T.R.

HACETTEPE UNIVERSITY

INSTITUTE OF HEALTH SCIENCES

NEW APPROACH TO UNSUPERVISED BASED CLASSIFICATION ON MICROARRAY DATA

Erdal COŞGUN

Biostatistics Program

PHILOSOPHY OF DOCTORATE THESIS

ANKARA

2013

T.R.

HACETTEPE UNIVERSITY

INSTITUTE OF HEALTH SCIENCES

NEW APPROACH TO UNSUPERVISED BASED CLASSIFICATION ON MICROARRAY DATA

Erdal COŞGUN

Biostatistics Program

PHILOSOPHY OF DOCTORATE THESIS

ADVISOR

Prof.Dr.Ergun KARAAĞAOĞLU

ANKARA

2013

Anabilim Dalı	**:Biyoistatistik**
Program	**:Doktora**
Tez Başlığı	**:New Approach to Unsupervised Based Classification on Microarray Data**
Öğrenci Adı-Soyadı	**:Erdal Coşgun**
Savunma Sınavı Tarihi	**:13.12.2013**

Bu çalışma jürimiz tarafından yüksek lisans/doktora tezi olarak kabul edilmiştir.

Jüri Başkanı:	**Prof.Dr.Osman SARAÇBAŞI** (Hacettepe Üniversitesi)
Tez danışmanı:	**Prof.Dr.Ergun KARAAĞAOĞLU** (Hacettepe Üniversitesi)
Üye:	**Doç.Dr.Mehtap AKÇİL OK** (Başkent Üniversitesi)
Üye:	**Doç.Dr.Erdem KARABULUT** (Hacettepe Üniversitesi)
Üye:	**Doç.Dr.Pınar KARAGÖZ** (Ortadoğu Teknik Üniversitesi)

ONAY

Bu tez Hacettepe Üniversitesi Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliğinin ilgili maddeleri uyarınca yukarıdaki jüri tarafından uygun görülmüş ve Sağlık Bilimleri Enstitüsü Yönetim Kurulu kararıyla kabul edilmiştir.

Prof.Dr. Ersin FADILLIOĞLU

Müdür

TEŞEKKÜR

Tez çalışmamın ve akademik kariyerimin başlangıcından itibaren bana desteğini esirgemeyen ve her zaman yanımda olan Prof.Dr.Ergun Karaağaoğlu'na,

Her zaman bana destek ve ilgilerini esirgemeyen Biyoistatistik Ab.D. öğretim üyeleri Prof.Dr.Osman Saraçbaşı ve Prof.Dr.Reha Alpar'a,

Beni akademik kariyere yönlendiren ve her zaman destekleyen Prof.Dr.Tülay Saraçbaşı'na,

Her türlü sorun ve sorumda sabırla bana destek ve cevap veren, veri madenciliği konusunda çalışmama yaptığı katkılardan dolayı Doç.Dr.Erdem Karabulut'a,

Değerli fikirleri ile her zaman yanımda olan Doç.Dr.Pınar Özdemir'e,

Genetik alanındaki çalışmalara yönelmemi sağlayan kıymetli hocam Prof.Dr.Nurten Akarsu'ya

Tez çalışmam sürecinde desteklerinden dolayı Doç.Dr.Pınar Karagöz'e

Teze değer katan yazılımın kullanıcı arayüzünün oluşturulmasına katkılarından dolayı Ar.Gör.Aydın Kaya, Ar.Gör.Ali Seydi Keçeli ve Ar.Gör.Hüseyin Temuçin'e,

Sağladıkları burs ile çok özel bir AR-GE merkezinde çalışmamı sağlayan Pfizer ilaç firmasına,

Genetik istatistik ve Biyoinformatik alanında bana öğrettikleri için Prof.Dr.David Allison, Prof.Dr.Murat Tanık ve Doç.Dr.Christine W. Duarte'ye,

Uyumlu ve verimli bir çalışma ortamı içinde araştırmalarımı yaptığım Biyoistatistik Ab.D.'nın akademik ve idari personeline,

Doktora eğitimim boyunca her türlü desteği veren Sağlık Bilimleri Enstitü'sü çalışanlarına,

Ve

Beni ben yapan değerlerin kaynağı sevgili Annem Meryem Coşgun'a,

Tezin dil kontrol ve yazım aşamalarında tüm hayatım boyunca olduğu gibi yanımda olan sevgili Babam Abidin Coşgun'a,

Kardeşim ve dostum Serdal Coşgun'a

Teşekkür ederim.

ABSTRACT

Coşgun, E., New Approach to Unsupervised Based Classification on Microarray Data, Hacettepe University Institute of Health Sciences, Biostatistics Program PhD Thesis, Ankara, 2013. Genetic studies have been an important part of medical researches in recent years. These studies have become essential for the development of personalized treatment options and discovery of new drugs. The majority of these researches have focused on obtaining gene expression data. Different methods have been developed for the analysis of gene expression data. The most important problem in the analysis of these data is that they are high dimensional to help find the expression levels for thousands of genes for the presence of a small number of individuals. Analyzing such data would not be possible with classical statistical methods because this type of data does not provide statistical assumptions. For this reason, data mining methods have been used for the analyses. According to the classical data mining approach, dimension reduction of high-dimensional data must be applied first by using Principal Component Analysis, Independent Component Analysis or Factor Analysis, then the classification, estimation or essential analysis methods such as clustering must be selected. Within the scope of this thesis, the solution has been suggested to the state of the factors of the reduced data to be similar, which is one of the missing points of this approach. In this context, the dimension has been reduced and factors have been obtained first in gene expression data, and then these structures have been analyzed by Random Forest, a most widely used tree-based method for the classification analysis. Results of this analysis were compared with the results of the use of cluster loadings obtained by size reduction proposed by the thesis study first, and then clustering factors with the Kohonen Self Organizing Map method in the Random Forest algorithm. One of the major advantages of the proposed approach is to send 1000 sub-samples selected by sampling method (bootstrap) to the Random Forest algorithm by replacing the factors clustered. In this way, both; data that could not be factorized were made more homogeneous by clustering analysis, and random selection criteria of the Random Forest method were further strengthened. The performance measures used in comparing these approaches are True Classification Rate, the F-score, Precision and Recall. Applications were carried out on two types of data: data publicly available based on 15 Gene Expression Omnibus database and 18 artificial data created for specific scenarios. The proposed method provided an average of 17.8% and 11.68% improvement for the true classification rate that is the most essential measure of comparison in data with 2 and 3 classes, and in artificial data an average of 14.5% improvement in data sets with 3 dimensions and have 3 classes with 50 individuals. The proposed method has increased the performance especially in data with less subjects and classes in terms of classification based on these findings. Software that can make all of these analyses more comfortable based on the R programming language has been developed within this thesis and the researchers will be able to carry out their own analysis.

Keywords: Gene expression, Data Mining, Random Forest, Kohonen Map, Bootstrap

ÖZET

Coşgun, E. Mikroarray Verilerde Danışmansız Öğrenmeye Dayalı Sınıflamada Yeni Yaklaşım, Hacettepe Üniversitesi Sağlık Bilimleri Enstitüsü Biyoistatistik Programı Doktora Tezi, Ankara 2013. Genetik araştırmalar son yıllarda tıbbi araştırmaların önemli bir parçası olmuştur. Kişiye özel tedavi yöntemlerinin geliştirilmesi, yeni ilaç keşifleri için bu çalışmalar olmazsa olmaz hale gelmiştir. Bu araştırmaların da büyük kısmı gen expresyon verilerinin elde edilmesine odaklanmıştır. Gen expression verilerinin analizi için farklı yöntemler geliştirilmiştir. Bu verilerin analizinde en önemli sorun yüksek boyutta olmalarıdır. Yani çok az sayıda bireye ait binlerce gen için expression düzeylerinin bulunmasıdır. Bu tip bir veriyi klasik istatistiksel yöntemler ile analiz etmek mümkün olmamaktadır. Çünkü bu veri türü istatistiksel varsayımları sağlamamaktadır. Bu nedenle Veri Madenciliği yöntemleri analizler için kullanılmaktadır. Klasik veri madenciliğine göre önce yüksek boyutlu verinin boyut indirgemesi Temel Bileşenler Analizi, Bağımsız Bileşenler Analizi ya da Faktör Analizi kullanılarak yapılmalı, sonrasında sınıflama, tahmin ya da kümeleme gibi ana analiz yöntemleri seçilmelidir. Tez çalışması kapsamından bu yaklaşımın eksik noktalarından biri olan indirgenen verideki faktörlerin benzer olması durumuna çözüm önerilmiştir. Bu kapsamda gen ekspresyon verilerinde önce boyut indirgenip faktörler elde edilmiş sonra bu yapılar sınıflama analizleri için en çok kullanılan ağaç tabanlı yöntem olan Random Forest yöntemi analiz edilmiştir. Bu analiz sonuçları tez çalışması ile önerilen önce boyut indirgeme sonra elde edilen faktörlerin Kohonen Self Organizing Map yöntemi ile kümelenmesi sonucu elde edilen küme yüklerinin Random Forest algoritmasında kullanılmasının sonuçları ile karşılaştırılmıştır. Önerilen yaklaşımın en önemli avantajlarından biri kümelenen faktörlerin yerine koyarak örnekleme *(bootstrap)* yöntemi seçilen 1000 adet alt örneklemi Random Forest algoritmasına göndermesidir. Bu sayede hem yeterince faktörleştirilemeyen veri kümeleme analizi ile daha homojen duruma getirilmiş hem de Random Forest yönteminin rastgele seçme kriteri daha da güçlendirilmiştir. Bu yaklaşımları karşılaştırırken kullanılan performans ölçüleri Doğru Sınıflama Oranı, F-Skoru, *Precision* ve *Recall*'dur. İki tip veri üzerinde uygulamalar yapılmıştır: 15 adet Gene Expression Omnibus veri tabanı üzerinden genel kullanıma açık veriler ve belirli senaryolara göre oluşturulmuş 18 adet yapay veri. Bu verilerin analizi sonucu elde edilen bulgulara göre önerilen yaklaşım gerçek verilerde 2 ve 3 sınıflı daha az sayıda örneklem genişliğine sahip verilerde en temel karşılaştırma ölçüsü olan doğru sınıflama oranı için ortalama %17,8 ve %11,68'lik artış, yapay verilerde ise her sınıfta 50 bireyin olduğu 3 boyutlu ve 3 sınıflı veri setlerinde ortalama %14.95'lik artış sağlamıştır. Bu bulgular ışığında özellikle daha az deneğe ve sınıfa sahip verilerde önerilen yöntem sınıflama açısında performansı arttırmıştır. Tüm bu analizlerin daha rahat yapılabilmesi için R programlama dilini esas alan bir yazılım tez kapsamında geliştirilmiş olup araştırıcıların kendi analizlerini yapabilmeleri sağlanmıştır.

Anahtar Kelimeler: Gen ekspresyon, Veri Madenciliği, Random Forest, Kohonen Map, Bootstrap

INDEX

	Page
APPROVAL PAGE	iii
TEŞEKKÜR	iv
ABSTRACT	v
ÖZET	vi
INDEX	vii
SYMBOLS AND ABBREVIATIONS INDEX	ix
FIGURES INDEX	xi
TABLES INDEX	xiv
1. INTRODUCTION	1
1.1. Proposed Method	2
2. GENERAL INFORMATION	4
2.1. Microarray Technology and Gene Expression Data	4
2.2. Fold Change Analysis	7
2.3. Data Mining Tools	8
2.4. [R] Software	8
2.5. Literature Review of Random Forest	9
2.6. Literature Review of Independent Component Analysis	10
2.7. Literature Review of Kohonen Map	11
2.8. Literature Review of Dimension Reduction Based Classification	13
3. MATERIAL AND METHODS	15
3.1. Dimension Reduction	16
3.1.1. Independent Component Analysis	16
3.2. Classification and Regression Trees	18
3.3. Random Forest	18
3.4. Selection of Generalization Methods	20
3.4.1. Bootstrap	20
3.5. Kohonen Map	21

Page

3.6. Performance Comparison Criteria 23

3.7. Gene3E 25

3.7.1. Objective of Tool 26

3.7.2. Methods of Tool 26

3.7.3. Application of Tool 27

3.7.4. RServe 28

3.7.5. Interface and System Requirement of Tool 29

3.7.6. Software Quality Attributes of Tool 31

3.7.7. Development of the Protocol of Communication Between Components 32

3.8. Materials 35

3.8.1. Gene Expression Omnibus Data Sets 35

3.8.2. Simulation Study 52

4. RESULTS 54

4.1. Results of GEO Data Sets 54

4.2. Results of Simulation Study 67

5. DISCUSSION 82

6. CONCLUSION 85

REFERENCES 89

APPENDICES

Appendix 1: Gene3E Software

SYMBOLS AND ABBREVIATIONS

AUC	Area Under Curve (ROC)
CA	Classic Approach
cDNA	Complimentary deoxyribonucleic acid
DM	Data Mining
DNA	Deoxyribonucleic acid
DAVID	The Database for Annotation, Visualization and Integrated Discovery
FA	Factor Analysis
G_{bg}	Green background
GEHP	The Gene Expression Data With Hidden Patterns
GEO	Gene Expression Omnibus
G_{fg}	Green foreground
GUI	Graphical User Interface
ICA	Independent Component Analysis
IP	Internet Protocol
IPCA	Independent Principal Component Analysis
KM	Kohonen Map
KNN	K-Nearest Neighborhood
LDA	Linear Discriminant Analysis
LLDE	Locally linear discriminant embedding
LOOCV	Leave One-Out Cross Validation
LPPO	Lagging Prediction Peephole Optimization
MAC	Media Access Control
MDS	Multi Dimensional Scaling
mRNA	Messenger Ribonucleic acid
mtry	Number of Parameter for Every Split
PCA	Principal Component Analysis
PLS	Partial least squares
PM	Proposed Method

R_{bg}	Red background
RF	Random Forest
RFE	Recursive Feature Elimination
R_{fg}	Red foreground
RNA	Ribonucleic Acid
SIR	Sliced inverse regression
SNP	Single Nucleotide Polymorphism
SVM	Support Vector Machine
SWT	Standart Widget Toolkit
TCP	Transmission Control Protocol
TCR	True Classification Rate

FIGURES

		Page
1.1.	Pipeline of Proposed Method	3
2.1.	The process from the cell samples to the microarray	4
2.2.	The reflection image of the basic expression data	5
2.3.	The description of gene expression data	6
2.4.	Pipeline of gene expression analysis	7
3.1.	Microarray Gene expression analysis flow chart on Data Mining	15
3.2.	ICA analysis pipeline	17
3.3.	Random Forest Algorithm Flowchart	19
3.4.	Visualization of Kohonen Map Clustering	22
3.5.	User interface of Gene3E	29
3.6.	User interface of Gene3E with web links	30
3.7.	Client structure of Gene3E	34
3.8.	Skeletal muscle response to insulin infusion data profile graph	37
3.9.	Lymph node and tonsil comparison data profile graph	38
3.10.	Atrial and ventricular myocardium comparison data profile	39
3.11.	Metastatic prostate cancer data profile graph	40
3.12.	Asthma and Atopy data profile graph	41
3.13.	Quercetin effect on the colonic mucosa data profile graph	42
3.14.	Tumor necrosis factor effect on macrovascular umbilical vein endothelial data profile graph	43
3.15.	Diabetic nephropathy data profile graph	44
3.16.	Liver response to a high cholesterol diet and phenobarbital data profile graph	45
3.17.	Hypothalamoneurohypophyseal system response to dehydration data profile graph	46
3.18.	Glioma cell migration: comparison of fast and slow invading cells data profile graph	47
3.19.	Dysferlin deficiency effect on skeletal and cardiac muscles data profile graph	48
3.20.	Treacher Collins' syndrome Tcof1 gene overexpression and knockdown effect on neuroblastoma cells data profile graph	49

3.21.	Visual cortex during the critical period for ocular dominance data profile graph	50
3.22.	Cigarette smoking effect on alveolar macrophage data profile graph	51
4.1.	Results of ICA+RF and ICA+KM+RF methods on public data-1	59
4.2.	Results of ICA+RF and ICA+KM+RF methods on public data-2	59
4.3.	Results of ICA+RF and ICA+KM+RF methods on public data-3	60
4.4.	Results of ICA+RF and ICA+KM+RF methods on public data-4	60
4.5.	Results of ICA+RF and ICA+KM+RF methods on public data-5	61
4.6.	Results of ICA+RF and ICA+KM+RF methods on public data-6	61
4.7.	Results of ICA+RF and ICA+KM+RF methods on public data-7	62
4.8.	Results of ICA+RF and ICA+KM+RF methods on public data-8	62
4.9.	Results of ICA+RF and ICA+KM+RF methods on public data-9	63
4.10.	Results of ICA+RF and ICA+KM+RF methods on public data-10	63
4.11.	Results of ICA+RF and ICA+KM+RF methods on public data-11	64
4.12.	Results of ICA+RF and ICA+KM+RF methods on public data-12	64
4.13.	Results of ICA+RF and ICA+KM+RF methods on public data-13	65
4.14.	Results of ICA+RF and ICA+KM+RF methods on public data-14	65
4.15.	Results of ICA+RF and ICA+KM+RF methods on public data-15	66
4.16.	Results of ICA+RF and ICA+KM+RF methods on simulated data-1	73
4.17.	Results of ICA+RF and ICA+KM+RF methods on simulated data-2	73
4.18.	Results of ICA+RF and ICA+KM+RF methods on simulated data-3	74
4.19.	Results of ICA+RF and ICA+KM+RF methods on simulated data-4	74
4.20.	Results of ICA+RF and ICA+KM+RF methods on simulated data-5	75
4.21.	Results of ICA+RF and ICA+KM+RF methods on simulated data-6	75
4.22.	Results of ICA+RF and ICA+KM+RF methods on simulated data-7	76
4.23.	Results of ICA+RF and ICA+KM+RF methods on simulated data-8	76
4.24.	Results of ICA+RF and ICA+KM+RF methods on simulated data-9	77
4.25.	Results of ICA+RF and ICA+KM+RF methods on simulated data-10	77
4.26.	Results of ICA+RF and ICA+KM+RF methods on simulated data-11	78
4.27.	Results of ICA+RF and ICA+KM+RF methods on simulated data-12	78

Page

4.28. Results of ICA+RF and ICA+KM+RF methods on simulated data-13 79
4.29. Results of ICA+RF and ICA+KM+RF methods on simulated data-14 79
4.30. Results of ICA+RF and ICA+KM+RF methods on simulated data-15 80
4.31. Results of ICA+RF and ICA+KM+RF methods on simulated data-16 80
4.32. Results of ICA+RF and ICA+KM+RF methods on simulated data-17 81
4.33. Results of ICA+RF and ICA+KM+RF methods on simulated data-18 81

TABLES

		Page
2.1.	R packages for Proposed Method	8
3.1.	Observed-Predicted class comparison table	23
3.2.	General information about public data sets	36
4.1.	Results of ICA+RF model on GEO data sets	55
4.2.	Results of proposed method on GEO data sets	56
4.3.	Gain of Proposed Method	57
4.4.	Descriptive statistics of gain of PM for bootstrap samples	58
4.5.	(ICA+ RF) Results of simulated data sets for 25 cases at each class	69
4.6.	(ICA+KM+RF) Results of simulated data sets for 25 cases at each class	69
4.7.	(ICA+ RF) Results of simulated data sets for 50 cases at each class	70
4.8.	(ICA+KM+RF) Results of simulated data sets for 50 cases at each class	70
4.9.	Gain of PM for 25 cases at each class scenerio	71
4.10.	Gain of PM for 50 cases at each class scenerio	71
4.11.	Descriptive statistics of gain of PM for different #Cases in each class	72

1. INTRODUCTION

Data Mining (DM) approaches have often been in use when handling multi-dimensional data in recent years. (6,8,12,30,36) The best example of this type of data are the ones obtained as the results of genetic research. Especially matrices containing the data of thousands of gene expression levels are obtained as a result of investigations of microarray data. However, it is not simply sufficient to obtain the data. Statistical models are required to analyze this data accurately and objectively.

Clinical trials in which DM methods are commonly used are the studies of microarray gene expression. (12,26,18,30,39) Given that the human genome contains approximately 40,000 genes, it is not possible to analyze so many genes one by one. But nowadays, it has become possible to analyze many genes at the same time by the developed systems which are based on automation. An important part of these analyses consists of the classification of genes and the exploration of important genes. In our country, the amount of resources devoted to research is gradually increasing. However, genetic researches are not to be practised on a great number of patients due to the cost and the difficulties in the recurrence of measurements as well.Therefore, some comments have had to be reached on thousands of genes of a small number of patients. Classical statistical methods face with problems in describing this type of data.Recently data mining methods such as Support Vector Machines (SVM), Decision Trees and Random Forest (RF) have been tried on this type of data and valuable results have been achieved. (12,30,38,40,52)

Clustering of genes has also been the starting point of many researches. The main objective of this approach is to bring together the genes with similar characteristics, using different means of distance measurement. This approach has gained importance especially in cancer researches, in ther sense that it provides information about the relationship between genes, in the treatment of disease. The most commonly used clustering methods in the literature are K-Means and Kohonen Map (KM) clustering methods. (9,14,44,46-48) Selection of important genes and classification of patients using gene expression data are also among the main

objectives of data mining analysis. (12,18,19,26,29,43) The main aim here is to find out a group of genes that affect the related phenotype. The selections of important genes to be determined affect the success and duration of the treatment directly. In particular, the identification of 'candidate'genes in drug development has become a standard approach.

1.1. Proposed Method (PM)

Machine learning methods are often categorized as supervised (outcome labels are used) and unsupervised (outcome label are not used) learning methods (26). In this study, these two methods were combined based on dimension reduction on microarray data sets` patient class prediction. The most important point is how to choose the method to be used while carrying out the classification of the patients with the assistance of gene expression data. It is not rational to expect a single method to have high true classification rate in every data set. For that reason, the method or the methods to give the highest performance under different scenarios must be chosen. Methods developed from the combination of several methods have been observed giving more generalized results, as the bias in the analysis can easily be eliminated in such combined methods. For this reason, methods will be brought together when the classification is done with gene expression inputs within the study.

First of these is the Independent Components Analysis (ICA) which is the dimension reduction method, the second is KM method which is the clustering method and the last one is the RF which is used as the classification method. Gene expression inputs are mostly those in which the number of subjects is far less than the number of the genes. There are thousands of genes belonging to a person. So, in these researches a lot of people cannot participate in the studies due to the cost restrictions. It means that such data do not meet the most important assumptions of many statistical methods. Therefore, using classical methods mostly lead to obtaining wrong or over-fited results. Gene expression inputs will primarily be generated by simulation model in the study. Afterwards, genes will be reduced to a smaller number of factors by the ICA to eliminate the problems of being multi-dimensional. The

purpose here is to form a common factor from the genes which have different expression levels. At the second stage, factors will be clustered by KM method. The aim here is to bring together the factors which have the same features. The reason for not doing the clustering at the first stage is the fact that the clustering methods may produce incorrect results when the number of dimensions increase. Consequently, similar factors which the independent genes constitute will be clustered in relation with the method the Independent Components Analysis (ICA) uses. At the last stage, the classification of the patients with RF is targeted by selecting a certain number of clusters among those selected randomly in 1000 trials randomly with bootstrap method. The reason for using the bootstrap method is to eliminate the bias in choosing the clusters to be used in classification. The reliability of the classification obtained consequently will be higher due to the fact that RF method uses the Bootstrap method in its own algorithms. For all data sets number of components for ICA is: (number of genes/3), number of cluster for KM is: (number of componenets/3). Parameters for RF are: 1000 tree and number of variables in each node split is at least (number of variables/3). (Figure 1.1)

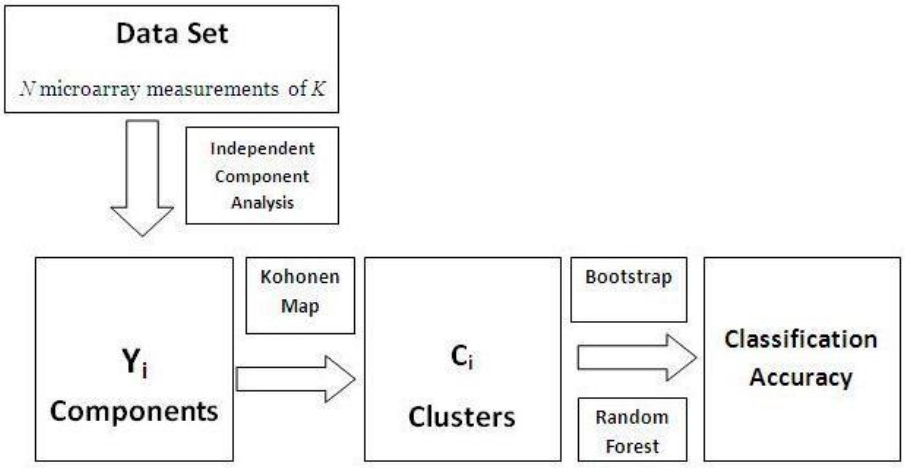

Figure 1.1. Pipeline of Proposed Method

2. GENERAL INFORMATION

2.1. Microarray Technology and Gene Expression Data

"In microarrays, thousands to millions of probes are fixed to or synthesized on a solid surface, being either glass or a silicon chip. The latter explains why microarrays are also often referred to as chips. The targets of the probes, the mRNA samples, are marked with fluorescent dyes and are hybridized to their matching probes. The hybridization intensity, which estimates the relative amounts of the target transcripts, can afterwards be measured by the amount of fluorescent emission on their respective spots. There are various microarray platforms differing in array fabrication, the nature and length of the probes, the number of fluorescent dyes that are being used, etc. The process from the cell samples to the microarray is shown in Figure.2.1" (55)

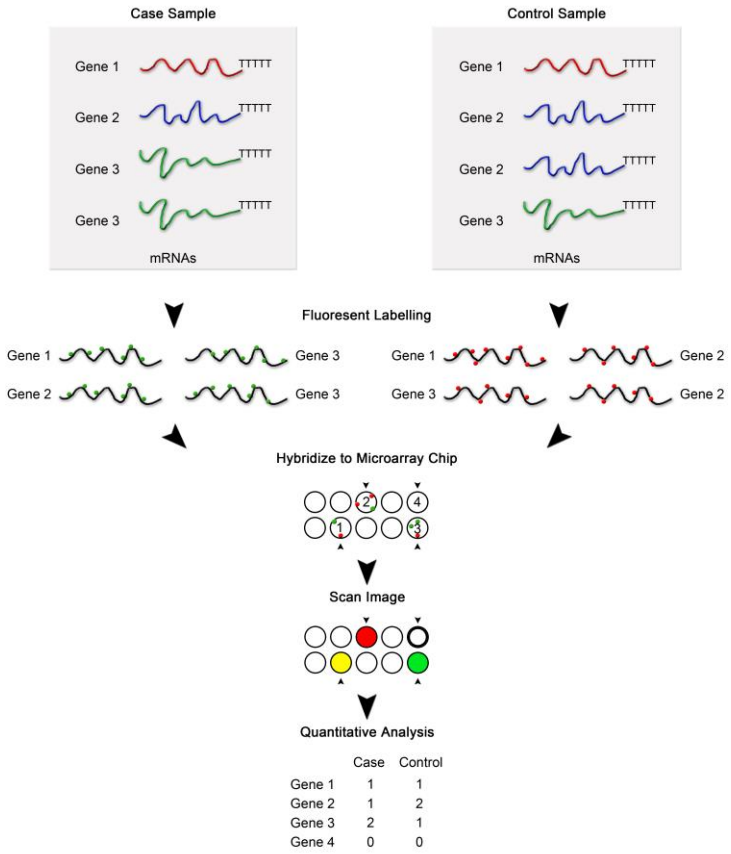

Figure 2.1. The process from the cell samples to the microarray

"Genomics and gene expression experiments can be used to identify new genes involved in a pathway, potential drug targets or expression markers that can then be used in a predictive or diagnostic fashion. Because the arrays can be designed and made on the basis of only partial sequence information, it is possible to include genes on an array that are completely uncharacterized. In many ways, the spirit of this approach is akin to that of classical genetics in which mutations are made broadly and at random (not only in specific genes), and screens or selections are set up to discover mutants with an interesting phenotype" (33).

In order to evaluate the expression level, the process uses the density of red and green point from image analysis. These densities are calculated by differences between foreground and background of specific probes on microarray chip. Microarray experiments focused on finding the log2 ratio of density of red and green. If this ratio is greater than zero: it means this gene is expressed and probe`s color is red but if this ratio is smaller than zero it means: gene is not expressed and probe`s color is green. In fact, if the ratio is equal to zero, this means: we cannot say anything about gene and probe`s color is yellow (Figure.2.2)

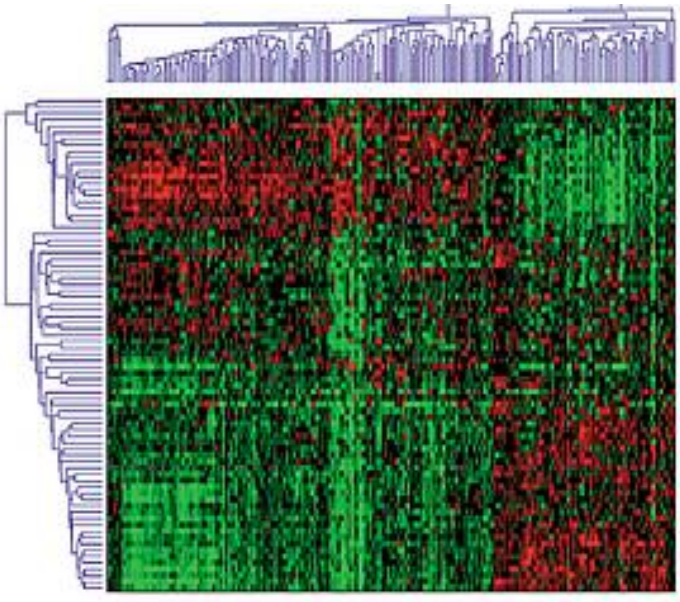

Figure 2.2. The reflection image of the basic expression data.

$$Density\ of\ Red = R_{fg} - R_{bg} \quad (2.1)$$

$$Density\ of\ Green = G_{fg} - G_{bg} \quad (2.2)$$

$$fg = foreground,\ bg = background \quad (2.3)$$

$$Expression\ Lev. = log2\ (Density\ of\ Red\ /\ Density\ of\ Green) \quad (2.4)$$

The description of gene expression data is shown in Figure.2.3 and the pipeline of gene expression analysis is shown in Figure.2.4.

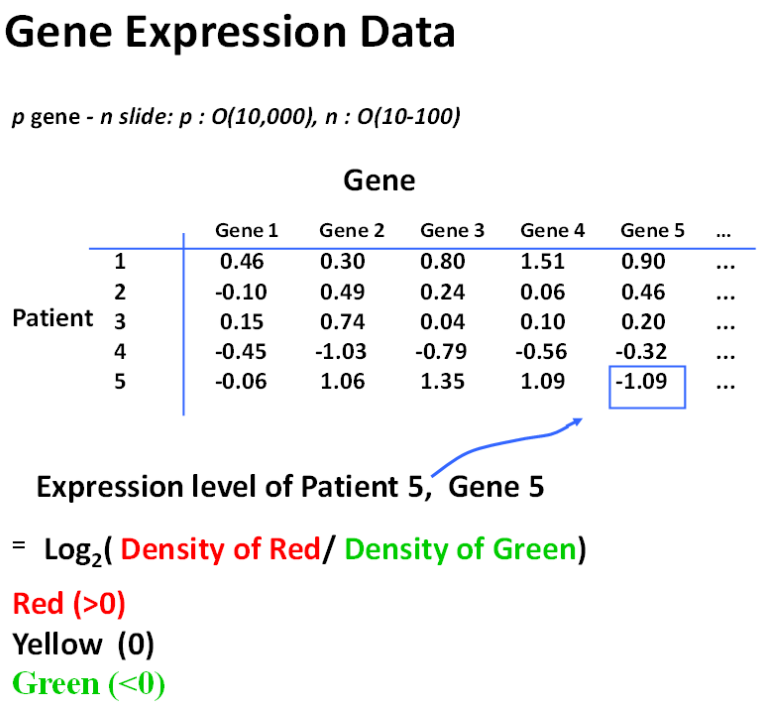

Figure 2.3. The description of gene expression data.

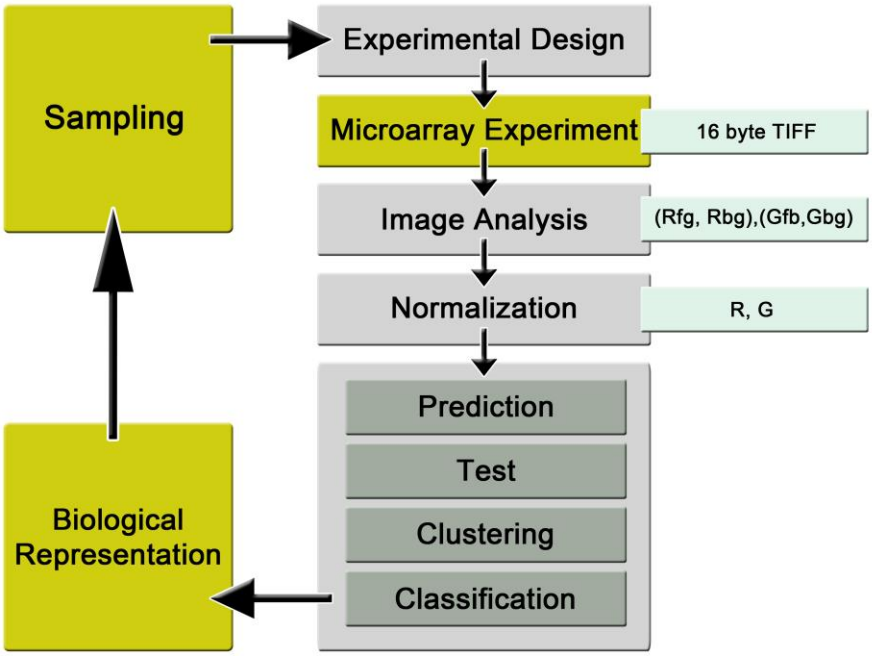

Figure 2.4. Pipeline of gene expression analysis.

2.2. Fold-Change Analysis

Fold-change analysis finds the expressed genes by calculating the ratios (or log ratios) or differences between two conditions and considering genes that differ more than an arbitrary cut-off value (10). Let x_{ij} and y_{ij} denote the expression level of i^{th} gene in j^{th} replicate and $\overline{x_{i.}}$ and $\overline{y_{i.}}$ denote the mean expression values of i^{th} gene in control and treatment groups respectively. Then, fold-change values can be calculated for ratios and differences as the following (56):

$$FC_i(ratio) = log_2 \frac{\overline{x_{i.}}}{\overline{y_{i.}}} \quad (2.1)$$

$$FC_i(difference) = log_2 \overline{x_{i.}} - log_2 \overline{y_{i.}} \quad (2.2)$$

After calculating fold-change values, when we choose the cut-off values as 2-fold difference, the genes whose fold-change values are greater than 2 (or less than -2) will be identified as different expressed genes.

2.3. Data Mining Tools

After the data is obtained and stored in gene expression patterns, the most important phase is to determine the effect of biological patterns on the observed phenotype. Analyses of genetic data require the use of data mining methods beyond the known classical statistical methods. The most important reason of this is that DM methods have the advantage of ` determining the non-linear relationship ` in very high-dimensional data. In recent years, open source softwares have been developed to make these advantageous methods possible to be used. (STATISTICA Data Miner, SAS-Enterprise Miner, ORANGE, Rapidminer, R) The most important of these is R software as follows:

2.4. [R] Software

R is an open source and free software which is standard for all researchers working in the DM. The logic of R program is that it has ` analysis packages` which can be used for each analysis. After downloading the package of the analysis, the users continue to conduct the analysis based on the guide. R`s most important advantage is that it is much faster than other packages. Very large data sets can be analyzed in a very short time by a standard computer configuration (ie, 2GB of RAM and a 2.13 GHz processor). All information about this program can be accessed in http://www.r-project.org/, and the complete list of packages can be accessed in: http://cran.r-project.org/src/contrib/PACKAGES.html.The R packages of DM methods within the scope of this study are provided in Table 2.1. Some other source packages, which can be used for the related methods, can also be found apart from these packages.

Table.2.1. R packages for Proposed Method

Analysis	**[R] Package**
Bootstrap	boot, bootstrap
Random Forest	randomForest
Kohonen Map	kohonen
Independent Component Analysis	fastICA

2.5. Literature Review of Random Forest

The use of the RF method in gene expression studies has become common since the 2000s. The majority of the studies in this area have been realized with the goal of classification of the patients or discovery of new genes. (12,30,38,40,52) In their studies, Moorthy et al. (36) indicated that the researchers would have advantage by identifying the important genes among a large number of genes and expressed that this might be guiding for further studies. They used 10 different data sets with different patients and number of genes for this purpose, and reported that the RF method provided a better discrimination compared with SVM, K-Nearest Neigborhood (KNN) and Diagonal Linear Discriminant Analysis (DLDA). In another study, Diaz-Uriarte (12) et al., reported that the RF method provided a very sufficient classification performance especially when the independent variables are in noise situation. They have implemented the RF in binary and multi-class classification problems for this purpose. They expressed that the RF had produced results comparable with DLDA, KNN and SVM methods according to the results. In this study, using sqrt (number of variables) for number of parameter for every split (mtry) parameter for the RF has been guiding for the use of this method. They have simulated the data with 2 and 4 classes, 25 cases in each class and 1 to 3 dimensions apart from 10 public data sets as data set. Okun et al. (40) performed the RF application on two real data sets (SAGE and Colon Cancer) and used Area Under Curve (AUC) value as the measure of performance in their studies. They observed that the AUC value increased as mtry increased, according to the results of six different RF models. In this way, it was revealed that the real data sets can be affected by mtry as well as the artificial data sets. Nannapaneni et al. (38) added "Transcriptional data for the B. subtilis wild-type strain 168 and its isogenic sigB mutant BSM29 were analysed using RF". They used two different RF models (expression RF and kinetic RF) according to their biological differences and demonstrated a significant gene set. (196 sigB Regulo members) Wang et

al. (52) have attempted to anticipate the cancer with a single gene, unlike other studies carried out. They have suggested that a single gene might be an important and simple way to classify patients. In this way, they have stated that a clearly understandable model of DM could exist instead of complex models. In this study, the RF was used with DLDA, KNN and SVM on eleven public data sets. They used Leave One-Out Cross Validation (LOOCV) accuracy as a measure of performance. Liu et al. (30) proposed a new method to be used in finding out new genes in their studies (Lagging Prediction Peephole Optimization-LPPO) and tested this approach on six real data sets using the methods of DM. In particular, they reported that methods such as the RF, SVM outperformed the other methods. They indicated that they chose the genes which give more information with the method proposed, and these genes produced better results with the RF. As can be seen in the studies related to literature, the RF is compared with other DM methods on gene expression data analysis and it is reported to have given better results than many of the others.

2.6. Literature Review of Independent Component Analysis

The most important issue in analyzing gene expression data is that they have high-dimensions. The dimension reduction is the most appropriate choice for a very small number of patients and for the analysis of the data which have a large number of genes. In this way, the relation of the biological factors not thought to be unrelated with each other will be put forth. PCA and ICA are the most commonly used methods for this purpose. The ICA has been reported to have been more successful in finding genes that are expressed with each other similarly especially in noisy data. (25,37,43) Because of the PCA projects the data will be collected into a new space spanned by the principal components. In contrast ICA aims at finding linear representation of

non-Gaussian data for these components are statistically as independent as possible. (25)

ICA was first used by Hori et al. (18,25,39) to classify the yeast gene expression data. Then Liebermeister (29) used the ICA to reveal the expression modes. In this way, the popular use of ICA has started. Lee et al. (26) have tested six different studies of ICA model and reported that the independent components that they obtained prior to the clustering analysis increased the performance. In the same study, they also indicated that the ICA produced better results compared with PCA prior to the cluster analysis. In another study, Suri et.al (43) indicated that ICA was more successful in finding periodicaly co-regulated genes' components than PCA. And Najarian et al. (37) factorized the data that they obtained in DNA microarrays with the ICA and then they identified the intra-cluster protein interaction analysis with the cluster analysis. The studies carried out show that the ICA is successful in the dimension reduction of the gene expression data. With the help of this success, the components created by ICA are obtained as possible as independent and create a source of data for further analysis.

2.7. Literature Review of Kohonen Map

One of the most widely used approaches in obtaining information from gene expression data is to reveal similar genes mathematically. Methods used for this purpose are generally called the clustering methods. Similarities between the genes are modelled by using certain distance measures (Euclidean, Euclidean Square, Manhattan, etc.). The KM is frequently used to analyze gene expression data in many studies in the literature. Törönen et al. (47) used the KM algorithm on yeast gene expression data set in their studies. They transformed the data into 2D data matrix model and tried to find out similar genes in an iterative way. They reported that the neurons in the KM were successful in bringing functionally similar genes together according to

the results they obtained. This approach has the qualification that may work as a manual to studies which particularly aim to cluster gene expression data. Dinger et al. (14) developed an algorithm that may be an alternative to KM, Fuzzy C-Means, K-Means, Hierarchical Clustering algorithms in their studies: The diffraction-based Clustering. Though their method provided better clustering in the study with five public data sets, the KM was reported to have given better results compared with other widely used algorithms.

In another important study, Covell et.al (9) have tried to distinguish normal and tumor expression groups for 14 different tumor types. This is especially an important study in understanding how the KM performs in different types of diseases. They have reached 80% correct classification on 14 different tumor types. Although it is well perceived as a classification work, the KM can be expected to give this performance without any dependent variable. While good estimates were obtained for Leukemia, Central Nervous System, Melanoma tumor types, they reported that they had obtained lower performance than expected for colocteral, overian, breast and lung tumors. In this respect, the study is an important guide for researchers to work on the type of the tumors on which better performance was obtained in the subsequent studies.

The KM is available as packs in much analysis software. The most important of these are MATLAB, R, and SAS. Apart from these, another analysis tool researchers prefer is GENECLUSTER. Tamoyo et al., (44) have used the KM visually with this package and demonstrated 3x2, 4x3, and 6x5 size KM models. This study can only be described as important for the KM by defining the differences in the selection of size. In general, size selection with the KM is an important problem. Although the researchers prefer to build smaller-sized models in general, Torkkola et al. (46) preferred a size of 80x80 in the KM models with artificially created 6400 elements. Moreover, they

reported that the clusters obtained can be interpreted in a biological point of view. The interpretation of the results obtained in such a large size is an important finding.

Trosset (48) held the study which compared the performance of the KM models in a different point of view. In this study, they reported that although the K-Means method alone cannot give better results than the KM, it gave better results when it is used together with a Multi Dimensional Scaling (MDS) method. They used 6220 yeast frames taken from Cell Cycle Expression Database for this purpose. In this respect, it can be regarded as an important study exploring the KM.

2.8. Literature Review of Dimension Reduction Based Classification

The first step in the analysis of microarray gene expression data is the dimension reduction. One of the main objectives of the thesis is the strengthening of the dimension reduction step that is to be done to achieve a better classification result. The classification and clustering analysis carried out after the dimension reduction have been indicated increased the performance in other studies in the literature. Dia. et al. (21) examined the effects of three different dimension reduction techniques on the classification analysis (Partial least squares-PLS, sliced inverse regression-SIR and PCA) in their studies. They did not work on the performances of the methods alone. They implemented the models they had proposed on colon cancer and leukemia data sets. They used the Correct Classification Rate (TCR) and the computational efficiency measures in comparing the performances. They reported that the models obtained with PLS and SIR models yielded more descriptive results than those of PCA here. They stated that the analysis based on PLS method proved better results in terms of computational efficiency and TCR. Released recently, in another study which uses size reduction and classification methods together, Bayer et al. (4) used different classification methods and the PCA method together. The work is modeled as follows: a)

to analyze the train set with PCA b) to use the first 5 components in SVM or RF HIGH algorithm. Interesting results have been obtained even though this method has not been used with such a small number of components in similar studies. The difference between the two approaches seems to be close as the number of samples increases in regression models when only root mean square error (RMSE) values of RF and PCA + RF models are compared.

In another study, Komura et.al (24) stated that the dimension reduction with PCA affect the classification studies in a bad manner as they are already known. To eliminate this they proposed a dimension reduction method based on the SVM which is a supervised method (Multidimensional SVMs). And in a similar study Lei et al. (16) had used PCA and the Recursive Feature Elimination (RFE) methods for the analysis of SVM. They indicated that the performance descends only with RFE and the modelling of RFE+PCA before the SVM is used demonstrates more consistent results. In this respect, it can be said that these two studies added innovation to the related field. Such a study that revises a method focused only on the classification in this way is not frequently observed in the literature a lot. H.Li et al. (28) referring to the importance of the reduction and suggesting a new dimension reduction method. (e.g: LDA) On the other hand they have stated that different approaches may be useful with a combination of different methods. B.Li et al. (27) used PLS, PCA, LDA, ICR (ICA algorithm used for regression), and dimension reduction methods such as Locally Linear Discriminant Embedding (LLDE) which is not used in the literature a lot with KNN and RF techniques in their studies. In particular, they reported that the TCR values which belong to the classification methods with LLDE were higher compared with other dimension reduction methods.

3. MATERIALS AND METHODS

Data mining methods and microarray data analysis consist of five basic steps. These are respectively indicated in Figure 3.1.

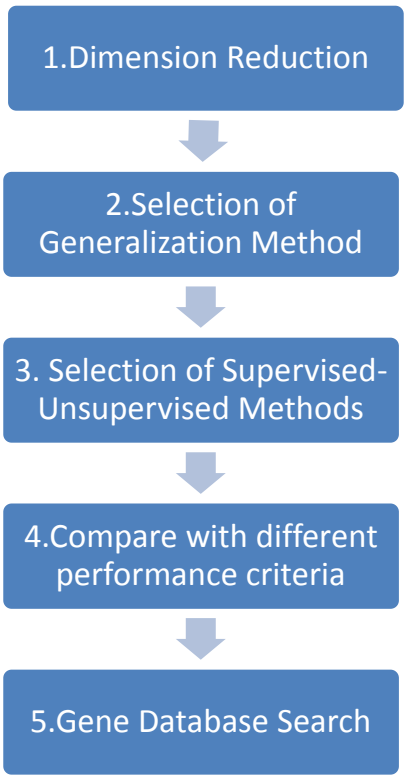

Figure 3.1. Microarray Gene expression analysis flow-chart on Data Mining

To know which class the individuals belong to in the study is called the Supervised Learning.For example: If the Patient / Not Patient information is known, algorithm can guess the model which will provide the classification more neutral and realistic. If this information does not exist, and if purely mathematical projections are made by using the distance measures as in clustering analysis, this approach is called the Unsupervised Learning.

3.1. Dimension Reduction

The first step in the analysis of these data types which are multi-dimensional due to their nature is to reduce the size of the data and to uncover linear or nonlinear relationships. The first step in the analysis of these data types which are multi-dimensional due to their nature is to reduce the size of the data and uncover linear or nonlinear relationships. We can explain this approach in the following way; for example, if we have the expression value of 6000 genes that belong to 20 patients; it is incorrect to analyze the data directly, for there are many known or unknown relationships between the genes. The results will be biased and inaccurate if analysis such as clustering and classification are performed without debugging these relationships. Therefore, in our example, it is required to reduce the information belonging to the 6000 gene to a smaller number of 'factor', usually up to the square root of the number of genes. Each factor obtained represents the information of a particular group of genes. After many studies on dimension reduction, it has been observed that one of the data mining methods, ICA, provides a better factorization instead of the well-known 'Principal Component Analysis'.

3.1.1. Independent Component Analysis

ICA is a statistical technique that aims to reveal hidden factors in the data sets, taking random variables, measurements or signals into consideration. In general, it focuses on creating a model in large data sets with the help of multiple variables. In the model, the variables uncover the hidden factors by coming together. The most important assumption for the hidden factors is that they do not show normal distribution and are completely independent of each other. ICA has common aspects with Principal Component Analysis (PCA) and Factor Analysis (FA). However, ICA gives more effective results in revealing the hidden factors especially with the expansion of the data sets.With the help of ICA, more effective analyses are performed by reducing the dimension in multi-dimensional data.Using it with classification and clustering methods has been started to be preferred, nevertheless it is used alone

most of the time.It especially gives more reliable results and alternative methods (PCA, FA), during the dimension reduction of microarray data sets.Together with dimension reduction,it has been used in eliminating bias (Whitening) and centralizing the data with self-value decomposition.

The best example to be given when explaining the ICA is the "Cocktail Party Problem". There are many sounds at a cocktail party. (music, sound from the outside, the sounds of people). If the voices of two people are demanded to be distinguished from other sounds, at least two microphones are placed at an equal distance from the individuals. Then, the sounds from each microphone are analyzed as one model. Factors are elicited including two variables in each model. We can generalize this example for gene expression data, too. Each gene carries different information. ICA is a method that generates very fast and accurate results to create a statistical model to be used in distinguishing these genes. (Figure 3.2., Formula 3.1)

$$\mathbf{r}_j = a_{j1} \times \mathbf{u}_1 + a_{j2} \times \mathbf{u}_2 + \cdots + a_{jn} \times \mathbf{u}_n$$

$$\text{ICA representation} = (a_{j1}, a_{j2,} \cdots a_{jn}) \qquad (3.1)$$

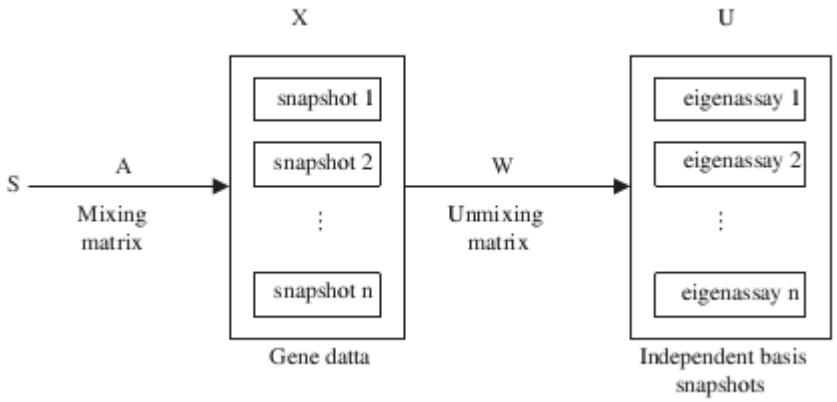

Figure 3.2. ICA analysis pipeline

3.2. Classification and Regression Trees

Classification and Regression Tree (CART) is the algorithm that uses RF for creating all the trees in forest. Therefore researchers should know the characteristic overview of this algorithm. CART is an algorithm which helps to predict the numerical or categorical variables under the impact of a group of numeric or categorical factors. It is among the methods used when the dose is desired to be estimated with the gene expression data. Being in the structure of the tree it reacts in response to the problem. As an information criterion it uses is the "Gini index". The input variables (p) can be divided only by 2 in the phase of division. For this reason it has a disadvantage in the input variable with the presence of more number of categories.

$$gini\,index(D) = 1 - \sum_{j=1}^{n} p_j^2 \qquad (3.2)$$

3.3. Random Forest

In recent years, one of the most widely used method in the analysis of gene expression data is RF.The most important advantage of it is that, although very necessary, if the researcher does not do dimension reduction, RF algorithm can constitute a great algorithm for the classification and prediction, using gene expression data of a large number of genes. (12,19,25,30,36,38,40,52) It has been proven that it has given much more successful results than a single decision tree algorithm in gene expression data. (12,19,25,30,36,38,40,52). RF is a structure which is composed of many (thousands) decision trees. Sample data is selected in the data set for each of the tree in the RF by the bootstrap method during this analysis, and 2/3 of the data selected is used to generate a tree and a classification is carried out. These classifications poll "votes".RF algorithm selects the tree which receives most votes of all the trees in the "forest" and uses its classification.The tree with low error rate is a better classifier.

Error rate in RF depends on two things:

i) The correlation between the two trees: The more the correlation increases, the more the error rate increases.

ii) The error rate of each tree.

And some advantages of RF are as follow:

1) Extreme cohesion does not exist in RF.

2) You can generate as many trees as you desire.

3) It is a fast algorithm.

4) The obtained RF value can be stored for use in other data sets.

5) It is a very effective method in the analysis of missing data, the TCR continues despite missing data.

6) Thousands of genes can be used without any elimination.

7) It can be used in unsupervised clustering method.

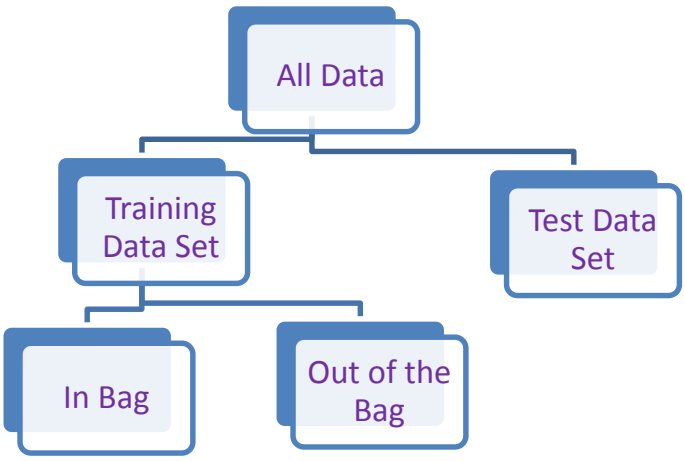

Figure 3.3. Random Forest Algorithm Flowchart

3.4. Selection of Generalization Methods

One of the most basic steps in the analysis of gene expression analysis is the generalization of the analysis. “Generalization" is used to state that DM methods, without any discrimination, are "model-based". If the classification is to be created, a classification model is constituted first and then the level of expression of a new incoming patient is expected to be estimated or the patient is expected to be assigned to a specific group according to this model. If all of the patients’ data from the entire data set is analyzed at once, the results being depended on chance are likely to be high. In other words, because the algorithm creates the model for the data set, it recognizes the patients’ group and creates a model that works only for this group of patients. It gives incorrect results in the patients who are from the same population but do not match this group of patients. Therefore, the analysis models are to be established by extracting some of the patients and the models obtained are to be tested with the party extracted. The most popular methods used to accomplish this are as follows:

3.4.1. Bootstrap

Bootstrap is a simple and reliable method which is used in statistics and non-parametric estimation problems such as standard deviation, confidence interval. This method is based on the sample drawing frequently by replacing it in a particular data set. Various data sets in amount and size can be created in a data set of any size by the re-sampling of the observations substituting depending on the chance. So that as much information as possible can be obtained from the existing data set. The mentioned method is defined as Bootstrap (Resampling) method. The bootstrap method has also some other advantages beside the ease of application and usefulness. In classical statistics, the estimation is based on the assumption that the studied variables show a normal distribution.

On the other hand, the statistical estimates are achieved by taking samples from the data set by chance in Bootstrap method. With this method, correct answers can be

taken even in very small data sets, and parallelism with classical statistical results is provided, and immediately, almost all statistics can be analyzed in large data sets.

3.5. Kohonen Map

Only KM clustering method within DM is focused on in this section. And the advantage of this method comparing with K-mean method which is more widely known and still used in some genetic researches is as follows: There is no obligation to determine the average number of clusters at least two and the maximum number of clusters not less than or equal to the number of observations in K-method. For this assumption of K-Means method is challenging especially due to the fact that in the genetic researches of “the less patient many genes” scenario. However, it has been reported in the previous studies that K-Means method could not be successful in gene expression data in which a large number of `contrary’ observations existed. (9,11,14,44,46-48)

KM, also known as Self Organizing Map, is a sort of neural network used for clustering purposes. This network algorithm is used in distinguishing the data whose group is not known prior to analysis and classifying into clusters independent from each other. Variances within the clusters are small, while they are great between each other. The main point in the analysis is the neurons. And these neurons consist of two layers: Input and Output Neurons. (Fig.3.4.) All the input neurons are connected with the output Neurons. These connections are expressed by measurements so-called "Power" or "Weight". Once the algorithm works the output neurons compete to connect the most neurons to themselfes. The Output "map" is a map of neurons which have a two-dimensional view; grid structure unrelated with each other. It is among the unsupervised learning methods as it does not need any target variable. (11,14,44,46-48) The algorithm randomly assigns weights to output neurons for input neurons first. And it assigns data to the output neurons according to the most powerful weight. At the end of the analysis, similar data diverge in the same place on the grid, and different ones detach the divergent grids.

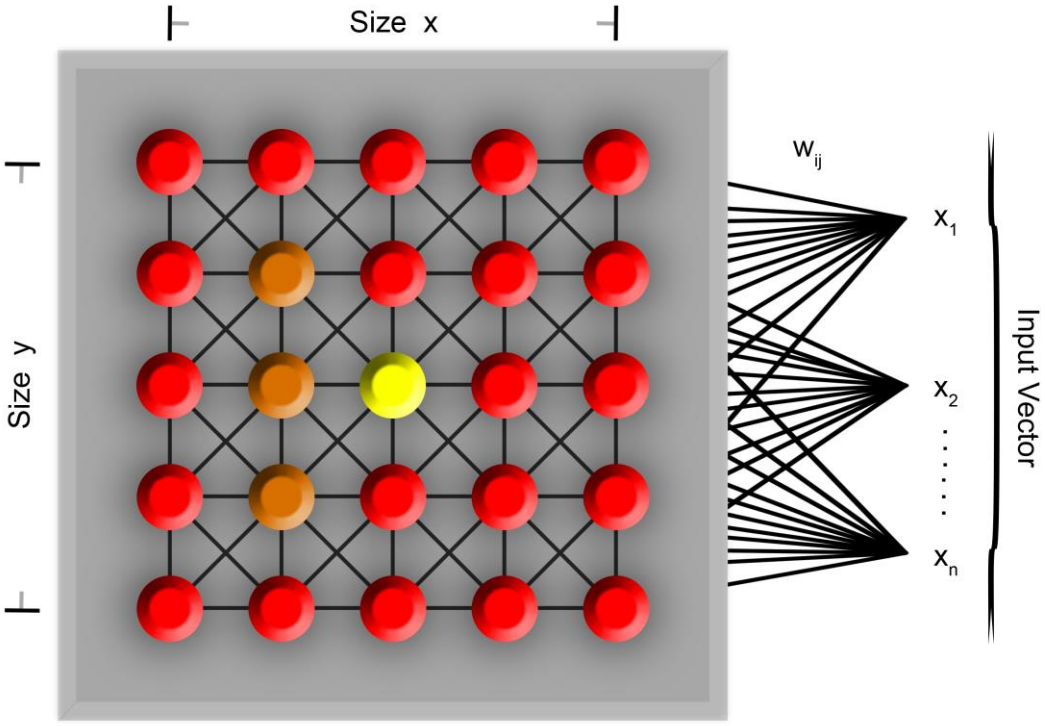

Figure 3.4. Visualization of Kohonen Map Clustering

3.6. Performance Comparison Criteria

Classification of individuals with gene expression data is among the common practices in DM.However, there is no such method which gives the best result of classification for each data set in DM.Taking the studied population, age group or disease, the number of genes in the data set analyzed, the number of patients, the mean and variance of the signal for the genes, and correlation between genes into consideration, different methods may give more reliable and accurate results. For this reason, all possible methods or the set of methods must be tested, and the method which gives the best result should be preferred. How can we distinguish these methods one from another then? The measure of True Classification Rate (TCR), Precision, Recall and F-Measure are recommended to be used in the studies in which more than two groups of patients (multi-class classification studies) take place.

Table 3.1. Observed-Predicted class comparison table

	Predicted Positive	**Predicted Negative**	**Total**
Actual Positive	TP	FN	AP
ActualNegative	FP	TN	AN
Total	PP	PN	N

TP: true positives (predicted positive, actual positive); TN: true negatives (predicted negative, actualnegative); FP: false positives (predicted positive, actual negative); FN: false negatives (predicted negative, actual positive), N: Number of Sample Size

1) **True Classification Rate (TCR)** measures proportion of actual positives and negatives which are correctly identified.[1]

$$TCR = \frac{TP+TN}{TP+TN+FP+FN} \quad (3.3)$$

2) **Recall (Sensitivity)** measures the proportion of actual positives which are correctly identified.

$$Recall = \frac{TP}{TP+FN} \quad (3.4)$$

3) **Precision (Positive Predictive Value)** is the proportion of positive test results that are true positives.

$$Precision = \frac{TP}{TP+FP} \quad (3.5)$$

4) **F-Measure** is a measure that combines precision and recall is the harmonic mean of precision and recall, the traditional F-measure.

$$F = 2\ x\ \frac{Precision\ x\ Recall}{Precision+Recall} \quad (3.6)$$

[1]Data sets for application part have 2 or more than 2 classes. Therefore *one vs all*, means select one class and merge other ones, approach accepted for evaluating the performance criteria.

3.7. Gene3E

With technological improvements we have a chance to collect and save high dimensional genetic data. These advantages arise with a potential difficulty of analysing. Many methods are useful to solve this problem. The main issue is which method has robust and unbiased results. We proposed a new tool which has four different “sub-analyze” steps: Dimension reduction, generalization, classification and clustering. Users are able to choose different combinations of methods and change their parameters easily. In this way, they are going to find the best approaches for analyzing data. We have developed a Java tool named “Gene 3E” which is essential to genetic research with data mining methods. Our tool runs R scripts using RServe package. Analyzing high-dimensional data (i,e: microarray gene expression, Single-Nucleotide Polymorphism (SNP)) is a common problem for biologists. In recent years many tools and packages were proposed by scientists. Some of them focused on limited analysis (only dimension reduction / only clustering/ only classification) and some of them are very hard to use by users whose background is biology or medicine**.** One of the most important [R] tool for data mining is “rattle”, “rattle” brings together a multitude of R packages that are essential for the data miner but often not easy for the novice to use,“rattle” (15) includes many analysis, (i,e: impute, rescale, transform, decision tree analysis). In recently published paper, Zhang et al. (58) they developed a novel machine-learning tool with [R] software, named miRD (microRNA Detection) for accurate and efficient detection of novel pre-microRNAs. There are two sets of features and each was used to build a support vector machines (SVM) model (50). Marc Johannes et al. (22) introduced a [R] package called “pathClass”:

"This package aims at providing the user with comprehensive implementations of these methods in a unified framework in order to allow easy and transparent benchmarking. It is the first package implementing several SVM-based algorithms that are capable of incorporating network knowledge into the classification process. These tool and packages are useful for analyzing data, But none of them focused on genetic data or genetic issues." (22)

3.7.1. Objective of Tool

The aim of this tool is to create a tool that includes all necessary methods for analyzing high dimensional genetic data. Our tool has four sub-dimensions;

1. Dimension Reduction
2. Generalization
3. Classification
4. Clustering

This tool was created on [R] software because [R] increasingly provides a powerful platform for data mining, (15).

3.7.2. Methods of Tool

All analysis share one user interface function: *bootstrap.* Tool currently supports ICA, RF and KM methods.

3.7.3. Application of Tool

Developed tool has a Java interface essential to genetic research with data mining methods, SWT (Standard Widget Toolkit). It is used to create a graphical user interface (GUI) in Java. GUI abstracts the statistical background and gives user the opportunity to use combinations of statistical analysis methods without profound knowledge of how these algorithms are implemented. User can perform several steps to analyze genetic data such as, dimension reduction, generalization, classification and clustering methods. These steps can be executed optionally, depending on whether they are necessary or not for the specific case user trying to analyse. For each step, several methods can be used by user and each of their performances can be observed. Methods can be seen in the Figure 3.5. Diversity of methods gives the ability to compare lots of combinations (48 combinations) to analyze genetic data. Without a GUI, a user must spend effort to combine all of the algorithms in R scripts which would take a long time to implement and require tough nerves to organize each of them. By using our tool, user can recreate and try each combination limitless times without too much effort.

Methods can be customized to fit genetic data by using the button next to the method selection drop-down menu. Important arguments of the R methods are listed in the pop-up window. Arguments on RF, SVM, and ICA methods can be seen in Figure 3.5. These customizations give the ability to test different settings and almost complete R functionality. Standard outputs of R methods are printed throughout the run time of analysis to track the progress. And also the progress bar shows the status and stage of the current run. The output data is currently stored in a data file to be used later on R. On the other hand, a summary of each method is printed in the results text box in the interface.

3.7.4. RServe

The tool runs R scripts using RServe package by Simon Urbanek (49). RServe is a TCP/IP server which allows creating remote R sessions. Each connection to RServe has a separate workspace and can be linked against R libraries which then be used by client-side implementations. This approach simplifies the communication between R and programming languages (Java, C, PHP). Simplicity of communication enables the implementation of many tools. Thus, users can run complex R scripts without editing R code or even having any R knowledge.

A developed tool is specific to genetic researches. Due to the nature of this kind of data analysis, a dimension reduction is the mandatory step. For this reason we have integrated dimension reduction step with our tool. Whether or not using this reduction step, users have many “analysis tracks”. This optional usage has many advantages. For instance, one is not limited to use “only clustering” or “only classification”. Some genetic studies have highly correlated genes. Therefore dimension reduction step is not enough to find the best prediction model. Probably, an extra clustering step solves this problem. In other words, “curse of dimensionality” is not a valid belief for this tool. This tool can run several methods and give the ability to run statistical analysis methods on genetic databases. However estimating the best method selection and optimum arguments for each method must be found by running the tool a number of times. We will implement two optimization tools for method selection and argument optimization respectively. With the method selection optimization, user can select more than one method in any step. Each of the combinations will be run, results will be given in separate files and a performance of each method can be compared by summaries on the GUI. Argument optimization will be done by specifying intervals for several arguments of methods. Results of each run will be printed in order of their performance. These optimization methods will facilitate the analysis of genetic data.

3.7.5. Interface and System Requirements of Tool

The interface templates have been created after the determination of the parameters the user benefits from and is not needed depending on R scripts prepared in advance and the parameters that the packages have at interface. The user interfaces can be divided into three segments:

1. General screen
2. Parameter screens
3. Results screens

General screen has been tried to be designed as simple as possible to help the user reach the functions he wants quickly. In genetic data mining, the orientation of the group of defined methods through a selection in a flow will facilitate the use without causing a misunderstanding in the users' minds. An overview screen template is shown as a representation in Figure.3.5.

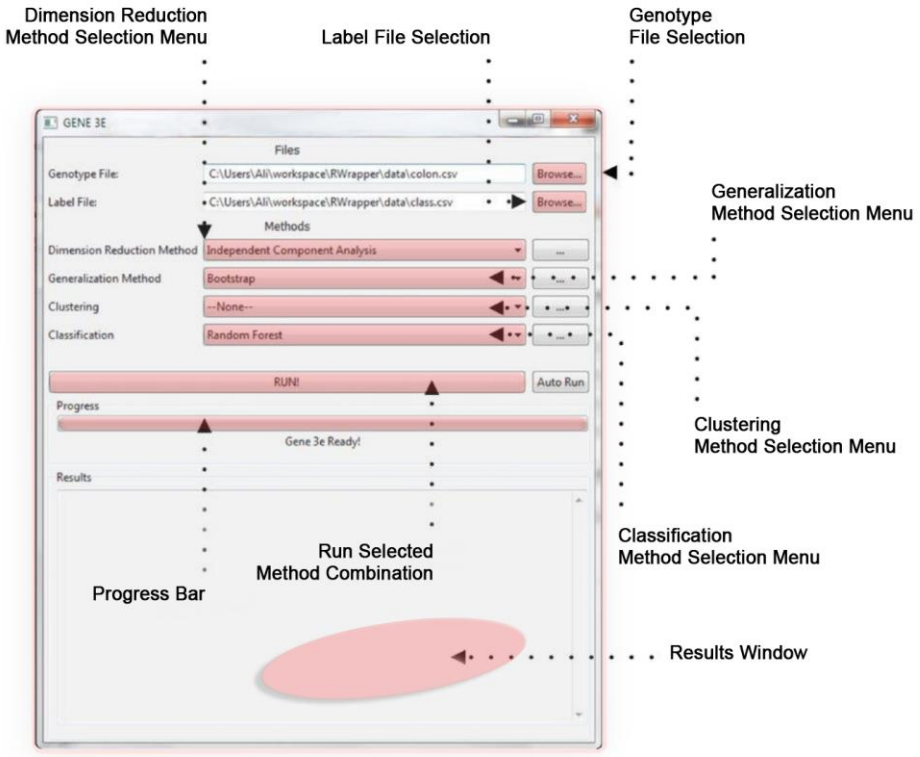

Figure 3.5. User interface of Gene3E

The menus in which the methods used take place in the file selection windows which will provide the selection and reading of the files in various formats including the genetic data are included in the interface. And the result screen which will include results of the analysis takes place at the lower part of the menu.

One of the greatest needs of users working in different disciplines is the matching process of the results obtained with the data in international databases. This has been considered as reporting and the result screens which allow circulate embedded into the system- on the Pub-Med. The Database for Annotation, Visualization and Integrated Discovery (DAVID) and so on, in which genetic and different biological / medical data is kept and the researchers who work in the field are taken as reference. The following figure includes a representative display of the above interface expanded for this situation.

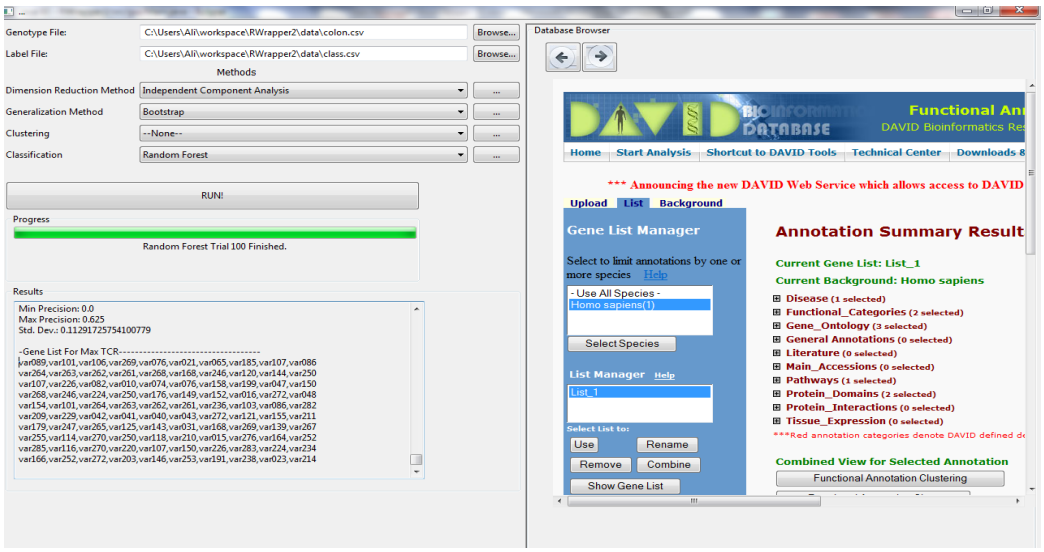

Figure 3.6. User interface of Gene3E with web links

The parameter screens for RF method have been designed to do the on-site optimization, the process of selecting the most effective parameters by the tool. This is also a requirement for the system workflow. Functions have been added to the analysis tool during the development process to be useful for the people who have no experience in analysis as well as those who have

experienced people. Certifications such as use case, collaboration diagram, sequence diagram, and so on, in which the components and the flow between the components are expressed for the flow of information between the screens of the tool and the realization of the reporting effectively, have been prepared.

3.7.6. Software Quality Attributes of Tool

The attributes of software such as usability, safety, effectiveness, flexibility, interoperability of different platforms and extensibility are referred as the quality attributes.

Although the information security is not a critical issue for local (the cases remote server connection is not needed to run the software) use, it should be handled sensitively in case of receiving the information on a server. The most important personal information and the most significant cost require ensuring the security of genetic data of one or more than one person. Software's providing data transfer and analysis in the local environment and not allowing remote connection except for the R packages and databases is to provide this reliability.

Interoperability of different platforms is a troublesome process if the analysis services are not presented in the form of web-based. To minimize these problems, the software has been taught to be prepared with Java and R database. Extensibility is a principle that should be considered for the upgradability of the software service in an efficient and easy way not to face inconsistency and drop in performance status in the next add-ons. This principle has been given importance in order to keep the ability of our tool, which allows the analysis of genetic data mining which is a constantly evolving academic field to scan on new methods, new files and communication format and the new reporting options and new databases alive.

3.7.7. Development of the Protocol of Communication Between Components

TCP / IP protocol is decided to be used to connect the improved tool to the R software platform-independent. TCP / IP is the abbreviation for Transmission Control Protocol / Internet Protocol. The expression, TCP / IP should not be considered as a single protocol. TCP / IP is a set of protocols. Each of which consists of a pile of protocols performing different jobs. In a computer network which is set up with TCP/IP is defined by three parameters, the computer name, IP address, MAC address (Media Access Control: Media Access ID). TCP / IP is a set of protocols that connect computers using these three parameters.

I have decided to write a middleware which uses the TCP/IP protocol for the analysis tool in connection with the definition given above. It provides connection from different programming languages with the R software by this layer.

In addition, a remote connection, the log on and file transfer are provided. With the help of this layer the place and R applications on the remote computers will be able to be connected. In addition, Rserve has been used in the development of the application. Rserve is an open source software which may be a model for the tool we developed.

Rserve's features:

- Fast and does not need to run R
- Binary transfer being able to transmit R object & binary R data
- Perform automatic data type conversions between the used programming language and the R
- Independent of the client
- Security can be configured to have the Rserve accept connections only from the local.
- File transfer allows the transfer of files between the client and the server.

An additional layer of communications appropriate for the software needs has been designed due to RServe's insufficient processing of large-scale genetic data.

The most important factor for the layer designed is to implement the most effective, fastest and safest approach into application in the established connection (remote or local) between the components(R software analysis tool).

Communication layer provided is integrated with the software base and R software. Opening session on R inside the code, the provided results received by operating R scripts embedded in the prepared source code were transferred to temporary displays. The results which R software transferred to are interpreted and presented to the user by the tool. The communication architecture of the tool is shown in the Figure.3.7.

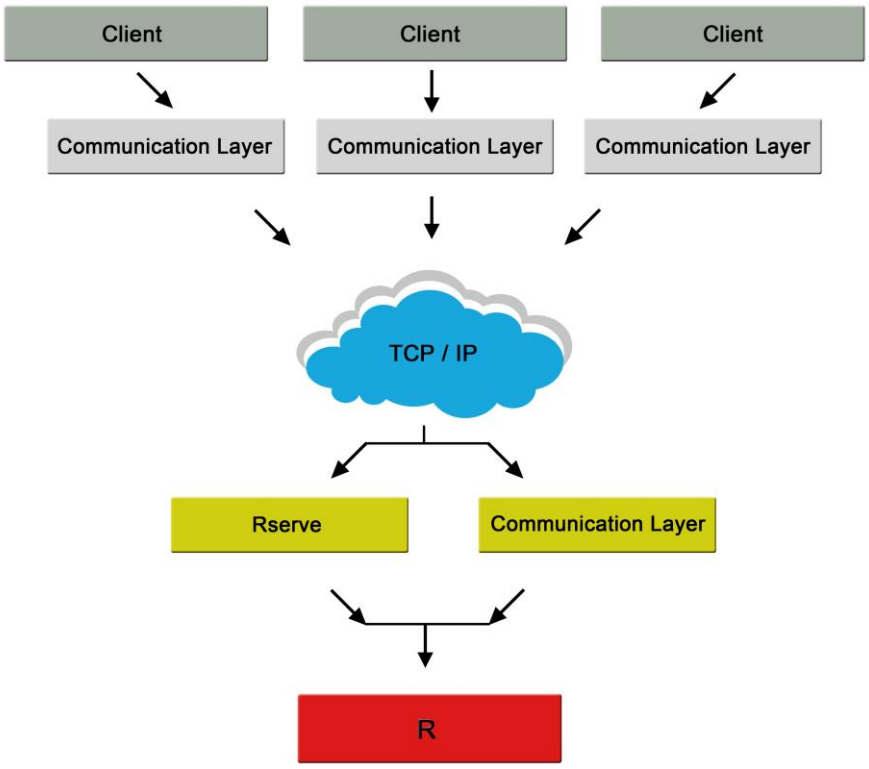

Figure 3.7. Client structure of Gene3E

To put the above figure briefly, when the clients (the segment of the analysis tool which is in interaction with the user) wish to use the R software during the analyze, within the scope of the process previously defined, the Rserve and layer special to our tool will be in use by using the TCP / IP as the communication layer to the remote or local resource access protocol.

3.8. Materials

3.8.1. Gene Expression Omnibus Data Set

In this study original public data sets from "The Gene Expression Omnibus (GEO)" were analyzed. *"GEO is a public repository that archives and freely distributes microarray, next-generation sequencing, and other forms of high-throughput functional genomic data submitted by the scientific community. In addition to data storage, a collection of web-based interfaces and applications are available to help users query and download the studies and gene expression patterns stored in GEO."* (53)

Two different approaches have been performed on these data sets:

1) ICA + RF

2) ICA + KM + RF

Comparisons of these approaches on these public data sets are shown below. These data sets have a raw data. Therefore data sets pre-processed then prediction models were built with application methods.

Pre-Process Steps:

1) Download full data sets from GEO
2) Select genes according to their fold change. (Select first 1000 genes according to fold change)
3) Create prediction models.

All results belong to mean performance of 1000-bootstrap samples for different performance criteria. (True Classification Rate, Precision,F-Measure,Recall)

Table 3.2. General information about public data sets

Data Sets	Organism	n	#Genes	#Class
Public Data-1: Skeletal muscle response to insulin infusion	Homosapiens	12	14,024	2
Public Data-2: Lymph node and tonsil comparison	Homosapiens	20	14,024	2
Public Data-3: Atrial and ventricular myocardium comparison	Homosapiens	25	17,011	2
Public Data-4: Metastatic prostate cancer	Homosapiens	167	10,386	4
Public Data-5: Asthma and Atopy	Homosapiens	29	17,011	5
Public Data-6: Quercetin effect on the colonic mucosa	Rattus norvegicus	8	24,606	2
Public Data-7: Tumor necrosis factor effect on macrovascular umbilical vein endothelial	Homosapiens	8	14,024	2
Public Data-8: Diabetic nephropathy	Homosapiens	6	9,461	2
Public Data-9: Liver response to a high cholesterol diet and phenobarbital	Mus musculus	12	14,102	3
Public Data-10: Hypothalamoneurohypophyseal system response to dehydration	Rattus norvegicus	30	24,606	5
Public Data-11: Glioma cell migration: comparison of fast and slow invading cells	Rattus norvegicus	7	6,058	4
Public Data-12: Dysferlin deficiency effect on skeletal and cardiac muscles	Mus musculus	20	9,726	4
Public Data-13: Treacher Collins' syndrome Tcof1 gene overexpression and knockdown effect on neuroblastoma cells	Mus musculus	9	14,102	3
Public Data-14: Visual cortex during the critical period for ocular dominance	Mus musculus	10	9,726	3
Public Data-15: Cigarette smoking effect on alveolar macrophage	Homosapiens	45	31,836	3

Public Data-1: Skeletal muscle response to insulin infusion

Summary: Analysis of skeletal muscles from non-diabetics after a 2 hours infusion of insulin. Glucose uptake by skeletal muscles in response to insulin is impaired in type 2 diabetes. Results provide insight into the molecular mechanisms regulating glucose homeostasis in response to insulin. (41)

Organism: Homo sapiens

Platform: GPL96: [HG-U133A] Affymetrix Human Genome U133A Array

Reference Series: GSE7146 **Sample count:** 12 **Value type:** count Series published: 2007/03/01

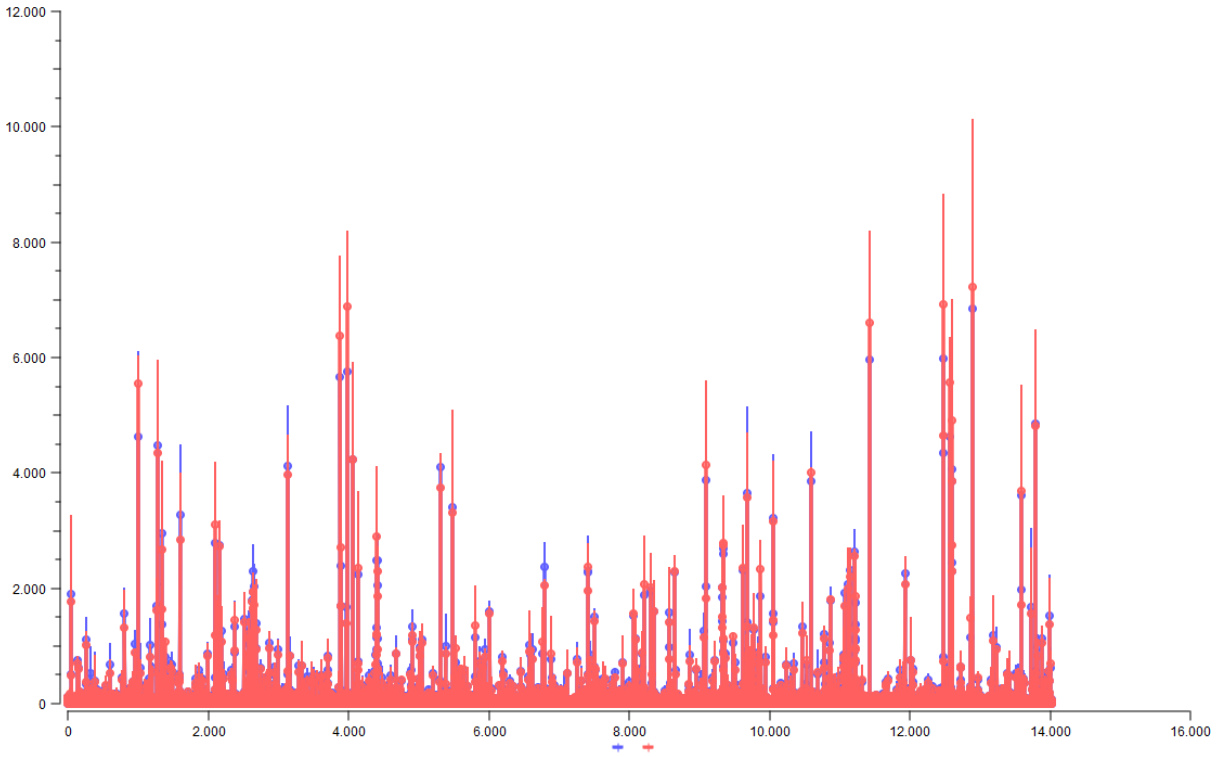

Figure 3.8. Skeletal muscle response to insulin infusion data profile graph

Public Data-2: Lymph node and tonsil comparison

Summary: Analysis of lymph node (sinuses) and tonsil (no sinuses), highly similar secondary lymphoid organs. Metastatic tumor cells are preferentially arrested in the lymph node sinuses. Results identify signature genes that are prime candidates for mediating adhesion of tumor cells to sinusoidal cells. (32)

Organism: Homo sapiens

Platform: GPL96: [HG-U133A] Affymetrix Human Genome U133A Array

Reference Series: GSE2665**Sample count:** 20 **Value type:** count **Series published**: 2005/12/31

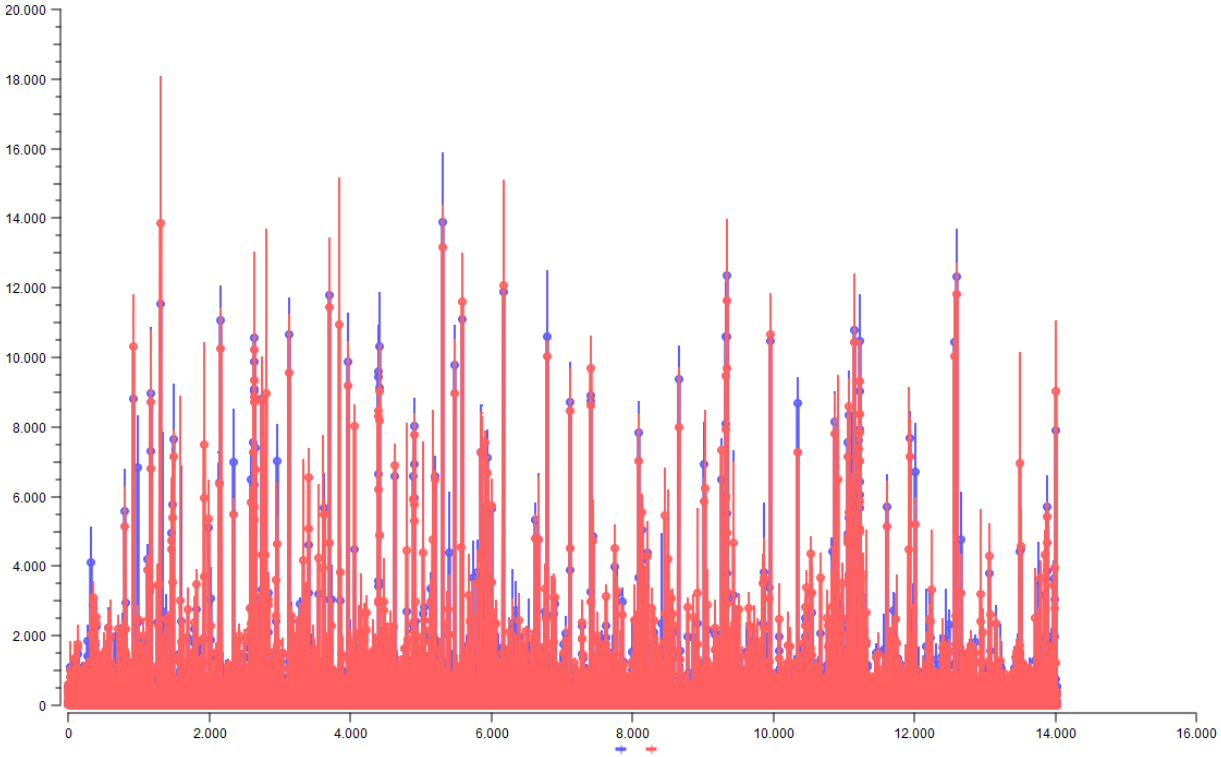

Figure 3.9. Lymph node and tonsil comparison data profile graph

Public Data-3: Atrial and ventricular myocardium comparison

Summary: Analysis of right atria and left ventricles of hearts. While both the ventricle and atrium contract, the atrium is also a source and target for neurohumoral signals. Results provide insight into the molecular basis for the ultrastructural and functional differences between the atrium and ventricle. (3)

Organism: Homo sapiens

Platform: GPL97: [HG-U133B] Affymetrix Human Genome U133B Array

Reference Series: GSE2240**Sample count:** 25 **Value type:** transformed count **Series published:** 2005/05/13

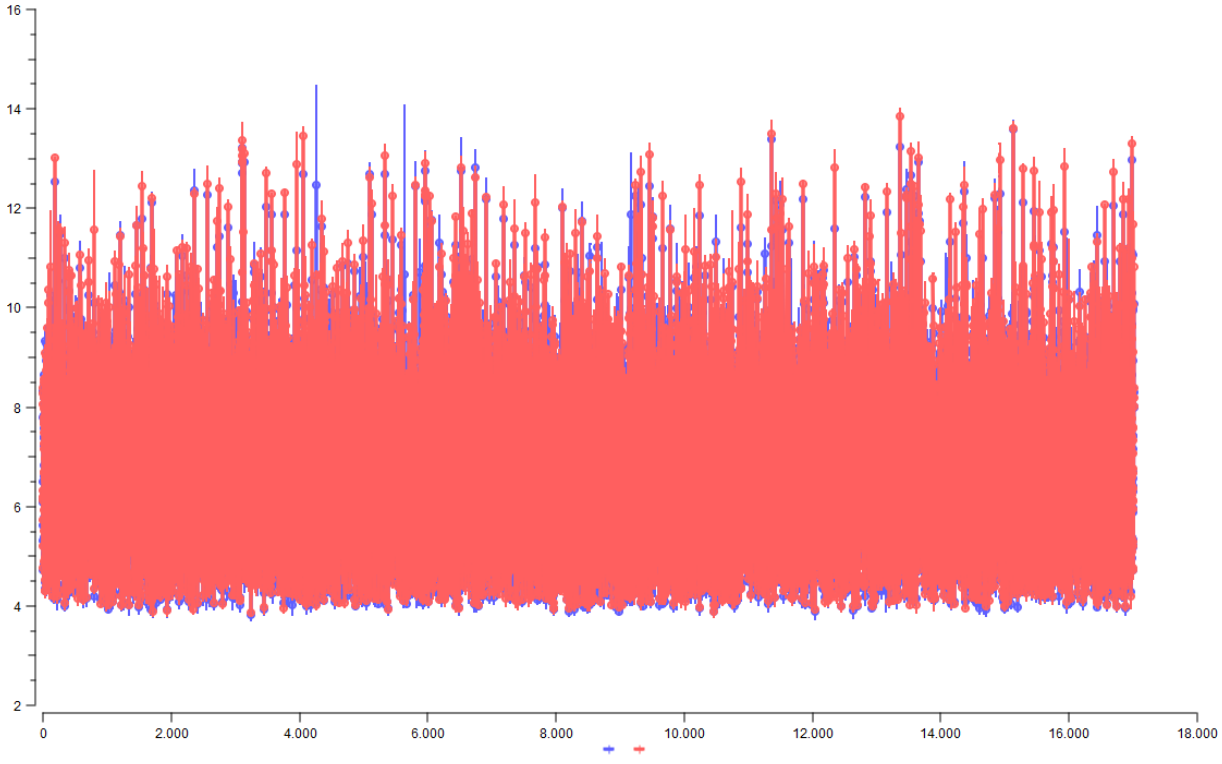

Figure 3.10. Atrial and ventricular myocardium comparison data profile graph

Public Data-4: Metastatic prostate cancer

Summary: Analysis of metastatic prostate tumors and primary prostate tumors. Normal tissue adjacent to the tumor and normal donor tissue also examined. Metastasis reflects the most adverse clinical outcome. Results provide insight into the molecular mechanisms underlying the metastatic process. (7)

Organism: Homo sapiens

Platform: GPL92: [HG_U95B] Affymetrix Human Genome U95B Array

Reference Series: GSE6919 **Sample count:** 167 **Value type:** count **Series published:** 2007/01/30

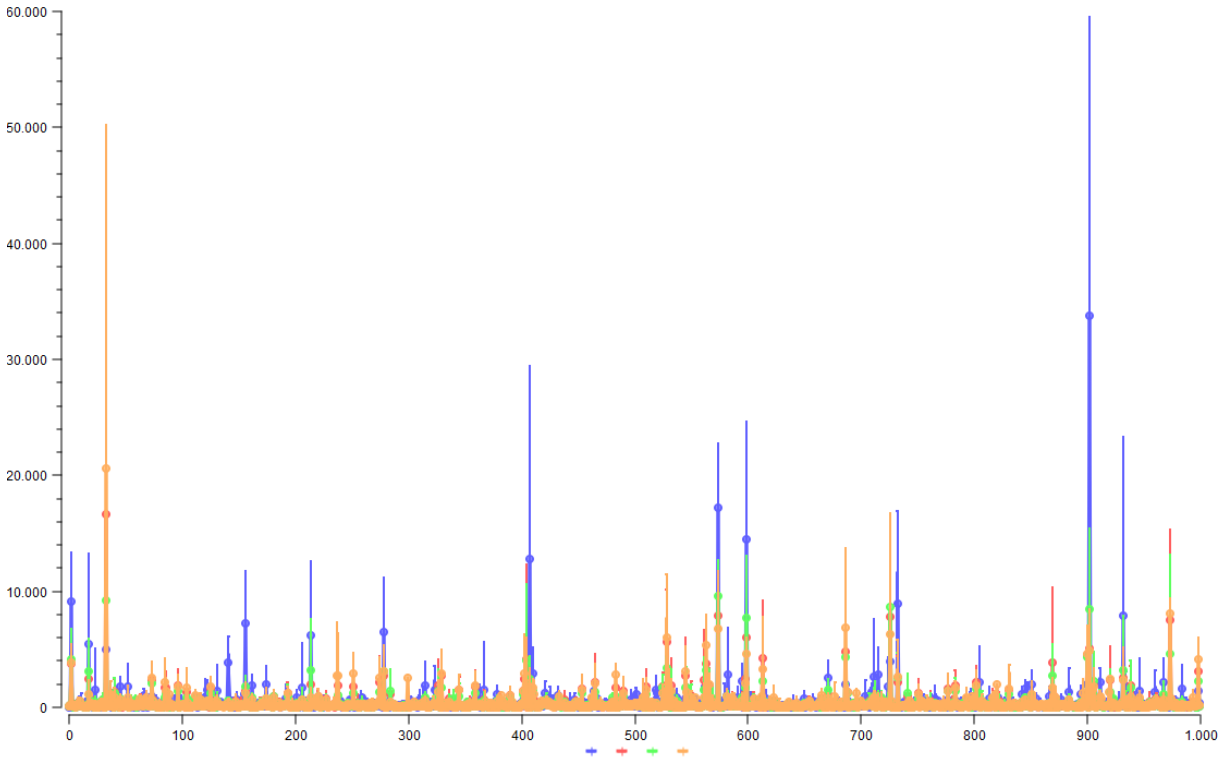

Figure 3.11.Metastatic prostate cancer data profile graph

Public Data-5: Asthma and Atopy

Summary: Investigation of CD4+ lymphocytes from patients with and without atopy, in combination with asthma. (34)

Organism: Homo sapiens

Platform: GPL97: [HG-U133B] Affymetrix Human Genome U133B ArrayCitation:

Reference Series: GSE473 **Sample count:** 29 **Value type:** count **Series published:** 2003/07/16

Figure 3.12. Asthma and Atopy data profile graph

Public Data-6: Quercetin effect on the colonic mucosa

Summary: Analysis of distal colonic mucosa scraping from animals fed a diet supplemented with quercetin. Quercetin is one of the major flavonoids found in fruits and vegetables and is a component of dietary supplements for humans. Results provide insight into the effect of quercetin on the distal colon. (13)

Organism: Rattus norvegicus

Platform: GPL1355: [Rat230_2] Affymetrix Rat Genome 230 2.0 Array

Dihal AA, van der Woude H, Hendriksen PJ, Charif H et al. Transcriptome and proteome profiling of colon mucosa from quercetin fed F344 rats point to tumor preventive mechanisms, increased mitochondrial fatty acid degradation and decreased glycolysis. Proteomics 2008 Jan; 8(1): 45-61. **Reference Series:** GSE7479 **Sample count:** 8 **Value type:** count **Series published:** 2008/01/01

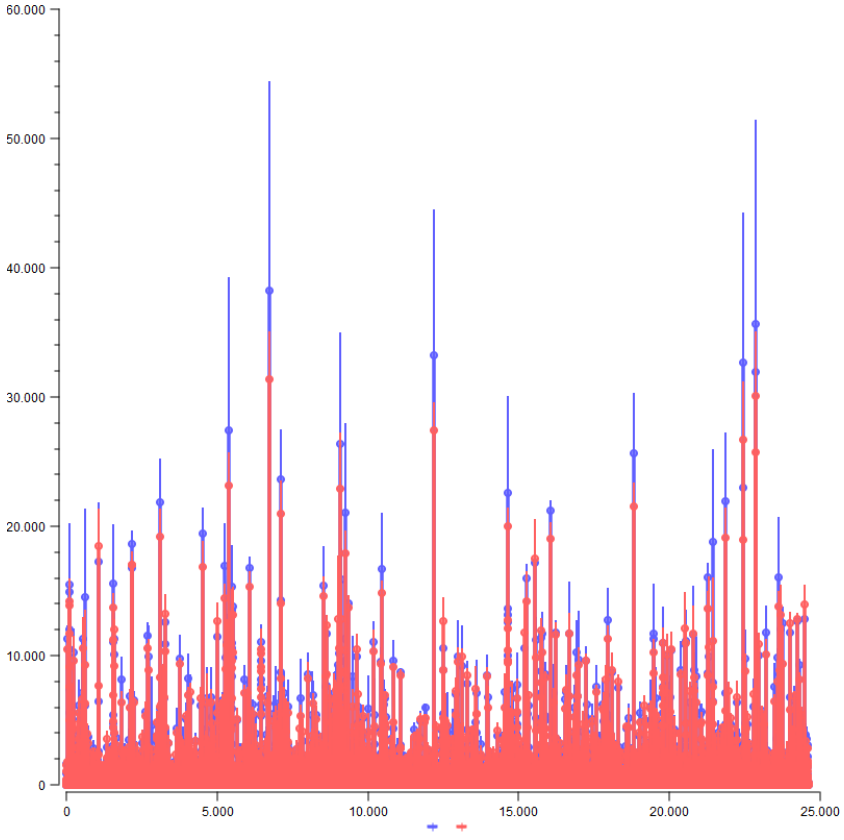

Figure 3.13.Quercetin effect on the colonic mucosa data profile graph

Public Data-7: Tumor necrosis factor effect on macrovascular umbilical vein endothelial

Summary: Analysis of macrovascular umbilical vein endothelial cells (HUVEC) stimulated with tumor necrosis factor (TNF)-alpha for 5 hours. TNF is a potent inflammatory stimulus. Results provide insight into the relevance of the diversity of endothelial cell subtypes for the response to inflammatory stimuli. (51)

Organism: Homo sapiens

Platform: GPL96: [HG-U133A] Affymetrix Human Genome U133A Array

Reference Series: GSE2639 **Sample count:** 8 **Value type:** count **Series published:** 2005/05/12

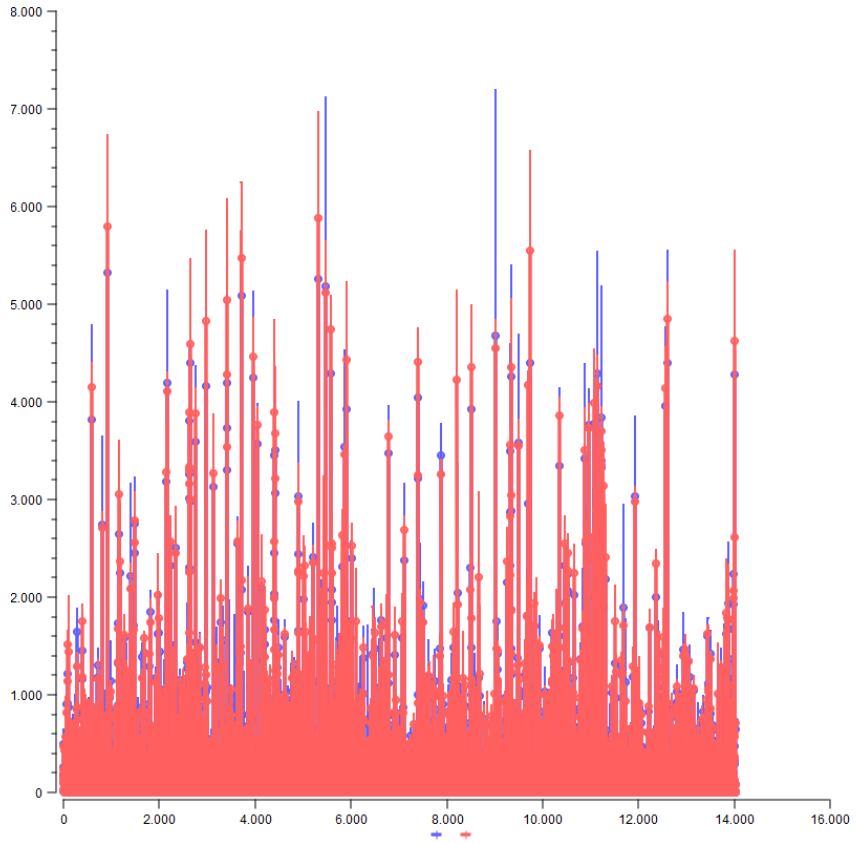

Figure 3.14. Tumor necrosis factor effect on macrovascular umbilical vein endothelial data profile graph

Public Data -8: Diabetic nephropathy

Summary: Comparison of glomeruli from kidneys with diabetic nephropathy (DN) and glomeruli from kidneys of healthy individuals. Progression of DN may be due to diminished tissue repair capability. (2)

Organism: Homo sapiens

Platform: GPL8300: [HG_U95Av2] Affymetrix Human Genome U95 Version 2

Reference Series: GSE1009 **Sample count:** 6 **Value type:** count **Series published:** 2004/02/02

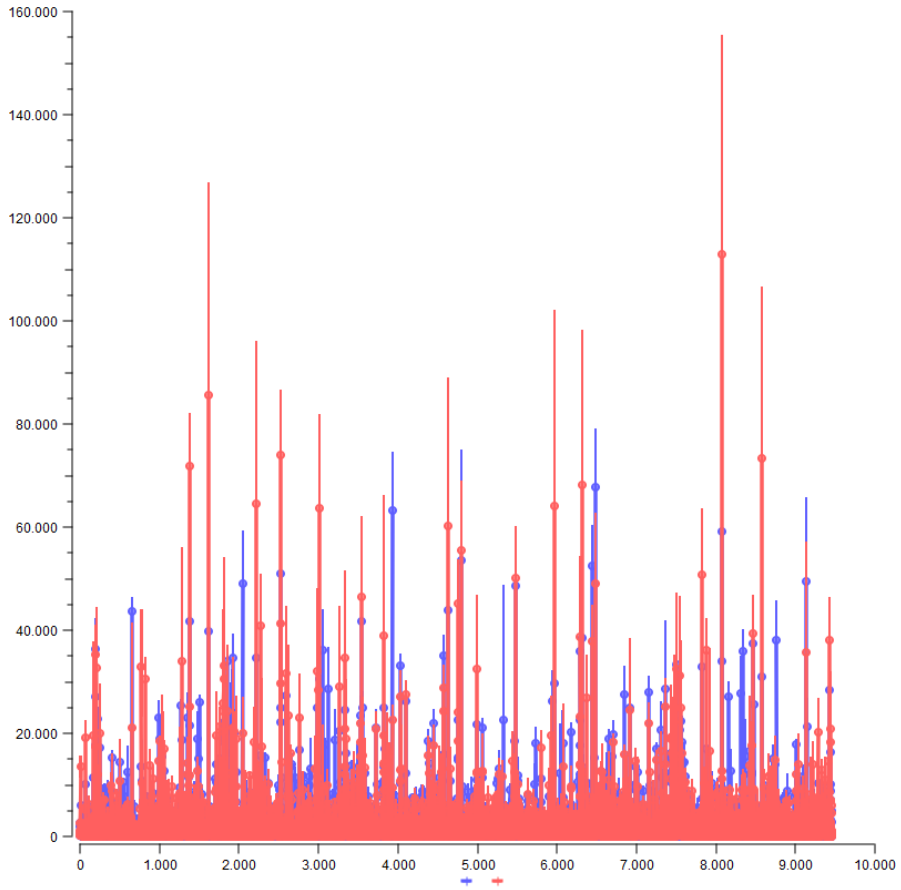

Figure 3.15.Diabetic nephropathy data profile graph

Public Data -9: Liver response to a high cholesterol diet and phenobarbital

Summary: Analysis of livers from animals treated with phenobarbital or fed a high cholesterol diet for 7 days. Results provide insight into the molecular mechanisms underlying the effects of these exogenous sources on cholesterol homeostasis and drug metabolism. (42)

Organism: Mus musculus

Platform: GPL339: [MOE430A] Affymetrix Mouse Expression 430A Array

Reference Series: GSE6721 **Sample count:** 12 **Value type:** transformed count

Series published: 2007/11/27

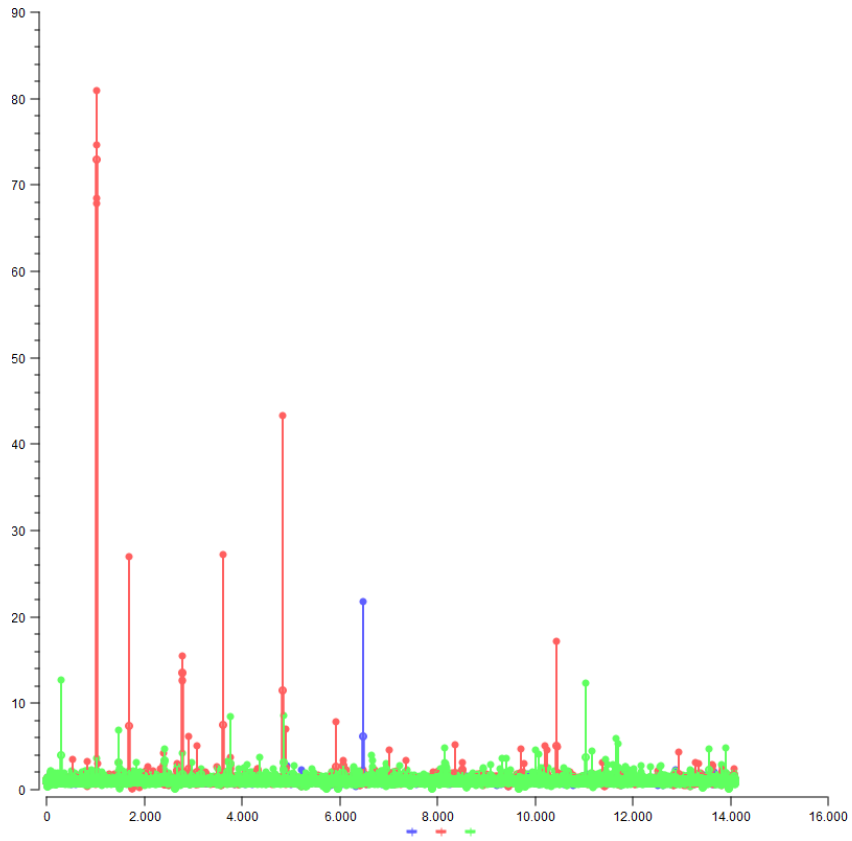

Figure 3.16.Liver response to a high cholesterol diet and phenobarbital data profile graph

Public Data-10: Hypothalamoneurohypophyseal system response to dehydration

Summary: Analysis of the hypothalamoneurohypophyseal system (HNS) in fluid-deprived males. The HNS components examined were hypothalamic paraventricular and supraoptic nuclei and neurointermediate lobe of the pituitary. Results identify candidate genes of HNS activity and remodeling induced by dehydration**.** (17)

Organism: Rattus norvegicus

Platform: GPL1355: [Rat230_2] Affymetrix Rat Genome 230 2.0 Array

Reference Series: GSE4130 **Sample count:** 30 **Value type:** count **Series published:** 2006/01/31

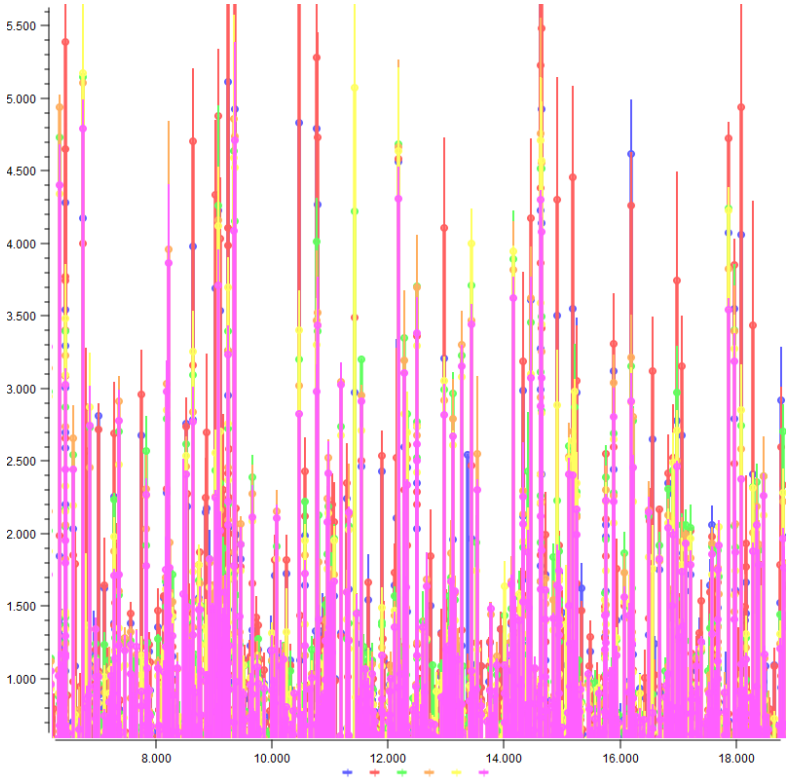

Figure 3.17.Hypothalamoneurohypophyseal system response to dehydration data profile graph

Public Data-11: Glioma cell migration: comparison of fast and slow invading cells

Summary: Analysis of glioma migration by comparing expression profiles of fast and slow invading C6 glial cells. Fast and slow cells separated using monolayer migration assay and by extent of migration of transplants in nude mice brains. Results provide insight into mechanisms underlying glioma invasion. (45)

Organism: Rattus norvegicus

Reference Series: GSE1139 **Sample count:** 7 **Value type:** count **Series published:** 2004/09/01

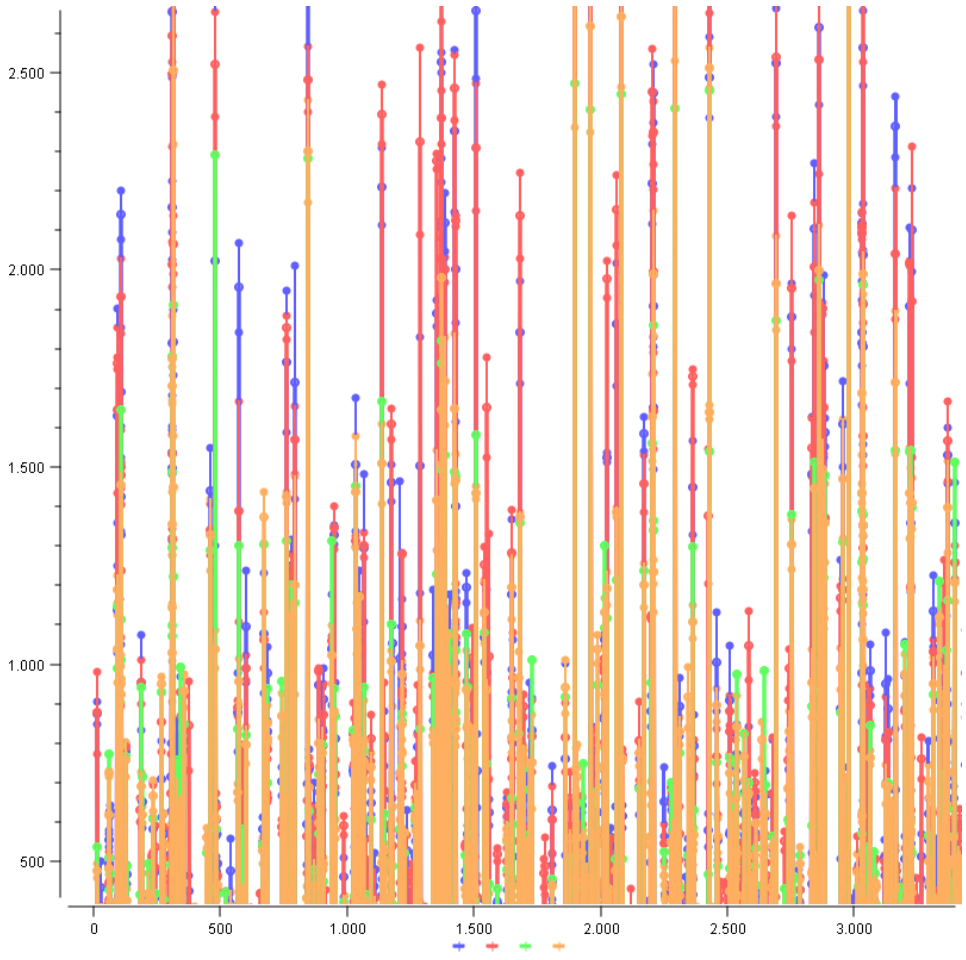

Figure 3.18.Glioma cell migration: comparison of fast and slow invading cells data profile graph

Public Data-12: Dysferlin deficiency effect on skeletal and cardiac muscles

Summary: Comparison of skeletal and cardiac muscles of dysferlin deficient SJL/J animals. Dysferlin deficiency results in skeletal muscle weakness, but does not affect the heart. Dysferlin mutations can result in the neuromuscular disorders Limb-Girdle muscular dystrophy type 2B and Myoshi myopathy. (54)

Organism: Mus musculus

Reference Series:GSE2507 **Sample count:** 20 **Value type:** count **Series published:** 2005/07/10

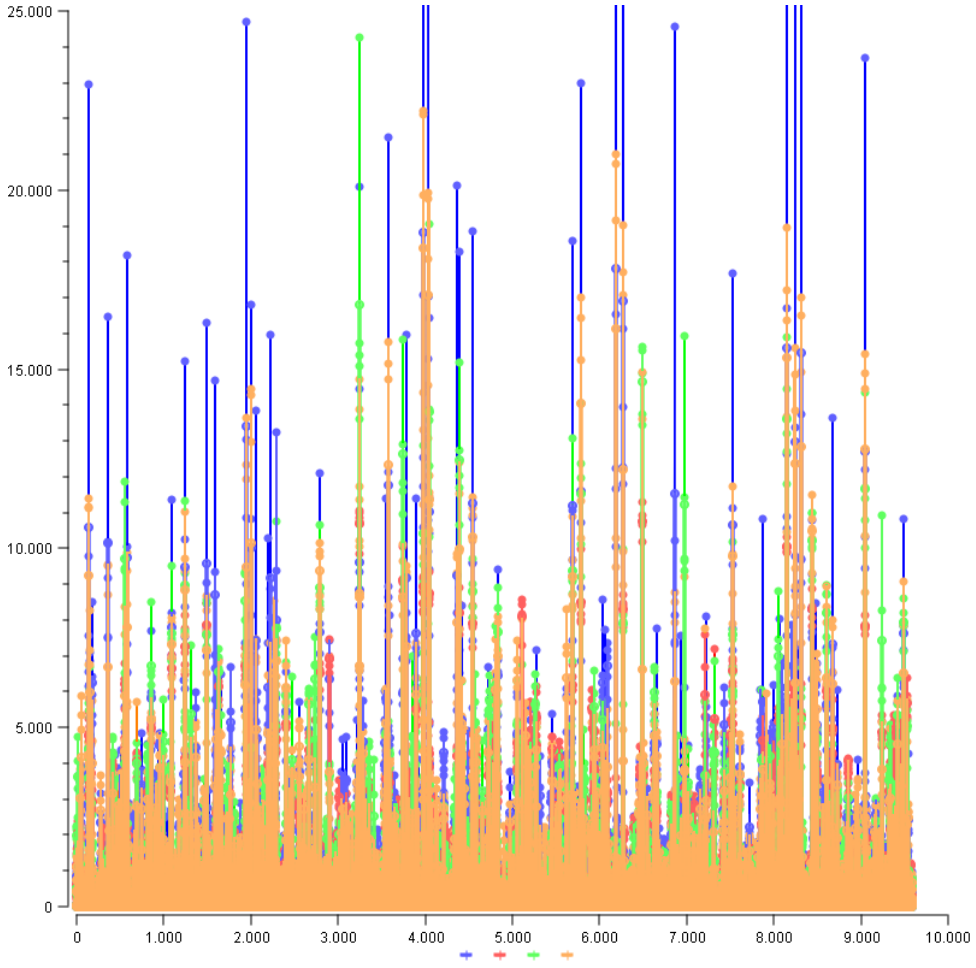

Figure 3.19.Dysferlin deficiency effect on skeletal and cardiac muscles data profile graph

Public Data-13: Treacher Collins' syndrome Tcof1 gene overexpression and knockdown effect on neuroblastoma cells

Summary: Expression profiling of neuroblastoma N1E-115 cells with the Treacher Collins' syndrome (TCS) Tcof1 gene overexpressed or inactivated. TCS is an autosomal dominant dysostosis of the face and lower jaw. Results provide insight into the role of Tcof1 in neuroblastoma cell proliferation. (35)

Organism: Mus musculus

Reference Series: GSE1956 **Sample count:** 9 **Value type:** count **Series published:** 2004/11/11

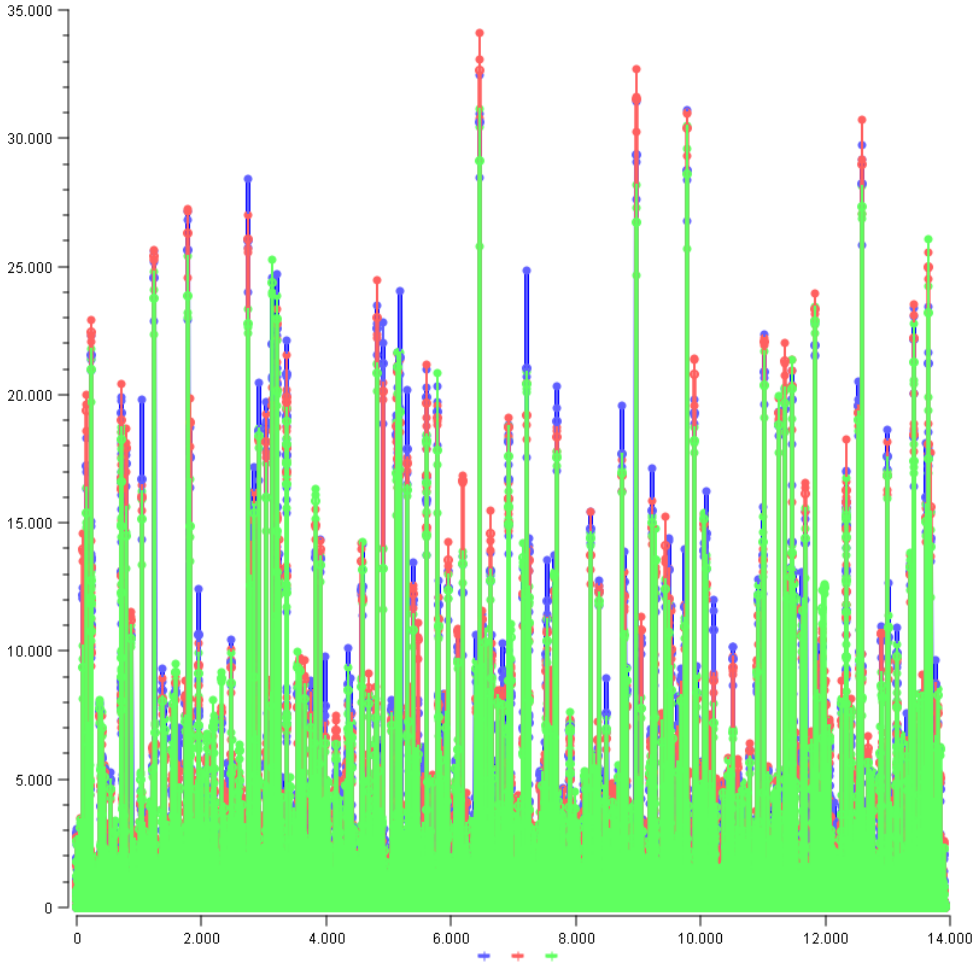

Figure 3.20.Treacher Collins' syndrome Tcof1 gene overexpression and knockdown effect on neuroblastoma cells data profile graph

Public Data-14: Visual cortex during the critical period for ocular dominance (MG-U74A)

Summary: Analysis of visual cortices before the critical period for ocular dominance plasticity opens [postnatal day 14 (P14)], at the peak sensitivity of the critical period (P28), and after the critical period (P60). Results provide insight into the molecular mechanisms associated with the critical period. (31)

Organism: Mus musculus

Reference Series: GSE11764 **Sample count:** 10 **Value type:** count **Series published:** 2008/06/12

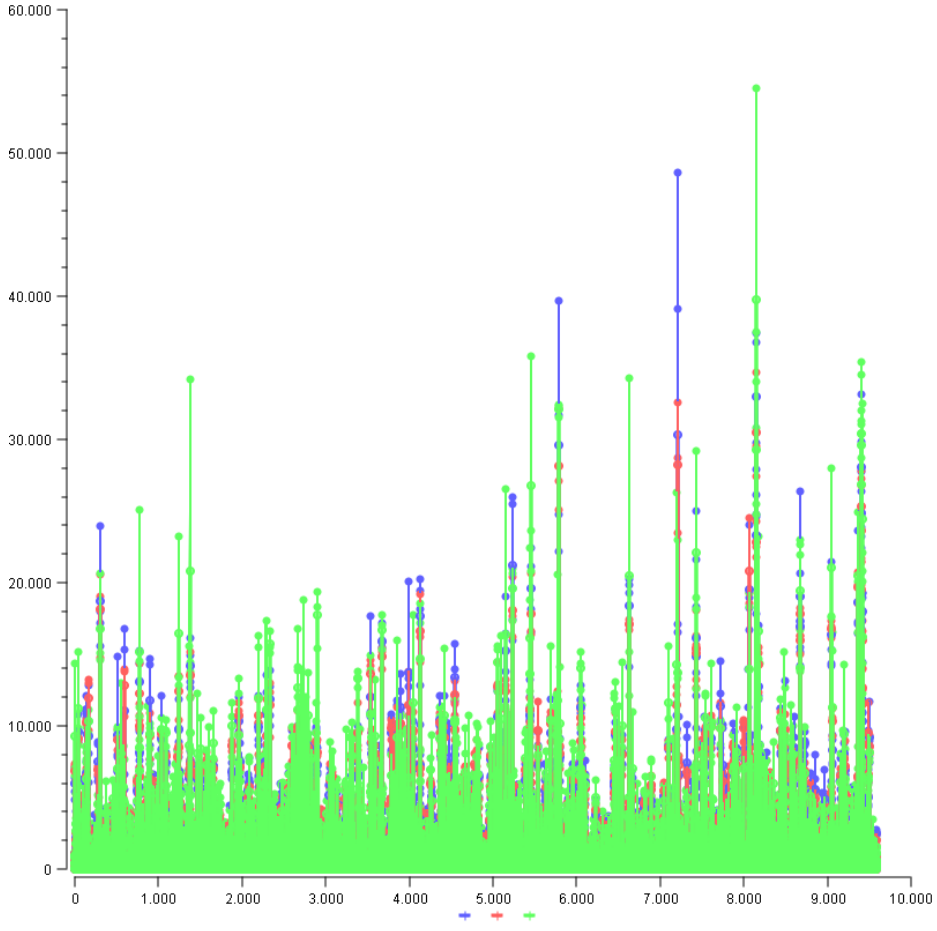

Figure 3.21.Visual cortex during the critical period for ocular dominance data profile graph

Public Data-15: Cigarette smoking effect on alveolar macrophage

Summary: Analysis of alveolar macrophages from 15 cigarette smokers, 15 non-smokers and 15 asthmatics. Results suggest that alveolar macrophage activation induced by smoking contributes to emphysema. (57)

Organism: Homo sapiens

Reference Series:GSE1859_**Sample count:** 45**Value type:** count **Series published:** 2005/10/04

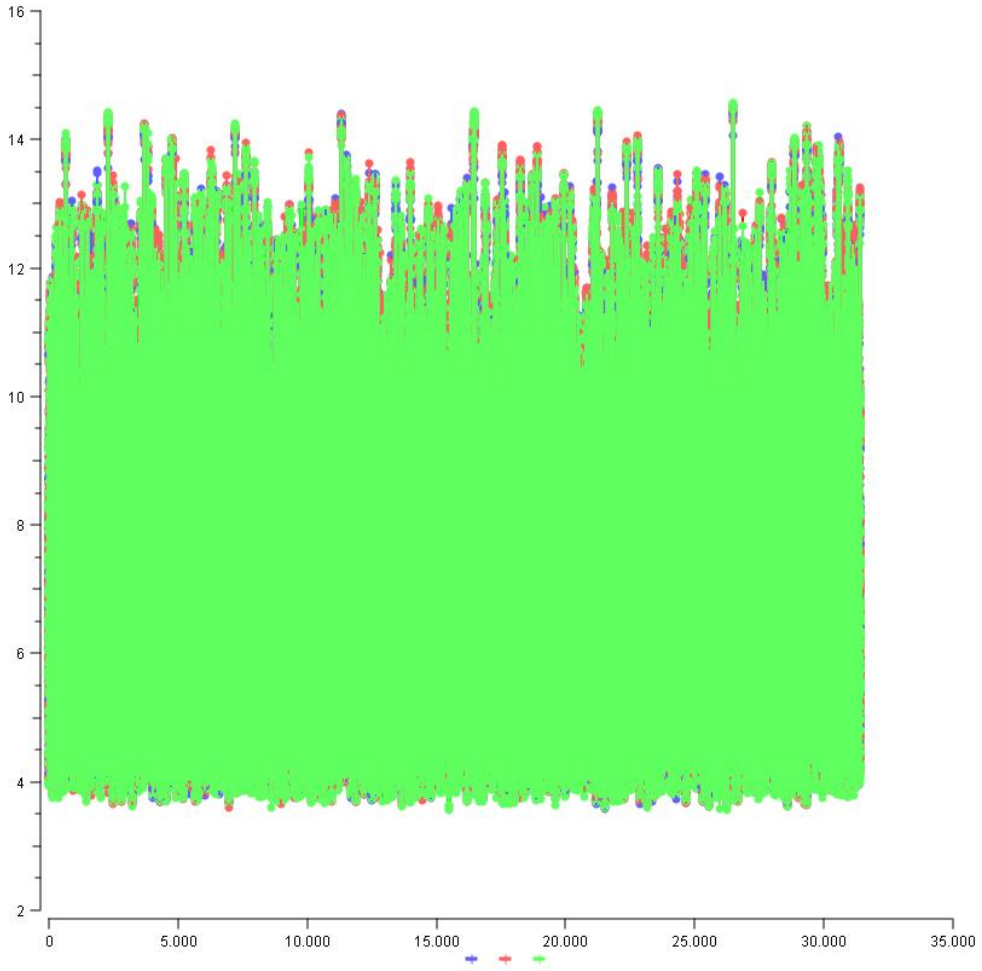

Figure 3.22. Cigarette smoking effect on alveolar macrophage data profile graph

3.8.2. Simulation Study

Application data have simulated with the model that described in Ram´on D´ıaz-Uriarte, et al. (12).

"Data have simulated under different number of classes of patients [2, 3, 4], number of independent dimensions [1 to 3], and number of genes per dimension [5, 20, 100], In all cases, the number of subjects per class has been set to 25 and 50 (a number which is similar to, or smaller than, that of many microarray studies), The data have been simulated from a multivariate normal distribution. All "genes" have a variance of 1, and the correlation between genes within a dimension is 0,9, whereas the correlation between genes among dimensions is 0. In other words, the variance-covariance matrix is a block-diagonal matrix": (12)

$$\sum = \begin{bmatrix} a & 0 & 0 & \cdots & 0 \\ 0 & a & 0 & \ldots & 0 \\ \vdots & . & . & . & \vdots \\ 0 & 0 & 0 & \ldots & a \end{bmatrix} \qquad (3.7)$$

$$a = \begin{bmatrix} 1 & 0.9 & \cdots & 0.9 \\ 0.9 & 1 & \ldots & 0.9 \\ . & . & . & . \\ . & . & . & . \end{bmatrix} \qquad (3.8)$$

"The class means have been set so that TCR, F-Measure, Recall, Precision of a proposed method using one gene from each dimension is approximately 5%; and each dimension has the same relevance in separation, Specifically, the class means used are:" (12)

• **One dimension:**

– Two classes: μ1 = −1, 65, μ2 = 1, 65,

– Three classes: μ1 = −3, 58, μ2 = 0, μ3 = 3, 58,

– Four classes: μ1 = −3, 7, μ2 = 0, μ3 = 3, 7, μ4 = 7, 4,

• **Two dimensions:**

– Two classes: μ1 = [−1, 18, −1, 18], μ2 = [1, 18, 1, 18],

– Three classes: μ1 = [0, 0], μ2 = [3, 88 cos (15), 3, 88 sin (15)], μ3 = [3, 88 cos (75), 3, 88 sin (75)],

– Four classes: μ1 = [1, 1], μ2 = [4, 95, 1], μ3 = [1, 4, 95], μ4 = [4, 95, 4, 95],

• **Three dimensions:**

– Two classes: μ1 = [−0, 98, −0, 98, −0, 98], μ2 = [0, 98, 0, 98, 0, 98],

– Three classes: μ1 = [2, 76, 0, 0], μ2 = [0, 2, 76, 0], μ3 = [0, 0, 2, 76],

– Four classes: μ1 = [2, 96, 0, 0], μ2 = [0, 2, 96, 0], μ3 = [0, 0, 2, 96], μ4 = [2, 96, 2, 96, 2, 96]

“After the genes that belong to the dimensions are generated, another 2000 N(0, 1) variable and 2000 U[−1, 1] variable are added to the matrix of “genes”.

4. RESULTS

4.1. Results of GEO Data Sets

15 public data sets were analyzed by two different ways. These data sets include different sample size and class numbers. Looking at Table 4.1, the highest increase was observed in TCR as 15.4%.

With the ICA+ RF based analysis, Table 4.1:

- The lowest true classification rate was found in the 7th data set as 52.1%, the highest true classification rate was found in the 15th data set as 75.8%.
- The lowest F-Score value was found in the 9th data set as 49.7%, the highest F-Score was found in the 3rd data set as 85.7%.
- The lowest Recall value was found in the 9th data set as 41.6%, the highest Recall value was found in the 5th data set as 90.8%.
- The lowest Precision value was found in the 2nd data set as 56.6%, the highest Precision value was found in the 3rd data set as 83.6%.

With the PM, Table 4.2.:

- The lowest true classification rate was found in the 2nd data set as 58.1%, the highest true classification rate was found in the 6thdata set as 92.7%.
- The lowest F-Score value was found in the 2nd data set as 53.4%, the highest F-Score was found in the 1th data set as 92.6%.
- The lowest Recall value was found in the 2nddata set as 61.1%, the highest Recall value was found in the 1stdata set as 97.6%.
- The lowest Precision value was found in the 2nddata set as 57.7%, the highest Precision value was found in the 1stdata set as 88.1%.

Average gain for all performance criteria is displayed in Table 4.3. According to this table, PM has increased all performance criteria. TCR has maximum gain as 12.19, Precision has minimum gain as 5.89.

Table 4.1. Results of ICA+RF model on GEO data sets

ICA + RF	TCR	F-Score	Recall	Precision
Public Data-1: Skeletal muscle response to insulin infusion	0.609	0.604	0.603	0.605
Public Data-2: Lymph node and tonsil comparison	0.566	0.583	0.613	0.556
Public Data-3: Atrial and ventricular myocardium comparison	0.701	0.857	0.879	0.836
Public Data-4: Metastatic prostate cancer	0.700	0.499	0.418	0.619
Public Data-5: Asthma and Atopy	0.538	0.855	0.908	0.807
Public Data-6: Quercetin effect on the colonic mucosa	0.612	0.619	0.614	0.625
Public Data-7: Tumor necrosis factor effect on macrovascular umbilical vein endothelial	0.521	0.572	0.587	0.557
Public Data-8: Diabetic nephropathy	0.745	0.848	0.874	0.824
Public Data-9: Liver response to a high cholesterol diet and Phenobarbital	0.714	0.497	0.416	0.617
Public Data-10: Hypothalamoneurohypophyseal system response to dehydration	0.578	0.850	0.901	0.804
Public Data-11: Glioma cell migration: comparison of fast and slow invading cells	0.678	0.728	0.745	0.712
Public Data-12: Dysferlin deficiency effect on skeletal and cardiac muscles	0.642	0.667	0.678	0.657
Public Data-13: Treacher Collins' syndrome Tcof1 gene overexpression and knockdown effect on neuroblastoma cells	0.715	0.698	0.657	0.745
Public Data-14: Visual cortex during the critical period for ocular dominance	0.712	0.701	0.699	0.704
Public Data-15: Cigarette smoking effect on alveolar macrophage	0.758	0.687	0.674	0.700

Table 4.2.Results of proposed method on GEO data sets

ICA + KM+ RF	TCR	F-Score	Recall	Precision
Public Data-1: Skeletal muscle response to insulin infusion	0.920	0.926	0.976	0.881
Public Data-2: Lymph node and tonsil comparison	0.581	0.594	0.614	0.577
Public Data-3: Atrial and ventricular myocardium comparison	0.855	0.890	0.944	0.842
Public Data-4: Metastatic prostate cancer	0.847	0.735	0.659	0.832
Public Data-5: Asthma and Atopy	0.664	0.761	0.827	0.704
Public Data-6: Quercetin effect on the colonic mucosa	0.927	0.844	0.879	0.812
Public Data-7: Tumor necrosis factor effect on macrovascular umbilical vein endothelial	0.645	0.623	0.622	0.625
Public Data-8: Diabetic nephropathy	0.874	0.896	0.942	0.854
Public Data-9: Liver response to a high cholesterol diet and Phenobarbital	0.845	0.731	0.657	0.823
Public Data-10: Hypothalamoneurohypophyseal system response to dehydration	0.661	0.746	0.794	0.704
Public Data-11: Glioma cell migration: comparison of fast and slow invading cells	0.702	0.738	0.754	0.722
Public Data-12: Dysferlin deficiency effect on skeletal and cardiac muscles	0.726	0.643	0.624	0.664
Public Data-13: Treacher Collins' syndrome Tcof1 gene overexpression and knockdown effect on neuroblastoma cells	0.812	0.707	0.668	0.750
Public Data-14: Visual cortex during the critical period for ocular dominance	0.799	0.730	0.745	0.716
Public Data-15: Cigarette smoking effect on alveolar macrophage	0.760	0.736	0.726	0.746

Table 4.3.Gain of proposed method (%)

Data Sets	#Group	TCR	F-Score	Recall	Precision
Public Data-1: Skeletal muscle response to insulin infusion	2	31.1	32.21	37.3	27.6
Public Data-2: Lymph node and tonsil comparison	2	1.5	1.04	0.01	2,1
Public Data-3: Atrial and ventricular myocardium comparison	2	15.4	3.31	6.5	0,6
Public Data-4: Metastatic prostate cancer	4	14.7	23.64	24.1	21.3
Public Data-5: Asthma and Atopy	5	12.6	-9.40	-8.1	-10.3
Public Data-6: Quercetin effect on the colonic mucosa	2	31.5	22.47	26.5	18.7
Public Data-7: Tumor necrosis factor effect on macrovascular umbilical vein endothelial	2	12.4	5.19	3.5	6.8
Public Data-8: Diabetic nephropathy	2	12.9	4.76	6.8	3.0
Public Data-9: Liver response to a high cholesterol diet and phenobarbital	3	13.10	23.37	24.1	20.6
Public Data-10: Hypothalamoneurohypophyseal system response to dehydration	5	8.30	-10.34	-10.7	-10.0
Public Data-11: Glioma cell migration: comparison of fast and slow invading cells	4	2.40	0.95	0.9	1.0
Public Data-12: Dysferlin deficiency effect on skeletal and cardiac muscles	4	8.40	-2.40	-5.4	0.7
Public Data-13: Treacher Collins' syndrome Tcof1 gene overexpression and knockdown effect on neuroblastoma cells	3	9.70	0.84	1.1	0,5
Public Data-14: Visual cortex during the critical period for ocular dominance	3	8.70	2.87	4.6	1.2
Public Data-15: Cigarette smoking effect on alveolar macrophage	3	0.2	4.91	5.2	4.6
Average Gain %		12.19	6.90	7.75	5.89

Table 4.4. Descriptive statistics of gain of PM for bootstrap samples

#Class		TCR	FScore	Recall	Precision
2	Mean	17.80	12.21	14.35	10.00
	Standard Deviation	11.60	16.17	17.59	14.67
	Median	15.05	12.89	15.30	10.40
	Minimum	1.50	-9.40	-8.10	-10.30
	Maximum	31.50	32.21	37.30	27.60
3	Mean	11.68	5.75	5.93	5.10
	Standard Deviation	2.27	13.79	14.30	12.59
	Median	12.65	4.98	5.15	4.90
	Minimum	8.30	-10.34	-10.70	-10.00
	Maximum	13.10	23.37	24.10	20.60
4	Mean	6.83	-0.20	-1.13	0.73
	Standard Deviation	3.89	1.90	3.70	0.25
	Median	8.40	0.84	0.90	0.70
	Minimum	2.40	-2.40	-5.40	0.50
	Maximum	9.70	0.95	1.10	1.00
5	Mean	4.45	3.89	4.90	2.90
	Standard Deviation	6.01	1.44	0.42	2.40
	Median	4.45	3.89	4.90	2.90
	Minimum	0.20	2.87	4.60	1.20
	Maximum	8.70	4.91	5.20	4.60

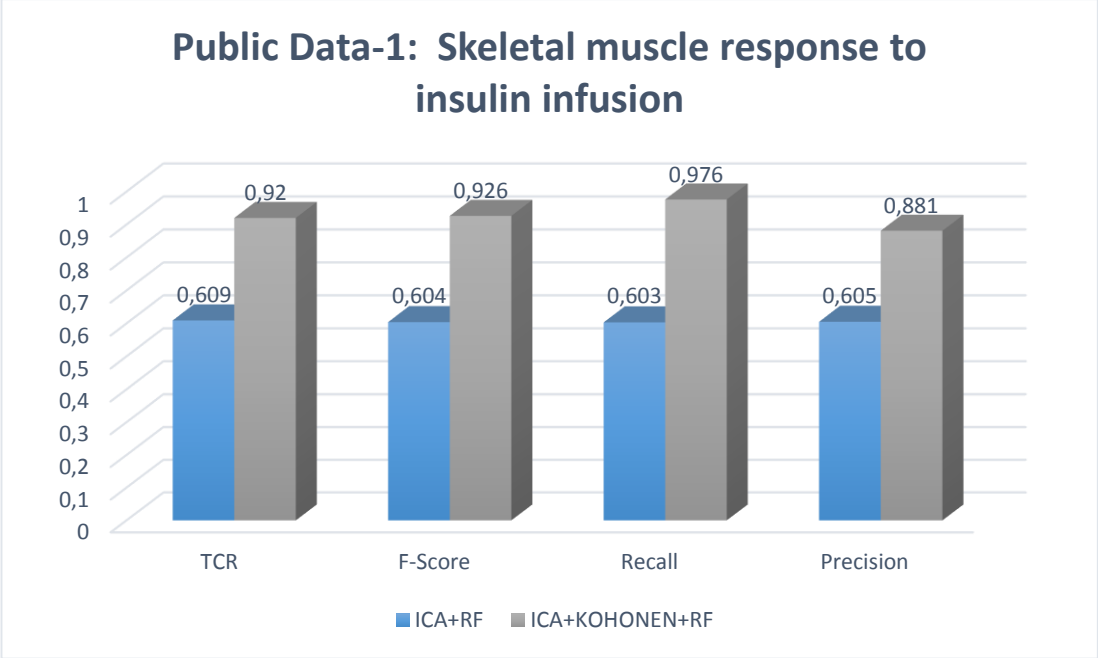

Figure 4.1. Results of ICA+RF and ICA+KM+RF methods on Public Data-1

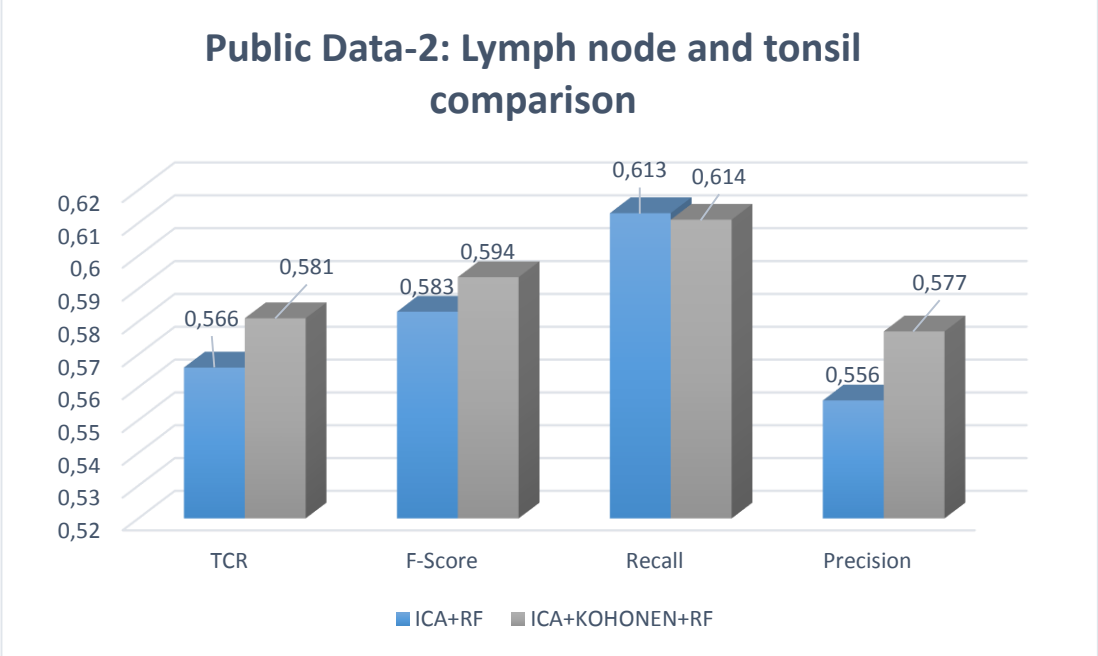

Figure 4.2. Results of ICA+RF and ICA+KM+RF methods on Public Data-2

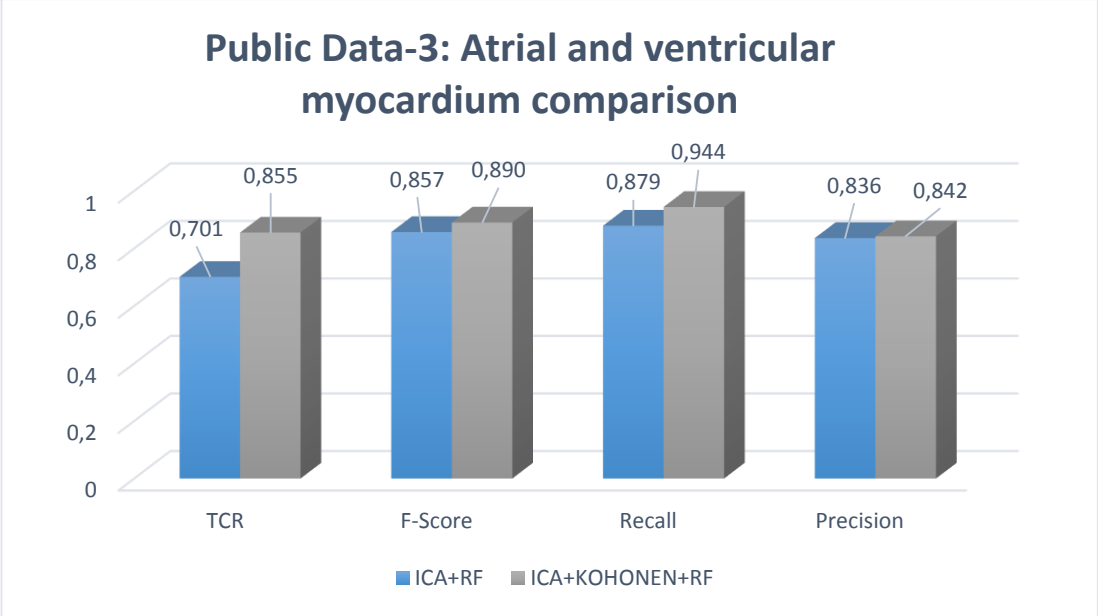

Figure 4.3. Results of ICA+RF and ICA+KM+RF methods on Public Data-3

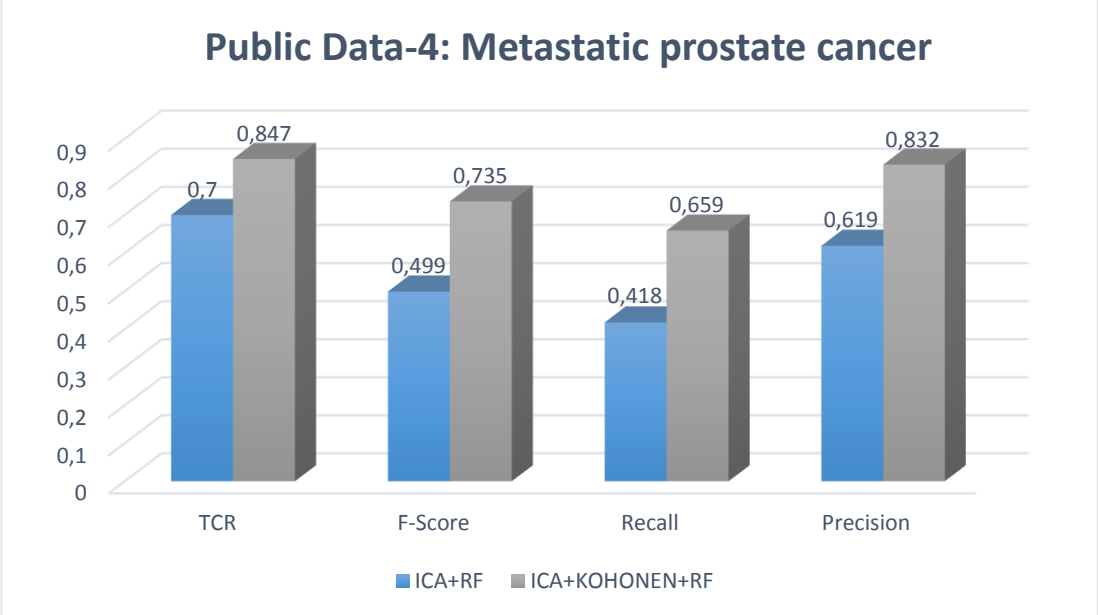

Figure 4.4. Results of ICA+RF and ICA+KM+RF methods on Public Data-4

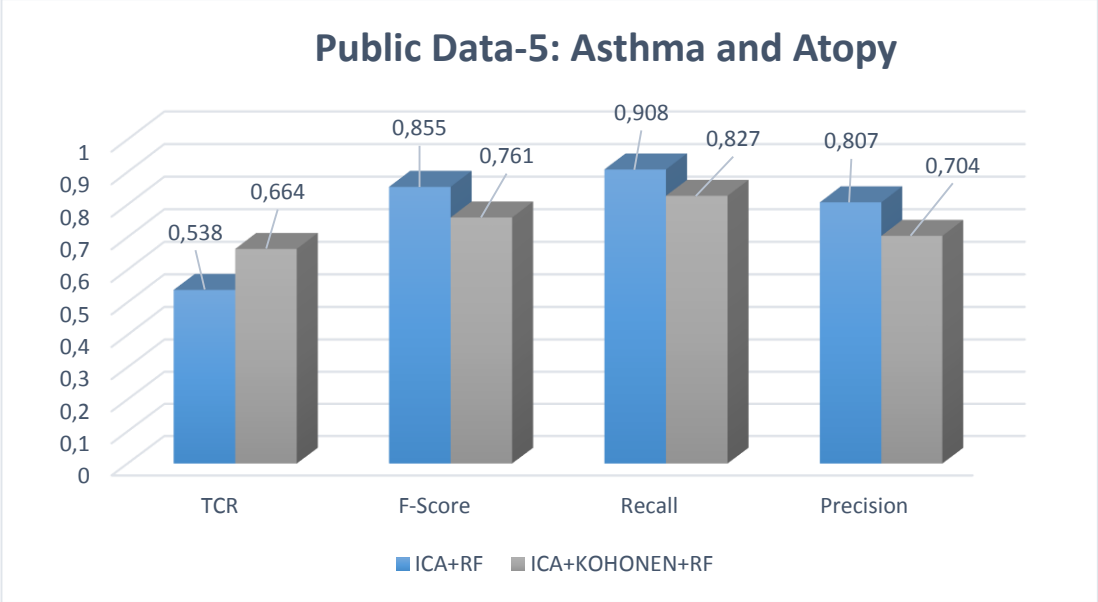

Figure 4.5. Results of ICA+RF and ICA+KM+RF methods on Public Data-5

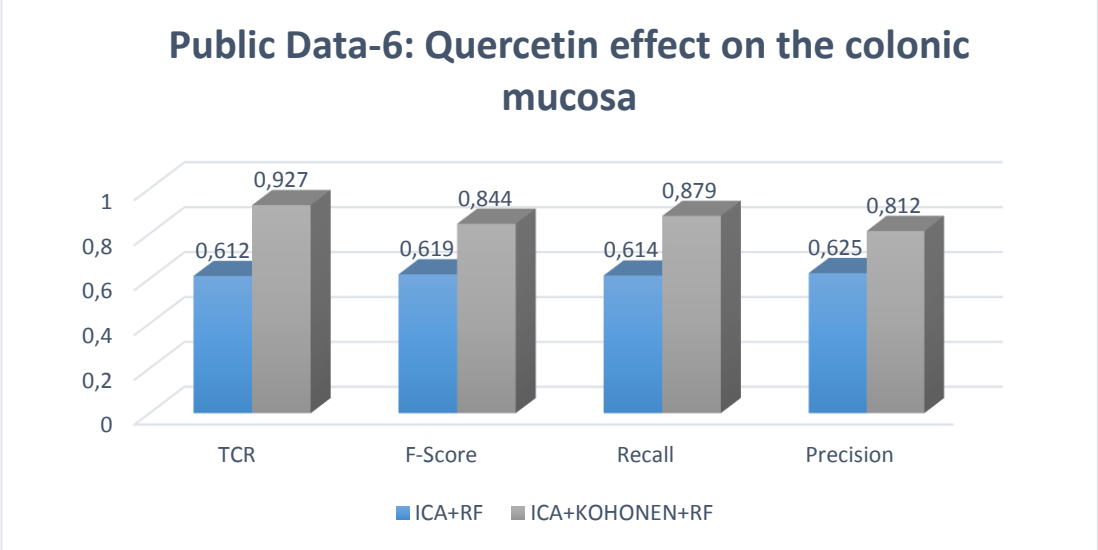

Figure 4.6. Results of ICA+RF and ICA+KM+RF methods on Public Data-6

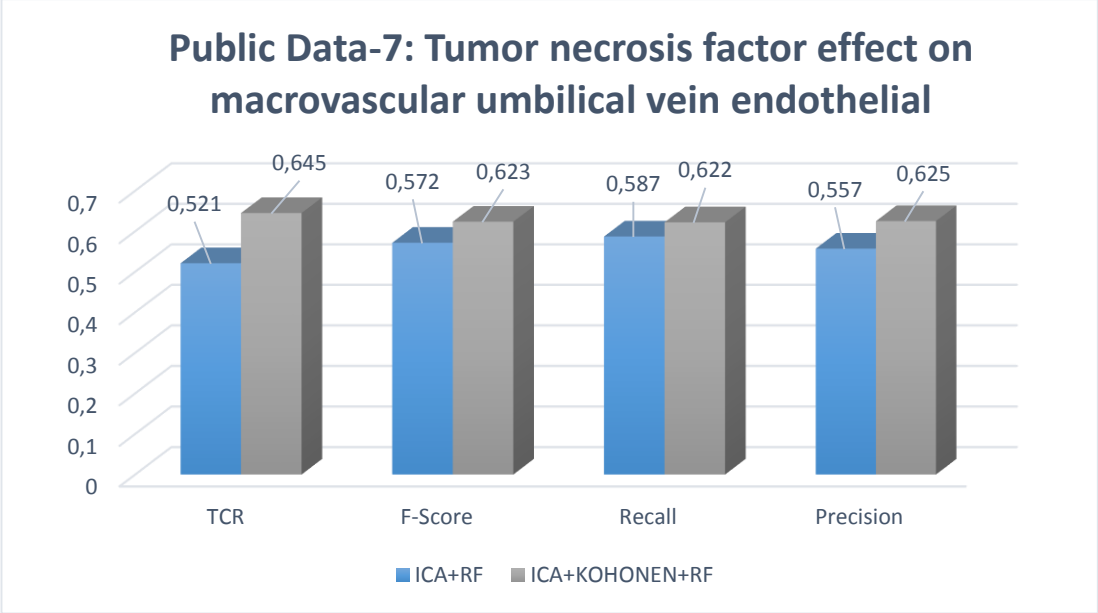

Figure 4.7. Results of ICA+RF and ICA+KM+RF methods on Public Data-7

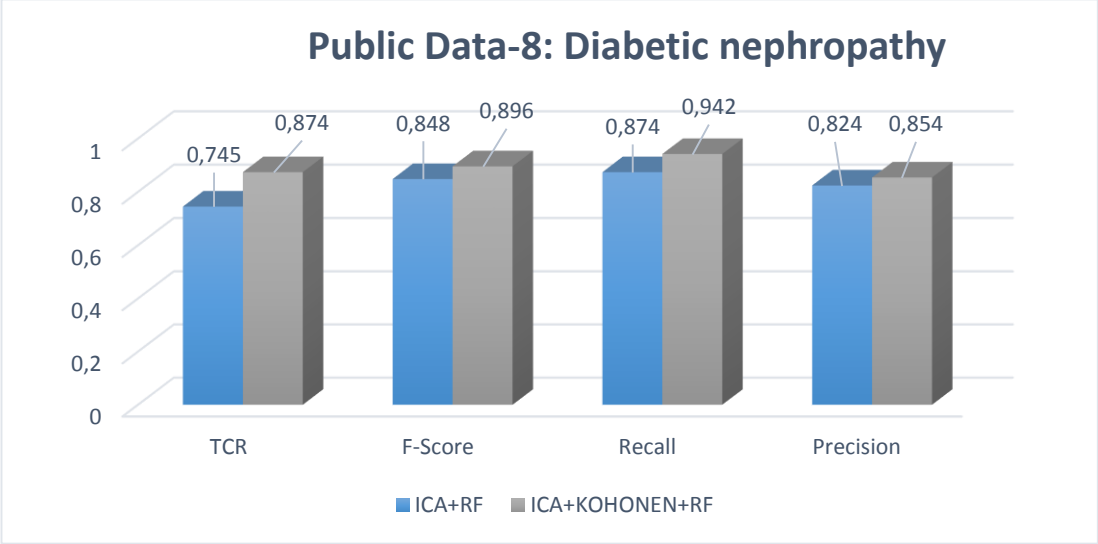

Figure 4.8. Results of ICA+RF and ICA+KM+RF methods on Public Data-8

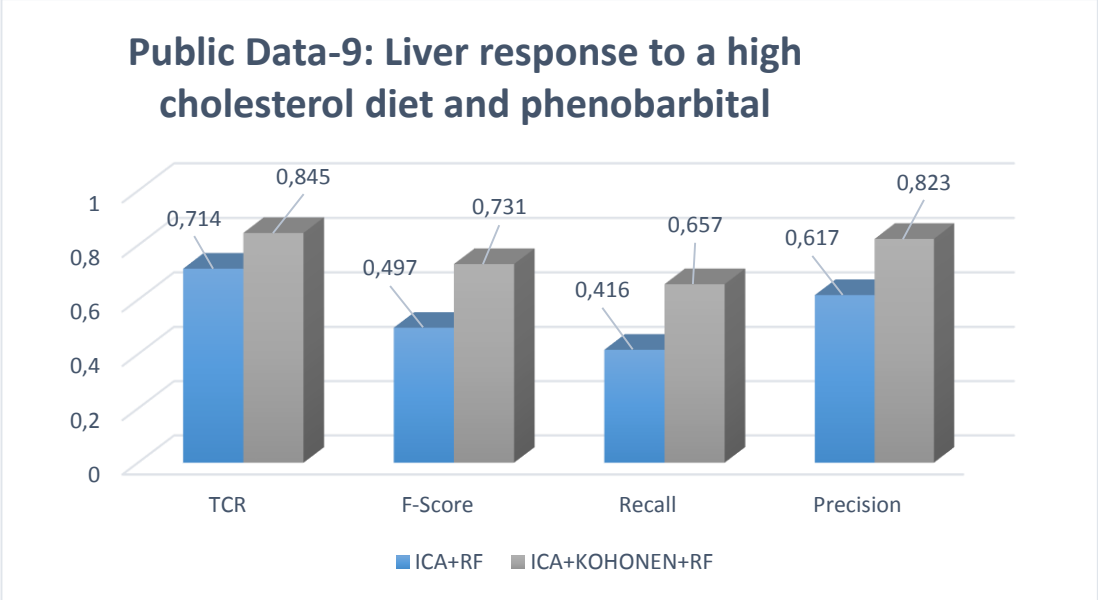

Figure 4.9. Results of ICA+RF and ICA+KM+RF methods on Public Data-9

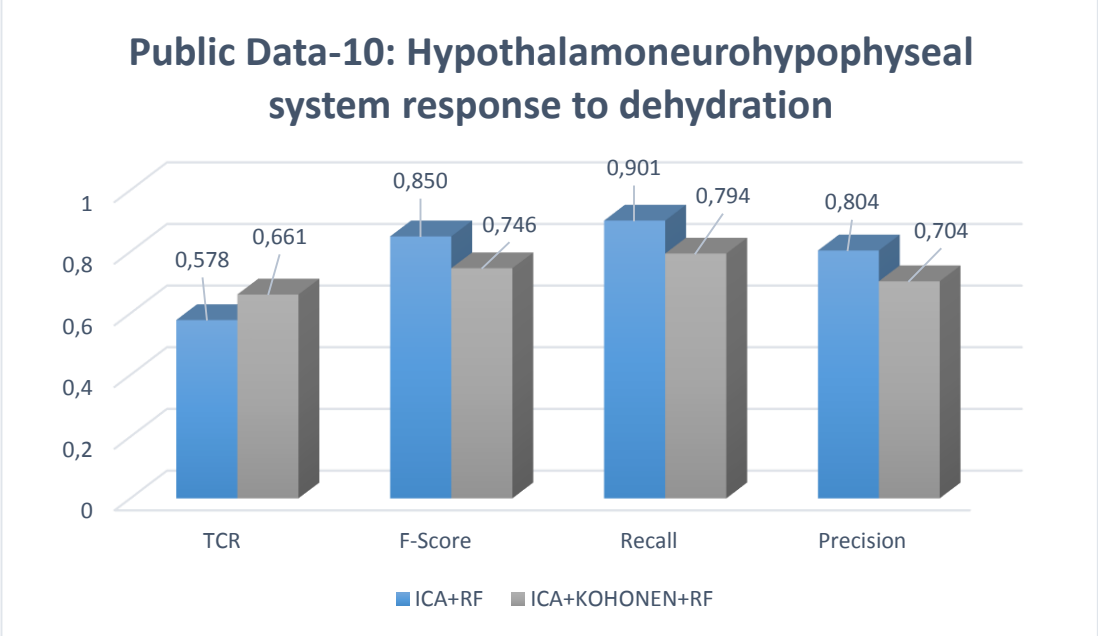

Figure 4.10. Results of ICA+RF and ICA+KM+RF methods on Public Data-10

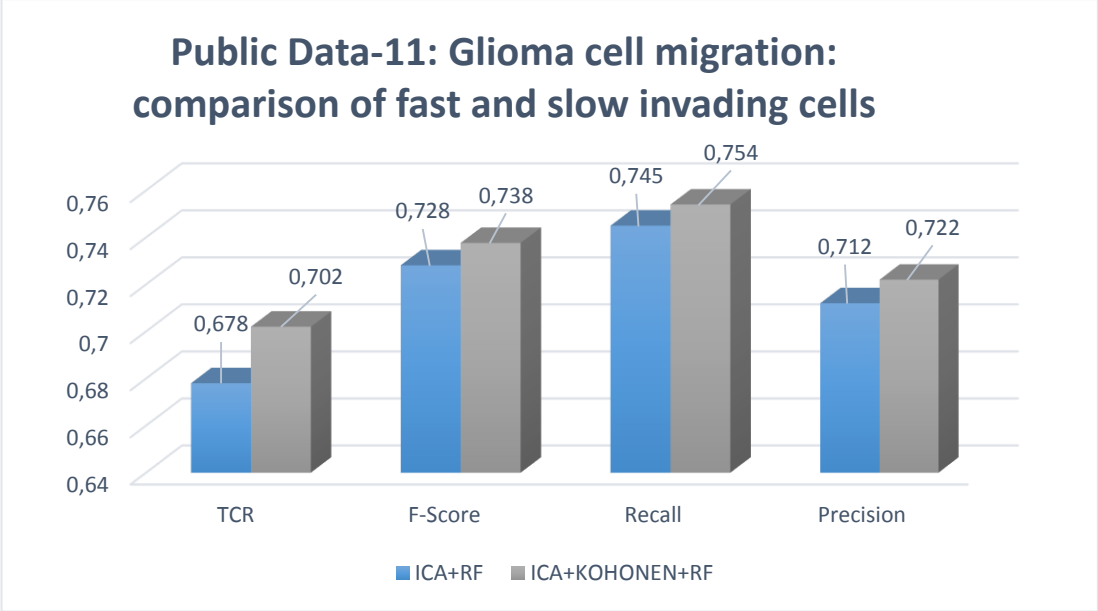

Figure 4.11. Results of ICA+RF and ICA+KM+RF methods on Public Data-11

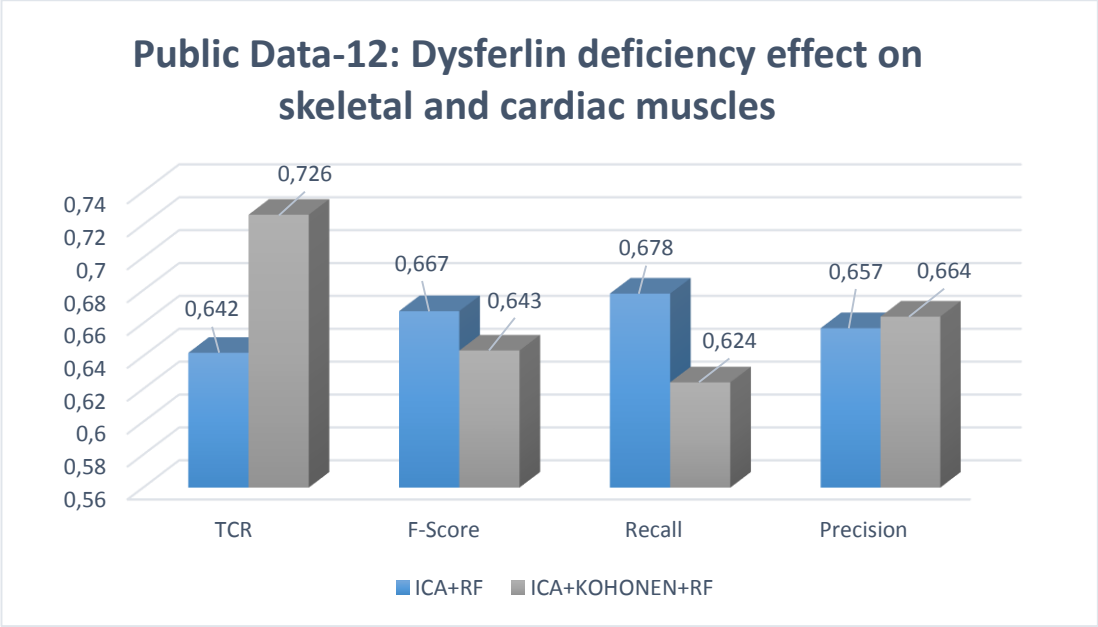

Figure 4.12. Results of ICA+RF and ICA+KM+RF methods on Public Data-12

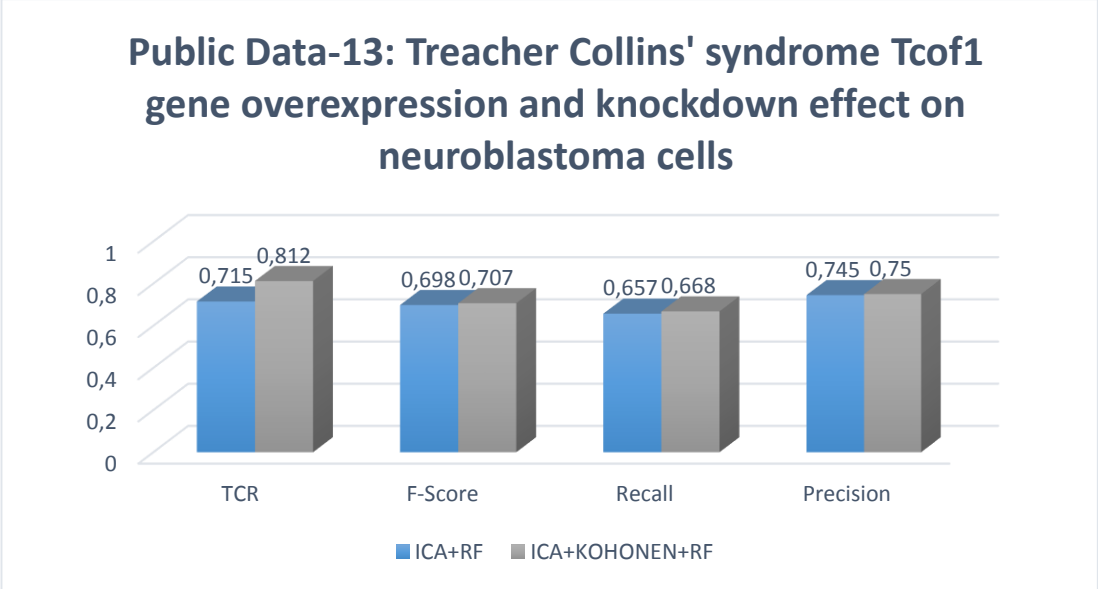

Figure 4.13. Results of ICA+RF and ICA+KM+RF methods on Public Data-13

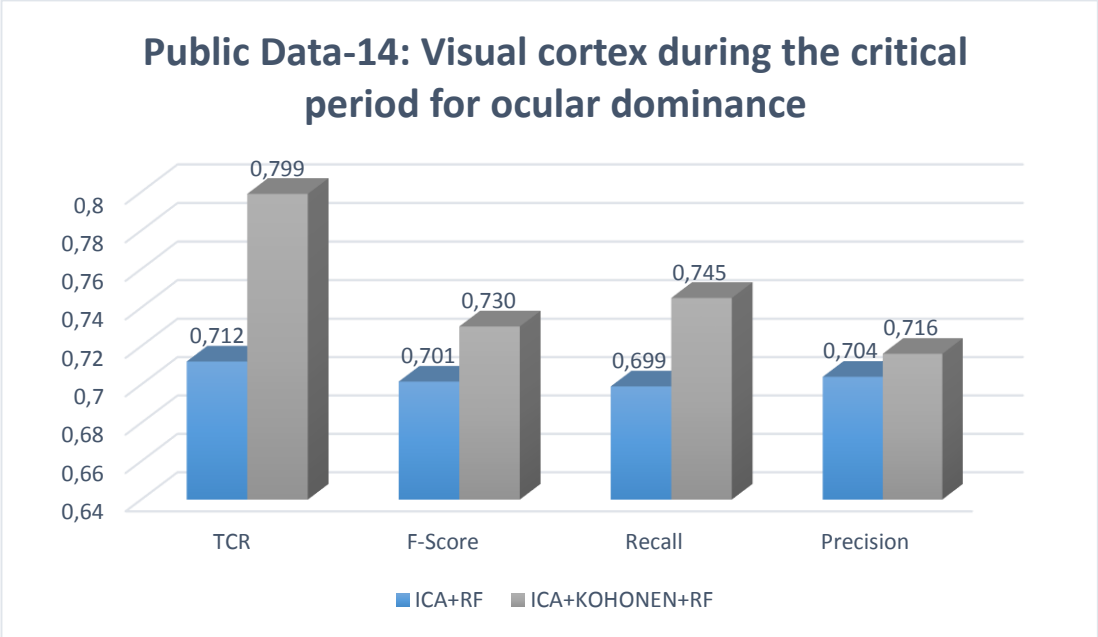

Figure 4.14. Results of ICA+RF and ICA+KM+RF methods on Public Data-14

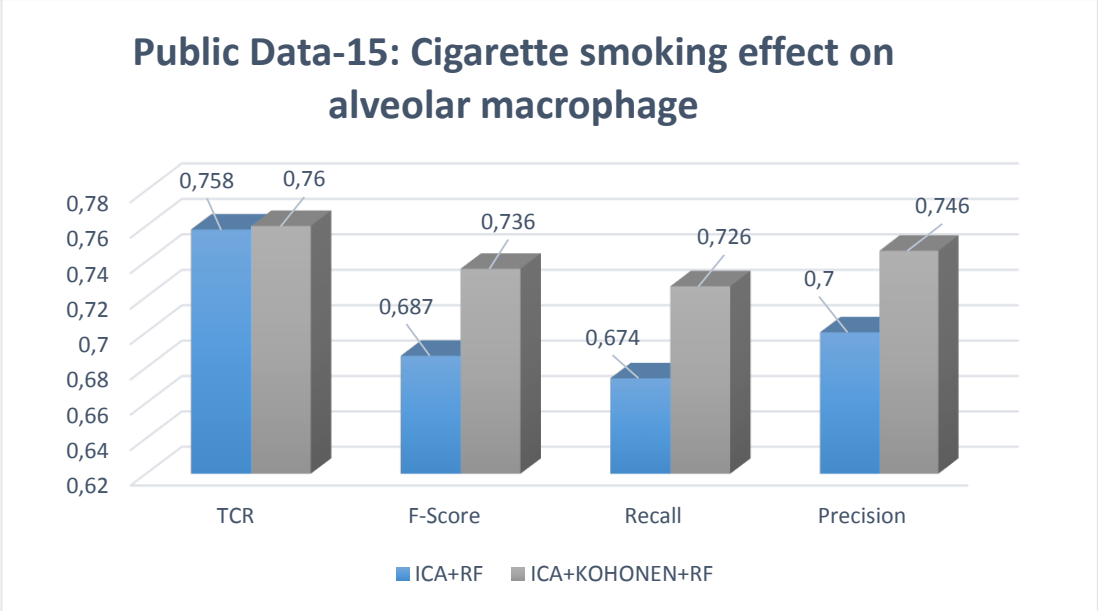

Figure 4.15. Results of ICA+RF and ICA+KM+RF methods on Public Data-15

4.2. Results of Simulation Study

As a result of the ICA+RF approach for simulated data sets which have 25 cases in each class, looking at Table 4.5, the highest increase was observed in the Precision value as 4.02%.

- The lowest TCR was found in # Dimension: 1 - # Class: 4 as 65.7%, the highest TCR was found in # Dimension: 1 - # Class: 2 scenario as 74.3%.
- The lowest F-Score value was found in # Dimension: 1 - # Class: 3 as 57.6%, the highest F-Score value was found in # Dimension: 1 - # Class: 2 scenario as 83.5%.
- The lowest Recall value in # Dimension: 1 - # Class: 3 as 53.2% the highest Recall value was found in Dimension: 1 - # Class: 2 scenario as 85.7%.
- The lowest Precision # Dimension: 1 - # Class: 3 as 62.8%, the highest Precision value was found in Dimension # 1 - # Class: 2 scenario as 81.4%.

As a result of the ICA+RF approach for simulated data sets which have 50 cases in each class, looking at Table 4.7, the highest increase was observed in TCR value as 8.24%.

- The lowest TCR was found in # Dimension: 2 - # Class: 2 as 61.5%, the highest TCR was found in Dimension: 3 - # Class: 3 scenario as 72.4%.
- The lowest F-Score value was found in # Dimension: 2 - # Class: 3 as 55.3%, the highest F-Score value was found in # Dimension: 1 - # Class: 2 scenario as 69.4%.
- The lowest Recall value was found in # Dimension: 2 - # Class: 3 as 49.9%, the highest Recall value was found in # Dimension: 1 - # Class: 2 scenario as 80.2%.
- The lowest Precision was found in # Dimension: 2 - # Class: 2 and # Dimension: 2 - # Class: 3 as 61.0%, the highest Precision value was found in Dimension # 1 - # Class: 3 scenario as 79.9%.

As a result of the PM for simulated data sets which have 25 cases in each class, looking at Table 4.6, the highest increase was observed in the Precision value as 4.02%.

- The lowest TCR was found in # Dimension: 2 - # Class: 4 as 65.6%, the highest TCR was found in # Dimension: 3 - # Class: 4 scenario as 84.2%.
- The lowest F-Score value was found in # Dimension: 2 - # Class: 2 as 60.5%, the highest F-Score value was found in # Dimension: 1 - # Class: 2 scenario as 88.3%.
- The lowest Recall value in # Dimension: 2 - # Class: 2 as 53.1% thge highest Recall value was found in Dimension: 1 - # Class: 2 scenario as 91.1%.
- The lowest Precision # Dimension: 2 - # Class: 4 as 61.1%, the highest Precision value was found in Dimension # 1 - # Class: 2 scenario as 85.6%.

As a result of the PM for simulation data sets which have 50 cases in each class, looking at Table 4.8, the highest increase was observed in TCR value as 8.24%.

- The lowest TCR was found in # Dimension: 2 - # Class: 2 as 62.5%, the highest TCR was found in Dimension: 3 - # Class: 3 scenario as 86.9%.
- The lowest F-Score value was found in # Dimension: 2 - # Class: 3 as 54.3%, the highest F-Score value was found in # Dimension: 1 - # Class: 2 scenario as 77.2%.
- The lowest Recall value was found in # Dimension: 2 - # Class: 3 as 48.9%, the highest Recall value was found in # Dimension: 1 - # Class: 2 scenario as 78.9%.
- The lowest Precision was found in # Dimension: 2 - # Class: 2 and # Dimension: 2 - # Class: 3 as 61.0%, the highest Precision value was found in Dimension # 1 - # Class: 3 scenario as 79.9%.

Average gain for all performance criteria, for 25 cases and 50 cases in each class, have been given in Table 4.9 and Table 4.10. According to these tables, PM has increased all performance criteria.

- Precision has maximum gain as 4.02, Recall has minimum gain as 3.10 for 25 cases in each class data set.
- TCR has maximum gain as 8.24, Recall has minimum gain as 1.42 for 50 cases in each class data sets.

Table 4.5. (ICA+ RF) Results of simulated data sets for 25 cases at each class

# Dimensions	#Class	**TCR**	**F-Score**	**Recall**	**Precision**
1	2	0.743	0.835	0.857	0.814
1	3	0.715	0.576	0.532	0.628
1	4	0.657	0.629	0.569	0.702
2	2	0.730	0.616	0.548	0.704
2	3	0.704	0.658	0.621	0.700
2	4	0.694	0.628	0.602	0.657
3	2	0.714	0.645	0.607	0.689
3	3	0.700	0.624	0.578	0.678
3	4	0.712	0.645	0.598	0.700

Table 4.6. (ICA+KM+RF) Results of simulated data sets for 25 cases at each class

# Dimensions	#Class	**TCR**	**F-Score**	**Recall**	**Precision**
1	2	0.765	0.883	0.911	0.856
1	3	0.723	0.653	0.612	0.699
1	4	0.69	0.663	0.597	0.745
2	2	0.72	0.603	0.531	0.697
2	3	0.712	0.642	0.604	0.684
2	4	0.656	0.605	0.600	0.611
3	2	0.812	0.715	0.718	0.713
3	3	0.793	0.700	0.614	0.814
3	4	0.842	0.694	0.604	0.815

Table 4.7. (ICA+ RF) Results of simulated data sets for 50 cases at each class

# Dimensions	#Class	**TCR**	**F-Score**	**Recall**	**Precision**
1	2	0.676	0.694	0.802	0.611
1	3	0.712	0.673	0.645	0.703
1	4	0.701	0.662	0.637	0.689
2	2	0.615	0.587	0.568	0.607
2	3	0.638	0.553	0.499	0.619
2	4	0.634	0.569	0.543	0.597
3	2	0.714	0.649	0.604	0.702
3	3	0.724	0.635	0.586	0.693
3	4	0.716	0.633	0.578	0.699

Table 4.8. (ICA+KM+RF) Results of simulated data sets for 50 cases at each class

# Dimensions	#Class	**TCR**	**F-Score**	**Recall**	**Precision**
1	2	0.798	0.772	0.789	0.755
1	3	0.815	0.745	0.698	0.799
1	4	0.798	0.651	0.600	0.711
2	2	0.625	0.603	0.597	0.610
2	3	0.637	0.543	0.489	0.610
2	4	0.716	0.587	0.562	0.615
3	2	0.815	0.696	0.625	0.784
3	3	0.869	0.672	0.605	0.756
3	4	0.799	0.664	0.625	0.708

Table 4.9. Gain of PM for 25 cases at each class scenario (%)

# Dimensions	#Class	**TCR**	**F-Score**	**Recall**	**Precision**
1	2	2.2	4.80	5.4	4.2
1	3	0.8	7.70	8	7.1
1	4	3.3	3.40	2.8	4.3
2	2	-1	-1.30	-1.7	-0.7
2	3	0.8	-1.60	-1.7	-1.6
2	4	-3.8	-2.30	-0.2	-4.6
3	2	9.8	7.00	11.1	2.4
3	3	9.3	7.60	3.6	13.6
3	4	13	4.90	0.6	11.5
Average Gain%		3.82	3.36	3.10	4.02

Table 4.10. Gain of PM for 50 cases at each class scenario (%)

# Dimensions	#Class	**TCR**	**F-Score**	**Recall**	**Precision**
1	2	12.2	7.80	-1.3	14.4
1	3	10.3	7.23	5.3	9.6
1	4	9.7	-1.12	-3.7	2.2
2	2	1	1.66	2.9	0.3
2	3	-0.1	-0.97	-1	-0.9
2	4	8.2	1.86	1.9	1.8
3	2	10.1	4.62	2.1	8.2
3	3	14.5	3.71	1.9	6.3
3	4	8.3	3.11	4.7	0.9
Average Gain%		8.24	3.08	1.42	4.76

Table 4.11. Descriptive statistics of gain of PM for different #Cases in each class (%)

		TCR		Fscore		Recall		Precision	
#Dimension	#Cases	25	50	25	50	25	50	25	50
1	Mean	2,10	10,73	5,30	4,64	5,40	,10	5,20	8,73
	Standard Deviation	1,25	1,31	2,19	4,99	2,60	4,66	1,65	6,15
	Median	2,20	10,30	4,80	7,23	5,40	-1,30	4,30	9,60
	Minimum	,80	9,70	3,40	-1,12	2,80	-3,70	4,20	2,20
	Maximum	3,30	12,20	7,70	7,80	8,00	5,30	7,10	14,40
2	Mean	-1,33	3,03	-1,73	,85	-1,20	1,27	-2,30	,40
	Standard Deviation	2,32	4,51	,51	1,58	,87	2,03	2,04	1,35
	Median	-1,00	1,00	-1,60	1,66	-1,70	1,90	-1,60	,30
	Minimum	-3,80	-,10	-2,30	-,97	-1,70	-1,00	-4,60	-,90
	Maximum	,80	8,20	-1,30	1,86	-,20	2,90	-,70	1,80
3	Mean	10,70	10,97	6,50	3,81	5,10	2,90	9,17	5,13
	Standard Deviation	2,01	3,19	1,42	,76	5,41	1,56	5,95	3,79
	Median	9,80	10,10	7,00	3,71	3,60	2,10	11,50	6,30
	Minimum	9,30	8,30	4,90	3,11	,60	1,90	2,40	,90
	Maximum	13,00	14,50	7,60	4,62	11,10	4,70	13,60	8,20

Graphical Representation of Simulation Study (25 Cases)

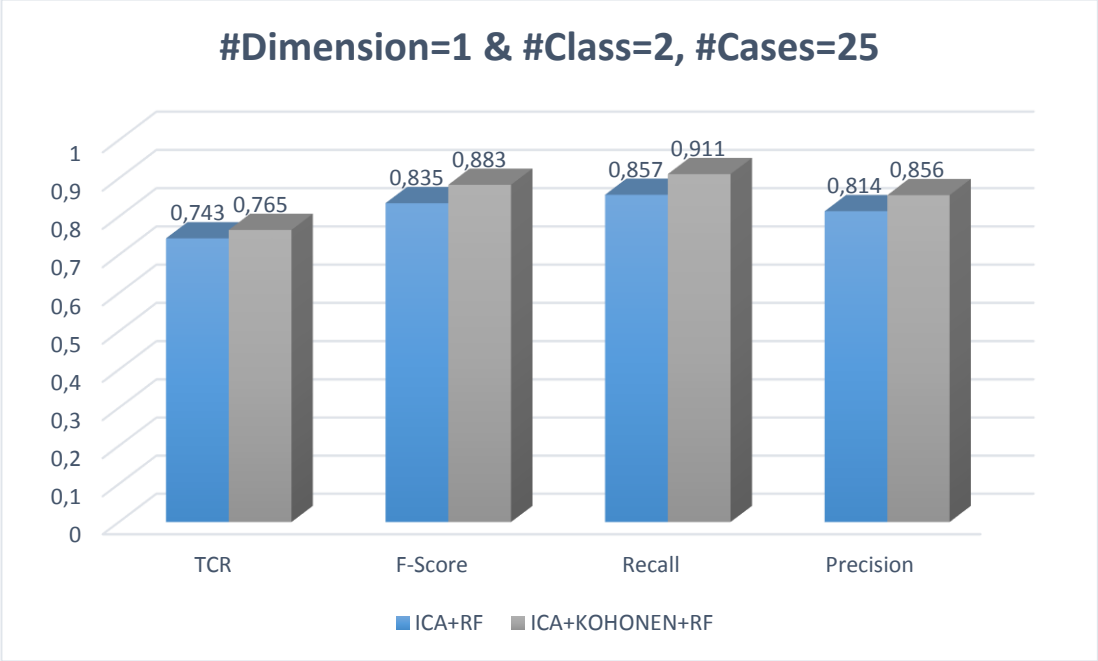

Figure 4.16. Results of ICA+RF and ICA+KM+RF methods on simulated data-1

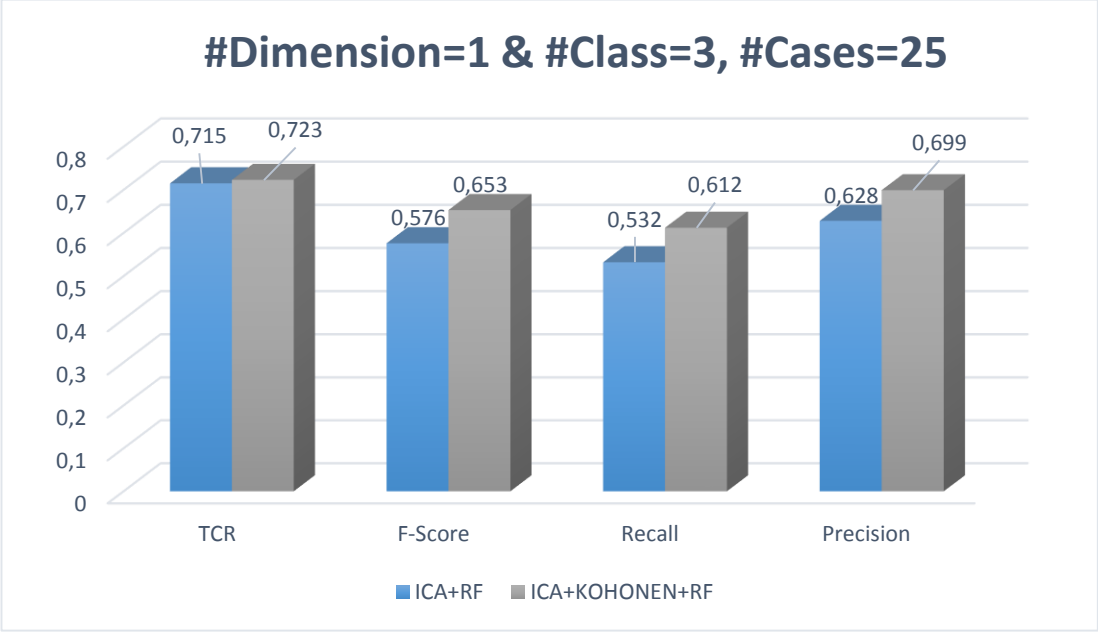

Figure 4.17. Results of ICA+RF and ICA+KM+RF methods on simulated data-2

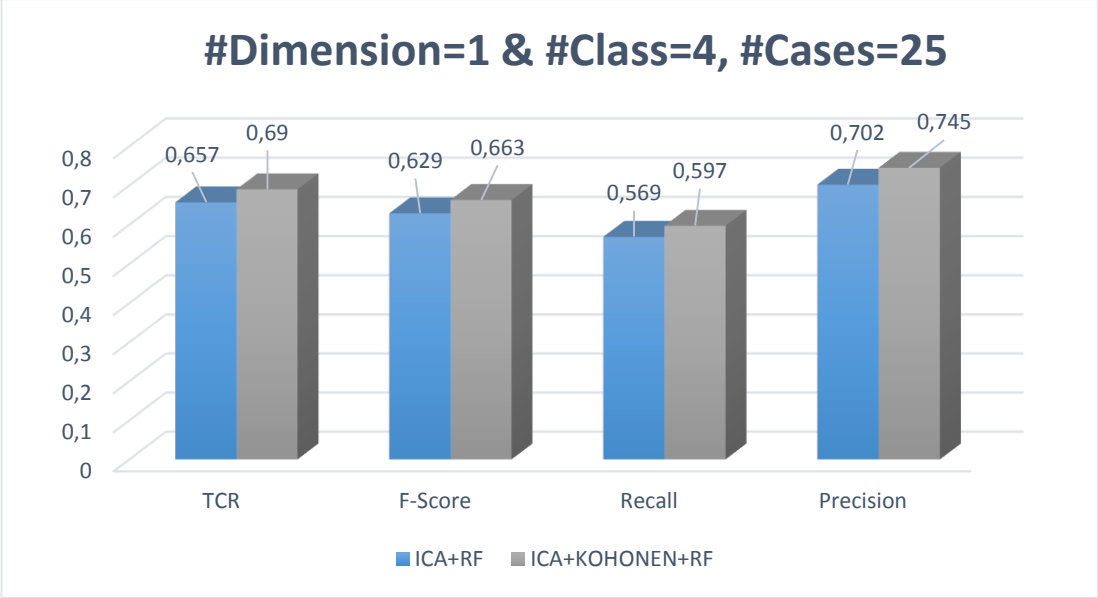

Figure 4.18. Results of ICA+RF and ICA+KM+RF methods on simulated data-3

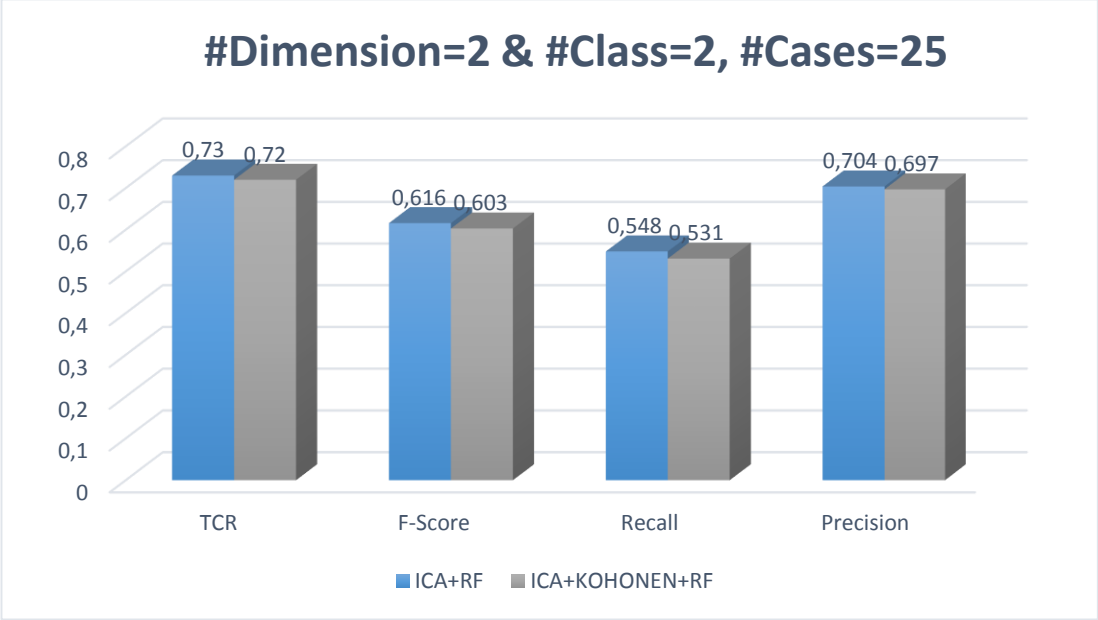

Figure 4.19. Results of ICA+RF and ICA+KM+RF methods on simulated data-4

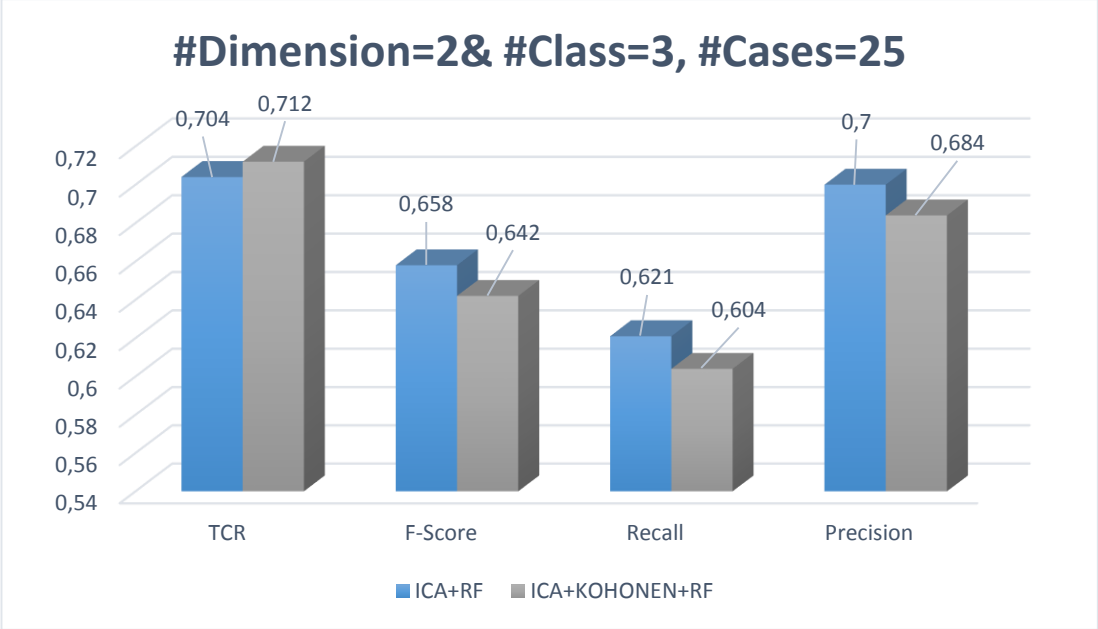

Figure 4.20. Results of ICA+RF and ICA+KM+RF methods on simulated data-5

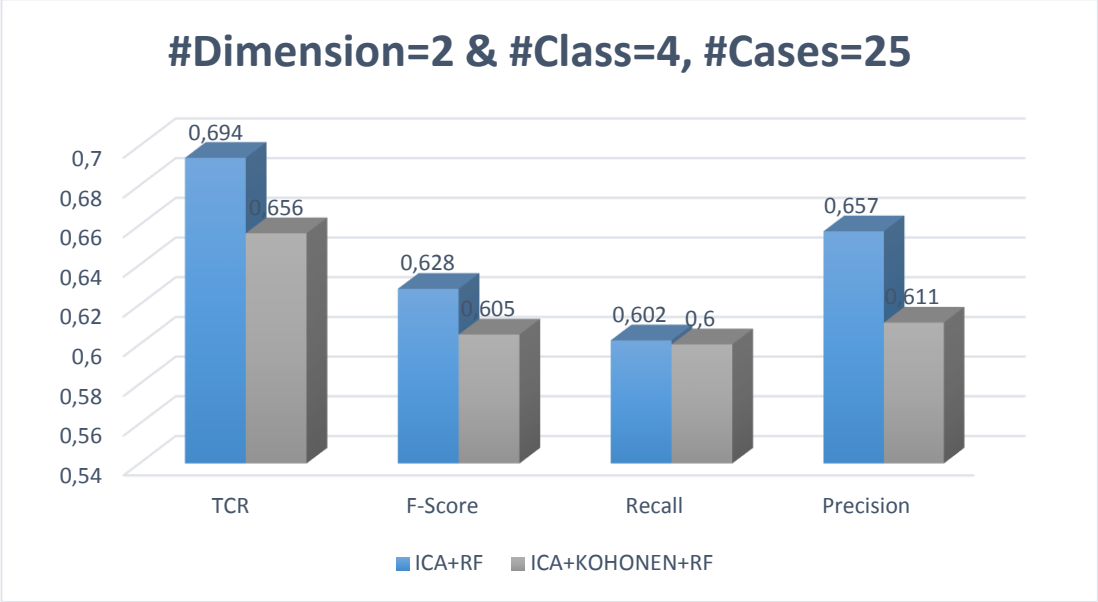

Figure 4.21. Results of ICA+RF and ICA+KM+RF methods on simulated data-6

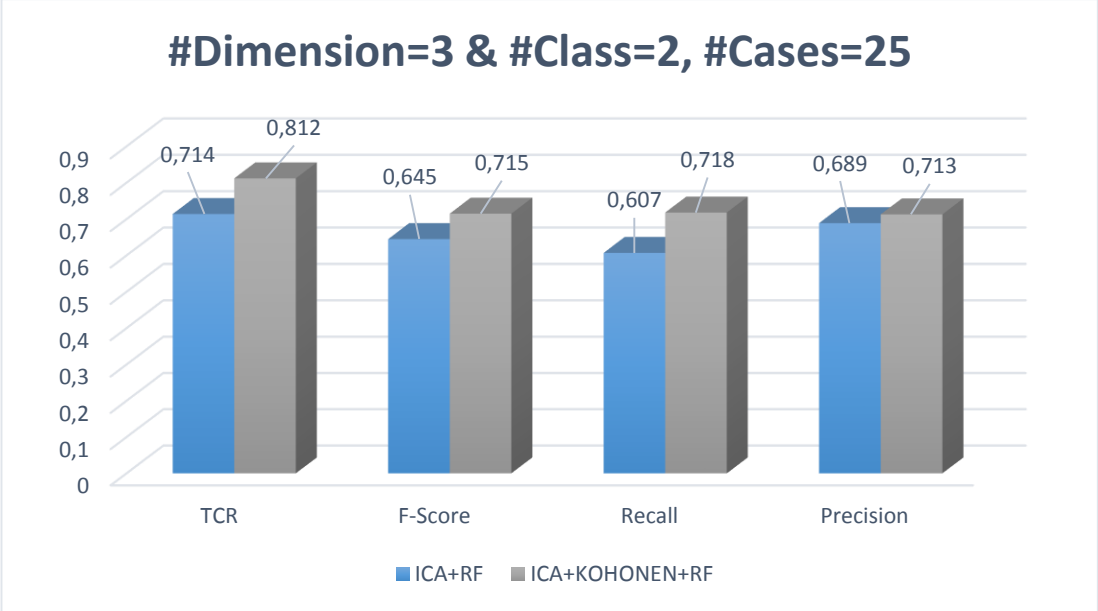

Figure 4.22. Results of ICA+RF and ICA+KM+RF methods on simulated data-7

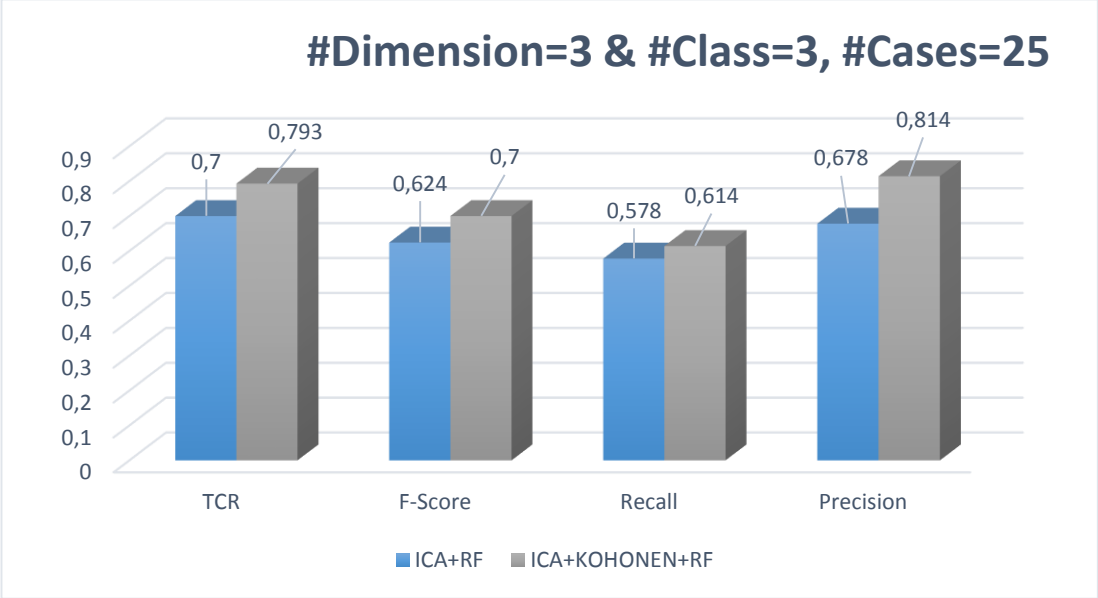

Figure 4.23. Results of ICA+RF and ICA+KM+RF methods on simulated data-8

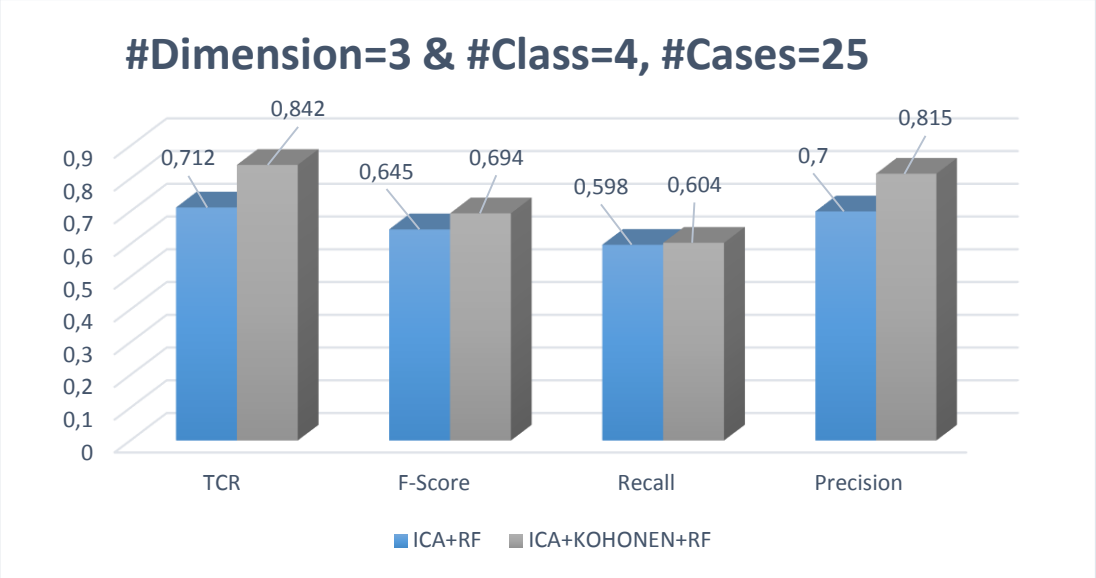

Figure 4.24. Results of ICA+RF and ICA+KM+RF methods on simulated data-9

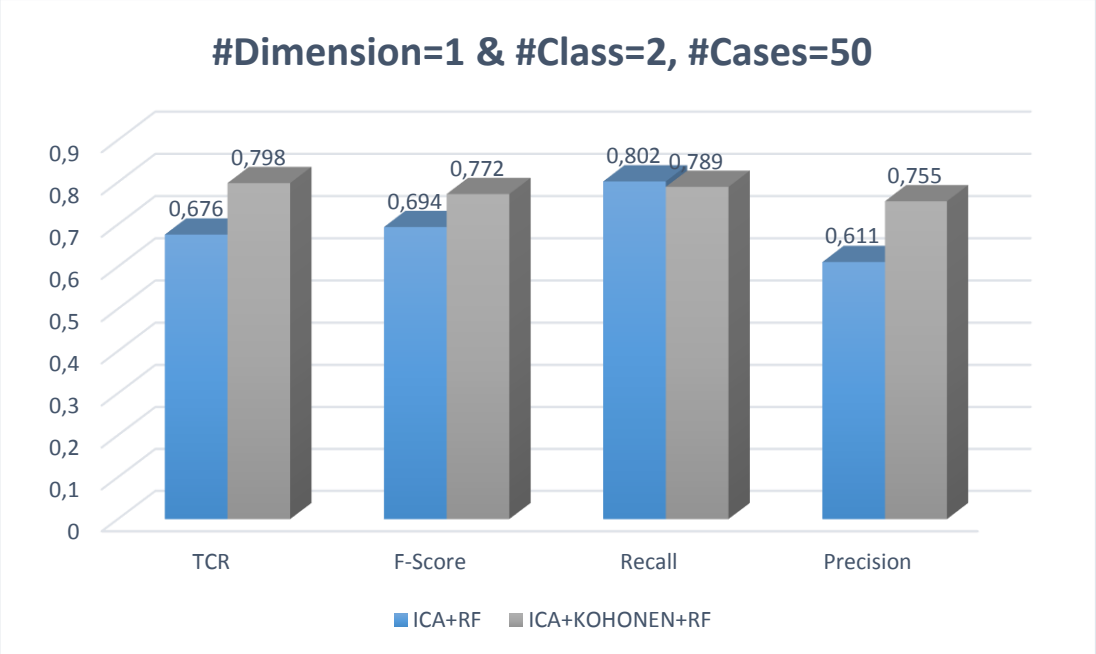

Figure 4.25. Results of ICA+RF and ICA+KM+RF methods on simulated data-10

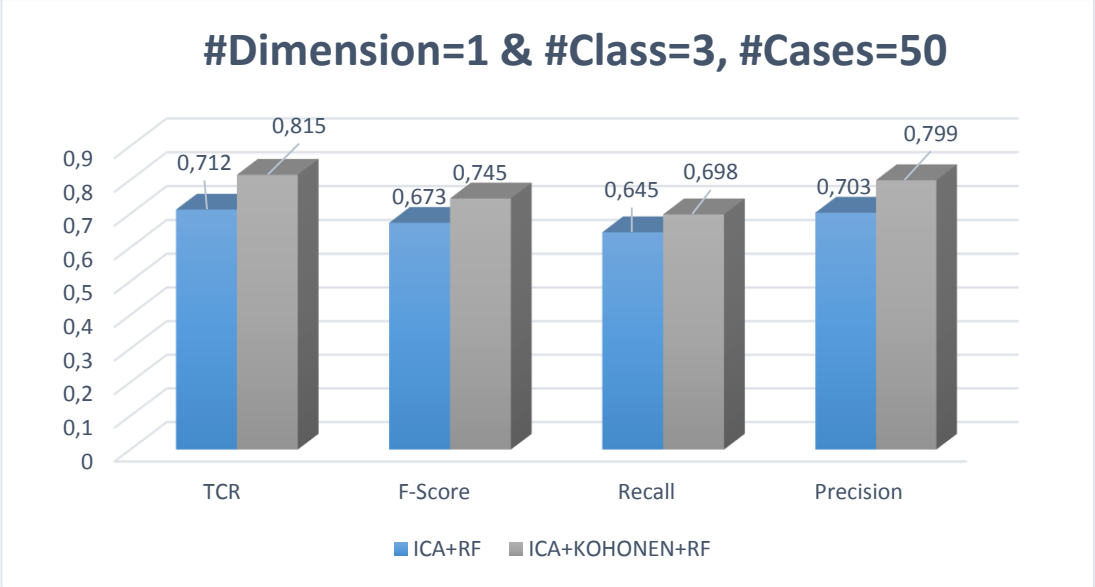

Figure 4.26. Results of ICA+RF and ICA+KM+RF methods on simulated data-11

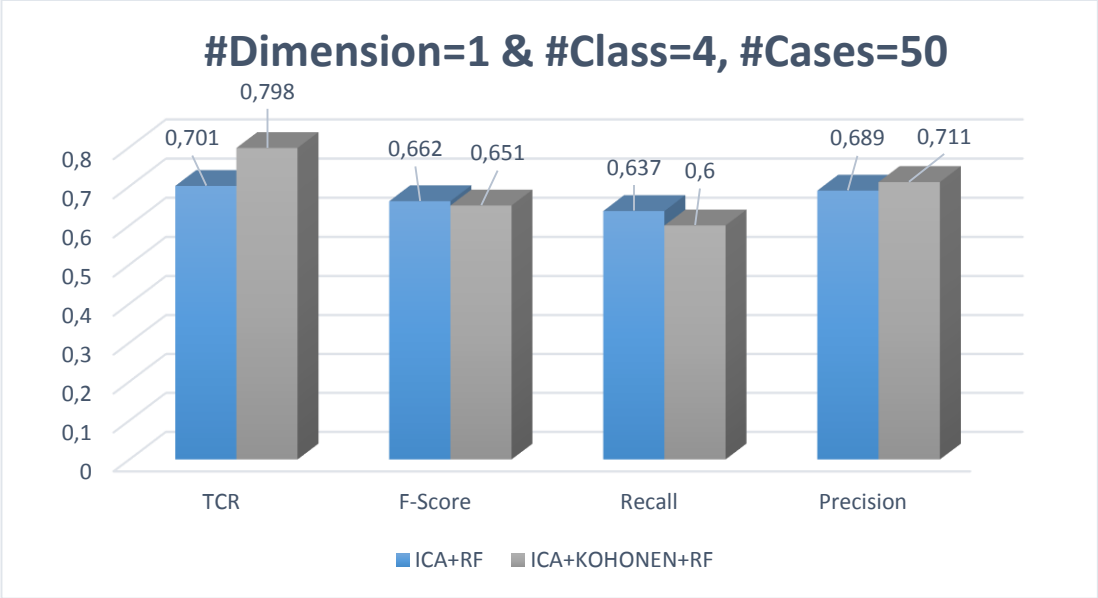

Figure 4.27. Results of ICA+RF and ICA+KM+RF methods on simulated data-12

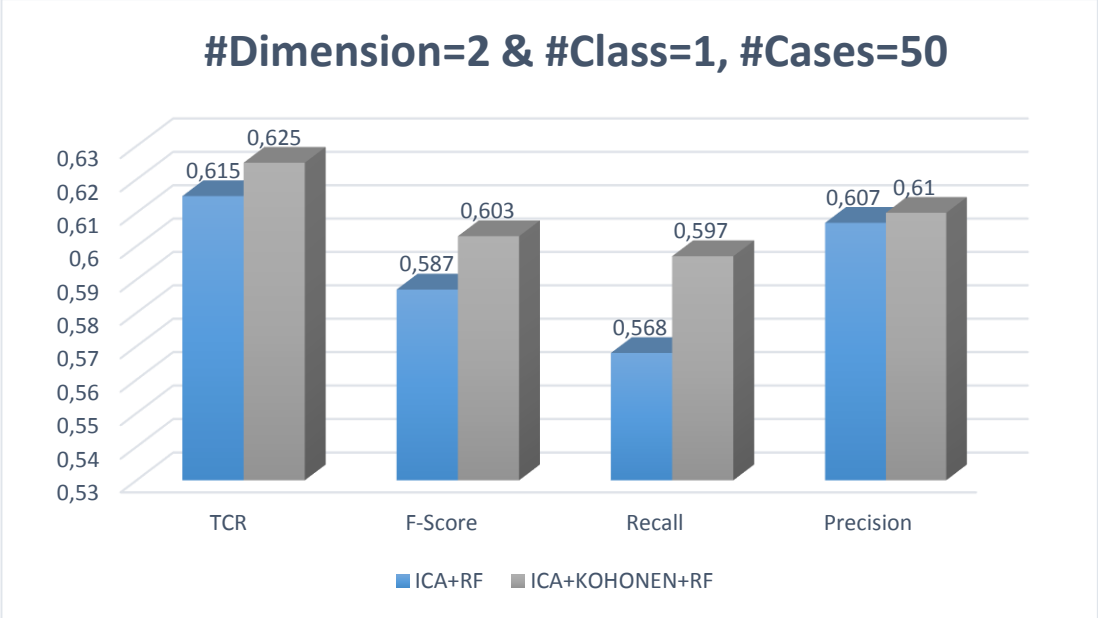

Figure 4.28. Results of ICA+RF and ICA+KM+RF methods on simulated data-13

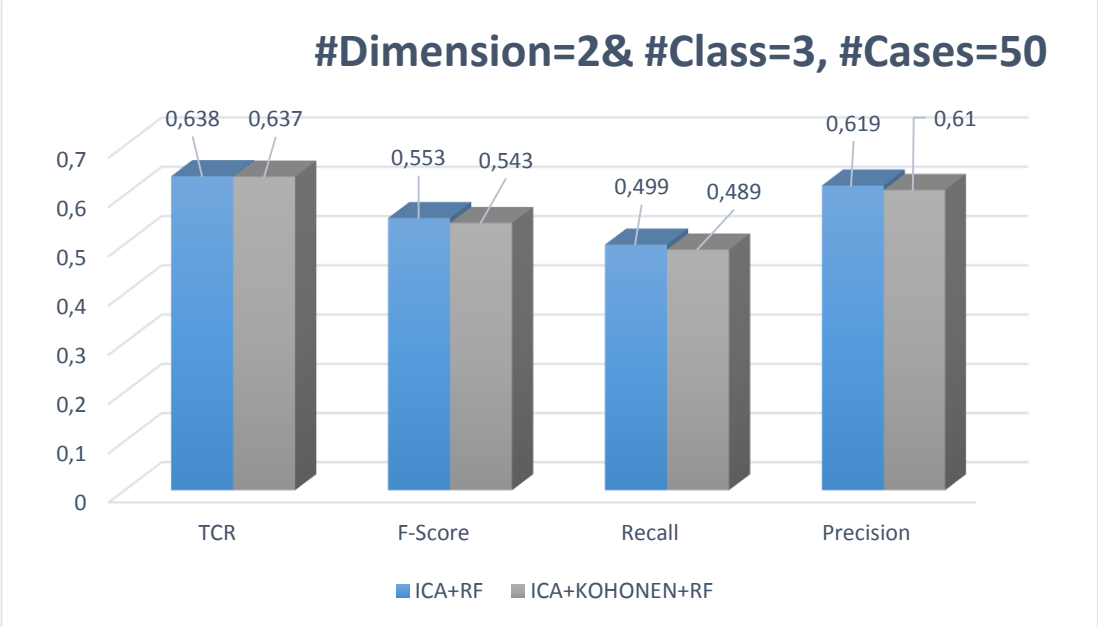

Figure 4.29. Results of ICA+RF and ICA+KM+RF methods on simulated data-14

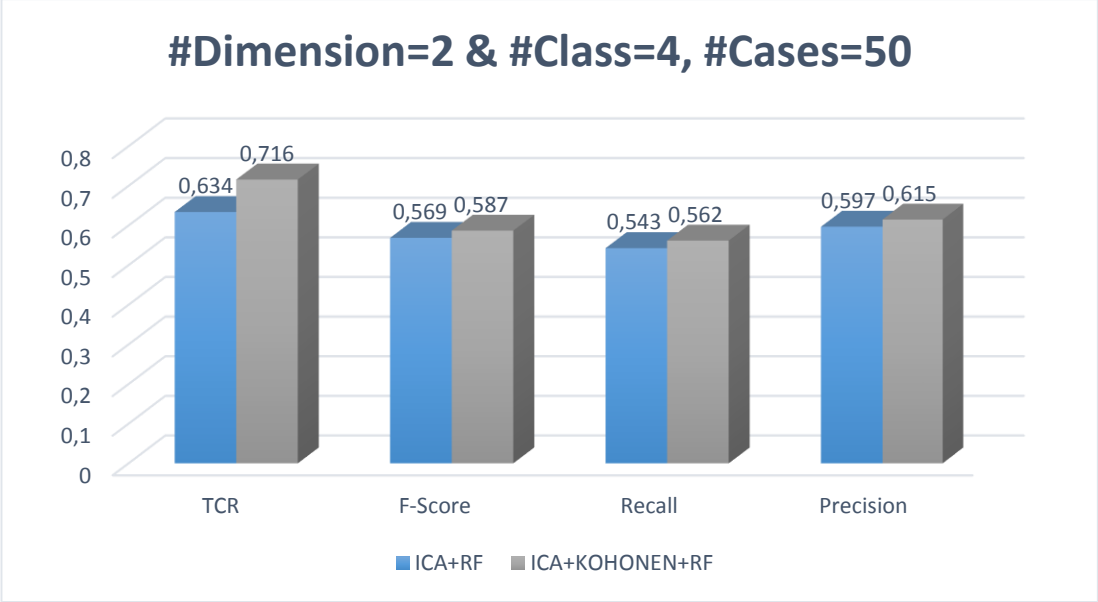

Figure 4.30. Results of ICA+RF and ICA+KM+RF methods on simulated data-15

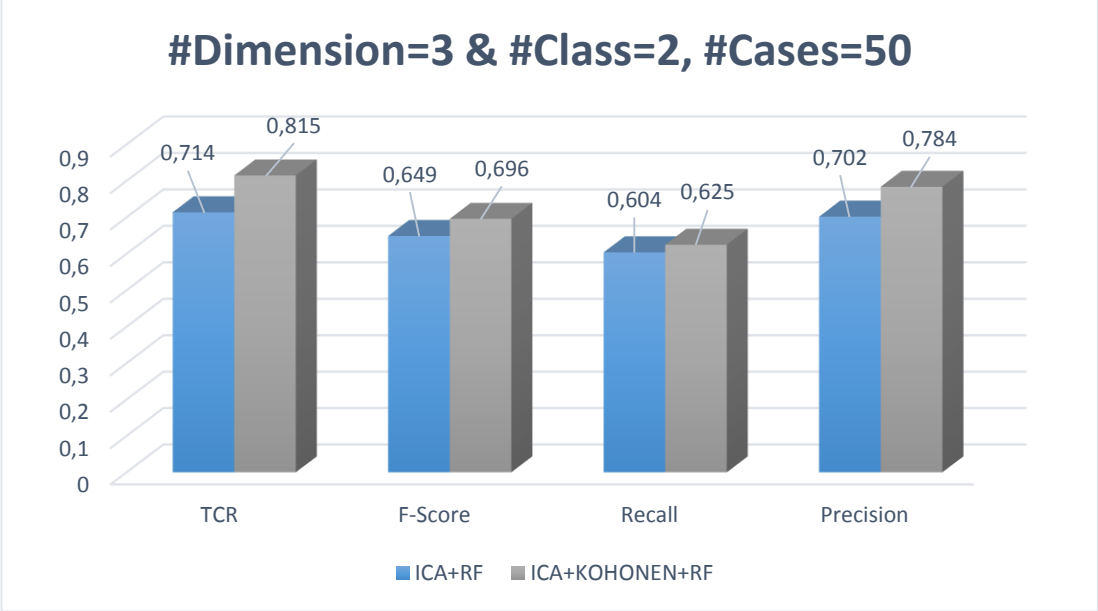

Figure 4.31. Results of ICA+RF and ICA+KM+RF methods on simulated data-16

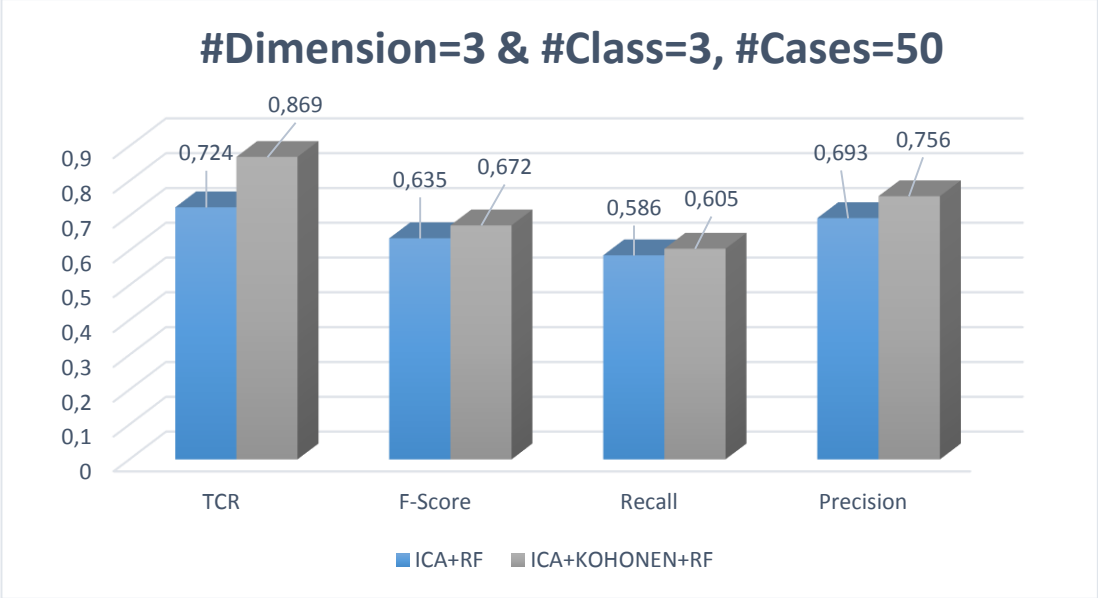

Figure 4.32. Results of ICA+RF and ICA+KM+RF methods on simulated data-17

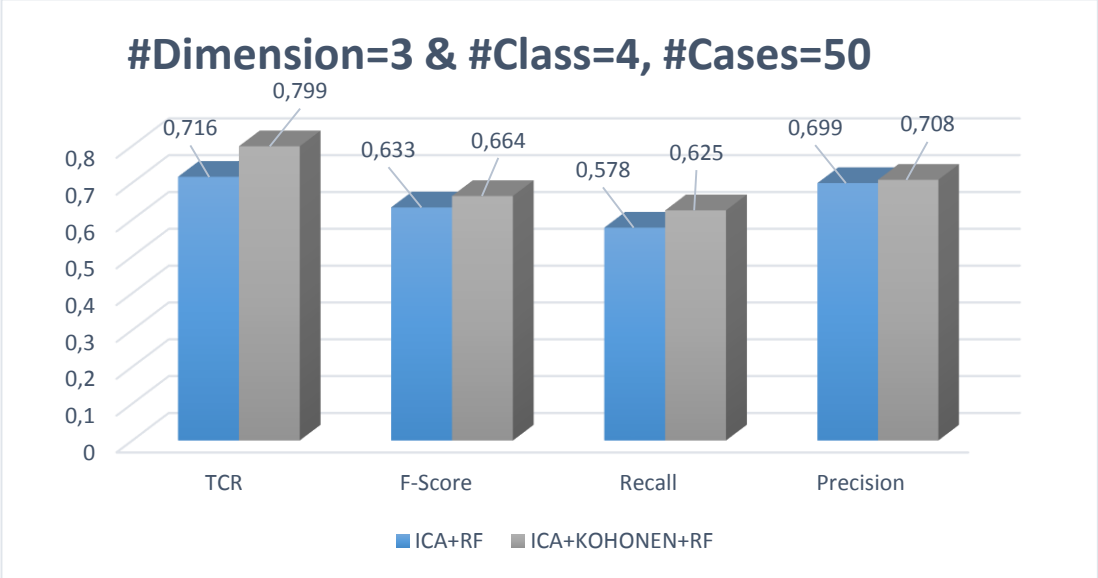

Figure 4.33. Results of ICA+RF and ICA+KM+RF methods on simulated data-18

5. DISCUSSION

The findings of this thesis were evaluated by four different performance measures. These measures are TCR, F-Score, Precision and Recall. The reason of choosing these performance measures is that each measure contributes from different point of views. This can be summarized as follows:

1) TCR: is a general performance measure as it shows the compliance of the real class value with the estimated class value at the end of the analysis.

2) Recall: is the measure to what extent accurate estimate has been done among the individuals whose real results are sick or positive.

3) Precision: shows to what extent the actual results of those individuals whose estimated results are sick or positive are also positive.

4) F-Measure: is an important measure that does not take TN values which take values between 0 and 1 into account, calculated as the weighted average of precision and recall.

When analysis of the 15 data sets obtained from the GEO database are examined (Table 4.1- 4.4), it is observed that an increase has been obtained in TCR, the F-Measure, Recall, Precision values in all data sets. The most distinguished increase was realized in data with 3 classes immediately after the data with 2 classes for all performance measures. The less number of increases obtained in the data with 4 and 5 class shows that the trend is toward reduction in the increase obtained in the performance measures as the number of classes increase. (Table 4.4) The lower values were obtained by ICA + RF approach for the 5th, 10th and 12th data sets with the proposed approach for calculated Recall, F-Measure values. The increase obtained for all other data sets indicates the TP rate is especially improving. When the information about these data sets are investigated (Table 3.2), it is seen that the data has 5, 5, 4 classes respectively. These comments do not apply only to the 12th data set according to the precision values. The precision value of the related data increased at a level of 0.7 according to the ICA + RF method. This case can be ignored and will not change the overall ratings as it is not a very high level. In addition, the

error that can be done by looking at the values of recall or precision alone with the F-Measure have become more reliable with weighted values. Simulated data sets have focused on two different scenarios. They are the cases on which the number of observations is 25 and 50 in each class. The reason why these numbers are chosen is that we do not often encounter more individuals used in these studies than these numbers in real studies due to the actual cost, time problems etc..

An increase has been provided in all performance measures in both 2 scenarios according to the findings obtained from these data sets. Higher values were obtained compared with the other scenario for TCR and precision measurements in the data with 50 cases in each class when Table 4.9 and Table 4.10 are examined. The scenario with 25 cases in each class showed better performance in Recall and F-Score values. It is difficult to relate this situation to any cause in terms of data mining. However, when it is examined in terms of TCR used as the overall performance measure, the increase in the number of cases in the classes may seem to be effective for the established RF model has given better results. As one of the most important criteria in the RF model is as follows: TCR is expected to increase as the number of variables used in the creation of the trees in each division (mtry) increases.

Studying the findings in terms of dimension: The PM provided an increase in performance measures for data with 50 cases in each class at all dimension levels. The highest increase for Recall the TCR was obtained at 3 dimensions and for the F-score and precision it was obtained at 1 dimension. The increase was less in two dimensional scenerios when compared with other dimension levels. The comment might be drawn from this is that when the higher TCR is targeted in data which has 1 and 3 dimensions with 50 or a similar number of cases in each class (both TP and TN), the proposed approach may be expected to give better results. In the same way, when a comment is required for such a weighted measure of the performance as the F-Score, further gain will be provided for the data that has 1 and 3 dimensions. When scenarios with 25 cases in each class are studied, 3 dimensions scenarios give better results in all other performance measures apart from the recall value. This is

especially important for the analysis of the data with a smaller number of individuals and better distinguishable groups, for they are performed on more clearly distinguished individuals such as case – control studies or lower number of individuals. When the artificial scenarios in general are examined, low yields were obtained compared with those obtained from the data taken from the GEO database. This situation can be explained with the fact that the scenarios established in artificial cases are more different and strict than those in real situations. The PM, which enhanced the performance also in biological knowledge, has realized a very considerable job in previously unidentified, even just based on mathematical models.

6. CONCLUSION

The main hypothesis of this thesis was that there had to be a hybrid approach which would eliminate the disadvantages of high dimensionality in gene expression studies. Here it was observed that the researchers reduced the dimension of the data (classification, clustering identifier) before carrying out any analysis (Classic Approach-CA) as a result of research of the literature in the field and the methods such as PCA, FA and ICA came to the fore for this purpose. Through these methods, the genes with similar structures are factorized in terms of the gene expression levels, and both the computation time had been saved and it had been tried to get unnecessary information averted in the subsequent analysis. However, this approach was modified just because taking the idea that dimension reduction does not contribute to the results at a sufficient level in this type of studies. With the change, the factorized genes were put into similar clusters in the second stage and then sent to the RF algorithm randomly with the bootstrap approach. (With the Proposed Method-PM) We can list the results which this approach provides as follows when Table 4.1-Table 4.11 in the results chapter are analyzed.

1) Original and simulated data sets have been used within the scope of this thesis. The aim here was to see how PM affects the simulated data with real examples. In this way, we aimed to make more accurate recommendations for the researchers. As it is seen in the results chapter, a certain level of knowledge gain was obtained in both data types.

2) For data containing 2 and 3 groups, an increase which might be important for genetic research provided in the PM the level of performance measures was compared with CA. This means that estimating model that the researchers to be obtained according to the PM will be more accurate when the number of groups is 2 and 3. This way, the researchers aiming to develop treatment special to the individual will be able to apply useful therapies on the convienient people-groups with the data available.

3) CA and PM performed similarly in the cases when the number of groups is 4 and 5. This can actually be considered as an expected condition. It is especially difficult to distinguish individuals with different clinical and genetic characteristics. The calculation of the variance-covariance matrix becomes more complex mathematically as the size increases. It becomes difficult to reveal correlations between genes and between individuals. Due to the fact that DM methods failed to put forth these relationships in multi-dimensional surfaces, it was observed that the PM prediction models have not produced much higher increase than the CA.

4) It was observed that a better distinctive result was obtained particularly in smaller numbers when the results are examined in terms of sample size since this type genetic researches require a considerable budget in developing countries in what seems to concern the financial situation and insufficient infrastructure. Therefore, these studies generally take place with a smaller number of individuals. PM is directly applicable in this respect.

5) Less performance increase was observed compared with real data sets when the analysis on simulated data was examined. Even decline was observed in some dimensions. This is thought to originate from synthetic data algorithm. Adding an extra step of clustering to the data set which has dimensions expected to have an ample way to analyze created the opposite effect in some cases. The conclusion to be drawn here is that there will not be a difference between using the CA or the PM if the data set is in a structure that the groups are well distinguished. However, the increase obtained in cases where the number of dimensions are 1 and 3 can be interpreted as a sign that better results can be obtained with new approaches on this issue.

6) When the result in Table 4.11 obtained with the analysis of simulated data, getting better results on the data with 50 cases in each class for TCR proves that the dimension effect declines as the number of cases increases and indicates that the PM can achieve a better classification performance.

7) The main objective of this thesis is not to test the impact of bootstrap. The objective of using this method is to send clustered factors to the RF algorithm as randomly as possible. If randomness had not been achieved at this point, there would have been some question marks that RF algorithm always uses the same clusters, and therefore always uses the same factors and the same genes at the most distant point. For this purpose 1000 bootstrap sample approach increased the reliability of the results and at the same time eliminated the problems that might originate from the RF. From this perspective, the only questionable aspect of the RF results was ended.

8) It is suggested to give the test and training set results separately when reporting. The reason of that is to check the results of the analysis one by one in overfitting cases and comment on the results originating from similarity of the test-train set results. However, with the help of innovations introduced by the PM, the bootstrap step theoretically eliminates this situation. 1000 different sets of data have been expressed with performance measures at each step, and similarity of test-train set has been eliminated.

9) In the “dimension reduction than use classification models” approach which is used in classical gene expression data analysis, selecting all of the components randomly with the PM and using in the classification algorithm is important for using all the available data instead of using the first 5 or 10 components as Bayer et al. (28) did.

10) The application of the proposed method (PM) in the thesis was carried out with the R programming language. The use of this software and the development of new approaches require a distinct professionalism. To develop a user interface for easy use of researchers was also within the scope of the thesis. With the help of this software, the PM has provided the opportunity to apply on desired data at the desired time. The tool, Gene3E, described in detail in Chapter 3.7 generally provides the user with the application of the program only by posting the parameters of pieces of codes prepared for R. At the same time, only the dimension reduction and classification analyzes can also be made by closing extra clustering phase proposed by the PM. In this way, it is clear that it will be a significant source of help for the researchers in our country.

REFERENCES

1 Ackermann, M., Sikora-Wohlfeld, W.,Beyer, A. (2013) Impact of Natural Genetic Variation on Gene Expression Dynamics. *Plos Genetics*, 9 (6).

2 Baelde HJ, Eikmans M, Doran PP, Lappin DW et al. (2004), Gene expression profiling in glomeruli from human kidneys with diabetic nephropathy. *American Journal of Kidney Disorders*.;43(4):636-50

3 Barth AS, Merk S, Arnoldi E, Zwermann L et al. (2005), Functional profiling of human atrial and ventricular gene expression. *Pflügers Archiv European Journal of Physiology* ; 450(4):201-8.

4 Bayer, I., Groth, P.,Schneckener, S. (2013) Prediction errors in learning drug response from gene expression data - influence of labeling, sample size, and machine learning algorithm. *PLoS One*, 8 (7), e70294.

5 Breiman, L. (1984), Classification and Regression Trees. Belmont CA, Wadsworth International Group

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

7 Chandran UR, Ma C, Dhir R, Bisceglia M et al. (2007), Gene expression profiles of prostate cancer reveal involvement of multiple molecular pathways in the metastatic process. B*MC Cancer*, 12;7:64.

8 Cosgun, E., Limdi, N.A.,Duarte, C.W. (2011) High-dimensional pharmacogenetic prediction of a continuous trait using machine learning techniques with application to warfarin dose prediction in African Americans. *Bioinformatics*, 27 (10), 1384-1389.

9 Covell, D.G., Wallqvist, A., Rabow, A.A.,Thanki, N. (2003) Molecular classification of cancer: Unsupervised self-organizing map analysis of gene expression microarray data. *Molecular Cancer Therapeutics*, 2 (3), 317-332.

10 Cui X., Churchill G.A. (2003) Statistical tests for differential expression in cDNA microarray experiments. *Genome Biology* 4(210), 1-10

11 David Rocke. Jian J. Dai. (2003), Sampling and Subsampling for Cluster Analysis in Data Mining: With Applications to Sky Survey Data, Data Mining and Knowledge Discovery, 01/2003; 7:215-232.

12 Diaz-Uriarte, R.,Alvarez de Andres, S. (2006) Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7, 3.

13 Dihal AA, van der Woude H, Hendriksen PJ, Charif H et al. (2008) Transcriptome and proteome profiling of colon mucosa from quercetin fed F344 rats point to tumor preventive mechanisms, increased mitochondrial fatty acid degradation and decreased glycolysis. *Proteomics*; 8(1):45-61.

14 Dinger, S.C., Van Wyk, M.A., Carmona, S.,Rubin, D.M. (2012) Clustering gene expression data using a diffraction-inspired framework. *Biomedical Engineering Online*, 11.

15 Graham Williams (2009). Rattle: A Data Mining GUI for R, Graham J Williams, *The R Journal*, 1(2):45-55

16 Hansheng Lei, V.G. (2005) *Speeding Up Multi-class SVM by PCA and Feature Selection. Feature Selection in Data Mining*, Workshop in conjunction with the 5th SIAM International Conference on Data Mining,California,US.

17 Hindmarch C, Yao S, Beighton G, Paton J et al. (2006), A comprehensive description of the transcriptome of the hypothalamoneurohypophyseal system in euhydrated and dehydrated rats. *Proceedings of the National Academy of Sciences*; 103(5):1609-14.

18 Hori, G., M. Inoue, S. Nishimura, and H. Nakahara. (2001) Blind gene classification based on ICA of microarray data. *International Conference on Independent Component Analysis and Signal Separation*, 3, 332-336.

19 Hyvärinen, A., J. Karhunen, and E. Oja. (2001) *Independent Component Analysis. A Volume in the Wiley Series on Adaptive andLearning Systems for Signal Processing, Communications, and Control.*

20 Ioannidis, J.P.A., Allison, D.B., Ball, C.A., Coulibaly, I., Cui, X.Q., Culhane, A.C. ve diğerleri. (2009) Repeatability of published microarray gene expression analyses. *Nature Genetics*, 41 (2), 149-155.

21 Jian J. Dai, L.L., David Rocke. (2006) Dimension Reduction for Classification with Gene Expression Microarray Data. *Statistical Applications in Genetics and Molecular Biology*, 5 (1), 6.

22 Johannes, M., Frohlich, H., Sultmann, H.,Beissbarth, T. (2011) pathClass: an R-package for integration of pathway knowledge into support vector machines for biomarker discovery. *Bioinformatics*, 27 (10), 1442-1443.

23 Kim, K., Zakharkin, S.O., Allison, D.B. (2010) Expectations, validity, and reality in gene expression profiling. *Journal of Clinical Epidemiology*, 63 (9), 950-959.

24 Komura, D., Nakamura, H., Tsutsumi, S., Aburatani, H.,Ihara, S. (2005) Multidimensional support vector machines for visualization of gene expression data. *Bioinformatics*, 21 (4), 439-444.

25 Kong, W., Vanderburg, C.R., Gunshin, H., Rogers, J.T., Huang, X.D. (2008) A review of independent component analysis application to microarray gene expression data. *Biotechniques*, 45 (5), 501-+.

26 Lee, S.I.,Batzoglou, S. (2003) Application of independent component analysis to microarrays. *Genome Biology*, 4 (11).

27 Li, B., Zheng, C.H., Huang, D.S., Zhang, L.,Han, K. (2010) Gene expression data classification using locally linear discriminant embedding. *Computers in Biology and Medicine*, 40 (10), 802-810.

28 Li, H., Zhang, K.,Jiang, T. (2005) Robust and accurate cancer classification with gene expression profiling. *Proc IEEE Computer Systems and Bioinformatics Conference,* 310-321.

29 Liebermeister, W. (2002) Linear modes of gene expression determined by independent component analysis. *Bioinformatics*, 18 (1), 51-60.

30 Liu, Q.Z., Sung, A.H., Chen, Z.X., Liu, J.Z., Chen, L., Qiao, M.Y. ve diğerleri. (2011) Gene selection and classification for cancer microarray data based on machine learning and similarity measures. *BMC Genomics*, 12.

31 Lyckman AW, Horng S, Leamey CA, Tropea D et al. (2008), Gene expression patterns in visual cortex during the critical period: synaptic stabilization and reversal by visual deprivation. *Proceedings of the National Academy of Sciences;* 105(27):9409-14. PMID: 18606990

32 Martens JH, Kzhyshkowska J, Falkowski-Hansen M, Schledzewski K et al. (2006) Differential expression of a gene signature for scavenger/lectin receptors by endothelial cells and macrophages in human lymph node sinuses, the primary sites of regional metastasis. *Journal of Pathology*.,Mar; 208(4):574-89.

33 McConnell, P., Johnson, K.,Lockhart, D.J. (2002) An introduction to DNA microarrays. *Methods of Microarray Data Analysis* Ii, 9-21

34 Mobini R, Andersson BA, Erjefält J, Hahn-Zoric M et al.(2009), A module-based analytical strategy to identify novel disease-associated genes shows an inhibitory role for interleukin 7 Receptor in allergic inflammation. *BMC System Biology*, 12;3:19.

35 Mogass M, York TP, Li L, (2004), Rujirabanjerd S et al. Genomewide analysis of gene expression associated with Tcof1 in mouse neuroblastoma. Biochemical and Biophysical Research Communications, 325(1):124-32. PMID: 15522210

36 Moorthy, K.,Mohamad, M.S. (2011) Random forest for gene selection and microarray data classification. *Bioinformation*, 7 (3), 142-146.

37 Najarian, K., Kedar, A., Paleru, R., Darvish, A.,Zadeh, R.H. (2004) Independent component analysis and scoring function based on protein interactions. 2004 2nd International Ieee Conference Intelligent Systems, Vols 1 and 2, Proceedings, 595-599.

38 Nannapaneni, P., Hertwig, F., Depke, M., Hecker, M., Mader, U., Volker, U. ve diğerleri. (2012) Defining the structure of the general stress regulon of Bacillus subtilis using targeted microarray analysis and random forest classification. Society for General *Microbiology*, 158, 696-707.

39 Nelson B, Nishimura S, Kanuka H, Kuranaga E, Inoue M, Hori G, Nakahara H, Miura M. (2005), Isolation of gene sets affected specifically by polyglutamine expression: implication of the TOR signaling pathway in neurodegeneration, *Cell Death and Differentiation*. Aug;12(8):1115-23.

40 Okun, O., Priisalu, H. (2007) Random forest for gene expression based cancer classification: Overlooked issues. *Pattern Recognition and Image Analysis, Pt 2, Proceedings*, 4478, 483-490.

41 Parikh H, Carlsson E, Chutkow WA, Johansson LE et al.(2007) TXNIP regulates peripheral glucose metabolism in humans. *PLoS Med*icine, May;4(5):e158.

42 Rezen T, Juvan P, Fon Tacer K, Kuzman D et al. (2008),The Sterolgene v0 cDNA microarray: a systemic approach to studies of cholesterol homeostasis and drug metabolism. *BMC Genomics*; 11; 9:76.

43 Suri, R.E. (2003) Application of independent component analysis to microarray data. *International Conference on Integration of Knowledge Intensive Multi-Agent Systems*, 375-378.

44 Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E. ve diğerleri. (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences* ,U S A, 96 (6), 2907-2912.

45 Tatenhorst L, Püttmann S, Senner V, Paulus W.(2005), Genes associated with fast glioma cell migration in vitro and in vivo. Brain Pathology; 15(1):46-54. PMID: 15779236

46 Torkkola, K., Gardner, R.M., Kaysser-Kranich, T.,Ma, C. (2001) Self-organizing maps in mining gene expression data. *Information Sciences*, 139 (1-2), 79-96.

47 Toronen, P., Kolehmainen, M., Wong, C.,Castren, E. (1999) Analysis of gene expression data using self-organizing maps. *Febs Letters*, 451 (2), 142-146.

48 Trosset, M.W. (2008) *Representing Clusters:K-Means Clustering, Self-Organizing Maps, and Multidimensional Scaling*. Technical Report 08-03, Indiana University, Bloomington,IN,US.

49 Urbanek, S. (2003), *"Rserve - A Fast Way to Provide R Functionality to Applications"* in Proc. of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003), ISSN 1609-395X, Eds.: Kurt Hornik, Friedrich Leisch & Achim Zeilei

50 Vapnik, V.,Chapelle, O. (2000) Bounds on error expectation for support vector machines. *Neural Computation*, 12 (9), 2013-2036.

51 Viemann D, Goebeler M, Schmid S, Nordhues U et al.(2006), TNF induces distinct gene expression programs in microvascular and macrovascular human endothelial cells*. Journal of Leukocyte Biology*;80(1):174-85

52 Wang, X.S., Simon, R. (2011) Microarray-based cancer prediction using single genes. *BMC Bioinformatics*, 12.

53 Web Page: http://www.ncbi.nlm.nih.gov/geo/info/faq.html-Reached 23.10.2012

54 Wenzel K, Zabojszcza J, Carl M, Taubert S et al. (2005), Increased susceptibility to complement attack due to down-regulation of decay-accelerating factor/CD55 in dysferlin-deficient muscular dystrophy. The *Journal of Immunology*, 175(9):6219-25. PMID: 16237120

55 Willem Talloen, Hinrich GѴðhlmann (2009) *Gene Expression Studies Using Affymetrix Microarrays. Gene Expression Studies Using Affymetrix Microarrays, Chapter 1, 1-15.*

56 Witten D., Tibshirani R. (2007) *A comparison of fold-change and the t-statistic for microarray data analysis*. Technical Report, Stanford University

57 Woodruff PG, Koth LL, Yang YH, Rodriguez MW et al. (2005), A distinctive alveolar macrophage activation state induced by cigarette smoking. American Journal of Respiratory and Critical Care Medicine, 72(11):1383-92. PMID: 16166618

58 Yuanwei Zhang, Y.Y., Huan Zhang, Xiaohua Jiang, Bo Xu, Xue Yue, Yunxia Cao, Qian Zhai, Yong Zhai, Mingqing Xu, Howard J. Cooke, Shi Qinghua. (2011) Prediction of Novel Pre-microRNAs with High Accuracy through Boosting and SVM. *Bioinformatics*, 27 (10), 1436-1437.