DETERMINATION OF THE EFFECT OF POLYADENYLATION SLR VALUES ON MICROARRAY DATA CLASSIFICATION

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

ÜMİT ASLAN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
COMPUTER ENGINEERING

SEPTEMBER 2014

Approval of the thesis:

DETERMINATION OF THE EFFECT OF POLYADENYLATION SLR VALUES ON MICROARRAY DATA CLASSIFICATION

submitted by **ÜMİT ASLAN** in partial fulfillment of the requirements for the degree of **Master of Science in Computer Engineering Department, Middle East Technical University** by,

Prof. Dr. Canan Özgen _____
Dean, Graduate School of **Natural and Applied Sciences**

Prof. Dr. Adnan Yazıcı _____
Head of Department, **Computer Engineering**

Assoc. Prof. Dr. Tolga Can _____
Supervisor, **Computer Engineering Department, METU**

Examining Committee Members:

Prof. Dr. Ferda Nur Alpaslan _____
Computer Engineering Department, METU

Assoc. Prof. Dr. Tolga Can _____
Computer Engineering Department, METU

Prof. Dr. Ahmet Coşar _____
Computer Engineering Department, METU

Assoc. Prof. Dr. A. Elif Erson Bensan _____
Biology Department, METU

Assist. Prof. Dr. Aybar Can Acar _____
Informatics Institute, METU

Date: 02.09.2014

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name: ÜMİT ASLAN

Signature:

ABSTRACT

DETERMINATION OF THE EFFECT OF POLYADENYLATION SLR VALUES ON MICROARRAY DATA CLASSIFICATION

Aslan, Ümit

M.S., Department of Computer Engineering

Supervisor: Assoc. Prof. Dr. Tolga Can

September 2014, 99 pages

Microarray data classification is generally used to predict unknown sample outcomes by the help of models created using the preprocessed and categorized microarray data that includes gene expression values. Preparation of microarray experiments, design of Affymetrix chips and availability of previous microarray experiments give the opportunity to extract a new kind of data; differential expressions of proximal and distal probes (Short to Long Ratio -SLR- values), which is used to predict the alternative polyadenylation (APA) events. In this thesis, we aim to integrate gene expression data and these SLR values and then determine how the microarray data classification is affected after this integration process. Because of the filtering operations applied while predicting the APA events, SLR values are not available for all the probe sets on a microarray sample. These missing values are not left out not only while integrating the data, but also while applying the classification techniques. Three types of classification techniques, Support Vector Machines (SVM), Decision Tree (J48) and Random Forest are applied to primary breast tumor microarray data

before and after integration of gene expression values with SLR values and the classification accuracies of metastasis are found out. The results show that; APA events have incontrovertible impact on gene expression classifications and mostly towards improvement of accuracies.

Keywords: microarray, gene expression data, alternative polyadenylation, classification, support vector machines, decision tree, random forest, tumor, metastasis

ÖZ

MİKROÇİP VERİSİNİN SINIFLANDIRILMASI ÜZERİNDE ÇOKLU ADENİN OLAYI SLR DEĞERLERİNİN ETKİSİNİN TESPİT EDİLMESİ

Aslan, Ümit
Yüksek Lisans, Bilgisayar Mühendisliği Bölümü
Tez Yöneticisi: Doç. Dr. Tolga Can

Eylül 2014, 99 sayfa

Mikroçip veri sınıflandırılması genel olarak, gen ifade değerlerini içeren önceden işlenmiş ve kategorize edilmiş mikroçip verisinin kullanılmasıyla oluşturulan modellerin yardımı ile bilinmeyen örnek sonuçlarının tahmin edilmesinde kullanılmaktadır. Mikroçip deneylerinin hazırlanışı, Affymetrix çiplerinin tasarımı ve önceki mikroçip deneylerinin kullanılabilirliği, alternatif çoklu adenin olaylarının (APA) tahmin edilmesinde kullanılan yeni bir çeşit veri olan yakın ve uzak ölçüm uçlarının diferansiyel ifadelerinin (Kısadan Uzuna Oran -SLR- değerlerinin) çıkarımı fırsatını sunmaktadır. Bu tezde, gen ifade değerlerini bu SLR değerleri ile birleştirmeyi ve ardından bu birleştirme işlemi sonrası mikroçip veri sınıflandırılmasının nasıl etkilendiğini belirlemeyi amaç edindik. APA olaylarının tahmin edilmesi esnasında uygulanan filtreleme operasyonları sebebiyle, bir mikroçip örneği üzerinde bulunan bütün ölçüm uçları için SLR değerleri mevcut değildir. Bu eksik değerler, sadece verilerin birleştirilmesinde değil sınıflandırma tekniklerinin uygulanması esnasında da dışarıda tutulmamaktadır. Destek Vektör Makinesi (SVM), Karar Ağacı (J48) ve Rastgele Orman olmak üzere üç tip

sınıflandırma tekniği, birincil meme tümörü mikroçip verisine gene ifade değerleri ile SLR değerlerinin birleştirilmesi öncesi ve sonrası uygulandı ve metastazların sınıflandırma doğrulukları bulundu. Sonuçlar gösteriyor ki; APA olayları gen ifade sınıflandırılması üzerinde yadsınamaz bir etkiye sahiptir ve bu etki çoğunlukla doğrulukların gelişmeleri yönündedir.

Anahtar Kelimeler: mikrodizi, gen ifade değeri, alternatif çoklu adenin olayı, sınıflandırma, destek vektör makinesi, karar ağacı, rastgele orman, tümör, metastaz

To my lovely wife Ayşe,
who motivated me a lot for this thesis.

ACKNOWLEDGMENTS

I owe my deepest gratitude to my supervisor; Assoc. Prof. Dr. Tolga Can. This thesis would not have been accomplished without his endless help, support, guidance and patience with me.

I deeply thank to Prof. Dr. Ferda Nur Alpaslan, Prof. Dr. Ahmet Coşar, Assoc. Prof. Dr. A. Elif Erson Bensan and Assist. Prof. Dr. Aybar Can Acar who accepted to join my defense committee.

I deeply thank to my mother for her support without complaining about any of her health conditions.

I would thank to my wife for her continuous encouragement and motivation that helped me a lot through this thesis.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

APA	Alternative Polyadenylation
DNA	Deoxyribonucleic Acid
GEO	Gene Expression Omnibus
NCBI	National Cancer for Biotechnology Information
RNA	Ribonucleic Acid
mRNA	Messenger Ribonucleic Acid
SLR	Short to Long Ratio
UTR	Untranslated Region
CPSF	Cleavage/Polyadenylation Specificity Factor
cDNA	Complementary DNA
IVT	In Vitro Transcription
MDL	Minimum Description Length
SVM	Support Vector Machine
ID3	Iterative Dichotomiser 3
C4.5	Successor of ID3
ER	Eustrogen Receptor
PCA	Principle Component Analysis

CHAPTER 1

INTRODUCTION

1.1 Problem Definition

In some infected or cancerous cells and tissues, the expression levels of some genes may differ when compared to normal/healthy cells and tissues as a result of many mutations occurring in genes which regulate the main cell activities and genome integrity while healthy cells evolve into cancerous cells. Such difference at gene expression values allows researchers to classify these tissues as cancerous/normal or according to different types or stages of chosen specific disease. At this point, microarray technology appears as a revolutionary hybridization technique that gives the chance to investigate thousands of genes' expressions concurrently in one experiment. Machine learning methods are needed to extract knowledge from these microarray gene expression profiles since these experiments produce a large number of gene expression values. The gene expression profiles from various microarray experiments and miscellaneous classification algorithms have been used for cancer classifications so far [1][11][35][36][56][57][58]. In order to improve the results of such classification problems different approaches are applied such as improving the selection of the most related genes from microarray data [59][60], combining the microarray data with clinical data (patient history, age, gender, clinical treatments, etc.) [48][49][50], to name only a few.

By taking into consideration the certain design criteria of Affymetrix chips which are used for microarray experiments, it is possible to detect alternative polyadenylation (APA) events; hence, a new kind of data -ratio of differential expressions of proximal and distal probes (SLR, Short/Long Ratio values)- is extracted by the APADetect tool [8].

In this thesis, the problem of how these SLR values affect the microarray data classification is questioned in both binary classification and multiclass classification issues. Integrating the SLR values with gene expression data and determining the classes in multiclass classification are other sub problems which are handled through this study.

1.2 Motivation and Contribution

In order to effectively analyze the experiments and determine the optimal treatment for cancer patients, identifying the genes which have different expression data under various experimental conditions and constructing the ideal classifiers according to these identified genes play crucial roles. This necessity raises the importance of microarray technology. Improvement of the accuracy of the classification in these experiments is a vital problem to state the right diagnosis or make the best prediction. By preprocessing the gene expression data effectively, integrating it with other kinds of informative data and applying correct classification techniques which are the most suitable for the data; researchers can make right predictions about existence of cancer, different stages and subtypes of cancer and even the sites of relapses in metastases of cancer.

The main motivation of this thesis is, determining the relationship between gene expression data and SLR values and figuring out whether SLR values improve the accuracy of microarray data classification or not, by an integrative approach.

Integrating the gene expression data with SLR values, applying different types of classification algorithms to these data separately and together, applying multiclass classification beside binary classification to figure out relationship effectively, can be considered as contributions of this thesis.

1.3 Organization

Apart from the introduction section, the rest of this thesis is organized as follows. Chapter 2 presents the biological and mathematical background of this study. Biological terms, gene selection topic and the classification techniques (SVM, J48

and Random Forest) are mentioned in detail in sub sections. In Chapter 3, the related work about microarray classification and integrating gene expression data with various kinds of data are presented. Chapter 4 describes the datasets which are used in this study and integration process of microarray data with SLR values. In Chapter 5, the classification experiments are described and their results are shown. Chapter 6 discusses the results, summarizes the thesis and concludes it with future work.

CHAPTER 2

BIOLOGICAL AND MATHEMATICAL BACKGROUND

2.1 Biological Background

Bioinformatics is defined as an interdisciplinary scientific area combining mainly biological and computer sciences with other contributing fields such as mathematics, statistics, chemistry, physics, linguistics and engineering [61]. It takes place at the crossing point of experimental and theoretical science. It applies informatics techniques which are derived from contributing fields to analyze, organize, understand, process, produce and store the information that is related with large scale biological data at molecular level [2][61]. Not only the scientific literature, patient clinical data, different experimental results are included in this biological data but also the information kept by genetic code is enclosed by it. It is very significant to investigate the biological data related to DNA, gene and protein in order to acquire the most reliable and useful information about genome processing. Bioinformatics has four main goals at this point: The first goal is storing and organizing the existing biological data so that the researchers can easily reach or expand it with new information which they extract from their researches. The second goal is implementing the required software and developing the needed tools which provide to analyze this data. The third goal is extracting biologically meaningful information from analysis of this data. The fourth and last goal is to help the practitioners in prediction of diseases and manufacturers in production of drugs.

In this section, brief explanations of basic concepts about molecular biology are presented beside mathematical background of attribute selection and classifications techniques.

2.1.1 DNA and Gene

Every living creature is composed of cells which contain the DNA molecules, mostly in their nucleus. DNA is basically formed by smaller molecules called *nucleotide*s, which are similar to each other in common part containing five-carbon sugar *deoxyribose* and *phosphate* group while they differ from each other by four type of *base*s Adenine (A), Guanine (G), Cytosine (C) and Thymine (T) as shown[1] in Figure 1. A and G are *purine* bases and C and T are *pyrimidine* bases. There are two sites on the sugar which are called 3' site and 5' site. Every single phosphate bonds to two sugars, one through 3' site and the other through 5' site. The backbones of the two chains are formed by these phosphate sugar links and the chains have directions from 5' to 3' and from 3' to 5' complementarily because of the asymmetrical structure of sugars. Bases are weakly bonded to one another as Adenine-Thymine or Cytosine-Guanine pairs. These two long chains of nucleotides swirl around on a common axis and create the double helix structure of DNA.

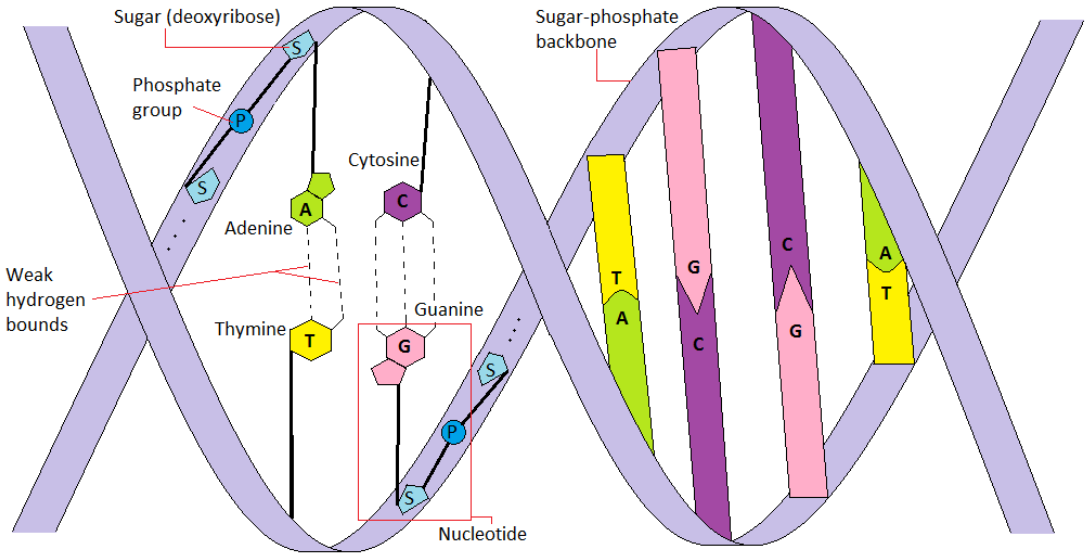

Figure 1: DNA structure

The double helix structure of DNA acts like a zipper at replication process. After the DNA polymerase enzyme catalyzes the replication, the helix structure unzips and complementary nucleotides available in the cell structure, bonds to the previously existing chains in order to create new two chains.

[1] http://ehrig-privat.de/ueg/images/dna-structure.jpg

Genes are the basic units of heredity those are different certain segments of DNA. Each gene has a different special task to do. They act by conducting the RNA production that settles the production of proteins which have many roles in cells such as; catalyzing metabolic reactions, transporting molecules, copying the DNA, etc. These proteins produced by genes of DNA, determine the specific physical features such as, skin color of a human or appearance of a tree's leaves. Different expressions of these features on the individuals are determined by *alleles* which are different forms of the genes.

2.1.2 RNA, mRNA and Central Dogma

Ribonucleic Acid (RNA) is a polymeric molecule which is synthesized from DNA with the use of RNA polymerase enzyme by the transcription process. It has many important roles in living organisms' cells such as; coding, decoding, regulation and expression of genes. Despite its chemical structure is similar to DNA, it has some differences as; being single stranded (only complementary base paring on itself in some cases), including shorter chain of nucleotides, containing *ribose* sugar instead of deoxyribose, being more unstable than DNA and more prone to degradation and containing the Uracil (U) base instead of Thymine as complementary to Adenine. There are three types of RNA those take place in translation process; ribosomal RNA (rRNA), transfer RNA (tRNA) and messenger RNA (mRNA).

Central Dogma of Molecular Biology, reconstructed by Francis Crick in 1958 and extended in 1970 [3], states that genetic information flows from DNA to DNA (by replication process) or DNA to RNA (by transcription process) and from RNA to proteins (by translation process), but there is not a reverse flow from protein to RNA or DNA as shown[2] in Figure 2.

[2] http://chsweb.lr.k12.nj.us/mstanley/outlines/dna/image14.gif

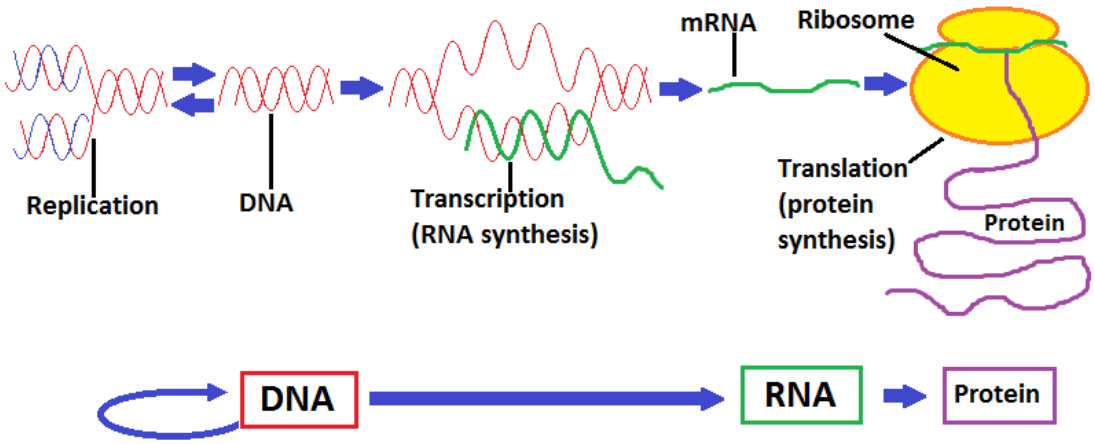

Figure 2: Central dogma of genetics

The process of RNA production from DNA is called transcription and takes places in nucleus. In this process the genetic information of DNA is written into primary transcript RNA. This pre-mRNA contains *intron*s which are non-coding sequences and *exon*s which are coding sequences. After splicing process, in which exons are remained and introns are removed, the mRNA is created as stated[3] at Figure 3. There are many different combinations of this splicing event so that a single gene can encode several proteins. This alternative splicing generally depends on the tissue type in which the transcription process occurs.

The structure of a typical human protein coding mRNA is shown[4] in Figure 4. The 5' cap plays crucial role at the recognition and attachment of mRNA to the ribosome and consists of an altered guanine nucleotide. 5' UTR is the section that takes place after this cap, before the start codon and is not translated in protein synthesis. Coding sequence is formed by *codon*s, groups of three nucleotides which are supposed to be translated into corresponding amino acids respectively. 3' UTR is the other untranslated section that takes place after the stop codon and before the poly(A) tail. Localization and stability of the mRNA, translation efficiency and Polyadenylation can be affected by regulatory parts placed in this 3' UTR section [4]. The poly(A) tail contains several Adenine bases and is placed at the end of mRNA just after the 3' UTR region.

[3] http://www.rnaxploration.com/_Media/pre-mrna_maturation-2_med.png
[4] http://upload.wikimedia.org/wikipedia/commons/b/ba/MRNA_structure.svg

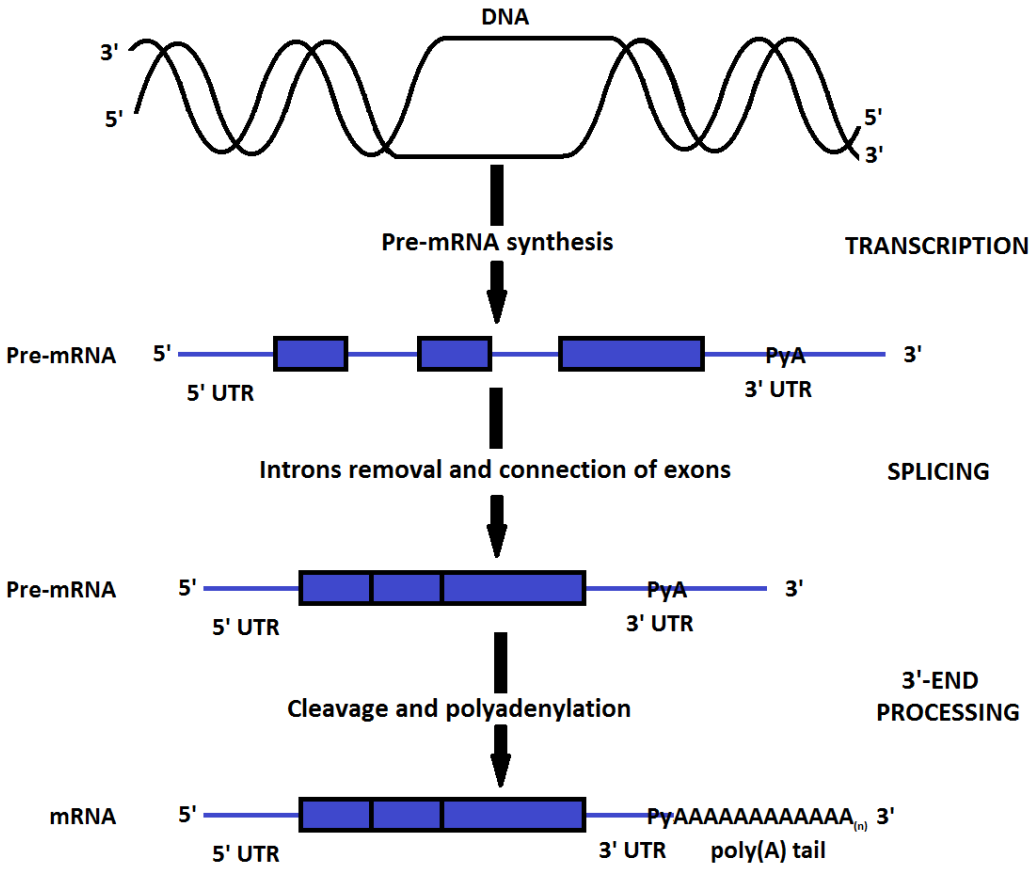

Figure 3: Pre-mRNA processing.

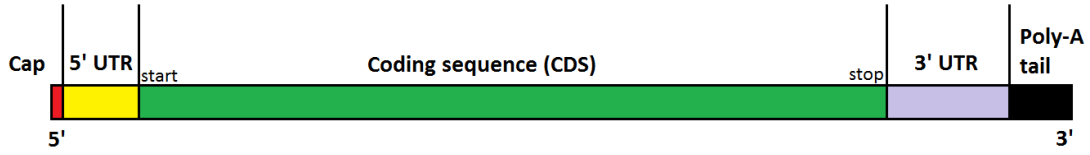

Figure 4: The structure of an mRNA

2.1.3 Polyadenylation and Alternative Polyadenylation

While the transcription process finishes, polyadenylation process occurs which means bonding of the poly(A) tail at the 3' end of the mRNA. As described in the previous topic, regulatory parts in 3' UTR section arrange this polyadenylation process. Figure 5 summarizes[5] the mechanism of polyadenylation process.

[5] http://www.biochemistry.ucla.edu/biochem/Faculty/Martinson/images/Slide1.jpg

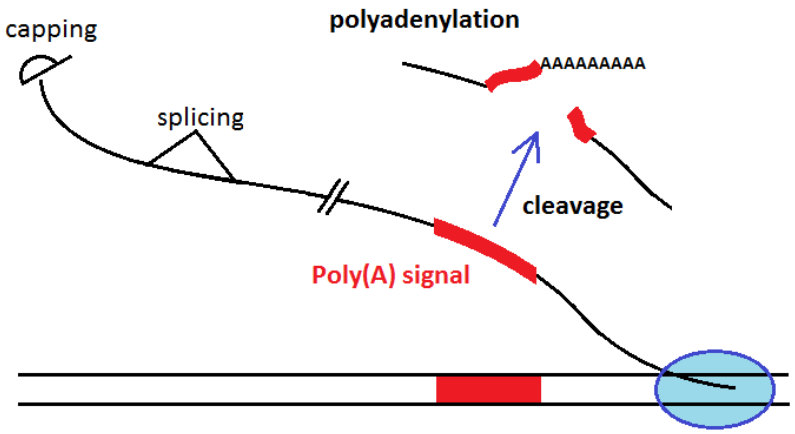

Figure 5: Polyadenylation process

A multiprotein complex cleaves off the 3' UTR segment and then produce the poly(A) tail. The enzyme CPSF catalyzes the cleavage which occurs in 10-35 nucleotides after the binding site, mostly the sequence AAUAAA on the mRNA.

There can be more than one polyadenylation site in the genes used in protein coding, which makes it possible to have more than one mRNAs differing in their 3' UTR sections from a single gene [5] like alternative splicing. This situation is called as alternative polyadenylation (APA) and has two types of effects; it can change the coding potential of the mRNA or it can change only the length of 3' UTR section while coding potential remaining same [6]. In second case it can change the availability of the binding sites on mRNA for proteins and microRNAs. According to researches about this subject [7] 3' UTR shortening or lengthening cases by APA events, in other words preferring the proximal or distal APA sites, may cause different biological results such as rapid proliferation of cells by escaping from microRNA binding sites.

2.1.4 DNA Microarray

A DNA Microarray is a set of microscopic DNA spots, named as features, attached to a solid slide. It is also commonly known as DNA chip, genome chip or gene array. By using DNA microarrays, large number of genes' expression levels can be measured simultaneously. There exists a specific sequence of DNA on each spot of this chip and they are called as *probes*. Depending on the structure of the solid

surface DNA microarrays can be silicon chip or glass chip. When an Affymetrix chip is used, DNA microarrays are also known as Affy chip [55]. The preparation steps of glass slide array and Affymetrix chip is given[6] in Figure 6.

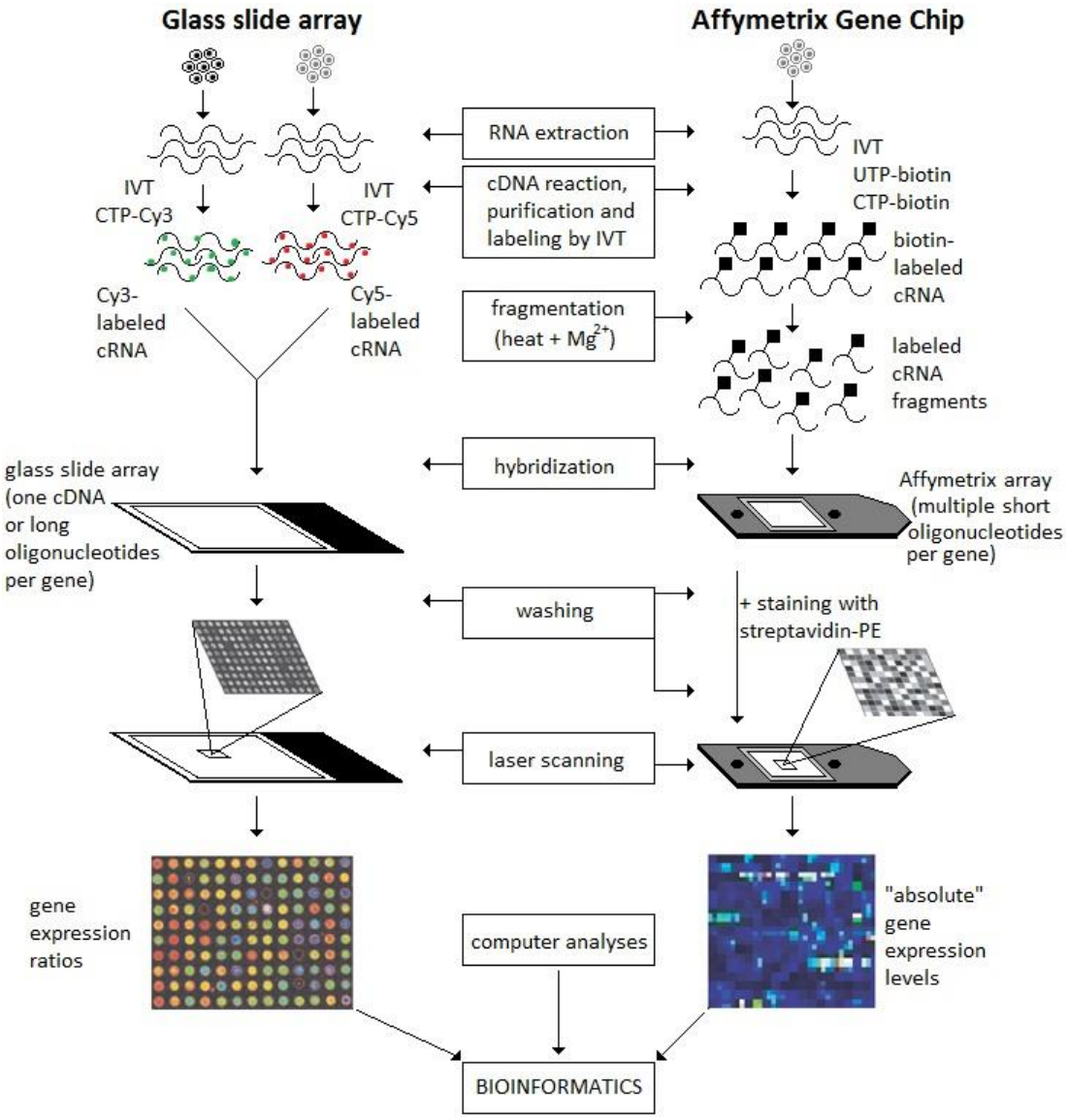

Figure 6: DNA microarray preparation

In glass slide arrays, mRNAs are extracted from two different tissue samples, for instance from cancerous and normal cells. Complementary DNA (cDNA) is produced and used in vitro transcription (IVT) with labeled nucleotides, usually red Cy5 and green Cy3. These two labeled cDNAs are mixed and hybridized with the DNA spots on the chip. By means of image analysis and fluorescent microscopes,

[6] http://www.nature.com/leu/journal/v17/n7/images/2402974f1.jpg

gene expression data is quantified by calculating the log ratio of the two dyes' intensities.

In Affymetrix chips, total mRNA is extracted from one population. cDNA is produced and by use of this cDNA in an IVT reaction, biotinylated cRNA is prepared. This cRNA is hybridized with the DNA spots on the chip, after it is fragmented. After washing and staining processes the chip is scanned on a laser scanner and absolute gene expression levels are extracted.

The collections of probes (25mers), which are prepared to analyze a certain DNA sequence, are called probe sets. There are perfect match (PM) probes beside mismatch (MM) probes in an individual probe set. As illustrated[7] in Figure 7, PM probes are constructed as exact match of the transcript while MM probes are constructed as no match of the transcript with an altered middle residue.

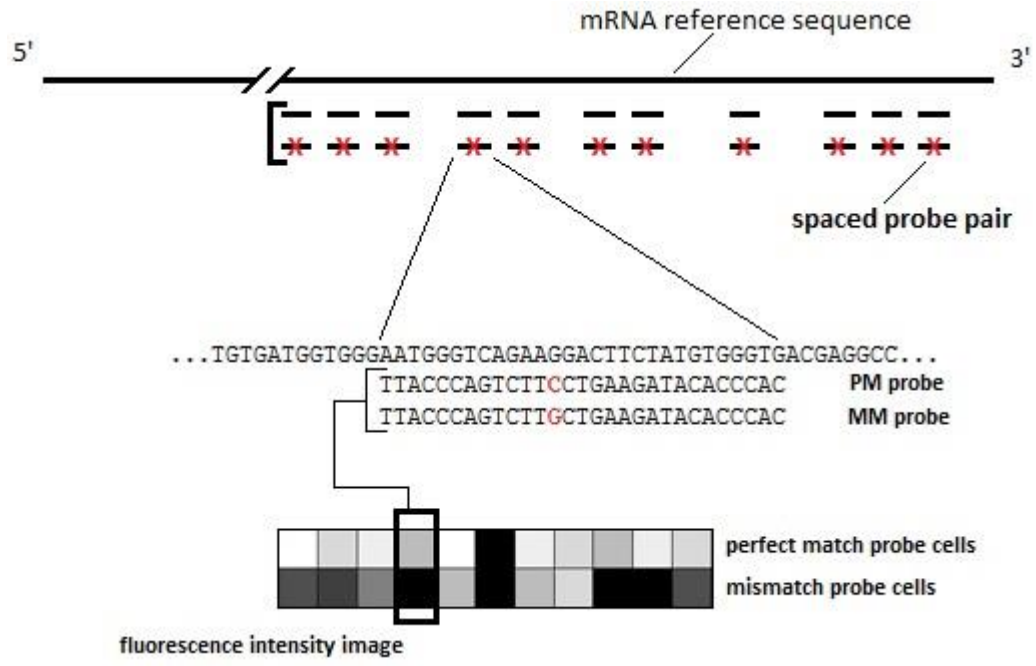

Figure 7: Probe set

[7] http://www.dkfz.de/gpcf/fileadmin/_migrated/pics/GeneChipDescription_01.jpg

2.2 SLR Values and APADetect

As a design issue, most part of the probes on Affymetrix microarrays are prepared to target 3' UTR sections. Since some probes target the regions arranged by APA events, this design issue gives the opportunity to analyze different 3' UTR isoforms which are created by APA events.

The probes in the probe set can be separated as two groups, proximal probes and distal groups, according to the location of the polyadenylation site. While proximal probes' sequences take place upstream of the splitting poly (A) site, distal probes' sequences take place downstream of it as shown[8] in Figure 8.

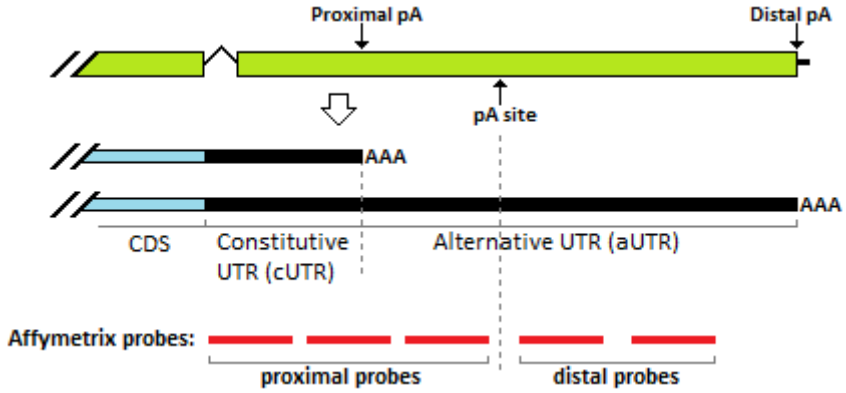

Figure 8: Proximal and distal probes

In order to identify expression levels of different 3'UTR isoforms a probe level analysis tool, APADetect [8], was developed. Known poly(A) sites available in PolyA_DB [9] are used to detect the target sites of transcripts in order to split the probe sets into two groups as proximal and distal probe sets. Then, in order to determine differentially expressed proximal and distal probe sets, signal intensities of these probe sets are analyzed. While doing this, raw intensities of the probe are used. Average signal intensities of these proximal and distal probe sets are calculated one by one for each individual gene and referred as *Short* and *Long*. For each sample, the proximal to distal ratio is calculated as division of these average probe intensities and named as *SLR* (Short/Long Ratio). A larger SLR value for a sample implies that a shorter 3' UTR isoform of the transcript is observed in that sample. By comparing

[8] http://www.pnas.org/content/106/17/7028/F1.large.jpg

the control group and test groups SLR values, isoform changes caused by APA events are identified. In order to detect the outliers, four different filtering processes are applied through the calculation of these SLR values. According to their orders, these filters are;

- *Size Filter*: Transcripts having only one proximal or one distal probe are discarded.
- *Outlier Probe Filter*: Individual outlier probes are discarded by use of Iglewicz and Hoaglin's median based outlier detection method[9]. After this second filter, since some samples can become single probed (proximal or distal) *Size Filter* is applied again.
- *Outlier Sample Filter*: Outlier samples, which have remarkable deviated SLR values when compared to their control groups, are determined with same Iglewicz and Hoaglin's method and eliminated.
- *Distal Filter*: Because both short and long isoforms can be recognized by proximal probes, probe sets having significantly larger distal intensities than proximal intensities are discarded.

Due to these filtering processes, at last point, SLR values may not be available for all the probe sets in a microarray sample.

2.3 Feature Selection

From different microarray experiments and conditions (let the number of sample type experiments be *j*) a gene expression data matrix, with dimensions *i* x *j*, can be derived as given in Table 1.

Table 1: Gene expression matrix

	Feature 1	**Feature 2**	**…**	**Feature i-1**	**Feature i**	**Class**
Sample 1	3852.528	7131.48	…	5510.58	4647.816	class A
Sample 2	1103.4	803.868	…	904.68	1397.376	class A
…	…	…	…	…	…	…
Sample j-1	160.8	389.544	…	232.104	186.54	class B
Sample j	699.888	557.952	…	348.348	480.564	class B

[9] http://www.itl.nist.gov/div898/handbook/eda/section3/eda35h.htm

In such a matrix, each row stands for a sample. Each sample contains *i* features (genes) from a certain experiment and results as a class (class A or class B). Because of this structure of gene expression matrix, having thousands of features while including low number of samples, gene expression data classification problem is a high dimensional low sample problem. Despite expression levels of a large set of genes are measured, most of these of genes are not directly related with class labels and not needed in classification. A subset of this large dataset contains the mostly related genes. This subset must be extracted before the application of the classification to the samples. Classifying a large gene expression dataset increases the cost of computations because of the addition of unnecessary noise and also decreases the accuracy of the classification and increases the risk of over-fitting [10]. Using a small subset of features brings more accurate diagnosis and the chance to analyze the nature of disease [11]. That transforms the microarray classification problem to optimization problem between minimization of the number of selected features and maximization of the classification accuracy.

This process of extracting the most related features from microarray data, which supplies not only the improved generalization by reducing over-fitting and much shorter training times but also increasing the interpretability of the model, is called gene selection or feature selection. The simplest algorithm, exhaustive search, is to handle every possible subsets of the feature space one by one and determine the one that gives the best classification result. However it is computationally expensive especially for large-sized feature spaces. Therefore different methods and evaluation techniques are used to optimize this problem. In feature selection, there are mainly three groups of these evaluation techniques; filter methods, wrapper methods and embedded methods [12].

In filter methods, features in the space are handled individually. A single feature's quality is measured by use of different statistical procedures, according to the relationship between the feature and the class label. After all features' scores are calculated, the top ranked ones are selected for preparation of the model. Because filter methods are isolated from building of the classification model and they require only *i* calculations -where *i* denotes the size of feature space-, they are the least computationally expensive methods. Note that in filter methods the features, which

are not informative as alone but can be informative with a certain subset, are not taken into consideration.

In wrapper methods, determined subsets (mostly by greedy search strategies) of the features are used in chosen classifiers and the subsets are scored according to their power of prediction. The classifier is used in these methods as black box and generally the ones having low running times are chosen such as; decision trees, naïve Bayes or SVMs etc. They are creating the model from scratch for each specific subset and retrain the classifier. Despite they can handle the features, which are informative only with a subset; wrapper methods are computationally too expensive especially for large feature spaces.

In embedded methods, selecting the features takes place in the creation of the model. They are similar to wrapper methods by this property but they are more efficient in some ways. They do not separate the data as training group and test group. The model is not created from scratch and the classifier is not retrained for every subsets. They are computationally more expensive than filter methods and less expensive than wrapper methods.

In this study, since we have large-sized microarray data sets a filtering method is chosen for feature selection instead of wrapper of embedded ones in order to avoid computational overhead. More specifically, we use Information Gain attribute evaluator for feature evaluation and Ranker for ranking and selection of the features.

2.3.1 Information Gain

The impurity of a set of samples is measured as *entropy*. Given a feature set *S* which contains *c* different class labels, the entropy of this set according to this c-labeled classification is given [13] as $Entropy(S) \equiv \sum_{i=1}^{c} - p_i \, log_2 p_i$ where p_i denotes proportions of class *i* samples respectively in set *S*.

Power of prediction (or significance) of a feature in classifying these samples can be determined as the diminution of the entropy induced by separating the samples due to this certain feature. This is called by *information gain* and defined as in (1);

$$Gain(S,A) \equiv Entropy(S) - \sum_{a \in Values(A)} \frac{|S_a|}{|S|} Entropy(S_a) \quad (1)$$

where *Values(A)* denotes the possible values of feature *A* and S_a denotes the subset of *S* in which feature *A* is valued as *a*. First part of this equation is the entropy of the main set *S* and the second part is sum of entropies of all subsets multiplied by fractions of these subsets in *S* when samples are separated according to the feature *A*'s values.

This *Gain(S, A)* can be used in the experiments as well. However, in this study *information gain ratio* is used which takes into consideration the number and size of the subsets formed by different values of the feature *A*, so that information gain's strong bias can be reduced [15]. This gain ratio is obtained [16] by dividing *Gain(S, A)* by the entropy of the feature value distribution as given (2);

$$GainRatio(S,A) \equiv \frac{Gain(S,A)}{-\sum_{i=1}^{c} \frac{|S_i|}{|S|} log_2 \frac{|S_i|}{|S|}} \quad (2)$$

where the denominator is the entropy of *S* due to different values of feature *A*.

The calculated information gain ratio values for all features in microarray data are ranked and top ranked ones (number differs according to experiment cases) are selected for preparation of the classification model. However, before going into classification methods, it is necessary to mention two main processes applied to gene expression and SLR values.

2.3.2 Principle Component Analysis

According to classifiers used in this study, selecting features depending on their gain ratios may cause some pitfalls. One of them is ignorance of some features because they are not informative individually, while topmost features are selected in

construction of classifiers, especially decision tree and random forests. The other one is specific to decision trees; some possibly useful features may be excluded because of pruning process. In order to handle these pitfalls we use *principle component analysis* (PCA) as a separate case to make reliable comparison in our experiments.

PCA, which is discovered by Karl Pearson in 1901 [62], is a statistical procedure used as dimension reduction technique in high dimensional data mining problems. It does this dimension reduction by transforming the original dataset to a new set of features encapsulating the original features. Newly formed features are called *principle components* and they include the maximal variance. For detailed information about eigenvectors, eigenvalues, how PCA works and how these principle components are derived can be reached via the linked tutorial[10].

In our study, because of the high dimensionality of datasets (>50000 features) and the computational complexity of PCA, we do not apply PCA directly the original data. Firstly we apply our two main preprocess normalization and discretization, secondly we extract the features which do not have single interval as discrete value, lastly we apply PCA to these extracted features. Number of extracted features varies in data groups; ~1000 features for gene expression data in binary classification, ~1800 features for gene expression data in multiclass classification and ~100 features for SLR data. Principle component are formed by use of at least 5 of these features.

2.4 Normalization

In microarray technology, gene expression values of each one of features are calculated relatively. Their raw values are meaningful while comparing features with each other in a single experiment. While handling the gene expression values from different samples, using these raw values can be less meaningful, since they have various ranges from feature to feature. This variety can be caused by not only the variation in the technology, but also the differences among the printed probes or differences among RNA samples.

[10] http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf

In order to overcome this issue, some data transformations are used to have gene expression values stay within the same interval for each feature. This process is called *normalization* and takes place before construction of the model as a part of data preprocessing. In our study, all gene expression values and SLR values are normalized into [-1, 1] interval.

2.5 Discretization

The other data preprocessing procedure is *discretization*, known as transforming the continuous values of the features into discretized values by constructing cut-points or contiguous intervals which cover the values of the features. A discretization process has four main steps [17];

- sorting the continues values of the feature,
- evaluating adjacent intervals for combining or a cut-point for partitioning,
- partitioning or combining the intervals continuous value due to some criterion,
- stopping at some point finally.

Discretization techniques can be described in two main groups, *unsupervised* and *supervised*. In unsupervised discretization, correlation between the feature's values and class label is not considered. Feature's value is partitioned by equal-width interval binning or equal frequency binning in which the number of bins mostly determined by the user. This situation causes a drawback as loss of knowledge when the values are not evenly distributed. Supervised discretization makes use of class membership knowledge during the discretization process. Basically, if a feature has no or very less correlation with the class label supervised discretization sets that feature values to a single interval, giving the opportunity to discard the features which are not informative. Previous studies show that supervised discretization is more useful than unsupervised in classification problems [18].

One of the most widely used supervised discretization technique is entropy based one, in which the class information entropy of possible partitions is used to determine possible cut-points for discretization. It starts with a single wide interval containing all values and recursively splits into subintervals and then stops if a stopping criterion such as minimum description length (MDL) principle is met.

There are two suggestions for this MDL principle. One is Fayyad and Irani's MDL method that is established on determining the most favorable point according to entropy measure and applying recursive binary partitioning on such points [19]. The second one is Kononenko's MDL method that is similar to Fayyad and Irani's, but has a regulation for the bias the entropy measure has towards a feature with many values [20]. In this study, supervised discretization method with Kononenko's MDL criterion is used to discretize both gene expression values and SLR values.

2.6 Classification

The gene expression researches can be classified in four main categories. In first category describing the subtypes of cancer is aimed by clustering. In second category diagnostic and prognostic studies are aimed by classification. In third category possible cancer signatures are determined by feature selection. In fourth and last category modeling the gene regulatory networks is aimed. This study is mainly interested in second category, classification. Among the lots of machine learning methods those are performed for classification, Support Vector Machines (SVM) and Random Forests are one step further than the others since they are robust to high dimensionality of gene expression data [21]. Three types of classification techniques are used in this study, SVM, Decision Tree and Random Forests.

2.6.1 Support Vector Machines

SVMs are supervised learning models used for classification and regression analysis, originally conceived by Vladimir N. Vapnik and transformed to present version by Vapnik and Corinna Cortes in 1995 [22]. In SVM, a non-probabilistic linear classifier is built as a model according to training set of features and used to predict the results of test set of features. Depending on the number of class labels, one or more hyperplanes are constructed in high-dimensional space to separate classes while maximizing the distance between hyperplanes and the nearest vectors –called as *margin*- of the classes. SVM is actually planned for binary classification problems. However it can be used in multiclass problems in two ways; first, using one-versus-all approach in which a hyperplane is constructed for each single class versus other classes, second, using one-versus-one approach in which a hyperplane is

constructed for every pair of two classes. Let's describe SVM on binary classification [23].

Let S be a set of points x_i in d dimensional space R^d where $i = 1, ..., m$. Let y_i {-1, +1} are two class labels which contain the respective x_i points. If there are a vector w in d dimensional space R^d and a constant b where $y_i (\boldsymbol{w} \bullet \boldsymbol{x_i} + b) \geq 1$, the set of x_i points S can be linearly separated. The separating hyperplane of this set S can be defined in terms of w and b as $\boldsymbol{w} \bullet \boldsymbol{x} + b = 0$. It follows that $\frac{1}{||w||}$ is the closest distance from the separating hyperplane to points of classes as shown[11] in Figure 9.

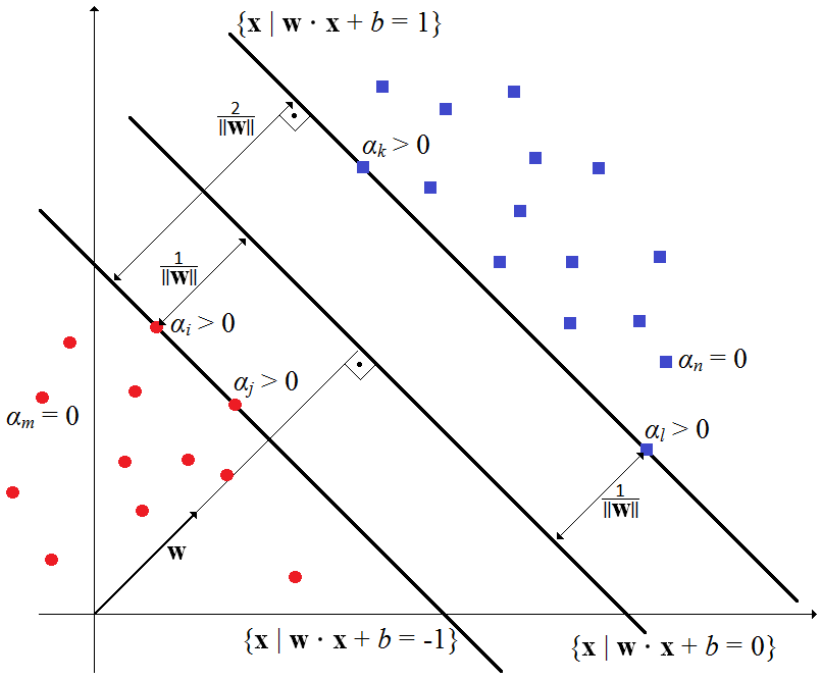

Figure 9: Hyperplane in SVM

In such a linearly separable set of points S, optimal separating hyperplane maximizes $\frac{1}{||w||}$ by maximizing the distance to the closest x_i points from each classes. The optimization problem is minimization of $\frac{1}{2}||w||^2$ with respect to $y_i(\boldsymbol{w} \bullet \boldsymbol{x_i} + b) \geq 1$. This problem needs to be solved by only quadratic optimization methods for the cases dimension is only low ($<10^3$) since it is a constrained quadratic problem. With the help of Lagrange multipliers, α_i, the primal form of this problem can be transformed into coequal dual form as in (3);

[11] https://www.sec.in.tum.de/assets/lehre/ws0910/ml/slideslecture8.pdf

$$L_D(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \, \boldsymbol{x_i} \bullet \boldsymbol{x_j} \qquad (3)$$

The dual form of the optimization problem becomes maximization of L_D (α) with respect to $\sum_{i=1}^{m} y_i \alpha_i = 0$ for $\alpha_i \geq 0$, i = 1, ..., m.

A variant of SVM, Soft-margin SVM is introduced in the case of absence of linearly separating hyperplane. The main idea is, if there is not a hyperplane separating the classes exactly; construct the best one which separates classes as clean as possible. This method brings two new variables; the non-negative slack variables ξ_i measuring the level of the x_i's misclassification and a constant C in order to regulate misclassification cost. The dual form of the problem transforms into maximization of L_D (α) in (3) with respect to $\sum_{i=1}^{m} y_i \alpha_i = 0$ for $0 \leq \alpha_i \leq C$, i = 1, ..., m. C is chosen large when it is aimed to minimize the number of misclassified error and small when it is aimed to maximize the margin $\frac{1}{||w||}$.

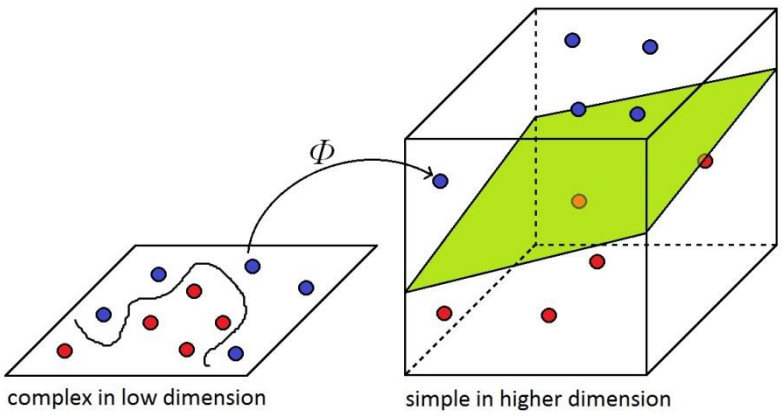

Figure 10: Dimension projection in SVM

Because the dot product is computationally expensive and it is very likely to construct a linear separator in a higher dimensional space, we can map the data

points from the input space R_d into a higher space R_n (where $n > d$) using a function $\Phi : R_d \rightarrow R_n$ as shown[12] in Figure 10.

Thus training algorithm can be described by the dot products of the form $\Phi(x_i) \bullet \Phi(x_j)$. If a *kernel function K* as $K(x_i, x_j) = \Phi(x_i) \bullet \Phi(x_j)$ is defined, this *K* function can be used in training algorithm. There are several kernel functions being used with respect to the classification problem but two popular ones are; *Polynomial kernel function* with the structure $K(x_i, x_j) = (x_i \bullet x_j + 1)^d$ and *Radial bases function (Gaussian)* with the structure $K(x_i, x_j) = e^{-\gamma||x_i - x_j||^2}$. In this study, both kernel functions and sequential minimal optimization (SMO) that is a java solution for soft-margin SVM problem are used.

2.6.2 Decision Tree

This learning type is a supervised classification technique as one of the most widely used methods for machine learning, in which discrete-values target function is denoted by decision tree. It is capable of learning disjunctive expressions and robust to noisy data. In decision trees, a node denotes a test on some feature, a branch of this node denotes a possible value of that feature and a leaf node denotes a class as shown in the decision tree for the concept *PlayTennis* Figure 11 [24].

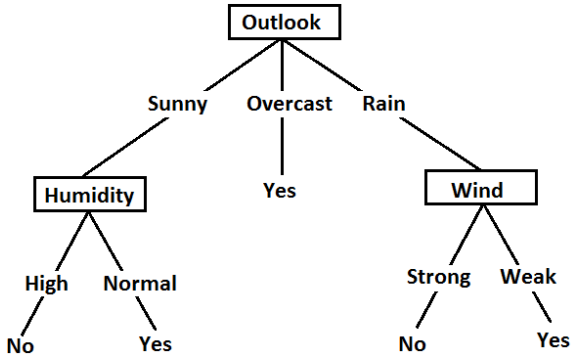

Figure 11: Decision Tree

For classification of a sample, starting from the root node, the feature represented by this node is tested and it is decided to move down over the selected branch due to this

[12] http://prateekvjoshi.files.wordpress.com/2012/08/2d-to-3d-projection.jpeg

test. By considering the new node as new root, this test and move operation is done in new subtree. When the leaf node is reached, the classification of that sample is done. The order of feature nodes of decision tree is determined according to the evaluation function, which the power of decision trees depends on, such as information gain described in 2.3.1.

Most popular decision tree algorithms are *Iterative Dichotomiser 3* (ID3) [25] and its successor C4.5 [26] which are invented and developed by J. R. Quinlan. ID3 has the following algorithm for binary classification (class labels: positive, negative) [27];

ID3 Algorithm

ID3(*Samples*, *TargetFeature*, *Features*)
Samples are training samples, TargetFeature is the feature that is wanted to be classified, Features are set of other features.

- Create a *Root* node for the tree.
- If all *Samples* are positive, return the single-node tree *Root* with label = +
- If all *Samples* are negative, return the single-node tree *Root* with label = –
- If *Features* is empty, return the single-node tree *Root* with label = most common value of *TargetFeature* in Samples.
- Else Begin
  - Let *A* be the feature from *Features* that best (highest information gain) classifies *Samples*
  - Assign decision feature for *Root* to *A*
  - For each possible value, *a*, of *A*,
    - Add a new tree branch below *Root*, corresponding to the test $A = a$
    - Let $Samples_a$ be the subset of Samples that has value a for A
    - If $Samples_a$ is empty
      - Then add a leaf node below this new branch with label = most common value of *TargetFeature* in *Samples*
      - Else add the subtree ID3($Samples_a$, *TargetFeature*, *Features* –{*A*}) below this new branch.
- End
- Return Root

C4.5 has many improvements compared to ID3. It handles both discrete-values and continuous data by extracting the threshold values to split feature values into several intervals [28]. This process is similar to discretization process that we mentioned in section 2.5. C4.5 handles the missing values by treating them as if they are separate values, which are not used in information gain and entropy calculations. It handles the features of differing costs. It solves the problem of over-fitting by use of pruning

to the tree after its construction as shown in Figure 12 [29]. It basically revises the tree and removes the branches which are not helpful to classification by replacing them with class labeled leaf nodes.

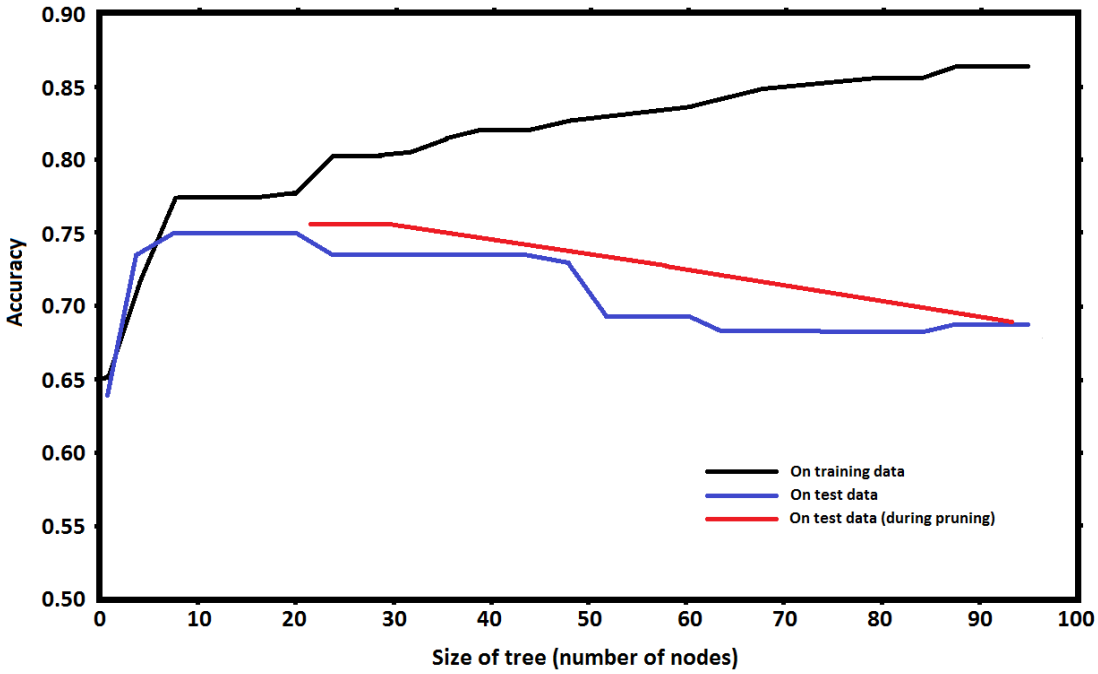

Figure 12: Over-fitting and pruning

Decision trees have many advantages such as; being simple to understand and interpret, not demanding complex data preprocessing, handling large-sized datasets efficiently, being robust to noise data and being able to handle both discrete-valued and continuous data. On the other hand, it has also some drawbacks such as; being based on heuristics since construction of optimal decision tree is a NP-complete problem and possibility of creation of over-complex trees (over-fitting) and requiring pruning to avoid this issue.

In this study a java version of the C4.5 algorithm (J48) is used.

2.6.3 Random Forest

Random forests developed by Leo Breiman [30] as a combination of Breiman's *bagging* idea [31] with Ho's random selection of features which is produced to build a set of decision trees with controlled variance [32][33].

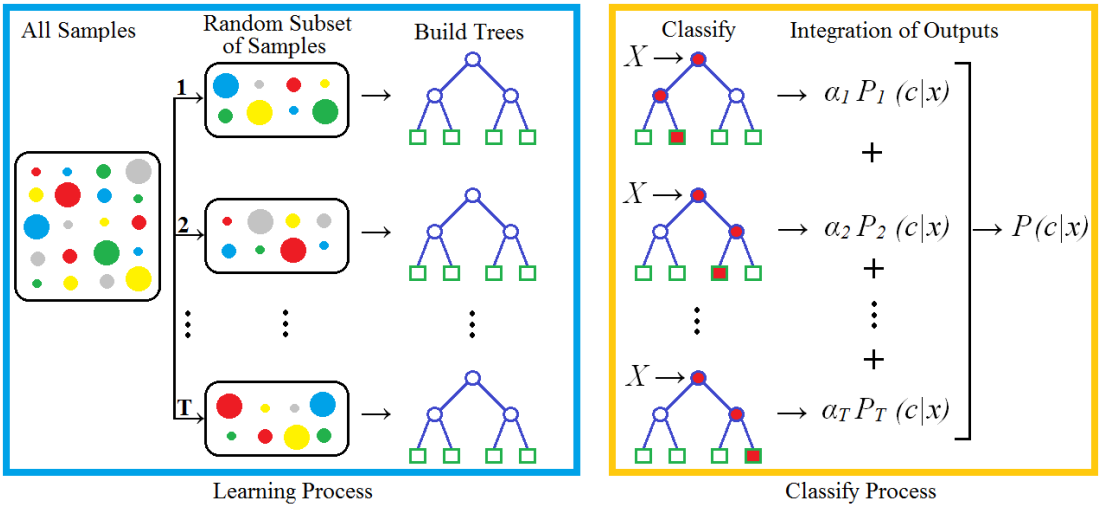

Figure 13: Random forest

Random forests are ensemble classifiers using many decision tree models. In this classifier, a different subset (mostly ~66%) of the training data is selected with replacement and a forest of trees built in which each tree is trained independently and possibly parallel. In order to estimate error and variable importance, remaining training data is used. According to the number of votes from all of the trees, the class label is determined. The learning and classifying processes of random forests are shown in Figure 13.

Every single random tree of the random forest is constructed with a given training set with *N* cases and *M* features as;

- By replacing from the original data *N* cases are sampled out of *S* samples randomly and they are used as training set at construction of tree.
- A number $m << M$ is specified and kept same through construction of tree. For each node, *m* features are chosen randomly from *M* features and the optimum splitting one of these features is used to split this node.
- Each tree is constructed as large as possible by omitting pruning.

According to Breiman [30], two main things determine the error rate of the forest:

- Forest error rate increases if the correlation among individual trees increases.
- Forest error rate decreases if individual trees become stronger classifiers with low individual error rates.

Strength and correlation will decrease while *m* decreases and will increases while *m* increases. Therefore, in random forests *m* is the only parameter needing regulation because of this optimization problem and it can be found quickly by out-of-bag (oob) error rate.

Cross-validation or separating test set from the training set are not needed in random forests so that an unbiased estimate of error is acquired. This error is estimated internally throughout run. About ~33% of the actual data are kept out during the growth of the each individual tree. By use of these left out ones in that built individual trees for classification, for each feature a test set classification and finally the *out-of-bag (oob) error rate* is obtained.

For each node in the tree, by permuting the features values, computing the oob errors, comparing to the original oob error and determining the increase gives the *variable's importance*. Overall oob error rate and variable importance measure are derived from aggregation of oob error and importance measures from all individual trees. The accuracy of forest can be increased by several constructions with only selected features according to their variable importance values.

Missing values are handled in random forests in two approaches. In the first one, missing values are filled by the majority of all observations with observed values. This approach is fast and computationally cheap. In second one, which is computationally more expensive but gives better results, missing values are handled by using surrogate decisions based on additional variables [34].

Random forests have many advantages such as; capability of not only binary classification but also multi-class classification, capability of handling high numbered features and low numbered samples without feature removal, running efficiently for high dimensional datasets with parallelization of processes of individual trees, robustness against irrelevant features and over-fitting and obtainment of variable importance calculated internally.

In this study, an open source java implementation of the Random Forest that sticks to Breiman's arguments [30] is used.

CHAPTER 3

RELATED WORK

As described in classification subtopic, the gene expression researches can be classified in four main categories. In first category describing the subtypes of cancer is aimed by clustering. In second category diagnostic and prognostic studies are aimed by classification. In third category possible cancer signatures are determined by feature selection. In fourth and last category modeling the gene regulatory networks is aimed. In this study our main focus is on the classification subject.

Lots of machine learning methods are performed for classification of gene expression data so far. In order to analyze the relationship between different genes, by ignoring the class labels of the microarray data, initially the unsupervised clustering methods are used. Most widely used ones are; hierarchical clustering, k-means clustering, self-organizing maps (SOM), cluster affinity search technique (CAST) and graph theoretic approaches [35].

By considering the class labels, many supervised learning methods are used to find best solutions for microarray classification problem. Most popular ones are; decision trees, random forests, SVMs, naïve Bayes, artificial neural networks and k-nearest neighbor (KNN). In some cases hybrid methods involving two or more methods are also used such as; hybridized KNN and SVM [36].

Apart from improvements on microarray data analysis, the impact of APA events on gene expressions creates a wide research area [37][38][39]. APA events can affect the gene expressions in three ways [40]. In first one, different protein isoforms that have different physiological features can be coded by alternatively polyadenylated mRNAs and by this way, in higher eukaryotes, APA events increases the protein

variety [41]. In second one, various mRNAs having different 3' UTRs are produced by APA events and many special regions take place in these 3' UTRs such as microRNA target sites [7]. Thus, translation efficiency and mRNA stability is negatively affected. Higher levels of proteins can be caused by mRNAs having shorter 3' UTRs [42]. Shorter 3' UTRs due to APA events may activate the oncogenes in cancer cells [43]. In third one, other gene expression phases can be affected by the variety of the APA events such as; alternative splicing and APA event may work together for their regulations [44]. These effects of APA events on gene expression make it inevitable and vital to understand how they are regulated.

Since a decisive tool or method does not still exist to make transcriptome level analysis of RNA polyadenylation event, APA event researches are precluded. Some of them focused on microarray analyses [45][46][47], in which there are some drawbacks [40]. As first drawback, microarray-based APA research depends on the design and obtainability of microchips. Relying on the microarray platforms, 3' UTR and downstream region coverages may show diversity. As second restriction; since poly(A) sites cannot be identified directly by microarray data, known poly(A) sites databases are used as reference. In third and last limitation; the probes that detect common sections shared by all isoforms and the other probes that detect the extended sections only placed in longer isoforms should be found and then it is needed to calculate the difference in average signal intensities between these two groups of probes. For this reason, analysis of the each APA event becomes so difficult and deceptive if there are more than two APA isoforms. Some previous studies are restricted by the genes which have only two APA isoforms [42][46]. It is mentioned in [40] that, PAS-Seq procedure, a deep sequencing-based method for analysis of RNA polyadenylation, is developed to handle these restriction.

Microarray analysis has a great potential of clinical management of various diseases, especially cancer. It has been used in many experiments for diagnostic, prognostic and even treatment processes of such diseases. Several different approaches are introduced in order to contribute to this potential of microarray analysis. One of the most widely referred one is integrating the clinical data (patient history, age, gender, clinical treatments, etc.), which constitute the basis of clinical decisions and treatments, with microarray data. Various methods and tools are introduced at this

field, clinical and microarray kernel integration (CMKIM) [48], clinical data partition (CliDaPa) [49] and the R package MAclinical [50] are only a few ones which have shown that integrating microarray data with clinical data improved the classifications, predictions and capability of understanding the disease behavior by taking into consideration the common patient behaviors or conditions.

In our case, we introduce a new perspective to these researches by integrating gene expression data with SLR values, which are produced by APADetect to predict APA events by analyzing microarray data. Our aim is to determine whether these SLR values have impact on accuracies of microarray data classification or not by applying on two different datasets.

CHAPTER 4

DATASETS AND INTEGRATION OF GENE EXPRESSION DATA WITH POLYADENYLATION EVENTS

We apply our approach to two datasets deposited into NCBI Gene Expression Omnibus (GEO) [51], GSE29271 and GSE2034. In this chapter these datasets and the integration process of each dataset with the SLR values are explained.

4.1 GSE29271: Expression data from primary breast tumors 2

One of the dataset that we apply our approach is GSE29271, gathered from NCBI GEO database[13]. Because accession GSE29271 is currently not available and scheduled to be released on Jul 10 2015, the gene expression matrix of this set can be accessed from this address[14] with its supplementary file for site of relapses[15]. It has been submitted by the contributions of J. A. Foekens, J. W. Martens and M. Smid in 2011.

As shown in Figure 14, liver metastases are common in breast cancer[16]. This dataset is used to identify the primary breast cancer patients who have tendency to develop liver metastases in order to supply improvement on understanding the breast cancer disease and better treatment for such patients.

This dataset includes microarray data of 210 primary tumors from breast cancer patients who got known site of relapses as liver or elsewhere as shown at Table 2. The sample IDs of dataset is given in (B.1). 61 of these samples have liver metastasis

[13] http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE29271
[14] http://www.ceng.metu.edu.tr/~e1521095/GSE29271_series_matrix.txt.gz (~42 MB)
[15] http://www.ceng.metu.edu.tr/~e1521095/GSE29271_site_of_relapse_info.txt.gz (~1 KB)
[16] http://www.nature.com/nrc/journal/v5/n8/images/nrc1670-f1.jpg

and 149 of them have no or different than liver metastasis. In GSE29271 series matrix, there are 54675 different probe-sets. Depending on the experiment cases, irrelevant ones are ignored according to their prediction power.

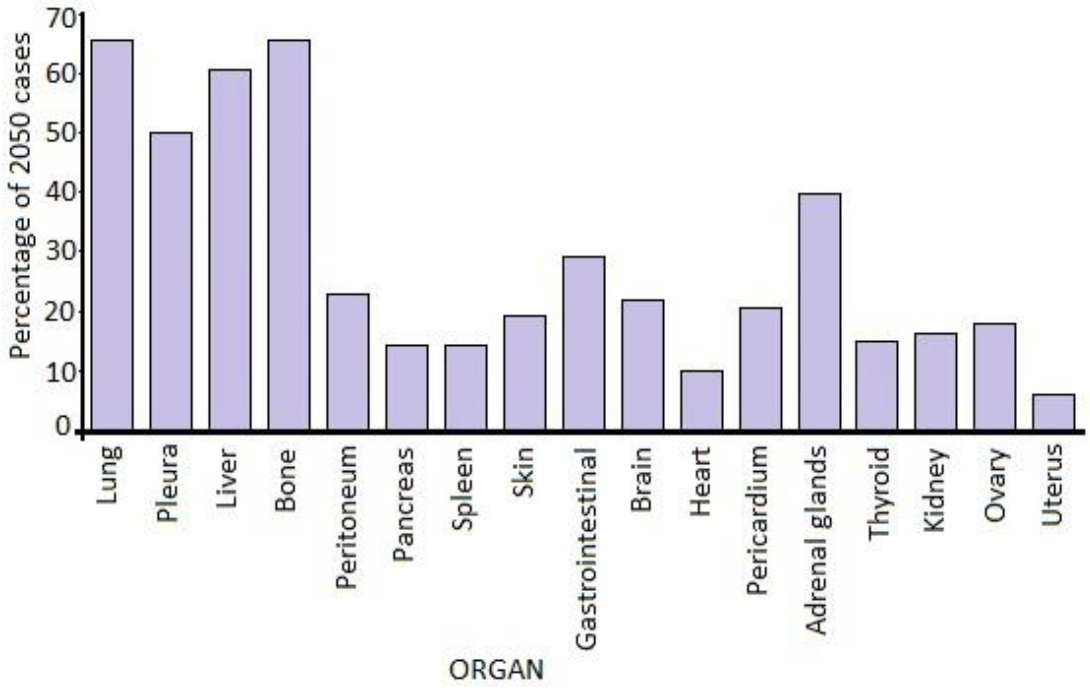

Figure 14: Breast cancer's most common metastasis sites

Table 2: GSE29271 dataset

Total Sample	210
Liver	61
Elsewhere	149
Probe Set Number	54675

For binary classification of this dataset, the class labels of "liver" and "elsewhere" are used as listed in supplementary file site of relapses.

195 out of these 210 sample cases have been subject to another study [52]. In this study's supplementary file[17], the site of relapse information is given in detail for these 195 samples as shown at Table 3. Not all samples are listed in this table. It is used to give idea about the format of detailed site of relapse information. A patient may have no relapse; may have a relapse that is not related with brain, lung or bone;

[17] http://www.nature.com/nature/journal/v459/n7249/extref/nature08021-s2.pdf

may have only one of brain, lung or bone relapse; may have only two of brain, lung or bone relapse; may have all of brain, lung or bone relapse.

Table 3: Detailed site of relapse information for GSE29271

Sample Number	Metastasis-free-survival (months)	All relapses (1=yes, 0=no)	Brain relapses (1=yes, 0=no)	Lung relapses (1=yes, 0=no)	Bone relapses (1=yes, 0=no)
GSM308256	3	1	0	0	1
GSM308257	25	1	0	0	1
GSM308258	23	1	0	0	0
GSM308259	13	1	0	0	1
GSM308260	20	1	0	0	1
GSM308261	18	1	1	1	0
...	...	...	...	...	...

This variety of these relapses gave an idea for multiclass classification. We tagged each one of these cases with four characters that is constructed by “1”s and “0”s as given at Table 4. First character determines whether there is any relapse or not. Second-third and fourth characters determine respectively whether there are brain, lung and bone relapses or not.

Table 4: Class tags and distributions for multiclass classification of GSE29271

Class Label	Site of Relapse	Distribution
0000	No Relapse	10
1000	Relapse - different than Brain, Lung or Bone	38
1001	Relapse - Bone	91
1010	Relapse - Lung	23
1011	Relapse - Lung and Bone	17
1100	Relapse - Brain	9
1101	Relapse - Brain and Bone	2
1110	Relapse - Brain and Lung	4
1111	Relapse - Brain, Lung and Bone	1

The list of site of relapses of GSE29271 can be accessed from this link[18]. As it is mentioned, these class labels are available for 195 samples and are not available for

[18] http://www.ceng.metu.edu.tr/~e1521095/GSE29271_site_of_relapse_info_multiclass.zip

15 samples (B.2). These samples are excluded from multiclass classification. According to these multiclass class labels, the distributions of samples are also given at Table 4 out of 195 samples.

4.1.1 SLR values for GSE29271

As we mentioned before, a sort of filtering processes are applied while the calculation of SLR values by the APADetect tool in order to remove outliers from data. Because of these filtering operations applied while predicting the APA events, SLR values are not available for all the probe sets in a microarray sample. And also SLR values for 15 samples are not available which are not excluded in binary classification but excluded from multiclass classification because of their absence on both gene expression values and SLR values. Sample ID for these samples are same as given in (B.2). There are 195 samples as given at Table 5.

Table 5: SLR values for GSE29271

Total Sample	195
Liver	56
Elsewhere	139
Probe Set Number	2336
Number of Probe Set ID – PolyA Site ID Pairs	2761

There are SLR values for 2336 probe sets. Since there can be more than one poly(A) site for a single probe set, we have SLR values for 2761 "probe set"-"poly(A) site" pairs. In other words we have 2761 features in SLR values. The list of these values for GSE29271 can be accessed from this link[19]. Multiclass labels and their distributions are same as given at Table 4.

4.2 GSE2034: Breast cancer relapse free survival

The other dataset that we apply our approach is GSE2034, gathered from NCBI GEO database[20]. The gene expression matrix of this set can be accessed from this address[21]

[19] http://www.ceng.metu.edu.tr/~e1521095/GSE29271_SLR_values.zip
[20] http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE2034
[21] ftp://ftp.ncbi.nlm.nih.gov/geo/series/GSE2nnn/GSE2034/matrix/GSE2034_series_matrix.txt.gz

and its supplementary file for patient clinical parameters from this address[22]. It has been submitted by the study [53] that aims to estimate distant metastasis events of lymph-node-negative primary breast cancer patients by help of gene expression profiles. According to study, two estrogen receptor (ER) statuses (ER+ and ER-) are defined by a threshold upon amount of estrogen receptors and the patients are cut-off to subgroups according to these statuses. At this study, a gene signature is developed that gives the opportunity to identify the patients at risk of distant metastasis. And it is shown that while identifying who had distant metastasis in five years, gene expression profile forms a reliable source of information.

This dataset includes 286 lymph-node-negative samples as shown at Table 6.

Table 6: GSE2034 dataset

Total Sample	286
ER+	209
ER-	77
Relapse (yes)	107
Relapse (no)	179
Probe Set Number	22283

According to patient clinical parameters, 80 of the samples, which have ER+ status, have distant metastasis and 129 of them do not have. In our study, these 209 samples with ER+ status are used in binary classification with class labels {*erpos_r1*, *erpos_r0*} where erpos_r1 represents having relapse and erpos_r0 represents not having relapse. Sample IDs used in binary classification are given at (B.3).

Again, according to patient clinical parameters, 107 samples of all 286 samples have distant metastasis. These samples are extracted with their "*time to relapse*" data (months). This time information is converted to year forms as shown at Table 7. In our study, these 107 samples are used in multiclass classification with these year forms as class labels {0, 1, 2, 3, 4, 5, 6}. Their distributions out of 107 samples are given at Table 7 too. Sample IDs used in binary classification are given at (B.4).

[22] http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?view=data&acc=GSE2034&id=40089&db=GeoDb_blob26

Table 7: Time to relapse conversion

month(s)	[0,12)	[12,24)	[24,36)	[36,48)	[48,60)	[60,72)	[72,84)
year(s)	0	1	2	3	4	5	6
Distribution	16	28	12	14	11	10	4

4.2.1 SLR values for GSE2034

As it is mentioned in 4.1.1, because of filtering processes there are not available SLR values for every probe set/sample in dataset GSE2034. The sample IDs and patient clinical parameters are same as GSE2034.

Table 8: SLR values for GSE2034

Total Sample	286
ER+	209
ER-	77
Relapse (yes)	107
Relapse (no)	179
Probe Set Number	1592
Number of Probe Set ID – PolyA Site ID Pairs	1892

There are SLR values for 1592 probe sets. Since there can be more than one poly(A) site for a single probe set, we have SLR values for 1892 "probe set"-"poly(A) site" pairs. In other words we have 2761 features in SLR values. The list of these values for GSE2034 can be accessed from this link[23]. Multiclass labels and their distributions are same as given at Table 7.

4.3 Integration of Gene Expression Data with SLR Values

While integrating gene expression values with SLR values, we considered SLR values as completely different features than gene expression features. Instead of merging one SLR value and one gene expression value based on one gene (that is also a hard process and may lead perversive results), we combined with as independent features per sample.

[23] http://www.ceng.metu.edu.tr/~e1521095/GSE2034_SLR_values.zip

When it is thought about a gene series matrix with *N* samples and *M* features. Each sample is represented as a point in *M*-dimensional space. If there exist *K* SLR values; at integration process, the *M*-dimensional space is extending to *M+K*-dimensional space and still each sample is represented as a point in this extended space.

Table 9: Integrated data matrix

	Gene 1	**…**	**Gene M**	**SLR 1**	**…**	**SLR K**	**Class**
Sample 1	3852.528	…	7131.48	1.277	…	1.604	class A
Sample 2	1103.4	…	803.868	?	…	0.234	class A
…	…	…	…	…	…	…	…
Sample n	160.8	…	389.544	-0.0318	…	1.328	class B
Sample n+1	699.888	…	557.952	?	…	?	class B
…	…	…	…	…	…	…	…
Sample N	214.645	…	1964.32	?	…	?	class A

Integrated data matrix has the structure given at Table 9. According to this structure, more than one SLR feature can be the SLR values of same gene feature, however they are treated as different two new features. For instance, for dataset GSE29271, while the feature named 1552765_x_at represents the gene expression of a gene, the new features named 1552765_x_at_Hs.116240.1.27, 1552765_x_at_Hs.116240.1.28 and 1552765_x_at_Hs.116240.1.30 represents the SLR values of the different poly(A) sites (Hs.116240.1.27, Hs.116240.1.28 and Hs.116240.1.30) of the same gene.

The missing SLR values are represented as "*?*" like *Sample2-SLR1* pair. Beside this, there cannot be available SLR values for some samples. Let samples 1, 2, ..., *n* have SLR values and samples *n+1*, ..., *N* do not have SLR values for some $n<N$ as given at Table 9. These missing SLR values for samples *n+1* to *N* are also represented as "*?*".

CHAPTER 5

CLASSIFICATION EXPERIMENTS AND RESULTS

Three classification methods; SVM, J48 and Random Forest are applied to each binary classification and multiclass classification of the datasets. Each experiment case is firstly applied to only gene expression values, then to only SLR values and lastly to the integrated dataset of the gene expression values and SLR values. 10-fold cross validation is used in all cases.

In this thesis, the three classifiers (J48[24], Random Forest[25] and SMO[26]) of popular data mining platform Weka[27] [54] are used with their default parameters. For detailed information about these classifiers' parameters, their Weka API pages can be referenced. Depending on the experiment case, the values in datasets are normalized as described in section 2.4 and/or discretized as mentioned in section 2.5.

Accuracy differences supply limited information for determination of the efficiency of the classification. Because of this reason we extracted true positive (TP) and false positive (FP) rates for experiments, which determine the sensitivity and the specificity of the results. These rates are presented in Appendix A.

5.1 Binary Classification of GSE29271 - Group 1

We have three test groups for binary classification of GSE29271. Group *Only GSE29271* and group *Integrated Data* are tested over 210 samples, group *Only SLR* is tested over 195 samples.

[24] http://weka.sourceforge.net/doc.stable/weka/classifiers/trees/J48.html
[25] http://weka.sourceforge.net/doc.stable/weka/classifiers/trees/RandomForest.html
[26] http://weka.sourceforge.net/doc.stable/weka/classifiers/functions/SMO.html
[27] http://www.cs.waikato.ac.nz/ml/weka/

5.1.1 Experiments for binary classification of GSE29271

Seven different experiment cases are prepared;

1. Values are not normalized, but discretized. 891 topmost features are selected for *Only GSE29271* tests. 102 topmost features are selected for *Only SLR* tests. 993 topmost features are selected for *Integrated Data* tests. While 891 of them come from *GSE29271*, 102 of them come from *SLR* at integration.

2. Values are normalized and discretized. 891 topmost features are selected for *Only GSE29271*. 102 topmost features are selected for *Only SLR*. 993 topmost features are selected for *Integrated Data* tests. While 891 of them come from *GSE29271*, 102 of them come from *SLR* at integration.

3. Values are normalized, but not discretized. 1000 topmost features are selected for all *Only GSE29271*, *Only SLR* and *Integrated Data* tests. At integration, 913 features come from *GSE29271*, 87 features come from *SLR*.

4. Values are normalized, but not discretized. 150 topmost features are selected for all *Only GSE29271*, *Only SLR* and *Integrated Data* tests. At integration, 145 features come from *GSE29271*, 5 features come from *SLR*.

5. Values are normalized, but not discretized. 50 topmost features are selected for all *Only GSE29271*, *Only SLR* and *Integrated Data* tests. At integration, all 50 features come from *GSE29271*.

6. Values are normalized, but not discretized. At first five experiments, features are selected by according to the process: all features from gene expression matrix and SLR matrix are combined, they ordered according to their gain ratios and then topmost informative features are selected from this sorted data. In this experiment a different way is chosen: gene expression values and SLR values are tested alone with their first 50 the most informative features, these features are extracted from their sets and integrated to construct the integrated data with 100 features.

7. Values are normalized and discretized. Features having only one discrete interval are discarded and PCA applied to remained features. Newly formed principle components are ordered and classification techniques are applied respectively to first 2 components, to first 3 components… and to first 100 components.

At first, second and seventh experiments, features discretized as having more than one interval are selected and other features (having one interval) are discarded.

5.1.2 Results for binary classification of GSE29271

According to first six experiment cases, six different result groups acquired as shown at Table 10. Green background in the cell of the result tables implies we have increase on the classification accuracy by integrating gene expression data with SLR values at that experiment and that classifier test. Orange background implies decreasing classification accuracy and no background implies no change. All values of classification accuracies are in percentage form. This style is applied through other result tables, too. At last column the P Value for difference on accuracy results is given which states whether the difference is statistically significant or not.

For this group, in 1st and 2nd experiments integrated data contains 993 features. 891 features (1.63% of 54675) of integrated data come from *Only GSE29271* and 102 features (3.69% of 2761) of integrated data come from *Only SLR*. Participation proportion of SLR features is higher than gene expression features which shows these SLR values are as informative as gene expression values. In 3rd, 4th and 5th experiments 87 features, 5 features and no feature come from *Only SLR* respectively. This shows that, despite SLR values' participation ratio is higher than gene expression values, the topmost informative features still come from *Only GSE29271*.

According to accuracy results, discretized values used in 1st and 2nd experiments provides better accuracy results when compared to other four experiments. This shows that discretization process of continuous values have direct impact on classification accuracies.

Table 10: Results for binary classification of GSE29271

Group 1		Only GSE29271	Only SLR	Integrated Data	P Value
Experiment 1	J48	83,3333	81,5385	82,8571	0.8977
	Random Forest	90,0000	82,5641	87,6190	0,4900
	SVM-Poly(d=1)	99,0476	88,7179	99,0476	1,0000
	SVM-Poly(d=2)	98,5714	89,2308	98,5714	1,0000
	SVM-G(c:10 g:0,0005)	99,0476	92,3077	99,0476	1,0000
Experiment 2	J48	83,3333	81,5385	82,8571	0.8977
	Random Forest	86,6667	84,6154	85,2381	0,6005
	SVM-Poly(d=1)	99,0476	88,7179	99,0476	1,0000
	SVM-Poly(d=2)	98,5714	89,2308	98,5714	1,0000
	SVM-G(c:10 g:0,0005)	99,0476	92,3077	99,0476	1,0000
Experiment 3	J48	67,1429	71,2821	70,4762	0,5161
	Random Forest	77,1429	69,7436	76,1905	0,8086
	SVM-Poly(d=1)	87,1429	73,8462	88,0952	0,7507
	SVM-Poly(d=2)	85,2381	74,3590	89,0476	0,2030
	SVM-G(c:10 g:0,005)	87,1429	74,8718	89,0476	0,5570
Experiment 4	J48	66,1905	71,2821	66,6667	0,8995
	Random Forest	79,0476	74,3590	80,9524	0,5480
	SVM-Poly(d=1)	70,9524	78,9744	70,4762	0,9199
	SVM-Poly(d=2)	75,7143	81,0256	75,7143	1,0000
	SVM-G(c:10 g:0,0045)	79,0476	82,0513	79,0476	1,0000
Experiment 5	J48	67,1429	67,1795	67,6190	0,8909
	Random Forest	83,3333	76,4103	84,2857	0,7886
	SVM-Poly(d=1)	77,6190	82,0513	74,2857	0,4045
	SVM-Poly(d=2)	77,1429	75,8974	77,1429	1,0000
	SVM-G(c:10 g:0,4)	81,9048	82,5641	79,0476	0,3680
Experiment 6	J48	67,1429	67,1795	70,4762	0,3219
	Random Forest	83,3333	76,4103	80,9524	0,4762
	SVM-Poly(d=1)	77,6190	82,0513	86,1905	0,0098
	SVM-Poly(d=2)	77,1429	75,8974	75,7143	0,6676
	SVM-G(c:10 g:0,4)	81,9048	82,5641	83,8095	0,6119

For this group with class labels {liver, elsewhere}, in 10 experiment/classifier pairs out of 30 (cases), accuracies are increased. For J48, accuracies are increased in 3 of 6 experiments. For random forest, accuracies are increased in 2 experiments and decreased in 4 experiments. For SVM, accuracies are increased in 5 cases, decreased in 4 cases and stayed same in 9 cases. Improvements on accuracies in 10 experiment/classifier pairs are mostly related with participation of SLR values as replacement of gene expression values which are less informative than joined SLR values.

Newly constructed J48 decision tree discards less informative features by pruning and contains the topmost informative features from both *Only SLR* and *Only GSE29271*. Differences on decision tree cases are affected by this pruning process.

Decrease on random forest could be explained by construction of random trees by selecting random ~11 features from feature space, which is larger than Only GSE29271 feature space. For instance in 6th experiment, ~83% accuracy is obtained by 50 *Only GE29271* features. After addition of 50 *Only SLR* features, which may include some features less informative than *Only GE29271* features, random tree constructions may include these less informative features randomly. So, the accuracy may be affected by these less informatively constructed random trees' votes and decrease to ~80%. In SVM case, the accuracy was already so close to maximum value by misclassifying only two samples in 1st and 2nd experiments and that caused "no change" on result. While SVM reveals slight changes on accuracies in 3rd, 4th and 5th experiments, at 6th experiment it has ~8.6% improvement by addition of 50 new dimensions from Only SLR.

Three types of kernel functions are used with SVM classifier. These are polynomial kernel (d=1), polynomial kernel (d=2) and Gaussian kernel (optimum c and g parameters due to the experiments). Gaussian kernel gave the highest accuracy results for many cases, which shows that SVM with this kernel function fits to this dataset better than the other classifier/kernel cases.

Accuracy differences are mostly slight changes. If we glance at the P values of differences, these slight changes are not statistically significant except the 6th experiment SVM-Polynomial Kernel (d=1) case, in which the improvement on accuracy is significant with P value 0.0098.

For this group, true positive (TP) and false positive (FP) rates are given at tables placed in (A.1). If we compare *GSE* (only gene expression data) rates with *Integrated Data* rates we can interpret the difference of *liver* metastasis prediction after integration process. In some cases, the sensitivity of this prediction decreased in parallel with accuracy such as random forests in 1st and 2nd experiments (Table 14 Table 15). In some cases, this sensitivity increased despite accuracy slightly decreased such as random forest in 4th experiment (Table 17). In some cases, this sensitivity decreased although accuracy slightly increased such as decision tree in 5th experiment. In some cases, this sensitivity increased in parallel with accuracy such as SVMs in 3rd experiment (Table 16). The largest increases of this sensitivity are

placed in SVM-Polynomial Kernel (d=1) in 5th experiment from 0.361 to 0.607 (Table 18) and in 6th experiment from 0.361 to 0.705 (Table 19). This shows that SLR values improve the power of liver metastasis prediction for some classifier /experiment cases.

The results for 7th experiment are grouped in three diagrams according to the classification techniques. In these diagrams, since 70 principle components are available for SLR, curve of SLR is given only for these principle components. In the first diagram, which is given in Figure 15, J48 decision tree accuracies are presented. According to J48 results, average accuracy result is 91.9% for gene expression data and 94.06% for integrated data. Two outcomes can be derived from these results. One is, while J48 has ~83% accuracy in 2nd experiment, J48 performed better accuracies with PCA. This shows that PCA contributes to decision trees by giving pruned features the opportunity to join classification model. Second outcome is that accuracy results of integrated data are higher than only gene expression data and gap between them gets larger while the number principle component increases.

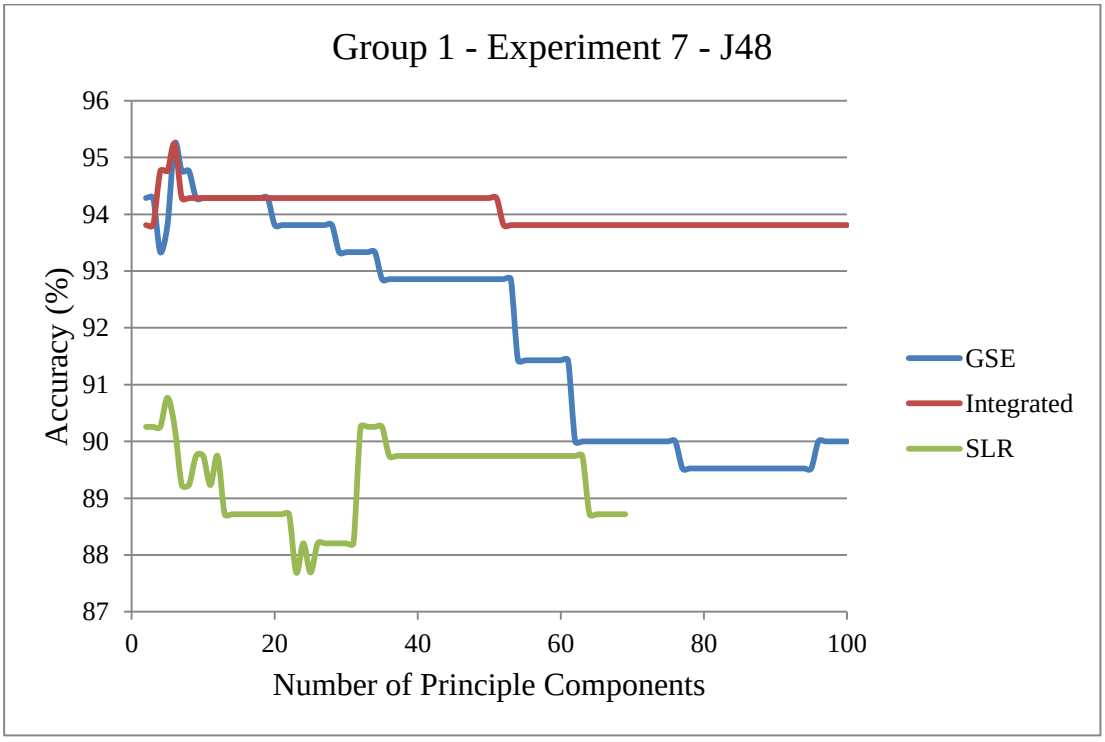

Figure 15: PCA-Accuracy results for Group 1 - J48

In the second diagram, which is given in Figure 16, random forest accuracies are presented. According to these results, average accuracy result is 93.34% for gene expression data and 93.97% for integrated data. While random forest has ~86% accuracy in 2nd experiment, it performed better accuracies with PCA as similar to decision tree. Not only the accuracies but also the differences between only gene expression data accuracies and integrated data accuracies are unsteady when compared to decision tree. The change on accuracy due to the integration process is slightly in favor of increase but not enough to make these differences statistical significant.

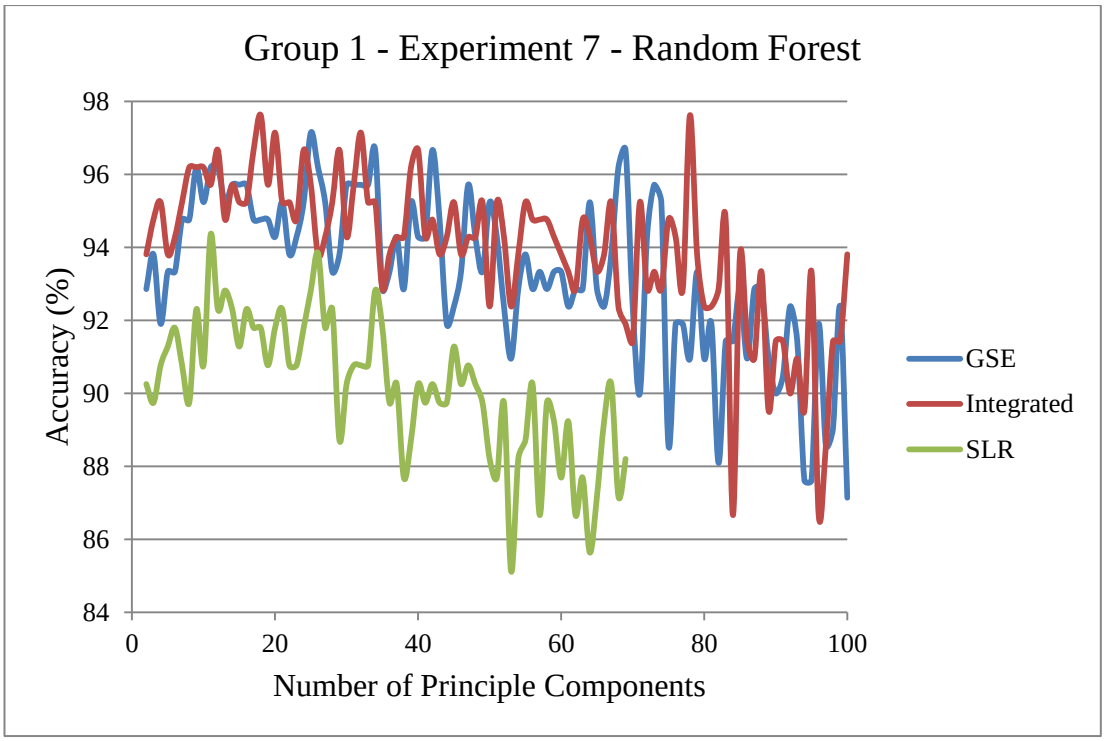

Figure 16: PCA-Accuracy results for Group 1 - Random Forest

In the third diagram, SVM accuracies are given in Figure 17. In SVM case, average accuracy result is 97.96% for gene expression data and 98.35% for integrated data. Two outcomes can be derived from SVM results with PCA. SVM accuracy in 2nd experiment is already close to maximum value by misclassifying two samples since SVM can work with high dimensional datasets well without reduction. One outcome is PCA can cause worse accuracy results for SVM case, by reducing the sensitivity of the dataset as a result of dimension reduction process. Results, especially for low number of components, support this idea. Despite this negative effect; with ideal

component subset, better accuracies can be achieved by PCA, too. With 18, 19, 20, 24, 31 and 35 components, integrated data gives 100% perfect classification. Other outcome is the difference on accuracy due to the integration process is slightly in favor of increase but not enough to be statistical significant as similar to random forest case.

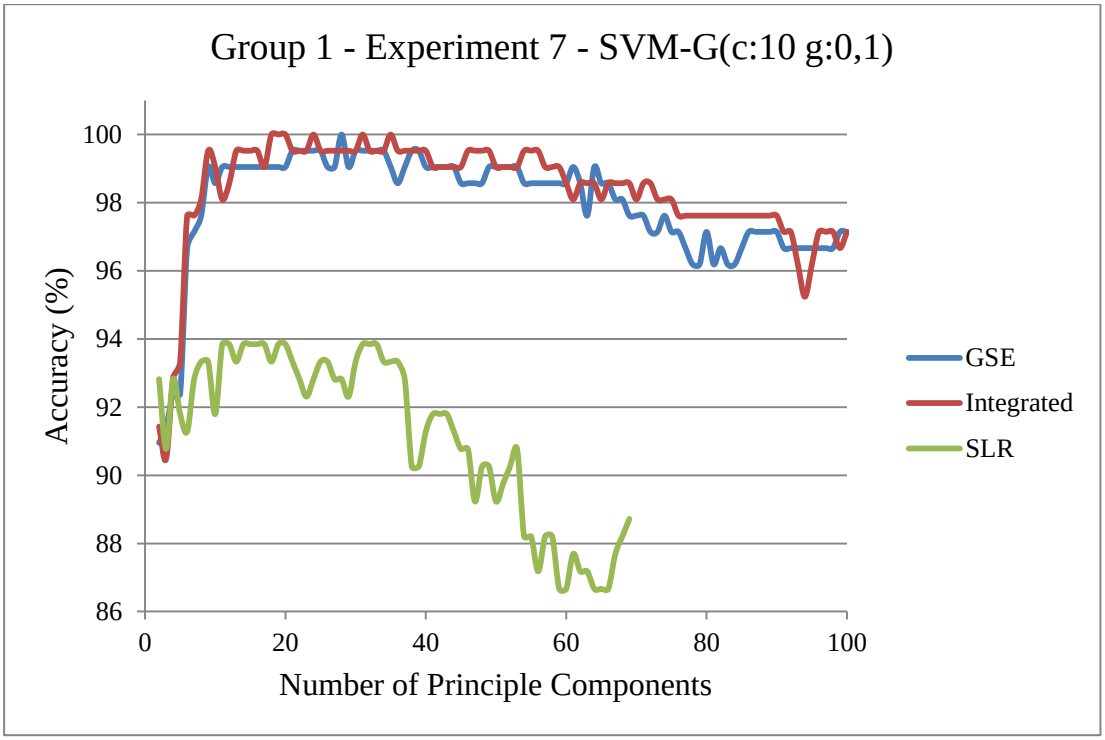

Figure 17: PCA-Accuracy results for Group 1 - SVM

5.2 Multiclass Classification of GSE29271 - Group 2

We have three test groups for multiclass classification of GSE29271 same as binary classification. The difference from binary classification is; all groups *Only GSE29271*, *Integrated Data* and *Only SLR* are tested over 195 samples.

5.2.1 Experiments for multiclass classification of GSE29271

Seven different experiment cases are prepared;

1. Values are not normalized, but discretized. 1788 topmost features are selected for *Only GSE29271* tests. 93 topmost features are selected for *Only SLR* tests.

1881 topmost features are selected for *Integrated Data* tests. While 1788 of them come from *GSE29271*, 93 of them come from *SLR* at integration.

2. Values are normalized and discretized. 1788 topmost features are selected for *Only GSE29271* tests. 93 topmost features are selected for *Only SLR* tests. 1881 topmost features are selected for *Integrated Data* tests. While 1788 of them come from *GSE29271*, 93 of them come from *SLR* at integration.

3. Values are normalized, but not discretized. 1000 topmost features are selected for all *Only GSE29271*, *Only SLR* and *Integrated Data* tests. At integration, 943 features come from *GSE29271*, 57 features come from *SLR*.

4. Values are normalized, but not discretized. 150 topmost features are selected for all *Only GSE29271*, *Only SLR* and *Integrated Data* tests. At integration, 146 features come from *GSE29271*, 4 features come from *SLR*.

5. Values are normalized, but not discretized. 50 topmost features are selected for all *Only GSE29271*, *Only SLR* and *Integrated Data* tests. At integration, all 50 features come from *GSE29271*.

6. Values are normalized, but not discretized. Features are selected like 6^{th} experiment of binary classification. 50 topmost features are selected for *Only GSE29271* and *Only SLR*. These features are integrated as 100 features for *Integrated Data*.

7. Values are normalized and discretized. Features having only one discrete interval are discarded and PCA applied to remained features. Newly formed principle components are ordered and classification techniques are applied respectively to first 2 components, to first 3 components… and to first 100 components.

At first, second and seventh experiments, features discretized as having more than one interval are selected and other features (having one interval) are discarded.

5.2.2 Results for multiclass classification of GSE29271

According to first six experiment cases, six different result groups acquired as shown at Table 11.

Table 11: Results for multiclass classification of GSE29271

Group 2		Only GSE29271	Only SLR	Integrated Data	P Value
Experiment 1	J48	57,4359	49,7436	55,8974	0,7476
	Random Forest	54,3590	52,3077	53,8462	0,9038
	SVM-Poly(d=1)	79,4872	64,1026	80,5128	0,6417
	SVM-Poly(d=2)	78,4615	64,6154	78,4615	0,9903
	SVM-G(c:20 g:0,0007)	77,9487	67,1795	78,4615	0,8029
Experiment 2	J48	57,4359	49,7436	55,8974	0,7476
	Random Forest	55,3846	58,4615	54,8718	0,8712
	SVM-Poly(d=1)	79,4872	64,1026	80,5128	0,6417
	SVM-Poly(d=2)	78,4615	64,6154	78,4615	0,9903
	SVM-G(c:20 g:0,0007)	77,9487	67,1795	78,4615	0,8029
Experiment 3	J48	40,5128	37,4359	36,4103	0,3700
	Random Forest	46,6667	45,6410	44,1026	0,5162
	SVM-Poly(d=1)	42,0513	47,1795	44,6154	0,5659
	SVM-Poly(d=2)	43,5897	45,6410	45,6410	0,5911
	SVM-G(c:10 g:0,005)	47,6923	48,7179	49,7436	0,4585
Experiment 4	J48	37,4359	45,6410	37,4359	1,0000
	Random Forest	45,6410	38,4615	50,7692	0,3084
	SVM-Poly(d=1)	45,6410	48,7179	47,1795	0,6324
	SVM-Poly(d=2)	46,6667	42,5641	47,6923	0,7522
	SVM-G(c:10 g:0,005)	49,7436	49,2308	49,7436	1,0000
Experiment 5	J48	43,5897	46,6667	43,5897	1,0000
	Random Forest	44,1026	46,6667	44,1026	1,0000
	SVM-Poly(d=1)	50,2564	46,6667	50,2564	1,0000
	SVM-Poly(d=2)	48,7179	45,1282	48,7179	1,0000
	SVM-G(c:10 g:0,01)	52,3077	48,2051	52,3077	1,0000
Experiment 6	J48	43,5897	46,6667	50,2564	0,1924
	Random Forest	44,1026	46,6667	43,0769	0,7700
	SVM-Poly(d=1)	50,2564	46,6667	53,8462	0,2107
	SVM-Poly(d=2)	48,7179	45,1282	47,6923	0,7766
	SVM-G(c:10 g:0,01)	52,3077	48,2051	54,3590	0,4445

For this group, in 1st and 2nd experiments integrated data contains 1881 features. 1788 features (3.27% of 54675) of integrated data come from *Only GSE29271* and 93 features (3.37% of 2761) of integrated data come from *Only SLR*. Participation proportion of SLR features is still higher than gene expression features but there is an increase on gene expression values' participation when compared to binary classification. This shows that SLR values are less informative in multiclass classification which is done according to different relapse sites and APA events are

more related with occurrence of relapse than site of relapse at least for this dataset. In 3rd, 4th and 5th experiments 57 features, 4 features and no feature come from *Only SLR* respectively. As in binary classification, the topmost informative features still come from Only *GSE29271*.

For this group with class labels {0000, 1000, 1001, 1010, 1011, 1100, 1101, 1110, 1111}, in 13 experiment/classifier pairs out of 30 cases, accuracies are increased. At this group increases of SVM come into prominence, accuracies are improved in 11 out of 18 cases and only one decrease observed. For J48, increase occurred in 1 experiment, accuracy stayed same in 2 experiments and decreased in 3 experiments. For random forest, improvement exists in 1 experiment, accuracies stayed same in 1 experiment and decreased in 4 experiments. According to accuracy results, discretized values used in 1st and 2nd experiments provides better accuracy results again when compared to other four experiments.

Because SLR values are less informative and show less contribution while integrating, they cause slight decrease on accuracies of mostly J48 and random forest cases or cause no change on accuracies of especially 4th and 5th experiments.

As a remarkable result in this multiclass classification, SVM performed better results than J48 and random forest. Beside this, SVM performed improved results in 11 cases with integrated data. This shows the adaptation power of SVM to newly added dimensions with multiclass classification. As similar to binary classification group, in this group Gaussian kernel gave the highest accuracy results for many cases too. That shows SVM with this kernel function fits to this dataset better than the other classifier/kernel cases.

Accuracy differences are mostly slight changes. If we look at the P values of differences, these slight changes are not statistically significant. The lowest P values are in 6th experiment J48 case with 0.1924 and SVM-Polynomial Kernel (d=1) case with 0.2107. Despite these values are not enough to make them statistically significant, they attract attention by occurring at cases in which accuracies are increased.

For this group, true positive (TP) and false positive (FP) rates are given at tables placed in (A.2). If we compare *GSE* (only gene expression data) rates with *Integrated Data* rates we can interpret the difference of *metastasis site* prediction of breast cancer after integration process. Sensitivity of this prediction can be thought as weighted average of all classes' rates, because sensitivity of a certain metastasis site may increase while sensitivity of another site decreases after integration process. In some cases, the averaged sensitivity of prediction decreased in parallel with accuracy such as random forests and decision trees in 1st, 2nd and 3rd experiments (Table 20 Table 21 Table 22). In 5th experiment, TP and FP rates stayed same as the classification accuracies (Table 24). In 4th and 6th experiments, mostly for SVM cases, sensitivity of prediction slightly increased in parallel with classification accuracies (Table 23 Table 25). Since we do not have significant differences on neither accuracies nor TP/FP rates, we can say that SLR values are less correlated with metastasis site than occurrence of metastasis.

The results for 7th experiment are grouped in three diagrams due to classifiers. In these diagrams, since 67 principle components are available for SLR, curve of SLR is given only for these principle components.

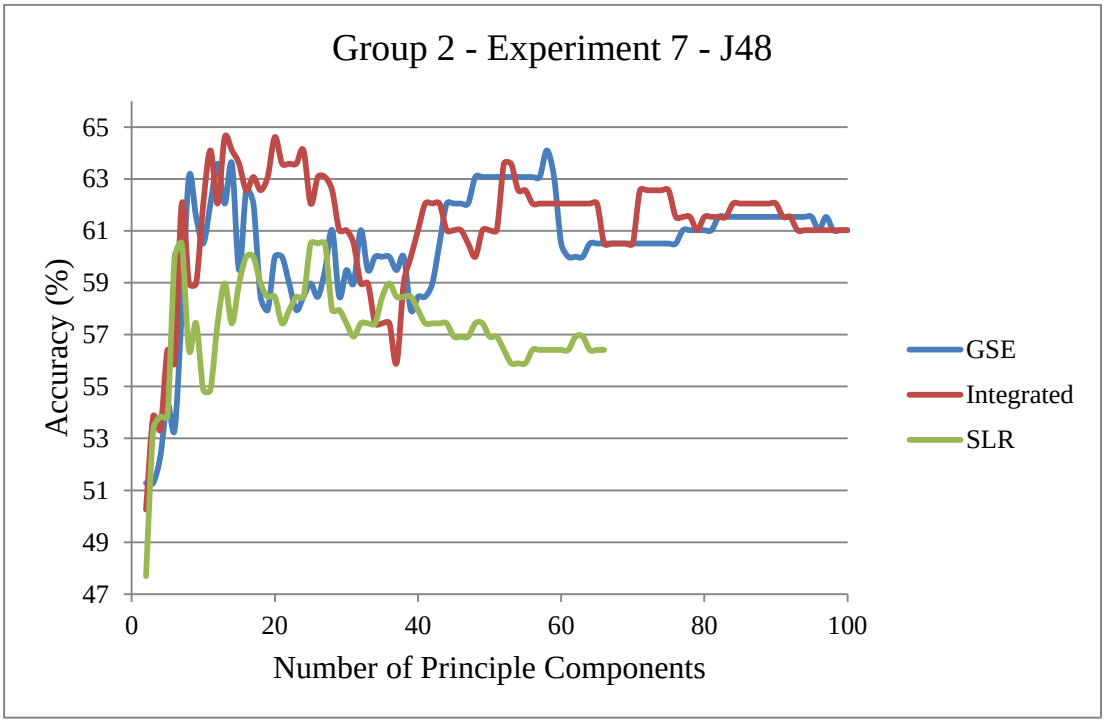

Figure 18: PCA-Accuracy results for Group 2 - J48

In the first diagram, which is given in Figure 18, J48 decision tree accuracies are presented. According to J48 results, average accuracy result is 60.48% for gene expression data and 61.20% for integrated data. While J48 has ~57% accuracy in 2nd experiment, J48 performed better accuracies with PCA. PCA contributes to decision trees in multiclass classification, too. Accuracy results of integrated data are slightly higher than accuracy results of only gene expression data depending on the number of principle components.

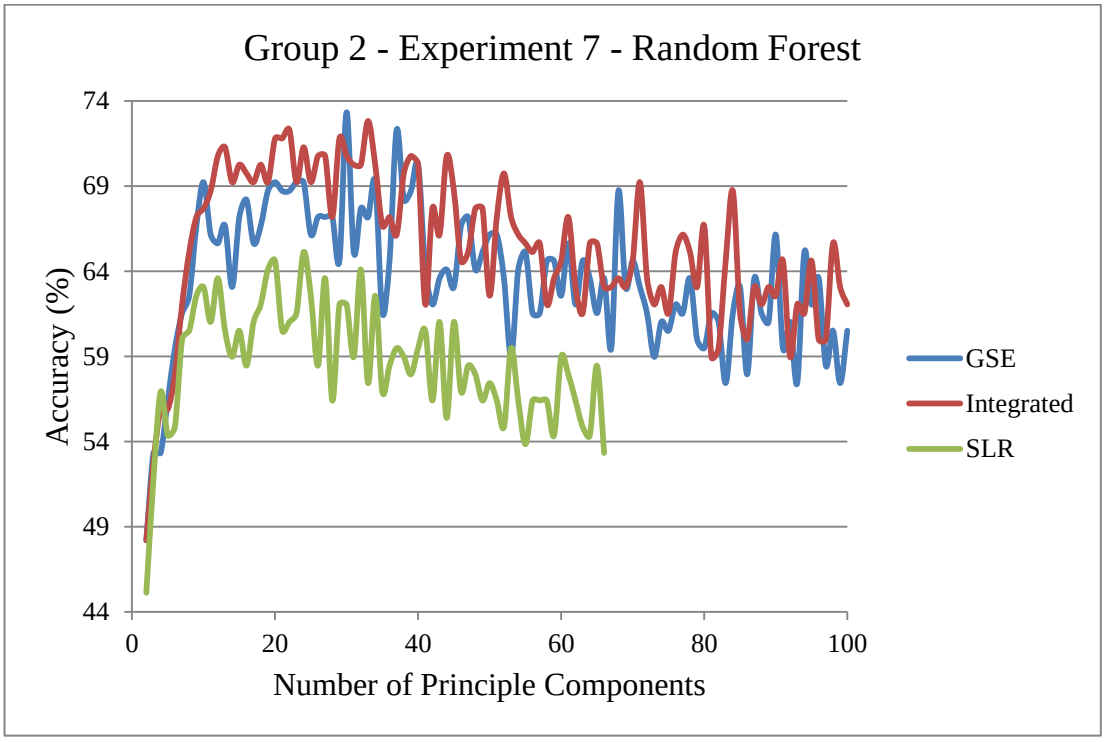

Figure 19: PCA-Accuracy results for Group 2 - Random Forest

In the second diagram, random forest accuracies are given in Figure 19. According to these results, average accuracy result is 63.66% for gene expression data and 65.51% for integrated data. While random forest has ~55% accuracy in 2nd experiment, it performed better accuracies with PCA. The accuracies and also the differences between only gene expression data accuracies and integrated data accuracies are unsteady for this case too. The difference of accuracies due to the integration process is slightly in favor of increase but not enough to be statistical significant.

In the third diagram, SVM accuracies are given in Figure 20. In SVM case, average accuracy result is 71.09% for gene expression data and 72.26% for integrated data.

SVM with PCA has less or equal accuracy results when compared to SVM result of 2nd experiment (~78%). PCA caused worse accuracy results for this case, by reducing the sensitivity of the dataset as a result of dimension reduction process. Especially for low number of components, accuracies decrease as result of this effect. Despite this outcome, the curve of integrated data results in figure shows increased accuracies for most of the cases where the number of principle components varies from 20 to 80.

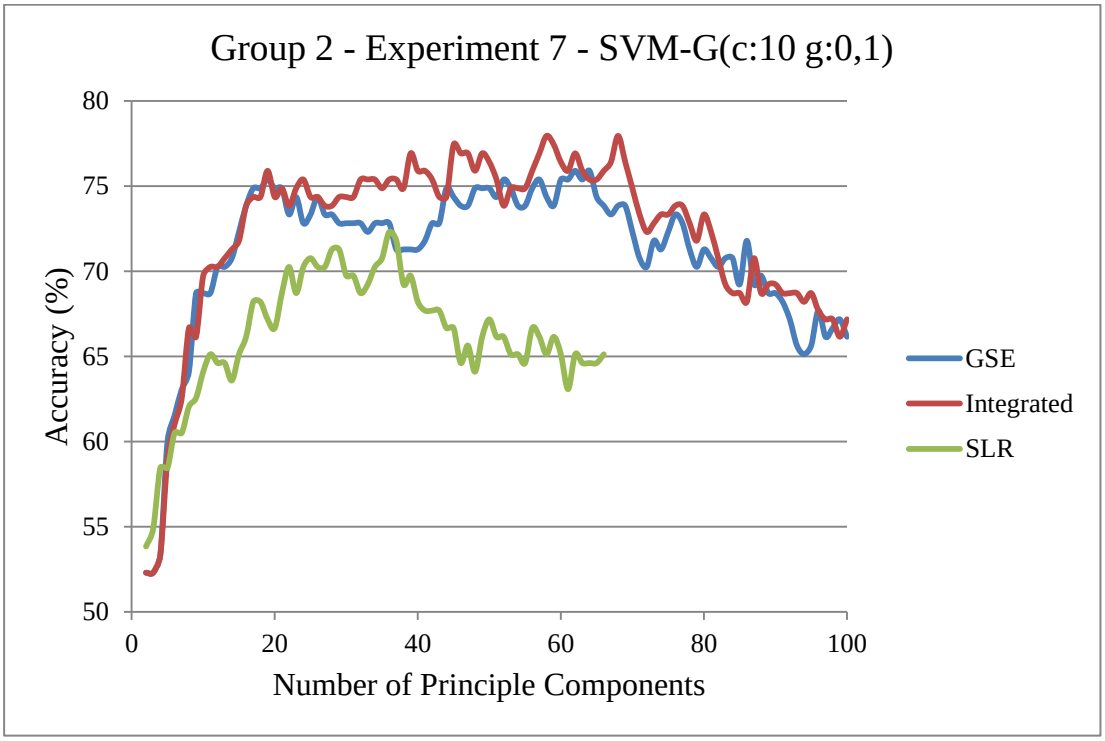

Figure 20: PCA-Accuracy results for Group 2 - SVM

5.3 Binary Classification of GSE2034 - Group 3

We have three test groups for binary classification of GSE2034; group *Only GSE2034*, group *Only SLR* and group *Integrated Data*. All of them are tested over 209 samples.

5.3.1 Experiments for binary classification of GSE2034

Seven different experiment cases are prepared;

1. Values are not normalized, but discretized. 556 topmost features are selected for *Only GSE2034* tests. 88 topmost features are selected for *Only SLR* tests. 644 topmost features are selected for *Integrated Data* tests. While 556 of them come from *GSE2034*, 88 of them come from *SLR* at integration.

2. Values are normalized and discretized. 556 topmost features are selected for *Only GSE2034* tests. 88 topmost features are selected for *Only SLR* tests. 644 topmost features are selected for *Integrated Data* tests. While 556 of them come from *GSE2034*, 88 of them come from *SLR* at integration.

3. Values are normalized, but not discretized. 1000 topmost features are selected for all *Only GSE2034*, *Only SLR* and *Integrated Data* tests. At integration, 923 features come from *GSE2034*, 77 features come from *SLR*.

4. Values are normalized, but not discretized. 150 topmost features are selected for all *Only GSE2034*, *Only SLR* and *Integrated Data* tests. At integration, 147 features come from *GSE2034*, 3 features come from *SLR*.

5. Values are normalized, but not discretized. 50 topmost features are selected for all *Only GSE2034*, *Only SLR* and *Integrated Data* tests. At integration, all 50 features come from *GSE2034*.

6. Values are normalized, but not discretized. Features are selected like 6^{th} experiment of binary classification of previous dataset. 50 topmost features are selected for *Only GSE2034* and *Only SLR*. These features are integrated as 100 features for *Integrated Data*.

7. Values are normalized and discretized. Features having only one discrete interval are discarded and PCA applied to remained features. Newly formed principle components are ordered and classification techniques are applied respectively to first 2 components, to first 3 components… and to first 100 components.

At first, second and seventh experiments, features discretized as having more than one interval are selected and other features (having one interval) are discarded.

5.3.2 Results for binary classification of GSE2034

According to first six experiment cases, six different result groups acquired as shown at Table 12.

Table 12: Results for binary classification of GSE2034

Group 3		Only GSE2034	Only SLR	Integrated Data	P Value
Experiment 1	J48	87,5598	64,5933	88,0383	0,8552
	Random Forest	82,2967	79,9043	82,7751	0,8969
	SVM-Poly(d=1)	98,5646	86,1244	98,0861	0,7505
	SVM-Poly(d=2)	98,0861	84,2105	98,0861	1,0000
	SVM-G(c:10 g:0,001)	98,0861	87,5598	98,0861	1,0000
Experiment 2	J48	87,5598	64,5933	88,0383	0,8552
	Random Forest	82,2967	79,4258	86,1244	0,3266
	SVM-Poly(d=1)	98,5646	86,1244	98,0861	0,7505
	SVM-Poly(d=2)	98,0861	84,2105	98,0861	1,0000
	SVM-G(c:10 g:0,001)	98,0861	87,5598	98,0861	1,0000
Experiment 3	J48	62,2010	61,7225	63,1579	0,7989
	Random Forest	70,3349	56,4593	70,3349	0,9949
	SVM-Poly(d=1)	77,0335	64,5933	80,8612	0,3414
	SVM-Poly(d=2)	76,5550	65,5502	79,9043	0,3457
	SVM-G(c:10 g:0,014)	77,0335	66,0287	80,3828	0,3508
Experiment 4	J48	66,9856	61,7225	66,0287	0,7846
	Random Forest	72,7273	65,0718	76,0766	0,3181
	SVM-Poly(d=1)	78,4689	69,8565	77,0335	0,7536
	SVM-Poly(d=2)	77,9904	68,8995	77,9904	0,9965
	SVM-G(c:10 g:0,005)	80,8612	70,8134	80,8612	1,0000
Experiment 5	J48	69,8565	61,7225	69,3780	0,8970
	Random Forest	75,5981	75,1196	76,5550	0,8233
	SVM-Poly(d=1)	77,0335	74,6411	77,0335	1,0000
	SVM-Poly(d=2)	75,5981	74,6411	75,5981	1,0000
	SVM-G(c:10 g:1,3)	81,3397	75,5981	81,3397	1,0000
Experiment 6	J48	69,8565	61,7225	71,2919	0,6661
	Random Forest	75,5981	75,1196	74,6411	0,7474
	SVM-Poly(d=1)	77,0335	74,6411	78,4689	0,6393
	SVM-Poly(d=2)	75,5981	74,6411	72,7273	0,3861
	SVM-G(c:10 g:1,3)	81,3397	75,5981	79,4258	0,6564

For this group, in 1st and 2nd experiments integrated data contains 644 features. 556 features (2.50% of 22283) of integrated data come from *Only GSE2034* and 88 features (4.65% of 1892) of integrated data come from *Only SLR*. Participation proportion of SLR features is higher than gene expression features in this classification too. In 3rd, 4th and 5th experiments 77 features, 3 features and no feature come from *Only SLR* respectively. As in previous dataset's classification experiments, the topmost informative features come from gene expression values.

For this group with class labels {erpos_r1, erpos_r0}, in 12 experiment/classifier pairs out of 30, accuracies are increased. For J48, accuracies are increased in 4 experiments, stayed same in 1 experiment and decreased in 1 experiment. For random forest, accuracies are improved in 4 experiments, decreased in 1 experiment and stayed same in 1 experiment. For SVM, increase is observed in only 4 cases out of 18. According to accuracy results, discretized values used in 1st and 2nd experiments provides better accuracy results again when compared to other four experiments.

J48 decision tree shows no change in 4th and decrease in 5th experiments. That may be caused by discarding of less informative features from integrated data as a result of pruning and constructing the tree with nearly same selected gene expression values. In other 1st, 2nd, 3rd and 6th experiments J48 shows improved results as a result of replacement of some less informative gene expression values by some more informative SLR values those are not pruned throughout tree construction. Random forest shows noticeable improvement in 4 experiments. Its accuracy decreased in only one case in which it can be caused by construction of random trees by selecting random ~11 features from feature space. For instance, in 6th experiment, the votes of random trees constructed by use of less informative 50 SLR features may affect classification accuracy negatively. For SVM, the accuracy was already so close to maximum value by ~98% in 1st and 2nd experiments, no change or slightly decrease occurred in these cases. Depending on the contribution of SLR values, SVM showed only slight differences on accuracies. Despite it has better results when compared to J48 and random forest results, it is limitedly affected by SLR values in this binary classification experiment.

As similar to GSE29271 classification groups, in this group SVM with Gaussian kernel function fits to dataset better than the other classifier/kernel cases, too.

Slight differences of accuracies are not statistically significant if we consider the P values of differences. The lowest P values are obtained in 4th experiment random forest case with 0.3181, 2nd experiment random forest case with 0.3266, 3rd experiment SVM-Polynomial Kernel (d=1) case with 0.3414 and SVM-Polynomial Kernel (d=2) case with 0.3457. Although these values are not enough to make these

differences statistically significant, these lowest P values are obtained in cases which have increased accuracies.

True positive (TP) and false positive (FP) rates are given at tables placed in (A.3) for this group. If we compare *GSE* (only gene expression data) rates with *Integrated Data* rates we can interpret the difference of *relapse occurrence* prediction of ER+ patients after integration process. In some cases, the sensitivity of this prediction decreased in parallel with accuracy such as SVM-Polynomial Kernel (d=1) in 1st, 2nd and 4th experiments (Table 26 Table 27 Table 29). In some cases, this sensitivity increased despite accuracy slightly decreased such as SVM-Polynomial Kernel (d=2) in 5th experiment (Table 30). In some cases, this sensitivity decreased although accuracy slightly increased such as decision tree in 3th and 6th experiments (Table 28 Table 31). The largest increase of this sensitivity is placed in SVM-Polynomial Kernel (d=1) in 6th experiment from 0.563 to 0.725 (Table 31) in which the P value of accuracy difference is 0.6393. We can see the effect of SLR integration on this group by increase and decreases on classification accuracies for relapse occurrence prediction on ER+ patients, but this effect is far from being statistically significant.

The results for 7th experiment are grouped in three diagrams depending on the classifiers. In these diagrams, since 61 principle components are available for SLR, curve of SLR is given only for these principle components.

In the first diagram, J48 decision tree accuracies are given in Figure 21. According to J48 results, average accuracy result is 86.95% for gene expression data and 87.60% for integrated data. While J48 has ~87% accuracy in 2nd experiment, it performs better accuracies with PCA for some number of principle components. For some others, especially for numbers <5 and >50 principle components, J48 performs worse or equal accuracies. PCA contribution to decision trees depends on the number of principle components used in classification model. Again, depending on the number of principle components, for intervals from 14 to 26 and from 48 to 100 components integrated data performed better accuracies. However these differences (max 2.5%) are not close to be statistically significant.

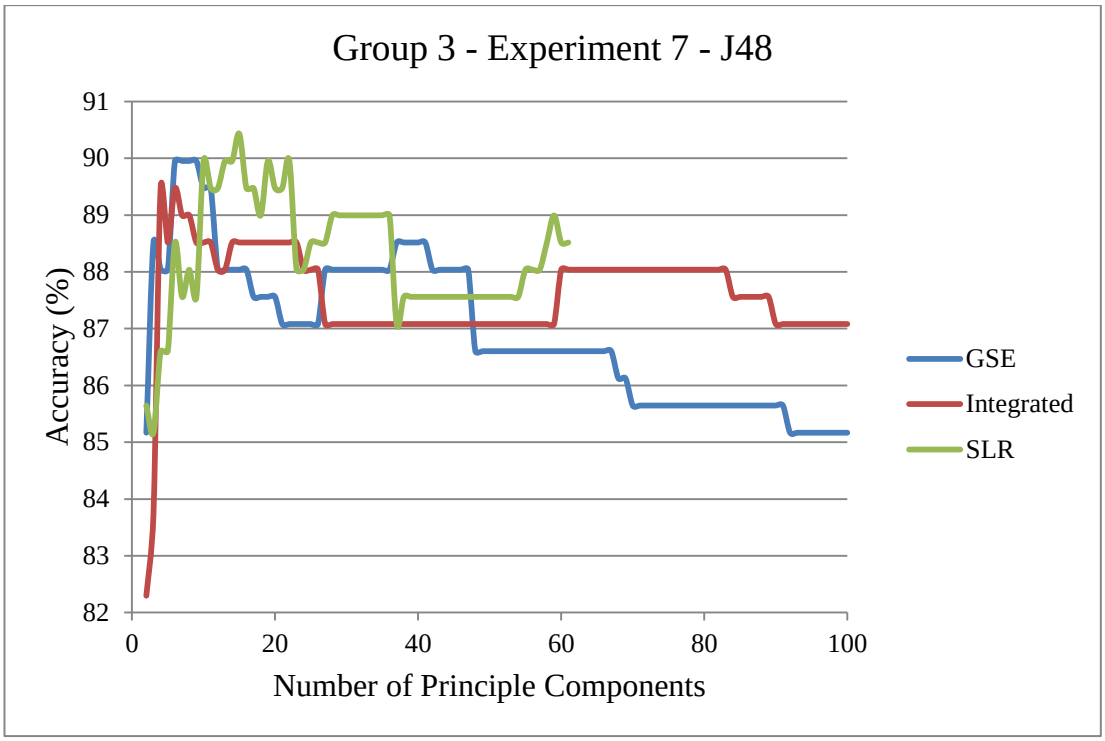

Figure 21: PCA-Accuracy results for Group 3 - J48

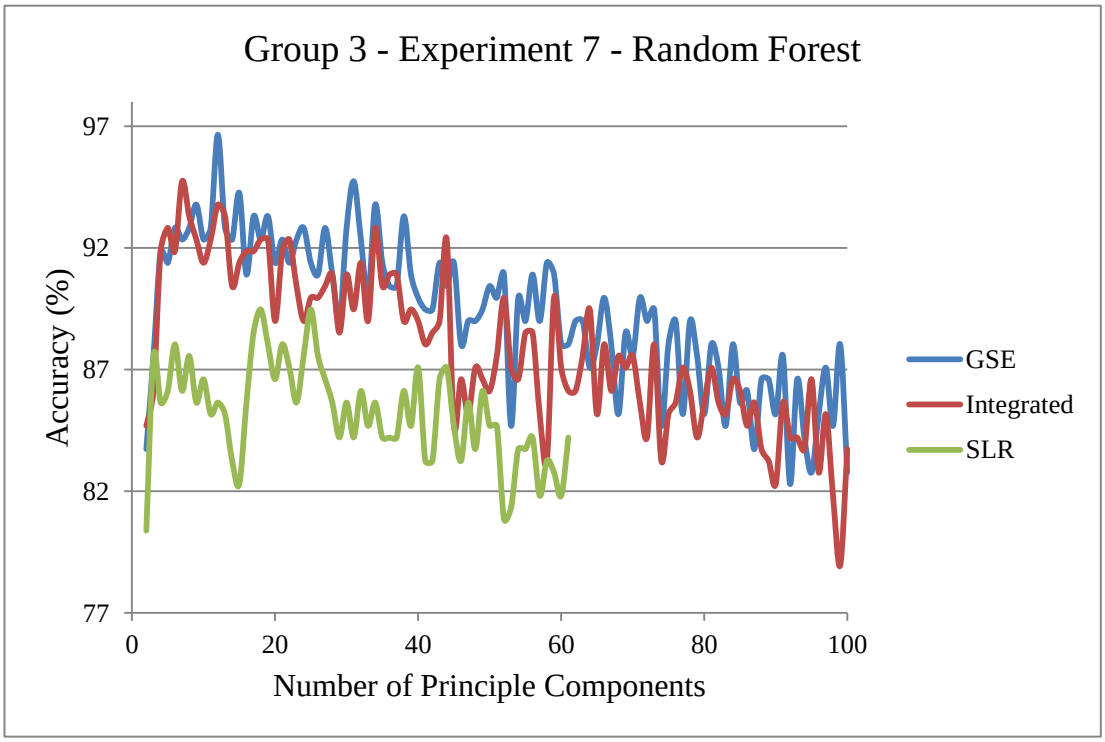

Figure 22: PCA-Accuracy results for Group 3 - Random Forest

In the second diagram, random forest accuracy results are given in Figure 22. According to these results, average accuracy result is 89.35% for gene expression

data and 87.93% for integrated data. While random forest has ~83% ~86% accuracies in 2nd experiment, it performed better accuracies with PCA for less than 50 principle components. Especially for some number of principle components, PCA contributes random forest well and gives high accuracy results. For instance gene expression data reaches 96.65% accuracy with 12 principle components and integrated data reaches 94.74% accuracy with 7 principle components. The accuracies and the differences between only gene expression data accuracies and integrated data accuracies are unsteady as first dataset. The change on accuracy due to the integration process is slightly in favor of decrease but not enough to be statistical significant.

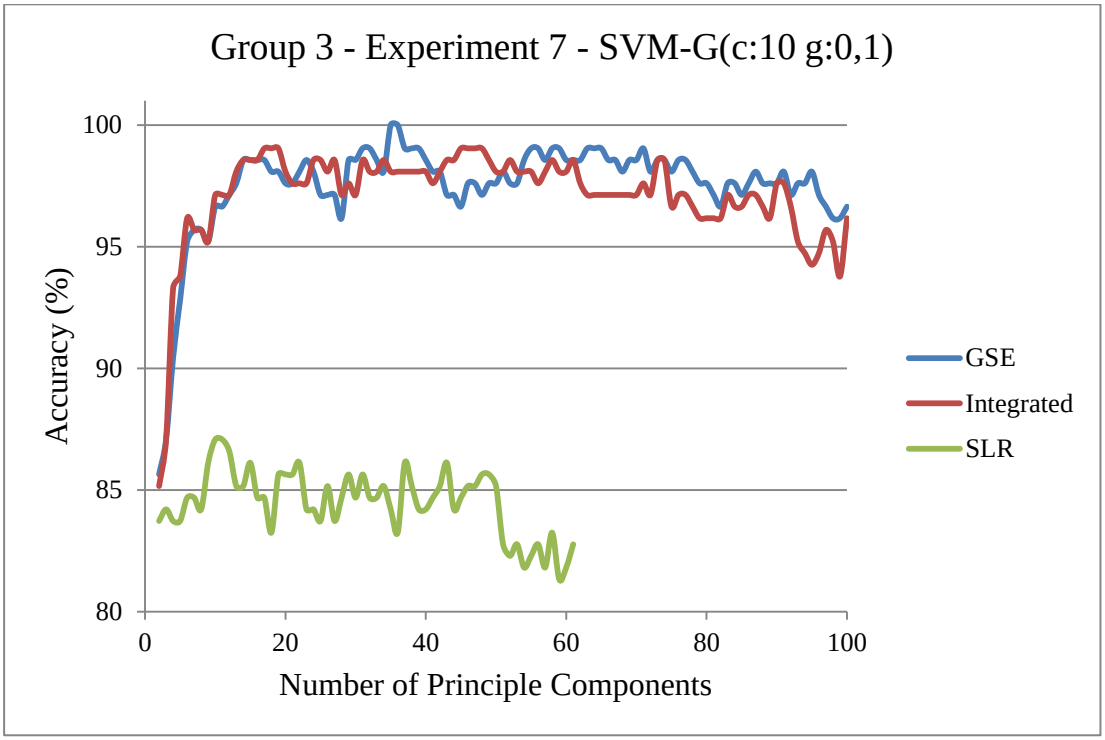

Figure 23: PCA-Accuracy results for Group 3 - SVM

In the third diagram, SVM accuracies are given in Figure 23. In SVM case, average accuracy result is 97.54% for gene expression data and 97.13% for integrated data. SVM with PCA has less or equal accuracy results when compared to SVM results of 2nd experiment (~98%). PCA caused worse or nearly equal accuracy results for this case, by reducing the sensitivity of the dataset as a result of dimension reduction process. Especially for low number of components, accuracies decrease as result of this effect. Despite this outcome, PCA contributes SVM in gene expression data by

giving 100% perfect classification results with 35 and 36 principle components. The difference on accuracy due to the integration process is slightly in favor of decrease but not enough to be statistical significant.

5.4 Multiclass Classification of GSE2034 - Group 4

We have three test groups for multiclass classification of GSE2034 same as binary classification and all of them are tested over are 107 samples.

5.4.1 Experiments for multiclass classification of GSE2034

Seven different experiment cases are prepared;

1. Values are not normalized, but discretized. 1654 topmost features are selected for *Only GSE2034* tests. 173 topmost features are selected for *Only SLR* tests. 1827 topmost features are selected for *Integrated Data* tests. While 1654 of them come from *GSE2034*, 173 of them come from *SLR* at integration.

2. Values are normalized and discretized. 1654 topmost features are selected for *Only GSE2034* tests. 173 topmost features are selected for *Only SLR* tests. 1827 topmost features are selected for *Integrated Data* tests. While 1654 of them come from *GSE2034*, 173 of them come from *SLR* at integration.

3. Values are normalized, but not discretized. 1000 topmost features are selected for all *Only GSE2034*, *Only SLR* and *Integrated Data* tests. At integration, 898 features come from *GSE2034*, 102 features come from *SLR*.

4. Values are normalized, but not discretized. 150 topmost features are selected for all *Only GSE2034*, *Only SLR* and *Integrated Data* tests. At integration, 48 features come from *GSE2034*, 102 features come from *SLR*.

5. Values are normalized, but not discretized. 50 topmost features are selected for all *Only GSE2034*, *Only SLR* and *Integrated Data* tests. At integration, 24 features come from *GSE2034*, 26 features come from *SLR*.

6. Values are normalized, but not discretized. Features are selected like 6th experiment of binary classification of previous dataset. 50 topmost features are selected for *Only GSE2034* and *Only SLR*. These features are integrated as 100 features for *Integrated Data*.

7. Values are normalized and discretized. Features having only one discrete interval are discarded and PCA applied to remained features. Newly formed principle components are ordered and classification techniques are applied respectively to first 2 components, to first 3 components… and to first 100 components.

At first, second and seventh experiments, features discretized as having more than one interval are selected and other features (having one interval) are discarded.

5.4.2 Results for multiclass classification of GSE2034

According to first six experiment cases, six different result groups acquired as shown at Table 13.

For this group, in 1st and 2nd experiments integrated data contains 1827 features. 1654 features (7.42% of 22283) of integrated data come from *Only GSE29271* and 173 features (9.14% of 1892) of integrated data come from *Only SLR*. Participation proportion of SLR features is still higher than gene expression features while participation from both sides increased according to binary classification. In 3rd, 4th and 5th experiments 102 features, 102 features and 26 features come from *Only SLR* respectively. Different from other three classification groups, SLR values are contributed in all experiment cases, even when topmost 50 features selected from integrated data, 26 features come from *Only SLR*.

For this group with labels {1, 2, 3, 4, 5, 6}, in 20 experiment/classifier pairs out of 30, accuracies are increased and they decreased only in 3 of them. The highest proportion of (increase observed case count/decrease observed case count) is acquired from this group with 20/3=6.67. For J48, accuracies are increased in 3 experiments stayed same in 3 experiments and no decrease observed. For random

forest, accuracies are increased in 3 experiments, decreased in 2 experiments and stayed same in 1 experiment. For SVM, accuracies are improved in 14 cases out of 18, stayed same in 3 cases and decreased in only 1 case.

Table 13: Results for multiclass classification of GSE2034

Group 4		Only GSE2034	Only SLR	Integrated Data	P Value
Experiment 1	J48	41,1215	47,6636	41,1215	1,0000
	Random Forest	57,9439	53,2710	59,8131	0,7899
	SVM-Poly(d=1)	99,0654	86,9159	99,0654	1,0000
	SVM-Poly(d=2)	93,4579	87,8505	96,2617	0,3473
	SVM-G(c:10 g:0,0005)	96,2617	87,8505	98,1308	0,4594
Experiment 2	J48	41,1215	47,6636	41,1215	1,0000
	Random Forest	53,2710	56,0748	61,6822	0,2832
	SVM-Poly(d=1)	99,0654	86,9159	99,0654	1,0000
	SVM-Poly(d=2)	93,4579	87,8505	96,2617	0,3473
	SVM-G(c:10 g:0,0005)	96,2617	87,8505	98,1308	0,4594
Experiment 3	J48	25,2336	25,2336	26,1682	0,7941
	Random Forest	22,4299	23,3645	15,8879	0,1744
	SVM-Poly(d=1)	26,1682	29,9065	26,1682	0,9865
	SVM-Poly(d=2)	25,2336	28,0374	31,7757	0,2361
	SVM-G(c:10 g:0,0027)	31,7757	32,7103	30,8411	0,8654
Experiment 4	J48	29,9065	24,2991	38,3178	0,0650
	Random Forest	35,5140	30,8411	29,9065	0,2918
	SVM-Poly(d=1)	27,1028	30,8411	36,4486	0,1753
	SVM-Poly(d=2)	29,9065	24,2991	37,3832	0,2238
	SVM-G(c:10 g:0,036)	35,5140	37,3832	44,8598	0,0539
Experiment 5	J48	38,3178	25,2336	39,2523	0,8854
	Random Forest	29,9065	28,9720	38,3178	0,4081
	SVM-Poly(d=1)	34,5794	38,3178	51,4019	0,0383
	SVM-Poly(d=2)	33,6449	31,7757	43,9252	0,1440
	SVM-G(c:10 g:0,13)	42,9907	41,1215	49,5327	0,4204
Experiment 6	J48	38,3178	25,2336	38,3178	1,0000
	Random Forest	29,9065	28,9720	29,9065	0,9923
	SVM-Poly(d=1)	34,5794	38,3178	42,9907	0,2006
	SVM-Poly(d=2)	33,6449	31,7757	42,0561	0,2429
	SVM-G(c:10 g:0,13)	42,9907	41,1215	45,7944	0,7241

According to accuracy results, discretized values used in 1st and 2nd experiments provides better accuracy results again when compared to other four experiments. Especially ~99% accuracy results of SVM in these experiments differs from other results, which indicates SVM can adapt easily to high dimensional feature spaced multiclass classification problem.

"No change" results of J48 decision tree in 1st, 2nd and 6th experiments may be -again- caused by discarding of less informative features from integrated as a result of

pruning and constructing the tree with same selected gene expression values. In other 3rd, 4th and 5th experiments improvement of J48 results may be caused by replacement of some less informative gene expression values by some more informative SLR values those are not pruned throughout tree construction. Two of our decreased accuracy results belong to random forest and probably caused -again- by construction of random trees with less informative features and their votes. For SVM, the accuracy was so close to maximum value from 96% to 99% in 1st and 2nd experiments and stayed same for polynomial kernel (d=1), slightly increased for polynomial kernel (d=2) and Gaussian kernel. Depending on the contribution of SLR values, SVM showed 8% - 17% improvements on accuracies of 4th, 5th and 6th experiments.

In four groups of classifications, this group takes attention with 20/30 increase cases and only 3/30 decrease cases. This impact of SLR values on accuracy results show itself on P values, too. Differing from other groups, these P values are less than 0.3 in 12 cases and less than 0.07 in 3 cases, which shows that accuracy differences in this group are statistically more significant than other three groups. The lowest P values are obtained in 5th experiment SVM-Polynomial Kernel (d=1) case with 0.0383, 4th experiment SVM-Gaussian Kernel case with 0.0539 and 3rd experiment decision tree with 0.0650.

For this group, true positive (TP) and false positive (FP) rates are given at tables placed in (A.4). By comparing *GSE* (only gene expression data) rates with *Integrated Data* rates we can interpret the difference of *time to relapse* prediction of breast cancer after integration process. Sensitivity of this prediction can be thought as weighted average of all classes' rates as other multiclass classification problem. Sensitivity of a certain *year* may increase while sensitivity of another *year* decreases after integration process.

On one hand, the averaged sensitivity of prediction decreased in parallel with classification accuracy such as random forests in 3rd and 4th experiments (Table 34 Table 35). On the other hand, prediction power of time to relapse classification is increased for many other cases such as SVMs in first six experiments (Table 32 Table 33 Table 34 Table 35 Table 36 Table 37).

By considering *Group 1* experiments and this group, it can be stated that SLR values are much related with relapse times and relapse occurrence than relapse sites, which may be used in related further studies.

The results for 7th experiment are grouped in three diagrams depending on the classifiers. In these diagrams, since 90, 91 and 68 principle components are available for gene expression data, for SLR and for integrated data, curves are given only for these principle components.

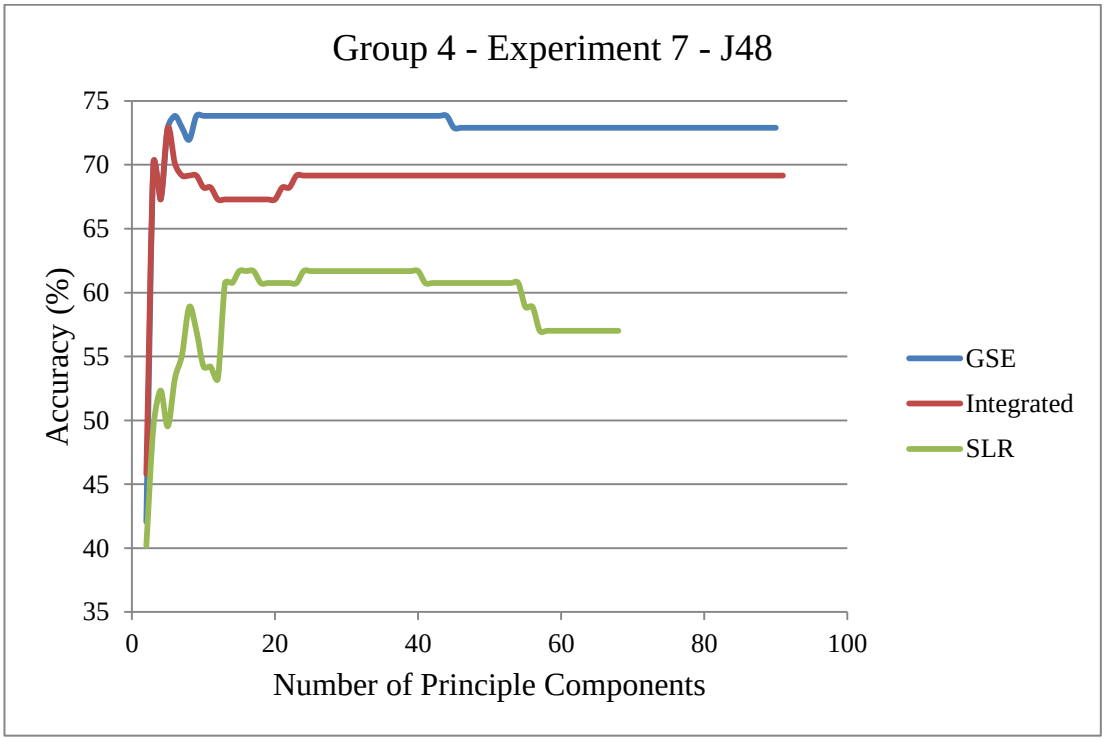

Figure 24: PCA-Accuracy results for Group 4 - J48

In the first diagram, J48 decision tree accuracies are given in Figure 21. According to J48 results, average accuracy result is 72.83% for gene expression data and 68.71% for integrated data. While J48 has ~41% accuracy in 2nd experiment, it performs significantly better accuracies with PCA. PCA contribution to decision tree reaches its maximum level at this experiment case when compared to other three experiment groups. As another outcome, differences on accuracies due to the integration process are slightly in favor of decrease and far from being statistically significant for most of the number of principle components.

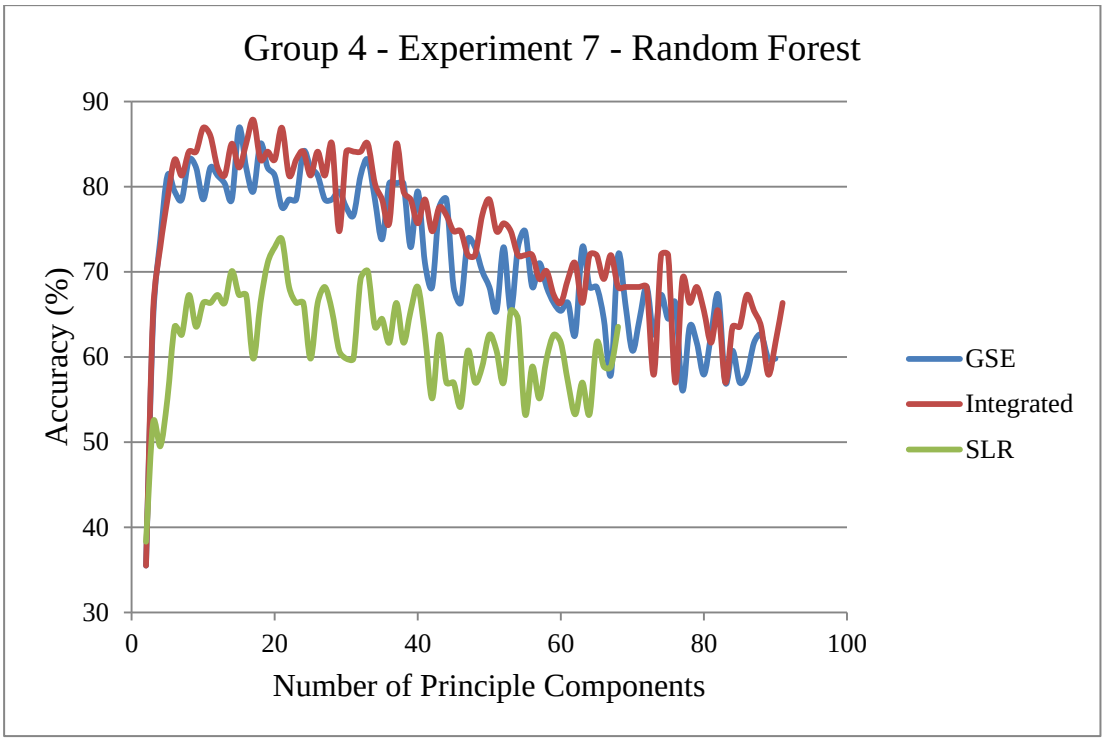

Figure 25: PCA-Accuracy results for Group 4 - Random Forest

In the second diagram, random forest accuracy results are given in Figure 25. According to these results, average accuracy result is 71.51% for gene expression data and 74.10% for integrated data. While random forest has ~53% ~61% accuracies in 2nd experiment, it performed better accuracies with PCA nearly for all numbers of principle components. Especially for some number of principle components, such as 10 to 30, PCA contributes random forest well and gives high accuracy results. For instance gene expression data reaches 85.05% accuracy with 18 principle components and integrated data reaches 86.92% accuracy with 21 principle components. For gene expression data and integrated data, the accuracies and the differences between them are unsteady as happened in other experiment groups. The change on accuracy due to the integration process is slightly in favor of increase but not enough to be statistical significant.

In the third diagram, SVM accuracies are given in Figure 26. In SVM case, average accuracy result is 84.18% for gene expression data and 87.00% for integrated data. SVM with PCA has less or equal accuracy results when compared to SVM result of 2nd experiment (~96%, ~98%). PCA caused worse accuracy results for some number of principle components, by reducing the sensitivity of the dataset as a result of

dimension reduction process. Especially for <10 and >40 number of components, accuracies decreases as result of this effect. Despite this outcome, gene expression data reaches 100% perfect accuracy with 23 principle components and integrated data reaches 100% perfect accuracy with 20, 25, 27 and 28 principle components. As another outcome, the curve of integrated data results in figure shows mostly increased accuracies especially for >50 number of components.

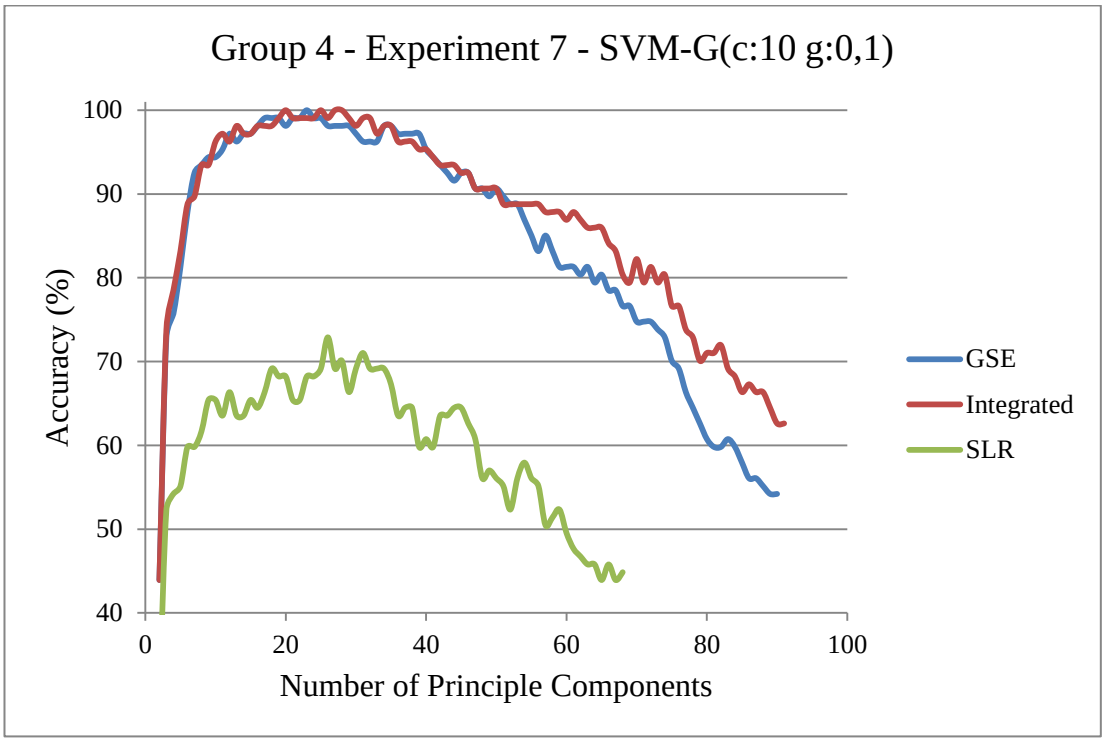

Figure 26: PCA-Accuracy results for Group 4 - SVM

CHAPTER 6

CONCLUSION

6.1 Summary and Discussion

In this thesis, we introduce a new perspective to gene expression data classification analysis by integrating gene expression values with SLR (ratio of differential expressions of proximal and distal probes) values calculated by APADetect. It is mentioned in Ch.3 that APA events have undeniable impact on gene expressions, thereby it is expected that there must be a relation between SLR values and gene expression values as well. Hence we integrated these values together on two different datasets and applied three different classification methods to them in order to identify how these SLR values affect classification accuracies. Integration of data and determining multiclass class labels are other sub problems solved throughout this study.

It is known that ideal feature selection problem is an NP-hard problem and according to literature some heuristics are used to find best way to select features depending on the classification problem. In our case we used information gain ratio, which is computationally cheap and fits to our classification methods. Despite it has some pitfalls such as ignoring the features which are informative only with a certain subset of feature space or the features which are informative only for a certain subset of sample space; it gives satisfactory results when each individual feature is considered alone. According to our experiments when feature selection is applied to integrated data, it is seen that many SLR features took place among the gene expression features which shows SLR features also can be informative depending on the class decisions. In our cases, depending on relapse existence, metastasis type or time to relapse. Especially in multiclass classification for GSE2034, 4^{th} and 5^{th} experiments;

majority of selected features come from SLR values that make SLR values can challenge gene expression values in terms of being informative, especially about *time to relapse* information. When APA events' impact on gene expression values is considered, being informative as much as gene expression values is not unexpected for SLR values. For all four groups, participation proportions of SLR features are higher than gene expression features.

According to results; for all four classification groups, discretized values used in 1st and 2nd experiments provided better accuracy results when compared to other four experiments in their groups. This consolidates the significance of discretization process of continuous values as preprocess before classification.

"Staying same" in accuracy results are mostly related with participation of SLR values into that specific experiment. If any SLR feature cannot join that experiment due to feature selection or pruning or etc., the result tends to stay stable cause of usage of same gene expression features.

"Decrease" in accuracy results of random forest experiments could be explained by construction of random trees by selecting random ~11 features from feature space as described in group results. Random tree constructions may include some newly added less informative features randomly. So, experiment accuracies may be affected by these less informatively constructed random trees' votes. A wise selection of features from both feature sets may lead better accuracies.

If we consider the classifier groups separately; improvements on J48 and Random Forest are mostly took place in binary classifications, while improvements on SVM mostly took place in multiclass classifications. This shows one-vs-one SVM can adapt to multiclass classification problems easily and it is open for improvement by addition of new features/dimensions.

If we consider the 7th experiments for all groups, we can state that PCA adapts well with decision trees and random forests despite it cannot contribute SVM classifier much. PCA's contribution to decision tree and random forests are mostly related with attendance of some pruned or discarded features in the classifier model with newly

created set of principal components. Beside this, PCA gave slightly decreased or equal accuracy results with SVM. SVM itself can adapt high dimensionality of dataset easily and it can handle classification without dimension reduction as it can be seen in Group 4, 1st and 2nd experiments. SVM gets its power from this characteristic. While PCA reduces dimensions, it also reduces the effect of SVM's this characteristic by decreasing sensitivity of the dataset. On the other hand, despite this negative effect SVM can get better accuracies by PCA with ideal principal component subset. Acquired 100% perfect accuracies by SVM with certain number of principle components, shows this result.

We are tried to compare our results with MammaPrint which is used to define a signature of gene expressions that accurately predicts distant metastasis developments within 5 years for the patients who have lymph-node-negative breast cancer [63]. However, dataset of MammaPrint's journal is not gathered from Affymetrix chip which is a prerequisite for obtainment of SLR values by APADetect. Because of dataset difference of these two approaches, we could not directly compare our results with MammaPrint.

When all four groups' results are considered as whole, improvements occurred mostly relating with time to relapse information and relapse occurrence. As stated above, we can express that SLR values, in other words APA events, are much related with relapse times and relapse occurrence than relapse sites.

Despite use of default parameters for classifiers and existence of decreases in some experiments; according to whole results in which there are 55 increased accuracy cases in 120 experiment/classifier pairs, it can be stated that APA events have irrefutable impact on gene expression classification. The P values for accuracy differences are extracted and analyzed in order to decide whether this impact is statistically significant or not. According to these P values, APA events have statistical significant effect on time to relapse classification of breast cancer and for only one or two cases in experiments APA events have this significant effect on relapse occurrence, too.

According to findings about the relationship between APA events and gene expression values, we can conclude that, SLR values produced by APADetect to predict the possible alternative polyadenylation events are highly correlated with gene expression values and the characteristic of the diseases. Accurate classification has a significant role in disease diagnosis and prognosis in order to find and apply best treatment for patients with least side effects. Therefore, we can conclude that, APA events have the potential to lead to more accurate classification results especially for time to relapse information, influencing diagnostic and prognostic decisions of diseases.

6.2 Future Work

SLR values are obtained by leveraging some design issues of Affymetrix chips. From this point, these values have limitations growing with Affymetrix chips' limitations. With a less limited tool or technology focusing on proximal/distal probe densities, more reliable and improved classification accuracies can be obtained.

Whole set of possible poly(A) sites are not fully known yet, that's why we are using the values produced by APA event prediction tool APADetect. Any further improvement on this tool, may affect the results of this study as well.

While we are integrating the datasets with their SLR values, more complicated selection algorithms can be applied. In order to avoid computationally complexity issues, we evaluated each feature individually, ignored the features which are informative only with a certain subset of feature space or the features which are informative only for a certain subset of sample space. We tried PCA in some experiments to reduce the effect of pruning/not selecting some related features in decision tree and random forest constructions. Results shows we managed this duty, especially for Group 4, however, in defiance of computational complexity, a kind of wrapper or embedded feature selection may end up with better results.

The classifiers, except SVM with Gaussian kernel function, are used with their default parameters in this study. Depending on the experimental conditions, use of ideal parameters may conclude with improved results.

REFERENCES

[1] T. R. Golub, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531-7, 15 October 1999.

[2] N. M. Luscombe, D. Greenbaum, M. Gerstein. What is bioinformatics? A proposed definition and overview of the field. *Method Inform Med*, 40(4):346-58, 2001.

[3] F. Crick. Central Dogma of Molecular Biology. *Nature*, 227(5258):561-3, 8 August 1970.

[4] L. W. Barrett, S. Fletcher, S. D. Wilton. Regulation of eukaryotic gene expression by the untranslated gene regions and other non-coding elements. *Cell. Mol. Life Sci.*, 69(21):3613-34, November 2012.

[5] Y. Shen, G. Ji, B. J. Haas, X. Wu, J. Zheng, G. J. Reese, Q. Q. Li. Genome level analysis of rice mRNA 3'-end processing signals and alternative polyadenylation. *Nucleic Acids Research*, 36(9), 3150-61, May 2008.

[6] D. C. Di Giammartino, K. Nishida, J. L. Manley. Mechanisms and consequences of alternative polyadenylation. *Molecular Cell*, 43(6):853-66, 16 September 2011.

[7] B. H. Akman, T. Can, A. E. Erson-Bensan. Estrogen-induced upregulation and 3'-UTR shortening of cdc6. *Nucleic Acids Research*, 40(21):10679-88, 2012.

[8] Y. İlgüner. Prediction of polyadenylation sites by probe level analysis of microarray data (APADetect). Computer Engineering Department, M. Sc. Thesis. Middle East Technical University, Ankara, August 2013.

[9] H. Zhang, J. Hu, M. Recce, B. Tian. PolyA_DB: a database for mammalian mRNA polyadenylation. *Nucleic Acids Research*, 33(Database issue):D116-20, 1 Jan 2005.

[10] M. Xiong, W. Li, J. Zhao, L. Jin, and E. Boerwinkle. Feature (gene) selection in gene expression-based tumor classification. *Molecular Genetics and Metabolism*, 73(3):239-47, July 2001.

[11] Z. Y. Wang, V. Palade, Y. Xu. Neuro-Fuzzy Ensemble Approach for Microarray Cancer Gene Expression Data Analysis. *Proceedings of the 2006 International Symposium on Evolving Fuzzy Systems (IEEE)*, 7-9:241-6, 2006.

[12] I. Guyon, A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157-82, 2003.

[13] Mitchell T. M. *Machine Learning*, McGraw-Hill, Inc. New York, 1st edition, ISBN: 070428072, Ch.3 p.57, 1997.

[14] Mitchell T. M. *Machine Learning*, Ch.3 p.58.

[15] C. Borgelt, J. Gebhardt, R. Kruse. Concepts of probabilistic and possibilistic induction of decision trees on real world data. *Proceedings of the 4th European Congress on Fuzzy and Intelligent Technologies*, 3:1556-60, 1996.

[16] Mitchell T. M. *Machine Learning*, Ch.3 p.74.

[17] S. Kotsiantis, D. Kanellopoulos. Discretization Techniques: A recent survey, GESTS International Transactions on Computer Science and Engineering, 32 (1):47-58, 2006.

[18] J. Dougherty, R. Kohavi, M. Sahami. Supervised and Unsupervised Discretization of Continuous Features. *Proceedings of the 12th International Conference on Machine Learning*, San Francisco, 194-202, 1995.

[19] U. Fayyad, L. Irani. Multi-interval discretization of continuous-valued attributes for classification learning. *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, 1022-7, 1993.

[20] I. Kononenko. On biases in estimating multi-valued attributes. *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 1034-40, 1995.

[21] J. L. Lustgarten, V. Gopalakrishnan, H. Grover, S. Visweswaran. Improving classification performance with discretization on biomedical datasets. *Proceedings of the Fall Symposium of the American Medical Informatics Association*, 445-9, 6 November 2008.

[22] C. Cortes, V. N. Vapnik. Support-vector networks. *Machine Learning*, 20(3): 273-97, September 1995.

[23] Christopher J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2:121-167, 1998.

[24] Mitchell T. M. *Machine Learning*, Ch.3 p.53.

[25] J. R. Quinlan. Induction of decision trees. *Machine Learning 1*, 81-106, 1986.

[26] J. R. Quinlan. C4.5: Programs for machine learning. *Machine Learning 16*, 235-40, 1994.

[27] Mitchell T. M. *Machine Learning*, Ch.3 p.56.

[28] J. R. Quinlan. Improved use of continuous attributes in C4.5. *Journal of Artificial Intelligence Research*, 4:77-90, 1996.

[29] Mitchell T. M. *Machine Learning*, Ch.3 p.70.

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

[31] L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123-40, 1996.

[32] T. K. Ho. Random Decision Forests. *Proceedings of the 3rd International Conference on Document Analysis and Recognition*, 278-82, 1995.

[33] T. K. Ho. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):832-844, 1998.

[34] L. Breiman, J. Friedman, C. J. Stone, R. A. Olshen. *Classification and regression trees.* Chapman & Hall/CRC, 1st edition, ISBN: 0412048418, 1984.

[35] R. Sharan, R. Elkon, R. Shamir. Cluster analysis and its applications to gene expression data. *Ernst Schering Res Found Workshop*, (38):83-108, 2002.

[36] L. H. Lee, C. H. Wan, T. F. Yong, H. M. Kok. A review of nearest neighbor-support vector machines hybrid classification models. *Journal of Applied Sciences*, 10(17):1841-58, 2010.

[37] C. S. Lutz. Alternative polyadenylation: a twist on mRNA 3' end formation. *ACS Chem Biol.*, 3(10):609-17, 17 October 2008.

[38] S. Millevoi, S. Vagner. Molecular mechanisms of eukaryotic pre-mRNA 3' end processing regulation. *Nucleic Acids Research*, 38(9):2757-74, 2010.

[39] J. R. Neilson, R. Sandberg. Heterogeneity in mammalian RNA 3' end formation. *Exp Cell Res.*, 316(8):1357-64, 1 May 2010.

[40] P. J. Shepard, Eun-A Choi, J. Lu, et al. Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. *RNA*, 17:761-72, 2011.

[41] Y. Takagaki, R. L. Seipelt, M. L. Peterson, J. L. Manley. The polyadenylation factor CstF-64 regulates alternative processing of IgM heavy chain pre-mRNA during B cell differentiation. *Cell*, 87(5):941-52, 1996.

[42] R. Sandberg, J. R. Neilson, A. Sarma, P. A. Sharp, C. B. Burge. Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites. *Science*, 320(5883):1643-47, 2008.

[43] C. Mayr, D. P. Bartel. Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell*, 138(4):673–84, 2009.

[44] E. T. Wang, R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S. F. Kingsmore, G. P. Schroth, C. B. Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470-6, 2008.

[45] S.W. Flavell, T. K. Kim, J. M. Gray, D. A. Harmin, M. Hemberg, E. J. Hong, E. M. Papadimitriou, D. M. Bear, M. E. Greenberg. Genome-wide analysis of MEF2 transcriptional program reveals synaptic target genes and neuronal activity-dependent polyadenylation site selection. *Neuron*, 60(6):1022-38, 2008.

[46] Z. Ji, B. Tian. Reprogramming of 3' untranslated regions of mRNAs by alternative polyadenylation in generation of pluripotent stem cells from different cell types. *PLoS One*, 4(12), 2009.

[47] Z. Ji, J. Y. Lee, Z. Pan, B. Jiang, B. Tian. Progressive lengthening of 3' untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. *Proc Natl Acad Sci.*, 106(17):7028-33, 2009.

[48] A. Daemen, O. Gevaert, B. De Moor. Integration of clinical and microarray data with kernel methods. *Conf Proc IEEE Eng Med Biol Soc.*, 5411-5, 2007.

[49] S. Gonzalez, L. Guerra, V. Robles, J. M. Pena, F. Famili. CliDaPa: A new approach to combining clinical data with DNA microarrays. *Intelligent Data Analysis*, 14(2):207-23, 2010.

[50] A. L. Boulesteix, C. Porzelius1, M. Daumer. Microarray-based classification and clinical predictors: on combined classifiers and additional predictive value. *Bioinformatics*, 24(15):1698-1706, 2008.

[51] T. Barrett, D. B. Troup, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, M. Holko, A. Yefanov, H. Lee, N. Zhang, C. L. Robertson, N. Serova, S. Davis, A. Soboleva. NCBI GEO: archive for functional genomics data sets-update. *Nucleic Acids Research*, 41(D1):D991-D995, January 2013.

[52] P. D. Bos, X. H. Zhang, C. Nadal, W. Shu, R. R. Gomis, D. X. Nguyen, A. J. Minn, M. J. van de Vijver, W. L. Gerald, J. A. Foekens, J. Massagué. Genes that mediate breast cancer metastasis to the brain. *Nature*, 459(7249):1005-9, 18 Jun 2009.

[53] Y. Wang, J.G. Klijn, Y. Zhang, A. M. Sieuwerts et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*, 365(9460):671-9, 2005.

[54] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten. The WEKA Data Mining Software: An Update, SIGKDD Explorations, 11(1), 2009.

[55] H. Auer, D. L. Newsom, and K. Kornacker. Expression profiling using affymetrix genechip microarrays. *Methods in Molecular Biology*, 509:35-46, 2009.

[56] J. Khan, J. S. Wei, M. Ringner, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson, P. S. Meltzer. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine*, 7:673-9, 2001.

[57] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, D. Haussler. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10):906-14, 2000.

[58] E. Marchiori, M. Sebag. Bayesian learning with local support vector machines for cancer classification with gene expression data. *Applications of Evolutionary Computing Lecture Notes in Computer Science*, 3449:74-83, 2005.

[59] C. Shi and L. Chen. Feature dimension reduction for microarray data analysis using locally linear embedding. *The Asia Pacific Bioinformatics Conference*, 211-7, 2005.

[60] L. Li, C. R. Weinberg, T. A. Darden, and L. G. Pedersen. Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. *Bioinformatics*, 17(12):1131-42, 2001.

[61] http://en.wikipedia.org/wiki/Bioinformatics as accessed on, July 2014.

[62] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(11):559-72, 1901.

[63] L. J. van't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, S. H. Friend. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871):530-6, 2002.

APPENDIX A

TRUE POSITIVE (TP) AND FALSE POSITIVE (FP) RATES

A.1 TP & FP Rates for Binary Classification of GSE29271 - Group 1

Table 14: TP and FP rates for Group 1 - Experiment 1

Group 1 Experiment 1	GSE		SLR		Integrated data		Class
	TP Rate	FP Rate	TP Rate	FP Rate	TP Rate	FP Rate	
J48	0.656	0.094	0.571	0.086	0.656	0.101	liver
	0.906	0.344	0.914	0.429	0.899	0.344	elsewhere
Random Forest	0.754	0.04	0.518	0.05	0.721	0.06	liver
	0.96	0.246	0.95	0.482	0.94	0.279	elsewhere
SVM-Poly(d=1)	0.967	0	0.75	0.058	0.967	0	liver
	1	0.033	0.942	0.25	1	0.033	elsewhere
SVM-Poly(d=2)	0.967	0.007	0.768	0.058	0.967	0.007	liver
	0.993	0.033	0.942	0.232	0.993	0.033	elsewhere
SVM-G(c:10 g:0,0005)	0.967	0	0.839	0.043	0.967	0	liver
	1	0.033	0.957	0.161	1	0.033	elsewhere

Table 15: TP and FP rates for Group 1 - Experiment 2

Group 1 Experiment 2	GSE		SLR		Integrated data		Class
	TP Rate	FP Rate	TP Rate	FP Rate	TP Rate	FP Rate	
J48	0.656	0.094	0.571	0.086	0.656	0.101	liver
	0.906	0.344	0.914	0.429	0.899	0.344	elsewhere
Random Forest	0.705	0.067	0.571	0.043	0.672	0.074	liver
	0.933	0.295	0.957	0.429	0.926	0.328	elsewhere
SVM-Poly(d=1)	0.967	0	0.75	0.058	0.967	0	liver
	1	0.033	0.942	0.25	1	0.033	elsewhere
SVM-Poly(d=2)	0.967	0.007	0.768	0.058	0.967	0.007	liver
	0.993	0.033	0.942	0.232	0.993	0.033	elsewhere
SVM-G(c:10 g:0,0005)	0.967	0	0.839	0.043	0.967	0	liver
	1	0.033	0.957	0.161	1	0.033	elsewhere

Table 16: TP and FP rates for Group 1 - Experiment 3

Group 1 Experiment 3	GSE		SLR		Integrated data		Class
	TP Rate	FP Rate	TP Rate	FP Rate	TP Rate	FP Rate	
J48	0.361	0.201	0	0	0.426	0.181	liver
	0.799	0.639	1	1	0.819	0.574	elsewhere
Random Forest	0.41	0.081	0.089	0.058	0.443	0.107	liver
	0.919	0.59	0.942	0.911	0.893	0.557	elsewhere
SVM-Poly(d=1)	0.754	0.081	0.375	0.115	0.787	0.081	liver
	0.919	0.246	0.885	0.625	0.919	0.213	elsewhere
SVM-Poly(d=2)	0.721	0.094	0.375	0.108	0.803	0.074	liver
	0.906	0.279	0.892	0.625	0.926	0.197	elsewhere
SVM-G(c:10 g:0,005)	0.738	0.074	0.321	0.079	0.787	0.067	liver
	0.926	0.262	0.921	0.679	0.933	0.213	elsewhere

Table 17: TP and FP rates for Group 1 - Experiment 4

Group 1 Experiment 4	GSE		SLR		Integrated data		Class
	TP Rate	FP Rate	TP Rate	FP Rate	TP Rate	FP Rate	
J48	0.328	0.201	0	0	0.344	0.201	liver
	0.799	0.672	1	1	0.799	0.656	elsewhere
Random Forest	0.41	0.054	0.214	0.043	0.443	0.04	liver
	0.946	0.59	0.957	0.786	0.96	0.557	elsewhere
SVM-Poly(d=1)	0.475	0.195	0.607	0.137	0.443	0.188	liver
	0.805	0.525	0.863	0.393	0.812	0.557	elsewhere
SVM-Poly(d=2)	0.525	0.148	0.643	0.122	0.541	0.154	liver
	0.852	0.475	0.878	0.357	0.846	0.459	elsewhere
SVM-G(c:10 g:0,0045)	0.328	0.02	0.625	0.101	0.344	0.027	liver
	0.98	0.672	0.899	0.375	0.973	0.656	elsewhere

Table 18: TP and FP rates for Group 1 - Experiment 5

Group 1 Experiment 5	GSE		SLR		Integrated		Class
	TP Rate	FP Rate	TP Rate	FP Rate	TP Rate	FP Rate	
J48	0.475	0.248	0.036	0.072	0.443	0.228	liver
	0.752	0.525	0.928	0.964	0.772	0.557	elsewhere
Random Forest	0.639	0.087	0.339	0.065	0.541	0.034	liver
	0.913	0.361	0.935	0.661	0.966	0.459	elsewhere
SVM-Poly(d=1)	0.361	0.054	0.607	0.094	0.607	0.094	liver
	0.946	0.639	0.906	0.393	0.906	0.393	elsewhere
SVM-Poly(d=2)	0.426	0.087	0.536	0.151	0.475	0.107	liver
	0.913	0.574	0.849	0.464	0.893	0.525	elsewhere
SVM-G(c:10 g:0,4)	0.639	0.107	0.607	0.086	0.59	0.128	liver
	0.893	0.361	0.914	0.393	0.872	0.41	elsewhere

Table 19: TP and FP rates for Group 1 - Experiment 6

Group 1 Experiment 6	GSE		SLR		Integrated data		Class
	TP Rate	FP Rate	TP Rate	FP Rate	TP Rate	FP Rate	
J48	0.475	0.248	0.036	0.072	0.311	0.134	liver
	0.752	0.525	0.928	0.964	0.866	0.689	elsewhere
Random Forest	0.639	0.087	0.339	0.065	0.492	0.06	liver
	0.913	0.361	0.935	0.661	0.94	0.508	elsewhere
SVM-Poly(d=1)	0.361	0.054	0.607	0.094	0.705	0.074	liver
	0.946	0.639	0.906	0.393	0.926	0.295	elsewhere
SVM-Poly(d=2)	0.426	0.087	0.536	0.151	0.475	0.128	liver
	0.913	0.574	0.849	0.464	0.872	0.525	elsewhere
SVM-G(c:10 g:0,05)	0.639	0.107	0.607	0.086	0.656	0.087	liver
	0.893	0.361	0.914	0.393	0.913	0.344	elsewhere

A.2 TP & FP Rates for Multiclass Classification of GSE29271 - Group 2

Table 20: TP and FP rates for Group 2 - Experiment 1

Group 2 Experiment 1	GSE		SLR		Integrated data		Class
	TP Rate	FP Rate	TP Rate	FP Rate	TP Rate	FP Rate	
J48	0	0.059	0.2	0.022	0	0.059	0000
	0.526	0.108	0.316	0.121	0.5	0.127	1000
	0.857	0.279	0.78	0.423	0.835	0.279	1001
	0.391	0.076	0.348	0.076	0.391	0.076	1010
	0.059	0.045	0.176	0.062	0.059	0.045	1011
	0.444	0.027	0	0.011	0.444	0.027	1100
	0	0	0	0.01	0	0	1101
	0	0	0.25	0.016	0	0	1110
	0	0	0	0	0	0	1111
Random Forest	0.1	0.027	0.1	0.016	0.2	0.022	0000
	0.289	0.134	0.211	0.134	0.184	0.121	1000
	0.89	0.481	0.89	0.49	0.901	0.519	1001
	0.435	0.064	0.478	0.087	0.478	0.07	1010
	0.118	0.011	0.059	0.017	0.059	0.006	1011
	0.111	0	0	0	0.222	0	1100
	0	0	0	0	0	0	1101
	0	0	0	0	0	0	1110
	0	0	0	0	0	0	1111
SVM-Poly(d=1)	0.5	0	0.4	0.005	0.5	0	0000
	0.711	0.057	0.474	0.108	0.711	0.038	1000
	0.967	0.212	0.857	0.24	0.978	0.24	1001
	0.783	0.047	0.696	0.087	0.783	0.041	1010
	0.647	0	0.294	0.034	0.647	0	1011
	0.444	0.005	0.222	0.032	0.556	0	1100
	0	0	0	0	0	0	1101
	0.5	0	0	0	0.5	0	1110
	0	0	0	0	0	0	1111
SVM-Poly(d=2)	0.5	0	0.5	0.005	0.5	0	0000
	0.684	0.064	0.421	0.121	0.711	0.051	1000
	0.956	0.221	0.868	0.279	0.967	0.24	1001
	0.783	0.047	0.739	0.07	0.783	0.047	1010
	0.647	0	0.294	0.022	0.529	0	1011
	0.444	0.005	0.222	0.022	0.444	0.005	1100
	0	0	0	0	0	0	1101
	0.5	0	0.5	0	0.5	0	1110
	0	0	0	0	0	0	1111
SVM-G(c:20 g:0,0007)	0.5	0	0.4	0.005	0.5	0	0000
	0.684	0.057	0.421	0.083	0.711	0.038	1000
	0.956	0.24	0.934	0.356	0.967	0.26	1001
	0.783	0.047	0.652	0.052	0.783	0.047	1010
	0.588	0	0.412	0.017	0.529	0	1011
	0.444	0.005	0.222	0.005	0.444	0.005	1100
	0	0	0	0	0	0	1101
	0.5	0	0.5	0	0.5	0	1110
	0	0	0	0	0	0	1111

Table 21: TP and FP rates for Group 2 - Experiment 2

Group 2 Experiment 2	GSE		SLR		Integrated data		Class
	TP Rate	FP Rate	TP Rate	FP Rate	TP Rate	FP Rate	
J48	0	0.059	0.2	0.022	0	0.059	0000
	0.526	0.108	0.316	0.121	0.5	0.127	1000
	0.857	0.279	0.78	0.423	0.835	0.279	1001
	0.391	0.076	0.348	0.076	0.391	0.076	1010
	0.059	0.045	0.176	0.062	0.059	0.045	1011
	0.444	0.027	0	0.011	0.444	0.027	1100
	0	0	0	0.01	0	0	1101
	0	0	0.25	0.016	0	0	1110
	0	0	0	0	0	0	1111
Random Forest	0.3	0.022	0.2	0.016	0.3	0.005	0000
	0.263	0.166	0.289	0.064	0.237	0.146	1000
	0.901	0.452	0.923	0.538	0.879	0.49	1001
	0.435	0.047	0.565	0.058	0.522	0.07	1010
	0.059	0.011	0.118	0.011	0.059	0.006	1011
	0.111	0	0	0	0.222	0	1100
	0	0	0	0	0	0	1101
	0.25	0	0.5	0	0	0	1110
	0	0	0	0	0	0	1111
SVM-Poly(d=1)	0.5	0	0.4	0.005	0.5	0	0000
	0.711	0.057	0.474	0.108	0.711	0.038	1000
	0.967	0.212	0.857	0.24	0.978	0.24	1001
	0.783	0.047	0.696	0.087	0.783	0.041	1010
	0.647	0	0.294	0.034	0.647	0	1011
	0.444	0.005	0.222	0.032	0.556	0	1100
	0	0	0	0	0	0	1101
	0.5	0	0	0	0.5	0	1110
	0	0	0	0	0	0	1111
SVM-Poly(d=2)	0.5	0	0.5	0.005	0.5	0	0000
	0.684	0.064	0.421	0.121	0.711	0.051	1000
	0.956	0.221	0.868	0.279	0.967	0.24	1001
	0.783	0.047	0.739	0.07	0.783	0.047	1010
	0.647	0	0.294	0.022	0.529	0	1011
	0.444	0.005	0.222	0.022	0.444	0.005	1100
	0	0	0	0	0	0	1101
	0.5	0	0.5	0	0.5	0	1110
	0	0	0	0	0	0	1111
SVM-G(c:20 g:0,0007)	0.5	0	0.4	0.005	0.5	0	0000
	0.684	0.057	0.421	0.083	0.711	0.038	1000
	0.956	0.24	0.934	0.356	0.967	0.26	1001
	0.783	0.047	0.652	0.052	0.783	0.047	1010
	0.588	0	0.412	0.017	0.529	0	1011
	0.444	0.005	0.222	0.005	0.444	0.005	1100
	0	0	0	0	0	0	1101
	0.5	0	0.5	0	0.5	0	1110
	0	0	0	0	0	0	1111

Table 22: TP and FP rates for Group 2 - Experiment 3

Group 2 Experiment 3	GSE		SLR		Integrated data		Class
	TP Rate	FP Rate	TP Rate	FP Rate	TP Rate	FP Rate	
J48	0.2	0.065	0	0.059	0.1	0.049	0000
	0.342	0.159	0.105	0.07	0.211	0.134	1000
	0.549	0.337	0.659	0.567	0.56	0.385	1001
	0.348	0.064	0.261	0.093	0.261	0.134	1010
	0.235	0.118	0.118	0.067	0.118	0.062	1011
	0.222	0.043	0	0.032	0.333	0.075	1100
	0	0.01	0	0.01	0	0.01	1101
	0	0.01	0.25	0.026	0	0.021	1110
	0	0	0	0	0	0	1111
Random Forest	0.1	0.022	0.1	0.022	0	0.032	0000
	0.132	0.166	0.132	0.166	0.211	0.178	1000
	0.857	0.567	0.857	0.567	0.813	0.51	1001
	0.304	0.07	0.304	0.07	0.174	0.105	1010
	0	0.011	0	0.011	0	0.011	1011
	0	0.005	0	0.005	0	0.011	1100
	0	0	0	0	0	0	1101
	0	0	0	0	0	0	1110
	0	0	0	0	0	0	1111
SVM-Poly(d=1)	0	0.027	0	0.005	0	0.016	0000
	0.184	0.172	0.158	0.217	0.237	0.159	1000
	0.692	0.442	0.824	0.51	0.747	0.5	1001
	0.435	0.099	0.478	0.052	0.304	0.122	1010
	0.059	0.062	0	0.034	0	0.017	1011
	0.111	0.027	0	0	0.333	0.016	1100
	0	0.005	0	0	0	0	1101
	0	0.005	0	0	0	0.005	1110
	0	0	0	0	0	0	1111
SVM-Poly(d=2)	0	0.022	0	0.011	0	0.016	0000
	0.132	0.134	0.105	0.197	0.184	0.14	1000
	0.769	0.529	0.824	0.548	0.791	0.538	1001
	0.348	0.099	0.435	0.058	0.304	0.122	1010
	0.059	0.039	0	0.034	0	0.006	1011
	0.111	0.027	0	0	0.333	0.011	1100
	0	0	0	0	0	0	1101
	0	0.005	0	0	0	0.005	1110
	0	0	0	0	0	0	1111
SVM-G(c:10 g:0,005)	0	0.005	0	0	0	0	0000
	0.237	0.172	0.105	0.153	0.211	0.102	1000
	0.824	0.481	0.89	0.654	0.879	0.606	1001
	0.304	0.087	0.435	0.041	0.348	0.105	1010
	0.059	0.028	0	0.006	0	0	1011
	0.111	0.016	0	0	0.111	0.005	1100
	0	0	0	0	0	0	1101
	0	0.005	0	0	0	0	1110
	0	0	0	0	0	0	1111

Table 23: TP and FP rates for Group 2 - Experiment 4

Group 2 Experiment 4	GSE		SLR		Integrated data		Class
	TP Rate	FP Rate	TP Rate	FP Rate	TP Rate	FP Rate	
J48	0.4	0.049	0	0.011	0.4	0.049	0000
	0.237	0.185	0.026	0.019	0.237	0.185	1000
	0.615	0.337	0.945	0.865	0.615	0.346	1001
	0.13	0.11	0.087	0.017	0.13	0.11	1010
	0	0.062	0	0.017	0	0.062	1011
	0.111	0.054	0	0.016	0.111	0.048	1100
	0	0.036	0	0	0	0.036	1101
	0	0.01	0	0.01	0	0.01	1110
	0	0	0	0	0	0	1111
Random Forest	0	0.038	0	0	0.1	0.011	0000
	0.184	0.236	0.026	0.166	0.184	0.127	1000
	0.78	0.462	0.769	0.74	0.879	0.5	1001
	0.348	0.07	0.087	0.076	0.478	0.099	1010
	0.059	0.006	0.118	0.022	0	0.017	1011
	0	0.005	0	0	0	0.011	1100
	0	0	0	0	0	0	1101
	0.5	0	0	0	0	0	1110
	0	0	0	0	0	0	1111
SVM-Poly(d=1)	0.1	0.016	0	0.032	0.1	0.022	0000
	0.289	0.191	0.289	0.178	0.316	0.185	1000
	0.714	0.365	0.802	0.423	0.725	0.356	1001
	0.391	0.134	0.391	0.076	0.435	0.134	1010
	0.059	0.034	0.059	0.045	0.059	0.028	1011
	0.222	0.022	0.111	0.005	0.222	0.016	1100
	0	0	0	0	0	0	1101
	0	0.01	0	0	0	0.01	1110
	0	0	0	0	0	0	1111
SVM-Poly(d=2)	0	0.022	0	0.049	0.1	0.027	0000
	0.342	0.197	0.316	0.191	0.289	0.172	1000
	0.692	0.327	0.648	0.327	0.714	0.346	1001
	0.391	0.116	0.435	0.099	0.478	0.099	1010
	0.176	0.039	0.059	0.09	0.118	0.045	1011
	0.333	0.032	0.111	0.027	0.333	0.038	1100
	0	0	0	0.005	0	0	1101
	0	0.01	0	0	0	0.01	1110
	0	0	0	0	0	0	1111
SVM-G(c:10 g:0,005)	0	0	0	0.011	0	0	0000
	0.026	0.038	0.237	0.146	0.026	0.038	1000
	0.923	0.615	0.868	0.596	0.923	0.615	1001
	0.522	0.163	0.348	0.052	0.522	0.163	1010
	0	0	0	0.017	0	0	1011
	0	0	0	0	0	0	1100
	0	0	0	0	0	0	1101
	0	0	0	0	0	0	1110
	0	0	0	0	0	0	1111

Table 24: TP and FP rates for Group 2 - Experiment 5

Group 2 Experiment 5	GSE		SLR		Integrated data		Class
	TP Rate	FP Rate	TP Rate	FP Rate	TP Rate	FP Rate	
J48	0.2	0.043	0	0	0.2	0.043	0000
	0.263	0.153	0	0	0.263	0.153	1000
	0.681	0.337	1	1	0.681	0.337	1001
	0.304	0.105	0	0	0.304	0.105	1010
	0.176	0.056	0	0	0.176	0.056	1011
	0.111	0.054	0	0	0.111	0.054	1100
	0	0.01	0	0	0	0.01	1101
	0	0.016	0	0	0	0.016	1110
	0	0	0	0	0	0	1111
Random Forest	0.1	0.027	0	0	0.1	0.027	0000
	0.158	0.166	0	0	0.158	0.166	1000
	0.802	0.5	1	1	0.802	0.5	1001
	0.261	0.105	0	0	0.261	0.105	1010
	0	0.017	0	0	0	0.017	1011
	0	0.022	0	0	0	0.022	1100
	0	0	0	0	0	0	1101
	0	0.005	0	0	0	0.005	1110
	0	0	0	0	0	0	1111
SVM-Poly(d=1)	0.1	0.005	0	0.005	0.1	0.005	0000
	0.184	0.083	0.132	0.025	0.184	0.083	1000
	0.879	0.635	0.945	0.904	0.879	0.635	1001
	0.435	0.087	0	0.006	0.435	0.087	1010
	0	0.006	0	0.011	0	0.006	1011
	0	0	0	0.011	0	0	1100
	0	0	0	0	0	0	1101
	0	0.005	0	0	0	0.005	1110
	0	0	0	0	0	0	1111
SVM-Poly(d=2)	0.3	0.054	0	0.016	0.3	0.054	0000
	0.263	0.185	0.263	0.096	0.263	0.185	1000
	0.78	0.404	0.835	0.673	0.78	0.404	1001
	0.348	0.058	0	0.064	0.348	0.058	1010
	0.059	0.028	0.059	0.011	0.059	0.028	1011
	0.111	0.011	0.111	0.022	0.111	0.011	1100
	0	0	0	0.005	0	0	1101
	0.25	0.01	0	0.005	0.25	0.01	1110
	0	0	0	0	0	0	1111
SVM-G(c:10 g:0,01)	0	0	0	0	0	0	0000
	0.158	0.032	0.184	0.038	0.158	0.032	1000
	0.945	0.683	0.956	0.885	0.945	0.683	1001
	0.435	0.099	0	0	0.435	0.099	1010
	0	0	0	0.011	0	0	1011
	0	0	0	0.005	0	0	1100
	0	0	0	0	0	0	1101
	0	0	0	0	0	0	1110
	0	0	0	0	0	0	1111

Table 25: TP and FP rates for Group 2 - Experiment 6

Group 2 Experiment 6	GSE		SLR		Integrated data		Class
	TP Rate	FP Rate	TP Rate	FP Rate	TP Rate	FP Rate	
J48	0.2	0.043	0	0	0.3	0.059	0000
	0.263	0.153	0	0	0.289	0.051	1000
	0.681	0.337	1	1	0.835	0.481	1001
	0.304	0.105	0	0	0.261	0.081	1010
	0.176	0.056	0	0	0.118	0.045	1011
	0.111	0.054	0	0	0	0.016	1100
	0	0.01	0	0	0	0	1101
	0	0.016	0	0	0	0.016	1110
	0	0	0	0	0	0	1111
Random Forest	0.1	0.027	0	0	0	0.038	0000
	0.158	0.166	0	0	0.132	0.178	1000
	0.802	0.5	1	1	0.791	0.577	1001
	0.261	0.105	0	0	0.304	0.076	1010
	0	0.017	0	0	0	0.006	1011
	0	0.022	0	0	0	0.011	1100
	0	0	0	0	0	0	1101
	0	0.005	0	0	0	0	1110
	0	0	0	0	0	0	1111
SVM-Poly(d=1)	0.1	0.005	0	0.005	0	0.005	0000
	0.184	0.083	0.132	0.025	0.368	0.121	1000
	0.879	0.635	0.945	0.904	0.835	0.452	1001
	0.435	0.087	0	0.006	0.565	0.093	1010
	0	0.006	0	0.011	0	0.028	1011
	0	0	0	0.011	0.111	0.005	1100
	0	0	0	0	0	0	1101
	0	0.005	0	0	0.25	0.005	1110
	0	0	0	0	0	0	1111
SVM-Poly(d=2)	0.3	0.054	0	0.016	0.3	0.049	0000
	0.263	0.185	0.263	0.096	0.421	0.185	1000
	0.78	0.404	0.835	0.673	0.648	0.26	1001
	0.348	0.058	0	0.064	0.435	0.093	1010
	0.059	0.028	0.059	0.011	0.176	0.084	1011
	0.111	0.011	0.111	0.022	0.111	0.016	1100
	0	0	0	0.005	0	0	1101
	0.25	0.01	0	0.005	0.25	0.016	1110
	0	0	0	0	0	0	1111
SVM-G(c:10 g:0,01)	0	0	0	0	0	0.011	0000
	0.158	0.032	0.184	0.038	0.342	0.115	1000
	0.945	0.683	0.956	0.885	0.868	0.452	1001
	0.435	0.099	0	0	0.522	0.099	1010
	0	0	0	0.011	0	0.017	1011
	0	0	0	0.005	0.111	0.005	1100
	0	0	0	0	0	0	1101
	0	0	0	0	0.25	0.005	1110
	0	0	0	0	0	0	1111

A.3 TP & FP Rates for Binary Classification of GSE2034 - Group 3

Table 26: TP and FP rates for Group 3 - Experiment 1

Group 3 Experiment 1	GSE		SLR		Integrated data		Class
	TP Rate	FP Rate	TP Rate	FP Rate	TP Rate	FP Rate	
J48	0.899	0.163	0.822	0.638	0.907	0.163	erpos_r0
	0.838	0.101	0.363	0.178	0.838	0.093	erpos_r1
Random Forest	0.969	0.413	0.915	0.388	0.93	0.338	erpos_r0
	0.588	0.031	0.613	0.085	0.663	0.07	erpos_r1
SVM-Poly(d=1)	0.984	0.013	0.899	0.2	0.984	0.025	erpos_r0
	0.988	0.016	0.8	0.101	0.975	0.016	erpos_r1
SVM-Poly(d=2)	0.984	0.025	0.884	0.225	0.984	0.025	erpos_r0
	0.975	0.016	0.775	0.116	0.975	0.016	erpos_r1
SVM-G(c:10 g:0,001)	0.984	0.025	0.915	0.188	0.984	0.025	erpos_r0
	0.975	0.016	0.813	0.085	0.975	0.016	erpos_r1

Table 27: TP and FP rates for Group 3 - Experiment 2

Group 3 Experiment 2	GSE		SLR		Integrated data		Class
	TP Rate	FP Rate	TP Rate	FP Rate	TP Rate	FP Rate	
J48	0.899	0.163	0.822	0.638	0.907	0.163	erpos_r0
	0.838	0.101	0.363	0.178	0.838	0.093	erpos_r1
Random Forest	0.93	0.35	0.899	0.375	0.969	0.313	erpos_r0
	0.65	0.07	0.625	0.101	0.688	0.031	erpos_r1
SVM-Poly(d=1)	0.984	0.013	0.899	0.2	0.984	0.025	erpos_r0
	0.988	0.016	0.8	0.101	0.975	0.016	erpos_r1
SVM-Poly(d=2)	0.984	0.025	0.884	0.225	0.984	0.025	erpos_r0
	0.975	0.016	0.775	0.116	0.975	0.016	erpos_r1
SVM-G(c:10 g:0,001)	0.984	0.025	0.915	0.188	0.984	0.025	erpos_r0
	0.975	0.016	0.813	0.085	0.975	0.016	erpos_r1

Table 28: TP and FP rates for Group 3 - Experiment 3

Group 3 Experiment 3	GSE		SLR		Integrated data		Class
	TP Rate	FP Rate	TP Rate	FP Rate	TP Rate	FP Rate	
J48	0.698	0.5	1	1	0.767	0.588	erpos_r0
	0.5	0.302	0	0	0.413	0.233	erpos_r1
Random Forest	0.86	0.55	0.791	0.8	0.93	0.663	erpos_r0
	0.45	0.14	0.2	0.209	0.338	0.07	erpos_r1
SVM-Poly(d=1)	0.86	0.375	0.729	0.488	0.876	0.3	erpos_r0
	0.625	0.14	0.513	0.271	0.7	0.124	erpos_r1
SVM-Poly(d=2)	0.868	0.4	0.752	0.5	0.868	0.313	erpos_r0
	0.6	0.132	0.5	0.248	0.688	0.132	erpos_r1
SVM-G(c:10 g:0,014)	0.86	0.375	0.767	0.513	0.86	0.288	erpos_r0
	0.625	0.14	0.488	0.233	0.713	0.14	erpos_r1

Table 29: TP and FP rates for Group 3 - Experiment 4

Group 1 Experiment 4	GSE		SLR		Integrated data		Class
	TP Rate	FP Rate	TP Rate	FP Rate	TP Rate	FP Rate	
J48	0.767	0.488	1	1	0.744	0.475	erpos_r0
	0.513	0.233	0	0	0.525	0.256	erpos_r1
Random Forest	0.922	0.588	0.822	0.625	0.884	0.438	erpos_r0
	0.413	0.078	0.375	0.178	0.563	0.116	erpos_r1
SVM-Poly(d=1)	0.853	0.325	0.767	0.413	0.86	0.375	erpos_r0
	0.675	0.147	0.588	0.233	0.625	0.14	erpos_r1
SVM-Poly(d=2)	0.868	0.363	0.775	0.45	0.884	0.388	erpos_r0
	0.638	0.132	0.55	0.225	0.613	0.116	erpos_r1
SVM-G(c:10 g:0,005)	0.938	0.4	0.806	0.45	0.938	0.4	erpos_r0
	0.6	0.062	0.55	0.194	0.6	0.062	erpos_r1

Table 30: TP and FP rates for Group 3 - Experiment 5

Group 3 Experiment 5	GSE		SLR		Integrated		Class
	TP Rate	FP Rate	TP Rate	FP Rate	TP Rate	FP Rate	
J48	0.806	0.475	1	1	0.806	0.488	erpos_r0
	0.525	0.194	0	0	0.513	0.194	erpos_r1
Random Forest	0.922	0.513	0.899	0.488	0.922	0.488	erpos_r0
	0.488	0.078	0.513	0.101	0.513	0.078	erpos_r1
SVM-Poly(d=1)	0.899	0.438	0.814	0.363	0.899	0.438	erpos_r0
	0.563	0.101	0.638	0.186	0.563	0.101	erpos_r1
SVM-Poly(d=2)	0.86	0.413	0.791	0.325	0.86	0.413	erpos_r0
	0.588	0.14	0.675	0.209	0.588	0.14	erpos_r1
SVM-G(c:10 g:1,3)	0.876	0.288	0.837	0.375	0.876	0.288	erpos_r0
	0.713	0.124	0.625	0.163	0.713	0.124	erpos_r1

Table 31: TP and FP rates for Group 3 - Experiment 6

Group 3 Experiment 6	GSE		SLR		Integrated data		Class
	TP Rate	FP Rate	TP Rate	FP Rate	TP Rate	FP Rate	
J48	0.806	0.475	1	1	0.93	0.638	erpos_r0
	0.525	0.194	0	0	0.363	0.07	erpos_r1
Random Forest	0.922	0.513	0.899	0.488	0.907	0.513	erpos_r0
	0.488	0.078	0.513	0.101	0.488	0.093	erpos_r1
SVM-Poly(d=1)	0.899	0.438	0.814	0.363	0.822	0.275	erpos_r0
	0.563	0.101	0.638	0.186	0.725	0.178	erpos_r1
SVM-Poly(d=2)	0.86	0.413	0.791	0.325	0.791	0.375	erpos_r0
	0.588	0.14	0.675	0.209	0.625	0.209	erpos_r1
SVM-G(c:10 g:1,3)	0.876	0.288	0.837	0.375	0.814	0.238	erpos_r0
	0.713	0.124	0.625	0.163	0.763	0.186	erpos_r1

A.4 TP & FP Rates for Multiclass Classification of GSE2034 - Group 4

Table 32: TP and FP rates for Group 4 - Experiment 1

Group 4 Experiment 1	GSE		SLR		Integrated data		Class
	TP Rate	FP Rate	TP Rate	FP Rate	TP Rate	FP Rate	
J48	0.5	0.088	0.438	0.033	0.5	0.088	0
	0.643	0.139	0.607	0.304	0.643	0.139	1
	0.208	0.145	0.375	0.145	0.208	0.145	2
	0.286	0.161	0.571	0.054	0.286	0.161	3
	0.545	0.115	0.455	0.042	0.545	0.115	4
	0.3	0.062	0.3	0.062	0.3	0.062	5
	0	0	0.5	0.019	0	0	6
Random Forest	0.813	0.099	0.188	0.066	0.75	0.077	0
	0.821	0.19	0.714	0.278	0.786	0.228	1
	0.5	0.133	0.5	0.145	0.625	0.157	2
	0.429	0.032	0.429	0.032	0.429	0.022	3
	0.273	0.031	0.636	0.042	0.364	0.021	4
	0.3	0.041	0.7	0.031	0.4	0.01	5
	0.5	0	0.5	0	0.25	0	6
SVM-Poly(d=1)	1	0	0.938	0.022	1	0	0
	1	0	0.964	0.051	1	0	1
	1	0.012	0.875	0.06	1	0.012	2
	0.929	0	0.786	0.022	0.929	0	3
	1	0	0.818	0	1	0	4
	1	0	0.9	0.01	1	0	5
	1	0	0.25	0	1	0	6
SVM-Poly(d=2)	0.938	0	0.938	0	1	0	0
	1	0.051	0.964	0.063	1	0.038	1
	1	0.036	0.875	0.06	1	0.012	2
	0.786	0	0.786	0.022	0.929	0	3
	0.727	0	0.818	0	0.727	0	4
	1	0	0.9	0.01	1	0	5
	1	0	0.5	0	1	0	6
SVM-G(c:10 g:0,0005)	0.938	0	0.938	0.011	1	0	0
	1	0.038	0.964	0.063	1	0.013	1
	1	0.012	0.917	0.06	1	0.012	2
	0.929	0	0.786	0.011	0.929	0	3
	0.818	0	0.818	0	0.909	0	4
	1	0	0.9	0.01	1	0	5
	1	0	0.25	0	1	0	6

Table 33: TP and FP rates for Group 4 - Experiment 2

Group 4 Experiment 2	GSE		SLR		Integrated data		Class
	TP Rate	FP Rate	TP Rate	FP Rate	TP Rate	FP Rate	
J48	0.5	0.088	0.438	0.033	0.5	0.088	0
	0.643	0.139	0.607	0.304	0.643	0.139	1
	0.208	0.145	0.375	0.145	0.208	0.145	2
	0.286	0.161	0.571	0.054	0.286	0.161	3
	0.545	0.115	0.455	0.042	0.545	0.115	4
	0.3	0.062	0.3	0.062	0.3	0.062	5
	0	0	0.5	0.019	0	0	6
Random Forest	0.563	0.044	0.313	0.044	0.5	0.055	0
	0.75	0.266	0.964	0.278	0.786	0.177	1
	0.542	0.169	0.458	0.145	0.708	0.12	2
	0.357	0.075	0.429	0.054	0.429	0.065	3
	0.545	0.021	0.545	0.031	0.455	0.031	4
	0.3	0.021	0.5	0.01	0.6	0.031	5
	0	0	0	0	0.5	0	6
SVM-Poly(d=1)	1	0	0.938	0.022	1	0	0
	1	0	0.964	0.051	1	0	1
	1	0.012	0.875	0.06	1	0.012	2
	0.929	0	0.786	0.022	0.929	0	3
	1	0	0.818	0	1	0	4
	1	0	0.9	0.01	1	0	5
	1	0	0.25	0	1	0	6
SVM-Poly(d=2)	0.938	0	0.938	0	1	0	0
	1	0.051	0.964	0.063	1	0.038	1
	1	0.036	0.875	0.06	1	0.012	2
	0.786	0	0.786	0.022	0.929	0	3
	0.727	0	0.818	0	0.727	0	4
	1	0	0.9	0.01	1	0	5
	1	0	0.5	0	1	0	6
SVM-G(c:10 g:0,0005)	0.938	0	0.938	0.011	1	0	0
	1	0.038	0.964	0.063	1	0.013	1
	1	0.012	0.917	0.06	1	0.012	2
	0.929	0	0.786	0.011	0.929	0	3
	0.818	0	0.818	0	0.909	0	4
	1	0	0.9	0.01	1	0	5
	1	0	0.25	0	1	0	6

Table 34: TP and FP rates for Group 4 - Experiment 3

Group 4 Experiment 3	GSE		SLR		Integrated data		Class
	TP Rate	FP Rate	TP Rate	FP Rate	TP Rate	FP Rate	
J48	0.188	0.099	0.125	0.077	0.188	0.11	0
	0.393	0.241	0.464	0.342	0.286	0.266	1
	0.25	0.108	0.208	0.205	0.292	0.169	2
	0.071	0.204	0.143	0.097	0.143	0.118	3
	0.364	0.115	0.182	0.083	0.545	0.083	4
	0.2	0.113	0.3	0.10	0.2	0.134	5
	0	0.019	0	0.019	0	0.019	6
Random Forest	0.313	0.143	0.188	0.154	0	0.209	0
	0.393	0.443	0.357	0.468	0.429	0.443	1
	0.208	0.253	0.25	0.145	0.167	0.229	2
	0	0.065	0.286	0.108	0	0.065	3
	0.182	0.042	0.182	0.073	0.091	0.083	4
	0.1	0.041	0	0.021	0	0.031	5
	0	0	0	0	0	0	6
SVM-Poly(d=1)	0.188	0.132	0.313	0.121	0.125	0.077	0
	0.321	0.405	0.571	0.329	0.464	0.405	1
	0.375	0.181	0.458	0.253	0.292	0.301	2
	0.143	0.097	0	0.075	0	0.075	3
	0.364	0.063	0	0.063	0.364	0.021	4
	0.1	0.052	0	0.041	0.2	0.062	5
	0	0	0	0	0	0	6
SVM-Poly(d=2)	0.063	0.121	0.313	0.121	0.125	0.088	0
	0.357	0.418	0.5	0.316	0.536	0.38	1
	0.458	0.205	0.458	0.301	0.458	0.265	2
	0	0.065	0	0.075	0	0.043	3
	0.364	0.083	0	0.052	0.364	0.031	4
	0.1	0.052	0	0.041	0.2	0.062	5
	0	0	0	0	0	0	6
SVM-G(c:10 g:0,0027)	0.063	0.088	0.063	0	0.063	0.055	0
	0.464	0.392	0.821	0.582	0.536	0.43	1
	0.542	0.301	0.458	0.277	0.417	0.277	2
	0.143	0.022	0	0	0.071	0.065	3
	0.364	0.052	0	0.021	0.364	0.031	4
	0.1	0.021	0	0.01	0.1	0.041	5
	0	0	0	0	0	0	6

Table 35: TP and FP rates for Group 4 - Experiment 4

Group 4 Experiment 4	GSE		SLR		Integrated data		Class
	TP Rate	FP Rate	TP Rate	FP Rate	TP Rate	FP Rate	
J48	0.25	0.11	0	0	0.313	0.132	0
	0.429	0.19	0.857	0.823	0.464	0.203	1
	0.25	0.193	0	0.096	0.583	0.133	2
	0.143	0.097	0	0	0.214	0.108	3
	0.455	0.146	0.091	0.042	0.455	0.104	4
	0.3	0.093	0	0.021	0.1	0.052	5
	0	0.019	0.25	0.019	0	0.019	6
Random Forest	0.5	0.132	0.188	0.099	0.313	0.088	0
	0.464	0.329	0.536	0.291	0.464	0.354	1
	0.333	0.169	0.5	0.277	0.25	0.217	2
	0.143	0.086	0.071	0.086	0.143	0.118	3
	0.364	0.073	0.182	0.073	0.182	0.063	4
	0.3	0.01	0	0.031	0.3	0.031	5
	0	0.01	0	0.01	0.25	0.01	6
SVM-Poly(d=1)	0.375	0.099	0.125	0.154	0.313	0.143	0
	0.429	0.329	0.536	0.203	0.643	0.291	1
	0.375	0.133	0.542	0.229	0.542	0.096	2
	0.143	0.108	0.071	0.075	0.071	0.032	3
	0	0.115	0.182	0.104	0.182	0.104	4
	0	0.103	0	0.072	0	0.103	5
	0	0.01	0	0.01	0	0.01	6
SVM-Poly(d=2)	0.313	0.121	0.125	0.22	0.438	0.121	0
	0.464	0.342	0.393	0.19	0.571	0.253	1
	0.375	0.12	0.375	0.169	0.542	0.133	2
	0.143	0.108	0.143	0.118	0.071	0.054	3
	0.182	0.094	0	0.073	0.182	0.083	4
	0.1	0.072	0.2	0.134	0.1	0.093	5
	0	0.01	0	0.01	0	0.029	6
SVM-G(c:10 g:0,036)	0.375	0.099	0	0.055	0.188	0.033	0
	0.571	0.342	0.786	0.443	0.893	0.494	1
	0.417	0.145	0.667	0.217	0.583	0.096	2
	0.214	0.086	0.143	0.075	0.143	0.054	3
	0.273	0.083	0	0.01	0.364	0.031	4
	0	0.052	0	0.01	0	0.01	5
	0	0	0	0	0	0	6

Table 36: TP and FP rates for Group 4 - Experiment 5

Group 4 Experiment 5	GSE		SLR		Integrated data		Class
	TP Rate	FP Rate	TP Rate	FP Rate	TP Rate	FP Rate	
J48	0.313	0.11	0	0	0.188	0.066	0
	0.536	0.241	0.964	0.987	0.464	0.177	1
	0.458	0.096	0	0.024	0.583	0.157	2
	0.286	0.14	0	0	0.357	0.204	3
	0.455	0.073	0	0	0.455	0.063	4
	0.1	0.082	0	0	0.2	0.062	5
	0	0.01	0	0	0	0.01	6
Random Forest	0.25	0.11	0	0	0.438	0.088	0
	0.464	0.304	0.821	0.759	0.5	0.304	1
	0.375	0.277	0.333	0.193	0.375	0.217	2
	0.143	0.065	0	0	0.143	0.075	3
	0.273	0.083	0	0	0.364	0.052	4
	0.1	0.031	0	0	0.5	0.021	5
	0	0.01	0	0	0	0.019	6
SVM-Poly(d=1)	0.25	0.121	0.125	0.022	0.5	0.077	0
	0.607	0.266	0.893	0.544	0.714	0.278	1
	0.417	0.205	0.5	0.181	0.625	0.108	2
	0.214	0.054	0.143	0.032	0.286	0.054	3
	0.091	0.115	0	0	0.273	0.052	4
	0.1	0.052	0	0.021	0.5	0.041	5
	0.25	0	0	0.01	0	0	6
SVM-Poly(d=2)	0.5	0.088	0.25	0.099	0.5	0.121	0
	0.393	0.278	0.429	0.354	0.571	0.215	1
	0.292	0.205	0.583	0.265	0.542	0.12	2
	0.357	0.097	0.214	0.011	0.214	0.054	3
	0.091	0.083	0.091	0.063	0.182	0.052	4
	0.2	0.062	0	0.062	0.3	0.082	5
	0.5	0.01	0	0.01	0.5	0.039	6
SVM-G(c:10 g:0,13)	0.5	0.132	0.125	0.022	0.563	0.088	0
	0.5	0.203	0.857	0.494	0.679	0.215	1
	0.458	0.181	0.583	0.181	0.542	0.108	2
	0.357	0.065	0.286	0.043	0.214	0.065	3
	0.182	0.063	0	0	0.364	0.073	4
	0.4	0.062	0	0.021	0.4	0.062	5
	0.5	0	0	0.01	0.25	0.01	6

Table 37: TP and FP rates for Group 4 - Experiment 6

Group 4 Experiment 6	GSE		SLR		Integrated data		Class
	TP Rate	FP Rate	TP Rate	FP Rate	TP Rate	FP Rate	
J48	0.313	0.11	0	0	0.25	0.099	0
	0.536	0.241	0.964	0.987	0.536	0.228	1
	0.458	0.096	0	0.024	0.5	0.12	2
	0.286	0.14	0	0	0.286	0.14	3
	0.455	0.073	0	0	0.455	0.073	4
	0.1	0.082	0	0	0.1	0.082	5
	0	0.01	0	0	0	0.01	6
Random Forest	0.25	0.11	0	0	0.25	0.154	0
	0.464	0.304	0.821	0.759	0.536	0.392	1
	0.375	0.277	0.333	0.193	0.25	0.193	2
	0.143	0.065	0	0	0.071	0.054	3
	0.273	0.083	0	0	0.273	0.042	4
	0.1	0.031	0	0	0.3	0.052	5
	0	0.01	0	0	0	0	6
SVM-Poly(d=1)	0.25	0.121	0.125	0.022	0.563	0.154	0
	0.607	0.266	0.893	0.544	0.5	0.177	1
	0.417	0.205	0.5	0.181	0.542	0.169	2
	0.214	0.054	0.143	0.032	0.286	0.065	3
	0.091	0.115	0	0	0.182	0.083	4
	0.1	0.052	0	0.021	0.3	0.052	5
	0.25	0	0	0.01	0.25	0	6
SVM-Poly(d=2)	0.5	0.088	0.25	0.099	0.625	0.132	0
	0.393	0.278	0.429	0.354	0.5	0.19	1
	0.292	0.205	0.583	0.265	0.417	0.108	2
	0.357	0.097	0.214	0.011	0.5	0.086	3
	0.091	0.083	0.091	0.063	0	0.104	4
	0.2	0.062	0	0.062	0.3	0.082	5
	0.5	0.01	0	0.01	0.25	0	6
SVM-G(c:10 g:0,13)	0.5	0.132	0.125	0.022	0.688	0.132	0
	0.5	0.203	0.857	0.494	0.536	0.177	1
	0.458	0.181	0.583	0.181	0.458	0.157	2
	0.357	0.065	0.286	0.043	0.429	0.065	3
	0.182	0.063	0	0	0.364	0.083	4
	0.4	0.062	0	0.021	0.2	0.052	5
	0.5	0	0	0.01	0	0	6

APPENDIX B

SAMPLE GROUPS

B.1 GSE29271 Sample IDs used for binary classification

GSM308256 GSM308257 GSM308258 GSM308259 GSM308260 GSM308261
GSM308262 GSM308263 GSM308265 GSM308266 GSM308267 GSM308268
GSM308269 GSM308270 GSM308271 GSM308272 GSM308273 GSM308274
GSM308275 GSM308276 GSM308277 GSM308278 GSM308279 GSM308280
GSM308281 GSM308282 GSM308283 GSM308284 GSM308285 GSM308286
GSM308287 GSM308288 GSM308289 GSM308290 GSM308291 GSM308292
GSM308293 GSM308294 GSM308295 GSM308296 GSM308297 GSM308298
GSM308299 GSM308300 GSM308301 GSM308302 GSM308303 GSM308304
GSM308305 GSM308306 GSM308307 GSM308308 GSM308309 GSM308310
GSM308311 GSM308313 GSM308314 GSM308315 GSM308316 GSM308318
GSM308319 GSM308320 GSM308321 GSM308322 GSM308323 GSM308324
GSM308325 GSM308326 GSM308327 GSM308328 GSM308330 GSM308331
GSM308332 GSM308333 GSM308335 GSM308336 GSM308338 GSM308339
GSM308340 GSM308341 GSM308342 GSM308343 GSM308344 GSM308345
GSM308346 GSM308347 GSM308348 GSM308349 GSM308350 GSM308352
GSM308353 GSM308354 GSM308355 GSM308356 GSM308358 GSM308359
GSM308360 GSM308361 GSM308362 GSM308363 GSM308364 GSM308365
GSM308366 GSM308367 GSM308368 GSM308369 GSM308370 GSM308371
GSM308372 GSM308373 GSM308374 GSM308375 GSM308376 GSM308377
GSM308378 GSM308379 GSM308380 GSM308381 GSM308382 GSM308383
GSM308384 GSM308385 GSM308386 GSM308387 GSM308388 GSM308389
GSM308390 GSM308391 GSM308392 GSM308393 GSM308394 GSM308395

GSM308396 GSM308397 GSM308398 GSM308399 GSM308400 GSM308401 GSM308402 GSM308403 GSM308404 GSM308405 GSM308407 GSM308408 GSM308410 GSM308411 GSM308412 GSM308413 GSM308414 GSM308415 GSM308416 GSM308417 GSM308418 GSM308419 GSM308420 GSM308421 GSM308422 GSM308423 GSM308424 GSM308425 GSM308426 GSM308427 GSM308428 GSM308429 GSM308430 GSM308431 GSM308432 GSM308433 GSM308434 GSM308435 GSM308436 GSM308437 GSM308438 GSM308439 GSM308440 GSM308441 GSM308442 GSM308443 GSM308444 GSM308445 GSM308446 GSM308447 GSM308448 GSM308449 GSM308450 GSM308451 GSM308452 GSM308453 GSM308454 GSM308455 GSM308456 GSM308457 GSM308458 GSM308459 GSM308460 GSM723449 GSM723450 GSM723451 GSM723452 GSM723453 GSM723454 GSM723455 GSM723456 GSM723457 GSM723458 GSM723459 GSM723460 GSM723461 GSM723462 GSM723463

B.2 GSE29271 Sample IDs excluded from multiclass classification

GSM723449 GSM723450 GSM723451 GSM723452 GSM723453 GSM723454 GSM723455 GSM723456 GSM723457 GSM723458 GSM723459 GSM723460 GSM723461 GSM723462 GSM723463

B.3 GSE2034 Sample IDs used for binary classification

GSM36777 GSM36778 GSM36779 GSM36781 GSM36782 GSM36783 GSM36784 GSM36785 GSM36786 GSM36787 GSM36789 GSM36790 GSM36792 GSM36794 GSM36796 GSM36799 GSM36801 GSM36802 GSM36803 GSM36804 GSM36805 GSM36806 GSM36807 GSM36810 GSM36811 GSM36813 GSM36814 GSM36815 GSM36817 GSM36818 GSM36819 GSM36820 GSM36821 GSM36823 GSM36824 GSM36825 GSM36826 GSM36829 GSM36830 GSM36831 GSM36832 GSM36834 GSM36836 GSM36837 GSM36838 GSM36839 GSM36840 GSM36841 GSM36842 GSM36843 GSM36844 GSM36845 GSM36848 GSM36849 GSM36850 GSM36851 GSM36852 GSM36853 GSM36856 GSM36857 GSM36858 GSM36859 GSM36860 GSM36861 GSM36864 GSM36866 GSM36867 GSM36868 GSM36869 GSM36870 GSM36871 GSM36872 GSM36873 GSM36874 GSM36877 GSM36878 GSM36880 GSM36881 GSM36882 GSM36883 GSM36884 GSM36885 GSM36887 GSM36888

GSM36889 GSM36890 GSM36892 GSM36893 GSM36894 GSM36895 GSM36896
GSM36897 GSM36898 GSM36899 GSM36900 GSM36901 GSM36902 GSM36903
GSM36907 GSM36908 GSM36910 GSM36911 GSM36913 GSM36914 GSM36916
GSM36917 GSM36919 GSM36920 GSM36921 GSM36922 GSM36924 GSM36925
GSM36927 GSM36928 GSM36929 GSM36930 GSM36931 GSM36932 GSM36933
GSM36934 GSM36936 GSM36938 GSM36939 GSM36942 GSM36943 GSM36944
GSM36945 GSM36946 GSM36947 GSM36948 GSM36950 GSM36951 GSM36954
GSM36956 GSM36957 GSM36958 GSM36960 GSM36962 GSM36963 GSM36965
GSM36967 GSM36970 GSM36971 GSM36972 GSM36973 GSM36974 GSM36975
GSM36976 GSM36979 GSM36980 GSM36982 GSM36983 GSM36984 GSM36985
GSM36986 GSM36987 GSM36988 GSM36989 GSM36990 GSM36992 GSM36993
GSM36994 GSM36995 GSM36996 GSM36997 GSM36998 GSM36999 GSM37000
GSM37001 GSM37003 GSM37004 GSM37005 GSM37006 GSM37007 GSM37008
GSM37009 GSM37010 GSM37011 GSM37012 GSM37013 GSM37014 GSM37015
GSM37018 GSM37019 GSM37024 GSM37025 GSM37026 GSM37027 GSM37028
GSM37029 GSM37030 GSM37031 GSM37032 GSM37033 GSM37035 GSM37036
GSM37037 GSM37038 GSM37039 GSM37041 GSM37044 GSM37046 GSM37047
GSM37051 GSM37057 GSM37058 GSM37059 GSM37060 GSM37062

B.4 GSE2034 Sample IDs used for multiclass classification

GSM36778 GSM36784 GSM36789 GSM36792 GSM36797 GSM36800 GSM36811
GSM36813 GSM36814 GSM36815 GSM36818 GSM36826 GSM36835 GSM36838
GSM36839 GSM36858 GSM36860 GSM36862 GSM36870 GSM36872 GSM36874
GSM36875 GSM36877 GSM36879 GSM36881 GSM36885 GSM36888 GSM36897
GSM36898 GSM36902 GSM36903 GSM36905 GSM36908 GSM36911 GSM36918
GSM36920 GSM36923 GSM36924 GSM36926 GSM36927 GSM36928 GSM36931
GSM36937 GSM36939 GSM36941 GSM36943 GSM36946 GSM36947 GSM36949
GSM36950 GSM36952 GSM36954 GSM36955 GSM36956 GSM36957 GSM36960
GSM36964 GSM36967 GSM36969 GSM36971 GSM36972 GSM36973 GSM36974
GSM36976 GSM36983 GSM36985 GSM36986 GSM36989 GSM36994 GSM36996
GSM36997 GSM36998 GSM36999 GSM37001 GSM37002 GSM37003 GSM37004
GSM37005 GSM37006 GSM37007 GSM37008 GSM37011 GSM37013 GSM37018
GSM37020 GSM37022 GSM37023 GSM37026 GSM37027 GSM37028 GSM37029

GSM37030 GSM37031 GSM37035 GSM37036 GSM37037 GSM37038 GSM37039 GSM37040 GSM37041 GSM37042 GSM37049 GSM37050 GSM37051 GSM37052 GSM37053 GSM37058