TAMPEREEN TEKNILLINEN YLIOPISTO
TAMPERE UNIVERSITY OF TECHNOLOGY

MD. FACIHUL AZAM

PREDICTIVE MODEL FOR BREAST CANCER SUB-STAGE CLASSIFICATION

Master of Science Thesis

Examiner: Prof. Frank Emmert Streib
Examiner and topic approved on
6th of April, 2016

ABSTRACT

MD. FACIHUL AZAM
Tampere University of Technology
Master of Science Thesis, 55 pages
May 2017
Master’s Degree Program in Information Technology
Major: Signal Processing
Examiner: Professor Frank Emmert Streib

Keywords: RNA_Seq, miRNA, Non-coding RNA, Breast Cancer sub-stage, SVM, PCA

Prediction accuracy of sub-stage classification varies with different staging methods of breast-cancer. In this Master’s thesis we investigate whether there are differences in the performance of machine learning models across different stages in two different staging systems with three different sets of RNA_Seq data for predicting the sub-stage of breast-cancer.

We applied Support Vector Machine method to classify the sub-stage of Surveillance, epidemiology and End Results and Tumor, Node and Metastasis staging system. We carried out tests to see whether the model performance differs in both the staging systems. We also used three different data sets with prior feature selection and investigated the performance of the model in both staging systems with different combinations of sub-stages. To make the performance result more accurate we used cross-validation with performance metric accuracy, sensitivity and specificity. In order to ensure the classifiers` ability of prediction we used three performance metric precisions, recall and F1 score. Finally, we compared the results between the two staging systems and investigated whether protein coding or non-coding RNA gives a better performance. We applied the principle component analysis to reduce the dimensionality and investigate the performance. Our results show that micro-RNAs give a better classification in both staging systems.

Preface

This thesis is made as a completion of the Master's in Information Technology. It is the product of the Master's period, which is the last part of the Master's in Information Technology at Tampere University of Technology, at Signal Processing Department.

Several people have contributed academically, practically and with support to this thesis. At first I would like to express my gratitude to my thesis supervisor, Prof. Frank Emmert-Streib, for his excellent guidance, topic selection, providing assistance and his quick feedbacks during the whole process of the experimenting stage and the writing of the Master's thesis. Thanks to Sailesh Tripati for his useful help in the coding problems and his useful discussions on the topics of cancer. Thanks to Aliyus Musa to his help during downloading the data from TCGA portal.

Finally, I would like to give thanks to my lovely wife for being helpful and supportive during my study time at the Tampere University of Technology.

Tampere, 01.05.2017

MD. Facihul Azam

Table of Contents

List of Figures

List of Tables

LIST OF SYMBOLS AND ABBREVIATIONS

ANN	Artificial Neural Networks
AUC	Area Under Curve
BRCA	Breast Invasive Carcinoma
BP	Biological Process
CC	Cellular Component
cDNA	complementary Deoxyribonucleic acid
DNA	Deoxyribonucleic acid
ER	Estrogen Receptor
fp	False Positive
fn	False Negative
GO	Gene Ontology
HER2	Human Epidermal growth factor Receptor 2
lncRNA	long non-coding Ribonucleic acid
miRNA	micro Ribonucleic acid
MF	Molecular Function
mRNA	messenger Ribonucleic acid
PCA	Principle Component Analysis
PR	Progesterone Receptor
ROC	Receiver operating Curve
RNA	Ribonucleic acid
RNA_Seq	RNA Sequencing

SCM	Subtype Classification Model
SEER	Surveillance Epidemiology and End Results
SVD	Singular vector Decomposition
SVM	Support Vector Machine
TCGA	The Cancer Genome Atlas
TNM	Tumor Node and Metastasis
tp	True Positive
tn	True Negative
UV	Ultra Violate
Σ	covariance matrix
λ	eigen value
σ	standard deviation

1. INTRODUCTION

Cancer is not a single disease but rather a group of diseases. Cancer cells grow inside the body involving abnormal growth of cells when genes deny to let some of the cells to die. As a result the cells are growing and forming tumor that turns into cancer in the human body. Breast cancer is the second most common cancer type after skin cancer in women in the United States. Approximately 1 in 11 women is developing the malignancy and 1 in 30 is dying from the disease [14]. Both men and women can have beast cancer but it is rare in men. There are many types of breast cancers, among which the most common one is called Breast Invasive Carcinoma (BRCA) [1]. This cancer grows from the ductal gland and spreads all over the breast and in later stage it spreads to other parts of the body. Breast cancer can be diagnosed by X-ray, blood test or other pathological tests. Many researchers are focusing on the gene based analysis breast cancer classification [2,21,22,19]. Cancer prediction and prognosis using machine learning algorithm is not novel in the cancer research area. More than 1500 papers have been published to identify, classify and prognoses cancer using machine learning [3]. Artificial Neural Network (ANN) and the state of art technique have been used in cancer prediction and prognosis for the last 20 years [4-6]. Those researches were mostly based on clinical laboratory data such as patients cancer cell image samples, ultra sound data or other factor of cancer such ER, relapse etc. as input to the system and the output determines whether a person has cancer and also determines the factor causing cancer. The primary goals of cancer prediction and prognosis is different from detection and diagnosis [3].

Many cancer predictions have been performed on the basis of subtype classification, cancer or not cancer patient classification or survivability analysis [9-11]. All of these research outcomes are based on the patients' clinical information. The researchers focused on the patients' age, sex, race, color, smoking habit (in the case of lung cancer), Estrogen, Progesterone receptor or HER2 as a key factor (in the case of breast cancer). The research outcome reveals that the key factors which cause the risk of a person developing breast cancer are ER+, PR+ or HER2. Also it shows the probability of survival after the positive diagnosis. In fact cancer prognosis involved different subsets of biomarkers and multiple clinical factors including the general health, the family history and the age of the patient. Typically the histological (cell-based), clinical and population based information needs to be integrated with the help of an attending physician to come up with a reliable prognosis. Prognostic and predicting breast-cancer or other cancer macro-scale factors including hormone receptor, family history, age, sex habit or environmental effects (uv radiation, radon-induced) [9] are not enough to make a robust prediction. Basically, what is needed is some very specific molecular details about the tumor or the patients` genetic information [13] such as micro-scale factors to be able to give a robust prediction of cancer.

The rapid growth of human cancer genome databases [36] allows researchers to predict cancer more robustly. The last 10 years many researchers have proved that DNA micro array can predict breast cancer better and more accurately than clinical data which in turn helps the physicians to offer a more efficient treatment [16]. About 90% of the breast cancer is always due to the abnormality in the genes and only 5% to 10% gene abnormality is inherited from families [15]. A study of breast tumor samples [17] shows that breast tumor samples can be classified according to DNA microarray gene expression. Hierarchical clustering was used to classify on the basis of similarity and expression.

Several researchers have found treasure in statistical analysis of microarray profile [18] and showed the gene signature that is associated with breast cancer. The author used principle component analysis (PCA) to sample the data and performed the famous hypothesis test to find the association of the random signature with breast cancer. Significant performance has been shown by using Affymetrix data with 70 gene signatures [19].

The comparison of breast cancer molecular subtype prediction has been done in [22]. The authors analyzed previous researches, thirty-six published samples (5715) and five published classifiers. It has been noticed in the article that the three-gene Subtype Classification Model (SCM) gives a significantly better performance than the others. The SCM model worked with micro array data and the author used three bio markers from the group of genes that are closely related to the molecular subtypes of the breast cancer. The SCM applied the gaussian mixture model to predict the class levels of the genes. The authors of [22] show that statistically the SCM method with three key genes gives a better performance. Hierarchical clustering model is another method which has been applied by many researches to predict subtype classification [17].

A recent discovery of the powerful next generation sequencing (RNA_Seq) drew the attention of researchers to investigate the class prediction and the prognosis of cancer. Monitoring the RNA_seq expression and the transcriptome analysis can detect and give the prognosis of cancer accordingly. The RNA_seq can also be used as a risk probability factor in patients` genome. It can also determine the variant of expression in the cancer cell. Recently micro-RNA plays the key role in metastatic cancer diagnosis [20]. Up to now more than 2000 miRNAs have been identified, among which some of the sets of miRNA play a basic role of signature in disease states. Those miRANs are called circulating RNAs as they can circulate over the blood in the blood vessel. These miRNAs are responsible for growing the cells. Several researchers claim that a few group of miRNA are sufficient enough to classify the lymph-node based subtype of cancer [21]. Their results show that these genes are more deferentially expressed in the high risk breast cancer patients than in the low risk breast cancer patients. Micro RNA is a small-non-protein-coding RNA that can act as the function of negative gene control. Recent research shows that they can control hundreds of genes. Gene mutation and mis-expression of certain miRNAs can cause different types of advanced level cancers in the human body. Finding the signature genes in metastasis cancer with the help of using statistical analysis has been done in [19,20,21,23]. These studies show micro-RNA and lncRNA in expression level plays a significant role in the metastasis breast cancer. The authors of [23] used cox regression model and kaplan-meire analysis using ROC as performance metric to classify sub-stages of breast-cancer and to identify bio markers. The author claims that three lncRNA (CAT104, LINC01234, and STXBP5-AS1) predicts metastasis breast-cancer with very good AUC and good sensitivity and specificity. They used TNM system as different class level of breast-cancer. A few researchers show that miR-139, miR-486 and miR-21 are the key genes for developing metastasis breast cancer [21].

In fact RNA_Seq gene data analysis has become more and more interesting to researchers as it helps in finding biomarkers and also for predicting the subtypes of cancer. Long sequence RNA and small sequence RNA both have potential influence in the progression to the advanced stage of cancer. The differentially expressed RNA, combining with noncoding RNA and miRNA can be a cause of the progression of cancer in human body.

2. THEORETICAL BACKGROUND

In this section we will discuss the background knowledge that is crucial to understand the breast cancer data analysis. At first we will discuss the are a of biology where our analysis method has been implemented. After that we will continue with the machine learning method that was used in our analysis such as the principal component analysis and the support vector machine. At last we will briefly explain the machine learning library libsvm used in R programing language and will finish the section with the gene ontology which we used in our study.

2.1 RNA

Ribonucleic acid (RNA) is one of the three major biological macromolecules that are necessary for all forms of life. It was believed for many years that RNA plays only three major roles in the cell- the roles as a DNA prototype (mRNA), as a coupler between genetic code and a protein building block and as a structural component of ribosome [26]. In the area of the genomic analysis many researchers dealing with RNA sequences are working with two types of RNAs: protein coding RNAs ie. simply RNAs and non- coding RNAs, for instance micro-RNAs and lnc-RNAs. Protein coding RNAs are those genes which encode proteins and non–coding RNA are the ones that do not encode proteins. The non-coding RNA is also known as the junk DNA. Many researchers identified a large number (5446) of human lnc-RNAs [26]. Small non-coding RNAs and microRNA are short, 18-22bp nucleotides [27,28]. Micro RNAs regulate gene expression often as gene silencers.

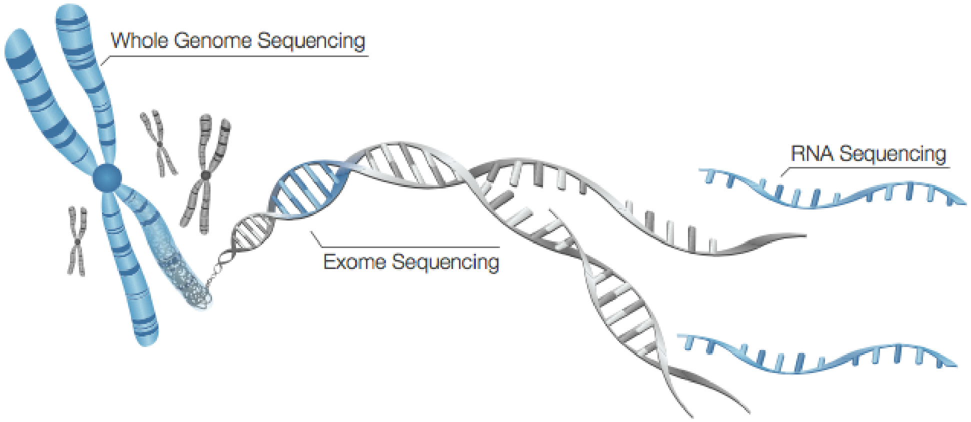


Figure 2.1.1 *RNA extraction from chromosome[27].*

2.2 RNA_Seq

RNA_Seq means using sequencing platforms such as Illumina [35]. It produces millions of short reads of cDNA sequence at low cost. RNA_Seq is a technique for NextSeq which enables rapid profiling and deep sequencing of whole transcriptome. The advantage of RNA_Seq over Gene expression array is – it does not require prior knowledge of transcriptome, provides quantitative and qualitative transcriptome analysis, better specificity, sequence and variant information [29].

A population of RNA fragmented, such as poly (A+) is converted to a library of cDNA with adapters in both ends (From figure 1-EST library with adapters). Each molecule is then sequenced to a high throughput manner and afterwards it obtains a short sequence read.

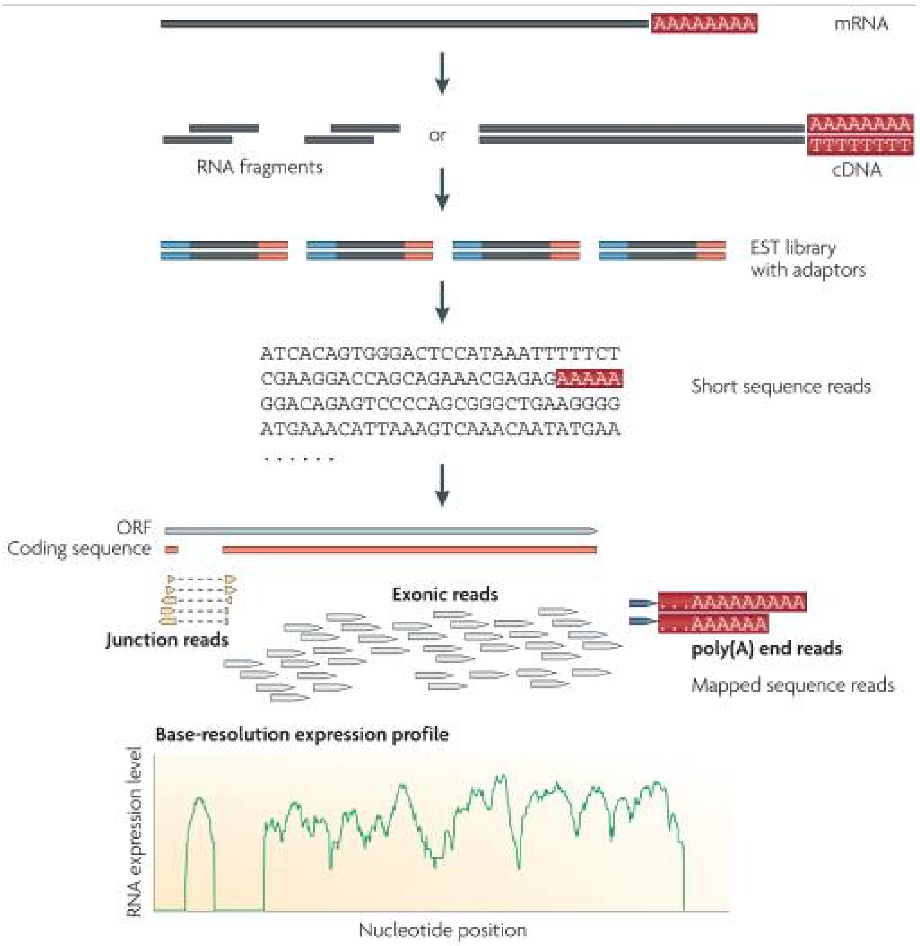


Figure 2.2.1 *RNA_Seq formation[27].*

2.3 Micro RNA

Micro RNA was first discovered in 1993. MicroRNAs (miRNAs) are small (10-20 bp) regulatory RNAs in humans and play a key role in cancer invasive [24]. They play the key role in directing mRNA and regulating protein binding with RNAs. Each miRNA can target more than 100 protein coding RNAs. Over expression of miRNA can cause various types of cancers such as breast cancer, lung cancer or abdominal cancer. Circulating miRNAs have also been discovered in peripheral blood circulation and are also identified in the higher stage of cancer. Most probably miRNAs have been the most extensively studied the last few years. These short nucleotide non coding RNAs regulate cellular transcriptome and proteome by destabilizing mRNA. Recently the miRNAs are being chosen as a good biomarker for different types of diseases including cancer because they are stable in the blood fluid of the human body and can be measured with high sensitivity since they are amplifiable.

These small noncoding RNAs (miRNAs) can be used as the therapeutic treatment for many diseases including the Alzheimer disease. In one research the miR-206 antagomir increased the brain derived neurotrophic factor levels and improved the memory function in rats with Alzheimer. Not only miRNA but also mRNA or other protein together with miRNA can cause the progressing stages of some cancers. RNA is involved in communicating intercellular functionalities in several steps [25]:

- ◆ 1. RNA can carry information in a simple and efficient way. For example mRNA is involved in the coding of proteins.
- ◆ 2. It can coordinate cellular activity in a fundamental and essential manner. For instance miRNA regulates transcriptome and proteome in a cellular level.

A cell can transport miRNA to different cells called recipient cells where miRNA can function the same way.

2.4 Breast Cancer Sub-stages

There are many different systems of breast cancer sub-stage classification. In this thesis only two systems will be described.

2.4.1 TNM Staging System

The most widely used tool to determine the stages of breast cancer is TNM staging system [39]. Staging is a way to determine where the cancer cell is located and how far it has spread all over the body. The abbreviation TNM comes from three terms: Tumor, Node and Metastasis.

In the TNM system, “T” stands for Tumor, T plus a number or letter is used to denote whether a tumor is identified or not in the patient`s body and determines how big it is. Tx means that tumor is primarily identified in the body, while T0 means that there is no tumor in the body. T1, T2, T3 and T4 determines the size of the tumor.

The “N” in the TNM system stands for the lymph node. The lymph nodes may contain the cancer cells at different parts of the body. N plus x or 0 indicates whether the cancer cells are present and N plus a number with the value of 1 to 3 shows that which part of the body the lymph nodes with the cancer cells are located. Nx means that it cannot be evaluated whether the lymph node contains cancer cells, N0 means that there is no cancer in the lymph node. N1-N3 can be subcategorized and the category given determines where the cancer cells are located in the lymph node.

Metastasis or in short “M” determines whether the cancer has spread to other parts of the body. M plus x, 0 or 1 determines whether the cancer has spread to other parts of the body. M0 means that it has not spread, Mx means that it cannot be evaluated and M1 means that cancer has already spread to other organs of the body. A fter examining these three values the doctors can identify the stage of the breast cancer. There are nine stages of breast cancer from 0 to IV, such as stage 0, stage IA, stage IB, stage IIA, stage IIB, stage IIIA, stage IIIB, stage IIIC and stage IV.

Stage I to IIA is the early stage and stage IIB to stage IIIC means that the cancer has advanced locally. Here we select only the four stages according to patient number. These stages are described as follows:

- **Stage IA:** Indicates that the tumor is small, invasive and has not spread to the lymph nodes (T1,N0,M0).
- **Stage IIA:** Indicates the following results: there is no evidence of tumor in the breast and it has not spread to the auxiliary lymph nodes (T0,N1,M0) / the tumor is 20 mm or smaller and has spread to the auxiliary lymph nodes (T1,N1,M0) / the tumor is 20mm to 50mm and has not spread to the auxiliary lymph nodes or to other parts of the body (T2,N0,M0).
- **Stage IIB:** Indicates that the tumor is 20mm to 50mm and has spread to one to three of the auxiliary lymph nodes (T2,N1,M0) / the tumor is larger than 50mm and has not spread to the lymph nodes (T3,N0,M0).
- **Stage IIIA:** Indicates any size of cancer or tumor larger than 50mm being present and that they have spread to 4 to 9 auxiliary lymph nodes.

2.4.2 SEER Staging System

The Surveillance, Epidemiology and End Results (SEER) is a more simplified system and is used for cancer registry and health research. The system can be explained as follows:

- **Local Stage:** It refers to the cancer identified in the breast and means that it stays only in the identified area. (In the TNM system stage I and stage IIA are equivalent to this stage.)
- **Regional Stage:** In this stage cancer has spread to the surrounding tissues and / or lymph nodes. Sometimes it is also being called as the advanced stage. (In the TNM system stage IIA and stage IIIA together belong to this stage.)
- **Distant Stage:** This stage refers to the cancer that has been spread to other parts of the body or lymph nodes above the collar bone. (In the TNM system stages IIIC and IV corresponds to this stage.)

2.5 Principle Component Analysis(PCA)

Principle Component Analysis (PCA) is a multivariate data analysis technique that analyzes the data and that which elements are intercorrelated and dependent variables. Presumably PCA is the most popular data dimension reduction technique used in many numerical data analysis and feature selection. In [18] authors select features from micro array data and apply hypothesis technique to the analysis of the breast cancer genes. The goal of the PCA is to extract the important information of the data variables and to map them into a similar observation called principle components. These principle components are holding the maximum information of the data.

"*Mathematically,PCA depends upon eigen-deomposition of positive semi-definite matrices and upon the singular value decomposition (SVD) of rectangular matrices*" [40]. Principle Component Analysis was described independently by many researchers [41,42]. It was first invented by a British mathematician called Pearson [41]. Later on

Hotelling [43] independently formalized it and termed it principle component analysis. The idea of the PCA is to extract important information from the data set and express it as a set of orthogonal variables. To get the principle component, PCA computes linear combination of orthogonal variables. The first principle component is the largest variance. The second principle component is being calculated the same way as the first principle component but with the largest variance under the constrain of orthogonal of the first principle component. The third, fourth and all the other principle components are being likewise computed. The most common application areas of the Principle Component Analysis are data dimensionality reduction, lossy compression, feature extraction and data visualization.

Principle Component Analysis was defined in two different ways by Pearson and Hotelling [41, 43]. First it can be defined as the orthogonal projection of the higher dimensional data onto a lower linear dimension called principle subspace where variance of the projected data is maximized [43]. Secondly, it can be defined as the linear projection where average projection error can be minimized [41]. The basic approach of principle component analysis is very simple. The first step is to compute d dimensional mean vector with dxd covariance matrix Σ for the whole data set. The second step is to compute eigen vector e and eigen value λ. The third step is to sort the eigen values in decreasing order. These eigen vectors with eigen values are the principle components. The mathematical description of the PCA algorithm is explained below.

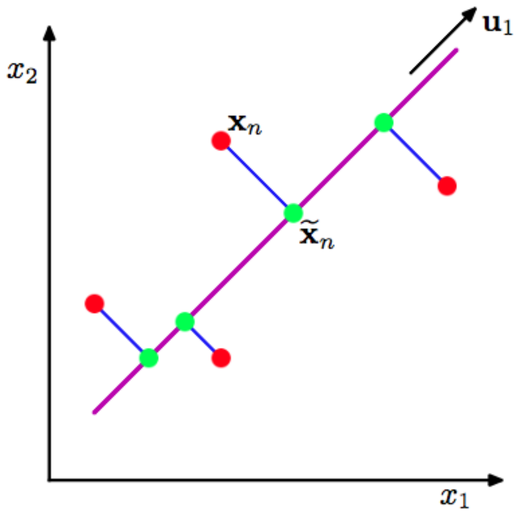


***Figure 2.5.1** Principle Component Analysis [44].*

Consider a data set **X$_n$** where n = 1,2,3 …….*N* . ***X$_n$*** is a euclidean variable with *D* dimension. The dimension of ***X$_n$*** to be reduced to *M<D*

Considering the projection of the data onto one dimensional space *M,* the direction of the data can be defined by *D* dimensional unit vector **u**. where $u^T u = 1$ and all the data from D dimension is projected on the value of $u^T\overline{x}_n$. Where $\overline{x}_n$ is sample mean

$$\overline{x} = \frac{1}{N}\sum_{n=1}^{N} x_n \tag{1}$$

The variance of the projected data is given by

$$\frac{1}{N}\sum_{n=1}^{N} \{u_1^T x_n - u_1^T \overline{x}_n\}^2 = u_1^T S u_1 \tag{2}$$

where S is the data covariance matrix is defined by

$$S = \frac{1}{N}\sum_{n=1}^{N} (x_n - \overline{x})(x_n - \overline{x})^T \tag{3}$$

To maximize the projected variance $u^T S u$ with respect to u_1 ,under the constrain

$uu_1 = 1$, a Lagrange multiplier λ_1 is introduced.

$$u_1^T S u_1 + \lambda_1 (1 - u_1^T u_1) \tag{4}$$

Taking the derivative with respect to u_1 ,equal to zero, we get the equation below. The equation says that u_1 is the eigen vector of S.

$$S u_1 = \lambda_1 u_1 \tag{5}$$

By left multiplying transpose of u^T in equation (5) we get

$$u_1^T S u_1 = \lambda_1 \tag{6}$$

The variance S will be maximum when u_1 eigen vector with eigen value equals to λ_1. This eigen vector is called first principle. All the principal components can be computed likewise by taking another direction of variance. For n dimensional vector we can get n different direction of projection eigen vector of eigen values λ_n.

where λ_1 , λ_2 , λ_3...............λ_n

2.6 Support Vector Machine

The Support Vector Machine (SVM) is a powerful machine learning technique for Two group classification. The idea behind the SVM is that the non-linearly separable input data is being transformed into high dimension feature space [32]. In this feature space a linear decision surface is being constructed. SVM can be implemented for both separable and non separable input data. Linear SVM is being implemented where input data can be linearly separable. The early SVM implementation was made by constructing optimal hyperplane between two separable input data. The Non-linear SVM is implemented where the input data is non separable.

Given a training set of input $(x_i, y_i), i=1,....l, y_i \in (-1,1), x_i \in R_d$ where x is input data and y is class label. Let us consider some hyperplane which separates two classes: the positive and the negative examples. The vector W is perpendicular to the plane in space. The points x which lie on the plane satisfy $W.x+b=0$.

The distance between origin and the plane is $|b|/\|W\|$. Let the minimum distance between the hyperplanes which separates the positive or the negative example be defined as the margin. SVM algorithm simply finds the hyperplane with the largest margin. Let us consider the training data that satisfies following constrains:

$$x_i.W+b \geq +1 \, for \; y_i=+1 \qquad (1)$$

$$x_i.W+b < -1 \, for \; y_i=-1 \qquad (2)$$

combining one set of equalities we have

$$y_i(x_i.W+b)-1 \geq 0 \, for \, all \, i \qquad (3)$$

The two hyperplanes are $H_1 \; : \; x_i.W+b=1$ and $H_2 \; : \; x_i.W+b=-1$, the margin is simply $2/\|W\|$. The two hyperplanes are parallel to each other and no training points fall in between those hyperplanes. Thus we can find the hyperplane that gives the maximum margin by minimizing the $\|W|^2|$

subject to constrain equation (3).

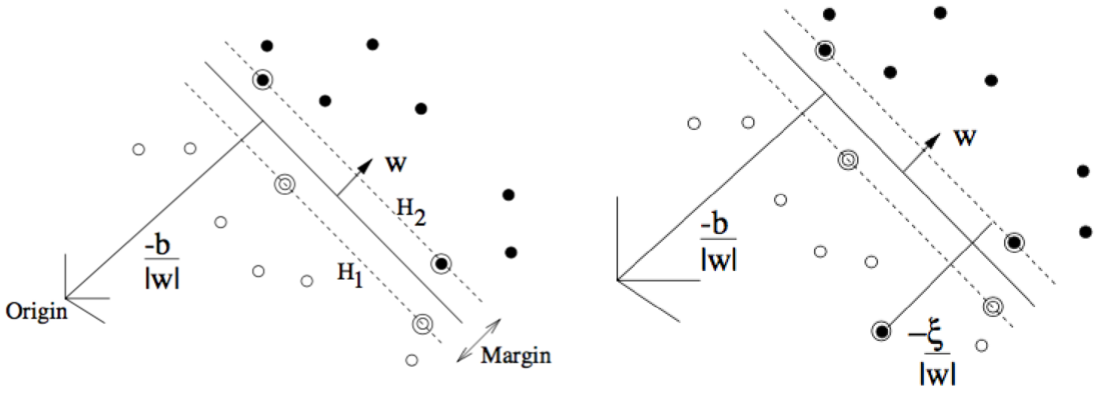


Figure 2.6.1 *left shows separable input data and right non-separable data in Support Vector Machine[45].*

Now let`s suppose that classes overlap in feature space so that data cannot be separable. One way to find the hyperplanes is by letting grow the margin such a way that some points can be located on the wrong side of the margin defined as slack variables $\xi_i, i=1...l$. So the equation (3) can be written as

$$y_i(x_i.W+b)\leq 1-\xi_i \tag{4}$$

$$\xi>0 \tag{5}$$

$$\Phi(\xi)=\sum_{i=1}^{l}(\xi_i) \tag{6}$$

This is the usual way of applying the support vector machine for non separable data [33]. Equation (6) describes the training error. To minimize the training error one can simply remove those data from the training set [34]. For this case the Lagrange function is

$$L_p=\frac{1}{2}\|W\|^2+\sum_{i=1}^{l}(\xi_i)-\sum_{i=1}^{l}\alpha_i[y_i(x_i.W+b)-(1-\xi)]-\sum_{i=1}^{l}\xi_i\mu_i \tag{7}$$

So far the support vector machine describes the linear boundary for classification. To make it more flexible the support vector classifier uses nonlinear kernel functions. There are three popular choices of kernel functions [34]:

dth- Degree polynomial: $(1+\langle x,x'\rangle)^d$

Radial basis: $\exp(-\Upsilon(x-x')^2)$

Neural network : $\tanh(k_1\langle x,x'\rangle+k_2)$

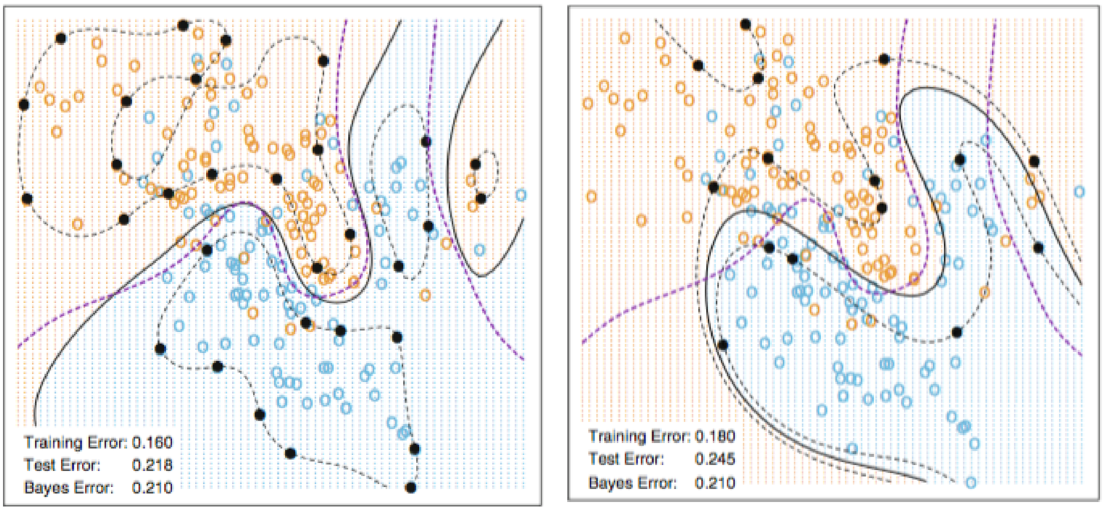


***Figure2.6.1** left shows 4th degree polynomial kernel SVM and right radial basis kernel SVM [34].*

2.7 LIBSVM

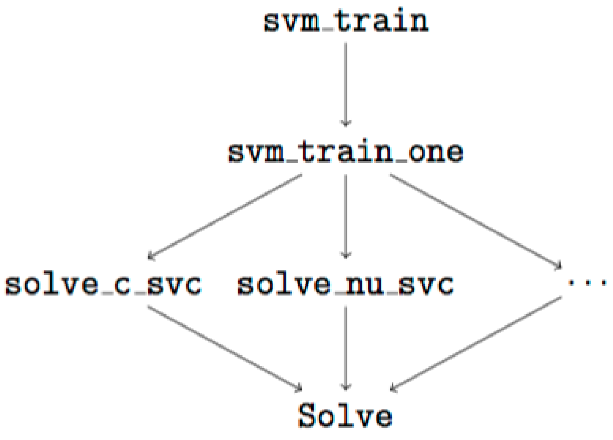
LIBSVM is a library for support vector machine. It can be used in many programming languages like c, c++, python, R etc. LIBSVM supports three different learning tasks[35] such as

1. Support vector classification(two class and multi class)

2. Support vector regression

3 One class support vector.

LIBSVM works in two steps, first it is training the system to obtain the model and then it is predicting the data using the model.



Main training sub-routine

Two class and one class SVM

Various SVM formulations

Solving optimization problems

Figure 2.7.1 *LIBSVM's code organization for training [35].*

2.8 Gene Ontology

Gene Ontology project [36] provides controlled, structured vocabularies, gene ontology (GO) and classifications that describes molecular and cellular biology, and which are freely available for use in annotation of genes, gene products and sequences. GO describes the ontology of gene in three non-overlapping domains of molecular biology. The vocabularies are organized based on three principles which are is-a and part-of-a relationship with (each) other [37]. The three GO terms are:

- Molecular function (MF) describes activities, such as catalytic or binding activities at the molecular level.

- Biological Process (BP) describes biological goals accomplished by one or more biological functions.

- Cellular component (CC) describes the location of sub cellular structures and macromolecular complexes.

In this project we use org.hs.eg.db package for gene mapping.

3. METHODOLOGY AND DATA

We collected the Level 3 RNA_Seq version 2 and the clinical data of BRCA cancer type from the TCGA portal. We chose four different pathological tumor stages of cancer from clinical information of all

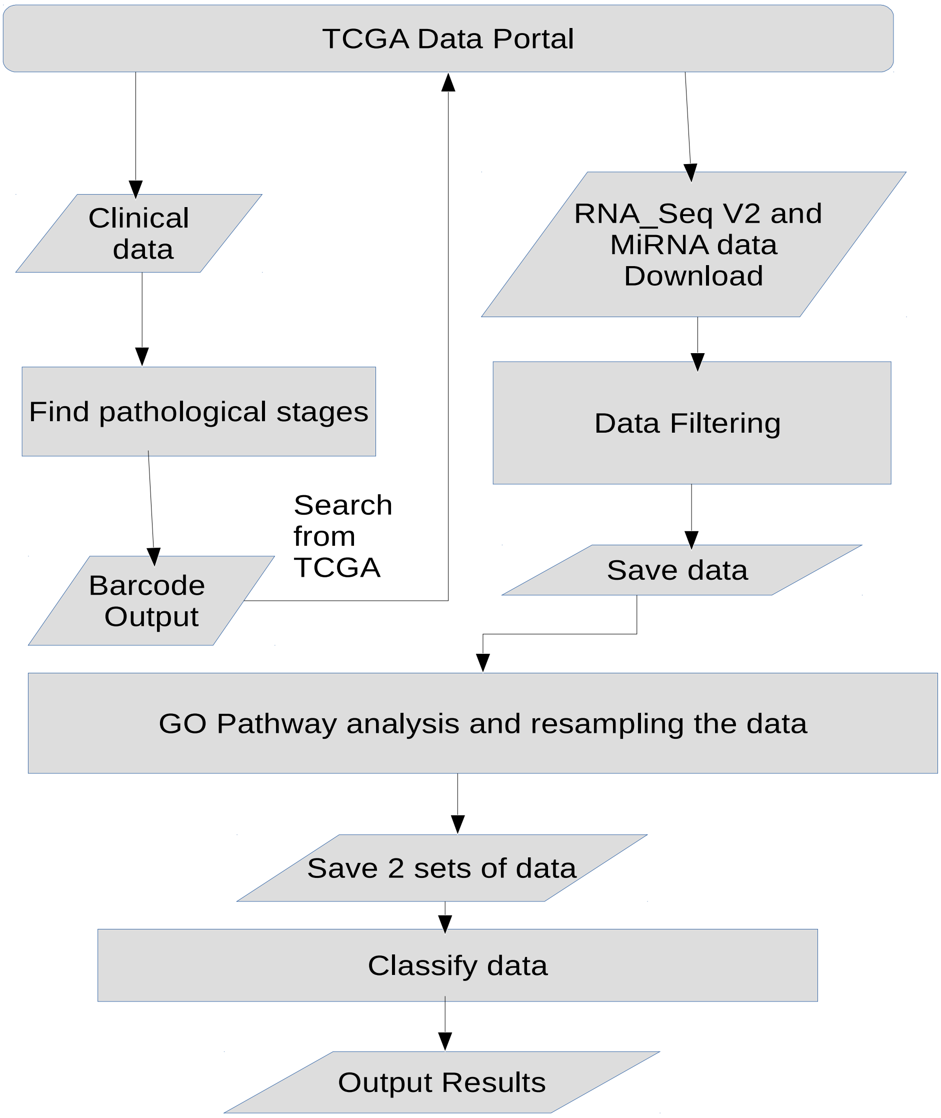


Figure 3.1 *Work flow diagram of the project.*

samples that is stage IA, IIA, IIB and IIIA (see Figure 3.2). The clinical data is filtered according to the stages, Estrogen receptor PR status and her2 status for further processing. We collected the corresponding barcodes of the patients according to the stages and made a subset of RNA_Seq data. The data downloaded from TCGA portal .rsem.-genes.results formate and rsem stands for RNA_Seq Expectation Maximization. The gene data obtained from the TCGA portal is normalized data. We also collected non-coding RNA seq data which is basically Micro RNA (miRNA_Seq) data. All the subsets of data were placed in a matrix separately as noncoding and coding RNA. We labeled the data according to stages such as 1,2,3 and 4 with respect to IA, IIA, IIB and IIIA. Finally we got 2 sets of data with 962 samples of RNA_Seq and 252 samples of miRNA data (see ***Table 3.1***).

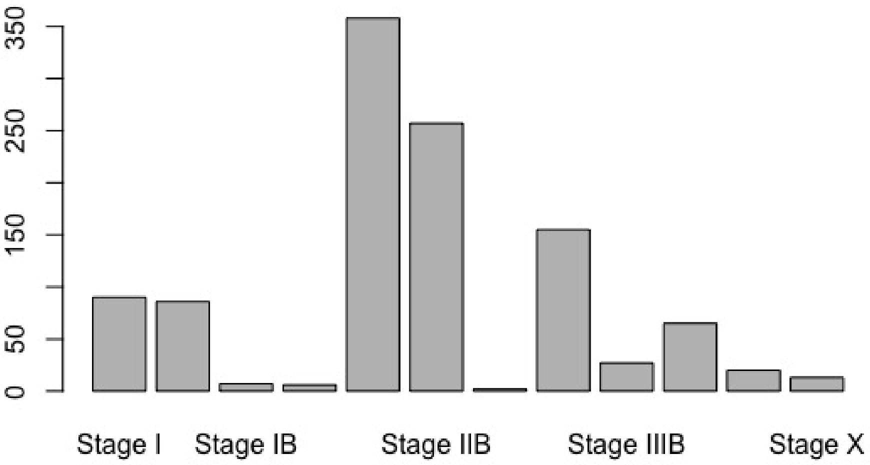


Figure 3.2 *Pathological stage of breast cancer patients of clinical data.*

Table 3.1: *Experimental data set*

Data set	Samples	Features
RNA_Seq V2	962	20,531
miRNA	255	812

RNA_Seq data was being mapped according to the Gene Ontology so that we could separate the significant genes from the large data sets. Those chosen genes are called feature vectors. We used org.Hs.eg.db library downloaded from the bioconductor repository. org.Hs.eg.GO which is an R object that maps entrenz IDs that are directly associated with GO identifiers. After mapping with GO db 18001 expressed genes were identified. RNA_Seq V2 data was divided into two parts, namely ProteinCoding RNA_Seq and NonProteinCoding RNA_Seq data. Those genes which were mapped with GO data base are called ProteinCoding RNA_Seq data and the ones which were not mapped with GO Data base are named NonProteinCoding RNA_Seq data. After GO mapping, obtain data sets were being given (see ***Table 3.2.).***

Table 3.2: *Experimental data set RNA_Seq V2 after GO analysis*

Data Sets	Balanced class samples	Unbalanced class Samples	Features
ProteinCoding RNA_Seq	424	962	18001
NonProteinCoding RNA_Seq	424	962	2240

Figure 3.3 shows clinical information about hormone receptor status that is being used mostly in molecular based analysis. The positive or negative values of Estrogen-Receptor(ER), Progesterone-Receptor(PR) and HER2 mean the presence or the absence of ER, PR, and HER2. These three markers are being used to identify the molecular subtypes of breast-cancer. From the figure it can be noticed that the number of triple negative of ER, PR and HER2 patients is around 200 which is known as basal-like breast-cancer and it is very common in women with BRCA1 gene mutation. This type of cancer is also common in young African-American women. Another type of cancer is called luminal B which is characterized by triple positive of ER, PR and HER2. The number of triple positive patients is around 150.

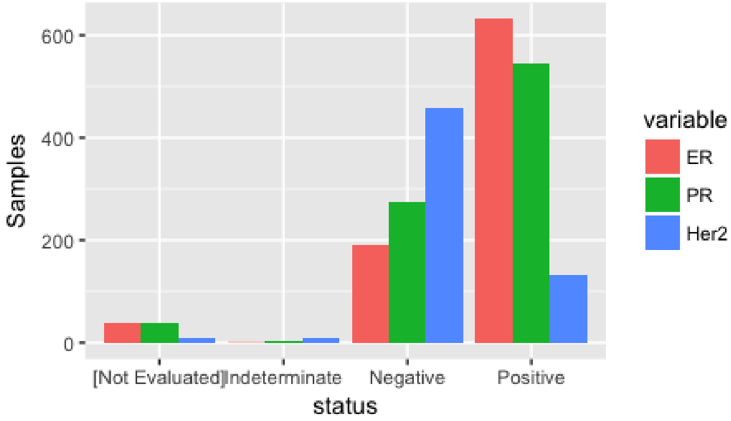


Figure 3.3 *Clinical information of estrogen progesterone and her 2 receptor status of breast cancer.*

Data was sampled according to variance of genes. The histogram of log of variance of three data sets are shown in Figure 3.4. We have extracted features by more frequently occurred variances for protein coding and non-coding RNA_seq and high variance for micro RNA data. The feature selection was performed manually trying different ranges of variances and then we selected the best feature.

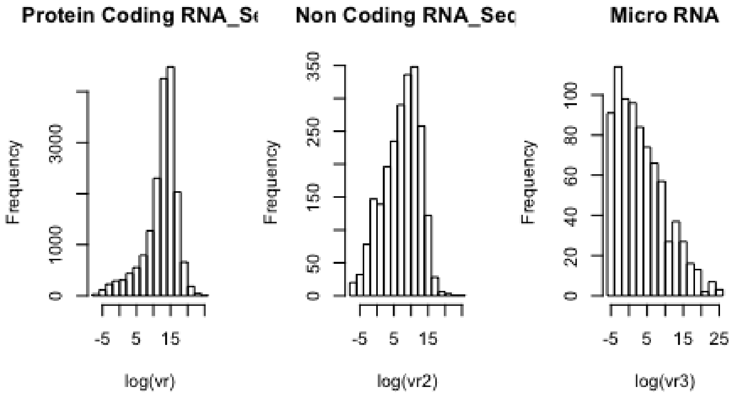
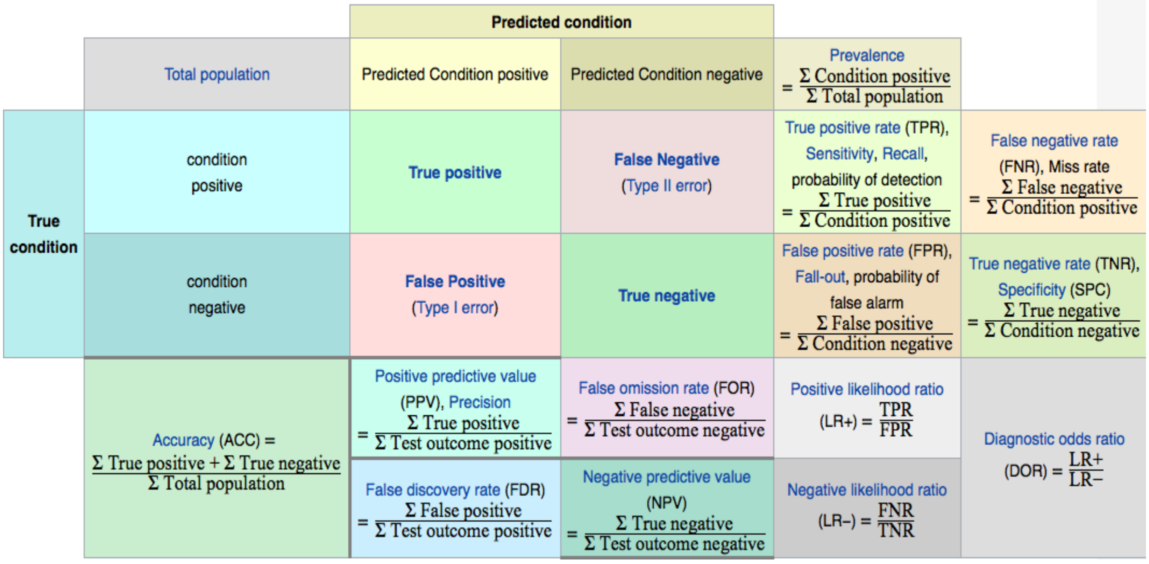


Figure 3.4 *histogram of log of variance of the data.*

3.1 Performance measure

There are four different classifications in machine learning area: binary, multi-class, multi-level and hierarchical classification[38]. In binary classification input is classified into one and only one class out of two non overlapping classes (C1 and C2). This is the most popular classification used in machine learning. Multi-class classification classifies input data into one and only one class out of *l* non overlapping classes. Multi-level classification classifies input data into several classes out of *l* non overlapping classes. In hierarchical classification input is classified into one out of l classes, which are divided into subclasses or subgroups of a superclass.

Our approach was to use multi-class classification as gene data were to be classified into four different subclasses or substages of breast cancer. There are two ways of performance measurements: one is the macro averaging which is the average of the same measures of all the classes C1 ….Cj and the other one is the micro averaging sum which is the sum of the cumulative True Positive (tp), True -negative (tr), False Positive(fp), false Negative(fn) and the measure`s performance.



		Predicted condition			
	Total population	Predicted Condition positive	Predicted Condition negative	Prevalence $= \frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	
True condition	condition positive	True positive	False Negative (Type II error)	True positive rate (TPR), Sensitivity, Recall, probability of detection $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$
	condition negative	False Positive (Type I error)	True negative	False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	True negative rate (TNR), Specificity (SPC) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$
	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	Positive predictive value (PPV), Precision $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Test outcome positive}}$	False omission rate (FOR) $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Test outcome negative}}$	Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR−}}$
		False discovery rate (FDR) $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Test outcome positive}}$	Negative predictive value (NPV) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Test outcome negative}}$	Negative likelihood ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$	

Figure 3.1.1 *Overview of binary classification(wikipedia).*

3.2 Performance measure calculation for Multi-class prediction

Let`s assume the number of the classes is denoted by *l*.

Sensitivity: The ability of the test is to correctly identify those samples which belong to the test class.

$$Sensi = \frac{tp}{tp+fn}$$

The Average per class effectiveness of the classifier to identify the class labels is

$$\frac{\sum_{i}^{l} Sensi}{l}$$

Specificity: The ability of the test to correctly identify those samples does not belong to the test class.

$$Speci=\frac{tn}{fn+tn}$$

The Average per class effectiveness of classifier to identify the class labels is

$$\frac{\sum_{i}^{l} Speci}{l}$$

Accuracy: Accuracy can be measured by calculating the summation of the True Positive over the whole data set.

$$\frac{\sum tp}{N}$$

where tp is true positive and true negative respectively of C(1...c) classes.

N is the number of the test samples.

Here in this project we calculated the standard error measure. The standard error is the standard deviation of sampling distribution of the mean. The formula can be written as:

$$\sigma=\frac{sd}{\sqrt{N}}$$

Where sd stands for the standard deviation of the original error of each class and N stands for the test sample. σ is the mean standard error. The larger the sample size is the smaller the standard error.

3.3 K- fold Cross Validation

Cross validation (also called model validation technique) is a testing method that tests the classifiers accuracy. Cross validation is used to select the best classifier in some independent data set.

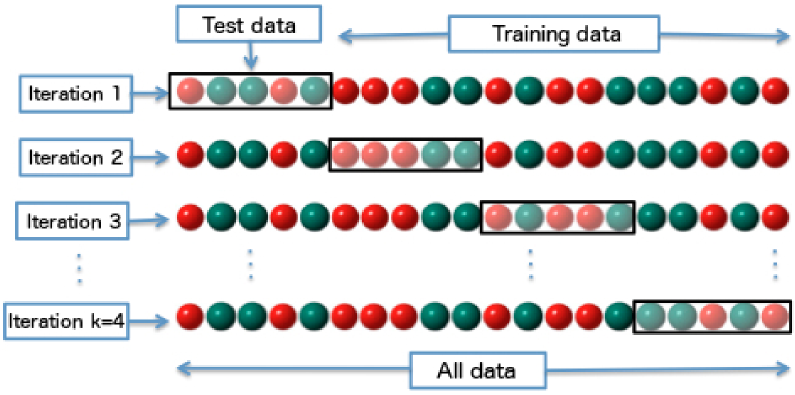


Figure 3.3.1 k-fold cross validation (wikipedia).

In the cross validation technique the whole data set is partitioned into k-sets: the classifier is being tested on one part of the data set and the rest is used for training. There are different types of cross validations. Here only k-fold cross validation will be discussed.

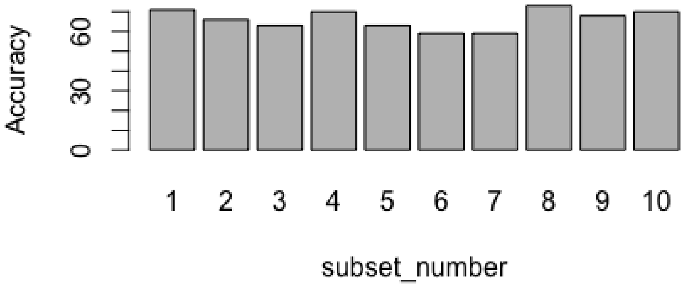


***Figure 3.3.2** Accuracy of different subsets of test data in 10 fold CV of Protein Coding RNA_seq.*

In k-fold cross validation (see Figure 3.5), in the first round, the data is partitioned into k fold and one fold is being kept for testing or validating the classifier. The rest of the data sets are used for training the classifier. To reduce the variability, multiple rounds of cross validation is being performed and the result is being averaged. Taking each accuracy of each fold or subset of test data and performing averaging on them makes the classifier accuracy more accurate. Figure 3.6 shows the different accuracy values in different subsets of test data. Accuracy varies with respect to the subset of the test data. It is clear from the bar graph that mean accuracy calculation reduces the variation of the accuracy and gives accurate results for the classifier. K-fold cross validation technique was used in this study to evaluate the prediction ability of the support vector machine (SVM) method in the breast cancer gene data classification. Here different numbers of cross folds were being used such as 4,5,6 to select the best number of folds. Performance measures reported in this study are the averages of 4,5 and 6 estimates obtained from 4,5 and 6 fold.

4. RESULTS AND DISCUSSION

In this chapter we compare different substages of breast cancer prediction classifications with different sets of RNA_seq data. The results are ordered according to the classifier performance in different data sets. The results are discussed in different assumptions such as balanced and unbalanced class labels and on the basis of how feature selection improves the accuracy and overall performance. We have analyzed the data with the help of two different breast cancer staging systems namely Surveillance, Epidemiology and End Results (SEER) and Tumor, Node and metastasis (TNM). Before the classification of the breast cancer into two different systems of staging we created two sets of RNA_seq data(see Table 4.1, 4.2.1 and 4.2.2). Each data set was classified them based on balanced class labels and unbalanced class labels. The balanced and unbalanced class labels are defined as follows:

- Balanced class labels: Equal size of samples from each stage of breast cancer genes were taken and we labeled them according to the stage name.
- Unbalanced class labels: Unequal size of samples from each stage of breast cancer genes were taken and we labeled them according to the stage name.

The idea behind creating balanced and unbalanced class labels was to see the effect of the classifier on equal and unequal sample sizes of data.

4.1 SEER Staging System sub-stage Classification

The three different data sets protein coding RNA, non-coding RNA and miRNA were classified according to the SEER staging system. It was investigated how the classifier performs in different combinations of stages. In this study we performed the classification task only for the primary and the regional stages. The four different substages of data namely stage IA, IIA, IIB and IIIA were being divided into 2 sets:

Table 4.1 *created data sets with combination of different stages*

Data	Stage	Sample number
Set 1	(Stage IA, Stage IIA) and (Stage IIB, Stage IIIA)	962
Set 2	Stage IA and Stage IIB	396

4.1.1 Unbalanced class labels

The following tables below show the classification results of unequal class labels with the sample size of 962 patients.

Table 4.1.1.1 *Summary of the results for the protein coding RNA_Seq data set1.*

Sensitivity/Specificity Table						
Fold	Accuracy(%)	Standard Error (acc)	Sensitivity (%)	Standard Error (Sensitivity)	Specificity(%)	Standard Error (Specificity)
4	58.67	0.0169	61.56	0.0255	55.48	0.0170
5	57.46	0.0044	59.14	0.0142	55.76	0.0154
6	56.16	0.0205	59.81	0.0176	52.25	0.0389

precision/recall Table			
Fold	precision	Recall	AF1
4	60.05	61.56	0.6076
5	59.14	59.14	0.5906
6	57.80	59.81	0.5858

The Protein Coding RNA_Seq data Set 1 (see section 4.1) prediction performance (table 4.1.1) shows that accuracy, sensitivity and specificity varies vary little amount in different cross folds while standard error varies significantly from 4 fold to 5 fold cross validation. In 5 fold cross validation the classifier gives lower standard error, almost similar accuracy, sensitivity and a good F1 score as well.

Now if we consider only a small portion of the data (for instance stage IA and Stage IIB as class 1 and class 2 respectively) (table 4.1.1.2) the performance differs from the table 4.1.1 significantly. In this prediction although sensitivity decreases 20%, prediction accuracy increases 10% together with 30% increment of Specificity. On the other hand AF1 score decreases with precision and recall.

Table 4.1.1.2 *Summary of the results for the protein coding RNA_Seq data set 2.*

Sensitivity/Specificity Table						
Fold	Accuracy(%)	Standard Error (acc)	Sensitivity (%)	Standard Error (Sensitivity)	Specificity(%)	Standard Error (Specificity)
4	71.18	0.0254	41.36	0.0944	81.83	0.0428
5	70.15	0.0310	34.72	0.0306	82.99	0.0188
6	69.09	0.0208	28.96	0.0423	84.08	0.0197

Precision /recall Table			
Fold precision	Precision	Recall	AF1
4	45.70	41.36	0.4196
5	42.58	34.72	0.3806
6	39.64	28.96	0.3275

The tables (Table 4.1.1.3 and Table 4.1.1.4) indicate prediction performance of noncoding data. By comparing results of the two tables for noncoding data we see that data set 1 gives overall better performance than data set 2 with 5 fold cross validation. In these two tables both accuracy and sensitivity is around 5 % higher in data set 1 than data set 2. Both tables show good AF1 score.

In noncoding data set although accuracy and AF1 score are almost the same, sensitivity is significantly better in data set 2 (table 4.1.4) than in data set 1(table 4.1.3). As compared to the protein coding data, overall prediction performance is better in noncoding data because here accuracy, sensitivity and specificity are in almost equal range, ie. in the range of 55% to 60%.

Table 4.1.1.3 *Summary of the results for the non-coding RNA_Seq data set 1.*

Sensitivity/Specificity Table						
Fold	Accuracy(%)	Standard Error (acc)	Sensitivity (%)	Standard Error (Sensitivity)	Specificity(%)	Standard Error (Specificity)
4	57.60	0.0240	57.35	0.0328	57.09	0.0659
5	61.08	0.0290	60.97	0.0598	60.20	0.0607
6	57.65	0.0222	60.52	0.0160	54.92	0.0394

precision/recall Table			
Fold	precision	Recall	AF1
4	58.13	57.35	0.5749
5	61.12	60.97	0.6046
6	53.11	60.52	0.5812

Table 4.1.1.4 *Summary of the results for the non-coding RNA_Seq data set 2.*

Sensitivity/Specificity Table						
Fold	Accuracy(%)	Standard Error (acc)	Sensitivity (%)	Standard Error (Sensitivity)	Specificity(%)	Standard Error (Specificity)
4	55.32	0.02733	63.22	0.0326	47.09	0.0234
5	55.58	0.05205	56.94	0.0764	53.18	0.0497
6	54.90	0.02834	57.18	0.0461	53.52	0.0428

precision/recall Table			
Fold	precision	Recall	AF1
4	54.01	63.22	0.5796
5	54.85	56.94	0.5548
6	54.84	57.18	0.5533

On the other hand, with miRNA sub-stage class prediction gives better classification accuracy and sensitivity than protein coding and non-coding RNA_seq. From table 4.1.5 the classification accuracy is the same all over cross fold while sensitivity and specificity varies 66% to 70% and 44% to 47% respectively. For the data set 2 with less samples miRNA class prediction performance is better than for data set 1. Table 4.1.1.6 shows that applying the cross fold has no significant effect on the class prediction model but using the data set 2 has a significant effect on accuracy, sensitivity and AF1 score.

Table 4.1.1.5 *Summary of the results for the miRNA_Seq data set 1.*

Sensitivity/Specificity Table						
Fold	Accuracy(%)	Standard Error (acc)	Sensitivity (%)	Standard Error (Sensitivity)	Specificity(%)	Standard Error (Specificity)
4	57.62	0.0366	66.60	0.0500	47.50	0.0629
5	57.17	0.0363	67.50	0.0439	49.83	0.1055
6	57.20	0.0412	70.93	0.0628	44.01	0.0630

precision/recall Table			
Fold	precision	Recall	AF1
4	57.37	66.60	0.6151
5	60.37	67.50	0.6118
6	57.23	70.93	0.6248

***Table 4.1.16** Summary of the results for the miRNA_Seq data set 2.*

Sensitivity/Specificity Table						
Fold	Accuracy(%)	Standard Error (acc)	Sensitivity (%)	Standard Error (Sensitivity)	Specificity(%)	Standard Error (Specificity)
4	63.60	0.0233	82.38	0.0315	35.95	0.0541
5	62.94	0.0375	83.47	0.0534	30.00	0.0670
6	63.96	0.0436	81.04	0.0578	35.99	0.0588

precision/recall Table			
Fold	precision	Recall	AF1
4	66.38	82.38	0.7313
5	64.49	83.47	0.7249
6	65.63	81.04	0.7215

4.1.2 Balanced class labels

In this section we evaluated the classifier performance in a balanced class labeled data. Here we used total four class labels with 424 samples each of which consists of 106 numbers of samples. The class prediction results of equal size of samples of different stages are listed below. In Table 4.1.2.1 for protein coding data set 1 (see Table 4.1) the accuracy, sensitivity, specificity and AF1 score are around the same values for all numbers of the cross fold.

***Table 4.1.2.1**: Summary of the results for the protein coding RNA_Seq data set1.*

Sensitivity/Specificity Table						
Fold	Accuracy(%)	Standard Error (acc)	Sensitivity (%)	Standard Error (Sensitivity)	Specificity(%)	Standard Error (Specificity)
4	56.42	0.0169	55.58	0.0315	57.53	0.01069
5	55.41	0.0221	56.65	0.0208	54.18	0.03561
6	55.55	0.0308	56.84	0.0475	53.38	0.01847

precision/recall Table			
Fold	precision	Recall	AF1
4	56.45	55.58	0.5587
5	55.46	56.65	0.5596
6	54.03	56.84	0.5539

The results for data set 2 are summarized in Table 4.1.2.2 in which the four fold cross validation gives a little bit better accuracy, sensitivity and higher AF1 scores than other cross fold numbers. Although we used the same features for the class prediction, the number of the training data does not have a significant effect on the classification results.

***Table 4.1.2.2**: Summary of the results for the protein coding RNA_Seq data set2.*

Sensitivity/Specificity Table						
Fold	Accu-racy(%)	Standard Error (acc)	Sensitivity (%)	Standard Error (Sensitivity)	Speci-ficity(%)	Standard Error (Speci-ficity)
4	59.06	0.0378	64.38	0.0599	53.80	0.0251
5	56.05	0.3898	62.11	0.0594	51.24	0.0394
6	56.75	0.0147	58.84	0.0349	55.78	0.04583

precision/recall Table			
Fold	precision	Recall	AF1
4	57.71	64.38	0.6075
5	56.27	62.11	0.5839
6	56.52	58.84	0.5690

From the tables (table 4.1.2.3 and table 4.1.2.4) it is clear that accuracy and AF1 of both data sets are around 60% while in the protein coding it was at a lower range. The non-coding data has a better performance as accuracy, sensitivity and specificity are at a similar range, i.e. around 60 % which is a little bit higher than for the protein coding RNA_seq data.

***Table 4.1.2.3**: Summary of the results for the non-coding RNA_Seq data set1.*

Sensitivity/Specificity Table						
Fold	Accuracy(%)	Standard Error (acc)	Sensitivity (%)	Standard Error (Sensitivity)	Specificity(%)	Standard Error (Specificity)
4	57.60	0.0240	57.35	0.0328	57.09	0.0659
5	61.08	0.0290	60.97	0.0598	60.20	0.0607
6	57.65	0.0222	60.52	0.0160	54.92	0.0394

precision/recall Table			
Fold	precision	Recall	AF1
4	58.13	57.35	0.5749
5	61.12	60.97	0.6046
6	53.11	60.52	0.5812

Table 4.1.2.4*: Summary of the results for the non-coding RNA_Seq data set2.*

Sensitivity/Specificity Table						
Fold	Accuracy(%)	Standard Error (acc)	Sensitivity (%)	Standard Error (Sensitivity)	Specificity(%)	Standard Error (Specificity)
4	55.32	0.02733	63.22	0.0326	47.09	0.0234
5	55.58	0.05205	56.94	0.0764	53.18	0.0497
6	54.90	0.02834	57.18	0.0461	53.52	0.0428

precision/recall Table			
Fold	precision	Recall	AF1
4	54.01	63.22	0.5796
5	54.85	56.94	0.5548
6	54.84	57.18	0.5533

4.1.3 Principle Component Analysis (PCA)

To reduce the high dimension of feature vector we applied the principle component analysis (PCA) to see the effect of the model on breast cancer data. The PCA was applied on the data set 1 (see Table 4.1) for the protein coding, non-coding RNA_Seq and the whole miRNA data. Only the first 20 PCA was taken as a feature with 5 fold cross validation for both balanced and unbalanced class label data. The performance result for the class prediction model is summarized in table 4.1.3. From the table below it can be stated that the principle component has a significant effect on accuracy, sensitivity and AF1 on protein coding for both balanced and unbalanced data sets. On the other hand we want to emphasize that it does not have a significant effect on non-coding and miRNA data sets.

Table 4.1.3*: Summary of the results of PCA for all the data set.*

x %	Balanced Levels			Unbalanced Levels		
	PCoding	Non-coding	miRNA	PCoding	Non-coding	miRNA
Accuracy	65.47	59.52	47.72	66.66	64.61	54.47
Sensitivity	70.04	71.42	70.00	74.45	73.91	79.42
Specificity	60.00	47.61	29.16	56.75	50.81	25.00
precision	65.95	57.42	45.16	68.62	63.21	56.54
Recall	70.04	71.42	70.00	74.46	73.91	79.42
F1 Score %	68.13	63.82	54.90	71.42	70.51	65.85

4.2 TNM Staging System sub-stage Classification

The following results depict the performance assessment of the tumor, the node and the metastasis system of staging in breast cancer classification. Unlike Surveillance, Epidemiology and End Results(SEER)

staging system, TNM staging classification is performed using multiple class classification. The performance measure is averaged per class basis. Likewise, SEER classification data set is divided into different sets with different combinations of stages. The best two sets (See table 4.2) are selected for the

assessment of the classifier. The prediction results for both the balanced and unbalanced class labeled data are presented below.

Table 4.2.1: *created data sets with combination of different stages of RNA_seq protein coding and non-coding data*

Data	Stage	Number of samples
Set 1	Stage IA, Stage IIA and Stage IIIA	672
Set 2	Stage IA, Stage IIA, Stage IIB and Stage IIIA	962

Table 4.2.2*: created data sets with combination of different stages of miRNA data*

Data	Stage	Number of samples
Set 1	Stage IIA, Stage IIB and Stage IIIA	241
Set 2	Stage IA, Stage IIA, Stage IIB and Stage IIIA	255

4.2.1 Unbalanced class

The total sample size is 962 for unbalanced data set. Class prediction results for protein coding RNA_Seq data are presented below. From the table (Table 4.2.1.1 and Table 4.2.1.1) it is evident that accuracy improves better in data set 1 than data set 2 but sensitivity and AF1 score do not show satisfactory results. The classifier cannot detect around 70% of the class samples perfectly on the other hand it can reject 66% of the samples that do not belongs to the target class. As an important finding of this study we have observed that by selecting classes including stage IIB we receive a large amount of false alarms.

***Table 4.2.1.1**: Summary of the results for the protein coding RNA_Seq data set 1.*

Sensitivity/Specificity Table						
Fold	Accuracy(%)	Standard Error (acc)	Sensitivity (%)	Standard Error (Sensitivity)	Specificity(%)	Standard Error (Specificity)
4	58.37	0.0145	33.24	0.0008	66.59	0.0006
5	58.30	0.0221	33.34	0.0000	66.66	0.0000
6	58.63	0.0110	33.34	0.0000	66.66	0.0000

precision/recall Table			
Fold	precision	Recall	AF1
4	19.56	33.24	0.2464
5	19.43	33.34	0.2452
6	19.54	33.34	0.2455

Similar performance to that of the protein coding RNA_Seq has been noticed for Non_coding RNA_Seq data. Table 4.2.1.3 and Table 4.2.1.4 show that accuracy, sensitivity and AF1 score remains the same even though the training data increases. In data set 2 specificity increases 10% more than in data set 1, while accuracy and sensitivity in data set 1 increases significantly more than in data set 2.

Table 4.2.1.2*: Summary of the results for the non-coding RNA_Seq data set2.*

Sensitivity/Specificity Table						
Fold	Accuracy(%)	Standard Error (acc)	Sensitivity (%)	Standard Error (Sensitivity)	Specificity(%)	Standard Error (Specificity)
4	42.77	0.022	27.31	0.0063	76.25	0.0034
5	43.15	0.014	27.30	0.0028	76.36	0.0019
6	42.18	0.016	26.73	0.0058	76.08	0.0030

precision/recall Table			
Fold	precision	Recall	AF1
4	19.82	27.31	0.2203
5	21,03	27,30	0.2373
6	20.05	26.73	0.2283

Table 4.2.1.3*: Summary of the results for the non-coding RNA_Seq data set1.*

Sensitivity/Specificity Table						
Fold	Accuracy(%)	Standard Error (acc)	Sensitivity (%)	Standard Error (Sensitivity)	Specificity(%)	Standard Error (Specificity)
4	58.52	0.0125	33.34	0.000	66.77	0.0010
5	58.46	0.0251	33.34	0.000	66.66	0.000
6	58.63	0.0110	33.34	0.000	66.66	0.000

precision/recall Table			
Fold	precision	Recall	AF1
4	19.56	33.34	0.2464
5	19.48	33.34	0.2455
6	19.54	33.34	0.2455

***Table 4.2.1.4**: Summary of the results for the non-coding RNA_Seq data set 2.*

Sensitivity/Specificity Table						
Fold	Accuracy(%)	Standard Error (acc)	Sensitivity (%)	Standard Error (Sensitivity)	Specificity(%)	Standard Error (Specificity)
4	42.36	0.0062	26.74	0.0036	76.02	0.0018
5	42.03	0.0200	26.82	0.0048	76.08	0.0028
6	41.98	0.0126	26.59	0.0063	75.90	0.0035

precision/recall Table			
Fold	precision	Recall	AF1
4	20.59	26.74	0.232
5	19.74	26.82	0.227
6	20.83	26.59	0.232

Micro RNA data set is divided into two different sets(see Table 4.2.2). The performance of the class prediction using data set 1 and data set 2 are listed in Table 4.2.1.5 and Table 4.2.1.5. It is evident from the table that the size of the training data effects on performance is being measured in miRNA data; for 5 fold cross validation accuracy increases 5% more than in the case of 4 and 6 fold cross validation.

Table 4.2.1.5: *Summary of the results for the miRNA_Seq data set 1.*

Sensitivity/Specificity Table						
Fold	Accuracy(%)	Standard Error (acc)	Sensitivity (%)	Standard Error (Sensitivity)	Specificity(%)	Standard Error (Specificity)
4	49.77	0.0221	37.80	0.0086	69.82	0.0074
5	54.10	0.0266	41.61	0.0173	71.64	0.0103
6	51.80	0.0351	39.71	0.0209	71.06	0.0152

precision/recall Table			
Fold	precision	Recall	AF1
4	39.61	37.80	0.3778
5	43.62	41.61	0.4189
6	35.67	39.71	0.3753

Table 4.2.1.6: *Summary of the results for the miRNA_Seq data set 2.*

Sensitivity/Specificity Table						
Fold	Accuracy(%)	Standard Error (acc)	Sensitivity (%)	Standard Error (Sensitivity)	Specificity(%)	Standard Error (Specificity)
4	45.51	0.0610	26.71	0.0182	76.39	0.0117
5	48.25	0.0307	28.74	0.0169	77.46	0.0083
6	44.32	0.0478	27.08	0.0195	76.24	0.0122

precision/recall Table			
Fold	precision	Recall	AF1
4	29.61	26.71	0.2729
5	24.07	28.74	0.2609
6	29.05	27.08	0.2763

4.2.2 Balanced class labels

In this section we will discuss about the effect of the classifier on equal size of class labels. Here only protein coding RNA_Seq and non-coding RNA_Seq will be discussed. From the protein coding and non-coding data set class labels are extracted in equal sizes. Each class contains 106 samples. Two different data sets are created as it can be seen from Table 4.2.1. The first data set contains three stages: Stage IA, Stage IIA and Stage IIIA with the total number of sample size of 318.

The second set of data contains all the stages with 424 samples (see Table 4.2.2). In the protein coding data set training size does not have effect on the performance measure unlike the unbalanced data set that was previously discussed. On the other hand the overall performance in the class prediction differs significantly between the two data sets.

The data set 1 has a better performance than the data set 2(see Table 4.2.21 and Table 4.2.2.1)

***Table 4.2.2.1**: Summary of the results for the protein coding RNA_Seq data set 1.*

Sensitivity/Specificity Table						
Fold	Accuracy(%)	Standard Error (acc)	Sensitivity (%)	Standard Error (Sensitivity)	Specificity(%)	Standard Error (Specificity)
4	38.55	0.0228	40.05	0.0152	69.94	0.0091
5	40.45	0.0230	40.93	0.0201	69.77	0.0096
6	38.69	0.0230	40.08	0.0201	69.77	0.0091

precision/recall Table			
Fold	precision	Recall	AF1
4	40.27	40.05	0.4015
5	40.27	40.93	0.4015
6	39.99	40.08	0.4001

From the tables(Table 4.2.2.1 and Table 4.2.2.2) it is obvious that the classifier cannot classify breast cancer sub-stages perfectly in the TNM staging system. Both data sets give poor performance.

***Table 4.2.2.2**: Summary of the results for the protein coding RNA_Seq data set 2.*

Sensitivity/Specificity Table						
Fold	Accuracy(%)	Standard Error (acc)	Sensitivity (%)	Standard Error (sensitivity)	Specificity(%)	Standard Error (Specificity)
4	24.81	0.0281	28.25	0.0276	75.88	0.0077
5	29.73	0.0225	30.32	0.0220	76.77	0.0066
6	29.16	0.0174	30.19	0.0214	76.78	0.0064

precision/recall Table			
Fold	precision	Recall	AF1
4	28.06	33.63	0.2936
5	31.62	30.32	0.3063
6	30.94	30.19	0.3039

The model was afterwards tested on non-coding RNA_Seq data. The classifier's performance is listed in the table (Table 4.2.2.3 and Table 4.2.24) with different cross fold numbers. From the tables (Table 4.2.2.3 and Table 4.2.24) the classifier's performance looks similar to that of the protein coding RNA (see Table 4.2.21 and Table 4.2.2.1).

Table 4.2.2.3*: Summary of the results for the non-coding RNA_Seq data set 1.*

Sensitivity/Specificity Table						
Fold	Accuracy(%)	Standard Error (acc)	Sensitivity (%)	Standard Error (Sensitivity)	Specificity(%)	Standard Error (Specificity)
4	38.59	0.0428	39.03	0.0373	69.53	0.0189
5	37.36	0.0134	37.11	0.0087	68.52	0.0043
6	35.30	0.0199	35.86	0.0174	68.18	0.0080

precision/recall Table			
Fold	precision	Recall	AF1
4	38.02	39.03	0.3850
5	37.62	37.11	0-3732
6	36.43	35.86	0.3613

***Table 4.2.2.4**: Summary of the results for the non-coding RNA_Seq data set 2.*

Sensitivity/Specificity Table						
Fold	Accuracy(%)	Standard Error (acc)	Sensitivity (%)	Standard Error (Sensitivity)	Specificity(%)	Standard Error (Specificity)
4	24.82	0.0045	25.58	0.0063	75.72	0.0014
5	26.64	0.0196	28.65	0.0217	76.09	0.0063
6	25.89	0.0255	27.55	0.0270	75.81	0.0078

precision/recall Table			
Fold	precision	Recall	AF1
4	24.50	25.58	0.2502
5	26.09	28.65	0.2724
6	27.88	27.55	0.2744

4.2.3 Principle Component Analysis (PCA)

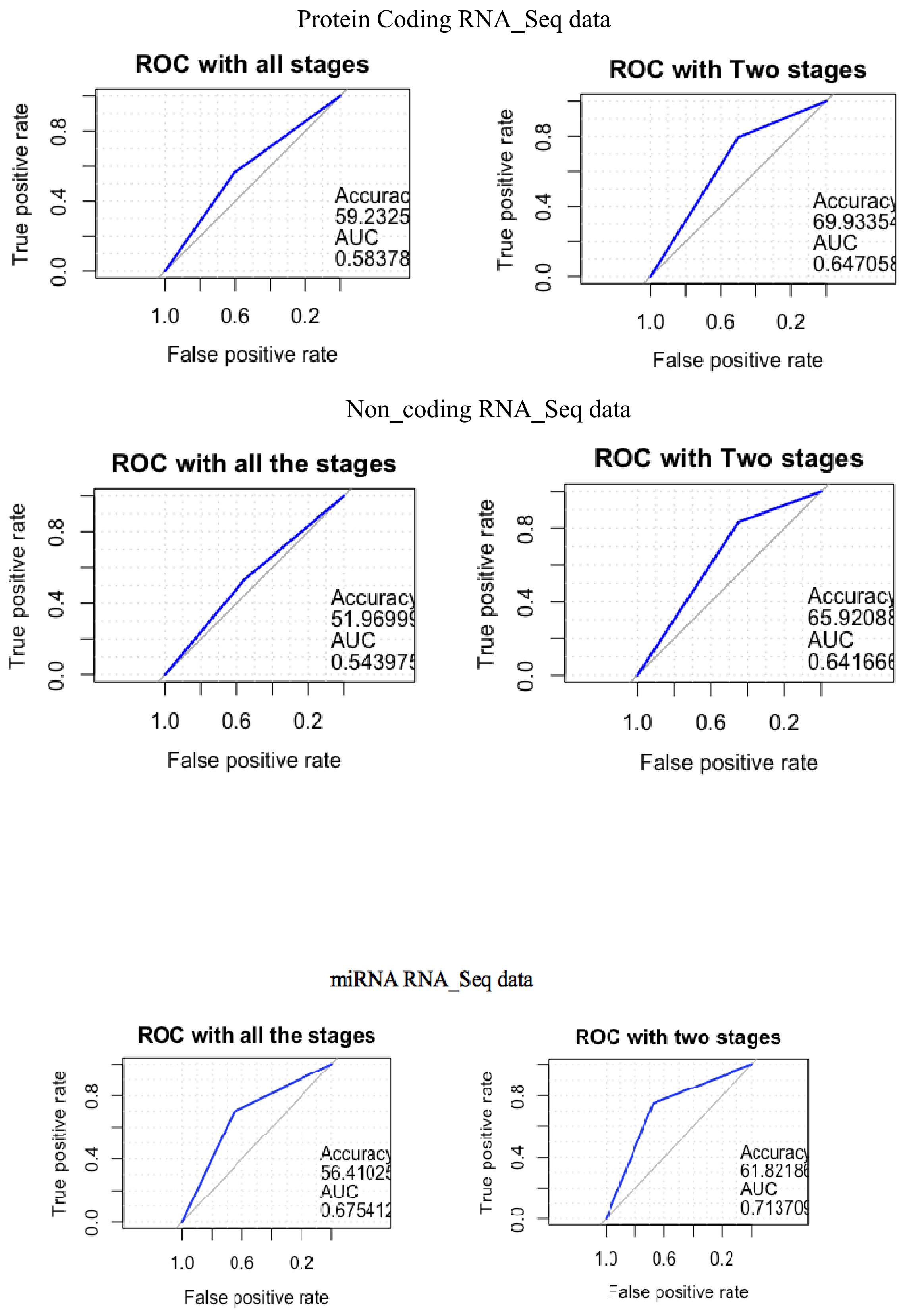
Application of principle components on feature vectors of both protein coding and non-coding RNA_seq in a balanced class label and PCA on protein coding RNA_Seq, non-coding RNA_Seq and miRNA in an unbalanced class label are listed in the Table 4.2.3 with 5 fold cross validation and the first 20 principle components. From the Table 4.2.3 it is quite clear that the principle component cannot increase the performance of the class prediction of the TNM staging system.

***Table 4.2.3**: Summary of the results of PCA for all the data set.*

	Balanced Levels		Unbalanced Levels		
	PCoding	Non-coding	PCoding	Non-coding	miRNA
Accuracy	26.40	28.08	41.97	41.59	46.36
Sensitivity	28.18	28.61	27.00	26.77	27.44
Specificity	76.05	76.27	76.07	76.07	76.65
precision	26.97	29.68	23.07	20.41	27.10
Recall	28.18	28.61	27.00	26.77	27.44
F1 Score %	0.275	0.290	0.245	0.231	0.264

4.3 Discussion on Results

Apart from creating different data sets with feature selection, the whole data set was used to classify the SEER staging system and TNM staging system with different cross fold.



Protein Coding RNA_Seq data

Non_coding RNA_Seq data

miRNA RNA_Seq data

***Figure 4.3.1** ROC of SEER sub-stage classification.*

The Figure 4.3.1 summarizes the performance of the classifier of the SEER sub-stage classification where Area Under the Curve (AUC) and accuracy are being considered as performance metrics. It is clear from the figure that taking into account only two stages (data set 1) gives a better performance than if we consider four stages (data set 2) for all the data sets.

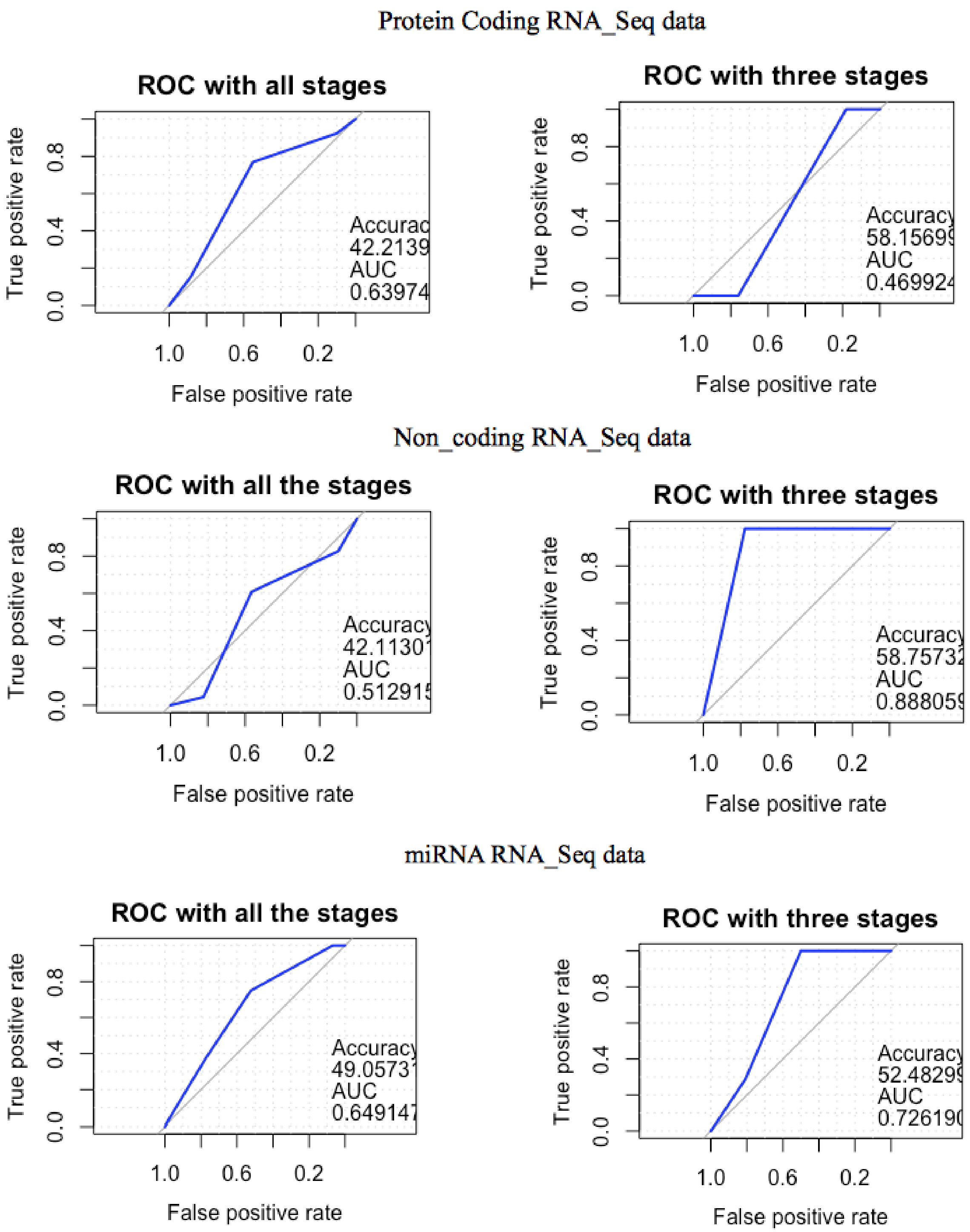


Figure 4.3.2 *ROC of TNM sub-stage classification.*

The classification results of the both protein coding and the non-coding RNA data sets show increase in accuracy and AUC in data set 2 as compared to data set 1(see Table 4.1). Both the protein coding and the non-coding RNA show poor AUC in the classifying data set 1. Here the classifier plays a better role in the miRNA sub-stage classification. In both data set 1 and data set 2 AUC is around 70% and accuracy is around 60%.

In TNM sub-stage classification system the classifier performance reduced significantly. Figure 4.3.2 shows the performance of the classifier for all the data sets(see Table 4.2.1 and Table 4.2.2). In the protein coding data, the accuracy is higher in the classification of the three different stages(data set 1) than in the classification of the four stages(data set 2) whereas AUC is lower. In the case of the non-coding RNA both AUC and accuracy show higher values in data set 1 than in data set 2. The sample size did not have signifiant effect on the performance of the protein coding and the non-coding data. In the case of the miRNA data classification, both accuracy and AUC improves from data set 2 to data set 1.

From the above discussion we see that if the number of the stages reduced then the classification performance improves despite the fact that the sample size reduced. Stage IIA and IIB have quite similar statistical distributions. This might be a reason why the classifier is more likely to give false predictions than true predictions. The results are being presented above only with the best features that we studied so far.

From the experiments we have performed the following conclusions can be drawn:

- ➢ The model cannot predict well the cancer sub-stage in the TNM staging system.
- ➢ The model has no effect on the sample size variation in the TNM system.
- ➢ The model can predict the cancer stages with satisfactory results in the SEER staging system using the protein coding RNA_Seq and the non-coding RNA_Seq data.
- ➢ The model can predict the stage of cancer with good result using the miRNA_Seq data.

- The model has no significant effect of principle component analysis on the miRNA in either of the staging systems.

- Further study on the miRNA by selecting the most significant genes may predict the cancer stages with a better result.

The reason is unknown why the Principle Component Analysis can give a better performance with the protein coding genes in the SEER staging system but has no signifiant effect on the non-coding genes(see section 4.1.3). With the help of the above investigation we have identified the 11 most significant miRNA genes that gives the classifier an output of more than 90% sensitivity, around 60% accuracy and an AUC of 0.72 in the SEER staging system (see Figure 4.3.1).

5. CONCLUSION

In this study we have investigated and compared the prediction performance of sub-stage classification of breast cancer in the SEER and the TNM staging system selecting from 10 to more than a thousand genes as a feature. We have identified that miRNA with the 11 most significant genes can offer a better prediction than the protein coding and non-coding genes. In the two data sets (see table 4.1) in unbalanced class labels and balanced class labels accuracy varies from 56% to 58% in data set 1 and 56% to 70% in dataset 2 for protein coding genes. On the other hand the non-coding genes have no variation in accuracy for balanced and unbalanced class labels. For the miRNA data we have applied only unbalanced class labels due to the small amount of available sample size in stage IA. A little variation in accuracy is present- from 57% to 62% for the 2 different data sets. Furthermore, we have also investigated the changes in the performance metrics using four stages. We have focused our attention on the best performing combination after checking the performance of all the possible combinations of two stages out of the four. The model behaves significantly different ways in the two different approaches. For the unbalanced class label accuracy varies 56% to 70% for the protein coding data while in the balanced class label it has no significant variation. In fact balancing the class labels does not give any significant improvements on the accuracy of the model for predicting the breast cancer sub-stage classification. However, the prediction model achieves 15% better accuracy by selecting only two stages out of the four stages. Moreover, miRNA performs more desirably with the accuracy of 62% and with an AUC of 0.72 which is a sign of a good classifier.

In addition, using the PCA model gives a better accuracy and sensitivity for protein coding and a little change in the non-coding genes but no significant accuracy gain on micro RNA genes. A few researchers have used PCA for micro array data to differentiate between tumor and non tumor genes [18]. The authors implement statistical hypothesis by selecting features using principal component analysis. In our study we selected features using variance of genes in different stages. Many researchers have predicted breast cancer on a molecular subtype basis [19,22].

Most of the subtype classification is performed by statistical analysis, while in our study we have used the machine learning method support vector machine (SVM) to predict breast cancer sub-stages in two different staging systems namely in SEER (Surveillance, epidemiology and END Results) and in TNM (Tumor, Node and Metastasis). The micro RNA is identified as an indicator of breast cancer progression and as a biomarker for sub-stage classification [21,22]. In our study we have also found that miRNA is the key gene for the breast cancer progression and sub-stage classification.

Feature selection was performed manually by checking beans of histogram of variance of all the genes and we fed the data to our model to predict sub-stages. After a lot of trial and error method we ended up with a certain range of beans of histogram of log of variances of genes that gives best performance with the SVM method. The feature dimension was high enough to give a perfect prediction. Then we applied the principle component analysis to reduce the feature dimension. Application of the PCA shows insignificant improvements in protein coding and non-coding data but no effect on miRNA. In fact PCA works on high dimensional data and they should be intercorrelated. In our study after the feature selection the miRNA feature vector was lower than the sample size. So by definition we cannot state that miRNA is high dimensional data. Protein coding and non-coding data has more than 7000 features and not all of them correlated. This may be one of the reasons why the principle component did not work on the miRNA.

Based on our work it can be concluded that for he TNM staging system merely the usage of the support vector machine is not sufficient enough but a more robust technique is being needed. SVM with PCA can be effectively used as a model providing possible further insights into the protein coding data analysis in SEER staging system in the future and SVM can be utilized in itself for the miRNA to predict class labels for the SEER staging system.

We used libsvm package for svm computation and prcomp library was used to compute pca; pROC, e1071 and caret for ROC, SVM and confusion matrix were used respectively. With corei5 and 8 GB RAM, the analysis of the three different data the computer took 0.5 min to 8 mins depending on the size of the data and the numbers of cross fold. This speed is pretty fast as we have not worked with any clusters.

REFERENCES

1. American Cancer Society. *Breast Cancer Facts & Figures 2013-2014*. Atlanta: American Cancer Society, Inc. 2013.

2. Sotiriou C, Neo S-Y, McShane LM, et al. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. Proceedings of the National Academy of Sciences of the United States of America. 2003;100(18):10393-10398. doi:10.1073/pnas.1732912100.

3. Joseph A. Cruz and David S. Wishart.Applications of Machine Learning in Cancer Prediction and PrognosisCancer Inform. 2006; 2: 59–77.

4. Simes RJ. Treatment selection for cancer patients: application of statistical decision theory to the treatment of advanced ovarian cancer. J Chronic Dis. 1985;38:171–86.

5. Maclin PS, Dempsey J, Brooks J, et al. Using neural networks to diagnose cancer. JMedSyst. 1991;15:11–9.

6. Cicchetti DV. Neural networks and diagnosis in the clinical laboratory: state of the art. Clin Chem. 1992;38:9–10.

7. Jerez-Aragonés JM1, Gómez-Ruiz JA, Ramos-Jiménez G, Muñoz-Pérez J, Alba-Conejo E.A combined neural network and decision trees model for prognosis of breast cancer relapse. Artif Intell Med. 2003 Jan;27(1):45-63.

8. Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining methods. Artif Intell Med. 2005;34:113–27.

9. Leenhouts HP. Radon-induced lung cancer in smokers and non-smokers: risk implications using a two-mutation carcinogenesis model. Radiat Environ Biophys. 1999;1999;38:57–71.

10. Bach PB, Kattan MW, Thornquist MD, et al. Variations in lung cancer risk among smokers. J Natl Cancer Inst. 2003;95:470–8.

11. Duffy MJ. Predictive markers in breast and other cancers: a review. Clin Chem. 2005;51:494–503.

12. Grumett S, Snow P, Kerr D. Neural networks in the prediction of survival in patients with colorectal cancer. Clin Colorectal Cancer. 2003;2:239–44

13. Colozza M, Cardoso F, Sotiriou C, et al. Bringing molecular prognosis and prediction to the clinic. Clin Breast Cancer. 2005;6:61–76

14. Duffy MJ. Biochemical markers in breast cancer: which ones are clinically useful? Clin Biochem. 2001;34:347–52.

15. Dumitrescu RG, Cotarla I. Understanding breast cancer risk--where do we stand in 2005? J Cell Mol Med. 2005;9:208–21.

16. Rajnish Kumar, Anju Sharma, and Rajesh Kumar Tiwari. Application of microarray in breast cancer: An overview J Pharm Bioallied Sci. 2012 Jan-Mar; 4(1): 21–26.

17. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey S, Rees CA, et al. Molecular portraits of human breast tumours. Nature. 2000;406:747–52.

18. Venet D, Dumont JE, Detours V (2011) Most Random Gene Expression Signatures Are Significantly Associated with Breast Cancer Outcome. PLoS Comput Biol 7(10): e1002240. doi:10.1371/journal.pcbi.1002240

19. Wang Y, Klijn JG, Zhang Y, Sieuwerts AM, Look MP, Yang F, Talantov D, Timmermans M, Meijer-van Gelder ME, Yu J, Jatkoe T, Berns EM, Atkins D, Foekens JA. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. The Lancet 2005; 365: 67 Available: http://www.ihes.fr/~zinovyev/princmanif2006/Wang_lancet_2005.pdf

20. McGuire, A., Brown, J. A. L., & Kerin, M. J. (2015). Metastatic breast cancer: the potential of miRNA for diagnosis and treatment monitoring. Cancer Metastasis Reviews, 34, 145–155. http://doi.org/10.1007/s10555-015-9551-7

21. Rask L, et al. Differential expression of miR-139, miR-486 and miR-21 in breast cancer patients sub-classified according to lymph node status. Cellular Oncology (Dordr) 2014;37:215. doi: 10.1007/s13402-014-0176-6.

22. Haibe-Kains, B., Desmedt, C., Loi, S., Culhane, A. C., Bontempi, G., Quackenbush, J., & Sotiriou, C. (2012). A Three-Gene Model to Robustly Identify Breast Cancer Molecular Subtypes. JNCI Journal of the National Cancer Institute, 104(4), 311–325. http://doi.org/10.1093/jnci/djr545

23. Wenna Guo, Qiang Wang, Yueping Zhan, Xijia Chen, Qi Yu, Jiawei Zhang, Yi Wang, Xin-jian Xu & Liucun Zhu.(2016).Transcriptome sequencing uncovers a three–long noncoding RNA signature in predicting breast cancer survival. Scientific Reports 6, Article number: 27931. doi:10.1038/srep27931

24. Frank Berger Corresponding address, Maximilian F. Reiser. Micro-RNAs as Potential New Molecular Biomarkers in Oncology: Have They Reached Relevance for the Clinical Imaging Sciences? Theranostics 2013; 3(12):943-952. doi:10.7150/thno.7445

25. Alton Etheridge, Clarissa P. C. Gomes, Rinaldo W. Pereira, David Galas and Kai Wang. The complexity, function, and applications of RNA in circulation.(2013)frontiers in Genetics. doi: 10.3389/fgene.2013.00115

26. Harvey Lodish, Arnold Berk, S Lawrence Zipursky, Paul Matsudaira, David Baltimore, and James Darnell., Molecular Cell Biology, 2000

27. NextSeq® Series RNA-Seq Solution, 2015 webwite https://www.illumina.com/content/dam/illumina-marketing/documents/products/appnotes/appnote-nextseq-rna-seq.pdf

28. Ann E. Loraine, Ivory Clabaugh Blakley, Sridharan Jagadeesan, Jeff Harper, Gad Miller, and Nurit Firon, Analysis and visualization of RNA-Seq expression data using RStudio, Bioconductor, and Integrated Genome Browser,

29. An Introduction to Next-Generation Sequencing Technology, website https://www.illumina.com/content/dam/illumina marketing/documents/products/illumina_sequencing_introduction.pdf

30. Simon Haykin, Neural Networks A comprehensive foundation, Pearson Education Inc. 2005

31. YU HEN HU and JENQ-NENG HWANG Handbook of Neural Network signal Processing, CRC press,2002

32. Boser, B.E., Guyon, I., & Vapnik, V.N., A training algorithm for optimal margin classifiers, 1992

33. Corinna Cortes, Vladimir Vapnik, Support-Vector Networks, 1995

34. Trevor Hastie, Robert Tibshirani and Jerome Friedman , The Element of Statistical Learning, 2009

35. Chih-Chung Chang and Chih-Jen Lin, LIBSVM: A Library for Support Vector Machines, 2013

36. The Cancer Genome Atlas, website http://tcga-data.nci.nih.gov/tcga/

37. Ontology Documentation, website http://www.geneontology.org/page/ontology-documentation

38. Marina Sokolova, Guy Lapalme "A systematic analysis of performance measures for classification tasks ", Information Processing and Management 45, Elsevier.

39. Breast Cancer: Stages, website http://www.cancer.net/cancer-types/breast-cancer/stages, Approved by the Cancer.Net Editorial Board, 02/2016

40. Abdi Hervé, Williams Lynne J. Principal component analysis. WIREs Comp Stat 2010, 2: 433-459. doi: 10.1002/wics.101

41. Pearson K. On lines and planes of closest fit to systems of points in space. *Philos Mag A* 1901, 6:559 – 572.

42. Grattan-Guinness I. The Rainbow of Mathematics. New York: Norton; 1997.

43. Hotelling H. Analysis of a complex of statistical variables into principal components. *J Educ Psychol* 1933, 25:417 – 441.

44. Christopher M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006

45. Christopher J.C. Burges, A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery 1998, doi:10.1023/A:1009715923555