DOKUZ EYLÜL UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

IDENTIFICATION OF BREAST CANCER SUB-TYPES BY USING MACHINE LEARNING TECHNIQUES

by

Yunus BURAKGAZİ

February, 2017

İZMİR

IDENTIFICATION OF BREAST CANCER SUB-TYPES BY USING MACHINE LEARNING TECHNIQUES

A Thesis Submitted to the
Graduate School of Natural and Applied Sciences of Dokuz Eylül University
In Partial Fulfillment of the Requirements for the Degree of Master of Sciences
in Computer Engineering

by

Yunus BURAKGAZİ

February, 2017
İZMİR

M.Sc THESIS EXAMINATION RESULT FORM

We have read the thesis entitled **"IDENTIFICATION OF BREAST CANCER SUB-TYPES BY USING MACHINE LEARNING TECHNIQUES"** completed by **YUNUS BURAKGAZİ** under supervision of **ASST. PROF. DR. ZERRİN IŞIK** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Asst. Prof. Dr. Zerrin Işık

Supervisor

Asst. Prof. Dr Derya BİRANT

(Jury Member)

Asst. Prof. Dr. Selma TEKİR

(Jury Member)

Prof.Dr. Emine İlknur CÖCEN

Director

Graduate School of Natural and Applied Sciences

ACKNOWLEDGEMENTS

I would like to thank my supervisor Asst. Prof. Dr. Zerrin Işık who allowed me to work in this project and I appreciate for her valuable ideas, support and guidance throughout this project.

Finally, I would like to offer my special thanks to my family and my fiancée Özge Akgün for their support, patience and help. It would not have been able to complete this thesis without their support and help.

This thesis is published as an article on the proceedings book of 1st International Mediterranean Science and Engineering Congress.

Yunus BURAKGAZİ

IDENTIFICATION OF BREAST CANCER SUB-TYPES BY USING MACHINE LEARNING TECHNIQUES

ABSTRACT

Recent years technological developments for DNA sequencing and publicly available patient databases reveal a new research field that proposes intelligent systems for biomedical domain. Such systems might help to predict future outcomes of patients. The main goal of this thesis is to predict breast cancer subtype of patients by using machine learning (ML) methods on RNA-sequencing data, which were extracted from the TCGA dataset. The significant genes were selected by applying fold change and t-test statistical methods. Support vector machine and Random forest models were trained with significant genes of 70% of the patient samples. The predictive performance of models were measured on the unseen test data. The overall performances of ML models varied between 86% to 98% of average accuracy.

We analyzed the best-predictive genes for each subtype to figure out which genes have more effect on the subtype classification of breast cancer. The relevant biological activities of these genes were found by applying a network-based analysis and gene enrichment analysis. The results revealed that some biological processes related with the cancer progression play a role for the classification of breast cancer subtypes.

Keywords: Breast cancer, machine learning techniques, gene expression, interaction network analysis, gene ontology

MAKİNE ÖĞRENMESİ TEKNİKLERİ KULLANARAK GÖĞÜS KANSERİ ALT TÜRLERİNİN TESPİT EDİLMESİ

ÖZ

Son yıllarda DNA dizilemesindeki teknolojik gelişmeler ve halka açık hasta veri tabanları biyomedikal alan için akıllı sistemler öneren yeni bir araştırma alanı ortaya çıkarmıştır. Bu tür sistemler hastaların gelecekteki sağlık durumlarını tahmin etmede yardımcı olabilirler. Bu tezin temel amacı, TCGA veri setinden çıkarılan RNA dizileme verileri üzerinde makine öğrenmesi yöntemlerini kullanarak göğüs kanseri hastaların alt tipini tayin etmektir. Önemli genler, kat değişikliği ve t-testi istatistiksel yöntemleri uygulanarak seçilmiştir. Destek vektör makinesi ve Rastgele orman modelleri, hasta numunelerinin %70'lık kısmından elde edilen genlerle eğitilmiştir. Modellerin tahminleme performansı, daha önce kullanılmayan test verileri üzerinde ölçülmüştür. Makine öğrenmesi modellerinin genel performansları ortalama olarak %86 ila %98 arasında değişmektedir.

Meme kanseri alt tip sınıflandırmasına hangi genlerin daha fazla etkisi olduğunu bulmak için, her bir alt tipi en iyi sınıflandıran genler analiz edilmiştir. Bu genlerin ilgili oldukları biyolojik aktiviteler, ağ tabanlı bir analiz ve gen zenginleştirme analizi uygulayarak bulunmuştur. Sonuçlar göğüs kanseri alt tiplerinin sınıflandırılmasında kanserin ilerlemesiyle ilgili bazı biyolojik süreçlerin rol oynadığını ortaya çıkarmıştır.

Anahtar kelimeler: Göğüs kanseri, makine öğrenmesi teknikleri, gen ifadesi, etkileşim ağı analizi, gen ontolojisi

CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

CHAPTER ONE
INTRODUCTION

1.1 Overview

Cancer is the disease of our age. There are more than 100 types of cancer known today and breast cancer is the most common type of cancer among women in many countries.

Early diagnosis and prognosis have huge effect on survival rates and prediction of recurrence for breast cancer. To select the best treatment, classification takes high priority. On traditional procedure, breast cancer classification regards to the tumor size, the stage of the tumor, histological grade and receptors status. However, prognosis can be different even similar clinical stage and pathological results. From another perspective, to explain complex genetic alterations and the biological events involved in cancer development and progression, histological appearance of the tumors is insufficient (Yersal & Barutca, 2014).

In the last two decades, DNA microarray technology has been improved by introducing genome wide sequencing. It allows the monitoring of expression levels in whole genome. By measuring the expression level of cancer-related genes, breast cancer can be analyzed more detailed and tests of prediction of recurrence outcome can become more accurate. Therefore, detailed biological classification can help to develop personalized treatment options for better survival rate and less toxicity.

1.2 Purpose

Recent years technological developments on RNA-sequencing, publicly available larger cancer databases, and machine learning (ML) methods led a new research field that develops intelligent systems for biomedical domain. These techniques can help to effectively predict future outcomes of a cancer patient. The application of ML methods could improve the accuracy of cancer susceptibility. The accuracy of cancer

prediction outcome has significantly improved by 15%-20% recently with the application of ML techniques (Cruz & Wishart, 2006).

The goal of our study is to classify breast cancer patients, extracted from the TCGA dataset, into four classes (Luminal A, Luminal B, HER2, Basal) by using RNA-sequencing data and ML methods. The novelty of our study is the biological analysis of signature genes that are commonly selected and provided higher accuracies in the discrimination of patient subtypes. We found out the relevant biological activities of these genes by applying network and gene enrichment analysis methods.

1.3 Organization of Thesis

This thesis includes five chapters and the rest of the thesis is organised as follows:

In Chapter 2, biological information about the terms that are mentioned in this thesis frequently and a summary of our literature review about the related works are presented.

In Chapter 3, we gave details about our dataset and its processing. Other details such as technical background of the processes, general information about the machine learning and feature selection methods used in our study, information about the tools used in evaluation of system such as confusion matrix, Monte Carlo cross validation, accuracy calculation are explained. Furthermore, in order to reveal biological functions of signature genes, we applied a protein-protein interaction based analysis. This network extraction is explained in detail in this chapter as well.

In Chapter 4, experimental results of this study presented.

Finally, in Chapter 5, the conclusion remarks and future works are given.

CHAPTER TWO
BIOLOGICAL BACKGROUND AND RELATED WORK

In normal human life, cells grow, divide to form new cells and die when old enough or useless for body in a controlled way. A tumor is abnormal mass of tissue which is caused by cells growth without purpose and stopping in out of control. Tumors can be categorized as benign, pre-malignant and malignant. Cancerous tumors are malignant. They have a potential to spread into or invade nearby tissues. Cancer cells may form new tumors on different places of the body by using blood vessels or lymph system. Normally, the immune system fights against abnormal cells and remove them, but cancer cells are able to hide from the immune system. Possible signs and symptoms include: a new lump, skin changes, abnormal bleeding, weight loss, coughing or chest pain, unexplained weight loss, and a change in bowel movements. While these symptoms may indicate cancer, they may also occur due to other issues. There are over 100 different known cancers that affect humans.

2.1 Breast Cancer

Cancer that develops from breast tissue is called breast cancer. Either men or women can get breast cancer. It affects one out of every eight women during their lives. On American women, breast cancer comes after lung cancer on the highest death rates, also comes after skin cancer on the most commonly diagnosed cancer among all cancer types. There are several high-risk factors for getting breast cancer:

- Gender (being a woman) and age (being old)
- Gene mutations (BRCA1 and BRCA2)
- Beginning periods on early age or going through menopause on late age
- Having a family history who has been diagnosed with breast cancer

Other risks include not having children or having first child after age 30, being overweight, long-term taking birth control pills, having dense breasts, continuous exposure to electromagnetic fields and radiation, hormone therapy after menopause

etc. Symptoms of breast cancer vary from person to person actually it may not cause any symptoms. Generally, the first sign is a new lump or mass in the breast. Here are some unusual changes that can be symptoms of breast cancer:

- Changes in the size or shape of the breast
- Breast pain
- Bloody or normal flow from the nipple not like breast milk
- A lump in the underarm area
- Swelling on the breast
- Forms such as redness, bruising, scarring, vasodilation, inward depression, common small swellings, orange peel appearance in breast skin

2.1.1 Breast Cancer Subtypes

Classification of breast cancer by applying new molecular techniques benefit more accurate tests for the prediction of recurrence, developing new therapies and personalized treatment. There are four major molecular subtypes of breast cancer that researchers focused on:

- **Luminal-A:** Most common subtype, 50-60% of all. Since Luminal A is under group of luminal, it is high expression of hormone receptors (estrogen-receptor and/or progesterone-receptor positive). Additionally, it is HER2 receptor-negative. Voduc et al. (2010) claims that Luminal A tumors tend to have the best prognosis among others. It has high survival rate and low relapse rate. Treatment of these tumors often based on hormone therapy, sometimes response to chemotherapy.

- **Luminal-B:** Comprises 10-20% of breast cancers. Like Luminal A, Luminal B is also hormone-receptor positive. But Luminal B can be either HER2 negative or HER2 positive. Compared to Luminal A tumors, Luminal B tumors have higher histological grade, higher recurrence rate and lower survival rate, thus worse prognosis. Its response to chemotherapy is higher than Luminal A.

- **HER2:** HER2 gene products HER2 proteins. Healthy breast cells division, growth and repairing are controlled by HER2 proteins. Protein overexpression and gene amplification of HER2 causes breast cells grow out of control. HER2 type stands for HER2-positive and accounts for 5-15% of breast cancers. HER2-positive tumors are both estrogen-receptor and progesterone-receptor negative, generally poor tumor grade and poor prognosis. HER2-positive breast cancers fairly have high recurrence rate and spread. There are targeted drug options for HER2-positive breast cancers treatment such as Herceptin, Kadycla, Perjeta and Tykerb.

- **Basal:** ER-negative, PR-negative and HER2-negative are triple-negative breast cancers. Kreike et al. (2007) defines triple-negative breast cancers as subtype of basal, they can be surrogate for clinical stage. Women having BRCA1 gene mutations tend to be basal breast cancers. About 15-20% of breast cancers are triple negative/basal. Since lack of ER/PR receptors triple-negative/basal tumors do not respond to hormone therapy. Targeted therapies except Herceptin, Kadycla or other medicines can be useful for treatment options.

2.2 The Cancer Genome Atlas (TCGA)

The Cancer Genome Atlas (TCGA), collaboration between the National Cancer Institute (NCI) and National Human Genome Research Institute (NHGRI), has generated comprehensive, multi-dimensional maps of the key genomic changes in 33 types of cancer (Tomczak, Czerwinska & Wiznerowicz, 2015). The TCGA dataset contains 2.5 petabytes of data describing tumor tissue and matched normal tissues from more than 11,000 patients. It is publically available and has been widely used in many research projects.

2.3 Gene Expression Profiling

In molecular biology, gene expression profiling (GEP) is a technique to measure the activities of thousands of genes related with a specific disease or biological process simultaneously.

Gene expression starts with transcription. Transcription refers to particular segment of DNA copying into RNA molecule, which is called messenger RNA (mRNA). In translation process with mRNA leading, gene product is synthesized. The synthesized product is not always (sometimes functional RNA, small nuclear RNA or transfer RNA) but usually proteins. Robertson (2014), expresses that a person's susceptibility to cancer can be easily determined by measuring the expression level of cancer causing genes (oncogenes) or tumor suppressor genes in a cell or tissue. Strong association with specific patterns in the cancer tumors gene activity helps accurate tumor classification or accurate predictions of recurrence. Useful for predictions of response to therapy with clinical outcome and a must for targeted therapy. It is a challenging technology for pharmaceutical companies.

There are different technologies for measuring gene expression level such as DNA microarray, massively parallel signature sequencing (MPSS), RNA sequencing (RNA-seq) and serial analysis of gene expression (SAGE). RNA-seq technology has advantages over others. It can work with any species and has bigger dynamic range beside microarray. It consists of whole transcript sequences, but MPSS and SAGE short parts of transcripts. In our study, our dataset includes RNA-seq data.

2.4 RNA-Seq Raw Data Generation

For building and maintaining of a cell, DNA is essential for the reason of containing the necessary instructions. These instructions are organized into genes. DNA must be copied into RNA to carry out these instructions. First DNA must be read and transcribed into RNA. These gene readouts are called transcripts. Transcriptome is the complete set of the transcripts under specific circumstances or in a specific cell (Young, Gordon & Voigt, 2016). Transcriptome profiling is practicable using next generation sequencing by applying high throughput methods. By comparison of transcriptomes, we can determine the genes, which are expressed in distinct cell populations differentially. And this yields personalized treatment options, high survival rates against disasters.

Next-generation sequencing platforms enable sequencing very rapidly in high accuracies. This makes RNA-sequencing technology a cheaper way, for comprehensive analysis of transcriptomes. SOLiD, Sanger, Roche 454 genome analyzer, Illumina Genome Analyzer, Helios, Pacific Biosciences and IonTorrent use different high throughput methods at lower cost. High throughput refers to parallelizing the sequence process, millions of sequences concurrently.

As seen on Figure 3.1, in a typical RNA-seq experiment, the first action to take is choosing the suitable samples (M. O. Griffith, Walker, Spies & Ainscough, 2015). In transcriptome analysis cDNA will be needed. For generating cDNA, RNA transcripts should be isolated and purified. The main concern on conversion of RNA to cDNA is cutting off Poly(A) tails of RNA transcripts. Next step is construction of fragment libraries. Sequencing adapters should be added to cDNA. Based on size selection, cDNA will be randomly fragmented. Then these cDNAs should be sequenced in chosen sequencing platform. Sequencing platform will produce hundreds of millions of short paired-end reads.

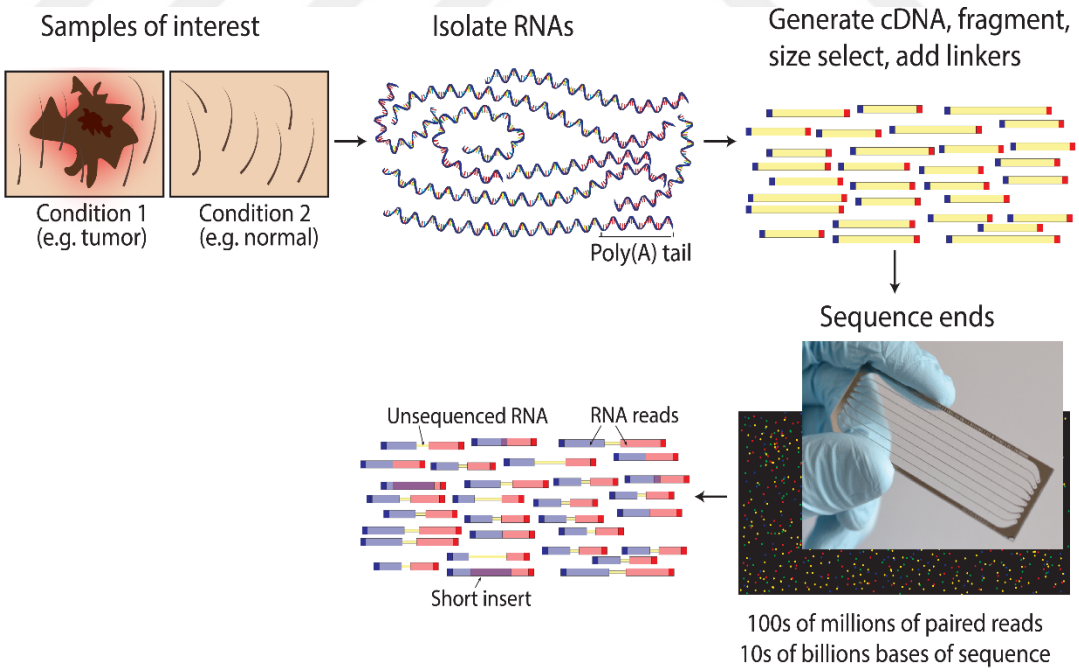

Figure 2.1 RNA-seq laboratory flowchart (M. O. Griffith, Walker, Spies & Ainscough, 2015).

2.5 Analysis of RNA-Seq Raw Data

Even there are many effective algorithms and bioinformatics tools, analysis of raw RNA-seq data is still open for development. Data analysis RNA-seq involves five steps:

- Quality check and preprocessing of raw reads
- Read mapping
- Read counting and gene quantification
- Data normalization and differential analysis
- Pathway enrichment analysis
- Visualization

2.5.1 Quality Check and Preprocessing of Raw Reads

Zhao et al. (2016) express that errors based on library preparation or sequencing steps or untrimmed adapter sequences may cause poor-quality data reads. And this will lead to low accuracies on data analysis. The reference genome data is stored in a format named FastA (Figure 3.2). Raw read data is generated by sequencer platform in a format named FastQ (Figure 3.3) which is sum of FastA and quality of reads. Via available tools such as PRINSEQ, FASTQC quality control can be checked in FastQ files and reads with low quality bases can be removed.

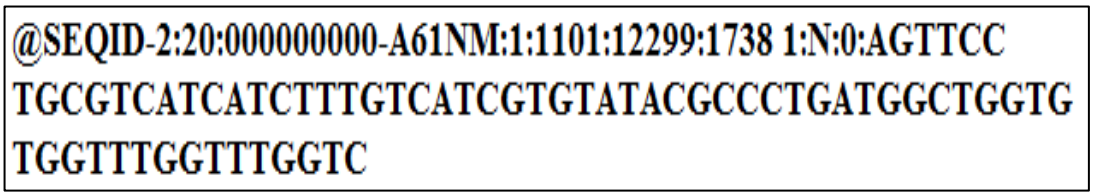

Figure 2.2 An example of FastA format.

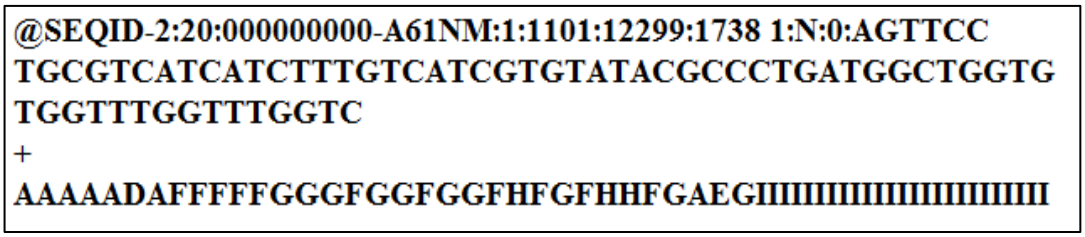

Figure 2.3 An example of FastQ format.

Tools like Cutadapt, Trimmomatic can be used for performing trim adapters, or other contaminating sequences.

Bioconductor provides tools and R packages (Section 2.6) for the analysis of genomic data. Quality control and preprocessing of raw reads can be done via Bioconductor package “systemPipeRdata”.

2.5.2 Read Mapping

Short sequence reads must be aligned with respect to reference transcriptome a genome assembly to find out their correct locations. There are many algorithms have been developed such as TopHat2 (Kim et al., 2013), STAR (Dobin & Gingeras, 2015), GSNAP (Wu & Nacu, 2010), OSA (Hu, Ge, Newman & Liu, 2012) and MapSplice (Wang et al., 2010).

There are different packages available on R-Bioconductor for alignment process, e.g., “TopHat2”, “Rsubread”.

2.5.3 Read Counting and Gene Quantification

To simply say, it is counting reads per feature/gene. There are two approaches for counting reads, one is transcript-based approach and the other one is union-exon based approach. Researchers indicate that for the reason of a gene expressed in one more transcript isoforms, transcript-based approach is more meaningful (Zhao, Xi et al., 2015). But it is also more difficult because of different isoforms of the gene

typically having a high proportion of genomic overlap. Different kinds of algorithms that can be performed such as RSEM (Li & Dewey, 2011), Cufflinks (Trapnell et al., 2011), ISOem (Nicolae, Mangul, Măndoiu & Zelikovsky, 2011), featureCounts (Liao, Smyth & Shi, 2014) and HTSeq (Anders, Pyl & Huber, 2014). "GenomicRanges" package is easy to use in R for these processes.

2.5.4 Data Normalization and Differential Analysis

Normalization is critical step to inference accurate gene expression after calculating read counts. Total number of aligned/mapped reads (library size) varies between different samples, for that reason we can not compare the different samples directly. We need to normalize library size. Scaling total number of read counts by the mean library size and then taking the $\log_2$, we can approximate library size. Therefore, genes of a sample having high differences compared to a sample evenly distributed will have lower expression and differentially expressed genes fail count will decrease. There are existing algorithms for identification of differentially expressed genes but there is still no optimal solution accepted among them.

"DESeq", "NOIseq", "NBPSeq", "rnaseqGene" are available packages in R for differential analysis.

2.5.5 Pathway Enrichment Analysis

By applying pathway enrichment analysis on differentially expressed genes, we can get more details about their biological activities. Annotation databases such as DAVID system (Huang, Sherman & Lempicki, 2009), Gene Ontology (Gene Ontology Consortium, 2004) and Kyoto Encyclopedia of Genes and Genomes pathways (Kanehisa, Goto, Kawashima, Okuno & Hattori, 2004) are available tools for functional enrichment analyses.

"topGO" package can be used in R for gene set enrichment analysis.

2.5.6 Visualization

RNA-seq data is very large, complex and abstract. The analysis of RNA-seq data is still open to development. There are available tools providing graphical user interface helping to visualize the dataset, plot the statistical results. Integrative Genomics Viewer is a high-performance visualization tool.

There are different packages on R for visualization such as “ggplot2”, “grammar of graphics”, “rgl”.

2.6 Literature Review

There are many research projects on breast cancer classification.

Kim et al. (2012) used Support Vector Machines (SVM) for training over 679 patients clinical, pathologic data; the method was validated via hold-out method and obtained 89% accuracy.

Researchers similarly trained a SVM over 547 patients, they obtained 95% average accuracy by applying 10-fold cross validation (Ahmad, Eshlaghy, Poorebrahimi, Ebrahimi, & Razavi, 2013).

Listgarten et al. (2004) used again SVM for training SNPs data of 174 patients, applied of 20-fold cross validation and got 69% accuracy.

In their study, researchers trained a Bayesian Network for 97 patients, 85% accuracy was reported by using hold-out method (Gevaert, De Smet, Timmerman, Moreau & De Moor, 2006).

TCGA (Cancer genome atlas network, 2012) investigated breast cancer subtypes by incorporating information from multiple platforms, i.e., genomic DNA copy number arrays, DNA methylation, exome sequencing, mRNA arrays, miRNA

sequencing and reverse-phase protein arrays. By classifying tumors using each individual platform and comparing results at different levels, they conclude that diverse genetic and epigenetic alterations converge phenotypically into four major breast tumor subgroups (i.e., luminal A, luminal B, HER2 positive, triple negative) using mRNA profiling. There exits research on TCGA dataset to predict patients' subtypes.

List et al. (2014) used gene expression and DNA methylation data from TCGA. Along with these data, TCGA provided a subtype classification of all gene expressions sample via the gold standard PAM50. Using a 543 patients data they created four models which include gene expression, DNA methylation, gene expression and DNA methylation combined (DNA methylation columns added to gene expression data) and PAM50 gene expression. They ran Random Forest on each model and after validation with 0.632 bootstrap error, the gene expression model performed best with a very low bootstrap error of less than 10% and an AUC of close to 100%, which was also the case for the combined model and the control model. The methylation model performed slightly worse, achieving a bootstrap error of 20% and an AUC of 88%.

Another publication using TCGA data proposed a subgroup-specific-gene-centering method to perform molecular subtyping on a study cohort that has a skewed distribution of clinicopathological characteristics relative to the training cohort (Zhao, Rødland et al., 2015). On such a study cohort, they center each gene on a specified percentile, where the percentile is determined from a subgroup of the training cohort with clinicopathological characteristics similar to the study cohort. They demonstrated their method using the PAM50 classifier and its associated University of North Carolina training cohort. On that training cohort, triple-negative cohort was subset of TCGA breast data. There were samples of 77 patients. Compared to the standard gene centering, subgroup-specific's overall prediction accuracy range was 17% to 33% across the five intrinsic subtypes. On the UNC triple-negative subgroup their method produced 11% (1/9) error rate on basal-like tumor classification, which standard-gene centering had a performance 56% (5/9).

Ali et al. (2014) have developed an expression-based method for the classification of breast tumors into the IntClust subtypes. They applied the method on public datasets of breast tumor transcriptomes to investigate the validity of IntClust. One of the dataset they used was containing gene expression data of 475 patients based on either RNA-seq or microarray from TCGA. Cross-tabulations of subtype assignment with Kappa-agreement statistics, by data type (RNA-seq or microarray) performed for each of the three classifiers (SCMGENE, PAM50 and IntClust). The agreement between classifiers was 93.1% for SCMGENE, 93.7% for PAM50 and 81.3% for IntClust.

CHAPTER THREE
DATA PROCESSING AND METHODS

3.1 Dataset

Dataset used in this thesis includes RNA-seq gene expression data of 418 breast cancer patients and 3 healthy people obtained from TCGA. Genes are coded by Entrez identifiers that allow us to get more information about a specific gene in the NCBI website (http://www.ncbi.nlm.nih.gov/gene). Patients' identities are hidden. We have the information of breast cancer subtypes (Luminal A, Luminal B, Her2, Basal) of each patient. In the data processing, we removed the genes whose expression could not be measured in the sequencing experiment. Those values might affect the performance of machine learning techniques negatively. Then the remaining gene count decreased from 20531 to 13259. Table 3.1 shows how looks like the last version of the dataset. Columns except gene symbol and entrezid are encoded names of the patients which are totally 421. And rows are gene expression value of these patients. Genes having negative value are downregulated, positive ones are upregulated. Downregulation means that the related gene product, protein, for the patient is decreased. Vice versa upregulation means protein increased. We used 70% of patient samples for training and 30% for testing.

Table 3.1 A snapshot of dataset extracted from TCGA.

GeneSymbol	EntrezID	TCGA.A1.A0SK.01A.12R.A084.07	TCGA.A1.A0SO.01A.22R.A084.07	TCGA.A2.A04P.01A.31R.A034.07
A1BG	1	6,43702E+14	8,19519E+14	6,76487E+14
GGACT	87769	8,65885E+14	6,67188E+14	6,47007E+14
A2M	2	1,20329E+14	1,19992E+14	1,31823E+14
A4GALT	53947	6,31715E+14	5,21095E+14	7,2328E+14
AAAS	8086	9,98745E+14	9,16413E+14	9,6676E+14
AACS	65985	1,12418E+14	8,98559E+13	9,82025E+13
AADAT	51166	8,73396E+14	9,64868E+14	6,22282E+14
AAGAB	79719	1,09364E+14	1,043E+14	9,15356E+13
AAK1	22848	8,99771E+14	9,40236E+14	9,4366E+14
AAMP	14	1,06036E+14	1,09862E+13	1,13824E+14
AARS	16	1,14801E+14	1,13731E+14	1,12736E+14
AARS2	57505	8,4556E+14	9,92146E+14	9,61392E+14
AARSD1	80755	8,72109E+14	8,51912E+13	9,30319E+14
AASDH	132949	8,08067E+14	8,74228E+14	7,6786E+14
AASDHPPT	60496	1,0649E+14	1,1339E+14	9,17572E+13
AASS	10157	5,63793E+14	7,8527E+14	7,64908E+14
AATF	26574	1,07435E+14	1,03759E+14	9,74377E+14
AATK	9625	5,90401E+14	7,13881E+13	5,62412E+14
ABAT	18	5,85771E+13	7,3797E+14	6,07526E+14
ABCA1	19	9,70375E+13	8,30749E+14	8,08632E+14
ABCA11P	79963	6,58946E+13	5,68133E+13	7,03263E+14

3.2 Feature Selection Methods

Feature selection, also named variable selection or attribute selection is a technique to select subset features that are most useful on predictive models. Features, columns in a tabular data, that are irrelevant or redundant may not contribute to the accuracy of predictive model but also may decrease the accuracy performance. There are three benefits of feature selection method. First is improving the predictive performance of selection models. Second is making more understandable of the process of generated data. Third is providing faster predictors for train sets (Guyon & Elisseeff, 2003). Feature selection should be performed on after model selection and before model is trained.

Feature selection methods are applicable on large datasets in supervised or unsupervised machine learning. In the fields of bioinformatics, feature selection technique has become a prerequisite to apply on large datasets for model building.

In our study, we applied two selection features to our predictive model, principle component analysis and t-Test & Fold Change separately. In fact, feature selection methods do not mean dimension reduction like principal component analysis does. But it is accepted as a means of feature selection.

3.2.1 Principal Component Analysis

Principal Component Analysis (PCA) is an algorithm used to reduce the dimensionality of the data by creating new variables, principal components, which are linear combination of existing features to explain maximum variance.

PCA is an orthogonal linear transformation (Jolliffe, 2002). It transforms data to a new coordinate system in which the new coordinates of the system are called principal components. The center of data points is the origin of new coordinate system. The first principle component points the direction of highest variance. The second points the direction of second highest variance and so on. Even most of the time dimensionality reduction is meaning loss of information, it is possible to just losing only a commensurately small amount of information by using only a few components.

Processing steps of PCA are:

1. Standardize the data
2. Subtract the mean
3. Calculate the covariance matrix
4. Calculate the eigenvalues and eigenvectors of the covariance matrix
5. Choose components and form a feature vector
6. Derive the new dataset

PCA has been applied in many fields such as taxonomy, biology, pharmacy, finance, agriculture, ecology, health and architecture (Sanguansat, 2012).

We used the "FactoMineR" R package to apply PCA. On each fold, we run PCA function on train data. Dimdesc function is performed after PCA function. For dimdesc function parameters, we set the significance threshold to 0.05, which means 95% confidence and we set correlation coefficients threshold big or equal to 0.7. With these options, the genes existing on the first five principal components are selected. Before applying PCA, train set had 13259 dimensions(genes). Even its count changes according to the train data, we were able to reduce the dimensions. On the first principal component we obtained most of the genes ~158. The genes count on the five principal component varied between 165 and 328 on different folds of cross validation.

3.2.2 Fold Change & t-Test

Fold change (FC) is the quantity change ratio between two values, shortly final value divided by initial value. Fold change is commonly used in bioinformatics mostly for gene expression change measurement but the ratio is calculated in $\log_2$. Consider there are 100 reads count in a control and 200 reads count in a treatment for a gene. The FC value of that gene becomes 2 FC. If it is bigger than 0, then this gene shows an over-expression in the treatment case. But when it is other way round, reads count in a control is higher than treatment, FC value always will be negative. FC value is calculated as:

$$\text{FC} = \log 2 \frac{\text{Patient Sample}}{\text{Healthy People Sample}} \tag{3.1}$$

t-Test is used to compare two sets of data samples whether they significantly differ from each other or not. t-Test makes this comparison by the means of two groups. t-Test tries to explain the mean of two groups statistically different or not.

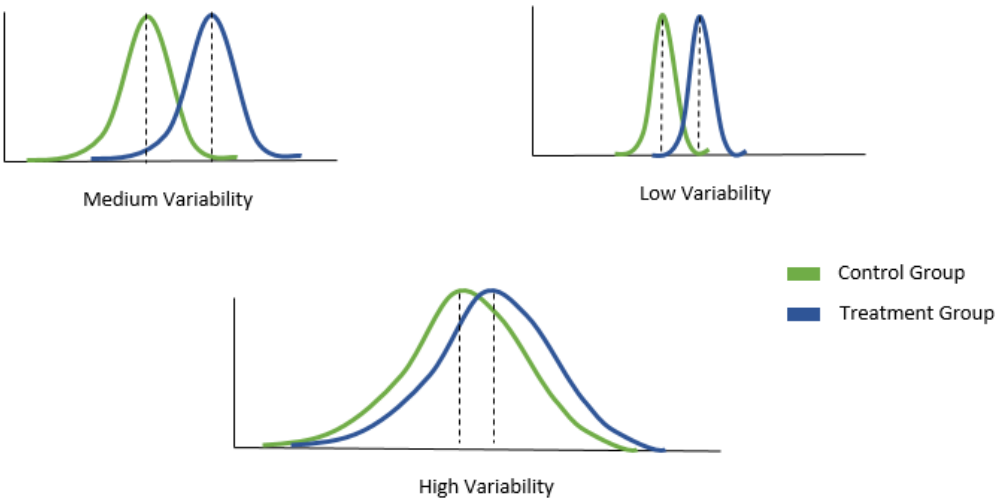

Figure 3.1 Control and treatment groups distributions in different scenarios.

As seen on Figure 3.1, the difference between the means are the same in all but their variation distributed differently. In high variability case, the groups difference is not much as others due to distribution overlap. t-Test considers the difference between the two groups means relative to their variance distribution (3.2).

$$t = \text{difference between groups / variability of groups} \tag{3.2}$$

We can tell that variability of groups is equal to standard error of the difference (Trochim, 2006). Standard error is calculated taking variance of each group and dividing it to the group's items count. Then we sum these values and take square root of this value (Equation 3.3).

$$t = \frac{\bar{X}_T - \bar{X}_C}{\sqrt{\frac{var_T}{n_T} + \frac{var_C}{n_C}}} \tag{3.3}$$

The result will be negative if the first group is less than the second group. The general concern is to define a confidence threshold for the value of the ratio which is 0.05. It means with confidence 95%, the groups significantly differ from each other.

We performed Fold Change and t-Test together as a feature selection method. On each fold, 70% of patient samples are randomly selected as a train set. We used "genefilter" R package and ran "rowTTest" function to apply t-Test. After selection of the train set, iteratively we applied t-Test to every patient on the train set. We created a matrix which consists of three healthy people and one of the patient's gene expression data. We passed two parameters to "rowTTest" function. One of them was matrix data itself and the second one was class labels. Healthy people are marked as "Class1" and the patient is labelled as "Class2" in that class labels. Since column count of the matrix and labels count are equal, we can perform t-Test on row level. Thus we were able to measure every gene expression change for each patient. On the results of "rowTTest" function, if the p-value (i.e., t-test outcome) is equal or smaller than 0.01, this gene was chosen as a significant one. On the same iteration, we applied FC by using "gtools" R package as well. We provided the mean values of healthy people samples and the cancer patients' samples to the FC calculation. The genes, whose absolute FC value is equal or bigger than 2, were marked as significant ones. We chose the genes which satisfy both criteria at the end of the iteration. Those genes were called as significant genes of the experiment. We reduced the dimension (i.e., total number of genes) of the train set by choosing only significant genes. In this way, it will be faster and more accurate to train machine learning methods.

3.3 Machine Learning Methods

By performing feature selection methods, we selected significant genes. Now we can train our data with machine learning methods to build our predictive model. Then we will test our model on unseen data and measure its accuracy performance.

3.3.1 Support Vector Machine (SVM)

SVM is a machine learning method that can be used in both classification and regression. Supervised learning refers to labelled data. Otherwise data without class labels requires unsupervised learning approaches such as support vector clustering.

A SVM model differentiates two classes. SVM tries to find an optimal hyper-plane to segregate the classes by using support vectors. The distance between classes is important to minimize the classification errors. According to the hard-margin or soft-margin concern, SVM chooses the hyper-plane with the convenient margin among the infinite separating hyper-planes.

For linearly separable training data, hard-margin SVM chooses the maximum-margin hyper-plane to segregate the classes. Hard-margin allows zero error, but it is open to over-fitting problems. On the other hand, it can be a better option to allow errors in the training set to have a more generalized model for working with new datasets. Generally working with non-linearly separable data, soft-margin SVM allows classification errors with a loss function. The "C" parameter is this loss function's multiplier. Users can increase error tolerance by setting C parameter to a higher value. Figure 3.2 illustrates linearly separable objects and two hyper-planes. Hyper-plane B is suitable for hard-margin SVM since there won't be any classification errors. When soft-margin SVM applied to training set, hyper-plane A will be selected. In that way, some misclassifications errors allowed but a more generalized model is selected.

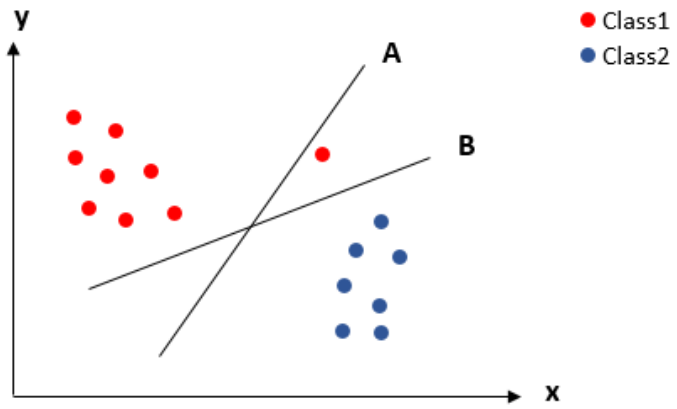

Figure 3.2 An example of linearly separable data with hyper-plane examples.

Most of real-world problems cannot be separated by a simple hyper-plane. One solution is to map the original data into a higher dimension and define a separating hyper-plane in the new space. This higher-dimensional space is called as feature space. If feature space has sufficient dimensionality, any kind of data sample can be separable with hyper-planes. However, separation of the data in such a hyper-space might lead over-fitting of the training data (Brown et al., 2000). SVMs nicely overcome these problems (Vapnik, 1998). They prevent over-fitting by finding a hyper-plane that separates classes. The decision function for classifying samples only requires dot products between feature vectors of samples. SVM finds the location of the optimal hyper-plane without explicitly representing the space, instead a kernel function performs a dot product in the feature space. Assume that SVM applies a kernel function to data points given in Figure 3.3. In the new coordinate system, data belongs to *class1* would be leading on higher value of *z* and data belongs to *class2* would be leading on lower value of *z* (Figure 3.4). So that, it would be easier to find a separating hyper-plane.

There are different kinds of methods for using kernels such as linear, polynomial, radial basis functions and sigmoid.

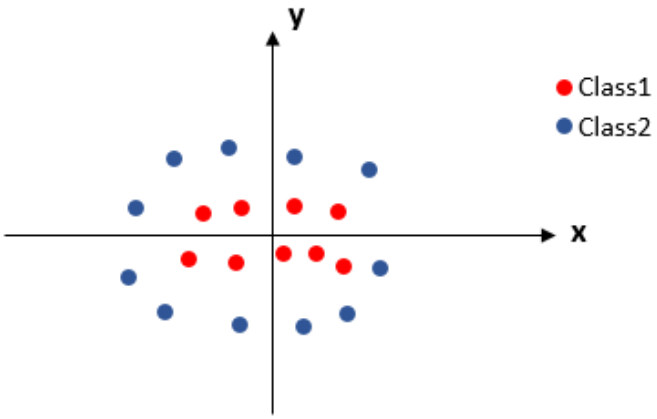

Figure 3.3 Input space of non-linearly separable data.

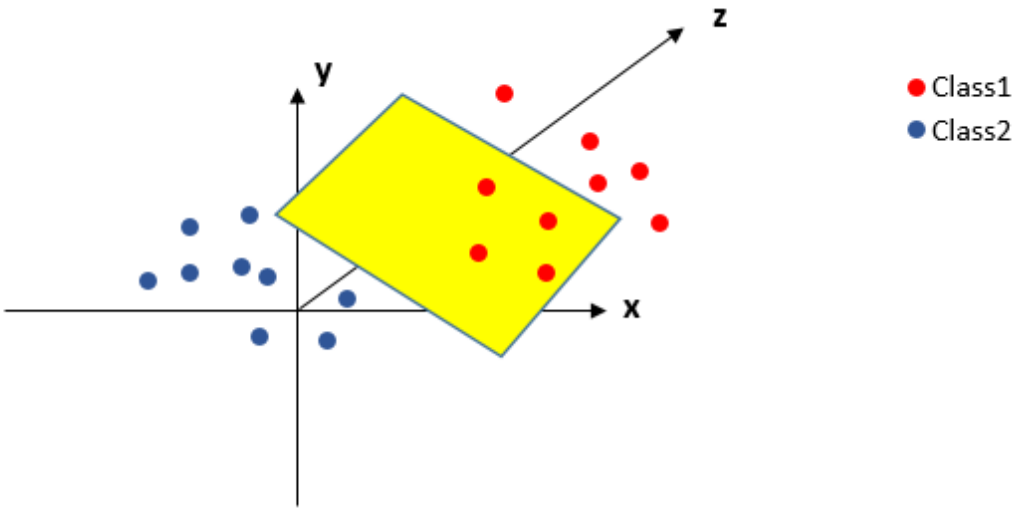

Figure 3.4 Feature space after applying kernel function.

Even though SVM is originally designed for binary classification, it can perform multi class classification. There are different approaches such as one-against-one and one-against-all. Due to the shorter training time, one-against-one is more suitable technique (Chih-Wei Hsu and Chih-Jen Lin, 2002).

SVM brings solutions to many real-world problems and is widely used in speech recognition, speaker identification, text categorization, image classification,

bioinformatics, hand-written character recognition etc. We used “e1071” R package to train support vector machine on training data.

3.3.2 Random Forest

The general method of random decision forests was first proposed by Ho in 1995 (Ho, 1995). An extension of algorithm was developed for classification purposes by Leo Breiman (Breiman, 2001), the algorithm uses an ensemble of decision trees. Each decision tree is constructed by applying bootstrap sampling of the data. The feature size for each node is represented by *m* that is quite smaller than the total number of features of the original data. So, *m* features are randomly chosen out of all possible features, the best split of *m* features are used to partition the node. Trees of the forest grow as much as they can without any pruning. To perform classification of data, each sample vector is threaded over each trees of the forest, each tree provides a decision, finally the forest makes a final decision by choosing the highest voted class. As summary, random forest uses both bagging (Breiman, 1996) and random feature selection (Friedman, Hastie & Tibshirani, 2001) in tree building.

Random forest is widely used in molecular biology, financial analysis, computer vision, astronomy etc. We used "randomForest" R package to train the random forest method on training data.

3.4 Evaluation of System

In our experiment, we performed 100-fold Monte Carlo cross validation. For each cross-validation process, the patient samples are randomly partitioned as 70% train and 30% test dataset (Figure 3.5). We applied both PCA and FC & t-Test feature selection methods on the train set separately. With the help of feature selection methods, we could filter the significant genes. Then we trained both SVM and random forest models independently with significant genes of the train set. We built our predictive models and performed cross validation on unseen data. To measure the accuracy performance of predictive models, we constructed a confusion matrix.

Since our predictive model is not a binary classification (four cancer sub-types), we constructed a confusion matrix with four classes. And overall accuracy is calculated for each subtype independently as given in Equation 3.4. For example, we are calculating the accuracy for the Basal subtype patients, TP (true positive) would be the number of correct predictions of Basal subtype, TN (true negative) would be the sum of correct predictions of other subtypes. FP (false positive) would be the sum of predictions that outcome Basal subtype when true class labels are from other subtypes. FN (false negative) would be the sum of predictions that outcome other subtypes when true class labels are Basal subtype.

$$Acc = (\mathrm{TP} + \mathrm{TN}) \,/\, (\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}) \tag{3.4}$$

This process is repeated for 100 times and the overall performance of each machine learning method is reported as the average accuracy of all iterations. The signatures that provided the best prediction performance are recorded for further biological validations.

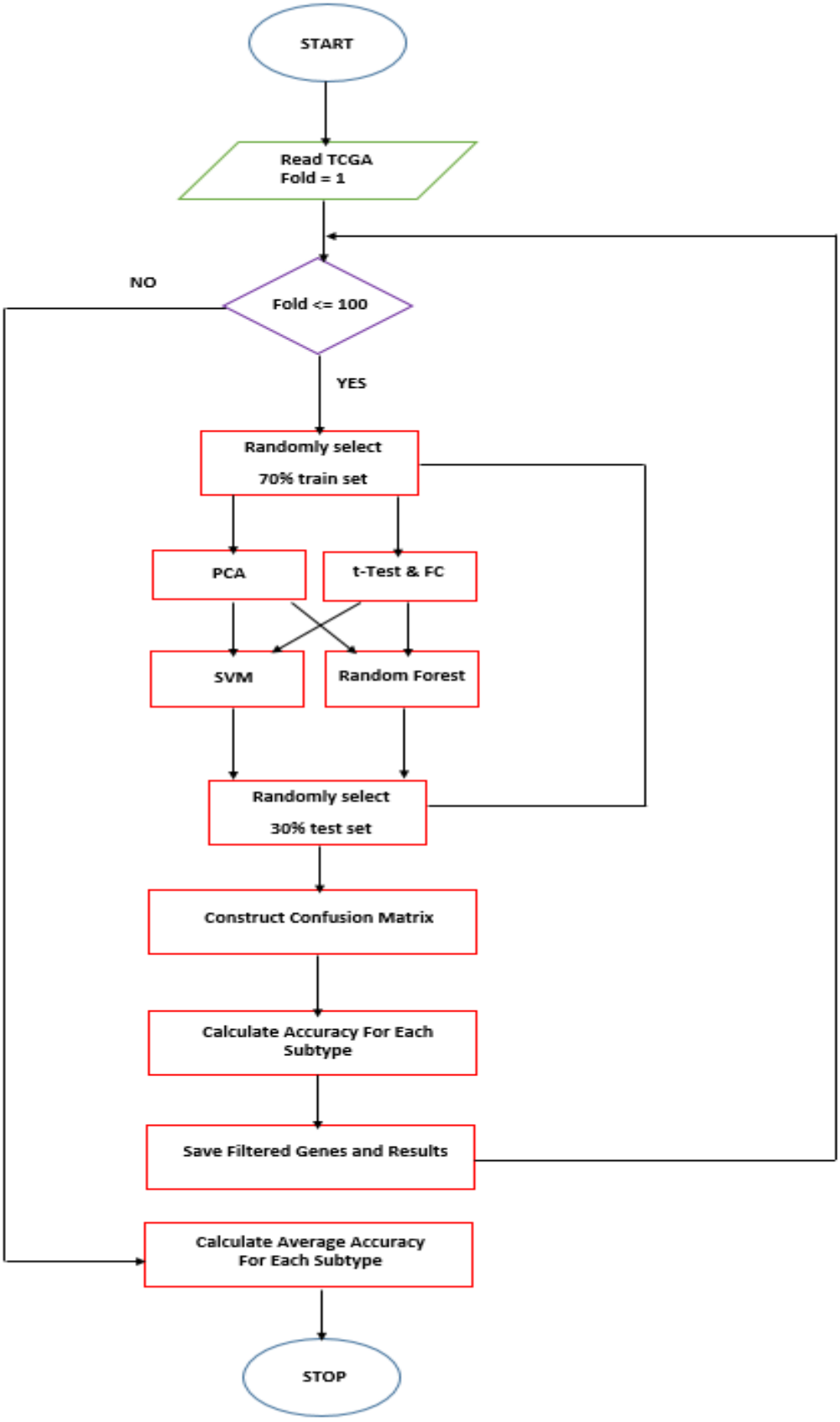

Figure 3.5 The workflow of the proposed method.

3.5 Biological Network Extraction

In order to reveal biological functions of best predictive signatures (i.e., genes), we applied a network based analysis. For this purpose, a functional protein-protein interaction network - STRING - was used (Szklarczyk et al., 2014). This network contains 10.579 proteins and 200.091 interactions. Using "igraph" R package we uploaded the input network as a graph on R-edge. We first identified the "core clusters" that are including the best predictive signature genes for each cancer subtype. For the genes which were not found in the core clusters, we applied a shortest path algorithm (Bread-First search) to interconnect these genes with the core cluster genes, these new clusters are called as "extended clusters". There were many interconnected genes; therefore, we only focused on some crucial genes, which have specific biological functions in cancer. For this purpose, we performed a functional enrichment analysis via DAVID tool (Huang, Sherman & Lempicki, 2009). We only focused on Gene Ontology annotations under the biological process functional class. We also mapped FC changes of genes that have cancer related annotations that were visualized by the Cytoscape tool for each subtype.

CHAPTER FOUR
EXPERIMENTAL RESULTS

4.1 Classification Performance of Models

Our dataset consists of 418 breast cancer patients' and 3 healthy people samples. 142 patients were Basal subtype, 67 patients were Her2 subtype, 105 patients were Luminal A subtype and 104 patients were Luminal B subtype. We performed 100-folds Monte Carlo cross validation in our experiments. For each cross-validation step, the patient samples of each subtype are randomly partitioned as 70% train and 30% test dataset. Before applying feature selection methods the genes total count was 13259, which is a high dimension to train machine learning models. PCA is performed with the options significance threshold equals to 0.05 and correlation coefficients threshold is bigger or equal to 0.7. We selected the genes covered in the first five principal components. The genes count varied between 165 and 328 on different folds of cross validation. On the other hand, Fold Change and t-Test were also applied together as a feature selection method. The amount of genes, whose absolute FC value is equal or bigger than 2 and t-Test p-value is equal or smaller than 0.01, varied between 510 and 667 on different folds of cross validation.

For each fold, after feature selection methods were performed, we trained our machine learning models and performed cross validation on unseen test data. Then we constructed confusion matrices for both SVM and RM predictive models. Since our predictive model contains multi-class (four subtypes) classification, we constructed the confusion matrix considering four classes. For this purpose, we used one-vs-all technique. This strategy trains a single classifier for each class by taking that class' samples as positive ones and others as negative ones. The base classifier produces a real-valued score for its decision. The prediction results (confusion matrix) of the SVM is given in Table 4.1 for each cross validation iteration.

Table 4.1 The confusion matrix of one-fold of Monte Carlo cross validation for four subtypes.

	Known (True) Subtype			
Prediction	**Basal**	**Her2**	**LumA**	**LumB**
Basal	41	0	0	0
Her2	2	17	0	1
LumA	0	0	26	6
LumB	0	4	6	25

This confusion matrix for Basal subtype is recalculated by applying one-vs-all technique as shown in Table 4.2. Basal subtype is labeled as positive class and the other subtypes are labeled as negative. According to the table, TP (true positive), which stands for the number of correct predictions of positive class (Basal subtype), is 41. TN (true negative), which stands for the number of correct predictions of negative class (Her2, LumA, LumB), is 68. FP (false positive), which stands for the sum of predictions that outcome positive class when true class labels are from negative, is 2. FN (false negative), which stands for the sum of predictions that outcome negative class when true class labels are positive, is 0. We calculated the accuracy for every iteration of cross validation as given in Equation 3.5. For every cross validation iteration, we calculated accuracy of both SVM and RF for each subtypes for both feature selection methods performed independently.

Table 4.2 The confusion matrix for only the Basal subtype

	Basal	**All**
Basal	41	0
All	2	68

When PCA was chosen as the feature selection method, the average accuracy of each subtype was reported as the prediction performance of each ML method (Figure 4.1). Basal and Her2 subtypes were classified with 0.97 and 0.94 average accuracies,

respectively. LumA and LumB subtypes were classified with 0.90 and 0.88 average accuracies, respectively.

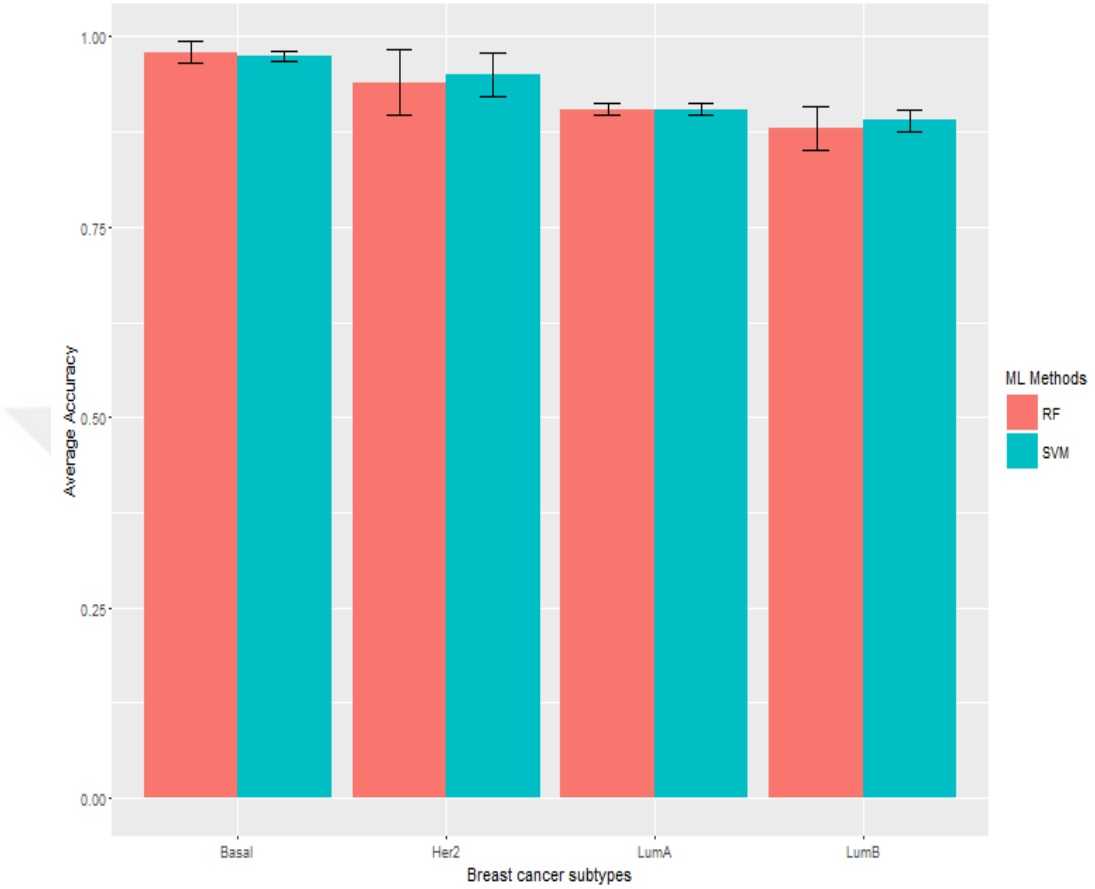

Figure 4.1 Average accuracy of 100-fold Monte Carlo cross-validation (error bar shows the standard deviation of 100 folds) when PCA was chosen as the feature selection method.

When FC & t-Test performed together as the selection feature method, the average accuracy of each subtype was reported as the prediction performance of each ML method (Figure 4.2). Basal and Her2 subtypes were classified with 0.97 and 0.95 average accuracies, respectively. LumA and LumB subtypes were classified with 0.89 and 0.87 average accuracies, respectively.

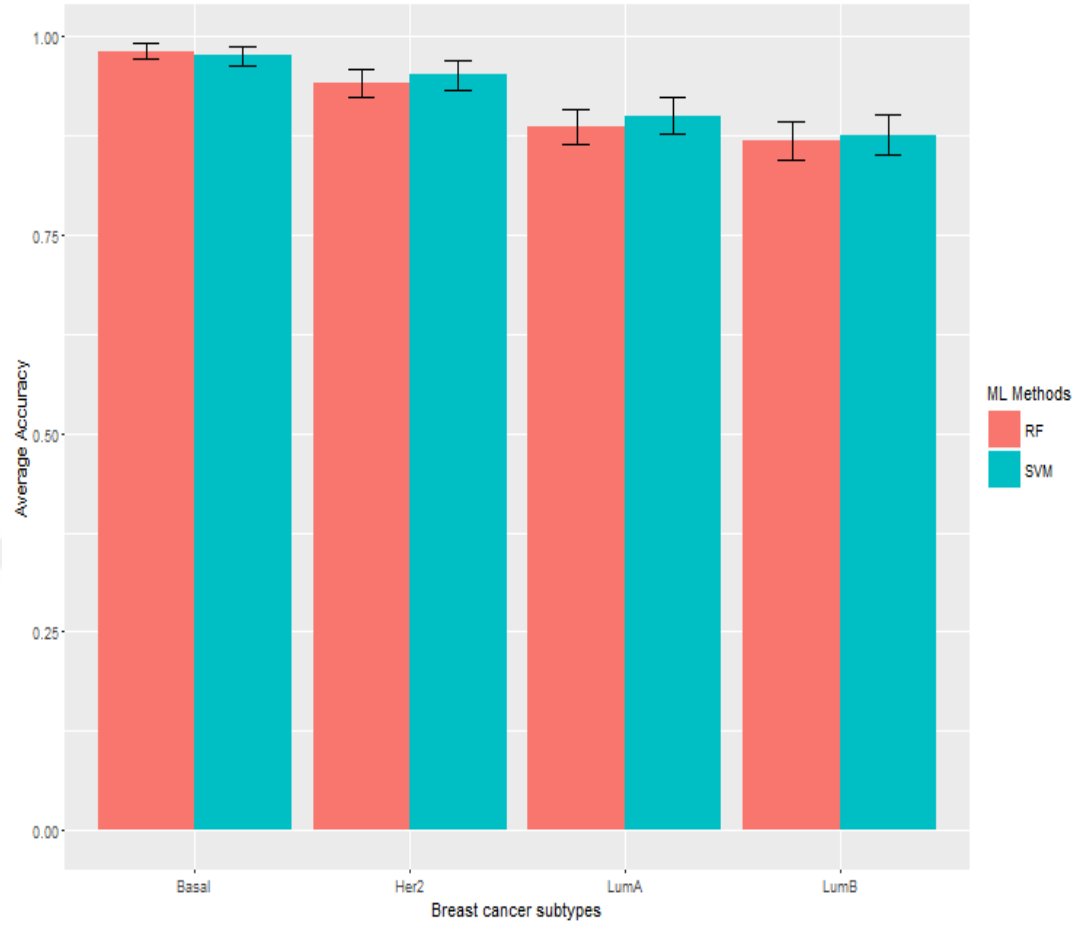

Figure 4.2 Average accuracy of 100-fold Monte Carlo cross-validation (error bar shows the standard deviation of 100 folds) when FC & t-Test were chosen as the feature selection method.

The results showed that applied feature selection methods (PCA and FC & t-Test) does not have an impact on the performance of machine learning methods in our experiments. Furthermore, ML methods are more successful for the discrimination of Basal and Her2 subtypes from others.

In a similar work (List et al., 2014), they also used data of TCGA breast cancer patients. They run 10 times 0.632 bootstrapping recursively for feature elimination and used RF (using varSelRF package) as the ML method. Their overall performance was nearly 100% accuracy. In their experiment, the Basal subtype prediction had the best performance and Luminal B had the worst prediction performance. Their results are also concordance with our prediction performances. So, the subtype classification

of breast cancer patients was not significantly affected by the different feature selection and machine learning techniques. Considering the other works on TCGA breast cancer data we can claim that this dataset is generally providing high accuracies for the subtype classification problem.

4.2 Biological Validation of Results

On every fold of Monte Carlo cross validation, we saved the accuracy and significant genes results into a file. FC is more widely used in gene expression analysis. So, instead of PCA, we chose the intersection of genes which had the best accuracies for both SVM and RF from the results of FC & t-Test performed together. We selected these best significant genes to find out the relevant biological activities, functions and pathways by applying a network analysis. We constructed a sub-network for each subtype that covers interactions (obtained from the STRING network) between the best predictive genes of that subtype. Some genes were not covered in the network, some of them created a core-connected cluster. The rest of them were not connected to these core-clusters. Therefore, we found the extended cluster that connects sub-clusters and the rest of the genes. Table 4.3 shows the amount of significant, covered by PPI network, core-cluster and extended cluster genes.

Table 4.3 The gene count based on the network topology.

Subtype	**Significant Genes**	**Network Coverage**	**Core Cluster**	**Extended Cluster**
Basal	547	263	83	1874
Her2	575	271	91	1814
LumA	545	250	105	1872
LumB	595	282	97	1937

For functional analysis, we investigated the Gene Ontology (GO) annotations of the member of extended core network by using DAVID tool. We only focused on the biological process GO terms. Table 4.4 shows the selected GO-terms and total number of genes that are annotated with these terms for each subtype. The most of the predictive genes of all subtypes are annotated with the cell proliferation and apoptosis terms that are well-known regulator processes for cancer development and progression. Immune system related genes are only shared by Basal and Her2 subtypes. A recent study (Jézéquel et al., 2015) also showed the high enrichment of immune system related genes for basal-like subtype. Hormone response related genes are only the member of LumA and LumB signatures, this result is well-known factor since LumA and LumB are ER+ cancer types.

Table 4.4 Total number of genes that were annotated with the given GO-term.

GO-term	**Basal**	**Her2**	**Luminal A**	**Luminal B**
Regulation of cell proliferation	106	97	97	105
Positive regulation of immune system process	39	39	-	-
Inflammatory response	54	47	52	55
Regulation of apoptosis	114	103	109	108
T-cell activation	23	22	20	23
Response to hormone stimulus	-	-	54	53

As seen on Table 4.5, most of the predictive genes are common among the significant genes of each subtype. These genes have important roles in subtype predictions, therefore we selected these genes for each subtype to find out the relevant biological activities, functions and pathways by applying a network analysis. We also mapped the fold change values and GO-annotations of genes in these sub-networks to understand the function and collaboration of these genes related with the specific biological process.

Table 4.5 Intersection of predictive genes between each subtype.

	Basal	Her2	Luminal A	Luminal B
Basal	547	504	474	499
Her2	504	575	471	511
LumA	474	471	545	489
LumB	499	511	489	595

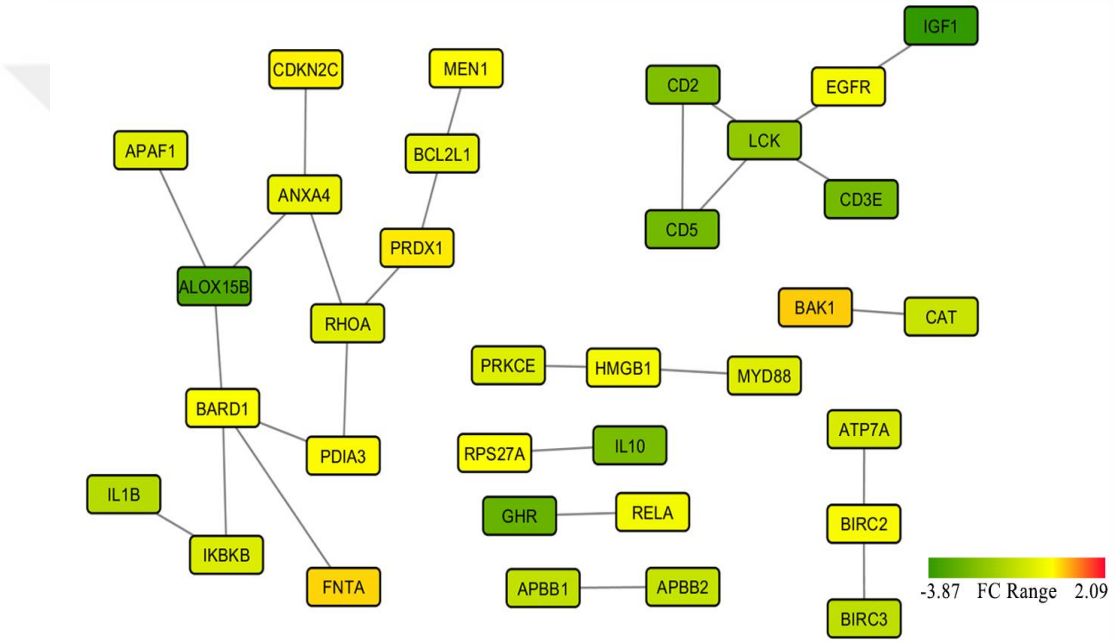

Figure 4.3 Basal subtype genes that were annotated with the "Regulation of apoptosis" GO-term.

In sub-networks, nodes are the genes that are taking place in the specific biological activity. Edges show the relationship between two nodes which can be identified in two ways. One way is these genes concurrently take place in the same biological process and the other way they take place in turn for the specific biological process.

When we analyzed the predictive genes of the Basal subtype, most of them (more than 106) were annotated with "Regulation of apoptosis" and "Regulation of cell proliferation" GO-terms. Figure 4.3 shows the network topology of "Regulation of

apoptosis" genes. There are two main connected clusters (on the left and up-right corners) of genes, the rest of them are loosely connected; almost all genes are either down-regulated or not expressed for this subtype. With a similar analysis, Figure 4.4 shows the network topology of "T-cell activation" genes for the HER-2 subtype. It contains quite small well-connected gene cluster, all genes were significantly down-regulated in this subtype. Figure 4.5 shows the network topology of "Response to hormone stimulus" genes for the LumA subtype. There is no a core-connected network and gene expressions are mixed for this subtype. Figure 4.6 shows the network topology of "Regulation of cell proliferation" genes for the LumB subtype. It is comparably larger network and contains several small clusters of genes from different gene expression level.

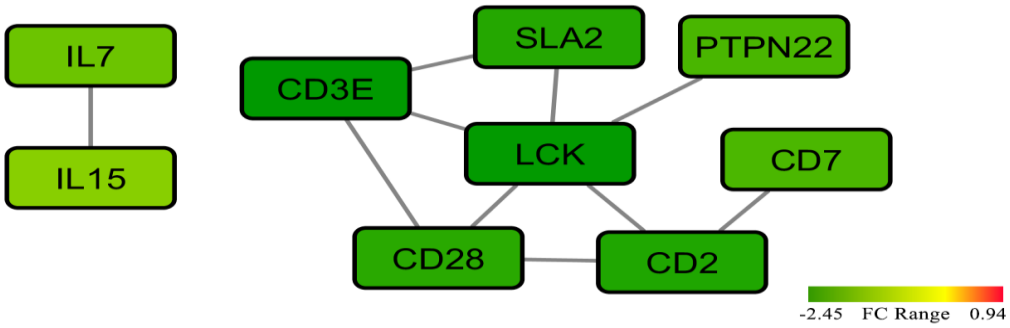

Figure 4.4 Her2 subtype genes that were annotated with the “T-cell activation” GO-term.

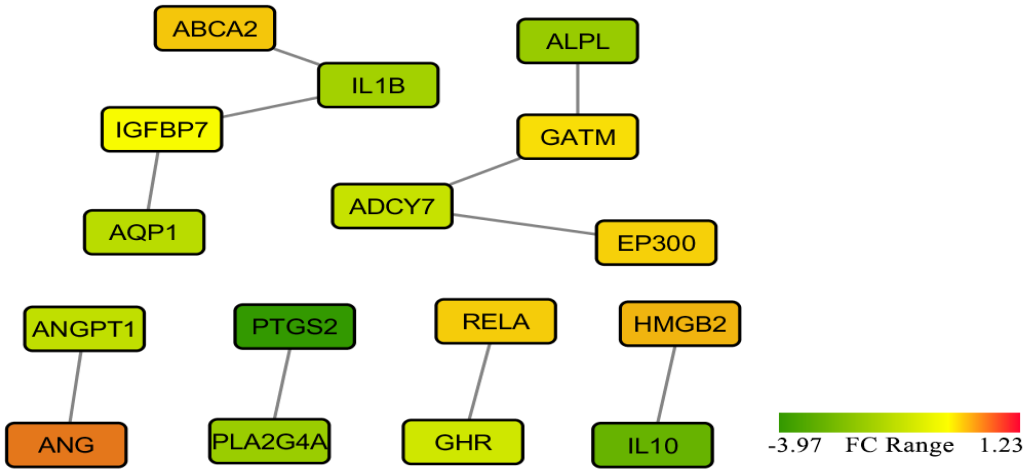

Figure 4.5 LumA subtype genes that were annotated with the "Response to hormone stimulus" GO-term.

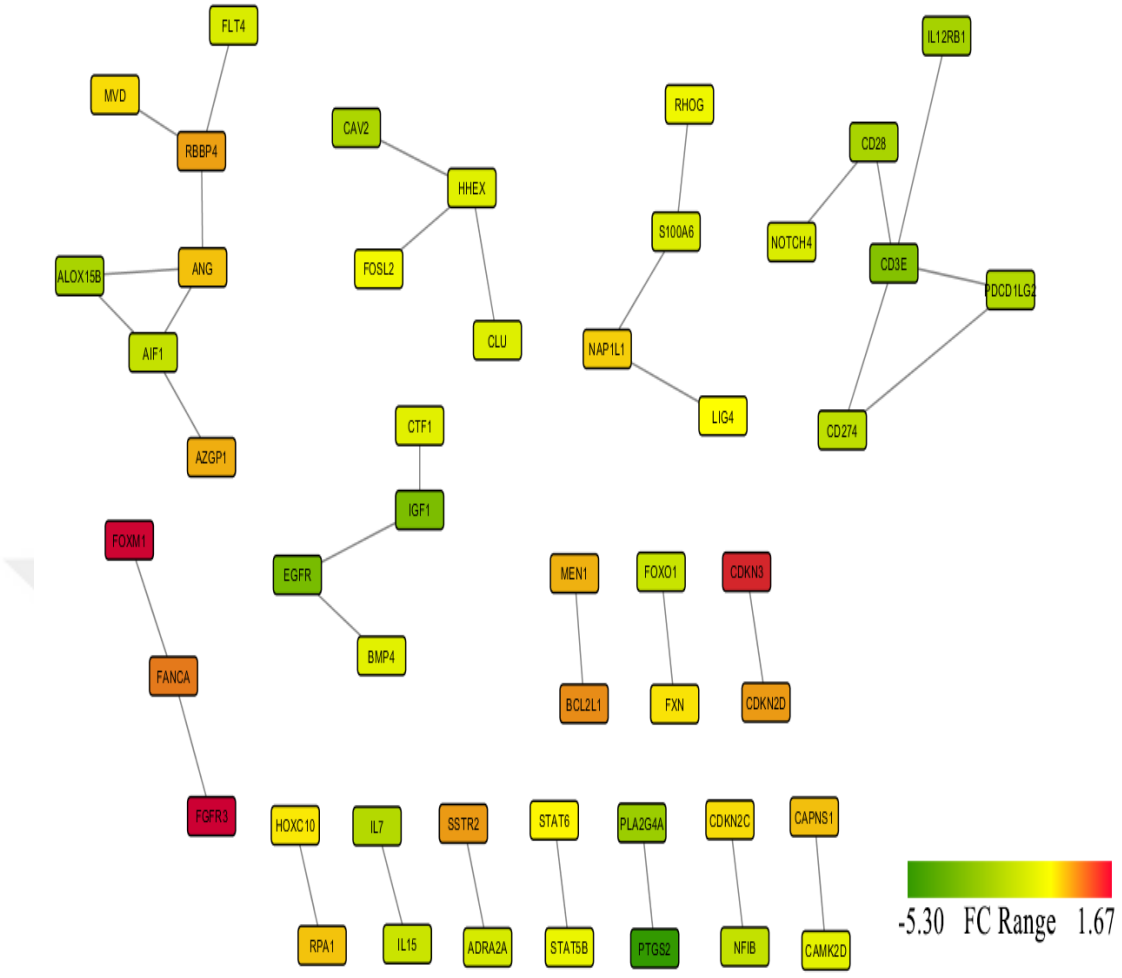

Figure 4.6 LumB subtype genes that were annotated with the "Regulation of cell proliferation" GO-Term.

By using TCGA RNA-seq data, first we selected significant genes. Then fold change values and GO-annotations of these genes were mapped in these sub-networks. Fold change values showed us down-regulation or up-regulation of the genes for different cancer subtypes. The significant genes in a specific sub-network with a certain up / down-regulation pattern or GO term annotation should be further analyzed on different breast cancer cohorts to see if they have important roles in these subtypes and biological processes. Another direction might be the highlighting the relations between specific GO-terms. The genes having an impact on breast cancer subtype classification can be determined on different patient datasets and be enriched by applying different functional annotations .

CHAPTER FIVE
CONCLUSION AND FUTURE WORK

For many years, histology and morphology had a lead role on breast cancer classification. It was seen that prognosis could be different even with the same treatment for the same histology and the clinical stage. The development of DNA microarray technology, measuring the expression levels of entire transcriptome simultaneously, can help molecular subtyping. Such technologies have positive effects on early diagnosis and prognosis of cancer.

In our study, we classified breast cancer patients by using RNA-sequencing data and ML methods. The overall performances of machine learning models varied between 0.88 to 0.97 average accuracy when PCA is applied as the feature selection method; and varied between 0.87 to 0.97 average accuracy when FC & t-Test is applied. It showed us that applying different feature selection methods do not have high impact on ML prediction performance. Basal and Her2 subtypes predictions had the highest accuracies compared to the LumA and LumB predictions. Cancer subtypes were predicted with similar accuracies by applying different machine learning methods. This study was presented in 1st International Mediterranean Science and Engineering Congress (Burakgazi & Işık, 2016).

The biological analysis of predictive genes revealed that most of them were shared by the other subtypes and such genes were commonly annotated with *cell proliferation* and *apoptosis* Gene Ontology terms that are well-known regulator processes for cancer development and progression. Based on our results, a network based analysis can identify more accurate predictive genes for cancer subtype classification, such genes might be suggested as biomarkers for the new diagnostic kits.

As a future work, the most predictive genes highlighted by our network analysis can be tested on different breast cancer datasets to investigate their classification capabilities in the subtype prediction. Beside over-representation analysis, there are

different pathway analysis methods in literature. Functional class-scoring and topology-based methods (Khatri, Sirota & Butte, 2012) can be researched to have different pathway analysis and compare our results with them.

REFERENCES

Ahmad, L. G., Eshlaghy, A. T., Poorebrahimi, A., Ebrahimi, M., & Razavi, A. R. (2013). Using three machine learning techniques for predicting breast cancer recurrence. *Journal of Health & Medical Informatics*, 2013.

Ali, H. R., Rueda, O. M., Chin, S. F., Curtis, C., Dunning, M. J., Aparicio, S. A., & Caldas, C. (2014). Genome-driven integrated classification of breast cancer validated in over 7,500 samples. *Genome Biology*, *15*(8), 1.

Anders, S., Pyl, P. T., & Huber, W. (2014). HTSeq–a Python framework to work with high-throughput sequencing data. *Bioinformatics*, btu638.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123-140.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5-32.

Brown, M. P., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., Jr., M. A. & Haussler, D. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences*, *97*(1), 262-267.

Burakgazi, Y., & Işık Z. (2016). Classification of breast cancer subtypes by using TCGA data. *Proceedings of 1st International Mediterranean Science and Engineering Congress*, 1699-1706

Cancer genome atlas network. (2012). Comprehensive molecular portraits of human breast tumours. *Nature*, *490*(7418), 61-70.

Cruz, J. A., & Wishart, D. S. (2006). Applications of machine learning in cancer prediction and prognosis. *Cancer informatics*, 2.

Dobin, A., & Gingeras, T. R. (2015). Mapping RNA-seq Reads with STAR. *Current Protocols in Bioinformatics*, 11-14.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1). Berlin: Springer

Gene Ontology Consortium. (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research*, *32*(suppl 1), D258-D261.

Gevaert, O., De Smet, F., Timmerman, D., Moreau, Y., & De Moor, B. (2006). Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. *Bioinformatics, 22*(14), 1367-4811.

Griffith, M., Walker, J. R., Spies, N. C., Ainscough, B. J., & Griffith, O. L. (2015). Informatics for RNA sequencing: a web resource for analysis on the cloud. *PLoS Comput Biol*, *11*(8), e1004393.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*(2003), 1157-1182.

Jolliffe, I. (2002). *Principal component analysis*. New York: John Wiley & Sons, Ltd.

Ho, T. K. (1995). Random decision forests. In *Document Analysis and Recognition, 1995 IEEE., Proceedings of the Third International Conference on*, 278-282

Hu, J., Ge, H., Newman, M., & Liu, K. (2012). OSA: a fast and accurate alignment tool for RNA-Seq. *Bioinformatics*, *28*(14), 1933-1934.

Hsu, C. W., & Lin, C. J. (2002). A comparison of methods for multiclass support vector machines. *IEEE transactions on Neural Networks*, *13*(2), 415-425.

Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, *4*(1), 44-57.

Jézéquel, P., Loussouarn, D., Guérin-Charbonnel, C., Campion, L., Vanier, A., Gouraud, W., ... & Campone, M. (2015). Gene-expression molecular subtyping of triple-negative breast cancer tumours: importance of immune response. *Breast Cancer Research*, *17*(1), 1.

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., & Hattori, M. (2004). The KEGG resource for deciphering the genome. *Nucleic Acids Research*, *32*(suppl 1), D277-D280.

Khatri, P., Sirota, M., & Butte, A. J. (2012). Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Computational Biology*, *8*(2), e1002375.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., & Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, *14*(4), 1.

Kim, W., Kim, K. S., Lee, J. E., Noh, D. Y., Kim, S. W., Jung, Y. S., ... & Park, R. W. (2012). Development of novel breast cancer recurrence prediction model using support vector machine. *Journal of Breast Cancer*, *15*(2), 230-238.

Kreike, B., van Kouwenhove, M., Horlings, H., Weigelt, B., Peterse, H., Bartelink, H., & van de Vijver, M. J. (2007). Gene expression profiling and histopathological characterization of triple-negative/basal-like breast carcinomas. *Breast Cancer Research*, *9*(5), 1.

Li, B., & Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, *12*(1), 1.

Liao, Y., Smyth, G. K., & Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, *30*(7), 923-930.

List, M., Hauschild, A. C., Tan, Q., Kruse, T. A., Mollenhauer, J., Baumbach, J., & Batra, R. (2014). Classification of breast cancer subtypes by combining gene expression and DNA methylation data. *Journal of Integrative Bioinformatics*, *11*(2), 236.

Listgarten, J., Damaraju, S., Poulin, B., Cook, L., Dufour, J., Driga, A., Mackey, J., Wishart, D., Greiner, R., & Zanke, B. (2004). Predictive models for breast cancer susceptibility from multiple single nucleotide polymorphisms. *Clinical Cancer Research*, *10*(8), 2725-2737.

Nicolae, M., Mangul, S., Măndoiu, I. I., & Zelikovsky, A. (2011). Estimation of alternative splicing isoform frequencies from RNA-Seq data. *Algorithms for Molecular Biology*, *6*(1), 1.

Robertson, S. (2014). *Gene expression measurement.* Retrieved August 12, 2016, from http://www.news-medical.net/life-sciences/Gene-Expression-Measurement.aspx

Sanguansat, P. (Ed.). (2012). *Principal component analysis-multidisciplinary applications*. Croatia: InTech.

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., ... & Kuhn, M. (2014). STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*, *43*, Database issue D447-D452.

Tomczak, K., Czerwinska, P., & Wiznerowicz, M. (2015). The cancer genome atlas (TCGA): an immeasurable source of knowledge. *Contemporary Oncology (Pozn),19*(1A), A68-A77.

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., Van Baren, M. J., ... & Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, *28*(5), 511-515.

Trochim, W. M. (2006). *The t-test*. Retrieved December 20, 2016, from http://www.socialresearchmethods.net/kb/stat_t.php

Vapnik, V. N., & Vapnik, V. (1998). *Statistical learning theory* (Vol. 1). New York: John Wiley & Sons, Ltd.

Voduc, K. D., Cheang, M. C., Tyldesley, S., Gelmon, K., Nielsen, T. O., & Kennecke, H. (2010). Breast cancer subtypes and the risk of local and regional relapse. *Journal of Clinical Oncology, 28*(10), 1684-1691.

Wang, K., Singh, D., Zeng, Z., Coleman, S. J., Huang, Y., Savich, G. L., ... & MacLeod, J. N. (2010). MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Research*, *38*(18), e178.

Wu, T. D., & Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, *26*(7), 873-881.

Yersal, O., & Barutca, S. (2014). Biological subtypes of breast cancer: Prognostic and therapeutic implications. *World of Journal Clinical Oncology. World, 5*(3), 412-424

Young, E. M., Gordon, D. B., & Voigt, C. (2016). *U.S. Patent No. 20,160,083,722*. Washington, DC: U.S. Patent and Trademark Office.

Zhao, S., Xi, L., & Zhang, B. (2015). Union exon based approach for RNA-Seq gene quantification: To Be or Not to Be?. *PloS one*, *10*(11), e0141910.

Zhao, S., Zhang, B., Zhang, Y., Gordon, W., Du, S., Paradis, T., ... & von Schack, D. (2016). Bioinformatics for RNA-Seq Data Analysis. *Bioinformatics-Updated Features and Applications*, 125.

Zhao, X., Rødland, E. A., Tibshirani, R., & Plevritis, S. (2015). Molecular subtyping for clinically defined breast cancer subgroups. *Breast Cancer Research*, *17*(1), 1.