



TECHNICAL UNIVERSITY OF CRETE  
DEPARTMENT OF ELECTRONIC AND COMPUTER ENGINEERING  
DIGITAL SIGNAL & IMAGE PROCESSING LAB

---

**Genetic Data analysis for Classification of bipolar disorders**

---

**Diploma Thesis**

**Leska Valsamo**

*Chania, 2015*

**Thesis Committee**

Professor Michael Zervakis, *Thesis Supervisor*

Professor Costas Balas

Professor Euripides Petrakis

## **Abstract**

*In the recent years DNA microarray analysis has become a widely used tool for gene expression profiling and data analysis. This technology can be useful in the classification of complex diseases such as bipolar disorder, providing useful information for its genetic background. Bipolar disorder is a common, heritable mental illness characterized by recurrent episodes of mania and depression that manifests from multiple genetic and environmental factors. There are four basic types of bipolar disorder; bipolar I disorder, bipolar II disorder, Bipolar Disorder Not Otherwise Specified (BP-NOS) and Cyclothymic. The ability to classify bipolar disorders may have a major impact on our understanding of disease pathophysiology and may provide important opportunities to investigate the interaction between genetic and environmental factors involved in pathogenesis. Also this ability may be essential to guide appropriate therapy and determine prognosis for successful treatment. The aim of this diploma thesis is to extract a significant genomic signature for which biological knowledge already exists or discover novel genomic information, which might stand as the motivation for further analysis. Under this genomic signature we classify the bipolar disorders using gene expressions from two different populations.*

*Microarray analysis normally leads to datasets which contain a small number of samples which have a large number of gene expression levels as features. In order to extract useful informative sets of genes that can reduce dimensionality and maximize the performance of classifiers, feature selection algorithms were used. Another aim of this study is to achieve stable performance assessment of feature selection and classification methods. In that manner, the genetic evaluation framework named "Stable Bootstrap Validation" (SBV), introduced by Nick Chlis, is presented. The SBV utilizes bootstrap resampling of the original dataset and an explicit criterion that determines the stability of the observed classification accuracy and the biological interpretation of genes, also called genomic signature. Moreover, methodologies for evaluating the discrimination, consistency and generalization ability of the observed results are also introduced. In this diploma thesis a unified "32 common gene signature" was extracted, which is closely associated with several aspects of bipolar disorders.*

## Περίληψη

Τα τελευταία χρόνια η ανάλυση μικροσυστοιχειών DNA έχει γίνει ένα ευρέως χρησιμοποιούμενο εργαλείο για μέτρηση των τιμών έκφρασης χιλιάδων γονιδίων και για την ανάλυση δεδομένων. Η τεχνολογία αυτή μπορεί να φανεί χρήσιμη για την κατάταξη πολύπλοκων ασθενειών, όπως η διπολική διαταραχή, παρέχοντας χρήσιμες πληροφορίες για το γενετικό τους υπόβαθρο. Η διπολική διαταραχή είναι μια κληρονομική διανοητική νόσος η οποία χαρακτηρίζεται από επαναλαμβανόμενα επεισόδια μανίας και κατάθλιψης τα οποία πηγάζουν από πολλούς γενετικούς και περιβαλλοντολογικούς παράγοντες. Υπάρχουν τέσσερις βασικοί τύποι διπολικής διαταραχής: διπολική διαταραχή I, διπολική διαταραχή II, διπολική διαταραχή που δε μπορεί να καθοριστεί και κυκλοθυμία. Η κατάταξη των διπολικών διαταραχών, μπορεί να επιδράσει στην κατανόηση της παθολογίας της ασθένειας. Με τον τρόπο αυτό παρέχονται σημαντικές δυνατότητες στην έρευνα της αλληλεπίδρασης μεταξύ γενετικών και περιβαλλοντολογικών παραγόντων, που σχετίζονται με την παθογένεια. Επίσης, αυτή η δυνατότητα μπορεί να φανεί απαραίτητη για την καθοδήγηση της σωστής πρόγνωσης και θεραπείας για την πετυχημένη αντιμετώπιση της νόσου. Στην εργασία αυτή προσπαθήσαμε να εξάγουμε μια γονιδιακή υπογραφή για την οποία υπάρχει ήδη βιολογική γνώση για την σημαντικότητα τους η να παρουσιάσουμε μια νέα γονιδιακή υπογραφή, η οποία μπορεί να χρησιμοποιηθεί για περαιτέρω ανάλυση. Βάση της γονιδιακής υπογραφής προσπαθούμε να κατατάξουμε τις διπολικές διαταραχές χρησιμοποιώντας τις εκφράσεις γονιδίων δύο διαφορετικών πληθυσμών.

Η ανάλυση μικροσυστοιχειών συνήθως οδηγεί σε σύνολα δεδομένων που περιέχουν ένα μικρό αριθμό δειγμάτων με έναν πολύ μεγάλο αριθμό γονιδίων. Αρχικά, για να προκύψουν χρήσιμα πληροφορικά σύνολα γονιδίων, τα όποια είναι ικανά να μειώσουν την διαστατικότητα των συνόλων δεδομένων και να μεγιστοποιήσουν την απόδοση των ταξινομητών, χρησιμοποιείται μια μέθοδος φιλτραρίσματος. Στη συνέχεια, σημαντικός στόχος της παρούσας εργασίας είναι η εξαγωγή σταθερών αποτελεσμάτων των μεθόδων φιλτραρίσματος και των ταξινομητών. Για το λόγο αυτό, χρησιμοποιείται ένα πλαίσιο που ονομάζεται "Stable Bootstrap Validation" (SBV), το οποίο έχει παρουσιαστεί από τον Νίκο Χλή στη διπλωματική του εργασία. Το πλαίσιο SBV χρησιμοποιεί bootstrap αναδειγματοληψία του αρχικού συνόλου παράλληλα με ένα κριτήριο το οποίο καθορίζει την σταθερότητα της παρατηρούμενης απόδοσης του ταξινομητή και τη βιολογική ερμηνεία των γονιδίων, γνωστή στη βιβλιογραφία ως 'γονιδιακή υπογραφή'. Επίσης, παρουσιάζονται μεθοδολογίες που αφορούν την διαφοροποίηση, τη συνοχή καθώς και την ικανότητα γενίκευσης της γονιδιακής υπογραφής. Τέλος, στην παρούσα διπλωματική εργασία εξάγεται μια γονιδιακή υπογραφή 32 κοινών γονιδίων, η οποία συνδέεται στενά με πολλές πτυχές της διπολικής διαταραχής.

# Acknowledgements

I would like to thank my thesis supervisor, Professor Michalis Zervakis, for his guidance, support, constructive remarks, devoted time, as well as for giving me the opportunity to expand my knowledge in the field of bioinformatics. Moreover, I would like to thank Dr. Katerina Bei for her support and biological insight. I would also like to thank M.Sc. Stelios Sfakianakis and Nikolaos Chlis for sharing his knowledge. Furthermore I would like to thank Professors Costas Balas and Euripides Petrakis for their contribution and their participation in my Diploma Thesis committee. Last but not least, I would like to thank, not only my friends; Stauroula, Sofia and Marietta but also my family, my father Nikos, my mother Bayia ,my brother Jim as well as Vasilis for their love, support and encouragement.

# Table of Contents

List of Figures .....	8
List of Tables .....	12
Introduction.....	13
1.1 Introduction to Genome Analysis .....	13
1.2 Bipolar disorder .....	16
1.3 The Human Genome .....	19
1.4 Related Work .....	23
1.5 Thesis Outline and Innovation .....	25
Theoretical Background.....	26
2.1 Machine Learning and Pattern Recognition.....	26
2.1.1 Patterns –Classes – Features.....	27
2.1.2 Implementation of pattern recognition .....	28
2.2 Dataset (general) .....	29
2.3 Feature Subset Selection (FSS).....	29
2.3.1 Filter methods .....	29
2.3.2 Wrapper methods .....	30
2.3.3 Embedded methods.....	31
2.4 Classification .....	32
2.4.1 Classification Analysis.....	32
2.4.2 Classifiers .....	32
2.5 Classification Methods.....	33
2.5.1 Regularized Least Squares Classifiers .....	34

2.5.2 Support Vector Machine (SVM) .....	36
2.5.3 Relevance Vector Machine (RVM).....	40
2.6 Evaluation methods .....	41
2.6.1 Holdout Validation .....	41
2.6.2 K-Fold Cross Validation (K-Fold CV) .....	42
2.6.3 Leave One Out Cross Validation (LOOCV).....	43
2.6.4 Repeated Random Sub-Sampling Validation .....	43
2.6.5 Bootstrap Resampling Validation .....	44
2.7 Weak Law of Large Number (LLN).....	45
Methodology.....	46
3.1 Processing the dataset: SAM .....	48
3.2 Stable Bootstrap Validation .....	52
3.3 Evaluation of the Results .....	54
3.3.1 Evaluation of Discrimination of Genomic Signature .....	54
3.3.2 Consistency Evaluation of gene selection in the signature.....	55
3.3.3 Evaluation of Generalization Ability of Genomic Signature.....	56
Results.....	58
4.1 Original dataset .....	58
4.2 Processing the Dataset Results .....	59
4.2.1 SAM Parameters .....	59
4.3 SBV Results .....	65
4.3.1 RFE and LASSO parameters .....	66
4.3.2 Classifier Results .....	68

4.4 Evaluation Results .....	77
4.4.1 Classification Accuracy Comparison .....	77
4.4.2 Genomic Signature Significance .....	78
4.5. Biological Evaluation .....	84
Conclusion .....	89
References .....	92

## List of Figures

Figure 1.1: Overview of the proposed framework.....	15
Figure 1.2: Illustration of a cell, its nucleus, a chromosome and the double-helix DNA. ....	20
Figure 1.3: Overview of gene expression profiling. Messenger RNA is isolated from tissues or cells and copied, labeled, and hybridized onto microarrays, which are subsequently scanned by a confocal microscope. Computational methods are subsequently used to interpret the resulting image. ....	22
Figure 2.1: Pattern recognition process. (a) Set of class C, (b) set of pattern/samples P, (c) set of set of features/genes F.....	28
Figure 2.2: Filter Subset Selection Methods .....	30
Figure 2.3: (a) Linear classifier, (b) Non linear classifier .....	33
Figure 2.4: (a) Random data points, (b) Their linear regression, (c) Error for the pair $(x_i, y_i)$ : $e_i = y_i - wx_i$ .....	34
Figure 2.5: Binary classification. Samples on the margin are called the support vectors	37
Figure 2.6: Maximum - margin hyperplane and margins of a linear SVM. ....	40
Figure 2.7: Holdout validation method .....	42
Figure 2.8: K-Fold Cross Validation method .....	42
Figure 2.9: Leave One Out validation method.....	43
Figure 2.10: Repeated random sub-sampling validation method.....	44
Figure 2.11: Bootstrap resampling validation.....	45
Figure 3.1: Overview of the proposed methodology .....	47
Figure 3.2: Assign experiments to two groups (1,2) .....	48
Figure 3.3: (a) original grouping, (b) randomized grouping .....	49
Figure 3.4: Highlighting and invoking SAM .....	50
Figure 3.5: The SAM Dialog Box .....	50

Figure 3.6: The SAM Plot Controller on the front side, The SAM Plot sheet on the second side .....	51
Figure 3.7: Flowchart for Preprocessing the Dataset - SAM .....	52
Figure 3.8: Overview of Stable Bootstrap Validation approach .....	53
Figure 3.9: Flowchart corresponding to one iteration of the 10 - fold Cross Validation methodology. ....	56
Figure 3.10: Structure of the overall proposed methodology .....	57
Figure 4.1: Structure of the Original Dataset.....	60
Figure 4.2: The SAM Plot sheet of 1393 significant genes (Healthy controls – All bipolar disorder patients) .....	61
Figure 4.3: The SAM Plot sheet of 360 significant genes (Healthy controls – All bipolar disorder patients) .....	61
Figure 4.4: The SAM Plot sheet of 223 significant genes (Healthy controls – First episode bipolar disorder patients) .....	62
Figure 4.5: Structure of group 1: significant genes from healthy controls and medicated bipolar patient, group 2: significant genes from healthy controls and all bipolar patients and group 3: significant genes from healthy controls and first episode bipolar patients. ....	63
Figure 4.6: Left: Structure of significant genes from healthy controls and medicated bipolar patient (fold change=1.06). Right: Structure of genomic signature from healthy controls and all bipolar patients.....	64
Figure 4.7: Structure of the bootstrap datasets used in the first significant set (673genes). ....	65
Figure 4.8: Structure of the bootstrap datasets used in the second significant set (360genes). ....	66
Figure 4.9: Left: Stabilization of LASSO mean accuracy over all bootstrap datasets Right: Stabilization of LASSO mean signature size over all bootstrap datasets .....	67
Figure 4.10: Structure of the SAM as well as LASSO results .....	68
Figure 4.11: Structure of the SVM results .....	69

Figure 4.12: Left: Stabilization of SVM mean accuracy of 6 significant genes over all bootstrap datasets Right: Stabilization of SVM mean signature size of 6 significant genes over all bootstrap datasets .....	70
Figure 4.13: Left: Stabilization of SVM mean accuracy of 8 significant genes over all bootstrap datasets Right: Stabilization of SVM mean signature size of 8 significant genes over all bootstrap datasets .....	70
Figure 4.14 Structure of the RVM results - 23 common genes .....	71
Figure 4.15: Left: Stabilization of SVM mean accuracy of 78 significant genes over all bootstrap datasets Right: Stabilization of SVM mean signature size of 78 significant genes over all bootstrap datasets .....	72
Figure 4.16: Left: Stabilization of SVM mean accuracy of 73 significant genes over all bootstrap datasets Right: Stabilization of SVM mean signature size of 73 significant genes over all bootstrap dataset.....	72
Figure 4.17: Structure of the RVM results - 8 common genes .....	73
Figure 4.18: Left: Stabilization of SVM mean accuracy of 79 significant genes over all bootstrap datasets Right: Stabilization of SVM mean signature size of 79 significant genes over all bootstrap datasets .....	74
Figure 4.19: Left: Stabilization of SVM mean accuracy of 82 significant genes over all bootstrap datasets Right: Stabilization of SVM mean signature size of 82 significant genes over all bootstrap dataset.....	74
Figure 4.20: Structure of the RVM results - 1 common gene.....	75
Figure 4.21: Left: Stabilization of SVM mean accuracy of 78 significant genes over all bootstrap datasets Right: Stabilization of SVM mean signature size of 78 significant genes over all bootstrap datasets .....	76
Figure 4.22: Left: Stabilization of SVM mean accuracy of 82 significant genes over all bootstrap datasets Right: Stabilization of SVM mean signature size of 82 significant genes over all bootstrap dataset.....	76
Figure 4.23: Mean – Standard Deviation of 25 healthy control samples of 32 genes .....	79
Figure 4.24: Mean – Standard Deviation of 28 bipolar samples of 32 genes .....	79

Figure 4.25: Mean – Standard Deviation of 53 samples of 32 genes ..... 79

Figure 4.26: Frequencies of 32 significant genes..... 82

Figure 4.27: Standard Deviation of 7 multiply genes of the new dataset 83

# List of Tables

Table 2.1: Feature Subset Selection Methods .....	31
Table 4.1: SAM results from Healthy Control – Medicated Bipolar Disorder patients .....	60
Table 4.2: SAM results from Healthy Control – All Bipolar Disorder patients .....	61
Table 4.3: SAM results from Healthy Control – First Episode Bipolar Disorder patients .....	61
Table 4.4: SAM results from Healthy Control – All Bipolar Disorder patients removing one by one First Episode patient .....	62
Table 4.5: SBV results of LASSO. ....	67
Table 4.6: SBV results of SVM classifier for 673 significant genes. ....	70
Table 4.7: SBV results of SVM classifier for 360 significant genes. ....	70
Table 4.8: SBV results of RVM classifier from 673 significant genes.....	72
Table 4.9: SBV results of RVM classifier from 360 significant genes.....	72
Table 4.10: SBV results of RVM classifier from 650 significant genes.....	74
Table 4.11: SBV results of RVM classifier from 337 significant genes.....	74
Table 4.12: SBV results of RVM classifier from 642 significant genes.....	76
Table 4.13: SBV results of RVM classifier from 329 significant genes.....	76
Table 4.14a: Synopsis of SBV results from Healthy Control (25 HC) – Medicated Bipolar Disorder (25 MBD) samples.....	77
Table 4.14b: Synopsis of SBV results from Healthy Control (25 HC) –Bipolar Disorder (28 BD) samples.....	77
Table 4.15: Variance of 32 significant genes. ....	80
Table 4.16: Consistency of Gene Selection in the Signature .....	82
Table 4.17: Generalization Ability of Genomic Signature Results .....	84
Table 4.18: Mapping of Probe Set IDs to Gene Symbols and Entrez Gene IDs. Red highlighted are the eight genes with the highest variance among the groups. Purple highlighted is the gene NOG known for its association with BPD.....	85
Table 4.19: Enriched pathways by GATHER.....	86
Table 4.20: Enriched biological processes by GATHER.....	87

# 1

## Introduction

### 1.1 Introduction to Genome Analysis

Bipolar disorder [1] is a common, heritable mental illness characterized by recurrent episodes of mania and depression. Genetic studies have suggested that bipolar disorder has a genetic component, meaning the disorder can run in families. In that manner, the need arises for measurement of different genes expression levels in order to provide useful information for the genetic background of the disease. Genomic analysis is the technique needed to determine and compare the genetic sequence. One genome technique is DNA microarrays which can measure the expression of thousands of genes to identify changes in expression between different biological states. Through genome analysis using DNA microarrays, scientists can observe patterns in the data that can lead to different expression profiles among distinct classes of interest. Thus, the need arises for identification of sets of genes that strongly differentiate their expression levels among classes of interest. Moreover, scientists have the opportunity to use these sets of genes along with the observed patterns in order to design classification methodologies that assign class labels to an independent dataset. Finally, the classification of bipolar disorders may be essential to guide appropriate therapy and determine prognosis for successful treatment.

However, the genome analysis usually leads to datasets that normally contain a small number of samples which have a large number of gene expression levels as features. That leads to the problem known as “curse of dimensionality”, which implies significant decrease in classification performance as well as in statistical significance. In this study, which constitutes a preliminary study, the original dataset consists of 53 samples related to bipolar disorder, 25 of which correspond to patients with bipolar disorder who had previously received medication, 3 patients with bipolar disorder who were experiencing

their first episode and had not previously received medication and 25 matched control samples. For each sample, there are measurements of 54675 genes. We must note that the dataset composes from a small number of samples, thus we propose that in the future more power calculation can be performed in order to assess the significance of the specimen. In order to extract useful informative sets of genes that can reduce dimensionality and maximize the performance of classifiers, feature selection methods were used. The aim of FFS is to reduce the number of genes by keeping the most relevant set, which are also called “significant set”. Feature selection methods can be separated into three categories: filter methods, which follow a univariate approach that examine one feature at a time, wrapper methods and embedded methods, which are multivariate approaches that simultaneously examine different sets of features. Univariate methods ignore the interaction with the classifier and each feature is considered separately, since they select features which differentiate their behavior between the classes of interest. On the other hand, multivariate methods aim at the incorporation of feature dependencies to some degree, selecting a set which maximizes the classification performance. In this study, the original dataset has undergone feature subset selection using a filter univariate method which is called “Significance Analysis of Microarrays” (SAM) [2],[3]. SAM uses a modified t-statistic and permutations of the repeated measurements of the data in order to decide if the gene expression is strongly related to the class label.

Another important aspect of microarray analysis is the problem of classification of new samples, which can lead to new prognosis methodologies. While, the feature selection methods are used in order to counterfeit the curse of dimensionality by keeping a relatively small set of significant features, the classification approaches are used in order to classify new data into known class of interest. Various classification approaches have been proposed for this purpose. In this study, the methods we considered were; Least Absolute Shrinkage and Selection Operator (LASSO), Support Vector Machine (SVM) and Relevance Vector Machines. Through classification approaches a small set of significant features, which achieves high classification accuracy, arises. These lists of significant genes are often called “genomic signatures” in literature.

Moreover, another aim of this study is to achieve stable performance assessment of feature selection and classification methods. As already mentioned, a wide variety of machine learning methods have been proposed for classification tasks related to microarrays, including support vector machines (SVM), relevance vector machines (RVM), K-Fold Cross Validation and many others. However, the use of an arbitrarily fixed combination of FSS method and classifier can lead to significant variations not only in the training or testing dataset but also in the set of features selected as well as

classification accuracy. Thus, may sacrifice performance that could have been achieved with another model. In that manner, the generic evaluation framework named “Stable Bootstrap Validation” (SBV), introduced by Nick Chlis [4], is presented. The SBV utilizes resampling of the original dataset and an explicit criterion that determines the stability of the observed classification accuracy and the biological interpretation of genes, also called genomic signature.

Another fundamental aspect of microarray analysis is the evaluation of the results extracted from feature selection as well as classification methods. The results that are stable and reflect the biological model should also be consistent across different executions of the feature selection and classification methodologies. This aspect is achieved through cross validation methodology, which splits the dataset in fold, in order to estimate how accurately the predictive model will perform in practice. Finally, another aspect of evaluation is the generalization ability of the observed results. This aspect is also addressed in our methodological framework through cross validation, determining how the results of a statistical analysis will generalize to an independent data set. The overview of the proposed framework is presented as a block diagram in figure 1.1.

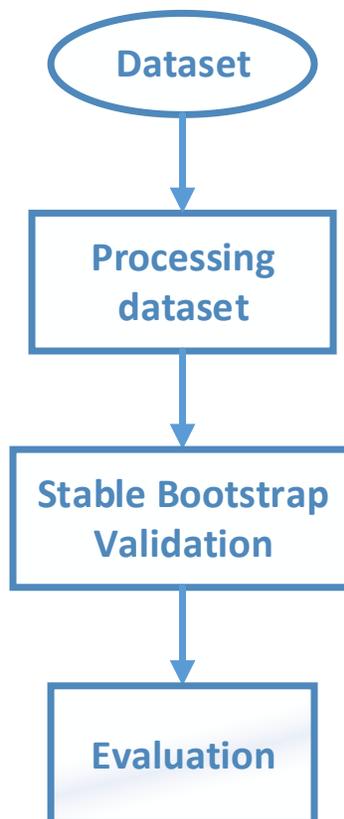


Figure 1.1: Overview of the proposed framework

## 1.2 Bipolar disorder

Bipolar disorder [1],[14], also known as manic-depressive illness, is a complex genetic disorder in which the core feature is pathological disturbance in mood ranging from extreme elation, or mania, to severe depression usually accompanied by disturbances in thinking and behavior. Symptoms of bipolar disorder are severe. They are different from the normal ups and downs that everyone goes through from time to time. Bipolar disorder symptoms can result in damaged relationships, poor job or school performance, and even suicide. But bipolar disorder can be treated, and people with this illness can lead full and productive lives. About 3% of people in the United States have bipolar disorder at some point in their life. Lower rates of around 1% are found in other countries. The most common age at which symptoms begin is 25. Rates appear to be similar in males as females.



### Types

There are four basic types of bipolar disorder:

1. *Bipolar I Disorder*: defined by manic or mixed episodes that last at least seven days, or by manic symptoms that are so severe that the person needs immediate hospital care. Usually, depressive episodes occur as well, typically lasting at least 2 weeks .
2. *Bipolar II Disorder*: defined by a pattern of depressive episodes and hypomanic episodes, but no full-blown manic or mixed episodes.
3. *Bipolar Disorder Not Otherwise Specified (BP-NOS)*: diagnosed when symptoms of the illness exist but do not meet diagnostic criteria for either bipolar I or II. However, the symptoms are clearly out of the person's normal range of behavior.
4. *Cyclothymic Disorder, or Cyclothymia*: a mild form of bipolar disorder. People with cyclothymia have episodes of hypomania as well as mild depression for at least 2 years. However, the symptoms do not meet the diagnostic requirements for any other type of bipolar disorder.

A severe form of the disorder is called *Rapid-cycling Bipolar Disorder*. Rapid cycling occurs when a person has four or more episodes of major depression, mania,

hypomania, or mixed states, all within a year. Rapid cycling seems to be more common in people who have their first bipolar episode at a younger age.

## Causes

The causes of bipolar disorder [15],[16] likely vary between individuals and the exact mechanism underlying the disorder remains unclear. Genetic influences are believed to account for 60–80% of the risk of developing the disorder indicating a strong hereditary component.

### Genetic

Genetic studies have suggested that many chromosomal regions and candidate genes are related to bipolar disorder susceptibility with each gene exerting a mild to moderate effect. The risk of bipolar disorder is nearly ten-fold higher in first degree-relatives of those affected with bipolar disorder when compared to the general population; similarly, the risk of major depressive disorder is three times higher in relatives of those with bipolar disorder when compared to the general population.

### Environmental

Evidence suggests that environmental factors play a significant role in the development and course of bipolar disorder and those individual psychosocial variables may interact with genetic dispositions. There is fairly consistent evidence from prospective studies that recent life events and interpersonal relationships contribute to the likelihood of onsets and recurrences of bipolar mood episodes, as they do for onsets and recurrences of unipolar depression. There have been repeated findings that 30–50% of adults diagnosed with bipolar disorder report traumatic/abusive experiences in childhood, which is associated on average with earlier onset, a higher rate of suicide attempts, and more co-occurring disorders such as PTSD.

### Physiological

Abnormalities in the structure and/or function of certain brain circuits could underlie bipolar. Functional magnetic resonance imaging findings suggest that abnormal modulation between ventral prefrontal and limbic regions, especially the amygdala, are likely contribute to poor emotional regulation and mood symptoms.

### Neurological

Less commonly bipolar disorder or a bipolar-like disorder may occur as a result of or in association with a neurological condition or injury.

### Evolutionary

Because bipolar disorder affects an individual's ability to function in society and has a high morbidity rate, evolutionary theory would suggest that the genes responsible would have been naturally selected against, effectively culling the disorder. Yet there continue to be high rates of bipolar disorder in many populations, suggesting the genes responsible may have an evolutionary benefit.

There are currently no biological tests that differentiate patients with bipolar disorder (BPD) from healthy controls. While there is evidence that peripheral gene expression differences between patients and controls can be utilized as biomarkers for psychiatric illness, it is unclear whether current use or residual effects of antipsychotic and mood stabilizer medication drives much of the differential transcription. We therefore tested whether expression changes in first-episode, never-medicated bipolar patients, can contribute to a biological classifier that is less influenced by medication and could potentially form a practicable biomarker assay for bipolar disorder.

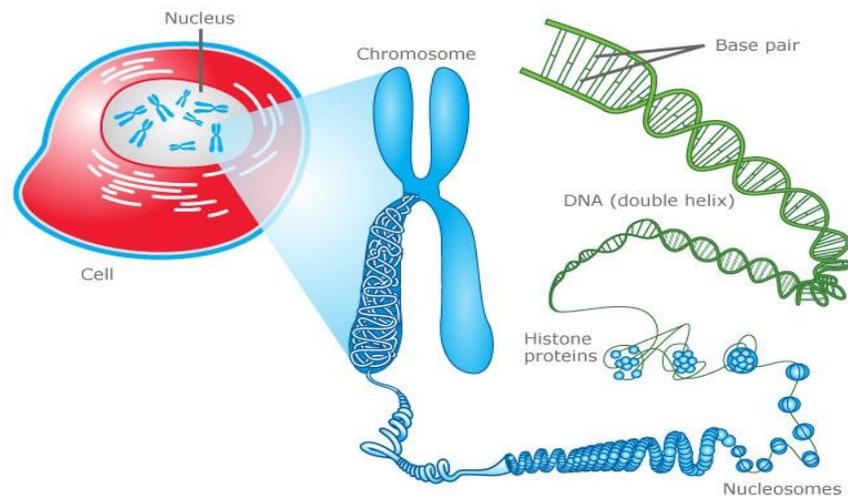
## 1.3 The Human Genome

The human genome refers to the complete set of human genetic information, the study, analysis and mapping of which, has been the subject of international scientific research project the “Human Genome Project”[5]. The project was proposed and funded by the US government; planning started in 1984, got underway in 1990, and was declared complete in 2003. The human genome is the complete set of nucleic acid sequence for humans, encoded as DNA within the 23 chromosome pairs in the nucleus of human cells and in a small DNA molecule found within individual mitochondria. Human cells have 23 pairs of chromosome, 22 pairs of autosomes and one pair of sex chromosomes, giving a total of 46 per cell. Each chromosome can be thought as a string of thousands of genes, which are in turn made of DNA. There are an estimated 20,000-25,000 human genes, most of them located in the nucleus, while only 37 refer to mitochondrial genes. The DNA which makes up the genes is called coding DNA, while the DNA “string” between each gene is called non-coding DNA. Coding DNA, which occupies only a small fraction of the genome (<2%), is defined as those sequences that can be transcribed into mRNA and translated into proteins during the human life cycle. On the other hand non-coding DNA is made up of all of those sequences (~ 98% of the genome) that are not used to encode proteins. The study of the human genome led to the genomic revolution since the notification of the first draft sequence of the genome had a huge impact on human disease research.

### DNA

As we already mentioned, each gene is made of DNA [6]. Deoxyribonucleic acid (DNA) is a molecule that carries most of the genetic instructions used in the development, functioning and reproduction of all known living organisms as well as many viruses. DNA is made of chemical building blocks called nucleotides and consists of two long complementary strands of nucleotides that take the form of a double stranded helix. This shape gives DNA the power to pass along biological instructions with great precision. The nucleotides are made of three parts: a phosphate group, a sugar group and one of four types of nitrogen bases. To form a strand of DNA, nucleotides are linked into chains, with the phosphate and sugar groups alternating. The four types of nitrogen bases found in nucleotides are: adenine (A), thymine (T), guanine (G) and cytosine (C). The sequence of these bases determines what biological instructions are contained in a strand of DNA. Each nucleotide of a strand is made up of two nitrogen bases, paired together by hydrogen bonds. Because of the highly specific nature of this type of

chemical pairing, base A always pairs with base T, and likewise C with G. So if the sequence of the bases on one strand of a DNA double helix is known, it is a simple matter to figure out the sequence of bases on the other strand. DNA's unique structure enables the molecule to copy itself during cell division. The discovery that DNA contains the code for life, gave impetus to a global effort to understand how the genome sequences of many organisms associated with their health.



**Figure 1.2:** Illustration of a cell, its nucleus, a chromosome and the double-helix DNA.

Source: 2011, the university of Waikato, [www.sciencelearn.org.nz](http://www.sciencelearn.org.nz)

## RNA

RNA stands for ribonucleic acid. It is an important molecule with long chains of nucleotides. As already mentioned DNA is defined as a nucleic acid that contains the genetic instructions used in the development and functioning of all known living organisms. RNA molecules are involved in protein synthesis and sometimes in the transmission of genetic information. Like DNA, RNA is assembled as a chain of nucleotides, but contrary to DNA is found not as a double-strand but as a single-strand folded on to itself. There are different types of RNA named according to the biological process in which they participate. First a type of RNA called messenger RNA (mRNA) carries information from DNA to structures called ribosomes. These ribosomes are made from proteins and ribosomal RNAs (rRNAs). The coding sequence of the mRNA determines the amino acid sequence in the protein that is produced. However, many RNAs do not code for protein. The most prominent examples of these non-coding RNAs are transfer RNA (tRNA) and ribosomal RNA (rRNA), both of which are involved in the

process of translation. Also RNA can act as enzymes (called ribozymes) to speed chemical reactions.

## **Genes**

Genes are subunits of DNA, the information database of a cell that is contained inside the cell nucleus. This DNA carries the genetic blueprint that is used to make all the proteins the cell needs. Every gene contains a particular set of instructions that code for a specific protein.

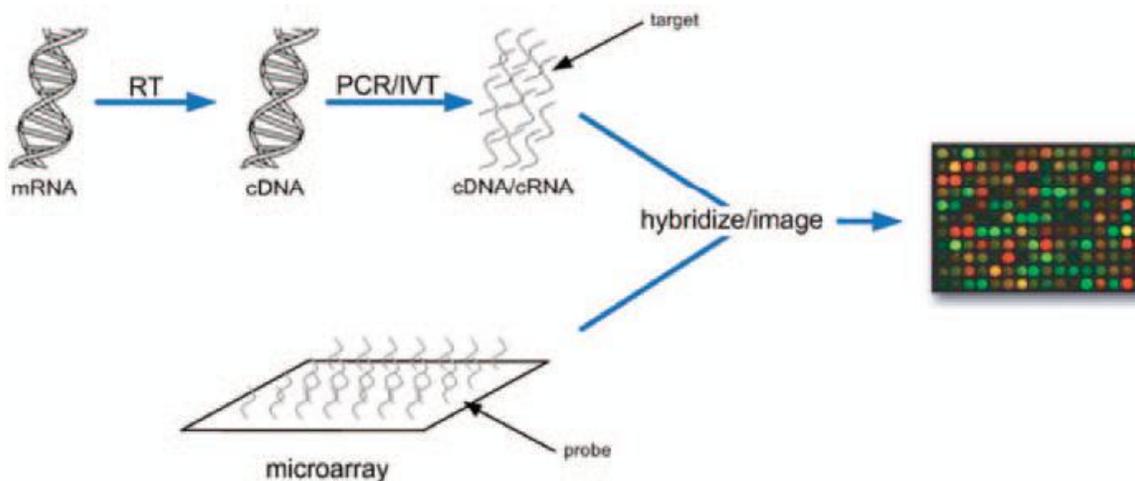
## **Gene Expression –DNA Transcription – DNA Translation**

Gene expression [7] is the process by which the genetic code, the nucleotide sequence, of a gene is used to direct protein synthesis and produce the structures of the cell. Genes that code for amino acid sequences are known as “structural genes”. The process of gene expression involves two main stages; transcription and translation. Transcription is the first step of gene expression and refers to the production of messenger RNA (mRNA) by the enzyme RNA polymerase, and the processing of the resulting mRNA molecule. On the other hand, translation is the use of mRNA to direct protein synthesis, and the subsequent post-translational processing of the protein molecule. Some genes are responsible for the production of other forms of RNA that play a role in translation, including transfer RNA (tRNA) and ribosomal RNA (rRNA).

## **DNA Microarray Analysis**

As already mentioned, the human genome contains approximately 21,000 genes. DNA Microarray technology [8] enables the researchers to investigate and address issues which were once thought to be non traceable. One can analyze the expression of thousand of genes in a single reaction quickly and in an efficient manner. DNA Microarray technology has empowered the scientific community to understand the fundamental aspects underlining the growth and development of life as well as to explore the genetic causes of anomalies occurring in the functioning of the human body. It is common knowledge that a mutation, or alteration, in a particular gene's DNA may contribute to a certain disease; there was an eager need for the development of a test that can trace these mutations. DNA Microarrays are tools that allow the measurement of the expression levels of different genes. A gene is considered to be expressed if it's DNA has been transcribed to RNA and gene expression refers to the level of transcription of the gene's DNA. Thousands of spotted samples known as probes are immobilized on a solid surfuse. The spots can be DNA, cDNA, or oligonucleotides. DNA

microarrays measure the levels of mRNA. DNA microarrays measure gene expression assessing the levels of mRNA present in the samples of interest indirectly. The assessment is indirect since DNA microarrays in reality measure the levels of cDNA, which derived from mRNA using a process called Reverse Transcription (RT). The sample has genes from both the normal as well as the diseased tissues. Spots with more intensity are obtained for diseased tissue gene if the gene is over expressed in the diseased condition. This expression pattern is then compared to the expression pattern of a gene responsible for a disease. Different types of microarray are in current use; they can be categorized by how the DNA probes are immobilized on the slide: the *in situ* synthesized Affymetrix GeneChips which utilizes photo-lithography for embedding cDNA probes on silicon chips, and the spotted cDNA (or oligonucleotide) microarrays developed at Stanford University which utilizes robotic spotting of aliquots of purified cDNA clones. In the recent past, microarray technology has been extensively used by the scientific community. Consequently, over the years, there has been a lot of generation of data related to gene expression. This data is scattered and is not easily available for public use. For easing the accessibility to this data, the National Center for Biotechnology Information (NCBI) has formulated the Gene Expression Omnibus or GEO. It is a data repository facility which includes data on gene expression from varied sources.



**Figure 1.3:** Overview of gene expression profiling. Messenger RNA is isolated from tissues or cells and copied, labeled, and hybridized onto microarrays, which are subsequently scanned by a confocal microscope. Computational methods are subsequently used to interpret the resulting image.

Source: Albert Hsiao et al., "High-throughput Biology in the Postgenomic Era", *J Vasc Interv Radiol* 2006; 17:1077–1085

## 1.4 Related Work

In the field of Bioinformatics, several studies focus on the genetic data analysis for the classification of complex diseases, such as bipolar disorder, based on their gene expression signatures, in order to provide useful information for the genetic background. Concerning the genetic analysis of bipolar disorder Nick Craddock et al. in [40] uses molecular genetic positional and candidate gene approaches for the genetic dissection of bipolar disorder. Moreover, Peter Holmans et al. in [41] has proposed a methodology for testing overrepresentation of biological pathways, indexed by gene-ontology terms, in lists of significant SNPs from genome-wide association studies. The method was applied to a meta-analysis of bipolar disorder, and it implicated the modulation of transcription and cellular activity, including that occurs via hormonal action, as an important player in pathogenesis.

Concerning the classification of the data, several classification approaches are used and mixed with feature selection algorithms in order to extract reliable sets of genes that can maximize the performance of classifiers. Georges Natsoulis et al. in [9] has proposed a methodology that aims to derive useful biological knowledge and readily interpretable drug signatures with high classification performance from a large database, using a variety of supervised classification algorithms, such as Support Vector Machines (SVMs) and Logistic Regression. Also, this approach proves that the combination of the results of these algorithms with feature selection techniques further reduce the length of the drug signatures. Osareh et al. in [10] have proposed a methodology that aims to develop an automated system for robust and reliable cancer diagnoses based on gene microarray data. They have presented a classification model which utilizes a subset of features chosen via information gain feature ranking for support vector machine classifier. Michael P. S. Brown et al. in [11] introduce a new method of functionally classifying genes using gene expression data from DNA microarray hybridization experiments. The method is based on the theory of support vector machines (SVMs). Moreover, several approaches have been proposed to evaluate the stability and reliability of results from feature selection and classification approach. Many studies focus on random sampling or splitting of the original dataset in order to infer stable performance estimates. Davis et al. in [12] notice that after a sufficiently large number of datasets have been generated by random splitting of the original dataset and are used to extract performance estimates, the average value of the classification accuracy tends to stabilize. The framework introduced by Armapanzas et al. in [13] requires an arbitrary number of 1000 bootstrap iterations followed by univariate filtering and training a k-Dependence Bayesian classifier, in order to result in a stable set of genes

selected in the model. Finally, Nick Chlis et al. in [4] have proposed a methodology that utilizes a formal criterion in order to extract robust estimates for the size of genomic signature as well as the classification accuracy and no further iterations are required. The stable estimates can be reproduced resulting in minimal variations during independent executions of the evaluation method. Our study is based on Nick Chlis [4] diploma thesis.

## 1.5 Thesis Outline and Innovation

The biological background concerning the human genome and biological concepts regarding the DNA microarrays is covered in chapter 1. While, the theoretical background concerning methodologies for the analysis of DNA microarray data in the field of bioinformatics is covered in chapter 2. This chapter includes feature selection and classification methods, while evaluation methods and the statistics theorem known as the “law of large numbers” are also presented. The proposed methodology for performing reliable feature selection and stable classification accuracy is presented in chapter 3. This chapter also includes the methodology for evaluating the discrimination, consistency and generalization ability of the observed results. Finally, in chapter 4 the results of the proposed methodology are presented, followed by a biological evaluation of the extracted signatures.

The innovative concept of this thesis involves the combination of the univariate filter method “Significance Analysis of Microarray” and a variety of supervised classification algorithms, such as Support Vector Machines (SVMs) and Relevance Vector Machines (RVMs) in order to derive useful biological knowledge for the bipolar disorder. Moreover, unlike similar studies, this thesis aims at testing whether expression changes in first – episode, never medicated bipolar patients, can contribute to a genomic signature that is less influenced by medication.

# 2

## Theoretical Background

In this chapter an introduction to the machine learning and pattern recognition as well as the implementations of pattern recognition are described in section 2.1, followed by a general presentation of the dataset in section 2.2. Then, feature subset selection methods are presented in section 2.3 including filter, wrapper and embedded methods. Moreover, the classification analysis and an introduction of classifiers are converted in section 2.4, including linear and non linear classifiers. In section 2.5 classification methods are examined in detail including regularized least squares, support vector machines as well as relevance vector machines. Finally different evaluation methods are described in section 2.6, such as holdout validation, k-fold cross validation, leave one out cross validation, repeated random sub-sampling validation and bootstrap resampling.

### 2.1 Machine Learning and Pattern Recognition

In machine learning, pattern recognition [17],[18],[19] is the process of discovering patterns and regularities in large amounts of data. There are three different approaches to pattern recognition, depending on machine learning: supervised learning, unsupervised learning and reinforcement learning.

- **Supervised learning**

The goal of supervised learning is to build a concise model of labeled samples. This set of labeled samples is called the training set. The resulting model is then used to assign class labels to the testing data where the value of the class label is unknown. Cases, in which the desired output is a continuous variable, are called

regression algorithms, while if the output falls within discrete values the task is called classification.

- **Unsupervised learning**

The aim of unsupervised learning is to discover groups in unlabeled data with similar attributes. This differentiates unsupervised learning from supervised learning and reinforcement learning.

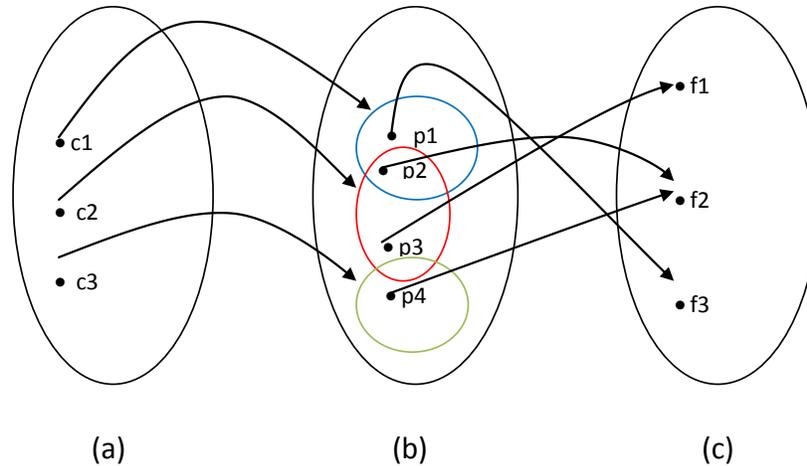
- **Reinforcement learning**

Finally, reinforcement learning is learning by interacting with an environment through a process of trial and error. The reinforcement learning agent receives a reinforcement signal, which constitutes a measure of how well the system operates, and tries to select actions that maximize the cumulative reward over time.

### **2.1.1 Patterns –Classes – Features**

DNA microarray analysis falls within supervised learning. In machine learning and pattern recognition [20], *patterns* are “physical” representation of the objects and we usually refer to them as objects, cases or samples. *Class or class label* is a set of patterns sharing common attributes and usually originated from the same source. *Features* are measurements or attributes derive from the patterns, which may be useful for their characterization. Features are numeric and usually the initial set of raw features is too large to be handled.

Pattern recognition can be also characterized as an information mapping process. There is a set of class  $C$  in which can be found a certain studied entity. Correspond to each class is a certain set of representation  $P$ , the patterns. Each class can be illustrated by a subset in the set of patterns. These subsets may overlap each other, allowing patterns of different classes to share same characteristics. Moreover, each pattern can be illustrated in the set of features  $F$ . Thus, each feature can be a member not only of different patterns but also different classes, as outlined in figure.



**Figure 2.1:** Pattern recognition process.  
 (a) Set of class C, (b) set of pattern/samples P, (c) set of set of features/genes F

## 2.1.2 Implementation of pattern recognition

As we mentioned above in machine learning, pattern recognition is the assignment of a label to a given input value. An example of pattern recognition is classification. However, it is a more general problem which encompasses other types of output as well. Other example is recognition, which assigns a real – valued output to each input and is presented in detail in section 2.4.

A classification [21] problem exists when an observation needs to be assigned into a class based on a number of features related to that sample. During classification given samples are assigns to prescribed categories. The algorithm that implements classification is known as a classifier and maps input data to class which performs classification. Particularly, a classifier is experienced on training data, adjusting his parameter to them and learns to recognize specific patterns. The result of classification process is based on the most significant characteristic of the classifier which is the ability of generalization. Generalization is the ability of a classifier to perform accurately on new, unseen data after having experienced a training data set. The classifier is designed effectively when he is able to correctly combine the characteristics of a sample in order to determine in which class it belongs. The best way to measure the generalization ability of a full trained classifier is to use a test data set which contains data that does not belong to the training set. Classification methods are also presented in detail in sections 2.4 and 2.5.

## 2.2 Dataset (general)

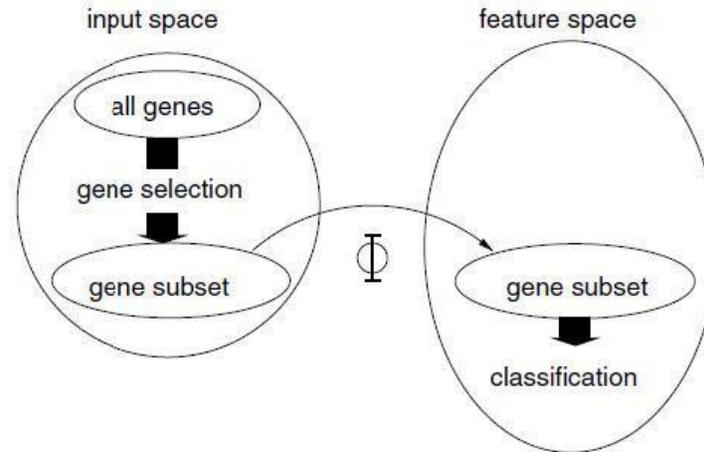
In this study, the data is composed of a set of **N samples/patterns**, where each sample contains the expression value of **K genes/features**. The dataset is expressed in array form as  $\mathbf{x} \in \mathbb{R}^{N \times K}$ , while the class labels of all samples are represented by a vector  $\mathbf{y} \in \mathbb{R}^N$ . Also, in the dataset, each sample  $N$  can be expressed as a vector  $\mathbf{x}_i \in \mathbb{R}^K$  where  $i = 1, \dots, N$ , while a class label  $\mathbf{y}$  is assigned to each of the samples. Particular, in the case of bipolar disorder/control binary classification  $\mathbf{y} \in \{-1, +1\}$ .

## 2.3 Feature Subset Selection (FSS)

Feature subset selection method (FSS) [22],[23],[24] is usually the crucial first step in microarray data analysis. DNA microarray data normally contains a small number of samples which have a large number of gene expression levels as features. The aim of FFS is to reduce the number of genes by keeping the most relevant set, which are also called "significant set", in order to extract useful information and reduce dimensionality. Then, this set of features is presented as input to the classification algorithm. Depending on how feature selection methods combine the feature selection search with the construction of the classification model can be separated into three categories: filter methods, wrapper methods and embedded methods.

### 2.3.1 Filter methods

Filter methods [22],[23],[24] evaluate the relevant set of features by looking only at the essential properties of the data. They calculate a feature relevance score and low-scoring features are removed. This subset of "significant" features is then used for classification. The advantages of these methods are that they are independent of the classification algorithm and they are computationally efficient. However, they are often *univariate* which means that they ignore the interaction with the classifier and each feature is considered separately. Thus, a number of *multivariate* methods are used, aiming at the incorporation of feature dependencies to some degree. *Univariate* filter techniques can be divided into two categories: parametric and model-free methods. In parametric methods the data is drawn from a given probability distribution while in model-free methods (or non parametric) the data may not follow a normal distribution. In microarray studies the most widely used techniques are t-test and ANOVA.



**Figure 2.2:** Filter Subset Selection Methods  
Source: [23]

### 2.3.1.1 Significance Analysis of Microarrays (SAM)

Significance Analysis of Microarrays (SAM) [2], [3] is a filter (univariate) method. It was proposed by Tusher, Tibshirani and Chu and the software was written by Balasubramanian Narasimhan and Robert Tibshirani. Particularly, SAM is a statistical technique, which identifies significant genes by assimilating a set of gene-specific  $t$  tests. Each gene is assigned a score on the basis of its change in gene expression relative to the standard deviation of repeated measurements for that gene. This analysis uses non-parametric statistics, since the data is drawn on the bases of some unknown distribution. Genes with scores greater than a threshold are potentially significant. SAM uses analyzing permutations of the repeated measurements to estimate the percentage of such genes identified by chance, the false discovery rate (FDR), which is calculated for each set. The threshold can be determined by a tuning parameter delta, chosen by the user based on the false positive rate. One can also choose a fold change parameter, to ensure that called genes change at least a pre-specified amount.

### 2.3.2 Wrapper methods

In comparison with filter methods, wrappers [22],[23],[24] have the ability to take into account feature dependencies, which means that they fall within multivariate approach. In wrapper methods, the feature subset selection algorithm exists as a wrapper around the classification models. They use the classifier itself as part of the function, evaluating

feature subsets according to their predictive power. The classifier is utilized as a black box. The feature subset with the highest evaluation is chosen as the final set on which to run the classifier. However, a common drawback of these techniques is that they have a risk of over fitting to the model and are very computationally intensive, since they need to evaluate different combinations of features.

## 2.3.3 Embedded methods

Embedded methods [22],[23],[24] are also a case of multivariate approach and use the classifier, evaluating feature subsets according to their predictive power. However, there are differences between embedded and wrapper methods. Particularly, embedded methods include the interaction with the classifier, while in wrappers the feature selection algorithm is independent of the classification model. Finally, embedded techniques are more computationally efficient than wrapper methods.

### 2.3.3.1 Recursive Feature Elimination (RFE)

Recursive Feature Elimination [25] is an embedded feature selection approach. The goal of RFE is to select features by recursively preserving smaller and smaller sets of features, maximizing the classification accuracy of a given classification method. The RFE eliminates a fixed number of least significant features and then reassessing the classification performance. That procedure is recursively repeated on the above set until the desired number of features to select is eventually reached. Then, the set of features across all iterations which maximizes the classification accuracy is chosen as the optimal feature set. The least significant feature is determined through a feature weighting scheme which can be the weight given to each feature by a linear classifier or by non-linear feature weighting methods.

Feature Subset Selection Methods		
<i>Univariate</i>	<i>Multivariate</i>	
<b>Filter Methods</b>	<b>Wrapper methods</b>	<b>Embedded methods</b>

**Table 2.1:** Feature Subset Selection Methods

## 2.4 Classification

### 2.4.1 Classification Analysis

Classification analysis [26] is one of the most crucial steps in machine learning and computer science. As we already mentioned the aim of classification is to find a rule, which, based on external observations, assigns a sample to one of several classes. Binary classification is the simplest case where the classifier categorizes the samples of given set into two different classes based on the aforementioned rule.

### 2.4.2 Classifiers

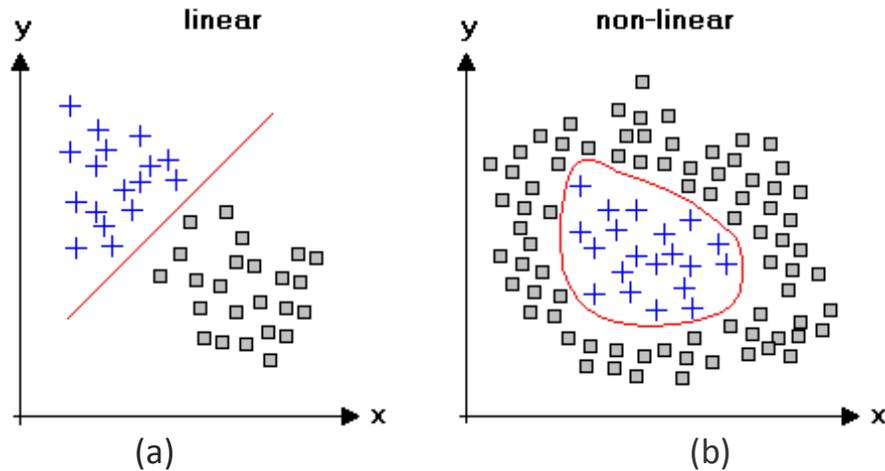
The algorithm that implements classification is known as a classifier. The main division of classifiers is to linear and nonlinear.

#### 2.4.2.1 Linear Classifiers

A linear classifier [27] can split two classes only when they are linearly separable. This means that there is a hyperplane which separates the data in both classes. The classification rule of a linear classifier is to assign a label  $\hat{y}$  to an unknown sample  $\hat{x}$  based on the formula  $\hat{y} = \mathbf{f}(\hat{x} \cdot \mathbf{w})$ , where  $\mathbf{w}$  is a real vector of weights and is produced during training process of the classifier. In this study, the linear classifiers that are examined are RLS methods like RR and the LASSO, linear SVM as well as RVM.

#### 2.4.2.2 Non – Linear Classifiers

While linear classifiers are simple and computationally efficient, for nonlinearly separable features, they might lead to very inaccurate decisions. Then simplicity and efficiency for accuracy are calculated through a nonlinear classifier [27]. An example of a nonlinear classifier is K Nearest Neighbor (K-NN) Classifier which classifies new samples depending on a set of samples closest to them, which are called their “nearest neighbors”.



**Figure 2.3:** (a) Linear classifier, (b) Non linear classifier  
Source: [27]

## 2.5 Classification Methods

### Linear Regression

Regression analysis is a statistical method for modeling the relationship between the observed and response variable of a system. The basic idea of linear regression [28], [29] is that, if there is a linear relationship between two variables, you can then use one variable to predict values on the other variable. Thus, given data set  $D$  of  $N$  samples of the form:

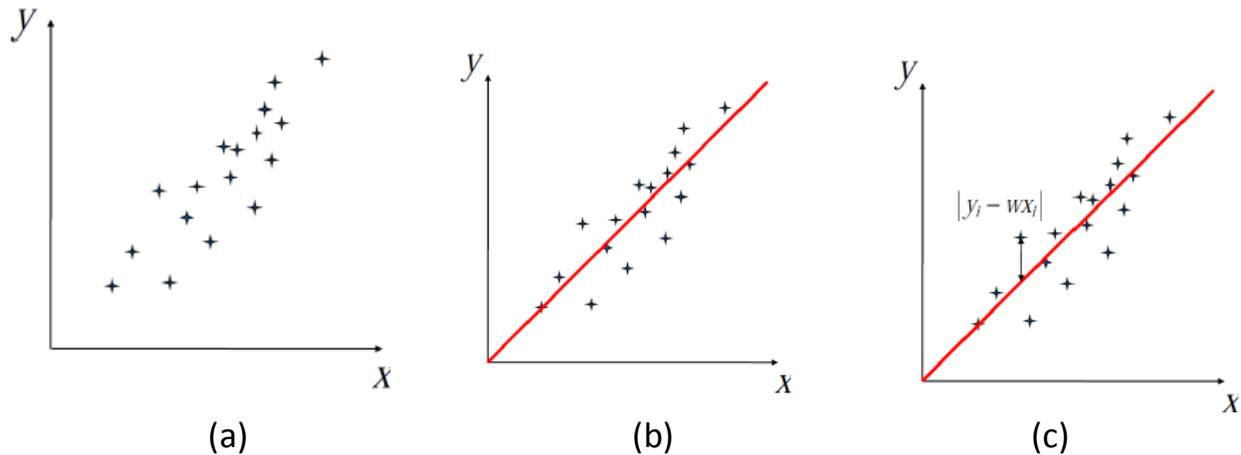
$$D = \{(x_i, y_i) \mid x_i \in \mathbb{R}^K, y_i \in \{-1, +1\}\}, i = 1, \dots, N$$

, the linear regression model assumes that the relationship between the response variables  $y_i$  and the observed variables  $x_i$  is linear. The aforementioned problem can be written in vector form as

$$\mathbf{y} = \mathbf{X} \cdot \mathbf{w} + \boldsymbol{\varepsilon}$$

In DNA microarray the observed variables are the expression values of  $K$  genes per sample represented as in matrix form  $\mathbf{X} \in \mathbb{R}^{N,K}$ , while the response variable  $\mathbf{y} \in \mathbb{R}^N$  is expressed as a vector of class labels of all samples (bipolar/control). The variable  $\varepsilon_i$  is

an error unobserved variable which adds random noise to the above linear relationship. Finally, the weight vector  $w \in \mathbb{R}^K$  is a vector of regression coefficients.



**Figure 2.4:** (a) Random data points, (b) Their linear regression, (c) Error for the pair  $(x_i, y_i)$ :  $e_i = y_i - wx_i$   
Source: [27]

## 2.5.1 Regularized Least Squares Classifiers

### 2.5.1.1 Ordinary Least Squares (OLS)

We already mentioned that the aim of regression is to describe the relationship between two variables with a line. Ordinary Least Squares (OLS) [30] (also known as “Least squares linear regression” or often just “least squares”) is a statistical method for finding the function which most closely approximates the data. Particular, it addresses to find the line which minimizes the total distance between the observed responses and the responses predicted by the linear approximation of the data. Given a training set  $X$  of  $N$  samples of the form:

$$X = \{(x_i, y_i) \mid x_i \in \mathbb{R}^K, y_i \in \{-1, +1\}\}, i = 1, \dots, N$$

, the goal of the ordinary least squares technique is to deduce a function that evaluate the labels  $\hat{y}$  of a new set of test samples  $\hat{X}$ . OLS regression assumes that there is a linear relationship between the two variables  $\hat{y} = \hat{X} \cdot w$ . According to the OLS formula,

the weight vector  $w$  is the one that minimizes the function and can be described by the equations:

$$w = \operatorname{argmin} F(w)$$

$$F(w) = \sum_{n=1}^N y_n - \hat{y}_n = \sum_{n=1}^N (y_n - x_n \cdot w)^2$$

### 2.5.1.2 Regularized Least Squares (RLS)

While Ordinary Least Squares [29] is one of the most basic prediction techniques which are able to give optimal estimates to the classification of new samples, is also known for often not performing well with respect to both prediction accuracy and model complexity. Particular, OLS can perform very badly when the number of variables in the linear system exceeds the number of observations, achieving low prediction accuracy. In such settings, Regularized Least Squares intends to use regularization to further constrain the resulting solution, improving the performance of the OLS approach. The aforementioned can be achieved by further restraining the weight vector  $w$ .

### 2.5.1.3 Ridge Regression (RR)

The Ridge Regression [29] is a continuous process which is a slight modification on the Ordinary Least Squares method and replaces the function  $F(w)$  by

$$F(w) = \sum_{n=1}^N (y_n - x_n \cdot w)^2, \text{ subject to } \sum_{k=1}^K w_k^2 < \tau$$

Here  $\tau \geq 0$  is a tuning parameter, which controls the strength of the penalty term and can be expressed as  $t = a \cdot \sum_{k=1}^K w_0^2$ ,  $a \in [0,1]$ . Thus, instead of  $t$ ,  $a$  needs to be estimated through cross-validation. By this limitation, RR shrinks the estimated coefficients towards zero, preserving the most important features. Hence is more stable in comparison with the case of OLS.

### 2.5.1.4 Least Absolute Shrinkage and Selection Operator (LASSO)

LASSO [31],[32], developed in 1996 by Tibshirani, is an alternative regularized version of least squares, which aims to improve model interpretability as well as prediction accuracy by combining the important features of Ridge Regression and subset selection.

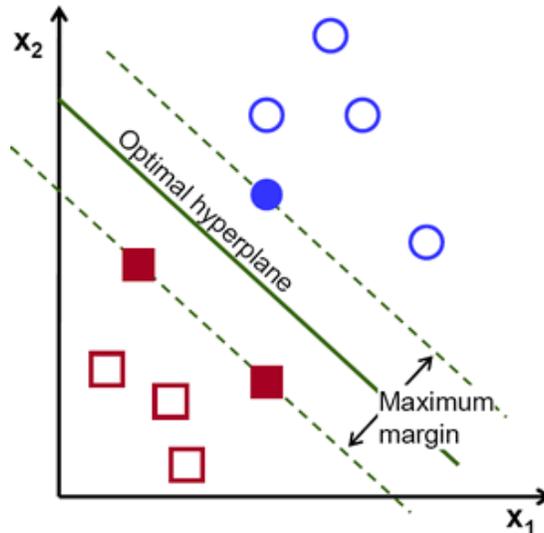
In particular, while RR is able to minimize the variability and improve the accuracy of linear regression models, it cannot perform variable selection in the linear model since it will never sets features to zero exactly. In such settings, Lasso not only reduces the variability of the estimates by shrinking the features but also produces interpretable models by setting a considerable amount of them at exactly zero. Thus, LASSO technique is a slight modification on the Ridge Regression method replaces  $F(w)$  with

$$F(w) = \sum_{n=1}^N (y_n - x_n \cdot w)^2, \text{ subject to } \sum_{k=1}^K |w_k| < \tau$$

Here  $\tau \geq 0$  is a tuning parameter, which is estimated with the same manner as RR. Since feature weights are small numbers the LASSO constraint is more limiting than the one of RR. Particularly, in the case of RR while the constraint is increased the distinct weight of each feature is reduced but still remaining non-zero. However, in LASSO process while the constrained is increased a large number of less important features being assigned weights that are exactly zero. As such, Lasso automatically selects more significant features and discards the others in comparison with RR which never fully discards any features.

## 2.5.2 Support Vector Machine (SVM)

Support Vector Machines (SVMs) [33],[34] are supervised learning algorithm that discover informative patterns and analyze data, applicable for both regression analysis and classification. In the case of binary classification, the SVMs aim at finding the optimal hyperplane that separates all samples between the two classes. The optimal hyperplane for an SVM is the one which maximizes the margin between the classes' closest samples. In general, the goal of maximizing the margin is to minimize the generalization error of the classifier. The samples which lie on the boundaries are called support vectors, and the middle of the margin is our optimal separating hyperplane.



**Figure 2.5:** Binary classification. Samples on the margin are called the support vectors  
 Source: OpenCV documentation, Willow Garage

The derivation of the SVM as presented so far assumed that the data is linearly separable. However, SVM algorithm can efficiently handle non – linearly separable data using a kernel function. In that case, the data are mapped into a higher dimension space, making the separation easier since the data become linearly separable. However, in most cases the data cannot be separated without error. Thus, a modified SVM algorithm, also known as “soft margin” is currently used, minimizing the misclassification rate. The “soft margin” ensures convergence even when the data is non-linearly separable. It creates a hyperplane that separate the samples as correctly as possible, while still maximizing the distance to the nearest classified samples of the two classes.

## Linear SVM

In this part of section we further explain the case of the simple linear SVM algorithm [22],[23] in order to be more clearly the concept of support vectors. Linear SVMs are particular linear discriminant classifiers.

Given a training set  $X$  of  $N$  samples of the form:

$$X = \{(x_i, y_i) | x_i \in R^m, y_i \in \{-1, +1\}\}, i = 1, \dots, N$$

where  $x_i$  the samples and  $y_i$  the class labels, the support vector method approach aims at constructing the maximum - margin hyperplane of dimension  $R^{(m-1)}$  that separate the

samples having  $y_i = +1$  from those having  $y_i = -1$ . Any hyperplane can be expressed as the set of samples  $x$  satisfying:

$$H : w \cdot x - b = 0$$

,where  $b$  a real constant and  $w$  the normal vector to the hyperplane. The offset of the hyperplane from the origin along the normal vector  $w$  can be expressed by the parameter  $\frac{b}{\|w\|}$ . If the data are linearly separable, there are two hyperplanes which can be described by the equations :

$$H_1: w \cdot x - b = 1$$

$$H_2: w \cdot x - b = -1$$

that fully separate the two classes without any samples between of them. The region bounded by these hyperplanes is called “the margin” and is equal to  $\frac{2}{\|w\|}$ . The aim is to maximize the margin, so  $\|w\|$  need to be minimized. Given the fact that  $\|w\|$  is minimized, samples of either class may fall into the margin, so in order to avoid it, extra constraints need to be applied:

$$w \cdot x_i - b \geq 1, \text{ for samples of class } y_i = +1$$

$$w \cdot x_i - b \leq -1, \text{ for samples of class } y_i = -1$$

The above equations can be expressed in one as:

$$y_i(w \cdot x_i - b) \geq 1, \text{ for } i = 1, \dots, N$$

Moreover, the previous constrained equation can be expressed as an optimization problem:

Minimize in  $w, b$

$$\|w\|$$

Subject to

$$y_i(w \cdot x_i - b) \geq 1, \text{ for } i = 1, \dots, N$$

This optimization problem is difficult to solve because it is necessary to calculate the norm of  $w$ , which involve a square root. Without changing the solution it is possible to substitute  $\|w\|$  with  $\frac{1}{2}\|w\|^2$ . So the optimization problem can be also expressed as:

Minimize in  $w, b$

$$\frac{1}{2}\|w\|^2$$

Subject to

$$y_i(w \cdot x_i - b) \geq 1, \text{ for } i = 1, \dots, N$$

By using the Lagrange multipliers  $\alpha$ , the aforementioned problem can be expressed as a problem of quadratic programming:

$$\text{arg min}_{w,b} \max_{a \geq 0} \left\{ \frac{1}{2} \|w\|^2 - \sum_{i=1}^n a_i [y_i (w \cdot x_i - b) - 1] \right\}$$

Then, conforming to the stationary Karush – Kuhn – Turkey condition, the solution can be expressed as a linear combination of the training input vectors:

$$w = \sum_{i=1}^N a_i y_i x_i$$

Only a few of the Lagrange multipliers  $\alpha$  will be greater than zero. These corresponding  $x_i$  are the support vectors and lie on the margin, satisfying :

$$y_i(w \cdot x_i - b) = 1$$

Solving the above equation for b can derive that the support vectors also satisfy:

$$w \cdot x_i - b = \frac{1}{y_i} \Rightarrow b = w \cdot x_i - y_i$$

The b depends on  $x_i, y_i$ , so it will vary among the samples. In that manner, a more stable approach for b is to average over all supports vectors:

$$b = \frac{1}{N_{SV}} \sum_{i=1}^{N_{SV}} (w \cdot x_i - y_i)$$

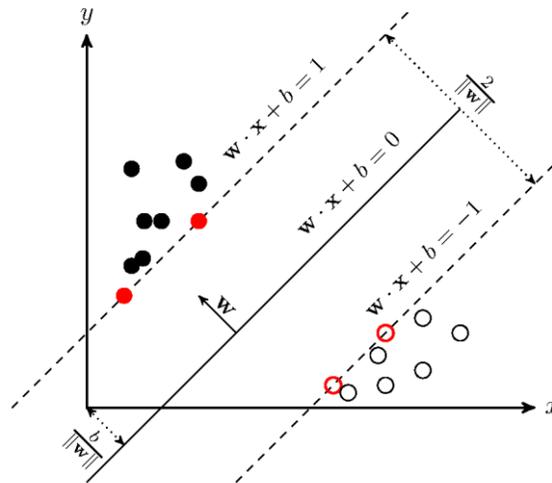
The optimization problem can also be expressed in its dual form, using the fact that  $\|w\|^2 = w \cdot w$  and  $w = \sum_{i=1}^N a_i y_i x_i$ . In dual form the classification task takes into account only a function of the supports vectors, which are a small subset of the set of the training samples that lie on the margin. Thus, the problem expressed in dual form is computationally efficient.

Maximize in  $a_i$

$$\begin{aligned} \tilde{L}(a) &= \sum_{i=1}^N a_i - \frac{1}{2} \sum_{i,j} a_i a_j y_i y_j x_i^T x_j = \\ &= \sum_{i=1}^N a_i - \frac{1}{2} \sum_{i,j} a_i a_j y_i y_j k(x_i, x_j) \end{aligned}$$

, subject to  $a_i \geq 0$ ,  $\sum_{i=1}^N a_i y_i = 0$

and the kernel function is defined by  $K(x_i, x_j) = x_i \cdot x_j$



**Figure 2.6:** Maximum-margin hyperplane and margins of a linear SVM.

Source: Yifan Peng, "Tikz example – SVM trained with samples from two classes", P.Guru, September 2013

### 2.5.3 Relevance Vector Machine (RVM)

The relevance vector machine [19],[35] is a sparse kernel technique for both regression and classification. It has an identical functional form to the state-of-art Support Vector Machine (SVM), but it is a special case of Bayesian Logistic Regression that utilizes a specific type of prior probabilities on the feature weights, called Automatic Relevance Determination (ARD) priors that automatically eliminate irrelevant features from the model. RVM is formed as a linear combination of data-centered basis functions, which are called relevance vectors. Compare to SVMs, RVMs are often found to be advantageous on several aspects including generalization ability and sparseness of the model. In particular, while the SVMs represent decisions, RVMs are based on a Bayesian formulation of a linear model with an applicable prior which is introduced over the weights governed by a set of hyperparameters and bring about a sparse performance.

As a consequence, they can generalize well and provide assumptions at low computational cost, since it typically uses dramatically fewer kernel functions.

RVM is a predictive model that directly models the posterior probability of a class  $C_k$ , given a sample  $p(C_k | x)$ . The RVM requires class labels of the form  $t \in \{0,1\}$ , where in the case of binary classification  $t_i = 1 \rightarrow x_i \in C_1, t_i = 2 \rightarrow x_i \in C_2$ . It computes a model which has the form  $y(w, x) = \sigma(w^T \cdot \varphi(x))$ , where  $\varphi(x)$  a basis function and  $\sigma(\cdot)$  the logistic sigmoid function. Thus according to the RVM procedure, each basis function  $\varphi(x) = k(x, x_n)$  is given by the kernel and each kernel is associated with one data point. The ARD priors have the form  $p(w|a) = \prod_{i=1}^M N(w_i | 0, a_i^{-1})$ . Many of the  $a_i$  are led to infinity and the corresponding features are removed from the model, during the ARD process.

## 2.6 Evaluation methods

Evaluation methods [36] are techniques for assessing how the results of statistical analysis will generalize to an independent data set. The main idea behind the evaluation methods is to split data, once or several times, for estimating how accurately a predictive model will perform in practice: Part of data, the training set, is used for training each model, and the remaining part, the test set, is used for estimating the accuracy of the model.

### 2.6.1 Holdout Validation

The holdout method [36],[37] is the simplest validation method. It partitions the data into two exclusive subsets called a training set and a test set, or holdout test. The training set consists of the majority of available samples and is used for training the model, while the test set conforms to a smaller percentage of the available samples and is used in order to assess the model's generalization ability. However, the holdout method has two basic drawbacks. Particularly, in problems where there is a sparse dataset we may not be able to afford the cost of setting aside a portion of the dataset for testing. Moreover, since it is a single train and test sample, the holdout estimate of error rate will be misleading if we happen to get an unfortunate split. These limitations of the simple holdout method can be overcome with other validation methods at the expense of higher computational cost.

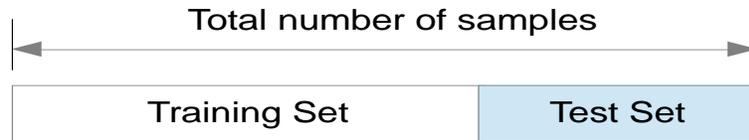


Figure 2.7: Holdout validation method

## 2.6.2 K-Fold Cross Validation (K-Fold CV)

In K-Fold cross-validation [36],[37] the dataset is randomly partitioned into  $k$  subsets of approximately the same size, which are called folds. Of the  $k$  subsets, a single subset is retained as the validation data for testing the model, and the remaining  $(k - 1)$  subsets are used as training data. This process is then repeated  $k$  times, with each of the  $k$  subset used exactly once as the validation data. Then the  $k$  results from the folds are averaged to produce a single estimation. In general,  $k$  remains an unfixed parameter but there are typical values used for it such as 3, 5 or 10. The advantage of this method is that all observations are used for both training and testing, and each observation is used for validation exactly once. Moreover, as the number of folds increases, the bias of the estimate reduces, so the estimation of performance is representative of the actual performance of the method. On the other hand, due to the large number of iterations, the discrimination of the estimation as well as the computational cost increase.

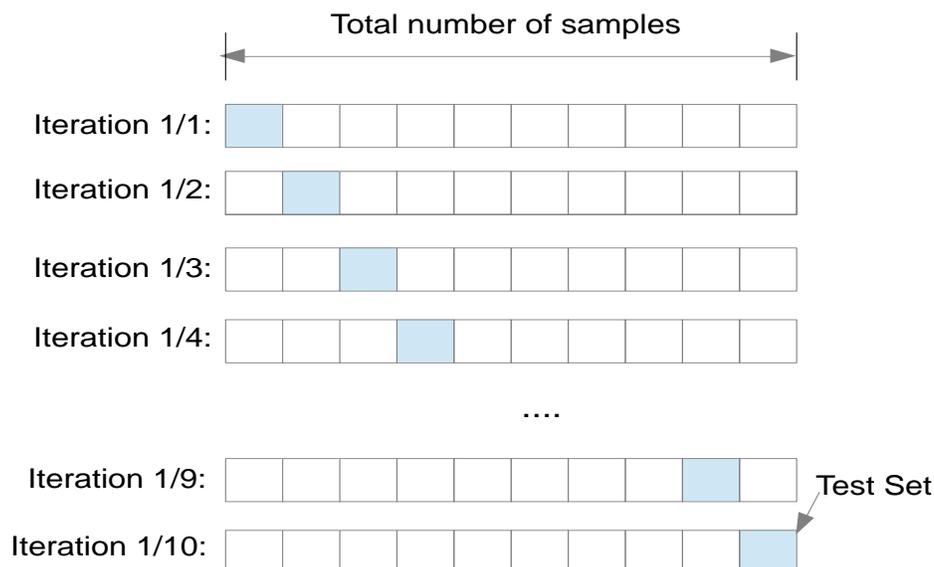


Figure 2.8: K-Fold Cross Validation method

### 2.6.3 Leave One Out Cross Validation (LOOCV)

Leave one out cross validation [36],[37] is the degenerate case of K-Fold cross validation, where K is chosen as the total number of samples in the dataset N. For each fold use N-1 samples for training and the remaining sample for testing. When the number of samples is large, the bias of the true error rate estimator will be small because the estimator will be very accurate, but the discrimination of the true error rate estimator as well as the computational time will be large.

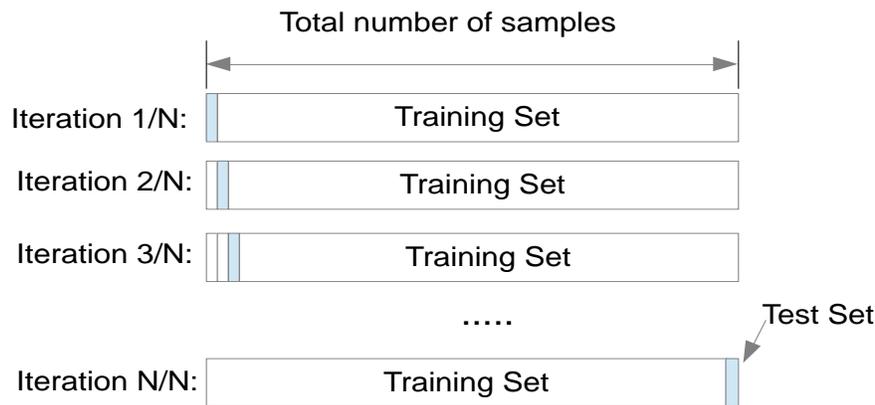
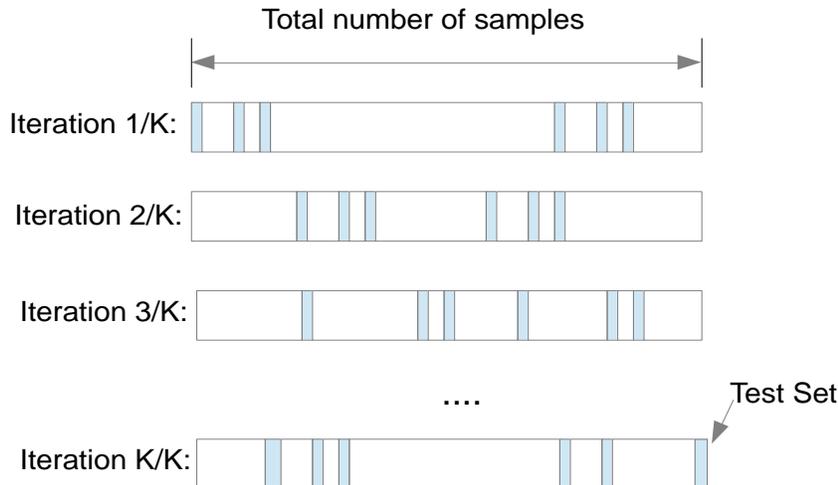


Figure 2.9: Leave One Out validation method

### 2.6.4 Repeated Random Sub-Sampling Validation

Repeated random sub-sampling validation [36],[37] performs K data splits of the dataset. Each data split randomly selects a fixed number of samples without replacement. For each such iteration, the model is fit to the training data, and predictive accuracy is assessed using the validation data. The results are then averaged over all iterations. The advantage of this method (over k-fold cross validation) is that the proportion of the training split is not dependent on the number of folds. While the drawback is that some observations may never be selected in the validation subsample, whereas others may be selected more than once.



## 2.6.5 Bootstrap Resampling Validation

The bootstrap resampling validation method [36],[37] which also called bootstrapping, is a random sampling technique with replacement. In particular, from a dataset with  $N$  samples randomly select with replacement a number of  $B$  bootstrap datasets of fixed size, usually the same number of  $N$  samples. Then, using the holdout method, each bootstrap dataset can be divided into training and test sets. At the end of the procedure in order to get a stable estimation, the statistics are calculated for each bootstrap dataset and are averaged over all bootstrap datasets.

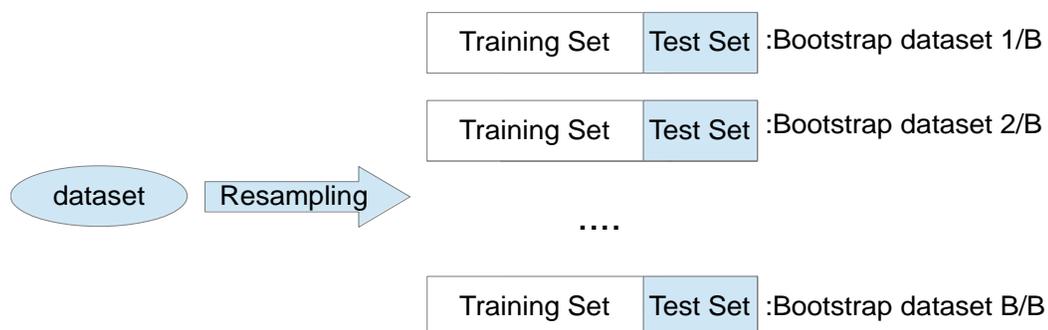


Figure 2.11: Bootstrap resampling validation

## 2.7 Weak Law of Large Number (LLN)

The weak law of large number [38], in probability theory is an approach which describes the result of executing a random experiment a sufficiently large number of times. Particularly, according to the aforementioned law the mean value of the obtained results from a large number of iterations will be closed to the expected value and will tend to become closer as more experiments are performed.

Let  $X_1 \dots X_N$  be a sequence of independent and identically distributed random variables, each having a mean  $\bar{X}_i = \mu$  and standard deviation  $\sigma$ .

Define a new variable  $\bar{X} \equiv \frac{X_1 + \dots + X_n}{n}$ .

Then as the number of experiments  $n \rightarrow \infty$  the sample mean  $\bar{X}$  equals the population mean  $\mu$  of each variable:

$$\bar{X} = \frac{\overline{X_1 + \dots + X_n}}{n} = \frac{\bar{X}_1 + \dots + \bar{X}_n}{n} = \frac{n \cdot \mu}{n} = \mu$$

In addition,

$$\begin{aligned} var(X) &= var\left(\frac{X_1 + \dots + X_n}{n}\right) = var\left(\frac{X_1}{n}\right) + \dots + var\left(\frac{X_n}{n}\right) = \left(\frac{\sigma^2}{n^2}\right) + \dots + \left(\frac{\sigma^2}{n^2}\right) \\ &= n \cdot \left(\frac{\sigma^2}{n^2}\right) = \frac{\sigma^2}{n} \end{aligned}$$

Moreover, by the Chebyshev inequality, for all  $\varepsilon > 0$ ,

$$P(|X - \mu| \geq \varepsilon) = var \frac{(X)}{\varepsilon^2} = \frac{\sigma^2}{n \cdot \varepsilon^2}$$

and as  $n \rightarrow \infty : \lim_{n \rightarrow \infty} P(|X - \mu| \geq \varepsilon) = 0$

The weak law of large numbers can be used in order to assess the stability of results in genomic datasets. In particular, bootstrap resampling can be utilized in order to generate a sufficiently large number of dataset. Then, under the perception that the observed results are independent and identically distributed random variables, according to LLN the average estimates for the classification accuracy and the size of the genomic signature will be stable.

# 3

## Methodology

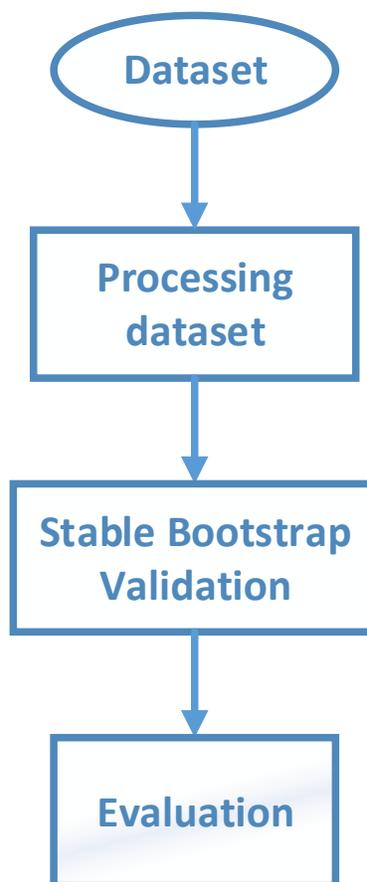
The goal of this section is to suggest a methodology for performing reliable feature selection and stable classification accuracy as well as for evaluating the consistency and generalization ability of the results.

A number of feature subset selection (FSS) methods have been developed for gene selection in microarray data. The first step of this methodology is proposed in section 3.1 and it has to do with the processing of the dataset. In this section, the data has undergone feature subset selection (FSS) using a filter univariate method (SAM). Several justifications for the use of filters for subset selection in DNA microanalysis have been put forward in this thesis.

Another significant aspect of microarrays analysis is the stability of performance assessments. A wide variety of machine learning methods have been proposed for classification tasks related to microarrays, including support vector machines (SVM), relevance vector machines (RVM), K-Fold Cross Validation and many others. However, the use of an arbitrarily fixed combination of FSS method and classifier can lead to significant variations not only in the training or testing dataset but also in the set of features selected as well as classification accuracy. Thus, may sacrifice performance that could have been achieved with another model. Hence, in order to extract robust performance estimates, a methodology that utilizes repeated resampling or splitting of the original dataset has been suggested. The Stable Bootstrap Validation methodology of Nikolaos-Kosmas Chlis in [4] is applied in section 3.2. This approach utilizes a formal criterion in order to extract robust estimates for the size of genomic signature as well as the classification accuracy and no further iterations are required. The stable estimates can be reproduced resulting in minimal variations during independent executions of the evaluation method.

The third step of our methodology is purposed in section 3.3 and includes the evaluation of the observed results which constitutes a fundamental aspect of microarray analysis.

The methodology of accessing the discrimination of genomic signature between the class labels is applied in section 3.3.1. Meanwhile, section 3.3.2 introduces a methodology concerning the assessment of consistency regarding the observed classification performance of a genomic signature. If a classification method is consistent, it should lead to considerable repeatability of results. Finally in section 3.3.3 the evaluation of generalization ability of genomic signature is proposed. This field examines how the results of the statistical analysis will generalize to an independent data set. The overview of the proposed methodology is presented as a block diagram in figure 3.1.



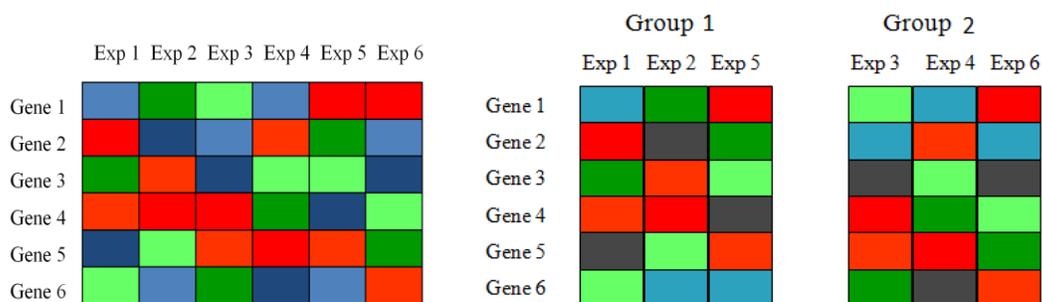
**Figure 3.1:** Overview of the proposed methodology

### 3.1 Processing the dataset: SAM

We have already mentioned in section 2.3 that the feature gene selection is one of the crucial steps in DNA microanalysis. Our original data is composed of a small number of samples (53 samples) which have a large number of gene expression (54675 genes) levels as features. Thus, our first step is to reduce the number of genes by keeping the most relevant set. That being the case, the original dataset it has undergone feature subset selection using a filter univariate method which is called “Significance Analysis of Microarrays” (SAM) [2], [3]. SAM uses a modified t-statistic and permutations of the repeated measurements of the data in order to decide if the gene expression is strongly related to the response. The theoretical background of SAM has been analyzed in detail in section 2.3.1.1 but the SAM procedure proceeds as follows.

The data should be put in an Excel spreadsheet and have a specific format. Particularly, the first row has information about the response measurement; all remaining rows have gene expression data, one row per gene. The columns represent the different experimental samples.

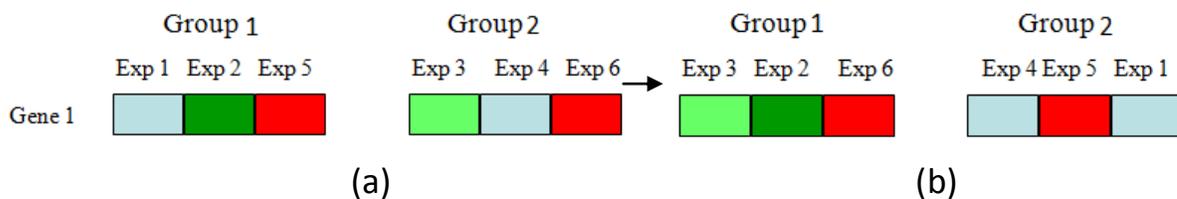
The input to SAM is gene expression measurements from a set of microarray experiments, as well as a response variable from each experiment. There are many different types of response such as quantitative, one class, two class (unpaired, paired), multiclass, survival data, time course and pattern discovery. In our case, gene expression measurements are separated into two class (unpaired) groups. These groups are two sets of measurements, in which the experiment units are all different. Particularly, we have two groups: healthy controls and medicated bipolar disorder patients (which also contains bipolar disorder patients in first episode in), with samples from different patients. Thus the response variable is grouped using numbers 1 (healthy control) – 2 (bipolar disorder patient).



**Figure 3.2:** Assign experiments to two groups (1,2)  
Source: [2]

The procedure of repeated permutations of the data which determine if the expression of any genes is significantly related to the response proceeds as follows:

1. For each gene, compute a statistic d-value, which is the observed d-value for that gene.
2. Order the genes according to their d- values.
3. Randomly shuffle the values of the genes between groups 1 and 2, such that the reshuffled groups 1 and 2 respectively have the same number of elements as the original groups 1 and 2. Compute the d-value for each randomized gene.



**Figure 3.3:** (a) original grouping, (b) randomized grouping  
Source: [2]

4. Order the genes according to their permuted d- values.
5. Repeat steps 3 and 4 many times. Thus, each gene has many randomized d-values corresponding to its rank from the observed (unpermuted) d-value (100 or 200 permutations are descent for initial exploratory analysis). Then, take the average of the randomized d-values for each gene which is the expected d-value of that gene.
6. Plot the observed d-values versus the expected d-values
7. For each permutation of the data, compute the number of positive and negative significant genes for a delta parameter, which is the cutoff for significance, chosen by the user based on the false positive rate. The median number of significant genes from these permutations is the median False Discovery Rate (FDR). Thus, any genes designated as significant from the randomized data are being picked up purely by chance. Therefore, the median number picked up over many randomizations is a descent estimate of FDR.

The procedure of running the SAM proceeds as follow: first, the area that represents the data should be highlighted. Then the SAM button in the toolbar must be selected and a dialog rises. The dialog box gives the opportunity to the user to select the type of response variable and to change any of values of the default parameters. Moreover the user should specify if the data are from (micro)array or a sequencing experiments and for two class and paired data, one has to specify if the data is in the logged (base 2) scale or not.

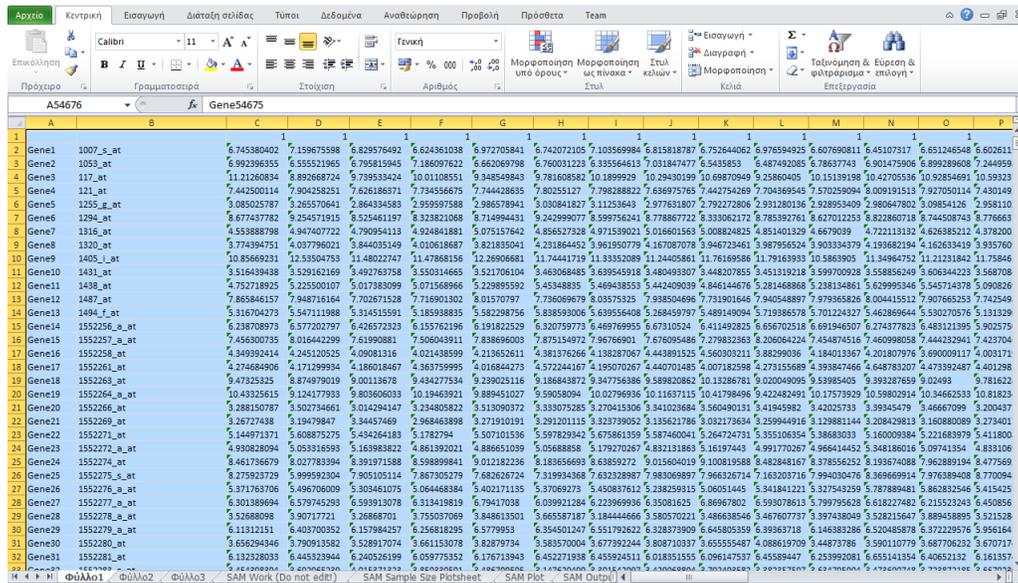


Figure 3.4: Highlighting and invoking SAM. Source [3]

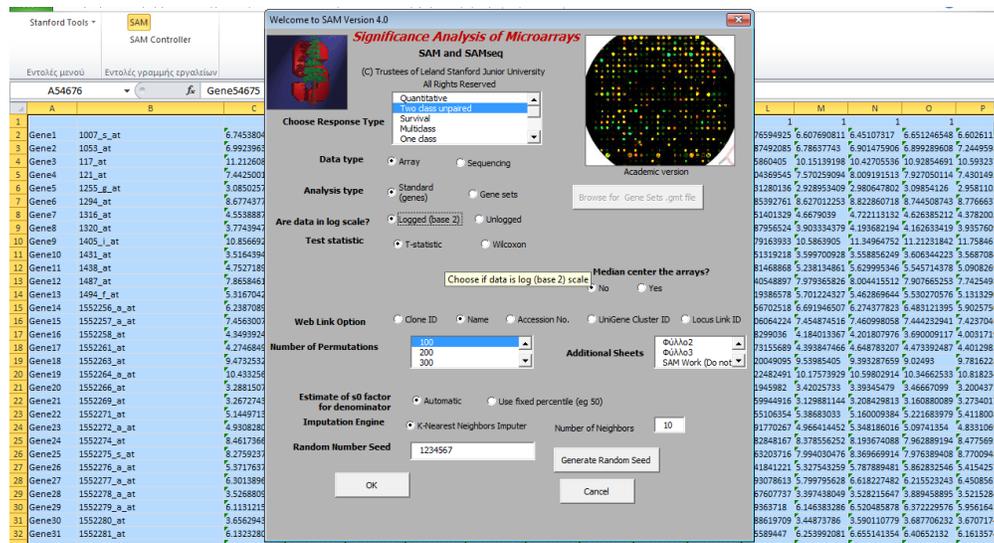
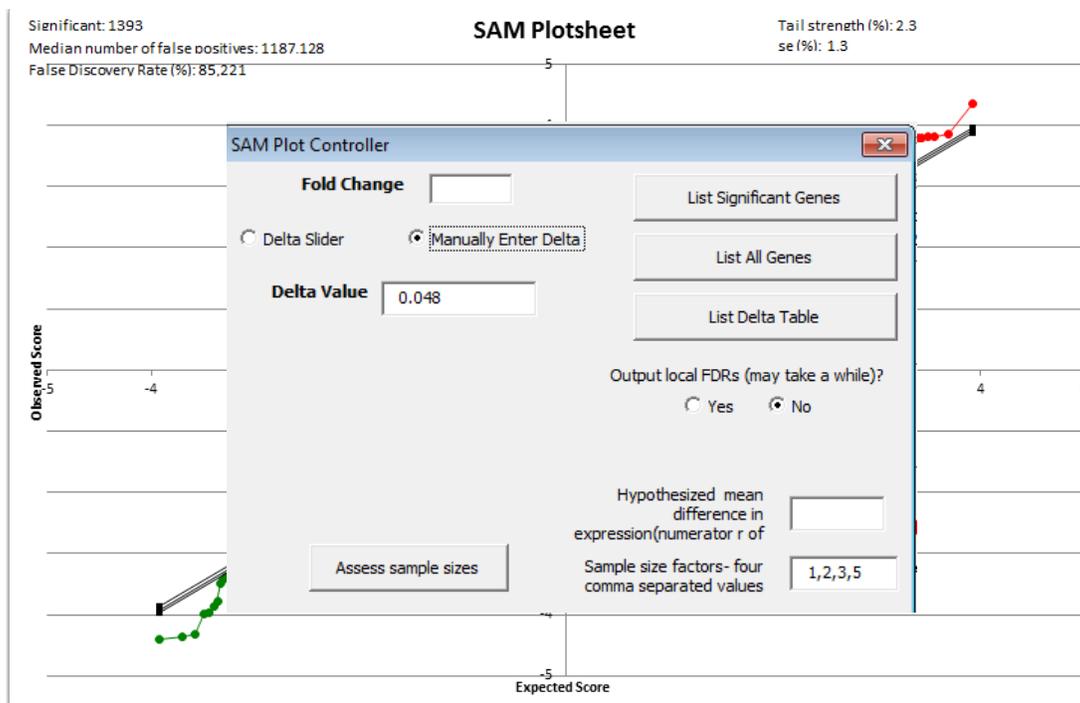


Figure 3.5: The SAM Dialog Box. Source [3]

While running the SAM, if there is any missing data in your spreadsheet, a new worksheet named SAM Imputed dataset containing the imputed dataset is added to the workbook, unless this worksheet is not added. Therefore, the software adds two more worksheets to the workbook. There is one which is hidden called SAM Plot data which contains the plot of the observed d-values versus the expected d-values and the user can interact with. Particular, a block dialog which is called Sam Plot Controller, shown in figure 3.6, gives the chance to the user to change the delta parameter and examine the effect on the false positive rate. If user wants a more stringent criterion, there is also a fold change parameter that he can select. Positive significant genes are labeled in red on the SAM plot, while negative significant genes are green. The List Delta Table button lists the number of significant genes and the false positive rate for a number of values of delta. The List All Genes prints out all genes in the dataset. After choosing the delta parameter a sheet named SAM Output is showed, including any output.



**Figure 3.6:** The SAM Plot Controller on the front side,  
The SAM Plot sheet on the second side  
Source: [3]

The output for list of significant genes has a specific format [14]. Particularly, it contains the row number, which is the row in the selected data rectangle, the gene name as well as the gene Id. It also contains the SAM score ( $d$ ), which is the t-statistic value with the numerator and the denominator ( $s + s_0$ ) of it. Moreover, the q-value, which is the lowest False Discovery Rate at which the gene is called significant as well as the local FDR, which is the false discovery rate for genes with scores  $d$  that fall in a window around the score for the given gene are also printed. Finally, in any testing problem, false positive rate (for example FDRs) are calculated, but is also important to consider false negative rates. Thus, a miss rate table is printed which gives the estimated false negative rate for genes that do not make the list of significant genes.

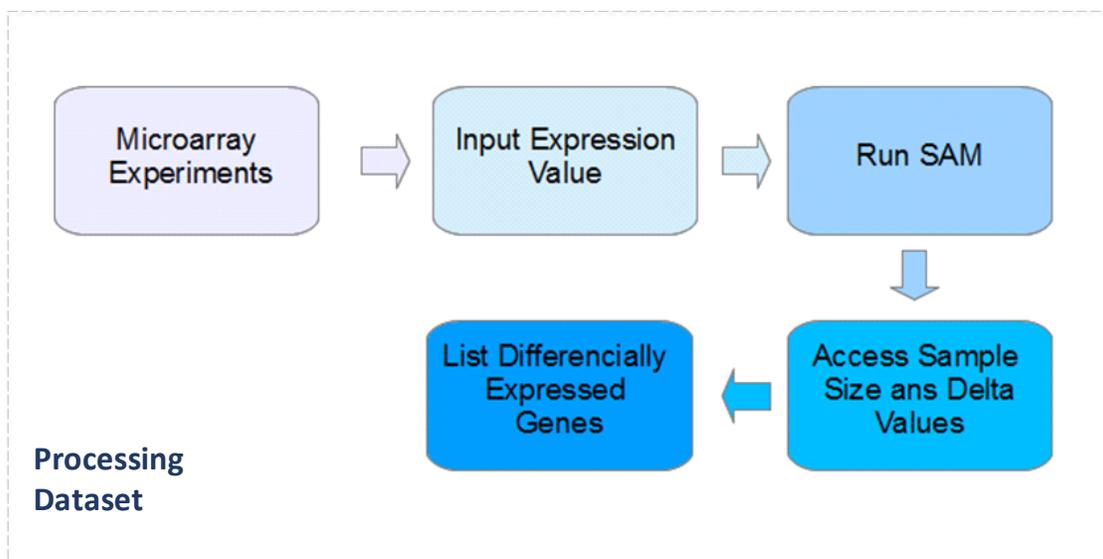
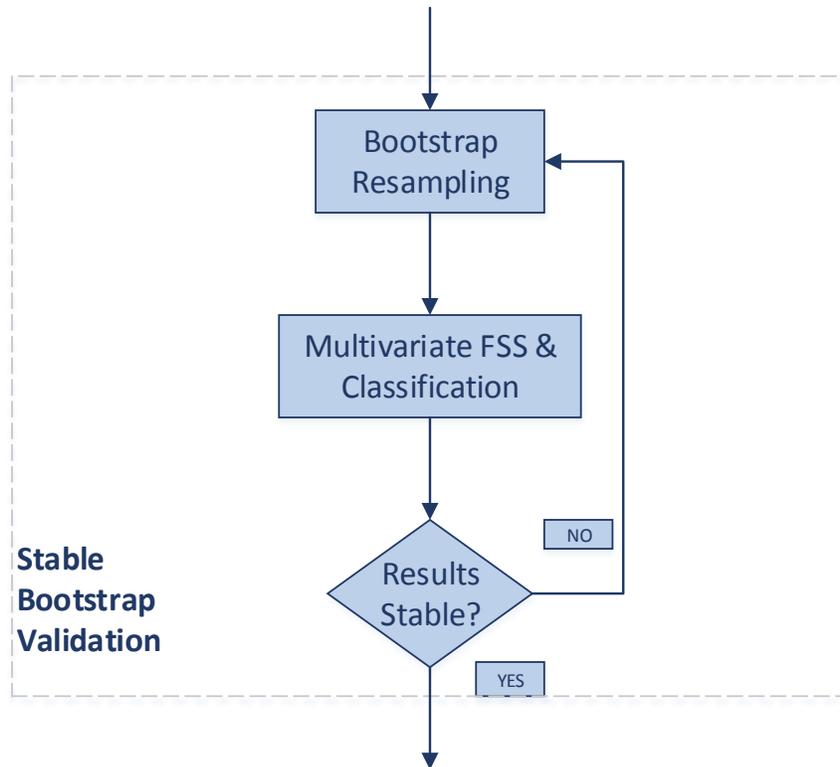


Figure 3.7: Flowchart for Preprocessing the Dataset - SAM

## 3.2 Stable Bootstrap Validation

As we mentioned above during the step of preprocessing, the dataset has undergone feature subset selection using a filter univariate method (SAM). Nevertheless, during the step of Stable Bootstrap Validation [4], which constitutes the second one of our methodology, the embedded multivariate feature subset selection approach of recursive feature elimination is applied. Since this approach uses both univariate and multivariate methods the observed results are expected to get the benefits of both schemes.



**Figure 3.8:** Overview of Stable Bootstrap Validation approach

The goal of the Stable Bootstrap Validation approach is to perform robust estimates for the classification accuracy and the size of the genomic signature. Thus, given a pair of feature selection subset and classification methods, SBV focus at utilizing a large number of datasets generated from bootstrap resampling of the observed dataset. These datasets will be used for the evaluation of feature selection as well as classification approaches. Thus, first of all a fix number of bootstraps datasets called “bootstrap window”  $B$  is defined. Then, a number of  $3B$  bootstrap datasets are generated from the original dataset by random sampling with replacement. . The feature subset selection as well as classification approach are then executed  $3B$  times, resulting in values  $A_1 \dots A_{3B}$  for the classification accuracy and  $G_1 \dots G_{3B}$  for the number of features selected. Next, assuming that  $A_i$  and  $G_i$  are sets of independent identically distributed random variables according to the LLN the average estimates over all samples  $\bar{A}$  and  $\bar{G}$  should converge towards the expected value of classification accuracy and the size of the genomic signature, respectively. That is, the average estimates can be used as a measure of stability. In order to determine when the sample size is large enough and no more bootstrap datasets are generated than necessary, SVM uses an

explicit dual criterion determining whether the stability of results have been reached for both signature size and average classification accuracy. Through batches of subsequent  $B$  trials, the above criterion determines the stability of observed results and assesses if the necessary level of stability has been reached. Otherwise, another set of  $B$  datasets is generated and the stability assessment is performed again for the 3 windows, which now extend to cover the additional datasets. The above steps are repeated until stability for the classification accuracy as well as the signature size is reached. In comparison to similar approaches, which utilize an unnecessary large number of evaluation iterations, SBV is a computationally efficient methodology since is only executed until the desired level of stability is reached. Therefore, as we already mentioned the majority of similar approaches on the one hand tend to extract stable estimates for the classification accuracy but on the other hand select an arbitrary number of genes. To address this issue, after the SBV procedure has been completed,  $\bar{A}$  is considered to be the stable assessment of classification performance, while  $\bar{G}$  is the stable assessment of the genomic signature extracted by the FSS method. The classification accuracy estimate  $\bar{A}$  is considered stable according to  $acc_{thresh}$ , which is a fixed threshold. While the corresponding threshold for the signature size is normalized by the largest signature size, which is called  $gen_{thresh}$ . When both  $\bar{A}$  and  $\bar{G}$  are found stable the SVM procedure terminates. Finally, the  $\bar{G}$  genes with the highest selection frequency over all iterations of the method are selected as the genomic signature of the specific combination of FSS & classification methods. The SBV procedure proceeds in detail in the diploma thesis of Nikolaos-Kosmas Chlis in [].

## 3.3 Evaluation of the Results

### 3.3.1 Evaluation of Discrimination of Genomic Signature

Another aspect of evaluation is the one of discrimination of genomic signature between the class labels. In particular, the expression value of each gene should be examined in order to access the dispersion between the class labels. In that manner, the mean as well as the variance and the standard deviation of each gene are performed. In Statistics, the mean gives a very good idea about the central tendency of the data being collected, while the variance and the closely-related standard deviation are measures of how spread out a distribution is. In other words, they are measures of variability. Variance describes how much a random variable differs from its expected value.

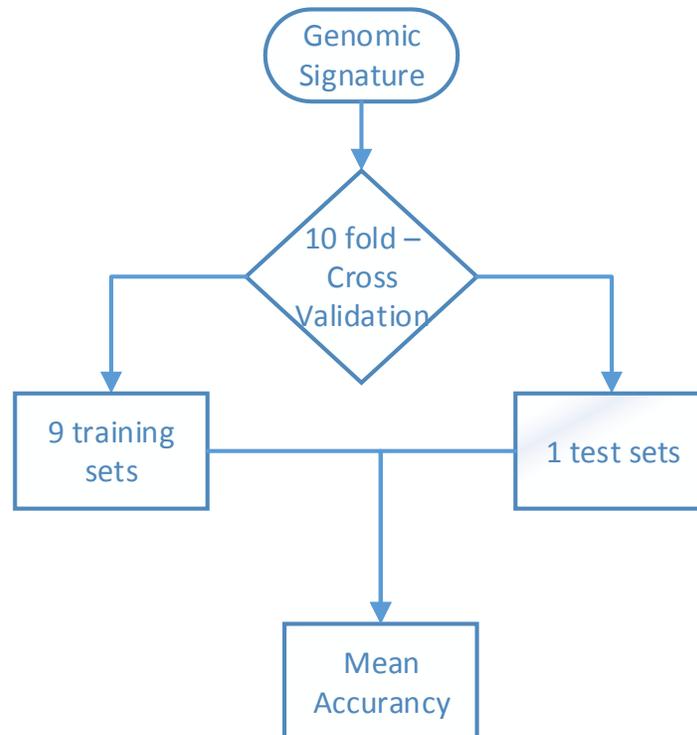
First the mean as well as the variance and the standard deviation of the each significant gene are calculated. The mean is known as a measure of location; that is, it tells us

where the data are. To calculate the mean we add up the observed values and divide by the number of them. The variance is defined as the average of the squares of the differences between the individual (observed) and the expected value, while standard deviation is calculated as the square root of variance. Then the standard deviation is added as well as reduced to the mean in order to evaluate the dispersion of the expression values of each significant gene between the class labels. Specifically, a standard deviation close to null indicates that the expression value of each gene tend to be very close to the mean, which also called the expected value of the set, while a high standard deviation indicates that the samples are extended over a wider range of values.

### **3.3.2 Consistency Evaluation of gene selection in the signature**

The evaluation of consistency of gene selection in the signature [39] is another significant aspect of microarray analysis and refers to the reliability of genomic signature. Particularly, the one refers to the ability of the genomic signature to yield similar performance when applied on a single test set multiple times, while using different training sets. There are many model validation techniques, which have been analyzed in section 2.6. In this thesis, the k – fold cross validation, specifically 10 fold CV is implemented, using the genomic signature. The main idea behind the 10 – fold cross validation is to divide the data into 9 training sets and 1 testing set, then train on the training set and use the testing set for estimating how accurately a predictive model will perform in practice.

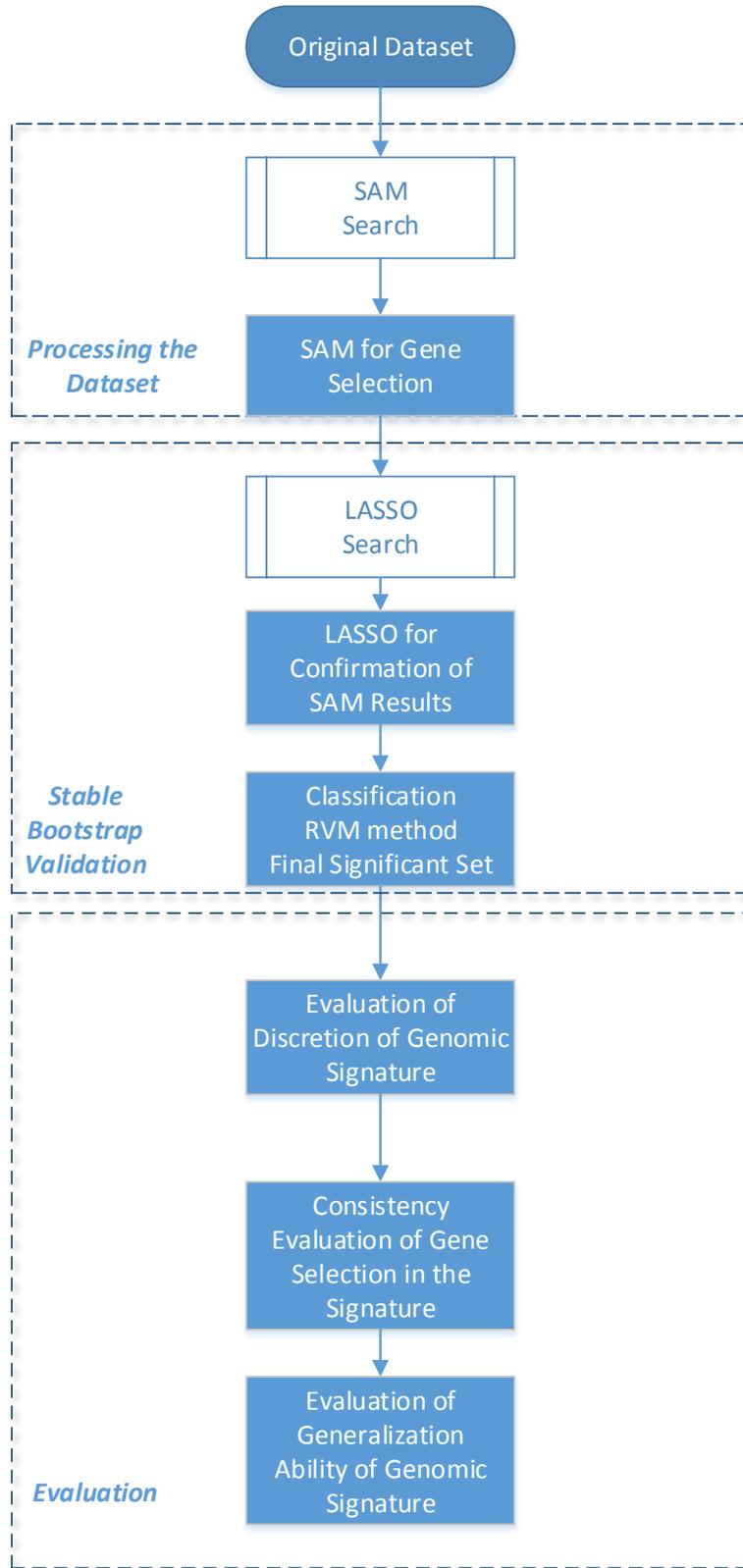
The procedure of k – fold cross validation proceeds as follow. First an integer k, which constitutes the parts that the dataset is divided, is chosen; specifically ten parts. Then, the original dataset is randomly partitioned into 10 subsets of approximately the same size. Of the 10 subsets, a single subset is retained as the testing set in order to access the strength as well as utility of the predictive relationship, and the remaining 9 subsets are used as training data. This process is then repeated 10 rounds. In each round, one of the folds is used for validation, and the remaining folds for training. Then, after training the classifier, its accuracy on the testing data is calculated. Finally, the k results from the folds are averaged to produce the final cross-validation accuracy as well as the corresponding variance, shown in figure 3.9. The above procedure is repeated a total of 200 iterations and the results are averaged to produce a more stable evaluation.



**Figure 3.9:** Flowchart corresponding to one iteration of the 10 - fold Cross Validation methodology.

### 3.3.3 Evaluation of Generalization Ability of Genomic Signature

Another significant aspect of microarray analysis is the evaluation of generalization ability of genomic signature. A good generalization performance is achieved when a genomic signature is able to predict the label of unseen samples correctly. Cross-validation is a widespread strategy because of its simplicity and its universality. Thus, the  $k$  - fold cross validation approach can also be used to assess how the results of a statistical analysis will generalize to an independent data set. In that manner, a new independent dataset is used and the aforementioned procedure of 10 - fold cross validation is repeated.



**Figure 3.10:** Structure of the overall proposed methodology

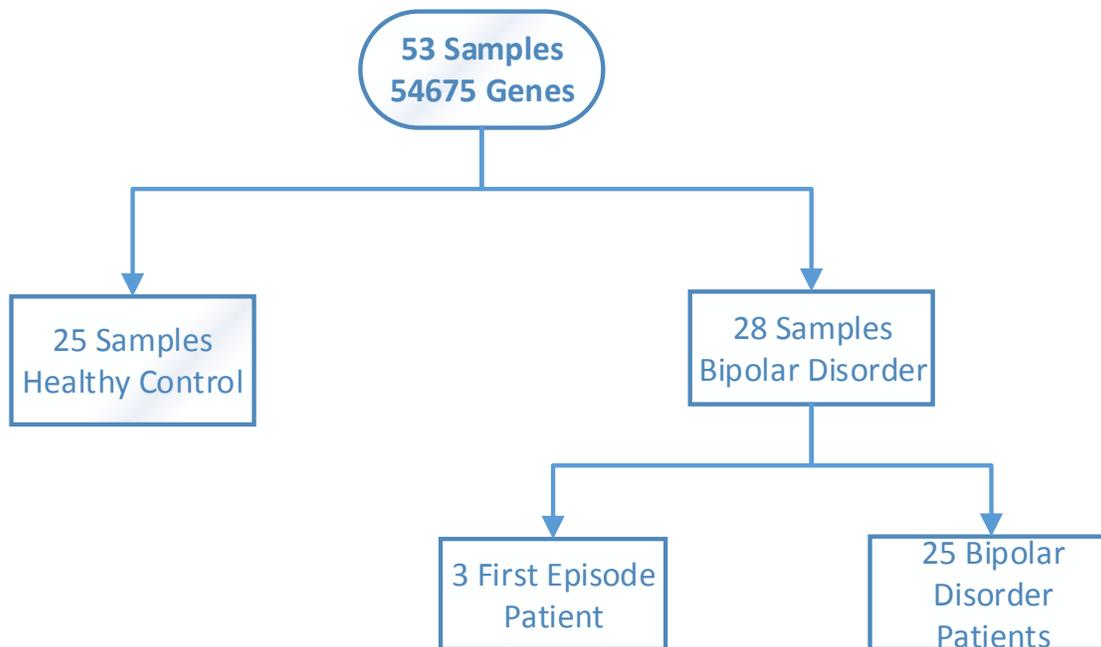
# 4

## Results

In this chapter the original dataset is introduced in section 4.1, followed by the results of feature subset selection methods in section 4.2. Moreover, the performance metrics of classification methods, including LASSO, SVM and RVM, extracted by SBV are presented in section 4.3. The statistical significance as well as the observed genomic signature significance of the above SBV results is then assessed in section 4.4.

### 4.1 Original dataset

The original dataset results from measurements of global leukocyte gene expression. Peripheral blood leukocytes from whole blood were collected from 25 patients with bipolar disorder who had previously received medication, 3 patients with bipolar disorder who were experiencing their first episode and had not previously received medication, and 25 matched control subjects. Thus the original dataset (GEO access number: GSE46449) consists of 53 samples related to bipolar disorder, 25 of which correspond to healthy control and 28 to bipolar samples. For each sample, there are measurements of 54675 genes.



**Figure 4.1:** Structure of the Original Dataset

## 4.2 Processing the Dataset Results

As mentioned in section 3.1, the aim of the step of processing the dataset is to reduce the number of genes by keeping the most significant. In this diploma this, the original dataset it has undergone feature subset selection using the filter univariate method, Significance Analysis of Microarrays (SAM). SAM uses a modified t-statistic and permutations of the repeated measurements of the data in order to decide if the gene expression is strongly related to the response. After the SAM method is run for a sufficient number of times and the relevant set is estimated according to the parameter delta the procedure terminates and returns the most significant set of genes.

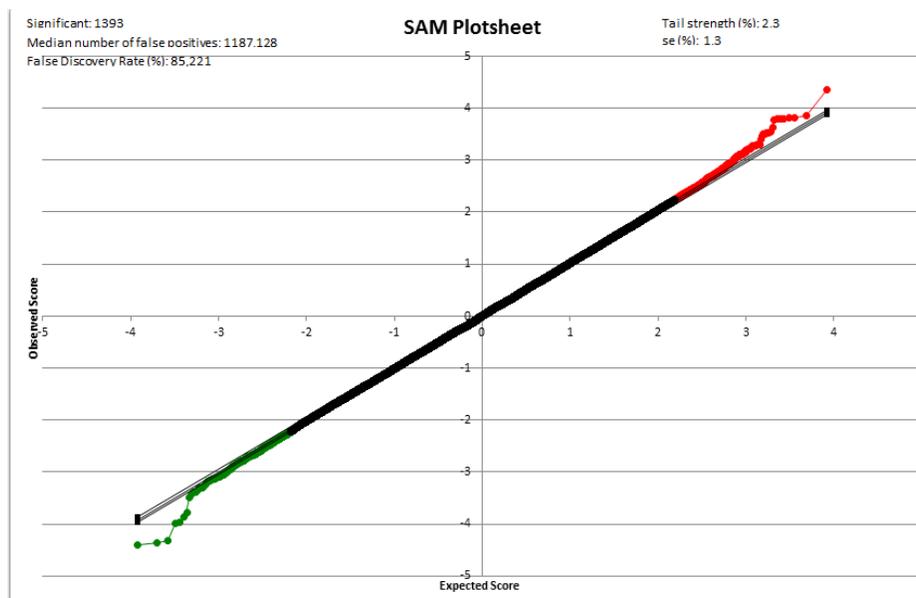
### 4.2.1 SAM Parameters

As already mentioned gene expression measurements are separated into two class (unpaired) groups, healthy controls (25 samples) and medicated bipolar disorder patients(25 samples), which also contains bipolar disorder patients in first episode (3 samples). Thus the response variable is grouped using numbers 1 (healthy control) – 2 (bipolar disorder patient). So, in the dialog box the two class univariate response is chosen, while the data is specified as (micro) array experiments in logged (base 2) scale.

In this diploma thesis, because of the small number of first episode bipolar patients the SAM procedure is run multiple times for different combinations of samples, in order to access the impact of medication in patients. The first group is composed of healthy controls and medicated bipolar disorder patients. The second one consisted of healthy controls and all bipolar disorder patients. The third one is composed of healthy controls and only the first episode bipolar disorder patients. In the Sam Plot Controller box the delta parameter, which is the cutoff of significance, was set to default values and the follow plots are appeared, figure 4.2, 4.3, 4.4.

Groups	Delta Parameter	FDR (%)	Genomic Signature Size
Healthy control (25 samples) – Medicated Bipolar disorder (25 samples)	0.043	85.2	1393

**Table 4.1:** SAM results from Healthy Control – Medicated Bipolar Disorder patients



**Figure 4.2:** The SAM Plot sheet of 1393 significant genes (Healthy controls – All bipolar disorder patients)

Response Data	Delta Parameter	FDR (%)	Genomic Signature Size
Healthy control (25 samples) – BD all (28 samples)	0.049	81.65	360

Table 4.2: SAM results from Healthy Control – All Bipolar Disorder patients

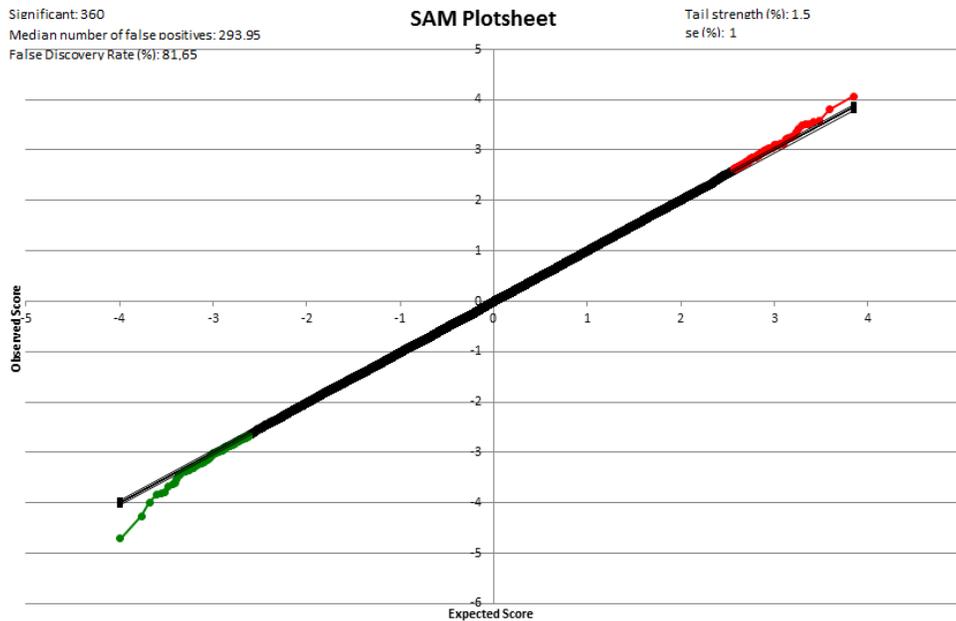
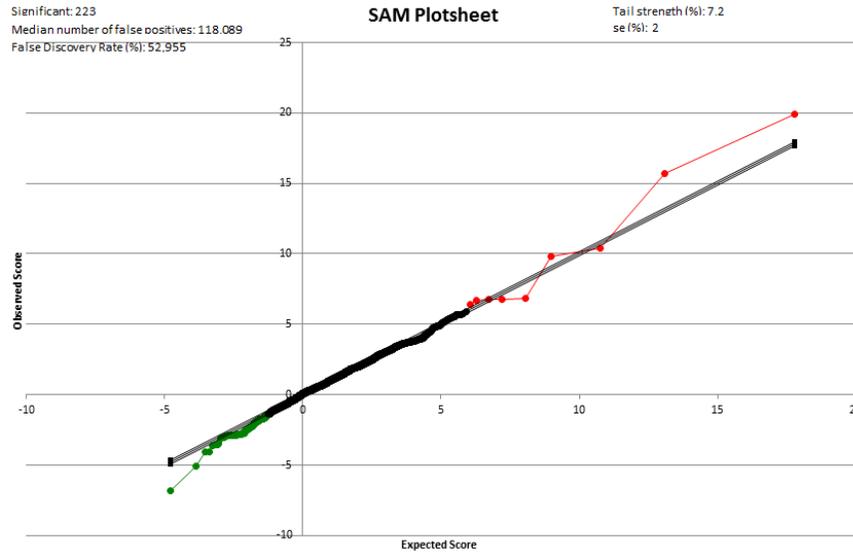


Figure 4.3: The SAM Plot sheet of 360 significant genes (Healthy controls – All bipolar disorder patients)

Groups	Delta Parameter	FDR (%)	Genomic Signature Size
Healthy control (25 samples) – First Episode Bipolar disorder patients (3 samples)	0.15	52.9	223

Table 4.3: SAM results from Healthy Control – First Episode Bipolar Disorder patients



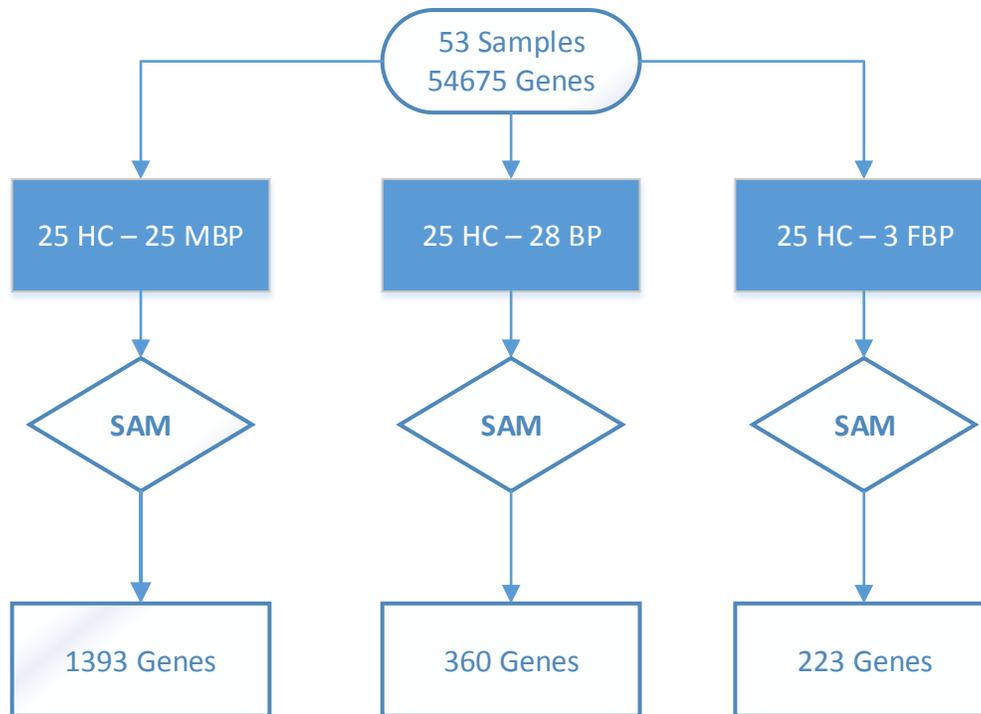
**Figure 4.4:** The SAM Plot sheet of 223 significant genes (Healthy controls – First episode bipolar disorder patients)

Given the fact that the population of first episode bipolar patients is small enough in comparison with the population of healthy controls we cannot come to efficient inferences concerning the disease. Thus, the SAM procedure is repeated three more times. Particularly, from the second group with all patients the first episode patients is removed one by one, shown in Table 4.4.

Response Data	Delta Parameter	FDR (%)	Significant Genes
Healthy control (25 samples) – Bipolar patient (25 samples)+ First Episode patient (3 samples)	0.049	81.65	360
Healthy control (25 samples) – Bipolar patient (25 samples)+ First Episode patients (2 samples)	0.049	80.48	459
Healthy control (25 samples) – Bipolar patient (25 samples)+ First Episode patient (1 samples)	0.049	82.58	598
Healthy control (25 samples) – Bipolar patient (25 samples)	0.049	84.51	969

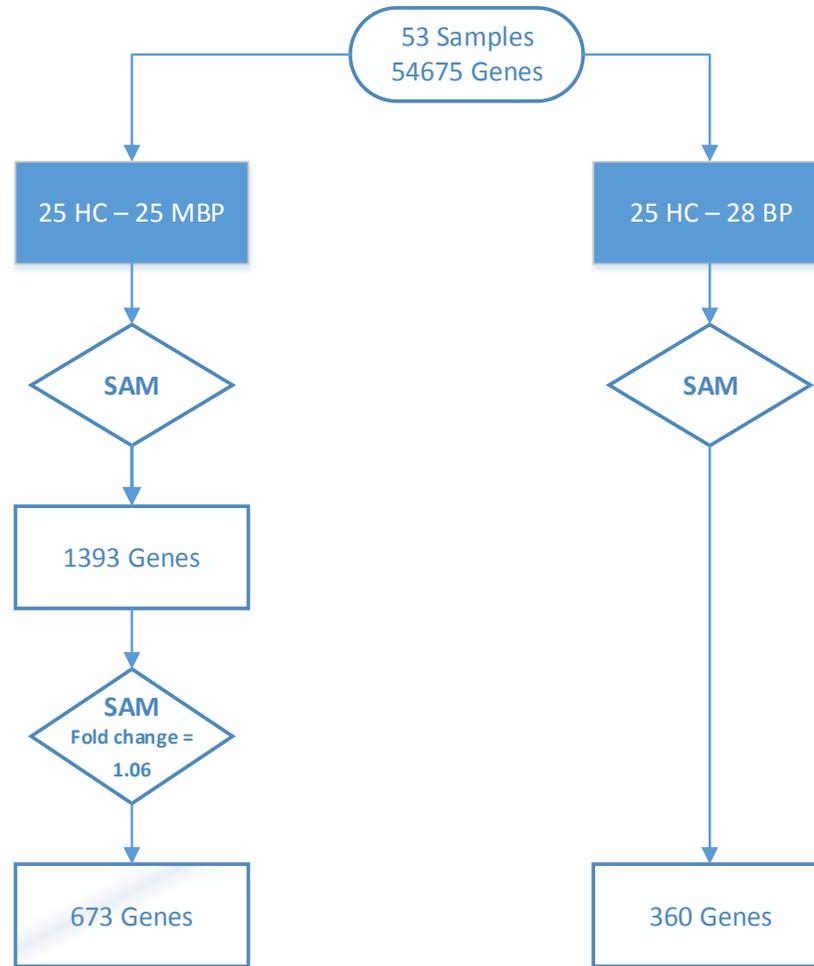
**Table 4.4:** SAM results from Healthy Control – All Bipolar Disorder patients removing one by one First Episode patient

In this study is observed that removing one by one the first episode patients from the group of all bipolar disorder patients, the number of significant genes tended to increase and the results of medication mitigated. Particularly, the set of 360 significant genes provides confidence due to the fact that it tends to represent highly diverse as well as variance.



**Figure 4.5:** Structure of group 1: significant genes from healthy controls and medicated bipolar patient, group 2: significant genes from healthy controls and all bipolar patients and group 3: significant genes from healthy controls and first episode bipolar patients.

Furthermore, because of the fact that the margin between set of significant genes which emerged from the first group (1393 genes) and the second group (360 genes) is large enough, the fold change parameter in the Sam Plot Controller of first group was set to 1.06, which is shown in figure 4.6.



**Figure 4.6:** Left: Structure of significant genes from healthy controls and medicated bipolar patient (fold change=1.06). Right: Structure of genomic signature from healthy controls and all bipolar patients

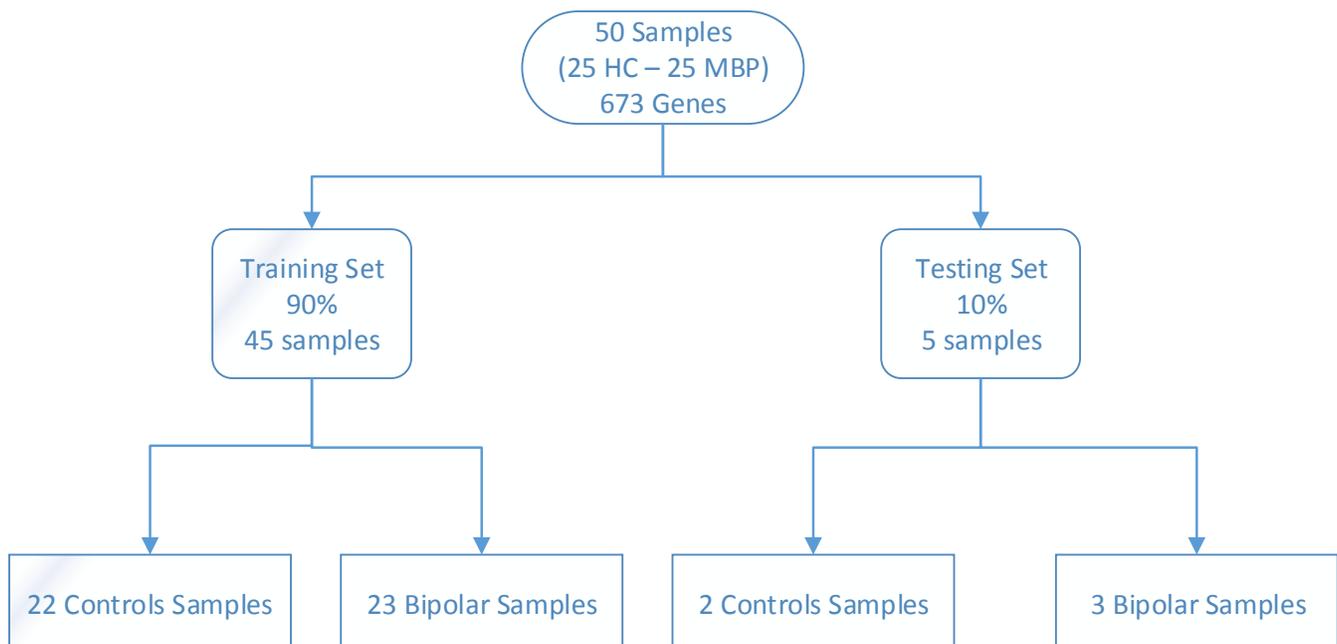
Finally, from the original dataset, which is composed of a small number of samples (53 samples) and a large number of gene expression (54675 genes), through Significance Analysis of Microarrays procedure the number of genes is reduced to a great scale by keeping the most relevant set, also called “significant genes”. Specifically, there are two significant sets which emerged from different combination of samples. The first relevant set consisted of 673 significant genes from 25 healthy controls and 25 medicated bipolar patients and the second one is composed of 360 significant genes from 25 healthy controls and 28 bipolar patients (including the 3 first episode bipolar disorder patients).

## 4.3 SBV Results

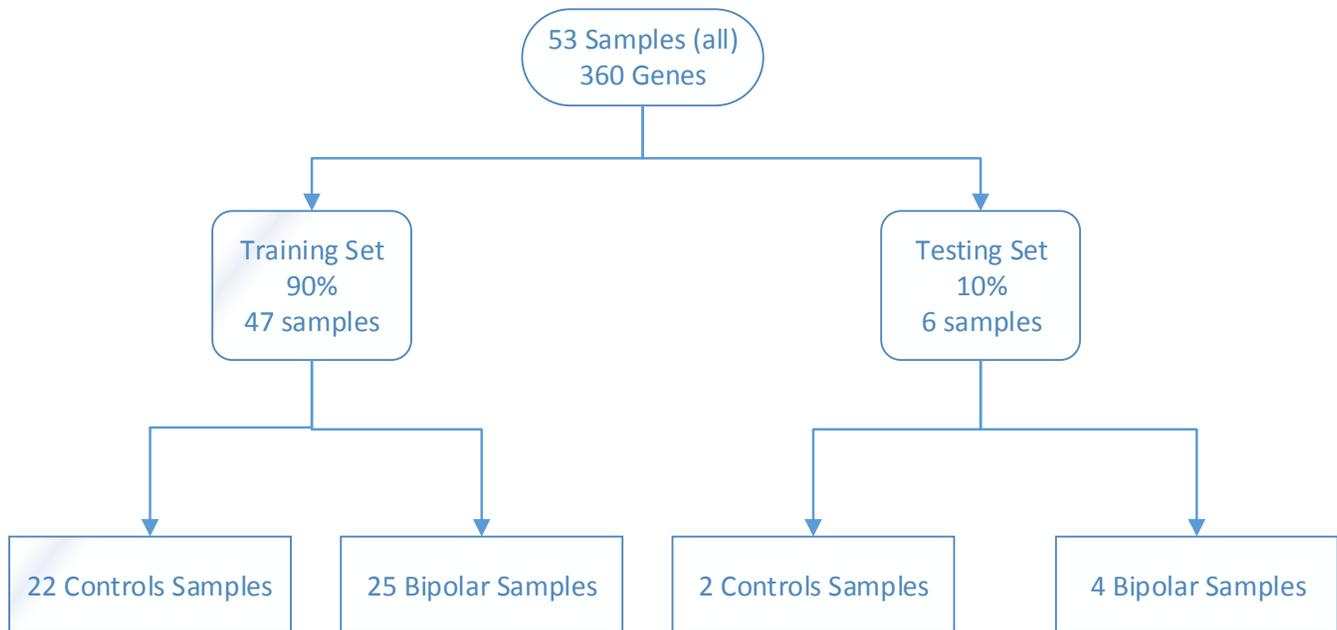
The SBV [4] approach performs robust estimates for the classification accuracy and the size of the genomic signature, extracted from a pair of feature selection subset and classification methods, on batches of bootstrap datasets, which has the same size, called “bootstrap window”  $B$ . After the SVM approach is run for a sufficient number of bootstrap windows, stabilizing the classification accuracy and genomic signature the procedure terminates and returns the stable performance estimates.

### SBV parameters

First of all, the bootstrap window  $B$  of SBV was set to 50 ( $B$ ) bootstrap datasets, the accuracy threshold was set to 0.02 ( $acc_{thresh}$ ) and the signature size threshold was set to 0.1 ( $gen_{thresh}$ ). As already mentioned, each bootstrap dataset have the same size as the original dataset and is divided into a training (90%) and a test set (10%). The SBV method was set to pause if no convergence had taken place at 1000 bootstrap datasets, a scenario that never took place as all methods converged at most 200 bootstrap datasets. In that manner, the results of the aforementioned procedure for the two different genomic signatures are shown in figure 4.7 and 4.8, respectively.



**Figure 4.7:** Structure of the bootstrap datasets used in the first significant set (673genes).



**Figure 4.8:** Structure of the bootstrap datasets used in the second significant set (360genes).

### 4.3.1 RFE and LASSO parameters

#### LASSO and SAM

The observed 673 - gene signature has been constructed using the SAM, which is a univariate selection method. The aim is to strengthen the SAM thesis, producing a reliable set of significant genes, which is easily to be assessed biologically. The solution is to combine subset feature selection and classification. Thus, LASSO regression is selected to improve model discrimination performance. As already mentioned, an advantage of this approach is that it produces interpretable models by setting a considerable amount of features at exactly zero. These represent genes that have no discriminatory power between the two classes, while those with nonzero coefficients represent genes that can separate classes of bipolar disorders successfully. LASSO tends to keep a large number of features, resulting in a genomic signature of large size, while accomplishes good classification accuracy. It also achieves similar discrimination performance to SAM, having a large number of common genes, specifically 531 common genes. Finally it requires a reasonable amount of running time and recursive feature elimination (RFE) was implemented in association with the embedded feature selection of the LASSO.

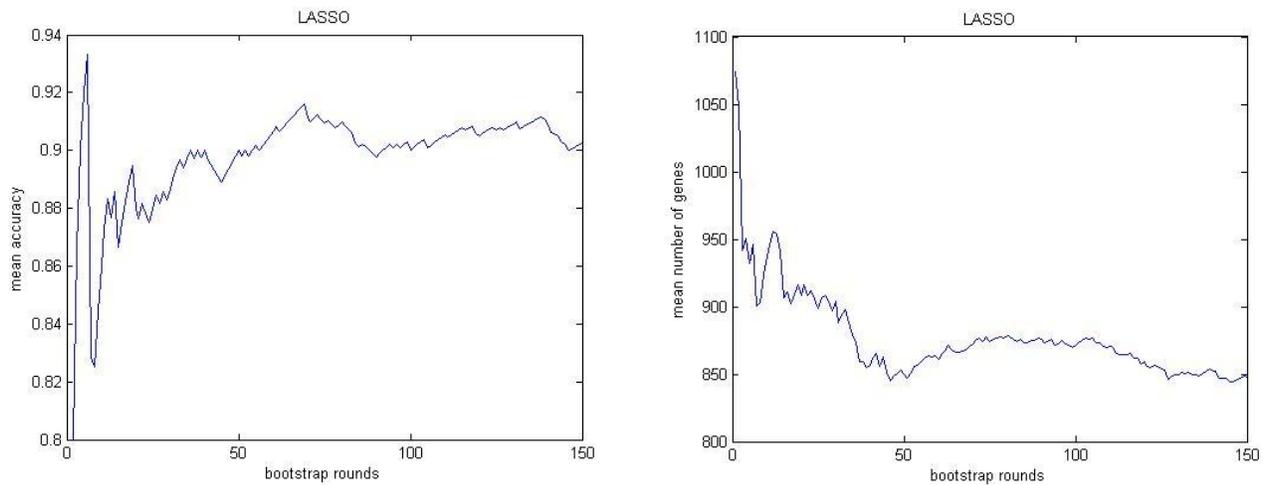
As mentioned in the section 2.1.5.3 the tuning parameter  $t$  was expressed as

$$t = a \cdot \sum_{k=1}^K w_0^2, a \in [0,1]$$

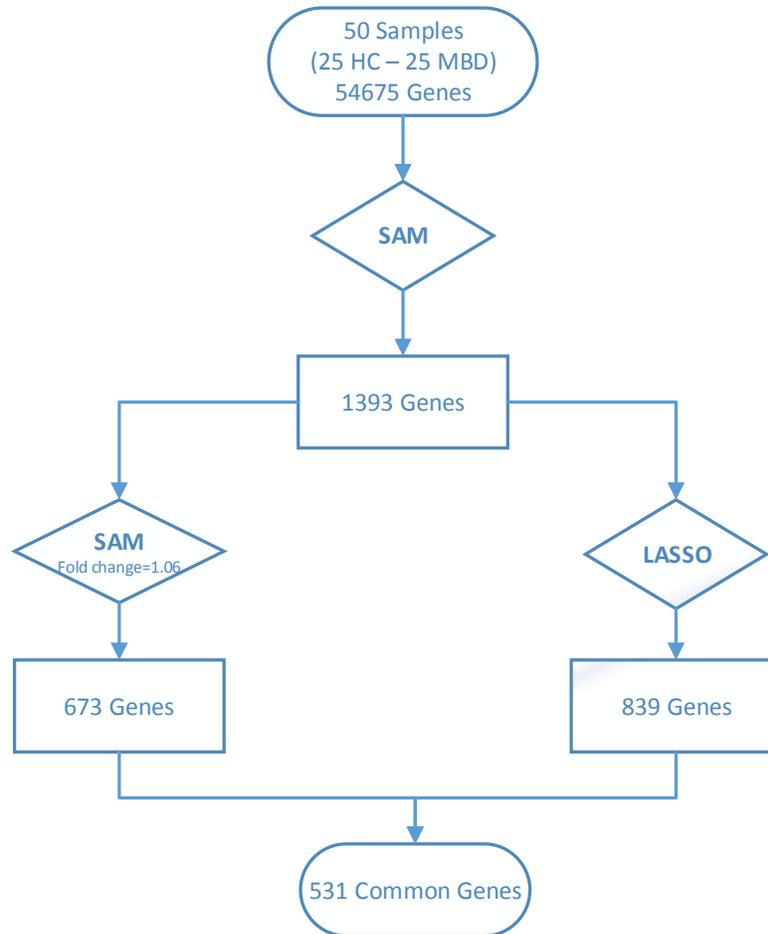
and estimated using 3 different executions 10-Fold CV on the original dataset. The value of  $\alpha=0.3$  proved to be best for classification performance of LASSO methods.

Classification Accuracy (%)	Genomic Signature Size	Time per bootstrap dataset (sec)
89.9	824	170.93

**Table 4.5:** SBV results of LASSO.



**Figure 4.9:** Left: Stabilization of LASSO mean accuracy over all bootstrap datasets  
Right: Stabilization of LASSO mean signature size over all bootstrap datasets



**Figure 4.10:** Structure of the SAM as well as LASSO results

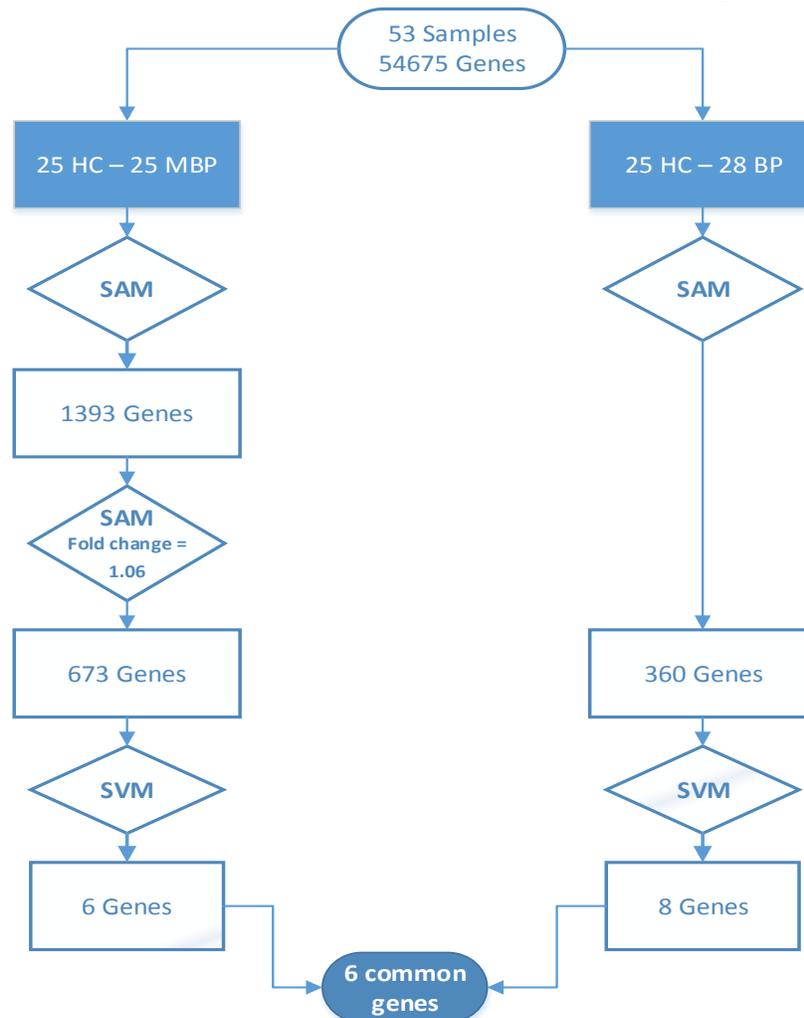
## 4.3.2 Classifier Results

As mentioned in the section 4.2, the SAM method results in two sets of significant genes from two different populations (25 HC – 25 MBD and 25 HC – 28 BD). The two populations have a large number of samples in common, as shown in section 5.3.2.2. Thus, instead of assessing the genomic signature of each population separately, the aim is to implement a unifying approach, comparing the observed results.

### 4.3.2.1 RFE and SVM results

The classification accuracy of the deterministic SVM algorithm is good enough, while resulting in considerable small genomic signatures size. Particularly, for the first group (25 Healthy Controls – 25 Medicated Bipolar Patients) the SVM method achieves accuracy of approximately 87% while the genomic signatures consist of a considerably

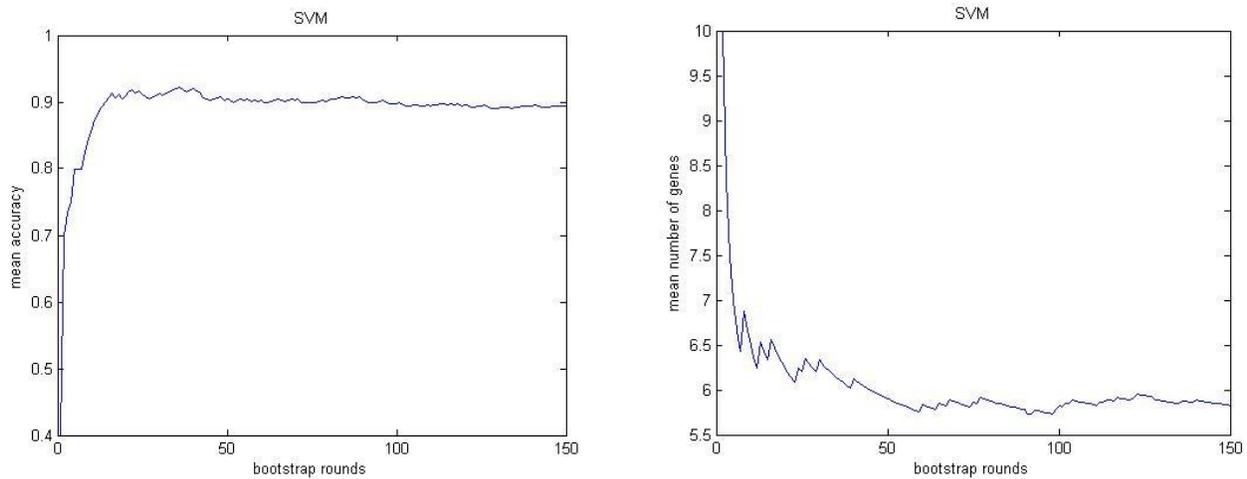
small number of genes, specifically 6 genes, shown in table 4.6. For the second group (25 Healthy Controls – 28 Bipolar Patients) the SVM classifier reaches accuracy of 89% for 8 genes selected, shown in table 4.7. Moreover, the two groups have 6 genes in common, shown in figure 4.11. The percentages mentioned above show that the statistical performance of the SVM classifier, although uses a very small number of significant genes, is good enough. But our main goal is to combine the statistical with the biological significance of the observed genomic signatures in order to extract a model which reflects the underlying biological system of the disease. That leads to the idea of using the RVM classifier, which is a probabilistic algorithm which stands for an improved prediction performance, shown in section 4.3.2.2. Moreover, SVM method requires a moderate amount of running time, while the RFE was implemented in association with the embedded feature selection of the SVM classifier. The results of the SVM procedure for the two different populations are shown in figure 4.11.



**Figure 4.11:** Structure of the SVM results

Classification Accuracy (%)	Genomic Signature Size	Time per bootstrap dataset (sec)
89.5	6	405.99

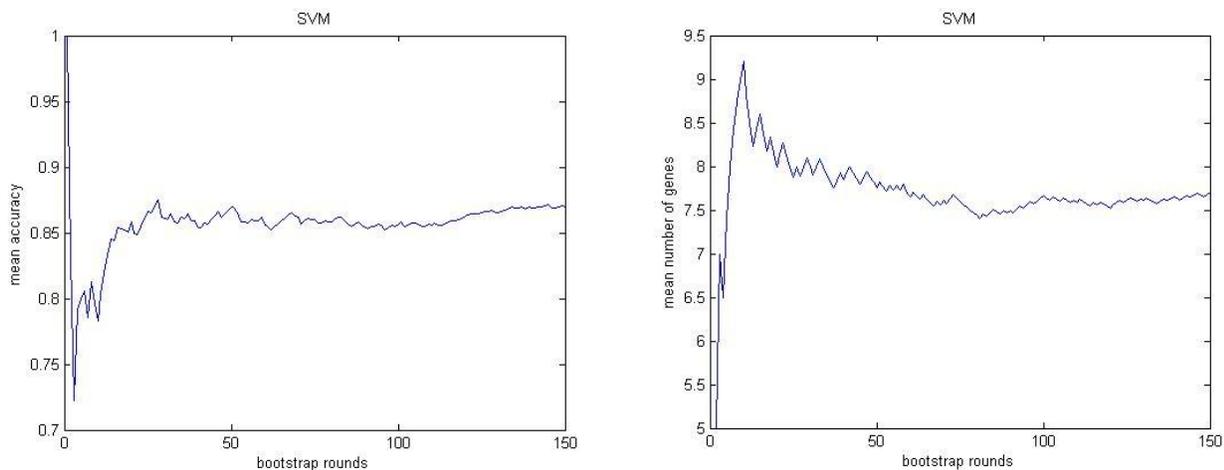
**Table 4.6:** SBV results of SVM classifier for 673 significant genes.



**Figure 4.12:** Left: Stabilization of SVM mean accuracy of 6 significant genes over all bootstrap datasets  
Right: Stabilization of SVM mean signature size of 6 significant genes over all bootstrap datasets

Classification Accuracy (%)	Genomic Signature Size	Time per bootstrap dataset (sec)
87	8	323.08

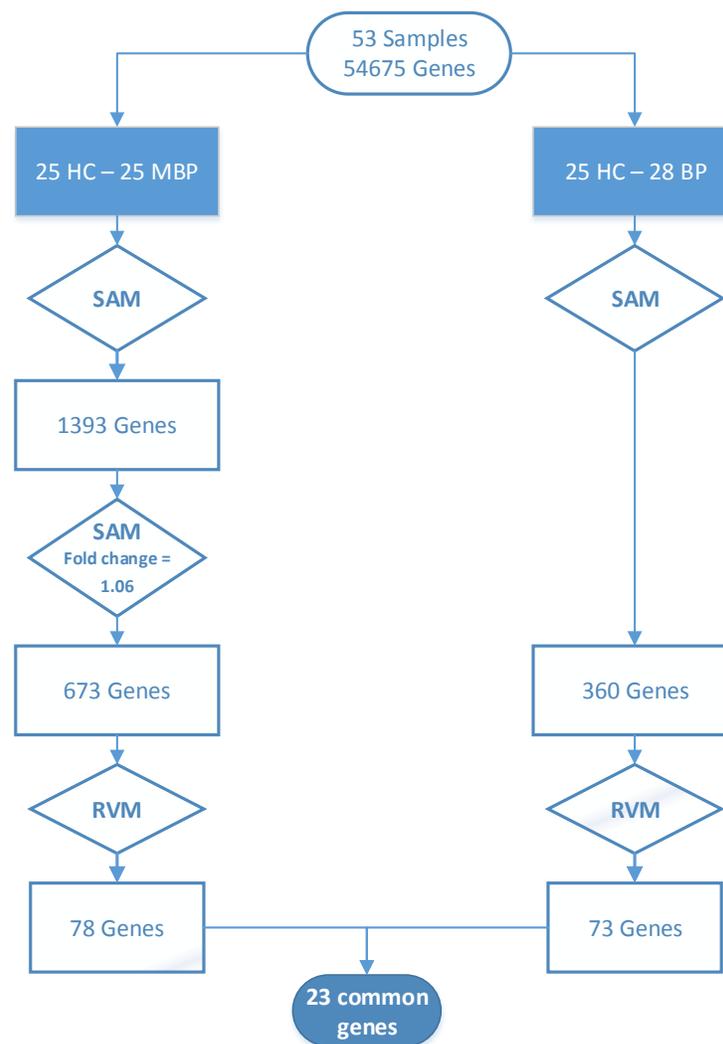
**Table 4.7:** SBV results of SVM classifier for 360 significant genes.



**Figure 4.13:** Left: Stabilization of SVM mean accuracy of 8 significant genes over all bootstrap datasets  
Right: Stabilization of SVM mean signature size of 8 significant genes over all bootstrap datasets

### 4.3.2.2 RFE and RVM results

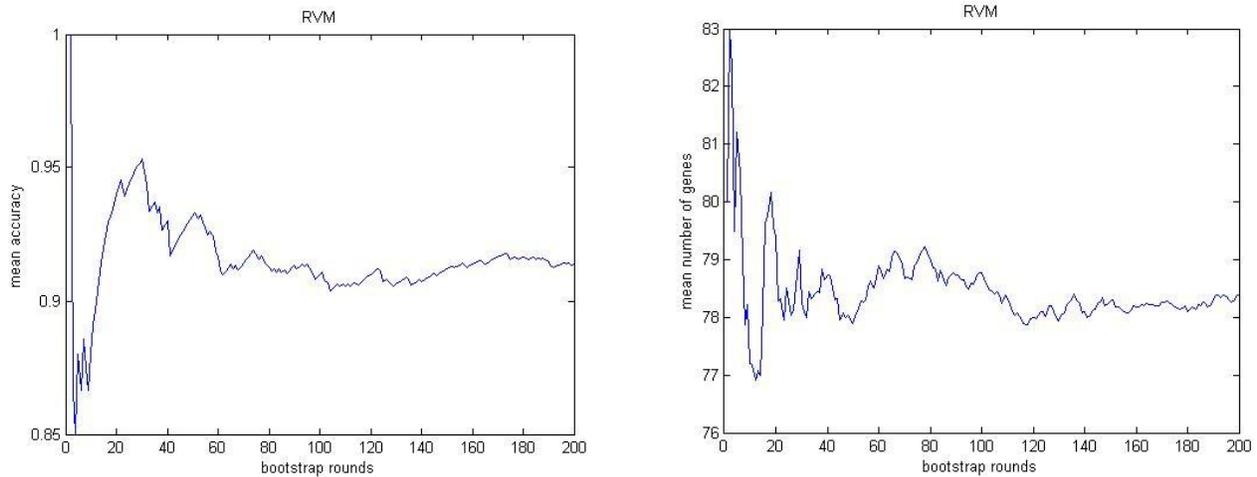
RVM method achieves similar classification accuracy to SVM approach, while the resulting genomic signatures as well as running time are considerably larger in size. Particularly, compared to SVM, for the first group (25 Healthy Controls – 25 Medicated Bipolar Patients) RVM reached accuracy of 91.4% for 78 genes selected, shown in table 4.8, while for the second group (25 Healthy Controls – 28 Bipolar Patients) achieves accuracy of 90% for 73 genes selected, shown in table 4.9. In that manner, it leads to a more easily interpretable model, but it requires an excessive amount of running time. RFE was also implemented in association with the embedded feature selection of the RVM classifier. The RVM procedure is repeated 100 times for each different set of significant genes and the overall results are averaged. The observed results of two different groups have 23 common genes, which are shown in figure 4.14.



**Figure 4.14** Structure of the RVM results - 23 common genes

Classification Accuracy (%)	Genomic Signature Size	Time per bootstrap dataset (sec)
91.4	78	5609.22

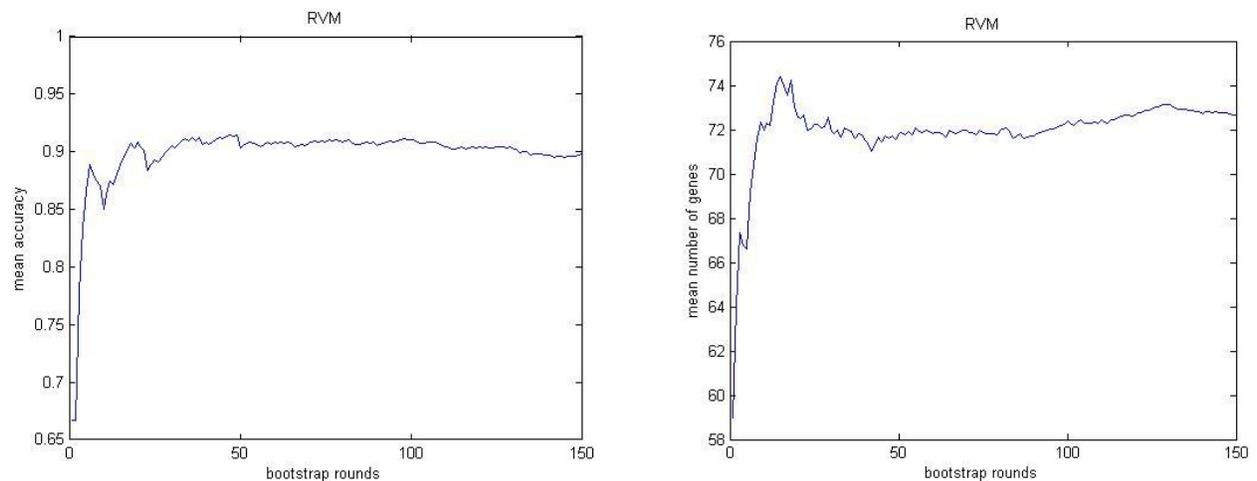
**Table 4.8:** SBV results of RVM classifier from 673 significant genes.



**Figure 4.15:** Left: Stabilization of SVM mean accuracy of 78 significant genes over all bootstrap datasets  
Right: Stabilization of SVM mean signature size of 78 significant genes over all bootstrap datasets

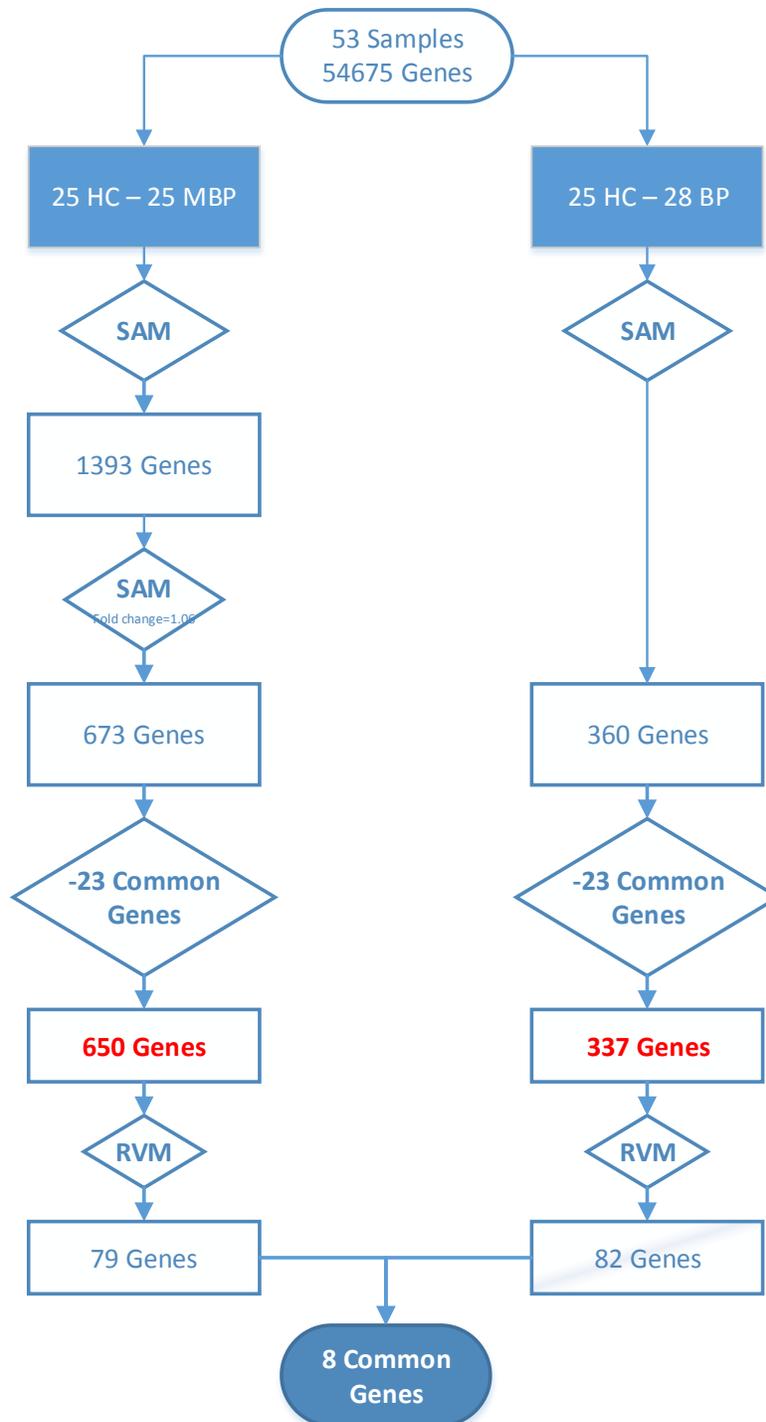
Classification Accuracy (%)	Genomic Signature Size	Time per bootstrap dataset (sec)
90	73	1220.01

**Table 4.9:** SBV results of RVM classifier from 360 significant genes.



**Figure 4.16:** Left: Stabilization of SVM mean accuracy of 73 significant genes over all bootstrap datasets  
Right: Stabilization of SVM mean signature size of 73 significant genes over all bootstrap dataset.

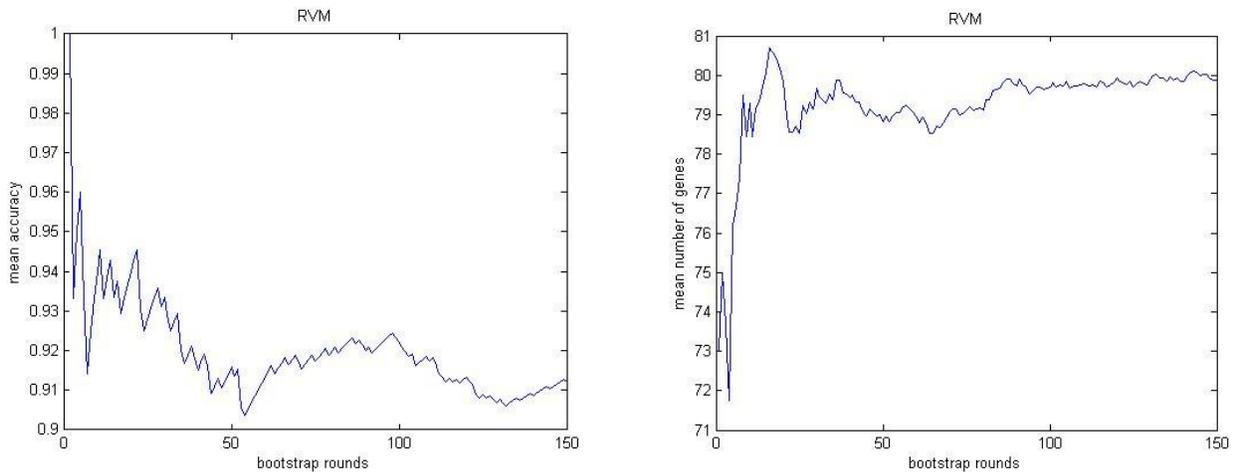
Then, the 23 observed common genes are removed from the initial sets of significant genes and the RVM procedure is repeated 100 times for each significant set and the overall results are averaged. The observed results also have 8 common genes, shown in figure 4.17.



**Figure 4.17:** Structure of the RVM results - 8 common genes

Classification Accuracy (%)	Genomic Signature Size	Time per bootstrap dataset (sec)
91.2	79	4218.89

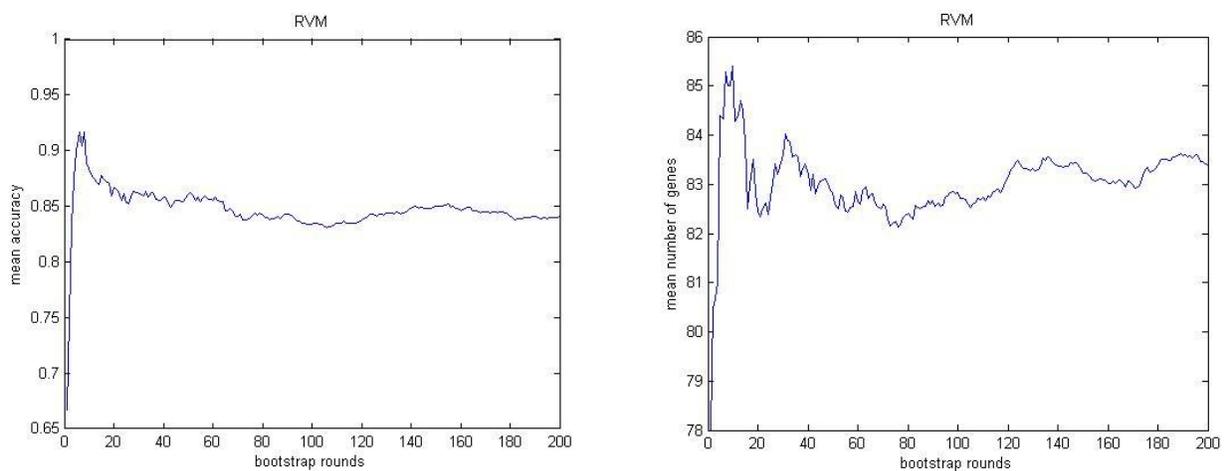
**Table 4.10:** SBV results of RVM classifier from 650 significant genes.



**Figure 4.18:** Left: Stabilization of SVM mean accuracy of 79 significant genes over all bootstrap datasets  
 Right: Stabilization of SVM mean signature size of 79 significant genes over all bootstrap datasets

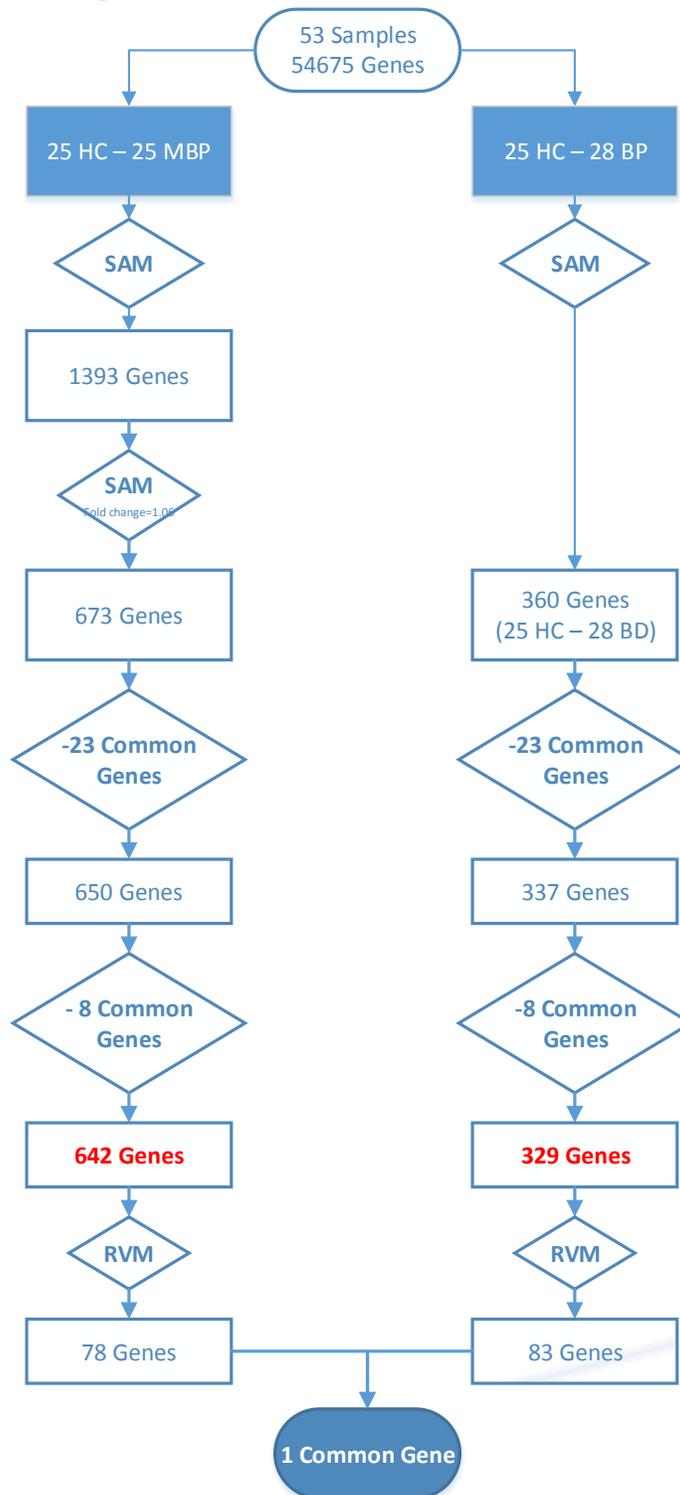
Classification Accuracy (%)	Genomic Signature Size	Time per bootstrap dataset (sec)
85.2	82	861.92

**Table 4.11:** SBV results of RVM classifier from 337 significant genes.



**Figure 4.19:** Left: Stabilization of SVM mean accuracy of 82 significant genes over all bootstrap datasets  
 Right: Stabilization of SVM mean signature size of 82 significant genes over all bootstrap dataset

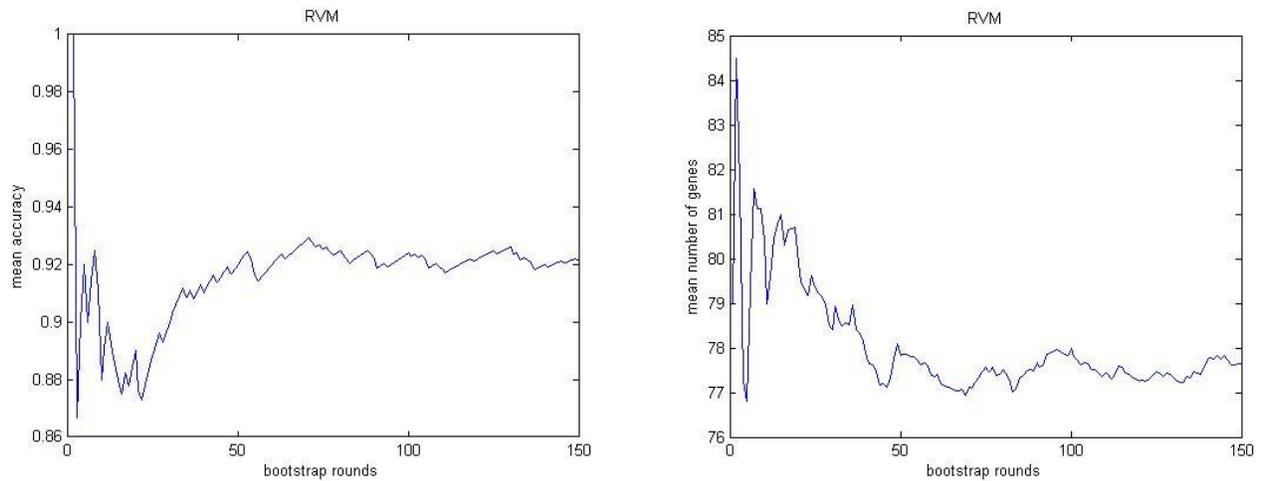
Consequently, the aforementioned procedure is repeated again, removing the 8 common genes from the rest sets of significant genes. The observed results also have 1 common genes, shown in figure 4.20.



**Figure 4.20:** Structure of the RVM results - 1 common gene

Classification Accuracy (%)	Genomic Signature Size	Time per bootstrap dataset (sec)
92.1	78	3955.70

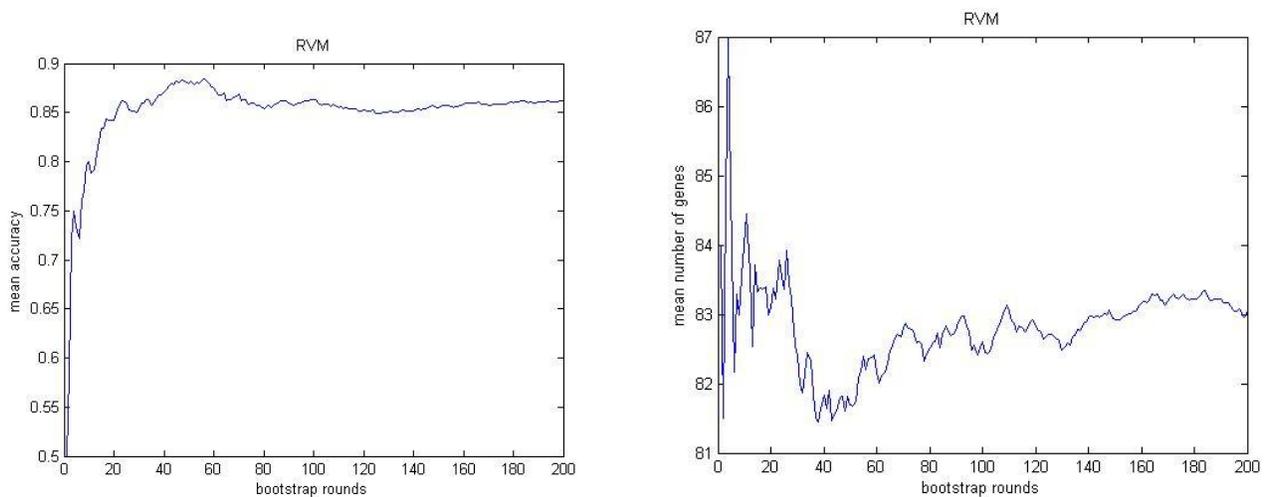
**Table 4.12:** SBV results of RVM classifier from 642 significant genes.



**Figure 4.21:** Left: Stabilization of SVM mean accuracy of 78 significant genes over all bootstrap datasets  
Right: Stabilization of SVM mean signature size of 78 significant genes over all bootstrap datasets

Classification Accuracy (%)	Genomic Signature Size	Time per bootstrap dataset (sec)
86.2	83	858.21

**Table 4.13:** SBV results of RVM classifier from 329 significant genes.



**Figure 4.22:** Left: Stabilization of SVM mean accuracy of 82 significant genes over all bootstrap datasets  
Right: Stabilization of SVM mean signature size of 82 significant genes over all bootstrap dataset.

## 4.4 Evaluation Results

### 4.4.1 Classification Accuracy Comparison

In the case of classification accuracy, the RVM approach outperformed the other classifiers, reaching accuracies of 90%. Moreover RVM extracted reliable genomic signatures and lead to models which are easily to access biologically. However, the execution time of RVM methods is 10 times larger than that of SVM methods and 20 to 40 times larger than that of LASSO. The LASSO classifier was second in terms of classification accuracy but it lead to an approximately ten times larger genomic signature than RVM approach. Finally, concerning SVM method achieves good classification accuracy but it kept a relatively small number of features.

Method	Groups	Original Genomic Signature	Classification Accuracy	Genomic Signature Size	Time/bootstr ap
<b>LASSO</b>	25 HC-25 MBD	1393	<b>89.9</b>	824	170.93
<b>SVM</b>	25 HC-25 MBD	673	<b>89.5</b>	6	405.99
<b>RVM</b>	25 HC-25 MBD	673	<b>91.4</b>	78	5609.22
<b>RVM</b>	25 HC-25 MBD	650	<b>91.2</b>	79	4218.89
<b>RVM</b>	25 HC-25 MBD	642	<b>92.1</b>	78	3955.70

**Table 4.14a:** Synopsis of SBV results from Healthy Control (25 HC) – Medicated Bipolar Disorder (25 MBD) samples.

Method	Groups	Original Genomic Signature	Classification Accuracy	Genomic Signature Size	Time/bootstr ap
<b>SVM</b>	25 HC- 28 BD	360	<b>87</b>	8	323.08
<b>RVM</b>	25 HC- 28 BD	360	<b>90</b>	73	1220.01
<b>RVM</b>	25 HC- 28 BD	347	<b>85.2</b>	82	861.92
<b>RVM</b>	25 HC- 28 BD	329	<b>86.2</b>	83	858.21

**Table 4.14b:** Synopsis of SBV results from Healthy Control (25 HC) –Bipolar Disorder (28 BD) samples.

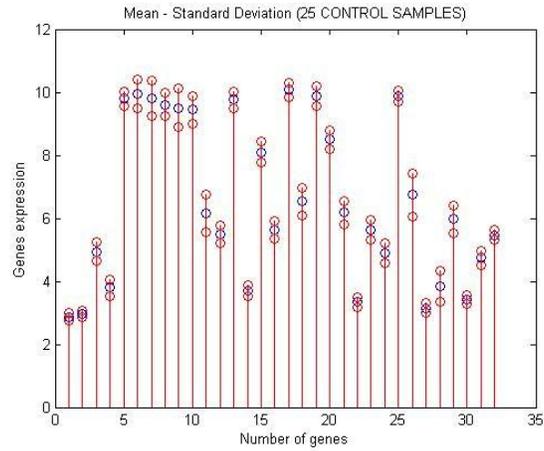
## 4.4.2 Genomic Signature Significance

### 4.4.2.1 Unifying the Genomic Signatures

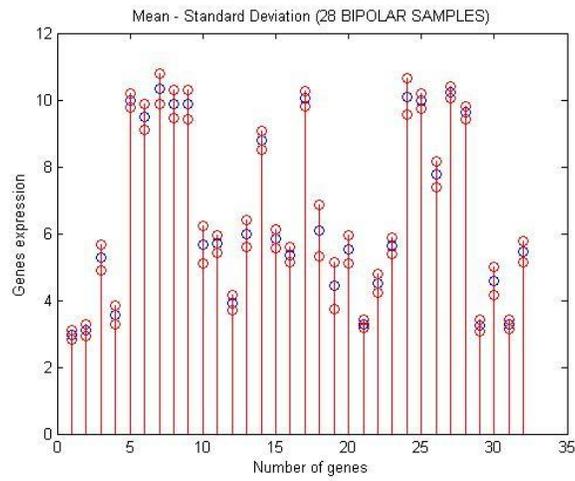
After assessing the genomic signature of each method separately, a unifying approach was implemented. Given the fact that the size of the genomic signature of the SVM approach is considerably smaller than the one of the RVM method, the SVM method is made practically unusable. On the other hand, the common genes existing in the signatures of all RVM procedures were selected as the unified common gene signature. Since there are 3 difference cases used for the RVM methods, 3 different unified signatures were extracted, the 23 gene, 8 gene and 1 gene signatures. In that manner, the final genomic signature is composed of 32 genes.

### 4.4.2.2 Discrimination of Genomic Signature

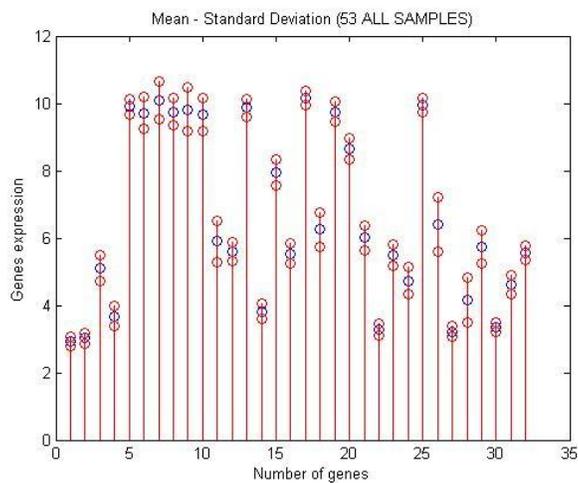
Resulting in the final genomic signature, the expression value of each gene should be examined in order to access the discrimination between the class labels. Thus, the standard deviation and mean of each gene are extracted, shown in figures 4.23, 4.24, 4.25. Then, we examined the genomic-wide expression variance distributions between the groups, shown in table 4.15. The variance and the closely-related standard deviation are measures of how spread out a distribution is. In other words, they are measures of variability. Particularly, a variance of zero indicates that all the values are identical. A small variance indicates that the genes expressions tend to be very close to the mean and hence to each other, while a high variance indicates that the genes expressions are very spread around the mean and from each other.



**Figure 4.23:** Mean – Standard Deviation of 25 healthy control samples of 32 genes



**Figure 4.24:** Mean – Standard Deviation of 28 bipolar samples of 32 genes



**Figure 4.25:** Mean – Standard Deviation of 53 samples of 32 genes

	Gene Name	Variance - Control	Variance - BD	Variance - All
1	1553864_at	0.0154	0.0175	0.0186
2	1557217_a_at	0.0109	0.0318	0.0268
3	1559117_at	0.0922	0.1511	0.1481
4	1559203_s_at	0.0714	0.0834	0.0893
5	201164_s_at	0.0478	0.0456	0.0550
6	203392_s_at	0.2056	0.1531	0.2280
7	205285_s_at	0.3143	0.1989	0.3162
8	206059_at	0.1393	0.1636	0.1675
9	208965_s_at	0.3604	0.1927	0.4060
10	211794_at	0.1972	0.3109	0.2376
11	212730_at	0.3420	0.0687	0.3791
12	213455_at	0.0783	0.0509	0.0820
13	213729_at	0.0645	0.1632	0.0689
14	217000_at	0.0313	0.0791	0.0530
15	218561_s_at	0.1067	0.0828	0.1503
16	219805_at	0.0818	0.0570	0.0946
17	220761_s_at	0.0597	0.0465	0.0468
18	221648_s_at	0.1907	0.6010	0.2466
19	222409_at	0.1006	0.7800	0.0836
20	223135_s_at	0.0870	0.1843	0.1016
21	224522_s_at	0.1313	0.0157	0.1342
22	224545_at	0.0253	0.0789	0.0301
23	228696_at	0.0988	0.0531	0.0937
24	230185_at	0.0991	0.2890	0.1671
25	231716_at	0.0306	0.0537	0.0464
26	231798_at	0.4861	0.1431	0.6473
27	234765_at	0.0254	0.0309	0.0274
28	235216_at	0.2371	0.0352	0.4476
29	236398_s_at	0.2033	0.0323	0.2387
30	237145_at	0.0216	0.1833	0.0220
31	238682_at	0.0505	0.0255	0.0800
32	244326_at	0.0233	0.0967	0.049

**Table 4.15:** Variance of 32 significant genes.

The first column contains number of each gene. The second one includes the name of each gene. While the others contain the variance of control samples, bipolar disorder samples and all samples,

respectively. Bold are the genes with the higher variance. Red are the genes that have the higher variance among the groups

The standard deviation, in association with mean, can show what is normal and what is under or over expressed, concerning the mean of expression values of each gene. According to the plots, we observed that in all cases the average mean is approximately 6.5. Thus, as can be seen, concerning the standard deviation there are significant difference between the control and BD group and the group of all samples. Particularly, in the first group the measures are mainly expressed under the mean, while in the group of all samples the measures are over the mean. This indicates that among each group, expression values of genes tend to be close to each other, while among the group of all samples (bipolar disorder and control samples) the expression values tend to spread below and above the mean.

Moreover, according to the table 4.15 we observed that the group of all samples also presents the higher variance among the groups. As already mentioned variance describes how much a random variable differs from its expected value. Thus, concerning the variance of each gene, we noted that the group of control samples contains low - variance genes, while the group of all samples contains high – variance genes. The genes with the higher variance among the groups are; 203392\_s\_at, 205285\_s\_at, 208965\_s\_at, 212730\_at, 221648\_s\_at, 231798\_at, 235216\_at, 236398\_s\_at. The expressions values of these genes are very spread around the mean and from each other.

According to the above plots and table, we came to the conclusion that the standard deviation as well as variance of 32 significant genes of all samples are far off null, indicating that the expression value of each gene tend to be far off the mean and hence to each other. In that manner, the genes expressions are extended over a wider range of values, representing high discrimination among the class of bipolar patients and healthy controls samples and leading to a more easily interpretable model.

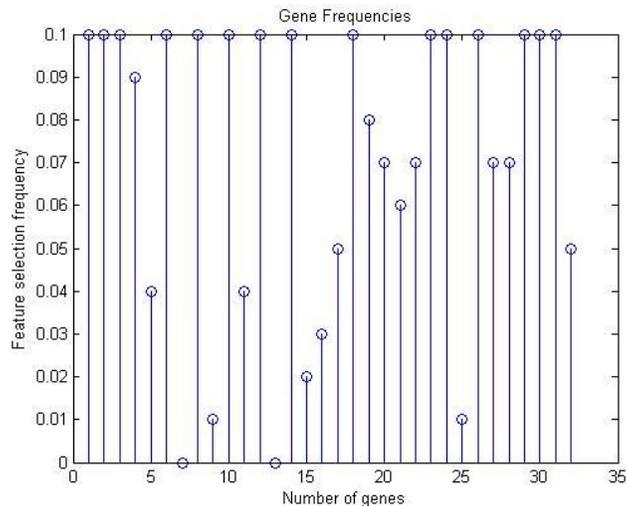
#### **4.4.2.3 Consistency of Gene Selection in the Signature**

As mentioned in section 3.3.3 the consistency of gene selection in the signature refers to the reliability of genomic signature. Specifically, it refers to the ability of the genomic signature to yield similar performance when applied on a single test set multiple times, while using different training sets. The final genomic signature, which is composed of 32 significant genes, was used. As presented in section 3.3.3, the 10 fold cross validation approach generates 9 training datasets and only one test set. This process is repeated 10 rounds. In each round, one of the folds is used for validation, and the other

9 folds for training. Then the RVM classification method is performed. The process is repeated 200 times and the overall results are averaged.

Genomic Signature Size	Mean Classification Accuracy (%)	Mean Genes	Variance
<b>32</b>	84.67	25	0.0318

**Table 4.16:** Consistency of Gene Selection in the Signature



**Figure 4.26:** Frequencies of 32 significant genes

According to the above observations, when using the RVM classifier the genomic signature lead to consistent results when applied multi times on one test set. Particularly, the 32 genomic signature achieves good classification performance of the RVM method and small variance of the observed classification accuracy, leading to a good signature consistency.

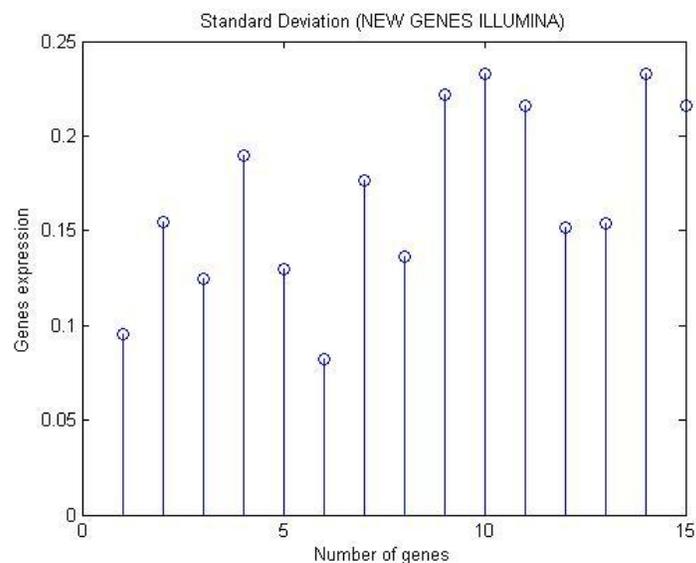
#### 4.4.2.4 Generalization Ability of Genomic Signature

##### New Dataset

The new dataset results from measurements of peripheral blood mononuclear cells (PBMC). Peripheral blood mononuclear cells from whole blood were collected from 8 patients with bipolar and 24 adult healthy control subjects. Thus the original dataset (GEO access number: GSE39653) consists of 32 samples related to bipolar disorder, 24 of which correspond to healthy control and 8 to bipolar samples. For each sample, there are measurements of 43117 genes.

## Generalization Ability

As already mentioned in section 3.3.4, the aim of this field is to access the generalization ability of the genomic signature. A good generalization performance is achieved when a genomic signature is able to predict the label of unseen samples correctly. The final genomic signature consists of 32 significant genes. Thus, these genes are selected from the new dataset in order to be used for accessing the generalization ability of the model to an independent dataset. However, there are 6 genes which are not detected in the new dataset, while 7 of them are appeared with more than one code and the other 19 have the same code to the original dataset. In that manner, concerning the 7 genes with multiply codes, standard deviation of each gene is extracted in order to decide which code is able to be used. The gene with the higher standard deviation is preferred, since these genes are extended over a wider range of values.



**Figure 4.27:** Standard Deviation of 7 multiply genes of the new dataset

Thus, the genomic signature of the new dataset is composed of 26 significant genes and is used to access the generalization ability of the model. As mentioned in section 3.3.3, the 10 fold cross validation approach generates 9 training datasets and only one test set. This process is repeated 10 rounds. In each round, one of the folds is used for validation, and the other 9 folds for training. Then the RVM classification method is performed. The process is repeated 200 times and the overall results are averaged.

Genomic Signature Size (new dataset)	Mean Classification Accuracy (%)	Mean Genes
26	74.17	13

**Table 4.17:** Generalization Ability of Genomic Signature Results

The observed mean classification accuracy is good enough, performing very good generalization performance when it comes to the classification of unknown samples.

## 4.5. Biological Evaluation

As presented in the methodology section, for the classification purpose, we used the dataset GSE46449 [42] obtained from GEO (Gene Expression Omnibus) repository [43], while the GEO Dataset GSE39653 has been used in order to evaluate the proposed methodology [44].

As shown in Table 4.18, the probe identifiers from the unified “32 gene signature” were mapped to unique Gene Symbols and Entrez Gene Ids, upon which pathway analysis has been performed.

Nr.	Affymetrix Probe Set ID	Entrez Gene Id	Gene Symbol	Description
1	1553864_at	N/A	N/A	Unknown
2	1557217_a_at	2187	FANCB	Fanconi anemia, complementation group B
3	1559117_at	N/A	N/A	Unknown
4	1559203_s_at	3845	KRAS	v-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog
5	201164_s_at	9698	PUM1	pumilio homolog 1 (Drosophila)
6	<b>203392_s_at</b>	<b>1487</b>	<b>CTBP1</b>	<b>C-terminal binding protein 1</b>
7	<b>205285_s_at</b>	<b>2533</b>	<b>FYB</b>	<b>FYN binding protein</b>
8	206059_at	7644	ZNF91	zinc finger protein 91
9	<b>208965_s_at</b>	<b>3428</b>	<b>IFI16</b>	<b>interferon, gamma-inducible protein 16</b>
10	211794_at	2533	FYB	FYN binding protein
11	<b>212730_at</b>	<b>23336</b>	<b>SYNM</b>	<b>synemin, intermediate filament</b>

				<b>protein</b>
<b>12</b>	213455_at	92689	FAM114A1	family with sequence similarity 114, member A1
<b>13</b>	213729_at	55660	PRPF40A	PRP40 pre-mRNA processing factor 40 homolog A (S. cerevisiae)
<b>14</b>	217000_at	442236	KRT18P50	keratin 18 pseudogene 50
<b>15</b>	218561_s_at	57128	LYRM4	LYR motif containing 4
<b>16</b>	219805_at	63932	CXorf56	chromosome X open reading frame 56
<b>17</b>	220761_s_at	51347	TAOK3	TAO kinase 3
<b>18</b>	<b>221648_s_at</b>	<b>N/A</b>	<b>N/A</b>	<b>Unknown</b>
<b>19</b>	222409_at	23603	CORO1C	coronin, actin binding protein, 1C
<b>20</b>	223135_s_at	56987	BBX	bobby sox homolog (Drosophila)
<b>21</b>	224522_s_at	79877	DCAKD	dephospho-CoA kinase domain containing
<b>22</b>	224545_at	N/A	N/A	Unknown
<b>23</b>	228696_at	85414	SLC45A3	solute carrier family 45, member 3
<b>24</b>	230185_at	79725	THAP9	THAP domain containing 9
<b>25</b>	231716_at	54542	RC3H2	ring finger and CCCH-type domains 2
<b>26</b>	<b>231798_at</b>	<b>9241</b>	<b>NOG</b>	<b>noggin</b>
<b>27</b>	234765_at	N/A	N/A	Unknown
<b>28</b>	<b>235216_at</b>	<b>114799</b>	<b>ESCO1</b>	<b>establishment of cohesion 1 homolog 1 (S. cerevisiae)</b>
<b>29</b>	236398_s_at	N/A	N/A	Unknown
<b>30</b>	237145_at	440275	EIF2AK4	eukaryotic translation initiation factor 2 alpha kinase 4
<b>31</b>	238682_at	257236	CCDC96	coiled-coil domain containing 96
<b>32</b>	244326_at	N/A	N/A	Unknown

**Table 4.18:** Mapping of Probe Set IDs to Gene Symbols and Entrez Gene IDs. Red highlighted are the eight genes with the highest variance among the groups. Purple highlighted is the gene NOG known for its association with BPD

Twenty-five probes - designed to interrogate a given sequence - were successfully mapped, whereas seven probes were not mapped. The biological significance underlying the unified “32 gene signature” was explored by enrichment analysis, while a gene-disease association within the signature was searched in the Database BDgene [45]. Moreover, aiming at a functional enrichment of KEGG pathways and biological processes in terms of Gene Ontology (GO), we utilized the annotation tool GATHER, which “*integrates various forms of available data to elucidate biological context within molecular signatures produced from high-throughput post-genomic assays*” [46]. GATHER uses the hypergeometric distribution or chi-square test in order to assign a p-value of >0.05 to genes that are important within the examined gene signature.

Pathways (KEGG IDs)	Genes	p Value
Ethylbenzene degradation (path:hsa00642)	ESCO1	0.001
Alkaloid biosynthesis II (path:hsa00960)	ESCO1	0.001
1- and 2-Methylnaphthalene degradation (path:hsa00624)	ESCO1	0.002
Phenylalanine metabolism (path:hsa00360)	ESCO1	0.003
Limonene and pinene degradation (path:hsa00903)	ESCO1	0.003
Valine, leucine and isoleucine degradation (path:hsa00280)	ESCO1	0.004
Notch signaling pathway (path:hsa04330)	CTBP1	0.004
Histidine metabolism (path:hsa00340)	ESCO1	0.005
Butanoate metabolism (path:hsa00650)	ESCO1	0.005
Lysine degradation (path:hsa00310)	ESCO1	0.005
Tyrosine metabolism (path:hsa00350)	ESCO1	0.006
Benzoate degradation via CoA ligation (path:hsa00632)	ESCO1	0.006
Glycerophospholipid metabolism (path:hsa00564)	ESCO1	0.007
TGF-beta signaling pathway (path:hsa04350)	NOG	0.008
Gap junction (path:hsa04540)	KRAS	0.008
Tight junction (path:hsa04540)	KRAS	0.01
Insulin signaling pathway (path:hsa04910)	KRAS	0.01
Wnt signaling pathway (path:hsa04310)	CTBP1	0.01

**Table 4.19:** Enriched pathways by GATHER

Biological Process (Gene Ontology IDs)	Genes	p Value
negative regulation of JNK cascade (GO:0046329)	TAOK3	0.001
positive regulation of JNK cascade (GO:0046330)	TAOK3	0.002
L-serine biosynthesis (GO:0006564)	CTBP1	0.004
regulation of JNK cascade (GO:0046328)	TAOK3	0.005
negative regulation of cell differentiation (GO:0045596)	NOG	0.006
autophosphorylation (GO:0046777)	TAOK3	0.006
NLS-bearing substrate-nucleus import (GO:0006607)	FYB	0.006
monocyte differentiation (GO:0030224)	IFI16	0.007
regulation of biological process (GO:0050789)	BBX CTBP1 IFI16 NOG PUM1 TAOK3 ZNF91	0.007
L-serine metabolism (GO:0006563)	CTBP1	0.007
protein amino acid phosphorylation (GO:0006468)	CTBP1 FYB TAOK3	0.008
serine family amino acid biosynthesis (GO:0009070)	CTBP1	0.008
myeloid blood cell differentiation (GO:0030099)	IFI16	0.01
cell differentiation (GO:0030154)	IFI16 NOG	0.01
viral genome replication (GO:0019079)	CTBP1	0.01
phagocytosis (GO:0006909)	CORO1C	0.01
negative regulation of signal transduction (GO:0009968)	TAOK3	0.01
phosphorylation (GO:0016310)	CTBP1 FYB TAOK3	0.01
protein kinase cascade (GO:0007243)	FYB TAOK3	0.01
mRNA metabolism (GO:0016071)	FNBP3 PUM1	0.01
viral infectious cycle (GO:0019058)	CTBP1	0.02
serine family amino acid metabolism (GO:0009069)	CTBP1	0.02
negative regulation of development (GO:0051093)	NOG	0.02
phosphorus metabolism (GO:0006793)	CTBP1 FYB TAOK3	0.02
phosphate metabolism (GO:0006796)	CTBP1 FYB TAOK3	0.02
JNK cascade (GO:0007254)	TAOK3	0.02
regulation of cellular process (GO:0050794)	CTBP1 NOG TAOK3	0.02
response to virus (GO:0009615)	IFI16	0.02
viral life cycle (GO:0016032)	CTBP1	0.02
regulation of cell differentiation (GO:0045595)	NOG	0.02
protein modification (GO:0006464)	CTBP1 FYB MNAB TAOK3	0.02
amino acid biosynthesis (GO:0008652)	CTBP1	0.03
cellular metabolism (GO:0044237)	BBX C6orf149 CTBP1 FNBP3 FYB IFI16 MNAB PUM1 TAOK3 ZNF91	0.03
nuclear import (GO:0051170)	FYB	0.03
protein-nucleus import (GO:0006606)	FYB	0.03
intracellular signaling cascade (GO:0007242)	FYB KRAS TAOK3	0.03
regulation of translation (GO:0006445)	PUM1	0.03

**Table 4.20:** Enriched biological processes by GATHER

As presented in Tables 4.19 and 4.20, the functional enrichment by GATHER assigned GO and KEGG terms to genes based on the features of their encoded products, that were statistically over-represented within the unified “32 gene signature”. Table 4.19 presents eighteen enriched pathways and Table 4.20 thirty-seven enriched biological processes. Interestingly, most of the genes in this signature were involved in four main GO categories, namely cellular metabolism, regulation of biological process, protein modification (phosphorylation) and intracellular signaling cascade (Table 4.20), while three of the eight genes with high variance among the groups (ESCO1, NOG, CTBP1) were implicated in the enriched KEGG pathways (Table 4.19). Of note, the gene *noggin* is involved in the 10-gene predictor set for bipolar disorder reported by Clelland et al (2013) [47], and the gene *CTBP1* has been reported as a putative biomarker gene able to discriminate between schizophrenia, BPD and control samples [48]. Finally, according to BDgene database the eukaryotic translation initiation factor 2 alpha kinase 4 (*EIF2AK4*) is linked to bipolar disease.

Even with poor knowledge about the directly association of the thirty-two genes with bipolar disorder, we consider them as potential classifiers of BPD; the majority of their assigned significant processes and pathways highlighted here, are studied by researchers regarding to their important role in pathophysiology and neurodevelopment of bipolar disorder [48],[49],[50],[51],[52],[53],[54].

Our methodology enables the classification of healthy controls from patients with bipolar disorder, where is less likely to be influenced by medication. We propose that the thirty-two genes of the “32 unified common gene signature” - validated on the independent dataset GSE39653 - represent powerful genes and might be considered as prediction genes for bipolar disorder. As a final point, we notice the usage of the proposed signature for the characterization of the unmapped/unknown probes (224545\_at, 221648\_s\_at, 244326\_at, 1553864\_at, 1559117\_at, 234765\_at, 236398\_s\_at).

The proposed signature could be easily validated experimentally in peripheral blood leucocytes.

# 5

## Conclusion

The aim of this diploma thesis is to provide a reliable and stable genomic signature that classifies the bipolar disorders and underlines the genetic background of the disease. Thus, gene expressions from two different populations are used.

The genome analysis usually leads to datasets that normally contain a small number of samples which have a large number of gene expression levels as features. In order to extract useful informative sets of genes that can reduce dimensionality and maximize the performance of classifiers, feature selection algorithms were used.

While, feature selection methods are used in order to counterfeit the dimensionality of the data by keeping a relatively small set of significant features, the classification approaches are used in order to classify new data into known class of interest. Through classification approaches a small set of significant features, which achieves high classification accuracy, arised.

Furthermore, an evaluation method called “Stable Bootstrap Validation” (SBV), introduced be Nick Chlis, is presented so as to achieve stable performance assessment of feature selection and classification methods. The SBV employs bootstrap resampling of the original dataset and an explicit stability assessment criterion in order to extract stable estimates of the classification accuracy as well as the genomic signature size; the number of genes selected in the signature.

Moreover, the discrimination, consistency and generalization ability of the observed results are also evaluated. The results that are stable and reflect the biological model should also be consistent across different executions of the feature selection and classification methodologies. Also, the ability of how the results of a statistical analysis will generalize to an independent data set should be evaluated.

The above methodology is performed on a dataset that is composed of two different populations. Particularly, the original dataset consists of 53 samples related to bipolar disorder, 28 of which correspond to patients with bipolar disorder, while 3 of them are

patient in first episode, and 25 matched control samples. The dataset is split into two different groups in order to assess the impact of medication in patients. The first group is composed of healthy controls and medicated bipolar disorder patients, while the second one consisted of healthy controls and all bipolar disorder patients. Each group is examined separately. Significance Analysis of Microarrays (SAM) is the filter univariate method used, while several categories of classification methods are implemented; Least Absolute Shrinkage and Selection Operator (LASSO), Support Vector Machine (SVM) and Relevance Vector Machines. Since SAM is a filter univariate approach, LASSO regression is selected to improve model discrimination performance, producing interpretable models by setting a considerable amount of features at exactly zero. LASSO achieves good classification performance while it tends to keep a large number of features, resulting in a genomic signature, which has a large number of common genes with SAM approach. Furthermore, while the SVM classifier leads to good classification performances, the size of genomic signatures is considerably small in size, leading to the idea of using the RVM classifier, which stands for an improved performance. Experimental results proved that SBV reached stable results after a maximum of 200 iterations on a worst case scenario. Moreover, observed estimates for the classification accuracy and the genomic signature were consistent across different and independent executions of SBV. According to the SBV results, RVM outperformed all other methods, reaching accuracies close to 90%. Specifically, concerning the two groups SVM reached accuracies of 89.5% for 6 genes selected and 87% for 8 genes selected, respectively. Compared to SVM, RVM reached accuracies of 91.4% for 78 genes selected and 90% for 73 genes selected, respectively. The RVM observed results of two different groups have 23 common genes. These 23 observed common genes are removed from the initial sets of significant genes and the RVM procedure is repeated 100 times for each significant set. The observed results also have 8 common genes. Consequently, the aforementioned procedure is repeated again, removing the 8 common genes from the rest sets of significant genes. The observed results also have 1 common gene. Since there are 3 difference cases used for the RVM methods, 3 different unified signatures were extracted, the 23 gene, 8 gene and 1 gene signatures. In that manner, the final genomic signature is composed of 32 genes.

Furthermore, in order to assess the discrimination of genomic signature between the class labels the expression value of each gene is examined by estimating the mean as well as variance and standard deviation. According to the observed results, the standard deviation as well as variance of 32 significant genes of all samples are far off null, indicating that the expression value of each gene tends to be far off the mean and hence to each other. In that manner, the genes expressions are extended over a wider range of

values, representing high discrimination among the class of bipolar patients and healthy controls samples and leading to a more easily interpretable model.

Moreover, the consistency of gene selection in the signature is evaluated using the 10 - fold cross validation method, which generates 9 training datasets and only one test set. This process is repeated 10 rounds. In each round, one of the folds is used for validation, and the other 9 folds for training. Then the RVM classification method is performed. The process is repeated 200 times and the overall results are averaged. The observed classification accuracy was approximately 84.67% for 26 mean genes selected.

Finally, a good generalization performance is achieved when a genomic signature is able to predict the label of unseen samples correctly. In that manner, a new independent dataset is used and the procedure of 10 – fold cross validation is repeated. The observed mean classification accuracy was approximately 74.17% for 13 mean genes selected, performing very good generalization performance when it comes to the classification of unknown samples.

Concerning the biological evaluation, the enriched processes and pathways that assigned to the thirty-two genes are important with respect to different aspects of pathophysiology and neurodevelopment of bipolar disorder.

Apart from the two genes, NOG and CTBP1, referred to be putative predictors, we support the notion that our “32 unified common gene signature” in its entirety can play a classification role in discriminating healthy controls from patients with BPD, and is a potent predictor without sound effects of medication.

# References

- [1] Anderson, IM; Haddad, PM; Scott, J (Dec 27, 2012). "Bipolar disorder.". *BMJ (Clinical research ed.)* 345: e8508. [doi:10.1136/bmj.e8508](https://doi.org/10.1136/bmj.e8508). PMID 23271744.
- [2] Gil Chu, Jun Li, Balasubramanian Narasimhan, Robert Tibshirani, Virginia Tusher "Significance Analysis of Microarrays", Users guide and technical document
- [3] V. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to transcriptional responses to ionizing radiation. *Proc. Natl. Acad. Sci. USA.*, 98:5116–5121, 2001.
- [4] Diploma Thesis Nikolaos-Kosmas Chlis. "Comparison of Statistical Methods for Genomic Signature Extraction", Chania (September 2013).
- [5] National Human Genome Research Institute, <http://www.genome.gov/>
- [6] Nuwer, Rachel (18 July 2015). "Counting All the DNA on Earth". *The New York Times* (New York: The New York Times Company). ISSN 0362-4331. Retrieved 2015-07-18
- [7] Danh V. Nguyen, A. Bulak Arpat, Naisyin Wang, Raymond J. Carroll, "DNA Microarray Experiments: Biological and Technological Aspects," *BIOMETRICS.*, vol. 58, pp. 701-717, 2002.
- [8] Musa H. Asyali, Dilek Colak, Omer Demirkaya, Mehmet S. Inan, "Gene Expression Profile Classification: A Review," *Current Bioinformatics*, vol. 1, no. 1, pp. 55-73, 2006.
- [9] Georges Natsoulis, Laurent El Ghaoui, Gert R.G. Lanckriet, Alexander M. Tolley, Fabrice Leroy, Shane Dunlea, Barrett P. Eynon, Cecelia I. Pearson, Stuart Tugendreich, and Kurt Jarnagin "Classification of a large microarray data set: Algorithm comparison and analysis of drug signatures", *Genome Research* 2005 May; 15(5): 724–736.
- [10] Osareh, A. Computer Eng. Dept., Islamic Azad Univ., Dezful, Iran Shadgar, B. , "Microarray data analysis for cancer classification," *Health Informatics and Bioinformatics (HIBIT)*, 2010 5th International Symposium on, pp 125-132, 2010
- [11] Michael P. S. Brownz, William Noble Grundyz, David Linz , Nello Cristianinix, Charles Sugne, Manuel Ares, Jr., David Hausslerz, "Support Vector Machine Classification of Microarray Gene Expression Data", *Proceedings of the National Academy of Sciences.* 97(1):262-267.
- [12] C. A. Davis, F. Gerick, V. Hintermair, C. C. Friedel, K. Fundel, R. Kóffner, R. Zimmer, "Reliable gene signatures for microarray classification: assessment of stability and performance," *Bioinformatics.*, vol. 22,no. 19, pp. 2356–2363, 2006.

- [13] R. Armapanzas, I. Inza, P. Larrapaga , "Detecting reliable gene interactions by a hierarchy of Bayesian network classifiers," *Computer Methods and Programs in Biomedicine.*, vol. 91, pp.110-121, 2008.
- [14] Anderson, IM; Haddad, PM; Scott, J. "Bipolar disorder." *BMJ (Clinical research ed.)* 2012; 345, Dec 27, 2012
- [15] Schmitt A, Malchow B, Hasan A, Falkai P., "The impact of environmental factors in severe psychiatric disorders". *Front Neurosci*, vol. 8, no. 19, February 2014
- [16] Kerner Berit "Genetics of bipolar disorder". *Application Clinical Genetis*, vol. 7, pp. 33–42, February 2014
- [17] Isabelle Guyon Clopinet, André Elisseeff (2003). "An Introduction to Variable and Feature Selection." *The Journal of Machine Learning Research*, Vol. 3, 1157-1182.
- [18] Szabo A. et al, Variable selection and pattern recognition with gene expression data generated by the microarray technology. *Math Biosci* (2002) 176(1), 71-98
- [19] Christopher M. Bishop, "Pattern Recognition and Machine Learning," Springer, 2006
- [20] J.P. Marques de Sa, *Pattern recognition, Concepts, Methods and Applications*, Springer, 2001.
- [21] Alpaydin, Ethem (2010). "Introduction to Machine Learning." MIT Press. p. 9. ISBN 978-0-262-01243-0.
- [22] Y. Saeys, I. Inza, P. Larrapaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics.*, vol. 23, no. 19, pp. 2507–2517, 2007. doi:10.1093/bioinformatics/btm344
- [23] I. Guyon, A. Elisseeff, "An Introduction to Variable and Feature Selection," *Journal of Machine Learning Research.*, vol. 3, pp. 1157-1182, 2003.
- [24] Yu Wanga,\*, Igor V. Tetkoa, Mark A. Hall, Eibe Frank, Axel Facius, Klaus F.X. Mayera, Hans W. Mewes,c. "Gene selection from microarray data for cancer classification—a machine learning approach", *Computational Biology and Chemistry* 29 (2005) 37–46
- [25] M. E. Blazadonakis , M. Zervakis, "The linear neuron as marker selector and clinical predictor in cancer gene analysis," *Computer methods and programs in biomedicine.*, vol. 91, pp. 22–35, 2008.
- [26] Klaous-Robert Muller, Sebastian Mika, Gunnar Ratsch, Koji Tsuda, Bernhard Scholkopf. "An Introduction to Kernel-Based Learning Algorithms", *IEEE Transactions On Neural Networks*, Vol. 12, No. 2., March 2001.
- [27] "Linear and nonlinear classifiers", [Online]. Available: <http://cs.joensuu.fi/pages/whamalai/expert/lecture6.htm> . [Accessed 14 April 2015].

- [28] Hilary L. Seal (1967). "The historical development of the Gauss linear model". *Biometrika* **54** (1/2): 1–24. doi:10.1093/biomet/54.1-2.1
- [29] C. Saunders, A. Gammerman, V. Vovk, "Ridge Regression Learning Algorithm in Dual Variables," Royal Holloway, University of London Proceedings of the 15th International Conference on Machine Learning, ICML '98, 1998.
- [30] "Ordinary least squares" [Online]. Available: [http://en.wikipedia.org/wiki/Ordinary\\_least\\_squares](http://en.wikipedia.org/wiki/Ordinary_least_squares) [Accessed 15 April 2015].
- [31] Tibshirani, R. (1996). "Regression shrinkage and selection via the lasso". *Journal of the Royal Statistical Society, B* 58, 267–288.
- [32] Zare, Habil (2013). "Scoring relevancy of features based on combinatorial analysis of Lasso with application to lymphoma diagnosis". *BMC genomics* 14: S14. doi:10.1186/1471-2164-14-S1-S14
- [33] C. Cortes, V. Vapnik, "Support-Vector Networks," *Machine Learning*, vol. 20, pp. 273-297, 1995.
- [34] "Support Vector Machine", 28 February 2012, [Online]. Available: [http://en.wikipedia.org/wiki/Support\\_vector\\_machine](http://en.wikipedia.org/wiki/Support_vector_machine) . [Accessed 29 April 2015]
- [35] Michael E. Tipping, "Sparse Bayesian Learning and the Relevance Vector Machine", *Journal of Machine Learning Research* 1 (2001) 211-244
- [36] Sylvain Arlot, Alain Celisse (July 2009). "A survey of cross-validation procedures for model selection": *Statistics Surveys* Vol. 4 (2010) 40–79 ISSN: 1935-7516 DOI: 10.1214/09-SS054 .
- [37] Ron Kohavi . "A study of Cross Validation and Bootstrap for accuracy estimation and model selection", Appears at the International Joint Conference on Artificial Intelligence (IJCAI), 1995
- [38]"Wolfram MathWorld - Weak Law of Large Numbers":  
<http://mathworld.wolfram.com/WeakLawofLargeNumbers.html>
- [39] Dennis Kostka , Rainer Spang , "Microarray Based Diagnosis Profits from Better Documentation of Gene Expression Signatures ," *PLoS Comput Biol* 4(2): e22. doi:10.1371/ journal.pcbi.0040022 , 2008.
- [40] Nick Craddock, Ian Jones, "Genetics of bipolar disorder", *Journal of Medical Genetics*, 1999; 36:585–594, doi:10.1136/jmg.36.8.585
- [41] Peter Holmans, Elaine K. Green, Jaspreet Singh Pahwa, Manuel A.R. Ferreira, Shaun M. Purcell, Pamela Sklar, The Wellcome Trust Case-Control Consortium, Michael J. Owen, Michael C. O'Donovan, Nick Craddock, "Gene Ontology Analysis of GWA Study Data Sets Provides Insights into the Biology of Bipolar Disorder," *Am J Hum Genet*, Volume 85, Issue 1, p13–24, 10 July 2009
- [42]=Clelland et al (2013) GSE46449

- [43] R. Edgar, M. Domrachev, and A.E. Lash, "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository," *Nucleic Acids Res.*, vol. 30, no. 1, pp. 207-210, 2002.
- [44] Savitz et al (2013) GSE39653
- [45] S.H. Chang, L. Gao, Z. Li, W.N. Zhang, Y. Du, and J. Wang, "BDgene: A Genetic Database for Bipolar Disorder and Its Overlap With Schizophrenia and Major Depressive Disorder," *Biol Psychiatry*, vol. 74, no. 10, pp. 727-733, 2013.
- [46] J.T. Chang and J.R. Nevins, "GATHER: a systems approach to interpreting genomic signatures," *Bioinformatics*, vol. 22, no. 23, pp. 2926-2933, 2006.
- [47] M.T. Tsuang, N. Nossova, T. Yager, M.M. Tsuang, S.C. Guo, K.G. Shyu, S.J. Glatt, C.C. Liew, "Assessing the validity of blood-based gene expression profiles for the classification of schizophrenia and bipolar disorder: a preliminary report," *Am. J. Med. Genet. B. Neuropsychiatr. Genet.*, vol. 133B, no. 1, pp. 1-5, Feb 2005.
- [48] R.B. Mansur, E. Brietzke, "The "selfish brain" hypothesis for metabolic abnormalities in bipolar disorder and schizophrenia," *Trends Psychiatry Psychother.*, vol. 34, no. 3, pp. 121-128, Sep 2012.
- [49] C.P. Johnson, R.L. Follmer, I. Oguz, L.A. Warren, G.E. Christensen, J.G. Fiedorowicz, V.A. Magnotta, and J.A. Wemmie, "Brain abnormalities in bipolar disorder detected by quantitative T1ρ mapping," *Mol Psychiatry*, vol. 20, no. 2, pp. 201-206, Feb 2015.
- [50] J.T. Coyle and R.S. Duman, "Finding the intracellular signaling pathways affected by mood disorder treatments," *Neuron*, vol. 38, pp. 157-160, April 2003.
- [51] J. Du, J. Quiroz, P. Yuan, C. Zarate, and H.K. Manji, "Bipolar disorder: involvement of signaling cascades and AMPA receptor trafficking at synapses. *Neuron Glia Biol.*, vol. 1, no. 3, pp. 231-243, Aug 2004.
- [52] R. Zanardi, G. Racagni, E. Smeraldi, and J. Perez, "Differential effects of lithium on platelet protein phosphorylation in bipolar patients and healthy subjects," *Psychopharmacology (Berl.)*, vol. 129, no. 1, pp. 44-47, Jan 1997.
- [53] B.N. Frey, M.M. Fonseca, R. Machado-Vieira, J.C. Soares, and F. Kapczinski, "Neuropathological and neurochemical abnormalities in bipolar disorder", (Article in Portuguese), *Rev Bras Psiquiatr.*, vol. 26, no. 3, pp. 180-188, Sep 2004.
- [54] V. Maletic and C. Raison, "Integrated neurobiology of bipolar disorder," *Front Psychiatry*, vol. 5, article 98, pp. 1-24, Aug 2014.