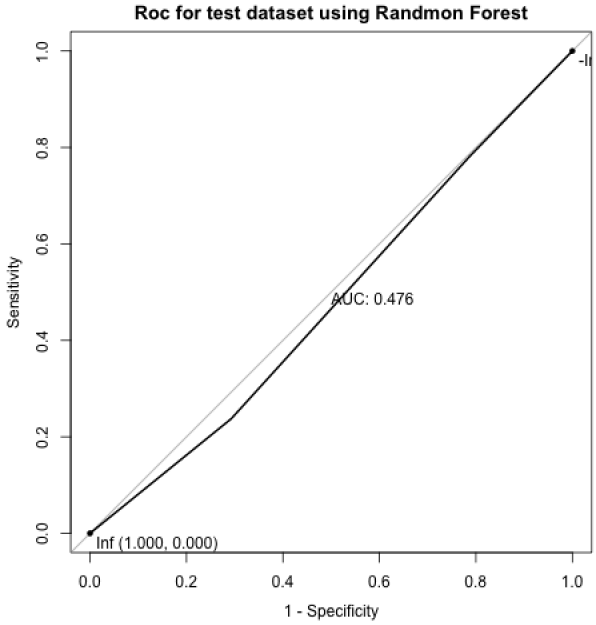


**Figure 3.8:** ROC curve - Random Forest model (Test)

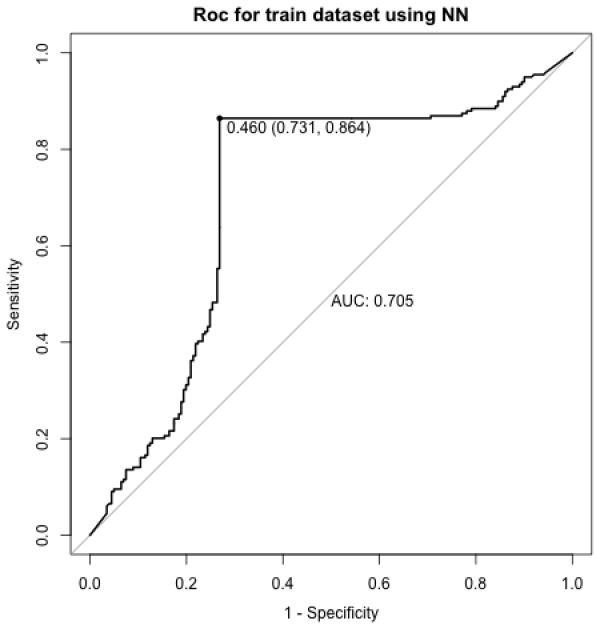


**Figure 3.9:** ROC curve - Neural Network model (Train)

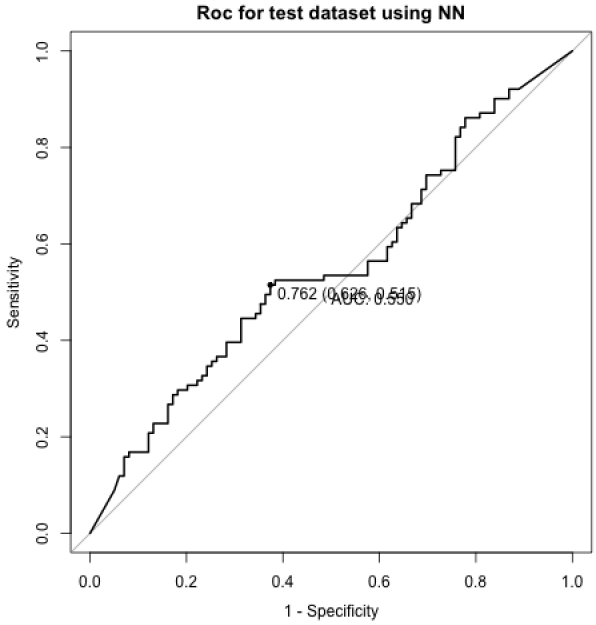


**Figure 3.10:** ROC curve - Neural Network model (Test)

In summary, the summary of AUC values was showed in Table 3.2. Among the models of LASSO, SVM, Random Forest, Neural Network, and PLS-DA, PLS-DA provided the best separation between cancer group and healthy control group in test dataset. The model had an AUC value of 0.614 (Figure 3.12) and an the optimal cut-off value of 0.524 with a specificity and sensitivity of 55.1% and 76.3%, respectively. We also noticed that the PLS-DA model in train dataset had an AUC value of 0.651 (Figure 3.11) with threshold value of 0.523, specificity of 57.5% and sensitivity of 82.6% which were very close to test set. It indicated the model was success in robustness. The loading weight of component 1 and component 2 of the PLS-DA model was shown in Figure 3.13. Especially, His,PC.aa.C30.2, lysope.a.c18.0, t4.OH.Pro were among the top 5 metabolites according to loading values for component 1. The classification error rate of PLS - DA model in test set was 34% based on 5-fold cross validation.

| Predictive Model | AUC - Train set | AUC - Test set |
|---|---|---|
| LASSO | 0.690 | 0.596 |
| SVM Linear | 0.667 | 0.546 |
| SVM Nonlinear | 0.858 | 0.574 |
| Random Forest | 0.893 | 0.476 |
| Neural Network | 0.705 | 0.515 |
| PLS-DA | 0.651 | 0.614 |

**Table 3.2:** AUC of ROC curve for models with all features

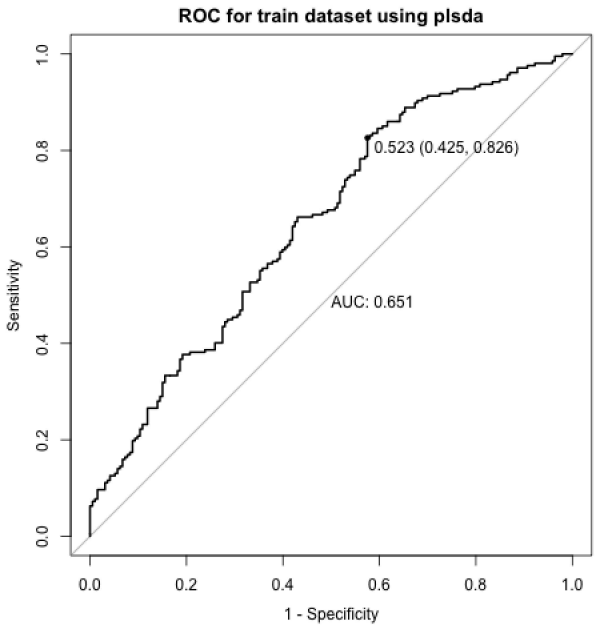


**Figure 3.11:** ROC curve - PLS-DA model (Train)

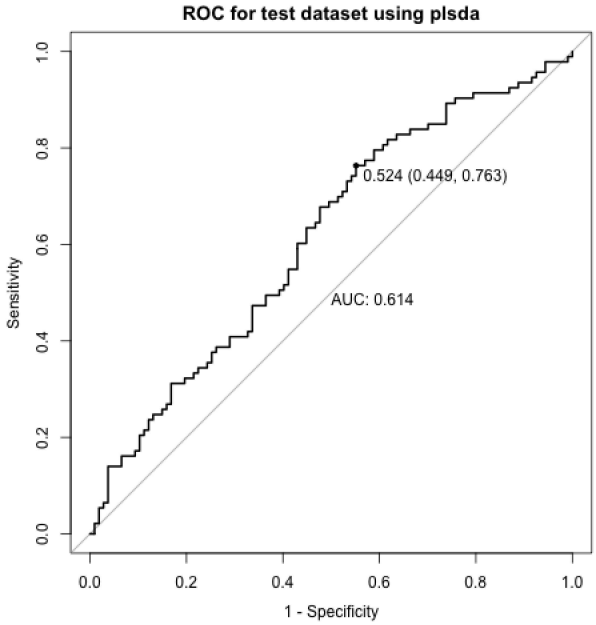


**Figure 3.12:** ROC curve - PLS-DA model (Test)

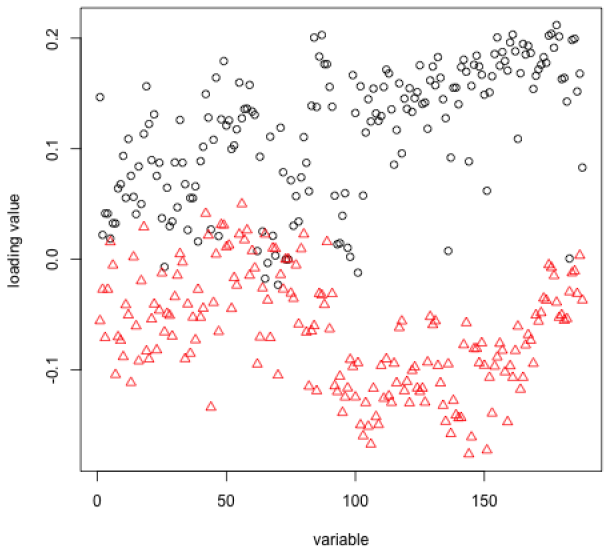


**Figure 3.13:** Loading weight of comp.1 and comp.2

## 3.5 Modelling with Selected Metabolites

In the previous section, we constructed predictive models using all metabolites. However, after analyzing metabolites, we noticed that some metabolites are not independent. In order to decrease the chance of collinearity and also reduce cost in lab, in this section, we build models based on the selected metabolites. To achieve the goal, we excluded those metabolites with variance inflation factor (VIF) value greater than 10. Thus, we kept 26 metabolites: Arg, His, Orm, Phe, Pro, Ser, Thr, Dopamine, Serotonin, t4.OH.Pro, lysoPC.a.C17.0, lysoPC.a.C18.0, lysoPC.a.C18.1, lysoPC.a.C18.2, lysoPC.a.C20.4, PC.aa.C30.2, PC.aa.C32.1, etc.

For logistic regression with LASSO penalty, we saw that the AUC value of test dataset was 0.646 with sensitivity of 58% compared to the AUC value

of 0.60 in the training set. There was no significant difference in the AUC values between model with all metabolites and with selected features. One of the reasons is that LASSO method is a good tool to remove unimportant features towards zero from the model. In that case, if several features were highly correlated, only one of them would be included in the model by applying LASSO penalty.

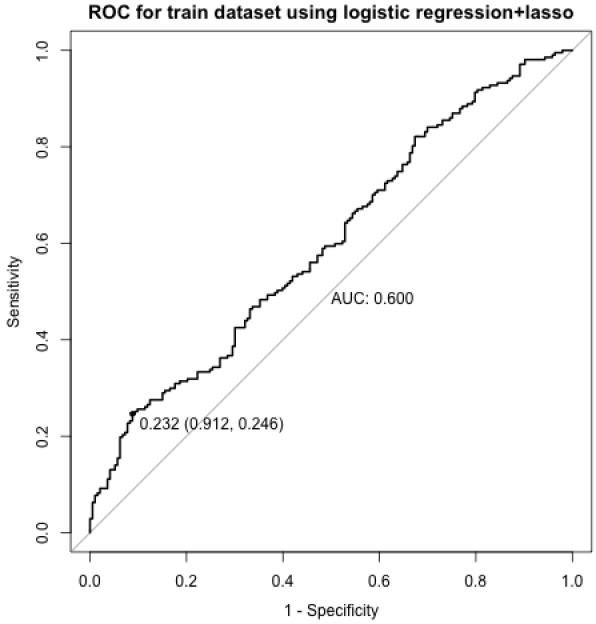


**Figure 3.14:** ROC curve - LASSO model with selected features (Train)

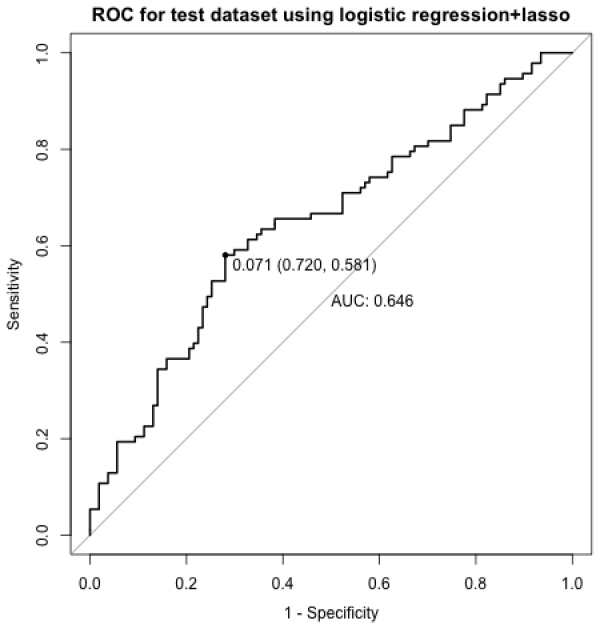


**Figure 3.15:** ROC curve - LASSO model with selected features (Test)

Next we looked at the results of SVM methods by using selected features. From the figures, it showed the non-linear model offered a better performance than the SVM model with linear boundary. For SVM-Nonlinear model, it had an AUC value of 0.667 for test set (Figure 3.19) with a specificity and sensitivity of 54% and 84.9%, respectively. The best kernel function was polynomial kernel with gramma of 0.1 and degree of 2. The AUC value of training set was 0.647 which was close to the value of test set.

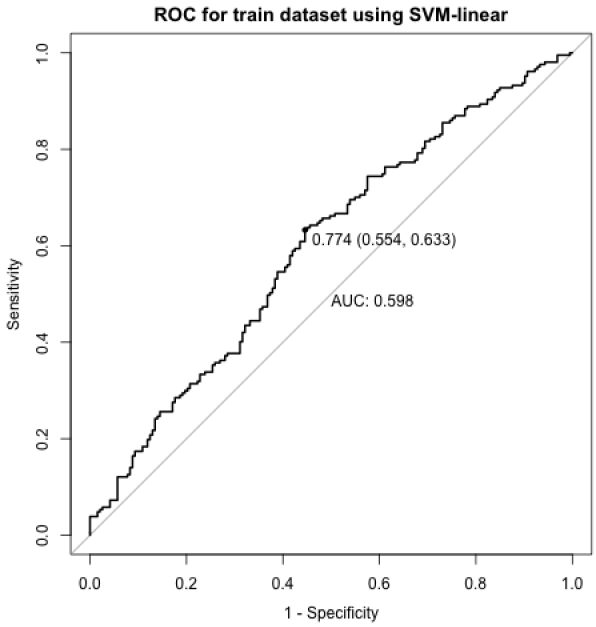


**Figure 3.16:** ROC curve - SVM Linear model with selected features (Train)

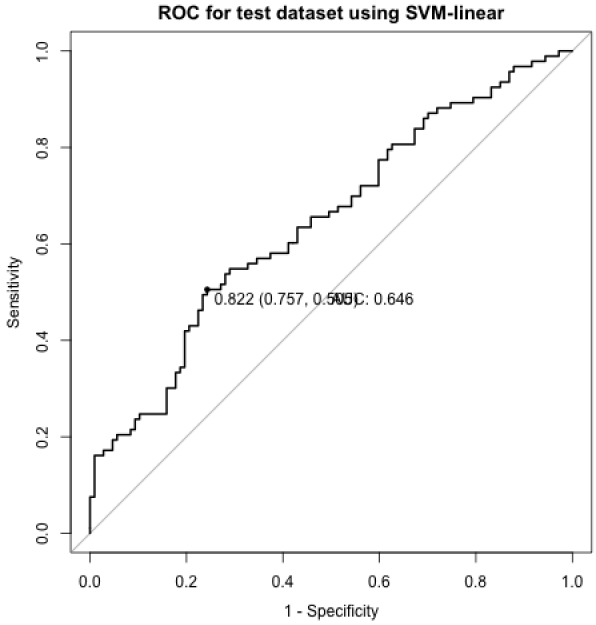


**Figure 3.17:** ROC curve - SVM Linear model with selected features (Test)

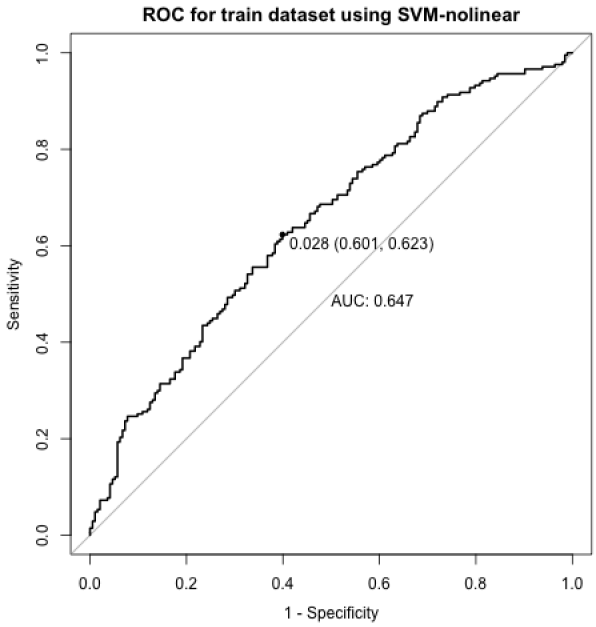


**Figure 3.18:** ROC curve - SVM Nonlinear model with selected features (Train)

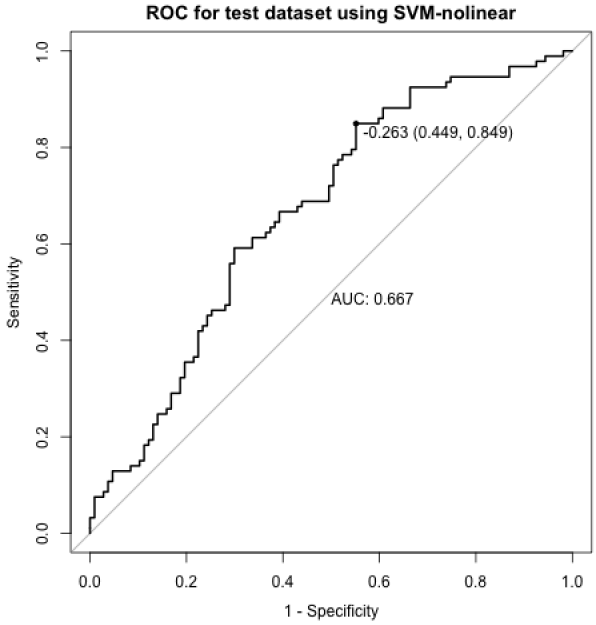


**Figure 3.19:** ROC curve - SVM Nonlinear model with selected features (Test)

For random forest, we still observed that the model was overfitting in training set. It illustrated that random forest approach was not a good method to work with small data set.

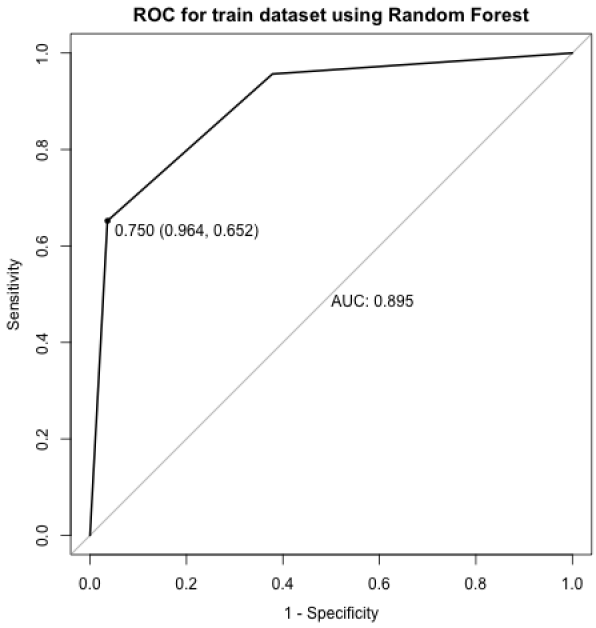


**Figure 3.20:** ROC curve - Random Forest model with selected features (Train)

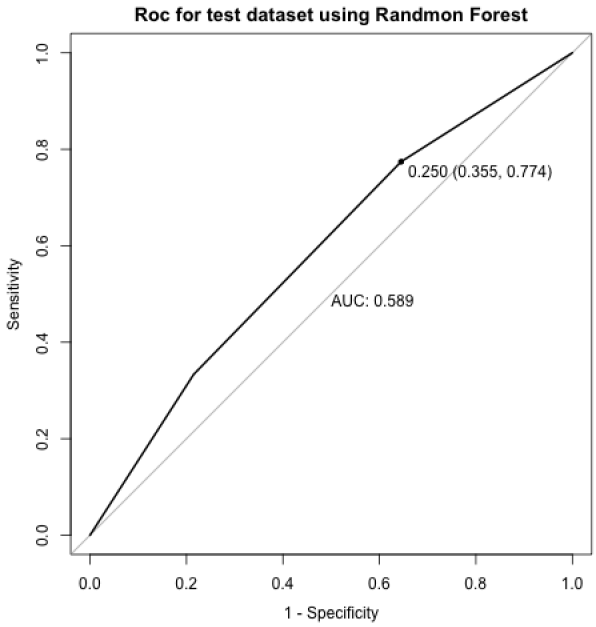


**Figure 3.21:** ROC curve - Random Forest model with selected features (Test)

PLS-DA had an AUC value of 0.669 for test set (Figure 3.23) with an optional cut-off value of 0.502 with a specificity and sensitivity of 39% and 73%, respectively. The loading weight of the component 1 and component 2 of the PLS-DA model was shown in Figure 3.24. The top 5 metabolites of component 1 were His, lysoPC.a.C18.2, PC.aa.C42.2, t4.OH.Pro and SM.C24.1. The classification error rate in test set was 31% by 5-fold cross validation.

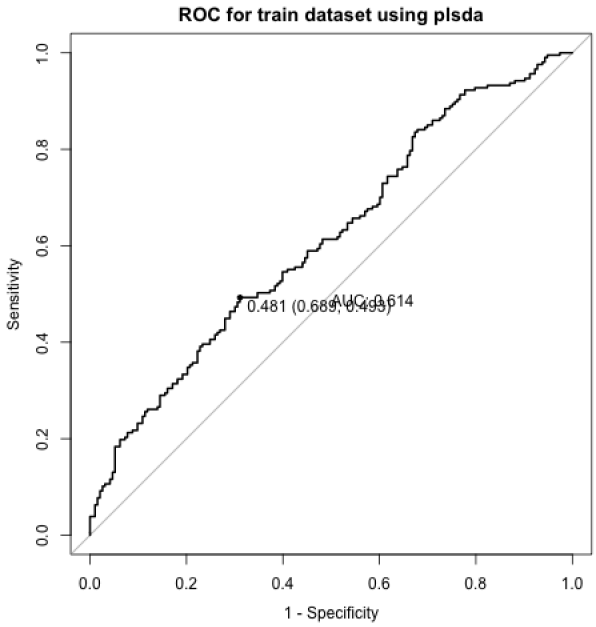


**Figure 3.22:** ROC curve - PLS-DA model with selected features (Train)

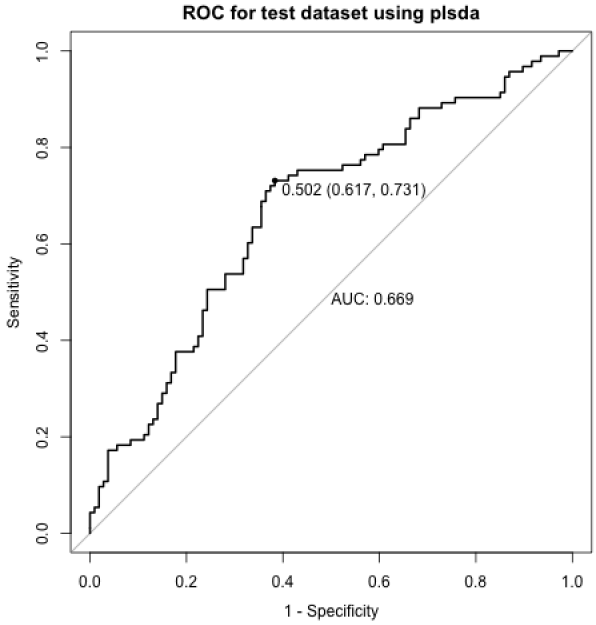


**Figure 3.23:** ROC curve - PLS-DA model with selected features (Test)

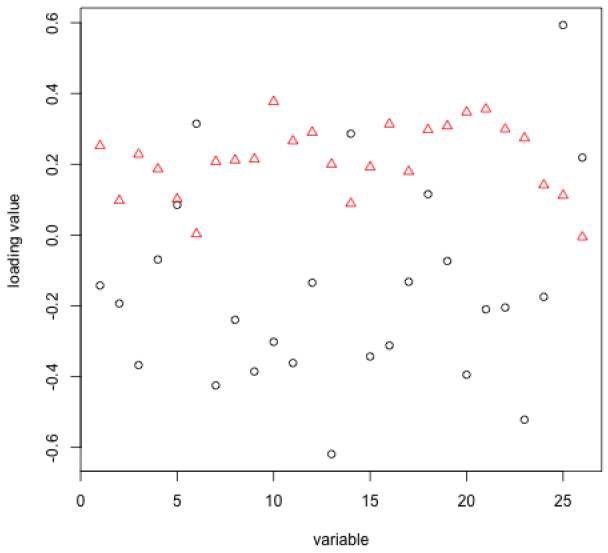


**Figure 3.24:** Loading weight of comp.1 and comp.2

For neural network, after reducing collinearity, we saw the model had a better performance with AUC values of 0.645 and 0.654 for training set and test set respectively, compared to the prior model with all features. It also showed the robustness of the model. In other words, the model worked well by being tested on the new independent dataset.

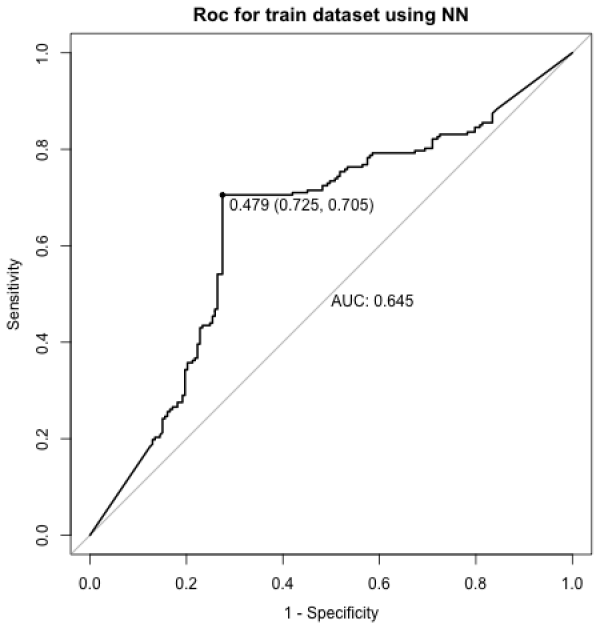


**Figure 3.25:** ROC curve - Neural Network model with selected features (Train)

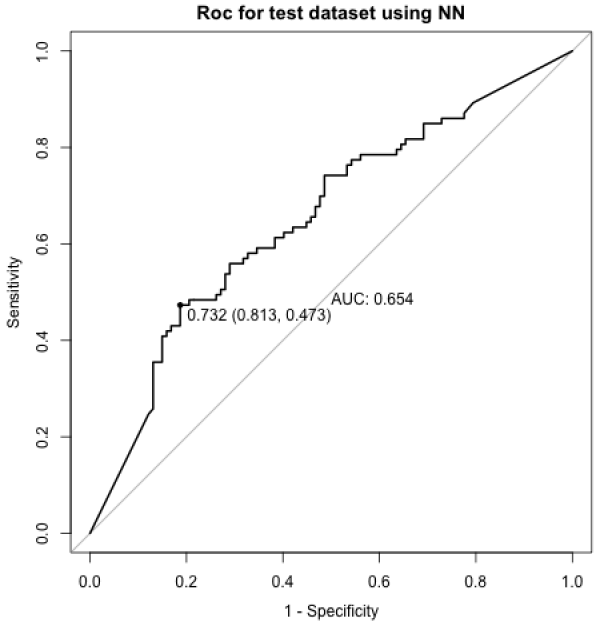


**Figure 3.26:** ROC curve - Neural Network model with selected features (Test)

In summary, the result of AUC values was shown in Table 3.3. Among those models, SVM Nonlinear model provided the best classification with AUC values of 0.667 and 0.647 in test and training set, respectively. In addition, after reducing the chance of collinearity, most machine learning methods such as SVM, Neural networks worked better than before.

| Predictive Model | AUC - Train set | AUC - Test set |
|---|---|---|
| LASSO | 0.600 | 0.646 |
| SVM Linear | 0.598 | 0.646 |
| SVM Nonlinear | 0.647 | 0.667 |
| Random Forest | 0.895 | 0.589 |
| Neural Network | 0.645 | 0.654 |
| PLS-DA | 0.614 | 0.669 |

**Table 3.3:** AUC of ROC curve for models with selected features

## 3.6 Discussion

In the previous sections, we built the predictive models by using different methods. We noticed that for some machine learning tools such as random forest, neutral network etc, the AUC values were much worse in test dataset compared to training dataset.

There were several reasons behind the situation. One of the most common reasons was overfitting. It means the predictions correspond too closely to a training dataset, and therefore fail to fit additional data. Covariance shift may also result in the worse performance on test dataset where predictor variables have different distribution in train and test data.

To solve the issue, increasing the sample size is a useful way. Due to funding limitation, in this study, we only involved in 600 samples. For covariance shift, dropping of drifting features is a simple method. As in this, we drop the features which are being classified as drifting. But it might result in loss of information.

# Chapter 4

# Conclusion

Predictive models on all features were first considered. The best model was PLS-DA with AUC value 0.614 and sensitivity of 82.6%. In order to manage laboratory budget and reduce the chance of collinearity, we excluded those metabolites with variance inflation factor (VIF) value greater than 10 and built predictive models based on selected features. The selected metabolites included lysoPC.a, Dopamine, PC.aa.C30.2, PC.aa.C42.1 and lysoPC.a.C18.2, etc. The best models with selected features was SVM non-linear model with AUC value varied from 0.647 to 0.667 highlighting the predictive power of metabolomics for prostate cancer detection. At an optional cut-off value of 0.502, the predictor's sensitivity value was 84%. It represented that 84% of prostate cancer patients who are correctly identified as having the condition. The kernel function of the final model was polynomial kernel with gramma of 0.1 and degree of 2. In summary, the final predictive model had a better performance than present market benchmark (combined PSA test and DRE).

While our method for diagnosing prostate cancer has shown promises, there

were a number of limitations to this study. First, there were only 600 participants. With this size and an even distribution of cancer stages, it was not possible to delve deeper into data set to predict the cancer stage. Second, metabolites vary with circadian rhythms, diet, age, sex, and weight [13],[14],[15] which can be difficult to control. The field of metabolomic research has significant impact on improving disease diagnosis, but it still has a long way to go.

# Bibliography

[1] http://www.cancer.ca.

[2] A. Dockser, The hidden toll of cancer testing: research shows that even benign results have emotional and financial consequences. *The Wall Street Journal Online* (2004).

[3] J. L. Spratlin, N. J. Serkova, and S. G. Eckhardt, Clinical applications of metabolomics in oncology: a review. *Clinical Cancer Research* (2009) vol. 15, no. 2, pp. 431-440.

[4] D. Kumar, A. Gupta, A. Mandhani, and S. N. Sankhwar, Metabolomics-derived prostate cancer biomarkers: fact or fiction. *Journal of Proteome Research* (2015) vol. 14, no. 3, pp. 1455-1464.

[5] C. Ding and H. Peng, Minimum redundancy feature selection from microarray gene expression data. *Journal of Bioinformatics and Computational Biology* (2005) vol. 3, no. 2, pp. 185-205.

[6] T. Huang, S. Wan, Z. Xu et al., Analysis and prediction of translation rate based on sequence and functional features of the mRNA. *PLoS ONE* (2011) vol. 6, no. 16036.

[7] R. Madsen, T. Lundstedt, and J. Trygg, Chemometrics in metabolomics?a review in human disease diagnosis. *Analytica Chimica Acta* (2010) vol. 659, no. 12, pp. 23-33.

[8] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, (1995).

[9] G. James, D. Witten, T. Hastie, *An introduction to Statistical Learning with Applications in R*, Springer, New York, (2013).

[10] https://www.asco.org.

[11] A. K. Kosmides, K. Kamisoglu, S. E. Calvano, S. A. Corbett, and I. P. Androulakis, Metabolomic fingerprinting: challenges and opportunities. *Critical Reviews in Biomedical Engineering* (2013) vol. 41, no. 3, pp. 205-221.

[12] S. Canu, SVM and kernel machines: linear and non-linear classification. *Ocean's Big Data Mining* (2014).

[13] M. E. Bollard, E. Holmes, J. C. Lindon, S. C. Mitchell, D. Branstetter, W. Zhang, and J. K. Nicholson, Investigations into biochemical changes due to diurnal variation and estrus cycle in female rats using high-resolution NMR spectroscopy of urine and pattern recognition. *Analytical Biochemistry* (2001) Vol. 295, no. 2, pp. 194-202.

[14] C. M. Slupsky, K. N. Rankin, J. Wagner, H. Fu, D. Chang, A. M. Weljie, E. J. Saude, B. Lix, D. J. Adamko, S. Shah, R. Greiner, B. D. Sykes, and T. J. Marrie, Investigations of the effects of gender, diurnal variation, and

age in human urinary metabolomic profiles. *Analytical Chemistry* (2007) vol. 79, no. 18, pp. 6995-7004.

[15] N. G. Psihogios, I. F. Gazi, M. S. Elisaf, K. I. Seferiadis, and E. T. Bairaktari, Gender-related and age-related urinalysis of healthy subjects by NMR-based metabonomics. *NMR in Biomedicine* (2008) vol. 21, no. 3, pp. 195-207.

[16] L. Deng, K. Ismond, Z. Liu, J. Constable, H. Wang, O. I. Alatise, M. R. Weiser, T. P. Kingham, D. Chang, and R. N. Fedorak, Urinary Metabolomics to Identify a Unique Biomarker Panel for Detecting Colorectal Cancer: A Multicentre Study. *Cancer Epidemiology, Biomarkers & Prevention* (2019) vol. 28, pp. 1283-1291.

[17] C. Guo, C. Xie, Q. Chen, X. Cao, M. Guo, S. Zheng, et al. A novel malic acid-enhanced method for the analysis of 5-methyl-2'-deoxycytidine, 5-hydroxymethyl-2'-deoxycytidine, 5-methylcytidine and 5-hydroxymethylcytidine in human urine using hydrophilic interaction liquid chromatography-tandem mass spectrometry. *Analytica Chimica Acta* (2018) vol. 1034, pp. 110-118.

[18] V. E. Tso, S. Macleod, K. P. Ismond, R. R. Foshaug, H. Wang, R. Joseph, D. Chang, N. Taylor and R. N. Fedorak, Consistency of Metabolite Determination from NMR Spectra over Time and Between Operators. *Metabolomics* (2015) vol. 5, no. 3.

[19] R. Eisner, R. Greiner, V. Tso, H. Wang, and R. N. Fedorak, A machine-learned predictor of colonic polyps based on urinary metabolomics. *BioMed Research International* (2013) vol. 2013, no. 303982.

[20] J. Chong, O. Soufan, C. Li, I. Caraus, S. Li, G. Bourque, et al. MetaboAnalyst 4.0: towards more transparent and integrative metabolomics analysis. *Nucleic Acids Research* (2018) vol. 46, no. W1, pp. 486-494.

[21] E. Altobelli, P. M. Angeletti, and G. Latella, Role of Urinary Biomarkers in the Diagnosis of Adenoma and Colorectal Cancer: A Systematic Review and Meta-Analysis. *Journal of Cancer* (2016) vol.7, no. 14, pp. 1984-2004.

[22] D. Stoessel, J. P. Stellmann, A. Willing, B. Behrens, S. C. Rosenkranz, S. C. Hodecker, et al. Metabolomic profiles for primary progressive multiple sclerosis stratification and disease course monitoring. *Frontiers in human neuroscience* (2018) vol. 12, no. 226.

[23] K. A. Cao, S. Boitard, and P. Besse, Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. *BMC Bioinformatics* (2011) vol. 12, no. 1.

[24] T. Sing, O. Sander, N. Beerenwinkel, and T. Lengauer, ROCR: visualizing classifier performance in R. *Bioinformatics* (2005) vol. 21, no. 20.

# Appendices

We prove that the output of (2.2) is a solution of (2.1) [9]. Define

$$\chi \in \{(\mathbf{w}, b) : y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) > 0 \text{ for each } i\}.$$

We see that

$$\max_{(\mathbf{w},b):\|\mathbf{w}\|=1} \min_{\forall i \in m} |\langle \mathbf{w}, \mathbf{x}_i \rangle + b| \in \chi.$$

Since $y_i \pm 1$ and $y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) > 0$ for all $i$, we have that $\forall (\mathbf{w}, b) \in \chi$,

$$\min_{\forall i \in m} |\langle \mathbf{w}, \mathbf{x}_i \rangle + b| = \min_{\forall i \in m} y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b).$$

Therefore, we can re-write (2.1) as

$$\begin{gathered} \max_{(\mathbf{w},b):\|\mathbf{w}\|=1} \min_{\forall i \in m} y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \\ \text{subject to: } y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) > 0 \text{ for each } i. \end{gathered} \tag{0.1}$$

Let $(\mathbf{w}^*, b^*)$ is a solution of (0.1) and the margin

$$M^* = \min_{\forall i \in m} y_i(\langle \mathbf{w}^*, \mathbf{x}_i \rangle + b^*).$$

Then $\forall i$,

$$y_i(\langle \mathbf{w}^*, \mathbf{x}_i \rangle + b^*) \geq M^*.$$

Equivalently,

$$y_i(\langle \frac{\mathbf{w}^*}{M^*}, \mathbf{x}_i \rangle + \frac{b^*}{M^*}) \geq 1.$$

We notice that $(\frac{\mathbf{w}^*}{M^*}, \frac{b^*}{M^*})$ satisfies the condition of (2.2). Therefore,

$$y_i(\langle \hat{\mathbf{w}}, \mathbf{x}_i \rangle + \hat{b}) = \frac{1}{\|\mathbf{w}_0\|} y_i(\langle \mathbf{w}_0, \mathbf{x}_i \rangle + b_0) \geq \frac{1}{\|\mathbf{w}_0\|} \geq \|\frac{M^*}{\mathbf{w}^*}\| = M^*,$$

since $\|\mathbf{w}^*\| = 1$. So $(\hat{\mathbf{w}}, \hat{b})$ is a solution of (2.1).