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

Machine Learning Methods for Genomic Signature Extraction

M.Sc. Thesis

Nikolaos-Kosmas Chlis

Thesis Committee
Professor Michael Zervakis, Thesis Supervisor
Professor Costas Balas
Associate Professor Katerina Mania

Chania, July 2015

Abstract

The application of machine learning methodologies for the analysis of DNA microarray data has become a common practice in the field of bioinformatics. DNA microarrays can be used in order to simultaneously measure the expression value of thousands of genes. Given the measurements of gene expression, machine learning methods can be employed in order to identify candidate genes that are related to a biological state or phenotype of interest, such as cancer. These lists of candidate genes are often called "genomic signatures" in literature. The application of machine learning methods for the extraction of genomic signatures is a necessity, since it is practically impossible for field experts to assess the importance of each gene individually by manual inspection due to the large size of the genome, which consists of approximately 25,000 genes.

Machine learning methods such as feature subset selection and classification algorithms are popular choices for the extraction of genomic signatures. Univariate feature selection methods filter genes according to difference in their gene expression profiles among samples belonging to different classes of interest, such as control and disease. Since they test each gene individually, univariate methods are computationally efficient and they select genes with high discrimination ability. However, they ignore associations among genes. On the other hand, multivariate methods simultaneously assess groups of genes and select candidate genes based on their predictive performance when used in conjunction with a classifier. As such, they are more efficient at capturing the latent associations among genes and select genes with high predictive capability, at the cost of being computationally expensive. While the applied feature selection and classification methodologies have matured and several state of the art algorithms have been established, the stability of the extracted genomic signatures is often overlooked. As a result, the genomic signatures extracted by many methodologies are unstable under sample variations. That is, the extracted signatures differ significantly under variations of the training data. Since result stability is related to generalization, this instability raises skepticism in the expert community and hinders the validity and clinical application of research findings extracted from such gene expression studies.

This thesis deals with the following three aspects of the selection and evaluation of gene signatures, namely stability, predictive capability and statistical significance. First, a framework for the extraction of stable genomic signatures, called Stable Bootstrap Validation (SBV) is introduced. The proposed methodology enforces stability at the validation step. As a result, it can be combined with any classification method, as long as it supports feature selection. Three publicly available gene expression datasets are used in order to test the proposed methodology. First the dimensionality of the datasets is reduced using a filtering method. Then, bootstrap resampling is utilized in order to generate a list of candidate signatures according to the selection frequency of genes across all bootstrap datasets. Then, a stable signature which has maximal predictive performance in terms of accuracy, sensitivity and specificity is extracted and the predictive performance of all candidate signatures is plotted in an elaborate manner for further inspection. Additionally, the application of random sampling methods for countering the negative effects of imbalanced datasets in classification was investigated, since imbalanced datasets are frequently found in DNA microarray studies where control samples are usually scarce. Moreover, a proper statistical framework was implemented that includes two separate statistical tests, in order to assess the statistical significance of the extracted signature in terms of classification accuracy as well as association to the response variable (phenotype/biological state). Finally, the robustness of the methodology is assessed by testing the degree of "agreement" among signatures extracted from independent executions of the methodology.

Περίληψη

Η εφαρμογή μεθόδων μηχανικής μάθησης για την ανάλυση δεδομένων από μικροσυστοιχίες DNA έχει γίνει κοινή πρακτική στον τομέα της βιοπληροφορικής. Μικροσυστοιχίες DNA χρησιμοποιούνται προκειμένου να μετρηθεί ταυτόχρονα η τιμή έκφρασης χιλιάδων γονιδίων. Λαμβάνοντας υπ'όψιν τις μετρήσεις της γονιδιακής έκφρασης, μέθοδοι μηχανικής μάθησης μπορούν να χρησιμοποιηθούν για τον εντοπισμό υποψήφιων γονιδίων που σχετίζονται με μία βιολογική κατάσταση ή φαινότυπο ενδιαφέροντος, όπως ο καρκίνος. Αυτές οι λίστες των υποψήφιων γονιδίων συχνά αποκαλούνται "γονιδιακές υπογραφές" στη βιβλιογραφία. Η εφαρμογή των μεθόδων μηχανικής μάθησης για την εξαγωγή γονιδιακών υπογραφών είναι αναγκαία, δεδομένου ότι είναι πρακτικά αδύνατο για τους εμπειρογνώμονες να αξιολογήσουν τη σημασία του κάθε γονιδίου ξεχωριστά, λόγω του μεγάλου μεγέθους του γονιδιώματος, το οποίο αποτελείται από περίπου 25.000 γονίδια.

Μέθοδοι μηχανικής μάθησης όπως μέθοδοι επιλογής χαρακτηριστικών και μέθοδοι ταξινόμησης αποτελούν δημοφιλείς επιλογές για την εξαγωγή γονιδιακών υπογραφών. Μονομεταβλητές μέθοδοι επιλογής χαρακτηριστικών φιλτράρουν τα γονίδια σύμφωνα με διαφορές στο προφίλ της γονιδιακής τους έκφρασής μεταξύ δειγμάτων που ανήκουν σε διαφορετικές κατηγορίες ενδιαφέροντος, όπως παθολογικά δείγματα και δείγματα αναφοράς. Εφόσον εξετάζουν κάθε γονίδιο ξεχωριστά, οι μονομεταβλητές μέθοδοι είναι υπολογιστικά αποδοτικές και επιλέγουν γονίδια με υψηλή διακριτικότητα. Ωστόσο, αγνοούν τις αλληλεπιδράσεις μεταξύ των γονιδίων. Από την άλλη πλευρά, οι πολυμεταβλητές μέθοδοι αξιολογούν ταυτόχρονα ομάδες γονιδίων και επιλέγουν υποψήφια γονίδια με βάση την προγνωστική απόδοσή τους όταν χρησιμοποιούνται σε συνδυασμό με έναν ταξινομητή. Ως εκ τούτου, είναι πιο αποτελεσματικές στο να λαμβάνουν υπ'όψιν τις λανθάνουσες σχέσεις μεταξύ των γονιδίων και επιλέγουν γονίδια με υψηλή προγνωστική ικανότητα, όμως έχουν υψηλό υπολογιστικό κόστος. Ενώ οι εφαρμοζόμενες μεθοδολογίες επιλογής χαρακτηριστικών και ταξινόμησης έχουν ωριμάσει και αρκετές αποδοτικές μέθοδοι έχουν δημιουργηθεί, η σταθερότητα των εξαγόμενων γονιδιακών υπογραφών συχνά παραβλέπεται. Ως αποτέλεσμα, οι γονιδιακές υπογραφές που εξάγονται από πολλές μεθοδολογίες είναι ασταθείς ως προς παραλλαγές των δειγμάτων εκπαίδευσης. Δηλαδή, οι εξαγόμενες υπογραφές τείνουν να διαφέρουν σημαντικά μεταξύ τους, όταν έχουν χρησιμοποιηθεί ελαφρώς διαφορετικά δεδομένα εκπαίδευσης. Δεδομένου ότι η σταθερότητα των αποτελεσμάτων σχετίζεται με την γενίκευση, αυτή η αστάθεια δημιουργεί σκεπτικισμό στην κοινότητα των εμπειρογνωμόνων, αμφισβητεί την εγκυρότητα και εμποδίζει την κλινική εφαρμογή των ερευνητικών ευρημάτων που προέρχονται από τέτοιου είδους μελέτες γονιδιακής έκφρασης.

Η παρούσα εργασία ασχολείται με τις εξής τρεις πτυχές της επιλογής και αξιολόγησης γονιδιακών υπογραφών: τη σταθερότητα, την προβλεπτική ικανότητα και τη στατιστική σημαντικότητα. Ένα πλαίσιο για την εξαγωγή των σταθερών γονιδιακών υπογραφών, που ονομάζεται Stable Bootstrap Validation (SBV) παρουσιάζεται. Η προτεινόμενη μεθοδολογία επιβάλλει σταθερότητα της εξαγόμενης γονιδιακής υπογραφής στο στάδιο της αξιολόγησης (validation). Ως αποτέλεσμα, μπορεί να συνδυαστεί με οποιαδήποτε μέθοδο ταξινόμησης, εφόσον αυτή υποστηρίζει επιλογή χαρακτηριστικών. Τρία ελεύθερα διαθέσιμα σύνολα δεδομένων γονιδιακής έκφρασης χρησιμοποιούνται για να αξιολογηθεί η προτεινόμενη μεθοδολογία. Αρχικά, η διαστατικότητα των συνόλων δεδομένων μειώνεται χρησιμοποιώντας μια μέθοδο φιλτραρίσματος. Στη συνέχεια, bootstrap αναδειγματοληψία χρησιμοποιείται για να δημιουργηθεί μια λίστα υποψήφιων υπογραφών, σύμφωνα με τη συχνότητα επιλογής των γονιδίων στο σύνολο των παραγόμενων bootstrap συνόλων δεδομένων. Στη συνέχεια, μία σταθερή υπογραφή που έχει τη μέγιστη ικανότητα πρόβλεψης όσον αφορά την ακρίβεια, την ευαισθησία και την ειδικότητα εξάγεται και η ικανότητα πρόβλεψης όλων των υποψήφιων υπογραφών συμπυκνώνεται και σχεδιάζεται σε ένα ευδιάκριτο διάγραμμα για περαιτέρω επιθεώρηση. Επίσης, εξετάζεται η εφαρμογή μεθόδων τυχαίας δειγματοληψίας για την αντιμετώπιση των αρνητικών επιπτώσεων της μη ισορροπημένης κατανομής των δειγμάτων σε παθολογικές και μη κατηγορίες στα σύνολα δεδομένων. Η μη ισορροπημένη κατανομή των δεδομένων αποτελεί συχνό φαινόμενο σε μελέτες μικροσυστοιχιών DNA, όπου τα δείγματα αναφοράς συνήθως είναι πολύ λιγότερα από τα παθολογικά. Επιπλέον, υλοποιήθηκε ένα κατάλληλο στατιστικό πλαίσιο, που περιλαμβάνει δύο ξεχωριστά στατιστικά τεστ, προκειμένου να αξιολογηθεί η στατιστική σημαντικότητα της εξαγόμενης υπογραφής όσον αφορά την ακρίβεια της ταξινόμησης, καθώς και τη σύνδεση της υπογραφής με την μεταβλητή απόκρισης (φαινότυπος/παθολογική κατάσταση). Τέλος, η ευρωστία της μεθοδολογίας αξιολογείται μέσω της εκτίμησης του βαθμού "συμφωνίας" μεταξύ των υπογραφών που προέρχονται από ανεξάρτητες εκτελέσεις της μεθοδολογίας.

Acknowledgements

I would like to thank:

My thesis supervisor, Professor Michalis Zervakis, for his guidance and for giving me the chance to expand my knowledge in the exciting field of bioinformatics.

Dr. Ekaterini S. Bei for her support and biological insight.

Professor Costas Balas and Associate Professor Katerina Mania, for their contribution as members of the thesis committee.

My high school biology teacher, Katerina Papadaki, for inspiring me to work hard and explore the field of biology.

Anastasia, my friends and my family.

Last but not least, I would like to thank the Onassis Foundation for supporting this work through a graduate studies scholarship.

This work was also supported by the "ONCOSEED" project funded by the NSRF2007-13 of the Greek Ministry of Development

Table of Contents

Table of Contents 5
List of Figures 7
List of Tables 9
1 – Introduction 10
1.1 Introduction to DNA Microarray Analysis and Related Challenges 10
1.2 The Problem of Genomic Signature Instability 11
1.3 Related Work 12
1.4 Thesis Outline, Innovation and Previous Work 13
2 - Theoretical Background 15
2.1 The Human Genome and DNA Microarrays 15
2.1.1 The Human Genome 15
2.1.2 DeoxyriboNucleic Acid – DNA 15
2.1.3 DNA Microarrays 16
2.1.4 Types of DNA Microarrays 17
2.1.5 Basics steps of a Microarray Experiment 18
2.2 Machine Learning Applied to DNA Microarray Data 20
2.3 Introduction to Classification 21
2.3.1 The K-Nearest Neighbors Classifier 21
2.3.2 The Support Vector Machine (SVM) Classifier 23
2.3.3 The Relevance Vector Machine (RVM) Classifier 25
2.4 Feature Subset Selection (FSS) 26
2.4.1 Filter Methods (Univariate) and Significance Analysis of Microarrays 26
2.4.2 Wrapper Methods (Multivariate) and RFE-SVM 26
2.4.3 Embedded Methods (Multivariate) and RVM 27
2.5 Gene Set Analysis Methods and Globaltest 28
2.6 Evaluation (Validation) Methods 28
2.6.1 Holdout Validation 28
2.6.2 K-Fold Cross Validation (K-Fold CV) 29
2.6.3 Leave One Out Cross Validation (LOOCV) 30
2.6.4 Repeated Random Sub-Sampling Validation 30
2.6.5 Bootstrap Resampling Validation 31
2.7 Random Oversampling and Undersampling for Classification of Imbalanced Datasets 32
2.8 The Law of Large Numbers 33
2.9 Correspondence At the Top (CAT) plots 35
3 – Proposed Methodology 36
3.1 Overview 36
3.2 Datasets 37
3.3 Preliminary Feature Selection 37
3.4 Stable Signature Extraction through Stable Bootstrap Validation 37

3.5 Estimation of Classification Performance 42
3.6 Assessment of Statistical Significance 43
3.7 Evaluation of Signature Consistency 44
4 – Results 45
4.1 Data Preparation and Preliminary Filtering 45
4.2 Classification (RVM and RFE-SVM) 45
4.2.1 GSE_Merged 46
4.2.2 GSE42568 48
4.2.3 GSE35974 50
4.3 Statistical Significance 52
4.4 Signature Consistency 54
4.5 Biological Evaluation 57
Conclusion and Future Work 64
References 67
SBV Source Code Availability 72
Appendix A - Classification Performance of All Candidate Genomic Signatures 72
Appendix B - Gene Lists of All Candidate Genomic Signatures 79

List of Figures

Figure 2.1: The DNA double helix (source: http://en.wikipedia.org/wiki/File:DNA_Structure%2BKey%2BLabelled.pn_NoBB.png) 16

Figure 2.2: Graphical illustration of hybridization between the probes and the target. (source: http://commons.wikimedia.org/wiki/File:NA_hybrid.svg) 19

Figure 2.3: Graphical illustration of a 2-channel DNA microarray image. (source: http://commons.wikimedia.org/wiki/File:DNA_microarray.svg) 20

Figure 2.4 The test sample (purple X) will be classified in the first class of green circles in the case of K=3. However, in the case of K=5 it will be classified in the second class of red rectangles 22

Figure 2.5 The class borders of an 1-NN classifier In the case of 3-way classification (source: http://commons.wikimedia.org/wiki/File:Map1NN.png) 22

Figure 2.6 The black hyperplane separates the two classes, resulting in the maximum margin between their closest samples, and thus is selected as the SMV separating hyperplane 23

Figure 2.7 Holdout validation method 29

Figure 2.8 5-Fold Cross Validation 29

Figure 2.9 Leave One Out Cross Validation 30

Figure 2.10 Repeated Random Sub-Sampling Validation 31

Figure 2.11 Bootstrap Resampling Validation. In this illustration, each bootstrap dataset is split using holdout validation as an example 31

Figure 2.12 Instantaneous values of the 300 rolls of a 6-sided die 34

Figure 2.13 Demonstration of the law of large numbers: the mean value over all rolls converges towards 3.5, the expected value of the experiment, as more repetitions of the experiment take place ... 34

Figure 2.14 CAT plot example 35

Figure 3.1 Steps of the proposed methodology 36

Figure 3.2 Flowchart of the SBV process for stable signature extraction. 40

Figure 3.3 Illustration of the convergence of the mean number of genes selected across all bootstrap datasets. For the generation of this figure RFE-SVM was used on the GSE35974 dataset. Notice that the average number of genes has converged and is stable at 200 datasets. A result that agrees with the proposed stability criterion 41

Figure 3.4 We zoom in and focus on the first 200 bootstrap datasets of the previous plot. The original bootstrap windows (step 2 of the pseudo-code above) are seen ranging from datasets 1 to 150 (or B to 3B for B=50). If convergence of average signature size has not achieved the required threshold, the windows are expanded (step 7 of the pseudo-code above) and the process is repeated until no further expansions and assessment of additional bootstrap datasets are necessary 41

Figure 4.1 RVM classification results for GSE_Merged (no sampling) 47

Figure 4.2 RFE-SVM classification results for GSE_Merged (no sampling) 47

Figure 4.3 RVM classification results for GSE42568 (no sampling) 49

Figure 4.4 RFE-SVM classification results for GSE42568 (no sampling) 49

Figure 4.5 RVM classification results for GSE35974 (no sampling) 51

Figure 4.6 RFE-SVM classification results for GSE35974 (no sampling) 51

Figure 4.7 CAT plot measuring the degree of “agreement” between the signatures extracted by RVM and RFE-SVM. 52

Figure 4.8 Consistency of the GSE_Merged signature 55

Figure 4.9 Consistency of the GSE42568 signature 55

Figure 4.10 Consistency of the GSE35974 signature 56

List of Tables

Table 4.1 Number of genes in each dataset. The number of genes is reduced after filtering with SAM. For two out of the three datasets further filtering using fold change was necessary …..... 45

Table 4.2 Overview of classification results …..... 46

Table 4.3 Predictive performance of the RVM and RFE-SVM classifiers, including over and undersampling results for GSE_Merged …..... 46

Table 4.4 Predictive performance of the RVM and RFE-SVM classifiers, including over and undersampling results for GSE42568 …..... 48

Table 4.5 Predictive performance of the RVM and RFE-SVM classifiers, including over and undersampling results for GSE35974 …..... 50

Table 4.6 Results of the two statistical significance tests for all datasets. The average difference in classification accuracy between the extracted signature and random signatures of the same size is displayed in parentheses at Test1-A …..... 53

Table 4.7 Results of Enrichment Analysis and Gene-Breast Cancer Association for GSE_Merged …..... 59

Table 4.8 Results of Enrichment Analysis and Gene-Breast Cancer Association for GSE42568 …..... 60

Table 4.9 Results of Enrichment Analysis and Gene-Breast Cancer Association for GSE35974 …..... 61

Table 4.10 Comparison of Gene Signatures in the context of their convergence by applying WebGestalt and GATHER …..... 62

1 - Introduction

1.1 Introduction to DNA Microarray Analysis and Related Challenges

While the mapping of the human genome has has been a subject of study for decades, it was until the more recent advent of DNA microarray technology that scientists have been given a valuable tool in measuring the expression levels of different genes in a biological system. The genomic analysis using DNA microarrays, serves a dual purpose. First, Scientists can observe patterns in the data that can lead to different expression profiles among distinct classes of interest. In that manner, the need arises for identification of sets of genes that strongly differentiate their expression levels among classes of interest. These sets of genes are also called “genomic signatures”. Second, using these sets of genes along with the patters that have been observed, scientists can design classification methodologies that assign class labels to new unknown samples. For example, when sets of genes that differentiate their expression levels between cancerous and non-cancerous tissue samples, they can be used to identify whether an unknown sample belonging to a patient corresponds to cancerous tissue or not. Moreover, these specific genomic signatures can be used to provide insight into biological processes, such as cancer and possibly lead to new methods of treatment.

Feature Selection

However, the analysis of genomic datasets is prone to the problem known as “curse of dimensionality” since typically the number of available samples is considerably smaller than the number of features (genes) used for classification. To be precise, the number of samples is usually in the order of a few hundred in a best case scenario, while there are thousands of genes, approximately 25,000 in the human genome. The effect of the “curse of dimensionality” implies significant decrease in classification performance, instability of the derived signature, as well as difficulties in generalizing the results. The above problems call for methods that perform dimensionality reduction by eliminating “irrelevant” sets of features, which are called feature selection methods. There are several categorizations of feature selection methods e.g.: filter methods, following a univariate approach that examines one feature at a time; wrapper and embedded methods, which are multivariate approaches for simultaneously examining different sets of features. Univariate methods select features that strongly differentiate their behavior between classes of interest and as such, they focus on features aimed at improving class separability. Multivariate methods, aim at selecting a set of features that maximizes the performance of a classification method and aim at selecting sets of features that improve class prediction of unknown samples. In this manner, feature selection as a methodology is often intertwined with the classification process of new samples.

Variability of Results and Imbalanced Datasets in Classification

Since classification methodologies are often mixed with feature selection to produce sets of informative features, the problem of classification of new samples is also an important aspect of microarray analysis by itself, since it can lead to new and efficient prognosis methodologies. Given that the effect of the “curse of dimensionality” can been counterfeited by some form of feature selection, and an informative and relatively small set of features has been extracted, classification methods are used in order to classify new data into known classes of interest. Moreover, in the case of multivariate (wrapper and embedded) methods, the extraction of the genomic signature is simultaneous to the classifier's training process. A challenge associated to classification is the variability of performance estimates. When the performance of classification methods is estimated, the variability of the observed results is typically ignored. That is, for

independent executions of the validation method the observed predictive performance of the classifier is always slightly different. In this manner, it is important to account for variability before deciding which classification method yields the best results [63] [64] [65]. In practice, it is reasonable to point out that there is not a single method among a group of classifiers that significantly outperforms all others and that the different methodologies yield comparable results, after accounting for the variability of classification performance estimates. Another issue in DNA Microarray studies is the fact that most available datasets are imbalanced. That is, the number of positive (disease) samples is typically much larger that the number of negative (control) samples available. This could lead to positively biased assessment of predictive performance if classification accuracy is the only metric utilized. For example, if there are 99 disease and only 1 control example and the control sample is consistently mis-classified as a disease sample, the resulting accuracy is still 99%, which is misleading.

Statistical Significance

Another need associated with biological problems is to determine whether the results extracted from feature selection and classification are observed as a result of the underlying biological system or are merely observed due chance (random noise). In this direction, statistical tests determining the randomness of results have been developed. Such tests often utilize permutation in order to measure the statistical significance of the observed results and assess their reliability [12] [21]. Results that are stable and reflect the biological model should also be consistent across different executions of the feature selection and classification methodologies. The proposed methodology assesses the statistical significance of the extracted genomic signature using two separate statistical tests.

1.2 The Problem of Genomic Signature Instability

Additionally to the challenges discussed in the previous section, another important aspect of DNA microarray analysis is the stability of the observed results, which is is usually overlooked, even though it is at least as important as classification accuracy [7] [9]. A genomic signature with good predictive capability is essential for domain experts in order to assess the prediction potential for clinical outcomes based on a targeted set of markers [8], besides the discrimination of the samples between classes of interest [55]. However, classification accuracy alone is not by any means proof that the extracted gene set is the only relevant subset of features [7], since many different gene subsets yield comparable predictive performance [61] [62], mainly due to high gene correlation, which leads to a large number of gene sets having comparable predictive capabilities. Most genomic signature extraction methods yield results that vary considerably when small variations take place in training and testing data, as well as in the algorithmic parameters. This instability raises skepticism in the expert community and hinders the validity and clinical applications of research findings. For example, different methodologies, or even the same methodology can extract substantially different genomic signatures under relatively small variations of the training data. As expected, the expert community will face results extracted from these methodologies with distrust. Result stability is linked to reproducibility and as such, methodologies that extract genomic signatures should yield result stability and robustness under sample variations. The most prominent causes of instability are [7] [8] [9]: (1) the small-N large-P problem, where the number of available samples is very small compared to the number of genes, (2) the redundancy of genes, which leads to the existence of multiple sets of “true” markers due to high correlation of genes and (3) the design of genomic signature extraction methodologies without considering stability.

The need for stability of results has lead to the development of methodologies aimed at extracting more stable, robust and generalizable performance estimates. A review of stable feature selection methodologies can be found in [7], while a review of stability metrics is reported in [10]. These methodologies

often rely on random sampling or splitting of the original dataset multiple times in order to generate a large number of training, as well as test sets, which are used to infer the performance estimates of a given feature selection and classification scheme. In accordance to this goal, Davis et al. in [11] perform random splitting of the original dataset a large number of times in order to extract stable feature selection and classification performance assessments over all datasets generated. Suzuki et al. in [13] generate multiple dataset using random sampling with replacement and take into account the results of leave one out cross validation over all datasets in order to extract performance estimates. Barrier et al. in [17] utilize Monte Carlo cross validation, splitting the dataset a large number of times in training and test sets of various sizes. Armañanzas et al. [15] propose bootstrap resampling as a means to extract a stable bayesian model of dependent genes. However, while these methodologies lead to stable results, they lack a formal definition of stability, as well as an objective criterion that defines when a sufficient level of stability is reached for the resulting genomic signature and the corresponding classification accuracy. The lack of such a criterion is bypassed using an arbitrary large number of bootstrap iterations in order to achieve stability, which range from 400 to thousands in the studies mentioned. Considering that feature selection and classification methods tend to be computationally intensive, performing such a large number of iterations can be impractical. Moreover, many of the studies mentioned utilize resampling methods to extract a stable genomic signature but assess classification performance based on typical cross validation techniques [15], [16]. Even if the genes in the signature are stable, the size of the signature itself (i.e. the number of selected genes) may differ considerable during the iterations [15] [16] [17]. This thesis aims at introducing a framework that utilizes an explicit definition of stability and objective criterion for determining when a sufficient level of stability is achieved for the extracted genomic signature and the classification accuracy, while performing a minimum number of bootstrap iterations. Moreover, the predictive capability of the extracted gene set is assessed using multiple performance metrics (accuracy, sensitivity, specificity) and appropriate confidence intervals are generated using a hybrid method.

1.3 Related Work

The evaluation of stability and reliability of results concerning genomic analysis has been the focus of several studies in the field of Bioinformatics. Many studies focus on random sampling of the original dataset in order to infer stable performance estimates. Bootstrap resampling, that is random sampling with replacement, as a method to estimate the sampling distribution of a random variable based on the observed data was first introduced by B. Efron in 1979 [18]. In the same study bootstrapping was compared to the Jackknife and standard leave one out cross validation, outperforming both methods. Davis et al. in [11] study the stability of genomic signatures and it's impact in the stability of classification accuracy. They also propose a methodology that utilizing random splitting for determining efficient combinations of feature selection and classification models depending on the stability of signatures as well as efficient classification performance. Soek Ying Neo et al. in [12] utilize Monte Carlo simulations in order to assess the quality of the selected genes as potential markers. Maglietta et al. In [21] rank each gene depending on performance of a ridge regression classifier when only that specific gene is used as a feature and also examine the statistical significance of that gene's observed classification accuracy. Suzuki et al. in [13] propose a model for the performance assessment of feature selection and classification methods, that takes advantage of the low bias of leave one out cross validation, while it aims to counter it's large estimation variance by utilizing bootstrap resampling. Haury et al. In [14] assess the influence in terms of stability, performance and interpretability, of different feature selection methods when used in conjunction with a set of classifiers. They also compare the performance of the genomic signatures to sets of randomly selected genes, a notion introduced by Ein-Dor et al. in [61]. Armañanzas et al. in [15] propose bootstrap resampling since it leads to reliability, robustness and few false positives in the observed results. They propose a scheme which utilizes bootstrap resampling in order to generate a large number of 1000 datasets and then univariate feature

selection method called “correlation feature selection” is performed on each dataset in order to reduce the dimensionality. A k-Dependence Bayesian classifier is then trained using each bootstrap dataset resulting in a directed acyclic graph where each arc represents statistical dependence between the connected nodes (genes). To achieve stability of the model, only arcs whose appearance frequency over all bootstrap datasets is over a fixed threshold, are included in the final model. To assess the classification performance, 5 fold cross-validation is performed. The same approach is followed by García-Bilbao et al. in [16] in order to construct a k-Dependence Bayesian classifier utilizing bootstrap resampling. However, instead of 5 fold cross validation on the constructed model, a set of 10 features selected by the model is used in conjunction with a set of different classification methods and their performance is evaluated using leave one out cross validation. The concept of using bootstrap resampling for the estimation of confidence in selecting a feature in a bayesian network was first introduced by Friedman et al. in [20] it was reported to lead to low rate of false positive rate for selected features and also achieve reliable conclusions about the selected features, even if the dataset used was relatively small. Barrier et al. in [17] propose Monte Carlo cross validation, which generates multiple random splits of the dataset using random sizes for the training and tests sets. That is, for each of the 16 different values for training test size, 100 datasets are performed by random splitting of the dataset leading into 1600 total datasets generated. Then, a filter feature selection method and a diagonal linear discriminant analysis classifier is trained on the training set, while classification performance is assessed using the corresponding test set. In that study it is also reported that many different signatures lead to similar classification performance, a result shared by [14] and [61] [17]. Kerr et al. in [19] perform bootstrap resampling from the original dataset in order to assess the stability of cluster analysis results. At the first level of bootstrapping, 10,000 bootstrap simulations are run in order to eliminate irrelevant features using a filter feature selection method. Then, at the second level of bootstrapping 499 additional datasets are generated from the filtered original dataset and each gene is clustered to one of 7 possible temporal patterns of yeast sporulation. Finally, the gene clusterings considered stable are only those being “95% stable” , that is they appear in at least 95% of the generated datasets, as well as in the clusters of the original dataset. Finally, a review additional stable feature selection methodologies is presented in [7], while a review of stability metrics is reported in [10].

1.4 Thesis Outline, Innovation and previous work

The necessary theoretical background concerning the human genome and methodologies concerning the analysis of DNA microarray data in the field of bioinformatics is covered in chapter 2. That includes the biological concepts regarding gene expression and DNA microarrays and machine learning fundamentals, such as feature selection and classification methodologies. Moreover, another category of genomic signature evaluation methods, called “gene set analysis” methods are introduced. Additionally, several evaluation methods including cross validation and bootstrap resampling are presented, followed by an introduction to oversampling and undersampling schemes for countering the effects of imbalanced datasets in classification. Then, the statistics theorem known as the “law of large numbers” is presented. Finally, a type of plot that measures the degree of “agreement” among different signatures, called the “correspondence at the top” plot is introduced. The proposed methodology for extraction of stable signatures and performance estimates, while assessing the statistical significance and consistency of results is covered in chapter 3. The results of the proposed methodology are presented in chapter 4, followed by a biological evaluation and interpretation of the extracted signatures.

The innovative concept of this thesis involves utilizing bootstrap resampling in order to generate a large number of datasets for training and testing feature selection and classification method and extracting a gene signature among a set of candidate signatures based on gene selection frequency. The extracted signature has maximal predictive performance in terms of accuracy, sensitivity and specificity and the results of all signatures are plotted in an elaborate manner for further inspection, using appropriate confidence

intervals generated by a hybrid method. The calculation of confidence intervals is usually omitted in similar methodologies, yet it is necessary since it accounts for the variability of the observed results and allows for the identification of statistically significant differences in predictive performance among different classification methods, as well as different candidate signatures. Moreover, another innovation is the formal statistical framework introduced for the assessment of the statistical significance of the extracted signature in terms of classification accuracy and association to the response variable (phenotype/class label). Through the assessment of statistical significance meaningful signatures that reflect the biological model are extracted, while signatures that reflect random noise are identified and discarded. The final innovative feature of the proposed methodology, is the assessment of signature stability based on correspondence at the top plots. Moreover, unlike similar methods that use an arbitrarily large number of bootstrap datasets, the proposed methodology employs an explicit criterion that determines when stability has been achieved for the genomic signature size. Under the assumption that the size of the signature extracted from each bootstrap dataset is an independent identically distributed random variable, according to the Law of Large Numbers the evaluation methodology is guaranteed to converge to a stable value for the average signature size, given that the number of bootstrap datasets used is large enough. As a result of this convergence, the computational burden of requiring additional bootstrap datasets to reach stability is minimized. In terms of previous work, an early concept of the stable bootstrap validation was introduced in [1], [2], [3] and some improvements which were developed as part of this thesis were first presented in [4], [5].

2 - Theoretical Background

In this chapter the necessary background concerning the human genome and the bioinformatics aspects of DNA microarray analysis are covered. The human genome and the technology of DNA microarrays are first briefly introduced. Next, the machine learning concepts necessary to understand the methods that were used for data analysis are presented.

2.1 The Human Genome and DNA Microarrays

DNA is contained in all living organisms and encodes all the information required for their development and function. DNA microarrays consist of a solid surface onto which DNA molecules are bonded and the abundance of specific types of DNA or RNA molecules is explicitly quantified. By using microarrays, scientists can measure the expression of thousands of genes simultaneously and extract useful biological information which may lead to discoveries about the functionality of these “building blocks” of living organisms. To be specific, new unknown functionalities of genes can be discovered, such as different behavior in different tissues or environments. As such, microarray experiments have multiple uses, such as discovery of gene functionality, co-regulation and predictive toxicology and cancer related studies. This project aims at collecting information about the theoretical background, manufacture and use of DNA microarrays up to the point of preprocessing the raw signal and before any machine learning algorithms are used for clustering or feature extraction and classification of the data.

2.1.1 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 the “Human Genome Project” [28]. The majority (~98%) of the human genome located in genetic material in the nucleus of human cells (with the exception of red blood cells), while the rest (~2%) is located in organelles called mitochondria which are responsible for converting the energy from food into a form usable by human cells. The genome located in the nucleus is organized into 23 pairs of chromosomes. These 46 chromosomes consist of 44 autosomes and 2 sex chromosomes, XX or XY for females and males respectively. Every chromosome has a constriction along it's length, called the centromere that divides the chromosome into a long and a short “arm”. Each chromosome can be thought as a string of thousands of genes, which are in turn made of DNA. The human genome is made of approximately 25,000 genes, most of them located in the nucleus, while only 37 refer to mitochondrial genes. Moreover, the genes located in the mitochondria are not organized in chromosomes. The DNA that makes up the genes is called “coding DNA”, while the DNA “string” between each gene is called “non-coding DNA”. Only a fraction of the genome refers to coding DNA, which is transcribed into RNA and then translated into proteins. Most of the genome consists of non-coding DNA that is associate with other known, or yet unknown, biological procedures.

2.1.2 DeoxyriboNucleic Acid - DNA

As mentioned above, each gene is made of DNA [29]. Deoxyribonucleic acid (DNA) consists of two long complementary strands of nucleotides that take the form of a double stranded helix. DNA consists of four primary types of nucleotide molecules. Each nucleotide consists of a phosphate, a sugar (deoxyribose) and one of four possible nitrogen bases, each represented by a letter: adenine (A), guanine (G), cytosine (C) and Thymine (T). These distinct nitrogen bases are also used to distinguish the four types of nucleotides from one another. Each nucleotide of a strand is connected by a hydrogen bond to it's complementary

nucleotide in the opposing DNA strand in order for the helix to maintain it's structure independent of the nucleotide sequence. These complementary nucleotide pairs are called the base pairs and correspond to G-C and A-T. The genetic information of each strand is read in the form of non-overlapping triplets of nucleotides. Given that there are 4 nucleotides, the possible number of different triplets is equal to $4^3=64$ combinations.

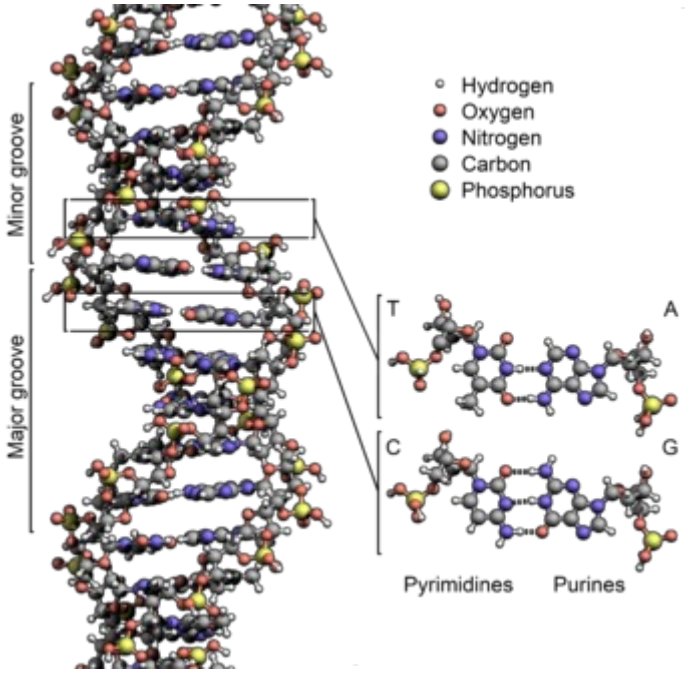

Figure 2.1: The DNA double helix.

2.1.3 DNA Microarrays

DNA Microarrays [29] [30] 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. During the process of transcription the DNA is used as a template for the enzyme RNA polymerase II to construct pre-mRNA utilizing complementary base pairing. However, since there is no Thymine in RNA, it is replaced by Uracile (U). Finally, the enzyme recognizes signals in the DNA chain that lead to the termination of the transcription process and the pre-mRNA chain is released into the nucleus where it is processed into mRNA. 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 is produced by mRNA using a process called Reverse Transcription (RT). The cDNA sequences used to bind target cDNA sequences of interest on the microarray are called “probes”. Probes bind target cDNA sequences by forming hydrogen bonds between complementary nucleotide base pairs, while multiple probes may be used to measure the same gene in order to reduce the noise present in the signal. The sequences bound by the probes are then detected using fluorescent dyes. If the genes of interest are found to be expressed, their expression levels are compared to those of known control samples in which the same genes are not expressed. Different technologies of DNA microarrays have been introduced. The “spotted cDNA microarray” developed at Stanford University utilizes robotic spotting of aliquots of purified cDNA clones, while category of microarrays developed by Affymetrix, Inc. Utilizes photo-lithography for embedding cDNA probes on silicon chips.

2.1.4 Types of DNA Microarrays

Spotted Microarrays

Spotted microarrays (also called deposition-based arrays [26]) were the first microarray platform created [24] [25] and they are still widely used. They consist of glass microscope slides on which PCR products or oligonucleotides are placed using robotic spotting. After the advent of microarrays it has become affordable for labs around the world to create their own spotted microarrays, fitting the needs of their experiments. While that may be a convenience, it also adds a degree of variability among the experimental platforms of different labs [25]. Moreover, the variability extends to the quality of features and as such dedicated image processing techniques have been developed in order to improve spot quality [24]. The spotted microarray is created in three main steps [24]:

1) Creating the DNA probes to be placed on the array
2) Spotting the DNA probes on the glass, using the spotting robot
3) Post-spotting processing of the glass slide (fixing) to avoid unwanted attachment of the target DNA on the glass during the hybridization step.

Affymetrix Chips

Affymetrix chips [24] [25] are the most popular commercial array platform in use. Their production relies on the light-directed synthesis techniques of photolithography and solid phase DNA synthesis [26]. They are constructed *in situ* on the surface of a chip using photolithography masks, similar to VLSI circuits. The photolithography mask is used in order to build up oligonucleotide chains on a solid substrate or glass chip. Each step of oligonucleotide synthesis requires a different mask and each mask is very expensive to produce. However, once produced a single mask can be used for the production of a large number of identical arrays, leading to a degree of standardization of the arrays used by the scientific community [24]. Each oligonucleotide probe consists of 25 bases per probe [27]. Contrary to spotted arrays, Affymetrix chips employ a set of probes in order to measure the expression of each gene. Each probe set consists of multiple probe pairs. Moreover, each pair consists of a perfect match (PM) and a mismatch (MM) probe. The PM probe matches the gene exactly, while the MM probe is different in one base in the center of the probe which strongly disrupts hybridization with the gene of interest. The purpose of the MM probe is to quantify background hybridization. The probes of a single gene are positioned randomly in order to protect against local hybridization artifacts. This process highlights another difference between spotted microarrays and Affymetrix chips. Since most spotted arrays use a single probe per gene, local hybridization artifacts may appear. Furthermore, in order to assess the expression level of each gene, Affymetrix uses a standard algorithm in order to merge the measurements of all probe pairs into a single number, while other alternative approaches are also available. Affymetrix chips are single sample (also called single color or single channel) microarrays. That is, each chip measures the gene expression of a single sample and if a comparison of two or more samples is to be performed (which is usually the case), a separate chip must be used for each sample [25] and the measurements must be scaled and normalized in exactly the same manner. Finally, Affymetrix chips suffer from light refraction caused by the masks, so that it leaks into overlapping features causing distortions in the signal being read. However, that “defect” is compensated by dedicated software so that it does not appear in the final signal presented to the user [24].

Other Microarray Technologies

Other, less popular microarray technologies exist. Maskless photodeprotection, is similar the methodology of Affymetrix but uses micromirror arrays instead of masks. Inkjet array synthesis [24] [25] was developed by Agilent (a spinoff of Hewlett Packard). Droplets of the desired base are placed on the appropriate spot of the glass slide via nozzles of inkjet printers, at each stage of synthesis. Instead of different colors of ink, the cartridges of the printer contain A, C, G and T nucleotides. Like micromirror arrays, inkjet array synthesis allows for a great degree of flexibility during microarray production, since the scientist can generate any oligonucleotide required for the experiment.

2.1.5 Basics Steps of a Microarray Experiment

Performing a microarray experiment consists of four basic steps [24] [25]

(1) Sample preparation and labeling
(2) Hybridization of target and reference samples
(3) Washing
(4) Image acquisition and processing

Sample Preparation and Labeling

The first step is always to extract the RNA from the tissue(s) of interest. This procedure can be difficult to precisely reproduce and can lead to variability among independent executions of the same experiment. The labeling step depends on the type of microarray used. While other platforms can be used to hybridize cRNA, it is a common practice to hybridize cDNA on other types of microarrays such as spotted arrays [24], due to RNAs inherent chemical instability [27]. Hybridizing cDNA requires denaturing, that is breaking up cDNA into its individual strands [27]. In the past DNA was radioactively labeled, but most laboratories use fluorescent labeling with two dyes: green Cyanine 3 (Cy3) and red Cyanine 5 (Cy5). Cy3 is excited at 550 nm using a green laser and its peak emission is at 581 nm [24]. Cy5 is excited at 649 nm using a red laser and its peak emission is at 670 nm [24]. The most common way of labeling is to directly incorporate reverse transcriptase in order to convert the mRNA into labeled cDNA. That is performed using the process of reverse transcription while adding some bases (usually Cytosine) that have already been marked by fluorescent dyes [24]. Another way is indirect labeling, which marks the Cdna at a subsequent step and has the advantage of incorporating Cy3 as efficiently as Cy5, contrary to direct labeling which is less efficient at incorporating Cy5

Hybridization

Hybridization refers to the step where the DNA probes on the glass slide are paired with the target mRNA or cDNA and for heteroduplexes via Watson-Crick base-pairing [24]. It is a process affected by many factors such as temperature, humidity, salt and formamide concentrations and volume of target solution. The two main methods for performing hybridization are manual and robotic, which leads to less variability of results.

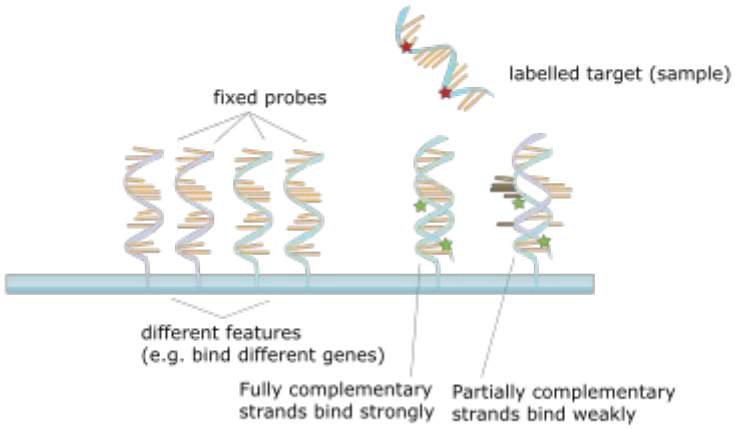

Figure 2.2: Graphical illustration of hybridization between the probes and the target.

Washing

After Hybridization, the slides are washed in order to remove excess hybridization solution from the array [24]. Washing ensures that only the labeled target strongly bound on the features is kept on the array, which represents the target that needs to be measured in the experiment. Moreover, washing reduces cross-hybridization. That is, labeled solution that has weakly bounded to the probes due to some degree of similarity with the target. Products of cross-hybridization lead to increased noise in the final signal measured [2]. Most automatic hybridization stations include a washing cycle as part of the standard process [24]. Finally, the microarray is dried using a centrifuge or by blowing clean compressed air [26].

Image Acquisition

The final step is scanning the array and producing an image of its surface [24]. The spots that are bound to target containing dye that fluoresces when excited with light of the appropriate wavelength. The microarray is placed in a scanner which reads the surface of the slide. The scanner contains one or more excitation lasers, depending on the number of colour channels supported (usually green and red). Each pixel of the resulting image represents the intensity of fluorescence induced by focusing the laser at a specific point on the array and exciting the dye present at that spot. The fluorescence is detected by a photomultiplier tube (PMT) in the scanner. In order to scan the whole array, the laser must subsequently excite each spot of the array.

Image Processing and Raw Data Preprocessing

After the scanned image has been acquired, it is processed in order to convert the light intensity of each spot into a gene expression matrix where each a numerical value corresponds to the gene expression level of each gene. Given the raw data, they are preprocessed before being passed on as input to bioinformatics algorithms. Typical preprocessing steps are transformation to log intensities, missing value estimation and normalization in order to make results from different microarray experiments comparable. All the above steps are explained in detail in [24] [26]

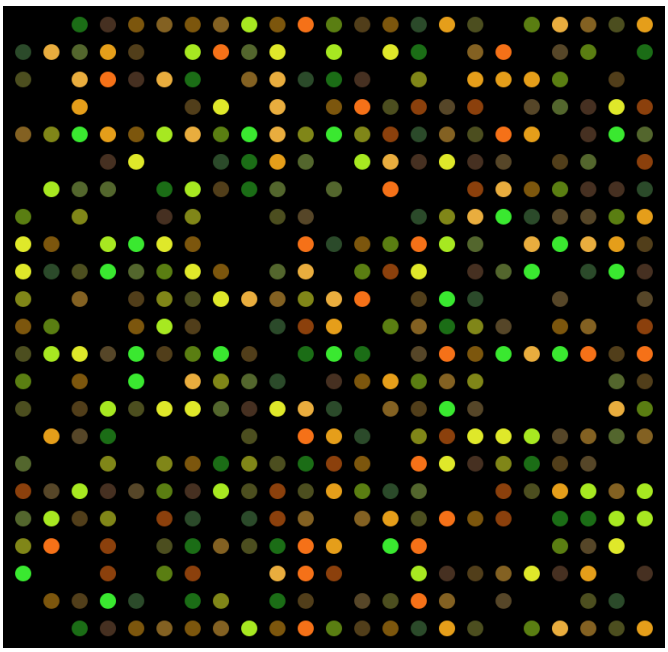

Figure 2.3: Graphical illustration of a 2-channel DNA microarray image.

2.2 Machine Learning Applied to DNA Microarray Data

Machine learning, also called pattern recognition, statistical learning and data mining is the act of performing an desired action based on the patterns observed in data. Exceptional books on the field of machine learning are [32] [33] [34] [35] [63]. Moreover, [32] is freely available online. At subsequent sections [32] will be usually cited since it is easily available, however there topic of interest will most likely also be covered in the other suggested books if the reader is interested in an additional view of the same topic.

Supervised learning aims to generate a function given a set of labeled samples. This function can then be used to assign labels to new unknown data. In regression, the label of each sample, is a continuous variable, often called the response variable. In the case of classification, the label can only take one among a set of discrete values. A typical scenario of supervised learning is the prediction of the value of a response variable given a set of values of known observed variables, if the response variable is continuous we face a regression problem. If the response variable is discrete, we face a classification problem. On the rare case that the response variable is ordinal, we face a problem of ordinal regression. Ordinary regression and classification methods should not be used for ordinal regression. If simple regression is used, the implicit assumption is made that different levels of the response variables have the same distance, which might not be the case. On the other hand, while a classification method could be used to map the ordinal variables into discrete levels, the model does not take advantage of the information that the response is ordinal, which could lead to reduced predictive performance [40].

Unsupervised learning aims to find groups of data that share similar properties. It differentiates from supervised and reinforcement learning, since the samples are unlabeled and there is no explicit feedback. A typical scenario of unsupervised learning is the clustering of a set of samples into groups/clusters according to the values of known observed variables, in order to find unknown “hidden” structure in the data.

Reinforcement learning is usually employed by Artificial Intelligence (A.I.) agents and aims to maximize a cumulative reward function, given a set of variables determining the environment and the actions available at a given time. Instead of labels, reinforcement learning utilizes a positive or negative reward signal sent to the agent after an action is completed.

In practice, both supervised and unsupervised methods are used for the analysis of DNA microarray data. The gene expression data are represented as a data matrix of N samples (rows) and P

features/predictors/genes (columns) which can be expressed in array form as $\boldsymbol{X} \in R^{N,P}$ where each row represents a sample containing the expression values of P genes. It is common for the transpose of $\boldsymbol{X}$ to be used some times in literature, which is just a simple matter of convention. The class labels of all samples are represented as a vector $\boldsymbol{y} \in R^{N}$. To each of the samples, a class label $y_i,\ i=1,\ \ldots,\ N$ is assigned. In the case of cancer/control binary classification, $\boldsymbol{y}$ has a binary encoding such as $y_i \in \{-1,+1\}$ or $y_i \in \{0,1\}$.

2.3 Introduction to Classification

Problem Formulation

As mentioned in the previous section, the problem of classification refers to the prediction of a value of a discrete response variable (class label), given a set of known observed variables, called features. Let us suppose we are given a data matrix $\boldsymbol{X} \in R^{N,P}$ of N samples and P features and a known vector of class labels $\boldsymbol{y} \in R^{N}$ for each of the samples. We will call these samples the "training data". Given the training data and their class labels we would like to predict the class label of a new sample $\hat{\boldsymbol{x}} \in R^{P}$. We know the values of each of the features for $\hat{\boldsymbol{x}}$ e.g. gene expression values, but the class label $\hat{y}_i$ of $\hat{x}$ is unknown e.g. "Cancer" or "Control" in a case of binary classification.

2.3.1 The K-Nearest Neighbors Classifier

The K Nearest Neighbors (K-NN), is a very simple and intuitive method to solve the problem of classification. K-NN utilizes a non-linear approach that classifies new samples depending on the set of samples closest to them, which are called their "nearest neighbors". Given a set of known training samples, K-NN classifies a new test sample depending on the class label of the majority of K samples nearest to it, according to a given distance metric. Supposing that in the case of binary classification the dataset is $D=\{(\boldsymbol{x_n}, y_n): \boldsymbol{x_n} \in R^{P}\ y_n \in \{-1,+1\}\}, n=1,\ldots,N$, then a new sample $\hat{\boldsymbol{x}}$ is given a class label $\hat{y}$ according to the formula $\hat{y}=sign(\sum_{i=1}^{K} \widetilde{y}_i)$. Where $\widetilde{y}_i$ is the class label corresponding to the i-th nearest neighbor of $\hat{\boldsymbol{x}}$. In the case Euclidean distance is used, the nearest neighbor of $\hat{\boldsymbol{x}}$ is expressed as $\widetilde{\boldsymbol{x}}=argmin\|\hat{\boldsymbol{x}}-\boldsymbol{x_i}\|, i=1,\ldots,N-1$.

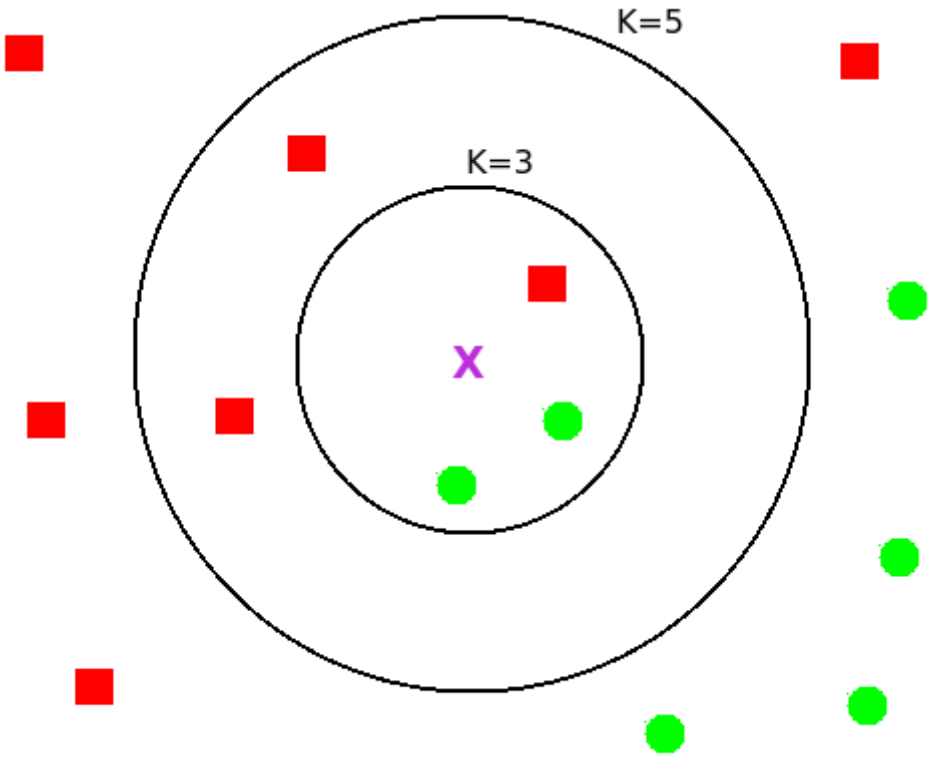

Figure 2.4 The test sample (purple X) will be classified in the first class of green circles in the case of K=3. However, in the case of K=5 it will be classified in the second class of red rectangles.

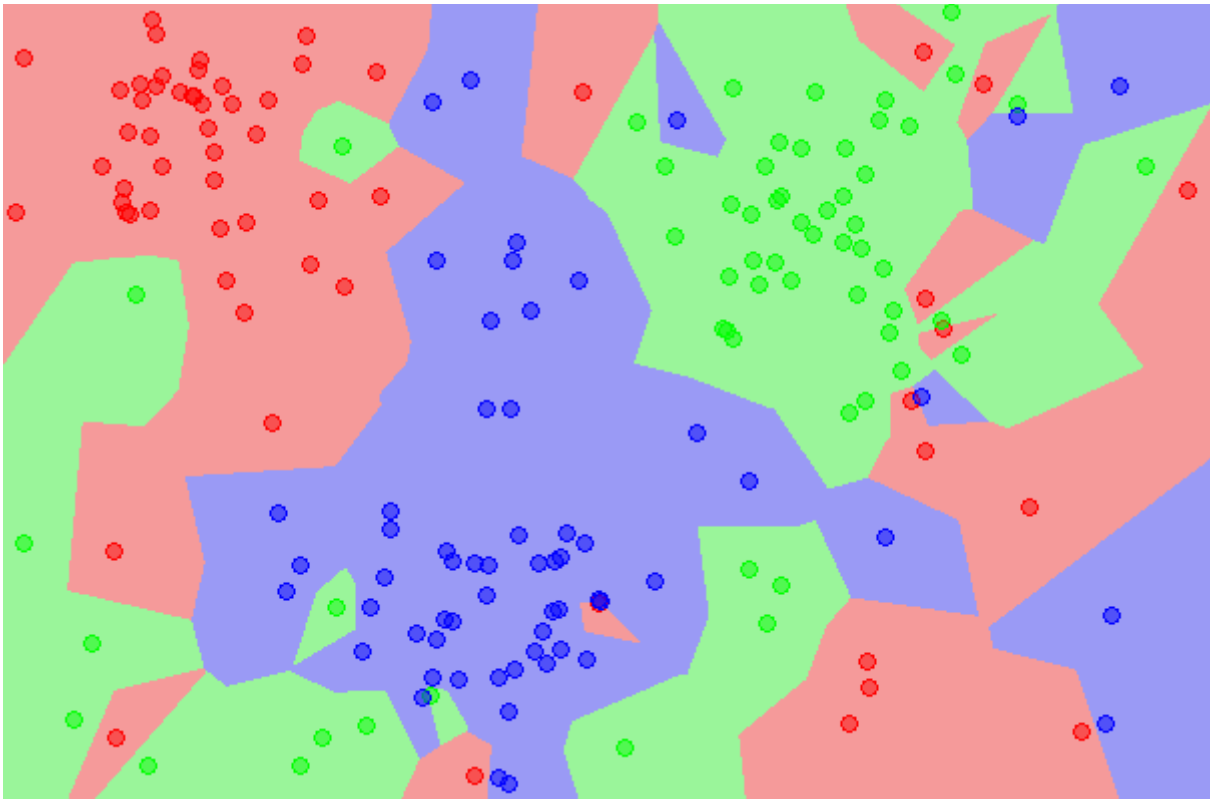

Figure 2.5 The class borders of an 1-NN classifier In the case of 3-way classification.

2.3.2 The Support Vector Machine (SVM) Classifier

Support Vector Machines [32] [33] [41] are a machine learning algorithm than can be used for regression, and classification purposes. In the case of two-way classification, the SVM computes the hyperplane separating the classes of interest with the maximum margin across the closest samples of the two classes. The aim of the utilization of the maximum margin hyperplane is to minimize the generalization error of the classifier. The original SVM algorithm assumes that the data are linearly separable. If that is not the case, using a kernel function the data are mapped to a higher dimension space in which they are found to be linearly separable. Moreover, the SVM algorithm has been extended to what is called the “soft margin” SVM, that makes no assumption about the linear separability of the classes. Instead it normally functions as a typical SVM but in case the data are not linearly separable, it utilizes “slack variables” and computes the hyperplane resulting in the lowest mis-classification rate, while it ensures the maximum margin between the closest correctly classified samples of the two classes. In order to understand the notion of the support vectors, the case of the simple SVM given linearly separable data is further explained. Further information, including the extension of the SVM for non-linearly separable data can be found in [33] the suggested reading books of section 2.2.1. Details concerning the convex optimization methods used to compute the SVM parameters can be found in [44]. In the case of the soft-margin SVM, the parameter C, which is an inverse regularization parameter, needs to be specified before training of the method. This is usually accomplished in practice using cross validation. The cross validation method will be explained in a subsequent section of this thesis.

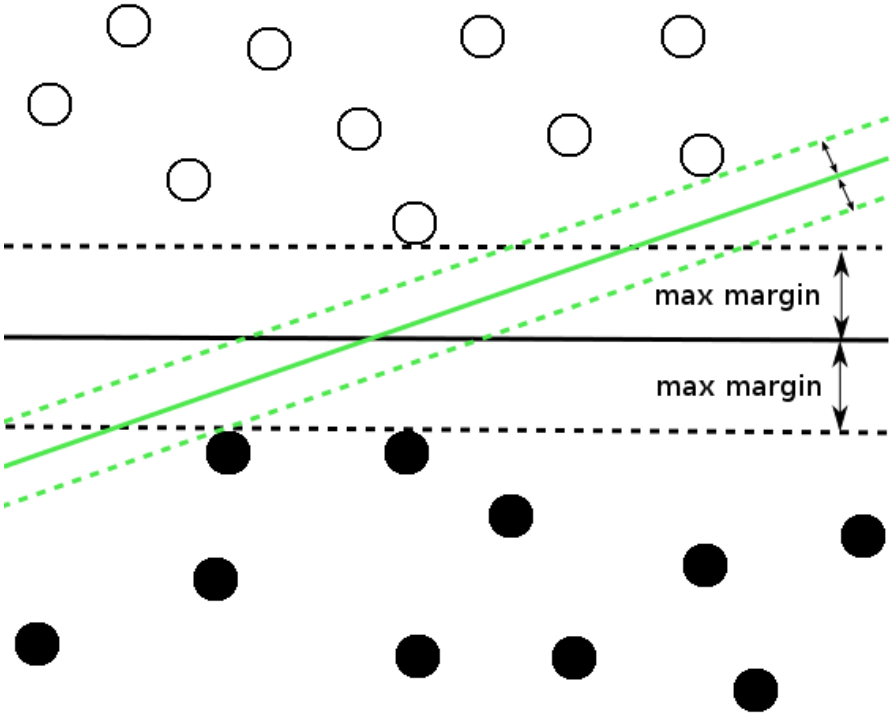

Figure 2.6 The black hyperplane separates the two classes, resulting in the maximum margin between their closest samples, and thus is selected as the SMV separating hyperplane.

Linear SVM

Given a dataset $D=\{(\boldsymbol{x_i}, y_i):\boldsymbol{x_i}\in R^P\ y_i\in\{-1,+1\}\}, i=1,..,N$ where $\boldsymbol{x_i}$ the samples and y_i the class labels, the goal of the SVM is to compute the hyperplane of dimension $R^{(P-1)}$ that separates all samples belonging to the class y=1 from those of y=-1, such as the margin of the closest samples of the two classes is maximized. If $\boldsymbol{x}\in R^P$ then any hyperplane can be expressed as $\boldsymbol{w}\cdot\boldsymbol{x}-b=0$, where **w** the normal vector to the hyperplane and b a real constant. Then the parameter $\frac{b}{\|\boldsymbol{w}\|}$ expresses the offset of the hyperplane from the origin, along the normal vector **w**. Given that the data are linearly separable, there exist two hyperplanes $H_1: w\cdot x-b=1$, $H_2: w\cdot x-b=-1$ that fully separate the two classes without any samples being misclassified. The region bounded by these two hyperplanes is called the "margin" between the two classes, which is equal to $\frac{2}{\|w\|}$. So in order to maximize the margin, $\|w\|$ needs to be minimized. While $\|w\|$ is minimized, samples of either class may appear inside the margin, for that to be avoided, further constraints need to be implemented:

$\boldsymbol{w}\cdot\boldsymbol{x_i}-b\geq 1$ for samples of class $y_i=1$ and $\boldsymbol{w}\cdot\boldsymbol{x_i}-b\leq -1$ for samples of class $y_i=-1$.

Both constraints can be expressed in one equation as $y_i\cdot(\boldsymbol{w}\cdot\boldsymbol{x_i}-b)\geq 1$ for i=1, ..., N. The above can be expressed as an optimization problem:

Minimize in w,b

$\|\boldsymbol{w}\|$

subject to $y_i\cdot(\boldsymbol{w}\cdot\boldsymbol{x_i}-b)\geq 1$ **, for i=1,...,N**

or to avoid calculating the square root:

Minimize in w,b

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

subject to $y_i\cdot(\boldsymbol{w}\cdot\boldsymbol{x_i}-b)\geq 1$ **, for i=1,...,N**

By introducing the Lagrange multipliers $\boldsymbol{\alpha}$, the above can be expressed as a problem of quadratic programming:

$\min_{w,b}\max_{\boldsymbol{\alpha}\geq 0}\{\frac{1}{2}\|\boldsymbol{w}\|^2-\sum_{i=1}^{N}\alpha_i[y_i\cdot(\boldsymbol{w}\cdot\boldsymbol{x_i}-b)-1]\}$ and then according to the stationary Karush-Kuhn-Tucker [44] condition, the solution can be expressed as a linear combination of the training input vectors $\boldsymbol{x_i}$:

$\boldsymbol{w}=\sum_{i=1}^{N}\alpha_i y_i \boldsymbol{x_i}$.

Only a few of the Lagrange multipliers α_i are greater then zero. These multipliers correspond to the closest samples of the two classes, the support vectors, that lie on the margin and satisfy $y_i\cdot(\boldsymbol{w}\cdot\boldsymbol{x_i}-b)=1$. Solving the previous equation for b we obtain $b=\boldsymbol{w}\cdot\boldsymbol{x_i}-y_i$ for a given support vector. In that manner, a more stable estimation of b is the mean value over all support vectors, given by the formula $\hat{b}=\frac{1}{N_{sv}}\sum_{i=1}^{N_{sv}}(\boldsymbol{w}\cdot\boldsymbol{x_i}-y_i)$.

Using the equations $\|\boldsymbol{w}\|=\boldsymbol{w}\cdot\boldsymbol{w}$ and $\boldsymbol{w}=\sum_{i=1}^{N}\alpha_i y_i \boldsymbol{x_i}$ the optimization problem can be expressed in it's dual form as:

Maximize in α_i

$L(\alpha)=\sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j \boldsymbol{x_i}^T\boldsymbol{x_j}=\sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j K(\boldsymbol{x_i},\boldsymbol{x_j})$

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

where $K(\boldsymbol{x_i},\boldsymbol{x_j})=\boldsymbol{x_i}\cdot\boldsymbol{x_j}$ a kernel function.

After the Lagrange multipliers α_i have been computed, $\boldsymbol{w}$ can be determined using $\boldsymbol{w}=\sum_{i=1}^{N}\alpha_i y_i \boldsymbol{x_i}$.

The problem expressed in dual form is computationally efficient, since the classification task takes into consideration only the support vectors, which generally are a small subset of the original set of training samples.

2.3.3 The Relevance Vector Machine (RVM) Classifier

The Relevance Vector Machine for classification is a special case of Bayesian Logistic Regression [43] [33] 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. It was proposed as an alternative to the SVM, having several advantages such as: (1) the RVM output is a posterior probability, instead of a "hard" decision, (2) it has a well defined extension for multiclass problems and (3) no parameter (such as C of the SVM) needs to be computed before training.

Being a discriminative model, the RVM directly models the posterior probability of a class C_k given a sample $p(C_k|\boldsymbol{x})$. The RVM requires class labels (targets) of the form $t\in\{0,1\}$, where $t_i=1\Rightarrow \boldsymbol{x_i}\in C_1,\ t_i=0\Rightarrow \boldsymbol{x_i}\in C_2$ in the case of binary classification. Since it is a special case of Bayesian Logistic Regression, it computes a model of the form $y(\boldsymbol{w},\boldsymbol{x})=\sigma(\boldsymbol{w^T}\cdot\boldsymbol{\varphi}(\boldsymbol{x}))$ where $\sigma(.)$ the logistic sigmoid function $\sigma(\alpha)=\frac{1}{1+\exp(-\alpha)}$ and $\varphi(\boldsymbol{x})$ a basis function [33]. According to the RVM model, each basis function $\varphi(\boldsymbol{x})=k(\boldsymbol{x},\boldsymbol{x_n})$ is given by a kernel and each kernel is associated with one data point. The basis functions may also include a constant term, usually called intercept or bias. The ARD priors are have the following form $p(\boldsymbol{w}|\boldsymbol{\alpha})=\prod_{i=1}^{M}N(w_i|0,\alpha_i^{-1})$. During the ARD process, many of the α_i are led to infinity and the corresponding features are eliminated from the model.

RVM for binary classification

The goal of the process is to calculate the probability $p(C_1|,\boldsymbol{x}')$ of an unknown sample $\boldsymbol{x}'$ belonging to C_1 . Since the problem is binary classification, the probability of $\boldsymbol{x}'$ belonging to the second C_2 class is $1-p(C_1|,\boldsymbol{x}')$. According to the bayesian framework, no single value is calculated for each weight bit the posterior $p(\boldsymbol{w}|\boldsymbol{t},\alpha)$ is estimated and is marginalization over w is performed. The predictive distribution of the unknown sample is $p(C_1|,\boldsymbol{x}')=\int p(C_1|\boldsymbol{x}',\boldsymbol{w})p(w|\boldsymbol{t},\boldsymbol{a})\cdot d\boldsymbol{w}=\int \sigma(\boldsymbol{w^T}\cdot\boldsymbol{x}')p(\boldsymbol{w}|\boldsymbol{t},\boldsymbol{a})\cdot d\boldsymbol{w}$. The RVM training procedure consists of the following steps:

Step 1) Select the form of the priors, which in this case are the ARD priors $p(\boldsymbol{w}|\boldsymbol{\alpha})=\prod_{i=1}^{M}N(w_i|0,\alpha_i^{-1})$

Step 2) Automatic Relevance Determination: find the optimal value for **α** through type-2 maximum likelihood, the steps 2a and 2b are repeated until the value of **α** converges.

Step 2a) Find the posterior distribution of the feature weights **w**. Since the posterior of w leads to intractable integrals in the predictive distribution and the marginal likelihood, it is approximated using the Laplace approximation [33]. To be precise, the mode **w*** and covariance matrix **Σ** of the posterior $p(\boldsymbol{w}|\boldsymbol{t},\boldsymbol{\alpha})$ are calculated. Then, a Gaussian approximation is fitted $q(\boldsymbol{w})=N(\boldsymbol{w}|\boldsymbol{w}^*,\boldsymbol{\Sigma})\approx p(\boldsymbol{w}|\boldsymbol{t},\boldsymbol{\alpha})$. In this thesis, Newton Raphson method with backtracking line search [44] is used to find the mode and covariance matrix of the posterior $p(\boldsymbol{w}|\boldsymbol{t},\boldsymbol{\alpha})$.

Step 2b) The marginal likelihood is calculated through the Laplace approximation $p(\boldsymbol{t}|\boldsymbol{a})=\int p(\boldsymbol{t}|\boldsymbol{w})p(\boldsymbol{w}|\boldsymbol{a})\cdot d\boldsymbol{w}\approx\int p(\boldsymbol{t}|\boldsymbol{w})q(\boldsymbol{w})\cdot d\boldsymbol{w}=p(\boldsymbol{t}|\boldsymbol{w}^*)p(\boldsymbol{w}^*|\boldsymbol{a})(2\pi)^{M/2}|\Sigma|^{1/2}$.

By setting $\nabla_\alpha p(\boldsymbol{w}|\boldsymbol{t},\boldsymbol{\alpha})=0$ a new value for **α** is calculated and step 2a can be repeated.

Step 3) Calculate the predictive distribution for the unknown sample $\boldsymbol{x}'$: $p(C_1|,\boldsymbol{x}')=\int p(C_1|\boldsymbol{x}',\boldsymbol{w})\,p(w|\boldsymbol{t},\boldsymbol{a})\cdot d\,\boldsymbol{w}=\int \sigma(\boldsymbol{w}^T\cdot\boldsymbol{x}')\,p(\boldsymbol{w}|\boldsymbol{t},\boldsymbol{a})\cdot d\,\boldsymbol{w}\approx\int \sigma(\boldsymbol{w}^T\cdot\boldsymbol{x}')\,q(\boldsymbol{w})\cdot d\,\boldsymbol{w}$.
The integral and by extent the predictive distribution can be approximated [33] by $P(C_1|,\boldsymbol{x}')\approx\sigma(\kappa(\sigma_\alpha^2)\mu_\alpha)$, where $\kappa(\sigma^2)=(1+\pi\sigma^2/8)^{-1/2}$, $\mu_\alpha=\boldsymbol{w}_{\boldsymbol{MAP}}^{\boldsymbol{T}}\cdot\boldsymbol{\varphi}=\boldsymbol{w}^{*T}\cdot\boldsymbol{\varphi}$, $\sigma_\alpha^2=\boldsymbol{\varphi}^T\boldsymbol{S}_N\boldsymbol{\varphi}=\boldsymbol{\varphi}^T\boldsymbol{\Sigma}\boldsymbol{\varphi}$ while the values of **w*** and **Σ** have been calculated using the Newton Raphson method in the previous step.

2.4 Feature Subset Selection (FSS)

Feature subset selection [49] [50] is an important aspect of microarray analysis, since it aims to counter the “curse of dimensionality” that is encountered in DNA microarray datasets. That is, classifier performance deteriorates when the number of features is larger than the number of available training samples. The goal of FSS methods is to reduce the number of features by keeping only the most “important” set which is considered to be the most relevant to the response variable (phenotype), while discarding all others. The set of kept features is then used for classification. In DNA microarray analysis, the set of kept features (genes) is usually referred to as “genomic signature” [31]. There are three different approaches to feature subset selection: filter, wrapper and embedded methods.

2.4.1 Filter Methods (Univariate) and Significance Analysis of Microarrays

Filter methods [49] [50] form univariate approaches, which act as a preprocessing step, independent of the classifier used. They rank each feature independent of others, based on its ability to discriminate between different classes of interest. They generally are simple to implement, computationally efficient and provide insight into class differences. However, filter methods produce a feature set that is not tuned to the performance of a specific classifier. Moreover, filter methods do not model the dependencies among the features (genes) in the dataset.

Significance Analysis of Microarrays (SAM)

Significance Analysis of Microarrays [52] is a popular filter method that utilizes a t-test like statistic along with a permutation test. It assesses the expression pattern of each gene separately and identifies which genes are significantly over or under expressed between classes of interest, such as between cancer and control samples. The genes showing no significant change in their expression pattern across the classes of interest are discarded and the differentiating genes are selected for subsequent analysis. One advantage of SAM is that it can adjust the thresholds according to which genes are considered significantly over or under expressed, in order to achieve an estimated False Discovery Rate (FDR) that is below a given threshold, e.g. 5%. The SAM procedure is as follows: First, the relative difference for each gene between the classes of interest is calculated using the formula provided in [52]. Then, the “null” relative difference is calculated for each gene using a given number of permutations, according to the standard procedure of a permutation test [47] [46]. Finally, genes are declared over or under expressed if their relative difference is significantly different from their “null” relative difference (called expected relative difference in [52]). That is, if the difference between the relative difference and expected relative difference is greater that a threshold which is implicitly defined by the desired FDR level. Finally, it has been implemented in the R package 'samr' [53] and is freely available.

2.4.2 Wrapper Methods (Multivariate) and RFE-SVM

Wrapper methods [49] [50] fall within a multivariate approach. They evaluate a feature subset based on the prediction accuracy of the classifier when that specific subset is used. In that manner, given a

classifier, they aim to find the set of features which maximizes the prediction performance. Yet, the classifier is perceived as a black box, independent of the feature selection method. Moreover, due to their multivariate nature, wrapper methods manage to model feature dependencies, unlike filter methods. However, since they need to evaluate different combinations of features, they can be computationally expensive. In that manner, greedy algorithms have been proposed in order to reduce the computational complexity, such as forward selection and backward elimination. As a result, wrapper methods are prone to overfitting due to their heuristic nature.

Recursive Feature Elimination (RFE) and RFE-SVM

Recursive Feature Elimination [49] [50] [51] [42] is a popular wrapper feature selection method that aims at preserving the minimal set of features maximizing the classification accuracy of a given classification method. RFE proceeds iteratively, eliminating a fixed number of least significant features during each iteration and then reassessing the classification performance. The elimination procedure stops when a predetermined small number of features are left. Then, the set of features across all iterations maximizing the classification accuracy is chosen as the optimal feature set, tuned for the specific classifier used. In order for the least significant feature to be determined, a feature weighting scheme is required. Such a weighting scheme can be the weight given to each feature by a classifier. A popular choice is to utilize RFE in conjunction with the SVM classifier and the resulting method is RFE-SVM [42], which is very popular in literature. Its popularity lies in its effectiveness, since it produces good genomic signatures and is computationally efficient (for a multivariate method). Especially, in the case where more than one features are eliminated during each round, which greatly reduces the number of required iteration at the cost of "resolution" of the gene sets being tested. In this thesis, RFE-SVM was set to eliminate half of the feature during each iteration, which lead to a very computationally efficient algorithm, contrary to removing the features one by one. While, removing half the features during each iteration limits the candidate gene sets during the RFE step, this limitation is overcome using the gene frequencies at the stable signature extraction step which is introduced later in the Methodology section. Another advantage that leads to the computational efficiency of RFE-SVM is that the SVM is a very popular state of the art classifier and very efficient implementations of it are publicly available for most programming languages, including R. This is contrary to less popular methods, where one has to write custom code, which in most cases is bound to be less efficient than popular packages maintained by groups of developers.

2.4.3 Embedded Methods (Multivariate) and RVM

Embedded [49] [50] methods also evaluate a feature subset based on the prediction accuracy of the classifier. They differentiate from wrapper methods however, since the search for the feature subset is embedded in the training of the classifier, while in wrapper methods the feature selection step is independent of the classifier used. Compared to wrapper methods, embedded methods usually are more computationally efficient than simple wrapper approaches. However, due to the embedding of feature selection in the training process, they can prove to be harder to implement. Moreover, since they are multivariate, embedded methods also successfully model the interactions between the features (genes). The Relevance Vector Machine introduced in a previous section implements embedded feature selection through by utilizing the Automatic Relevance Determination priors. As a result, the RVM selects a subset of features during training, excluding all other features from the final model. That is, each feature is assigned an ARD prior at the beginning of the training process. During training, the ARD priors of irrelevant features are assigned very large values and rise to infinity (in theory). In practice, for the code used in this thesis appropriate thresholds were set to define which features are considered irrelevant, based on the value of the ARD prior assigned to them during training.

2.5 Gene Set Analysis Methods and Globaltest

Gene set analysis methods [55] [54] [56] aim to assess whether a predetermined gene set as a whole is related to a pathological state of interest, instead of assessing the behavior of each gene independently. As a result a single p-value is associated with the gene set as a whole and the drawbacks of multiple testing are avoided. Gene set analysis methods are divided into two main categories: Competitive and Self Contained methods. Competitive methods aim to compare the gene set with its complement in terms of association, such as differential expression, with biological process of interest. Popular competitive methods include Gene Set Enrichment Analysis (GSEA) [58] and Gene Set Analysis (GSA) [59]. The competitive null hypothesis states that “the genes in the gene set are at most as differentially expressed as the genes in the complement”, yet the most popular competitive methods do not explicitly test this null hypothesis [54]. On the other hand, self contained methods assess the association of the gene set with the biological process of interest, focusing only on the genes in the gene set and ignoring the genes of the complement. The self contained null hypothesis states that “no genes in the gene set are associated with the phenotype”. It should be pointed out that gene set analysis methods assess the quality of predetermined gene sets and do not perform feature (gene) selection on their own.

Globaltest

Globaltest [55] is a popular self contained gene set analysis method. It tests the null hypothesis that the covariates of a generalized linear model (genes) are not associated with the response (phenotype) against the alternative that they are. By utilizing the general linear model framework, the Globaltest can be used for both linear regression and logistic regression (classification) scenarios. While it is a method that yields good results in practice, it has been observed that significance of gene sets according to Globaltest might be a result of a few genes being strongly associated with the phenotype [gt-review]. Finally, it has been implemented in the R package 'globaltest' [57] and is freely available.

2.6 Evaluation (Validation) Methods

Evaluation methods [32] [36] are used to estimate the ability of the model to generalize, that is to yield comparable results in unknown data as well data used during training. If all available data are used for training, there is no assessment of the FSS & classification performance on new data and as such, the generalization ability of the model remains unknown. In that manner, evaluation methods leave out a set of samples that are only used in order to assess the performance of the model on new data. That set of samples is called the test set, while the set of samples used while training the model is called the training set.

2.6.1 Holdout Validation

Holdout validation is probably the simplest validation method. It splits the available samples into two groups. The training set consists of the majority of available samples and is used for training the model while the test set corresponds to a smaller percentage of the available samples and is used in order to evaluate the model's generalization ability. However, excluding a portion of the dataset can be costly when the available samples are few. Moreover, the results obtained greatly depend on the random splitting of the dataset into training and test sets and the observed results are generally unstable and can be misleading if both splits are do not reflect the structure of the original dataset. Generally, holdout validation should only be used in practice if a very large number of samples is available, which is never the case in DNA microarray studies. To counter these drawbacks of the simple holdout method at the expense of computational load, other validation techniques have been proposed.

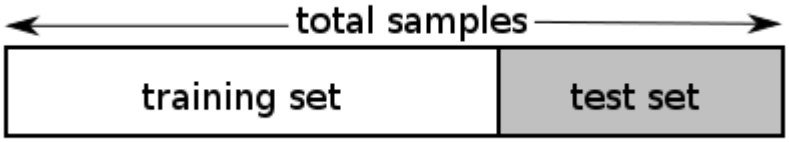

Figure 2.7 Holdout validation method.

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

K-Fold Cross Validation splits the dataset into K different subsets of approximately the same size, called folds. It then proceeds to iteratively use k-1 folds for training and 1 fold for testing the FSS & Classification model, using a different fold for testing during each iteration. At the end of the procedure, k different test statistics have been observed. The average statistics over all folds are then calculated. If for example the only test statistic examined is the classification accuracy, it is calculated using the following formula: $\bar{a} = \frac{1}{K}\sum_{k=1}^{K} a_k$. Typical values used for k are K=3, 5 or 10. As the number of folds increases, the bias of the estimate decreases, so the estimation of performance is representative of the actual performance of the method. However, the variance of the estimation as well as the computational cost increase due to the large number of iterations. If the cross-validation method is “stratified”, then the class ratio for all folds, is the same as in the original dataset. Moreover, in the case of multiple K-Fold CV, standard cross validation is repeated several times and the results are aggregated. For example, in 3x10-Fold CV, cross validation is performed three times while the overall results are aggregated.

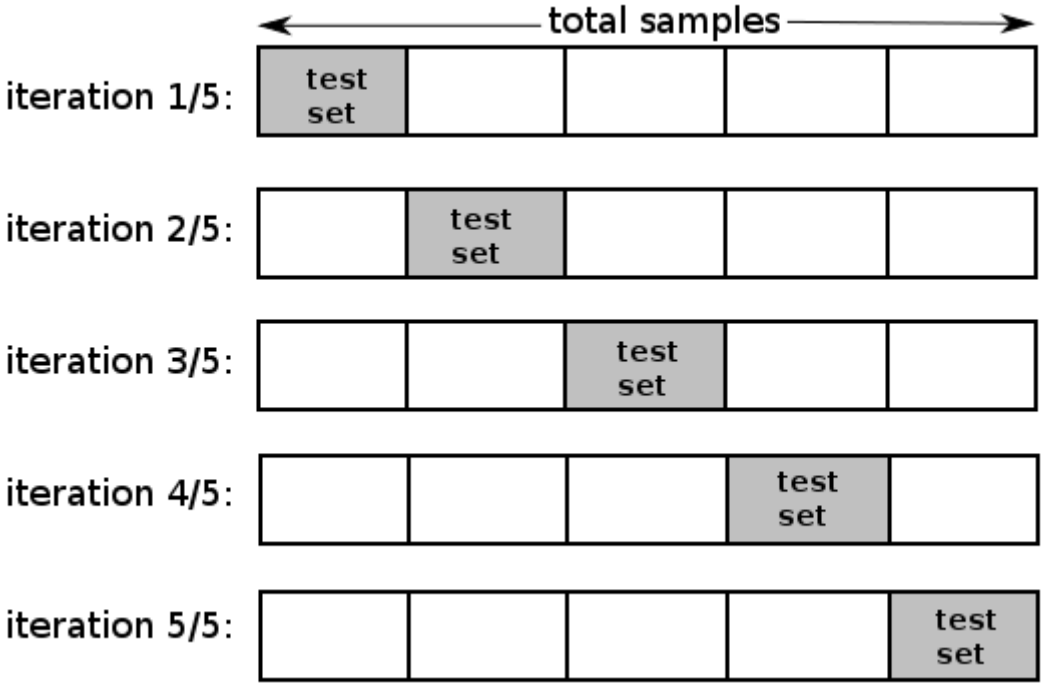

Figure 2.8 5-Fold Cross Validation.

2.6.3 Leave One Out Cross Validation (LOOCV)

Leave one out cross validation is a case of K-Fold CV where the number of folds K is equal to the number of samples in the dataset N. Since the number of samples is larger than the typical values of k used during simple K-Fold CV, LOOCV displays the characteristics of K-Fold CV when large K is utilized: small bias of the estimations accompanied by large variance of the test statistics as well as high computational cost. LOOCV is very useful in cases where only a few samples are available in a dataset. However, it is known to lead to performance estimates with large variance, due to the large overlap of the training datasets. As a result, it should only be utilized if the number of available samples is very limited.

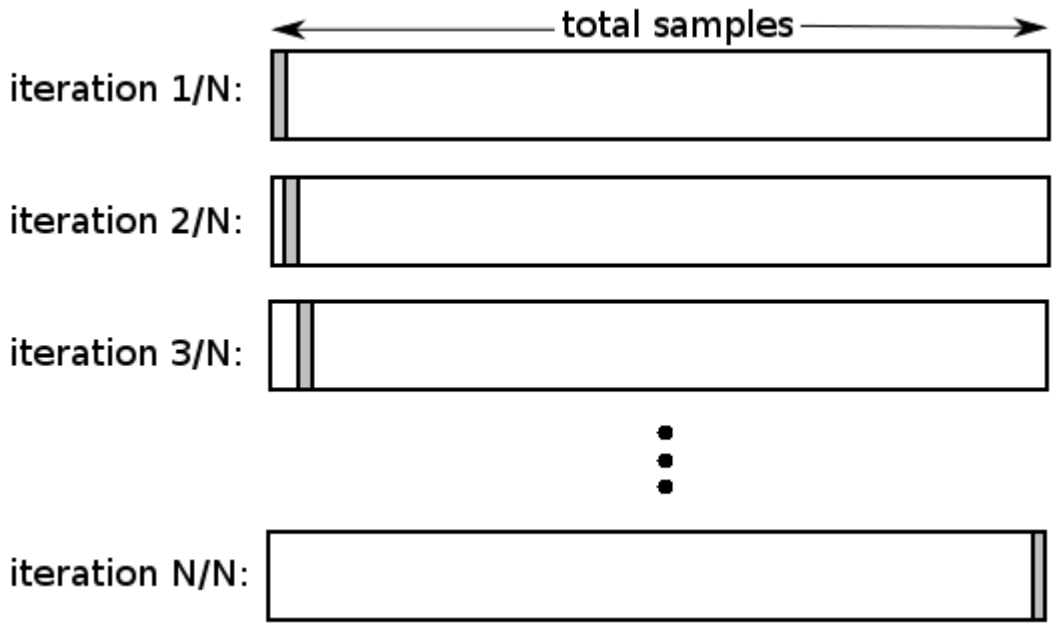

Figure 2.9 Leave One Out Cross Validation.

2.6.4 Repeated Random Sub-Sampling Validation

Repeated random sub-sampling validation is run for a fixed number of K iterations. During each iteration it utilized random sampling without replacement, in order to select a fixed number of S samples that make up the test set and are excluded from the training process of the model. The observed test statistics are then averaged over all iterations.

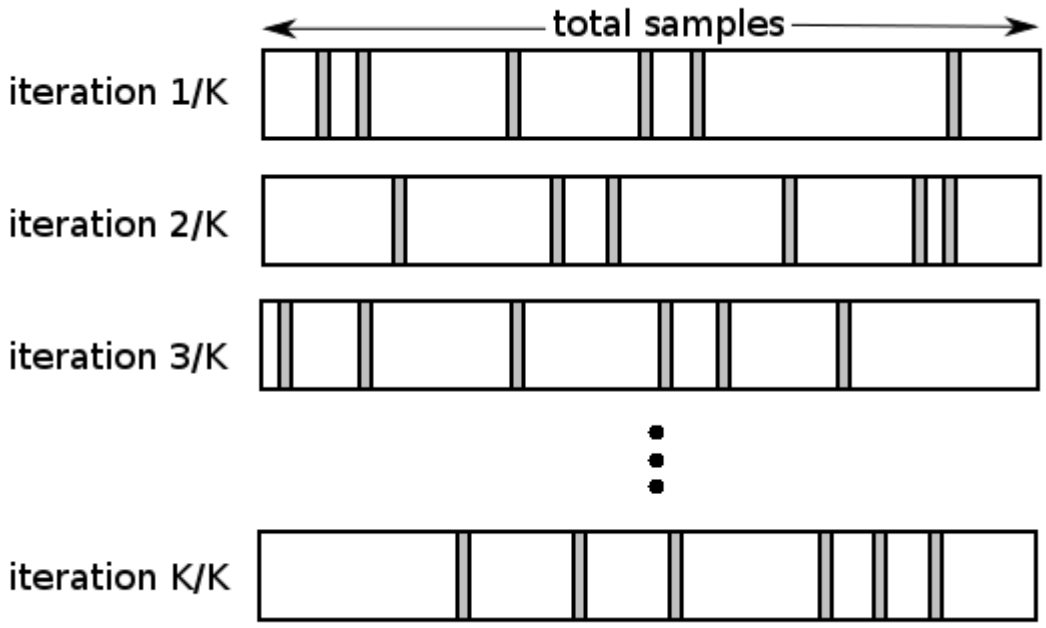

Figure 2.10 Repeated Random Sub-Sampling Validation.

2.6.5 Bootstrap Resampling Validation

Given an original dataset, bootstrap resampling, also called bootstrapping, utilizes random sampling with replacement in order to construct a number of B bootstrap datasets of fixed size, usually the same number of N samples as the original dataset. The class ratio in each dataset can either be random, or determined beforehand. Each bootstrap dataset can then be separated into training and test sets using the simple holdout method. The test statistics are then calculated for each bootstrap dataset and are averaged over all bootstrap datasets in order to get a stable estimation. The advantage of the bootstrap is its simplicity, however it may lead to overfitting and positively biased estimates of classification performance since some samples of the test set are also present in the training set due to the process of random sampling with replacement.

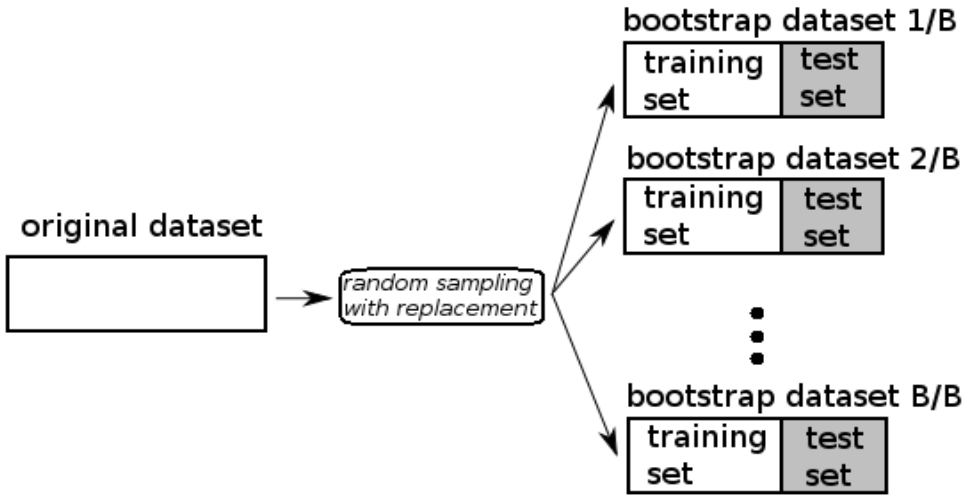

Figure 2.11 Bootstrap Resampling Validation. In this illustration, each bootstrap dataset is split using holdout validation as an example.

2.7 Random Oversampling and Undersampling for Classification of Imbalanced Datasets

In some classification problems, including DNA microarrays, there appears the problem of imbalanced datasets may yield misleading performance estimates [38] [39]. Classifiers are trained in order to minimize the overall error on the training set and as such, they may positively biased towards correctly classifying members of the majority class. If for example a dataset consists of 100 samples, 99 of which are cancerous and there is only 1 control sample, it is very likely that most classifiers will assign the 'cancer' label to the control sample, achieving 99% accuracy, which seems exceptional if one only checks the accuracy statistic. This raises the need for other statistics to be used in conjunction with accuracy. In this thesis, sensitivity (true positive rate – TRP, also called hit rate and recall) and specificity (true negative rate – TNR) are used. Accuracy is the total number of samples classified correctly, sensitivity is defined as the proportion of true positive samples across all samples classified as positive and specificity is defined as the proportion of true negative samples across all samples classified as negative. Let TP and TN refer to true positives and true negatives, that is those samples that have correctly been classified as positive or negative and FP and FN the false positive and false negative samples. Then these metrics can be expressed as [37]:

$$Accuracy = \frac{TP+TN}{P+N} = \frac{TP+TN}{TP+FN+TN+FP}$$

$$Sensitivity\,(TPR) = \frac{TP}{P} = \frac{TP}{TP+FN}$$

$$Specificity\,(TNR) = \frac{TN}{N} = \frac{TN}{TN+FP}$$

Another statistic frequently used in literature is the false positive rate (FPR) which is defined as FPR=1-TNR. Let us consider the previous example, where 99% accuracy was achieved. As a convention, let us assume that the cancerous samples are labeled as “positive” and the control sample is labeled as “negative”. Since the 1 control sample would be a false positive, it would result in Sensitivity 99/99 = 100% and specificity 0/1 = 0. Hence, the problematic behavior of the classifier is diagnosed thanks to the specificity metric.

In order to avoid the biased behavior of the classifier towards the class with the majority of samples, two simple sampling methods are common [38] [39]. Random oversampling, also called random minority oversampling (ROS) and random undersampling, also called random majority undersampling (RUS). When random oversampling is performed, random samples of the minority class in the dataset are duplicated until a predetermined class ratio is achieved between the classes. When random undersampling is performed, random instances of the majority class are excluded from the dataset until a predetermined class ratio is achieved between the classes, such as 1:1 (same number of samples for each class).

2.8 The Law of Large Numbers

The weak law of large numbers (LLN) [45] [46] [47] [48] is a theorem of probability theory which states that given that a random experiment is executed a sufficiently “large” number of times, the mean value of the observed results will be close to the expected value, and will continue to converge as more experiments are performed. Stated formally, the theorem suggests that given a set of independent identically distributed (i.i.d) random variables $X_{1,}\ldots,X_n$, each having a mean $\bar{X}_i=\mu$ and variance $var(X_i)=\sigma^2$.

A new random variable X can be defined, such as $X\equiv\frac{X_1+\ldots+X_n}{n}$.

Then, as the number of trials $n\to\infty$: $\bar{\bar{X}}=\frac{\overline{X_1+\ldots+X_n}}{n}=\frac{\bar{X}_1+\ldots+\bar{X}_n}{n}=\frac{n\cdot\mu}{n}=\mu$.

Moreover $var(X)=var(\frac{X_1+\ldots+X_n}{n})=var(\frac{X_1}{n})+\ldots+var(\frac{X_n}{n})=\frac{\sigma^2}{n^2}+\ldots+\frac{\sigma^2}{n^2}=n\cdot(\frac{\sigma^2}{n^2})=\frac{\sigma^2}{n}$

and by the Chebyshev inequality, for all ε>0:

$P(|X-\mu|\geq\varepsilon)=var\frac{(X)}{\varepsilon^2}=\frac{\sigma^2}{n\cdot\varepsilon^2}$ and for $n\to\infty$: $\lim_{n\to\infty}P(|X-\mu|\geq\varepsilon)=0$

For example, let $X_{1,}\ldots,X_n$ be the results of rolling a 6-sided die. Then each roll produces a result between that is one of the numbers 1, 2, 3, 4, 5 or 6 with equal probability. Then the expected value of the die roll is $\frac{1+2+3+4+5+6}{6}=3.5$. So, according to the weak law of large numbers, given a large enough number of repetitions, the average value of die rolls should converge towards 3.5. The results of such an experiment are shown in figures 2.6a and 2.6b.

The law of large numbers can be utilized in order to assess the stability of results in genomic datasets. First, bootstrap resampling can be used to generate a large number of datasets to be used for the evaluation of feature selection and classification methods. Then, under the assumption that the observed results are independent identically distributed random variables, the law of large numbers can guarantee the stability of the average estimates given that the sample size is sufficiently large. Thus, 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 required, an explicit criterion determining the stability of results can be used. The use of bootstrap resampling, in conjunction with the law of large numbers and a stability criterion constitute the concept behind the stable evaluation methodology proposed in this thesis.

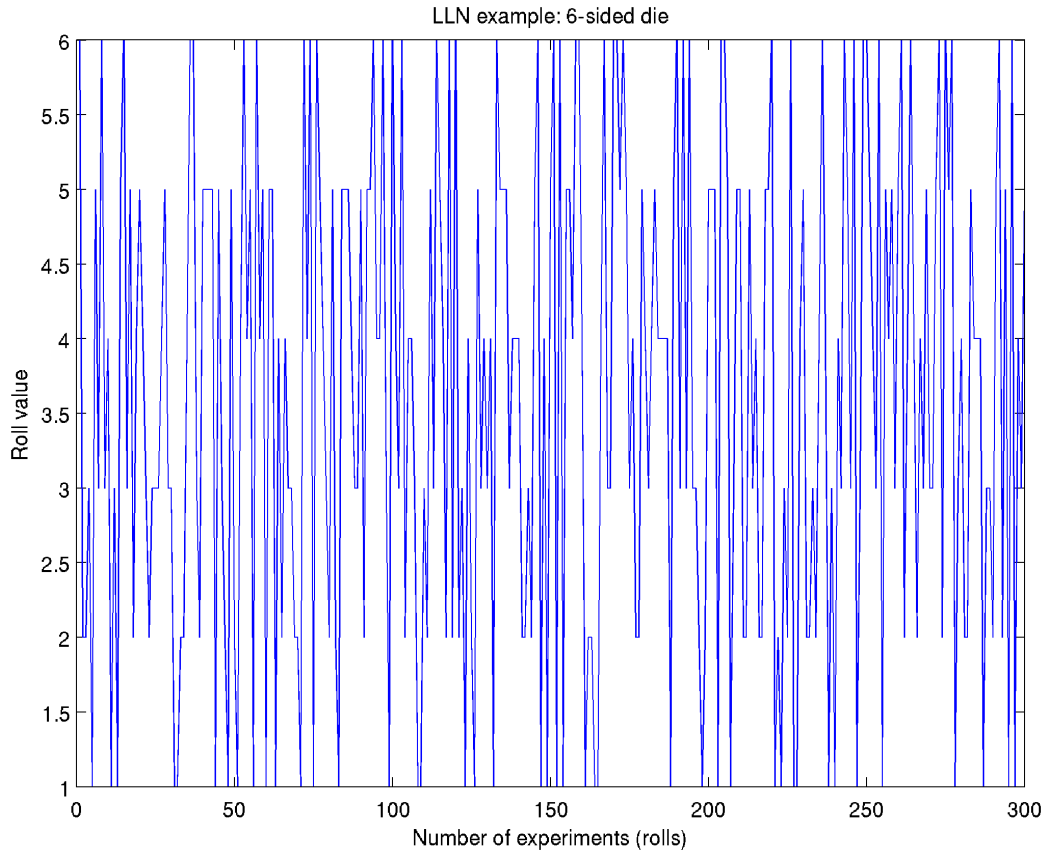

Figure 2.12 Instantaneous values of the 300 rolls of a 6-sided die.

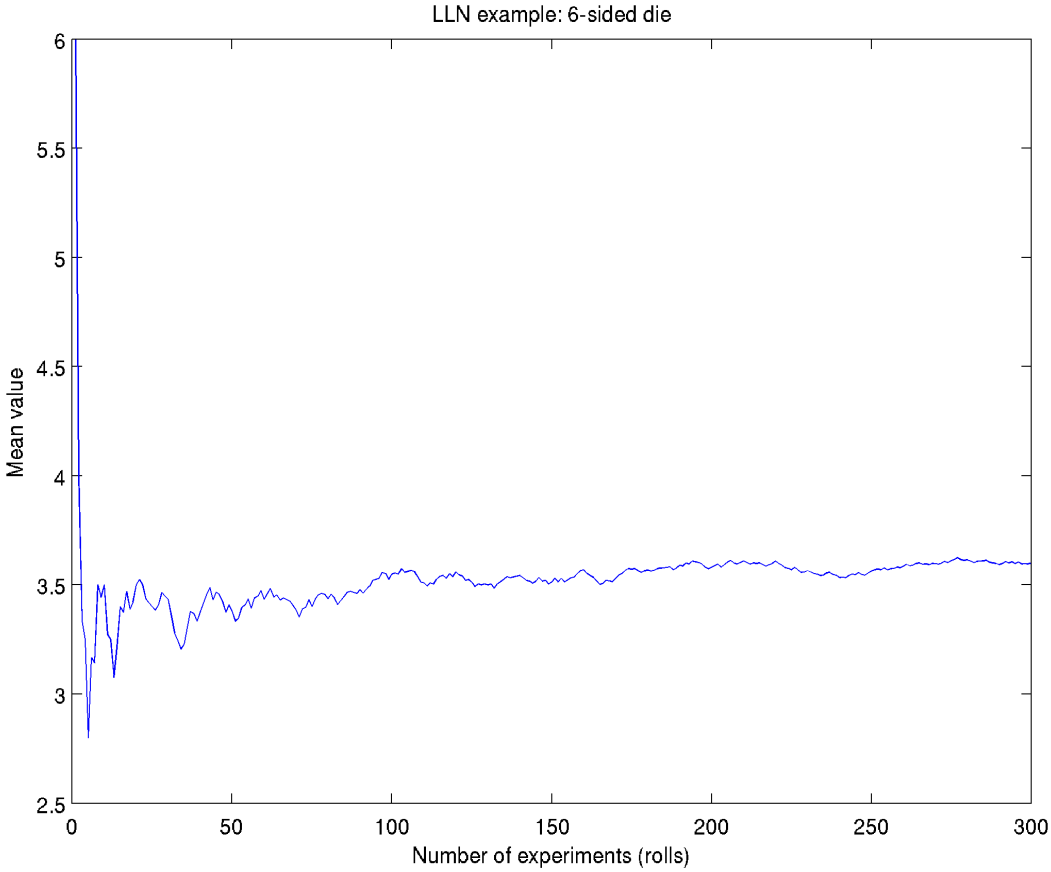

Figure 2.13 Demonstration of the law of large numbers: the mean value over all rolls converges towards 3.5, the expected value of the experiment, as more repetitions of the experiment take place.

2.9 Correspondence At the Top (CAT) plots

The correspondence at the top plot [60] is a convenient way to visualize the “degree of agreement” among lists of genes (features) extracted by different feature selection methods. It is expected that different FSS methods will lead to genomic signatures that are different to a certain degree. However, there should be significant overlap among these gene lists, especially among the top ranked genes of each list. In this manner, the CAT plot visualizes the “degree of agreement” among the different gene lists by plotting the percentage of common genes (y-axis), as the size of the list increases (x-axis). As a result, the CAT plot displays the degree of agreement for the top ranked genes at the beginning of the x-axis, and as the size of the list increases, the degree of agreement changes as genes that of lower rank (considered to be less important by FSS methods) are included. In the original paper where the CAT plot was introduced [60], it was used to assess the agreement between differential expression lists of genes among different labs and platforms.

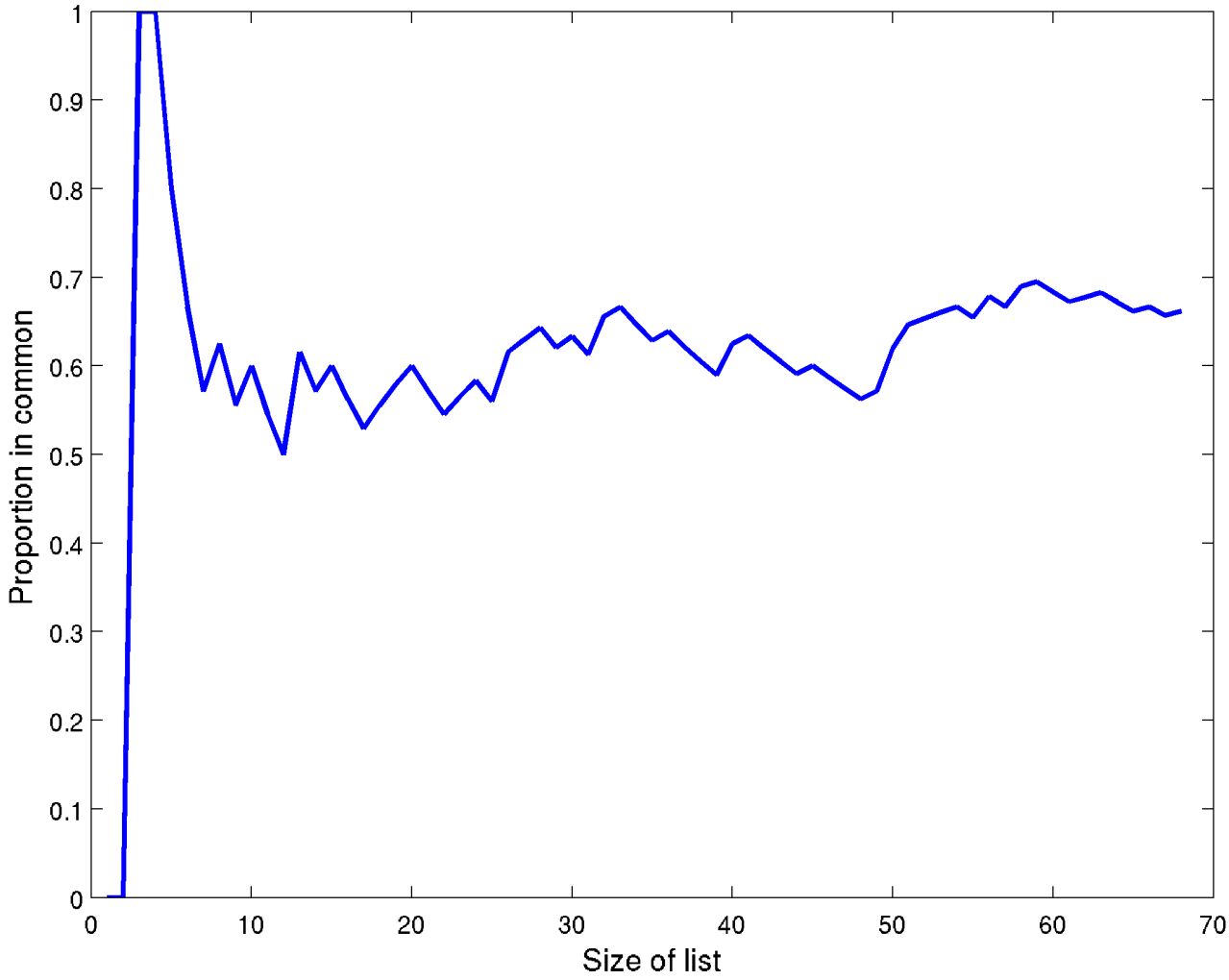

Figure 2.14 CAT plot example.

3 - Proposed Methodology

3.1 Overview

Given a DNA microarray dataset, the main steps of the proposed methodology are presented. First, the dimensionality of the original dataset is reduced as a preprocessing step, using the filter method SAM. Second, by utilizing bootstrap resampling a sufficiently (according to a stability criterion) large number of bootstrap datasets are generated and a classifier that supports feature selection is run on each of the bootstrap datasets. Then, an ordered list of genes is generated, according to gene selection frequency across all bootstrap datasets. This ordered list of genes serves as a basis for considering candidate genomic signatures according to the percentile of selection frequency, e.g. selecting genes above the 95th percentile, meaning the top 5% of genes according to selection frequency. According to a predetermined set of candidate percentiles, a set of candidate signatures is extracted. Then, the predictive performance of each candidate signature is estimated in terms of classification accuracy, sensitivity and specificity and the signature that yields maximal predictive performance is extracted as the stable genomic signature, also taking into account the number of genes in the signature. That is, appropriate confidence intervals are estimated for the predictive performance of each candidate signature and as such, if there is no significant difference in the predictive performance among candidate signatures, the one with the smaller number of genes is usually preferred. Moreover, the predictive performance of all candidate genes, along with confidence intervals is plotted in an elaborate manner for further inspection. As a subsequent step, the statistical significance of the extracted signature is assessed by two separate tests, concerning the significance of the achieved classification accuracy, as well as the signature's association to the response variable (phenotype). Finally, the stability of the extracted signature is estimated by comparing the signatures extracted by several independent executions of the methodology. That is, the degree of “agreement” among the independent signatures is visualized using CAT plots.

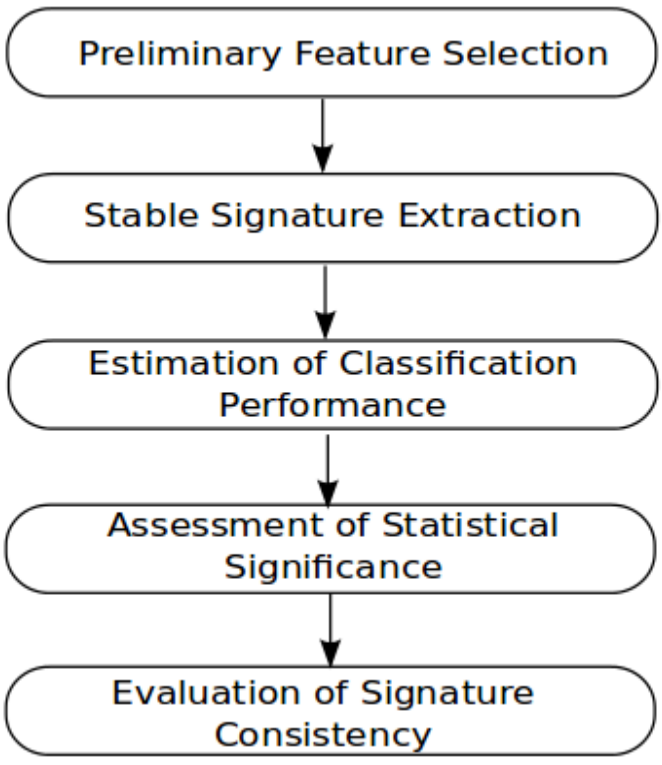

Figure 3.1 Steps of the proposed methodology.

3.2 Datasets

Three datasets were used in this study. The first dataset GSE_Merged was generated after merging several publicly available datasets (GSE22820, GSE19783, GSE31364, GSE9574, GSE18672), as described in [3] [6]. It consists of 529 samples (425 breast cancer, 104 controls) and 11928 features (genes). The second dataset, GSE42568, consists of 121 samples (104 breast cancer, 17 controls) and 54676 features (probesets). The third dataset, GSE35974, consists of 144 samples (94 bipolar, 50 controls) and 33297 features (probesets). In the last two datasets where the features are probesets, with more than one probeset corresponding to the same gene.

3.3 Preliminary Feature Selection

The dimensionality of the original datasets is too large and direct processing by multivariate feature selection and classification methods would be impractical and even practically impossible for some of the most computationally expensive methods [49]. As such, the datasets are first filtered using the univariate method SAM (see section 2.4.1) in order to discard irrelevant genes. While any univariate filtering method can be used, SAM was selected for its robustness. For two of the three datasets, further filtering using fold change was applied in order to filter out additional genes, since too many genes were kept by SAM. The goal of this step is to keep under 2000 features in the dataset, which is considered heuristically as a threshold that allows processing by multivariate feature selection and classification methods in “reasonable” time. The details of this step concerning algorithmic parameters for SAM and thresholds for fold change filtering are presented in section 4.1. This two step feature selection process aims to combine the advantages of both univariate and multivariate FSS methods, leading to selected features that differentiate their behavior among the classes of interest and achieve maximal predictive performance.

3.4 Stable Signature Extraction through Stable Bootstrap Validation

The main concept of the Stable Bootstrap Validation step is to extract a list of gene selection frequencies, which will in turn be used in order to identify candidate genomic signatures in a subsequent step. A number of bootstrap datasets are generated and a feature selection and classification method is run on each dataset, generating a distinct gene list for each bootstrap dataset. Then, each gene is assigned a selection frequency, based on the proportion of bootstrap datasets that it was selected by the feature selection and classification method. Moreover, a key idea is when enough bootstrap datasets have been generated, in order to extract a reliable list of gene frequencies. In that manner, a stability criterion is introduced that incorporates the average number of genes selected across all bootstrap datasets, which is bound to converge (stabilize) to a certain (initially unknown) value, due to the Law of Large Numbers. The stability of the average number of genes is assessed in three overlapping batches of bootstrap datasets which are called “bootstrap windows” and are expanded by generating additional bootstrap datasets until convergence has been achieved for the average number of genes. This average number of selected genes “**G**”, is also used in a subsequent step, in order to identify the top “**G**” genes in terms of selection frequency as an additional candidate signature. The stable signature extraction step is independent of the feature selection and classification method used. In this thesis, the RVM and RFE-SVM methods are utilized and their extracted gene signatures were compared.

Given a pair of FSS and classification methods, Stable Bootstrap Validation (SBV) aims at using a large number of datasets generated from bootstrap resampling of the original dataset, in order to extract a stable estimate the size of the genomic signature. If the FSS and classification method is evaluated on a sufficiently large number of bootstrap datasets, then according to LLN the average estimate of the genomic signature size (number of selected genes) will be stable. To ensure that no more bootstrap datasets than

necessary are generated, SBV utilizes an explicit criterion that determines whether stability has been reached for the average the signature size. The criterion assesses the stability of results over consecutive batches of bootstrap datasets and determines whether a desired level of stability has been reached, or generating another batch of datasets is required. Unlike similar methodologies which lack a stability criterion and are executed for an arbitrary number of iterations, SBV is only executed until the necessary level of stability is reached. As such, SBV is a more computationally efficient methodology.

The SBV procedure proceeds as follows. First, the number of datasets in each batch is associated with a variable called the “bootstrap window” size B, which is defined as a fixed number of bootstrap datasets. Then, a number of 3B bootstrap datasets are generated from the original dataset by random sampling with replacement. The size of the bootstrap datasets (number of samples) is arbitrary, however in most cases (including this thesis) it is selected to be the same as the size of the original dataset. The class ratio is also arbitrary and typical values include the same class ratio as in the original dataset, or equal class ratio for all classes. In this thesis the class ratio of the bootstrap datasets was set to be the same as in the original dataset. The FSS & classification method is then executed 3B times, resulting in values $G_{1,}\dots,G_{3B}$ for the number of features (genes) selected, also called the genomic signature size. Assuming that G_i is a set of independent identically distributed (i.i.d) random variables, then according to the weak law of large numbers the average value “**G**” should converge towards the expected value of the genomic signature size. The exact value of the expected signature size is unknown in practice, however it is not necessary to assess stability so it does not pose a problem. Next, the stability of the observed results is assessed in terms of the average value of genes selected across overlapping batches of gene sets, called bootstrap windows (see step 2 of the pseudo-code below) . Let $G_i, i=1,2,3$ be the mean genomic signature size of the fist, second and third bootstrap windows, while $\Delta G = max(|G_1-G_2|,|G_1-G_3|)$ the maximum difference of mean signature size between windows 1, 2 and 1, 3. However, there is a subtle detail concerning the assessment of signature size stability, since different FSS methods can lead to genomic signatures whose size differs in orders of magnitude. For this reason the corresponding threshold for the signature size is normalized by the largest signature size and is defined as $gen_{thresh}=\frac{|G_{wi}-G_{wj}|}{max(G_{wi},G_{wj})}$, where i, j the windows being compared. As a result, the normalized signature size takes values in the range [0,1] and the same threshold (e.g. 5%) can be used for all FSS and classification methods.

If “**G**” is found to be stable, the SBV procedure ends. 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 (see step 7 of the pseudo-code below). The above steps are repeated until stability for the signature size is reached. During each iteration, the following formula applies for the mean signature size:

$$G_{w_j}^{(n)}=\frac{1}{(n+j-1)B}\sum_{b=1}^{(n+j-1)} gen_b$$

Where n is the iteration number, j is the window being checked (1, 2 or 3), b runs all the bootstrap datasets and gen the number of genes selected (signature size).

After the SBV procedure has been completed, the list of gene selection frequencies is calculated. Each gene's selection frequency corresponds to the fraction of bootstrap datasets where it was selected by the FSS method. Finally, the “**G**” genes with the highest selection frequency across all bootstrap datasets are selected as candidate genomic signature to be evaluated in a subsequent step. A flowchart of SBV if shown in figure 3.2.

To summarize, During this step the selection frequency of each genes is calculated, as well as the number of genes “**G**”, selected on average. First, we define as a bootstrap window of size B a set (batch) of B bootstrap datasets. At the beginning of the algorithm, 3 bootstrap windows of size B are generated using bootstrap resampling. Then, a classifier that supports feature selection is evaluated on each bootstrap dataset, using stratified holdout validation to randomly split each bootstrap dataset into 90% training and

10% test sets. Assuming that the number of selected genes is and independent identically distributed (i.i.d.) random variable, we know from the law of large numbers that the mean of the number of selected genes will converge to a certain value. We exploit this convergence in order to create a stability criterion according to which we know that enough bootstrap datasets have been generated and no more are necessary to achieve stability of results. So, the mean value of the normalized genomic signature size (number of genes) in each of the three bootstrap windows is calculated. If the difference among the mean values among windows 1-2 and windows 1-3 is less than a predetermined threshold (5% in our case), then the results are considered stable. Otherwise, an additional set of B bootstrap datasets are generated, the three windows are extended to cover all datasets and the stability of results is assessed again. This procedure is repeated until the mean number of genes converges. The mean number of genes is normalized so that the same threshold can be used independent of the FSS and classification method, since different methods result in signature sizes that may differ in orders of magnitude. Typically, only a few (if any) additional iterations are required, depending on the inherent stability of the feature selection and classification method. The pseudo-code for the stable signature extraction step is, including the values used in this thesis is:

Pseudo-code for the stable signature extraction algorithm:

Input: A dataset D including class labels, a FSS and classification method, the bootstrap window size B = 50, the stability threshold thresh = 5% and i = 1 and index of additional generated batches of datasets
Step 1) Generate 3*B bootstrap datasets from D using bootstrap resampling
Step 2) Define the overlapping bootstrap windows
window1: bootstrap datasets 1 to B (1 to 50)
window2: bootstrap datasets 1 to 2*B (1 to 100)
window3: bootstrap datasets 1 to 3*B (1 to 150)
Step 3) Run the FSS and classification method on all bootstrap datasets
Step 4) Calculate the mean signature size for each window
mean1 = mean of genes selected in window 1
mean2 = mean of genes selected in window 2
mean3 = mean of genes selected in window 3
Step 5) Calculate the **normalized** differences between windows 1-2 and 1-3:
diff1_2 = abs(mean1-mean2)/max(mean1,mean2)
diff1_3 = abs(mean1-mean3)/max(mean1,mean3)
Step 6) Check for convergence:
if(max(diff1_2,diff1_3) < thresh) then **terminate** the algorithm
else continue to step 7
Step 7) Generate B additional bootstrap datasets and expand the windows by B positions “to the right”
window1: bootstrap datasets 1 to (B + i*B) (for i=1 this corresponds to bootstrap datasets 1 to 100)
window2: bootstrap datasets 1 to (2*B + i*B) (for i=1 this corresponds to bootstrap datasets 1 to 150)
window3: bootstrap datasets 1 to (3*B + i*B) (for i=1 this corresponds to bootstrap datasets 1 to 200)
i = i+1 (increase the index of additional generated batches of datasets)
Step 8) Run the FSS and classification method on the B additional datasets
Step 9) Go to step 6 (check for convergence again)
Output: A list of gene selection frequencies and “**G**”, the mean number of genes selected.

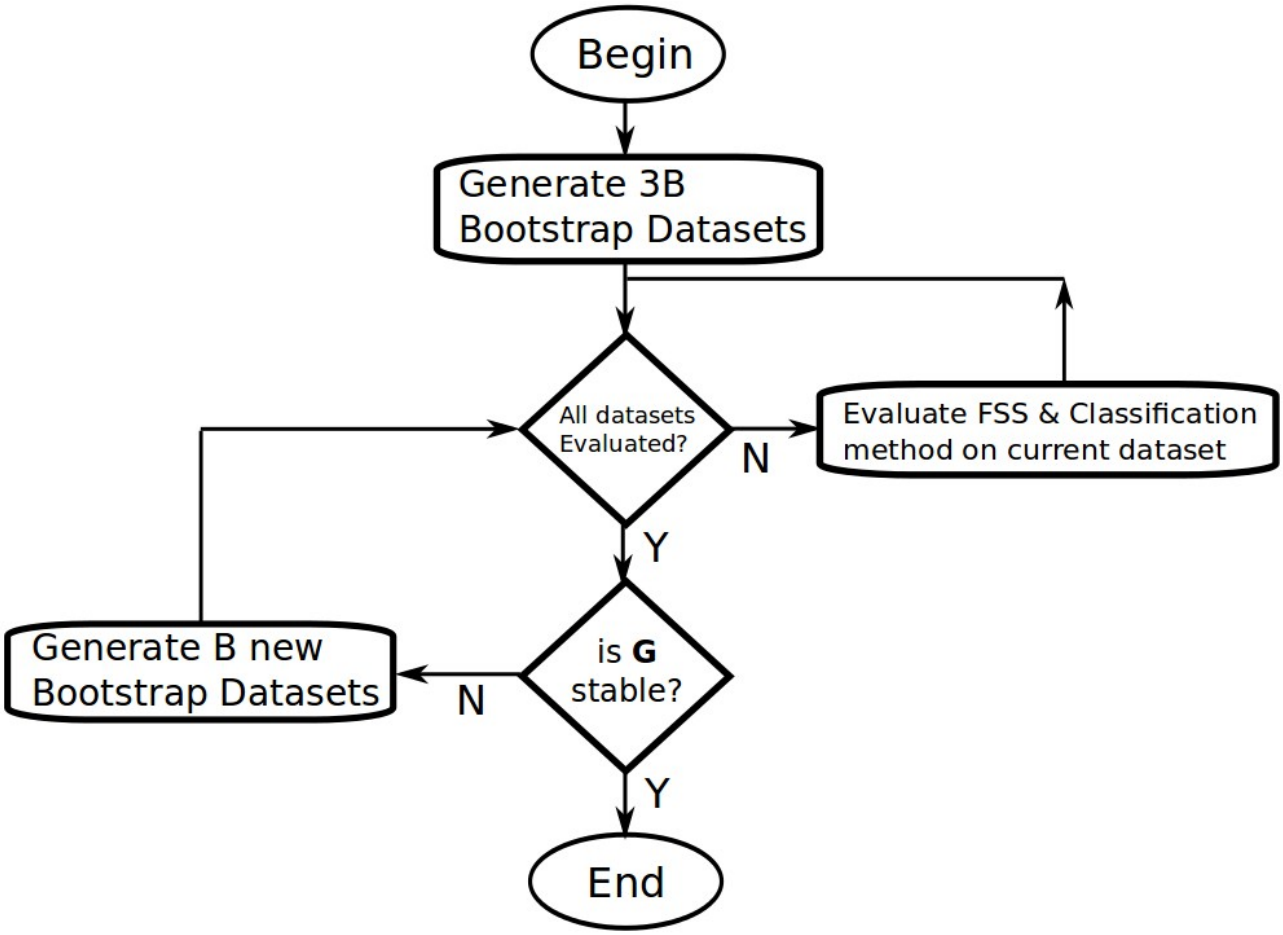

Figure 3.2 Flowchart of the SBV process for stable signature extraction.

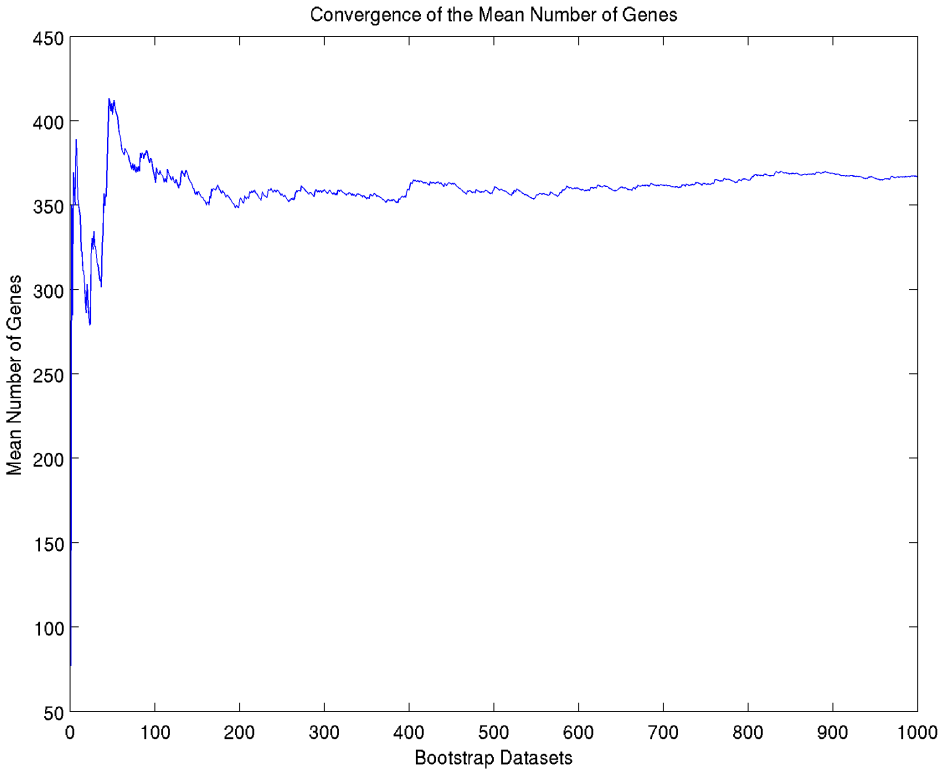

Figure 3.3 Illustration of the convergence of the mean number of genes selected across all bootstrap datasets. For the generation of this figure RFE-SVM was used on the GSE35974 dataset. Notice that the average number of genes has converged and is stable at 200 datasets. A result that agrees with the proposed stability criterion

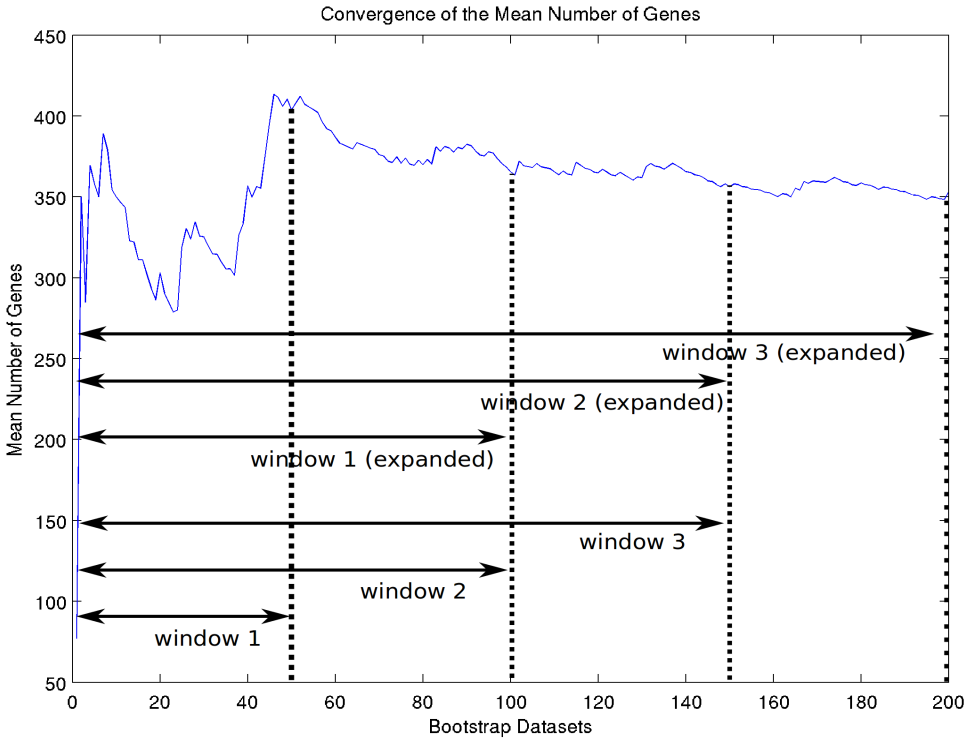

Figure 3.4 We zoom in and focus on the first 200 bootstrap datasets of the previous plot. The original bootstrap windows (pseudo-code-step 2) are seen ranging from datasets 1 to 150 (or B to 3B for B=50). If convergence of average signature size has not achieved the required threshold, the windows are expanded (pseudo-code step 7) and the process is repeated until no further expansions and assessment of additional bootstrap datasets are necessary.

3.5 Estimation of Classification Performance

At this point, the the selection frequency of each gene is known previous step of the algorithm, which results in the candidate gene signatures corresponding to genes above the 30^{th}, 40^{th}, 50^{th}, 60^{th}, 70^{th}, 80^{th}, 90^{th}, 95^{th} and 99^{th} percentiles of selection frequency. That is, the top 70%, 30%, 50%, 40%, 30%, 20%, 10%, 5% and 1% of genes in terms of selection frequency, respectively. During this step, the predictive performance of all these candidate signatures, as well as the top "**G**" genes is calculated and the best performing signature in terms of accuracy, sensitivity and specificity is selected as the stable extracted genomic signature. Any feature selection and classification method can be used in this step, yet it is recommended to use the same method utilized in the previous step of stable signature extraction. As such, the RVM and RFE-SVM methods are used for estimating the classification performance of the extracted signatures. Moreover, appropriate confidence intervals are calculated for accuracy sensitivity and specificity and as such, if a number of signatures achieve similar predictive performance (their confidence intervals overlap), usually the smallest signature is selected since it corresponds to a simpler and sparser model. Moreover, the results of all candidate signatures are plotted in a compact and elaborate manner for further inspection. Finally, we are free to choose whether to select the best performing signature in terms of accuracy, sensitivity or specificity, depending on which metric is considered more important in the specific application. However, usually accuracy is selected as the main metric and sensitivity and specificity are considered complementary to it.

The classification performance of all candidate signatures is estimated in terms of accuracy, sensitivity and specificity using multiple K-fold Cross Validation. That is, standard K-Fold CV is run several times and the results are aggregated. To be precise 3x10 Fold CV was run, meaning that standard 10-Fold cross validation was run 3 times, which was considered to be a good tradeoff between stability of classification performance estimates and execution time.

Calculating the corresponding confidence intervals [46] [47] for the accuracy, sensitivity and specificity values of each candidate gene signature was performed using a hybrid method. 95% confidence intervals were calculated, however the methodology can obviously be applied for the calculation of confidence intervals of any value. If the success failure condition is met, meaning that there are at least 10 "successes" and 10 "failures" when calculating the fraction corresponding each of the three metrics, then the confidence intervals are calculated using the standard methodology of z confidence intervals for proportions (large sample framework). However, the success failure condition may not be met when the metric (accuracy, sensitivity or specificity) is very close to either 0 or 100%. If the success failure condition is not met, then bootstrap t confidence intervals are calculated instead [46]. Since bootstrap t confidence intervals are symmetric, in some extreme cases they can slightly exceed either 0 or 100%. For example, if the observed metric has an average value of 98% $\pm$ 2.1% for 95% confidence intervals, the lower bound is 95.9% and the upper bound is 100.1%. However, since 100.1% accuracy, sensitivity, or specificity is impossible, if such a numeric artifact is generated, the resulting confidence interval is trimmed. In the previous example, this would lead to an asymmetric confidence interval from 95.9% to 100%. Another option would be to use bootstrap percentile confidence intervals [46], where the 95% confidence intervals are asymmetric by nature and correspond to the interval between the 2.5% and 97.5% percentiles of the bootstrap distribution of the test statistic.

Finally, the classification performance of the classification method for each dataset was reassessed, by using random oversampling and random undersampling (see section 2.7) in order to counter the effects of imbalanced class ratios in the original dataset. The aim of these sampling methods is to increase the predictive performance of the classification method on the underrepresented class (with less samples), possibly at the cost of mis-classifying a few more samples of the overrepresented class, correspoding to changes in accuracy, sensitivity and specificity.

3.6 Assessment of Statistical Significance

Having extracted a genomic signature, we would like to know whether it actually reflects the underlying biological process that is being studied and to what extent it reflects random noise. It is very important to test the statistical significance of the findings, especially since several studies question the validity of the genomic signatures extracted from DNA microarray data using certain methodologies [61] [62]. The main idea behind the proposed significance test, is to compare the extracted signature to random signatures of the same size.

Test 1: statistical significance of classification accuracy

The first test compares the extracted signature to random signatures of the same size in terms of predictive performance. The test consists of two steps. During the first step (Test1-A), a number of 10^4 bootstrap datasets are generated and the classification accuracy of the extracted signature, as well as the classification accuracy of a random signature of the same size are calculated for each of the bootstrap datasets. Each of the bootstrap datasets is split randomly into 90% training and 10% test sets using stratified holdout validation. As a convention, the same classifier is used for this step and estimating the classification performance of the extracted signature at the previous step of the proposed methodology (Estimation of Classification Performance). At the end of Test1-A the probability of a random signature performing better in terms of classification accuracy is calculated. Moreover, the average difference of classification accuracy between the extracted signature and random signatures across all bootstrap datasets is calculated and passed into the next step of the test. The second step (Test2-B) is a hypothesis test, where it is evaluated whether the increased predictive performance of the extracted signature compared to random signatures is statistically significant. That is, an appropriate null hypothesis is formed, that “the difference in accuracy during Test1-A is observed due to chance and the actual difference is zero”. This null hypothesis is then assessed by an appropriate permutation test [46] [47], resulting in a corresponding p-value. Briefly, the procedure of the permutation test is as follows: First, we assume that each observed classification accuracy has a corresponding label “stable” or “random”, depending on which signature it was generated from. Next, under the null hypothesis the labels are permuted and the difference between the “stable” and “random” sets of accuracy values is recalculated. This procedure is repeated several times (10^6 in our case) and the corresponding p-value is calculated as the fraction of permutations where the observed difference in classification accuracy between the two sets is larger than, or equal two the difference in classification accuracy observed during Test1-A.

Test 2: statistical significance association to the response variable (phenotype/clinical outcome)

The second test follows a similar two-step procedure, utilizing the self-contained gene set analysis method Globaltest [55] (see section 2.5) which is available through an R package of the same name [57] and is known to perform well in practice [54]. It tests the null hypothesis that the covariates (genes) are not associated with the response (phenotype), against the alternative that they are and calculates a single p-value for the whole gene set. During the first step (Test2-A), the Globaltest method is used to evaluate the extracted genomic signature and produce the corresponding p-value. However, that test is not enough on its own, especially since the response variable was used in the gene selection process. Next, as a subsequent step (Test2-B) a number of 10^4 bootstrap datasets are generated and the Globaltest method is run on each bootstrap dataset twice. Once using only the extracted signature and once using a random set of genes of the same size as the extracted signature. Finally, the empirical probability a random signature performing as well or better than the extracted signature is calculated as the fraction of random gene sets having an equal or smaller globaltest p-value than the extracted signature on the same bootstrap dataset. It should be pointed out that this empirical probability is not a p-value.

3.7 Evaluation of Signature Consistency

It is crucial that the resulting genomic signatures are stable and consistent. That is, very similar signatures must be extracted for independent executions of the methodology. To assess the consistency of the extracted signatures, the above methodology is run independently several times and the “degree of agreement” among different signatures is visualized through a Correspondence At the Top (CAT) plot [60] which was introduced in section 2.9. In brief, a CAT plot visualizes the proportion of common genes among the different lists (y-axis), against the size of the lists (x-axis). If there is strong overlap among the signatures extracted by the proposed methodology, then this will be reflected by strong “agreement” in the CAT plot. That is, there should be high degree of genes in common. If there is strong agreement only at the beginning of the CAT plot, it means that the different signatures “agree” for the importance of their top ranked genes only. If there is strong “agreement” (many genes in common) for all signature sizes in the CAT plot, then the different signatures strongly “agree” for all genes selected. If there is a weak degree of “agreement” among signatures, this is reflected as low degrees of overlap (common genes) in the CAT plot. As a result, depending on the “agreement scenario”, we can assess the stability and robustness of the extracted signatures and whether to further inspect the top ranked genes of the signatures, all genes, or no genes at all.

4 - Results

In this chapter, the results of applying the SBV methodology to the aforementioned datasets are displayed. Both, RVM and RFE-SVM achieved similar classification performance and on 2 out of 3 datasets showed significant overlap of gene signatures between the two feature selection and classification methods. The reason why no overlap is observed in the third dataset is also discussed. The results of the statistical significance tests, as well as the consistency test are also displayed.

4.1 Data Preparation and Preliminary Filtering

Since the dimensionality of the original datasets was too large for the data to be processed directly by the FSS and Classification methods, preliminary filtering of genes was performed by SAM as described in the methodology section. The FDR cutoff for SAM was set to 5% and the number of permutations to 500. The R package 'samr' [53] was used. Moreover, since the number of genes selected by SAM for 2 of the 3 datasets (GSE_Merged, GSE42568) was still too large, further filtering using the fold change was performed. The cutoff values were >1.5 for overexpressed and <0.67 for underexpressed genes in the first dataset (GSE_Merged) while the respective values for the second dataset (GSE42568) were >1.2 and <0.83. No further filtering using the fold change was necessary for the third dataset (GSE35974), as sufficiently few genes remained after filtering with SAM. The results of the preliminary filtering step can be found in the following table:

Dataset	Samples	Starting genes	Genes after SAM	Genes after Foldchange
GSE_Merged	529	11928	5565	1592
GSE42568	121	54676	21685	1397
GSE35974	144	33297	1246	1246

Table 4.1 Number of genes in each dataset. The number of genes is reduced after filtering with SAM. For two out of the three datasets further filtering using fold change was necessary.

4.2 Classification and Signature Extraction (RVM and RFE-SVM)

Both RVM and RFE-SVM methods had comparable predictive performance in all three datasets. Moreover, the signatures extracted by both RVM and RFE-SVM have a considerable overlap of common genes in two of the three datasets (Figure 4.7) Oversampling and undersampling improved specificity at the cost of sensitivity, as was expected. Moreover, random oversampling seemed to achieve better results compared to random undersampling. This is probably due to the fact that the performance of sophisticated classification methods such as RVM and RFE-SVM suffers when the number of samples available for training is reduced in the case of random undersampling. While the influence of over and undersampling is not statistically significant, since the confidence intervals for all three scenarios (no sampling, over and undersampling) overlap (see the appendix for confidence intervals), there seems to be a trend and more samples are necessary in order to further evaluate the influence of sampling methods in terms of statistical significance. After assessing the performance of all percentiles, as well as the G genes of SBV for each dataset, it is obvious that the assessment of the 30^{th} through 99^{th} percentiles is sufficient and also checking the performance of the "**G**" SBV genes is unnecessary and will probably be omitted in future studies. Finally, 95% confidence intervals were calculated for the three performance metrics, as described in section 3.5.

RVM	Genes	Accuracy	Sensitivity	Specificity
GSE_Merged	70	93.4%	96.6%	80.1%
GSE42568	16	98.9%	100%	92.2%
GSE35974	127	97.9%	98.9%	96%
SVM				
GSE_Merged	141	93.6%	99.1%	71.5%
GSE42568	16	96.4%	99%	80.4%
GSE35974	132	93.1%	96.1%	87.3%

Table 4.2 Overview of classification results.

The list of **all classification results**, which include the predictive performance of all percentiles along with the corresponding confidence intervals can be found in the **appendix**.

4.2.1 GSE_Merged

In the case of the RVM, the number of G=70 genes selected on average was selected as the size of the genomic signature and the top 70 genes were selected, as they outperformed all other genes in terms of predictive performance. However, the G gene list was almost identical to that selected by the 95th percentile in terms of size and predictive performance, which leads to the question whether assessing the G genes is really necessary and whether only the corresponding percentiles need to be assessed. The 141 genes of the 90th percentiles yielded optimum predictive performance in the case of the RFE-SVM method. The C parameter of the SVM was set to 10^-3 using cross validation. Among the two signatures of 70 (RVM) and 141 (RFE-SVM) genes, there is a substantial overlap of 54 common genes. As in can be seen in Figure 4.7, in the case of GSE_Merged the overlap of genes between RVM and RFE-SVM is consistently close 50% as the size of the signature increases in the CAT plot when we move from high ranking to low ranking genes (left to right).

Method	Genes	Accuracy	Sensitivity	Specificity
RVM	70	Normal: 93.4%	Normal: 96.6%	Normal: 80.1%
	75	Over: 91.9%	Over: 93.7%	Over: 84.6%
	49	Under 86.9%	Under: 86.7%	Under: 87.5%
SVM	141	Normal: 93.6%	Normal: 99.1%	Normal: 71.5%
	140	Over: 94.0%	Over: 96.2%	Over: 84.9%
	141	Under: 89.2%	Under: 90.0%	Under: 86.2%

Table 4.3 Predictive performance of the RVM and RFE-SVM classifiers, including over and undersampling results for GSE_Merged

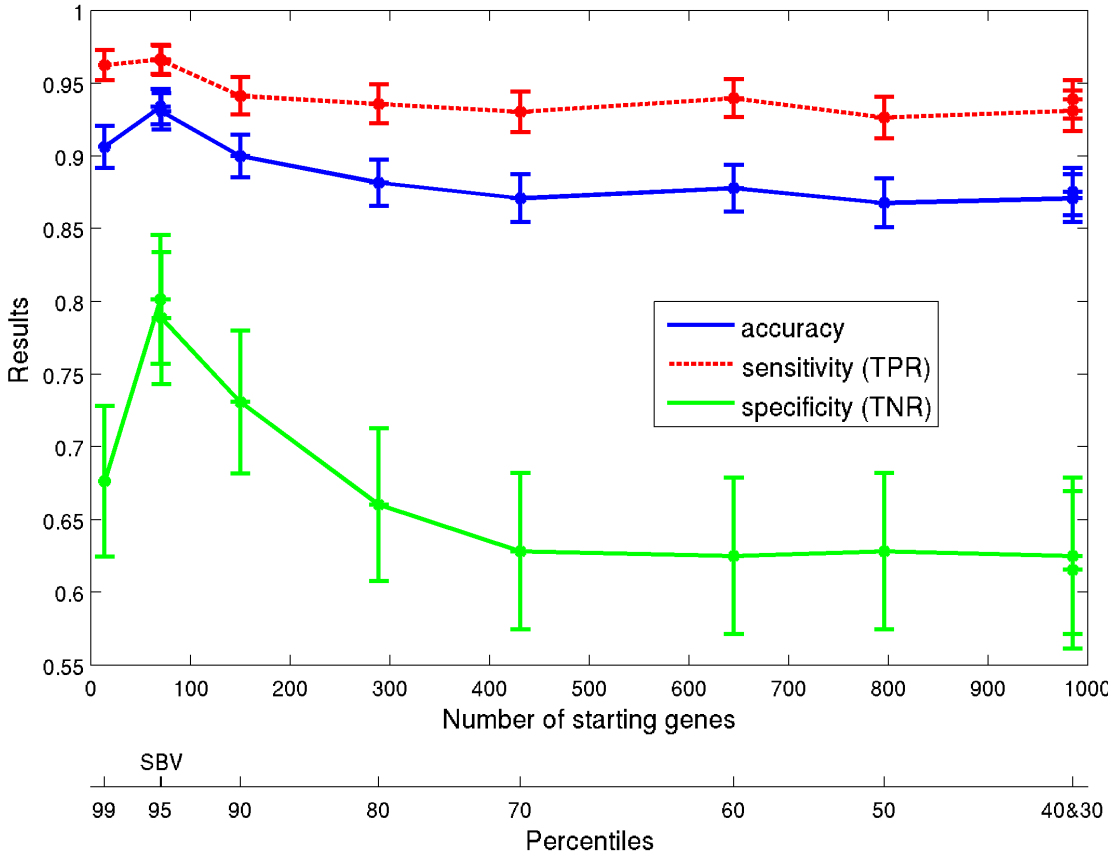

Figure 4.1 RVM classification results for GSE_Merged (no sampling).

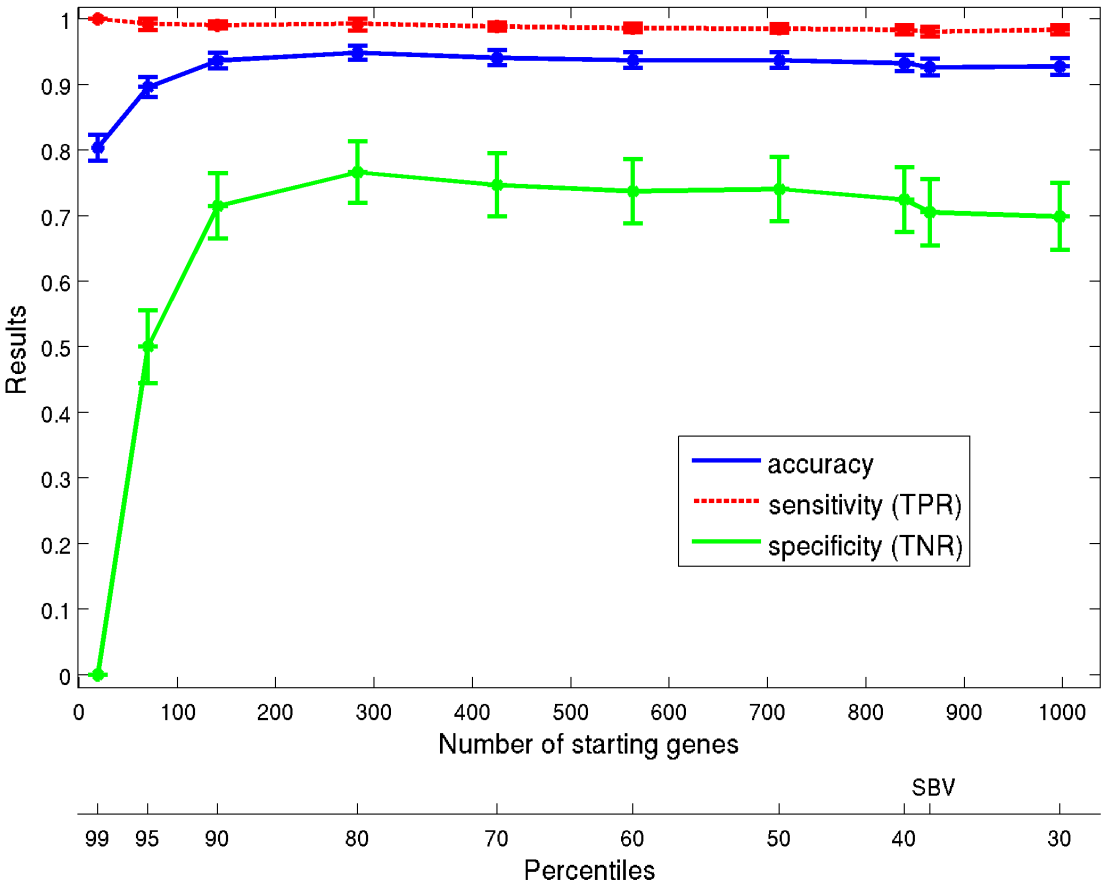

Figure 4.2 RFE-SVM classification results for GSE_Merged (no sampling).

4.2.2 GSE42568

GSE42568 is a very extreme case and an interesting dataset. In the case of GSE42568, almost any set or random genes achieves 100% accuracy, which makes selecting gene sets based on predictive performance a difficult task and produces gene lists of questionable importance. This “equivalence” of predictive performance among gene sets is reflected by the substantial degree of overlap of the confidence intervals of the predictive performance metrics (accuracy, sensitivity, specificity).

In the case of the RVM, the 16 genes of the 99th percentile maximized predictive performance. The 16 genes of the 99th percentiles yielded optimum predictive performance in the case of the RFE-SVM method, as well. The C parameter of the SVM was set to 10^3 using cross validation. It is a coincidence that the signatures extracted by both RVM and RFE-SVM methods consist of 16 genes. The signatures are completely different and show no overlap (no common genes) as it can be seen in Figure 4.7. This disagreement between the two signatures is another hint that selecting genes based on predictive performance is probably not a good idea in the extreme case where almost all gene sets have maximum predictive performance. This will also be confirmed by the statistical tests later.

Method	**Genes**	**Accuracy**		**Sensitivity**		**Specificity**	
RVM	16	Normal:	98.9%	Normal:	100%	Normal:	92.2%
	17	Over:	99.4%	Over:	100%	Over:	96.1%
	16	Under	96.4%	Under:	97.1%	Under:	92.2%
SVM	16	Normal:	96.4%	Normal:	99.0%	Normal:	80.4%
	18	Over:	97.8%	Over:	98.4%	Over:	94.1%
	17	Under:	90.1%	Under:	91.0%	Under:	84.3%

Table 4.4 Predictive performance of the RVM and RFE-SVM classifiers, including over and undersampling results for GSE42568

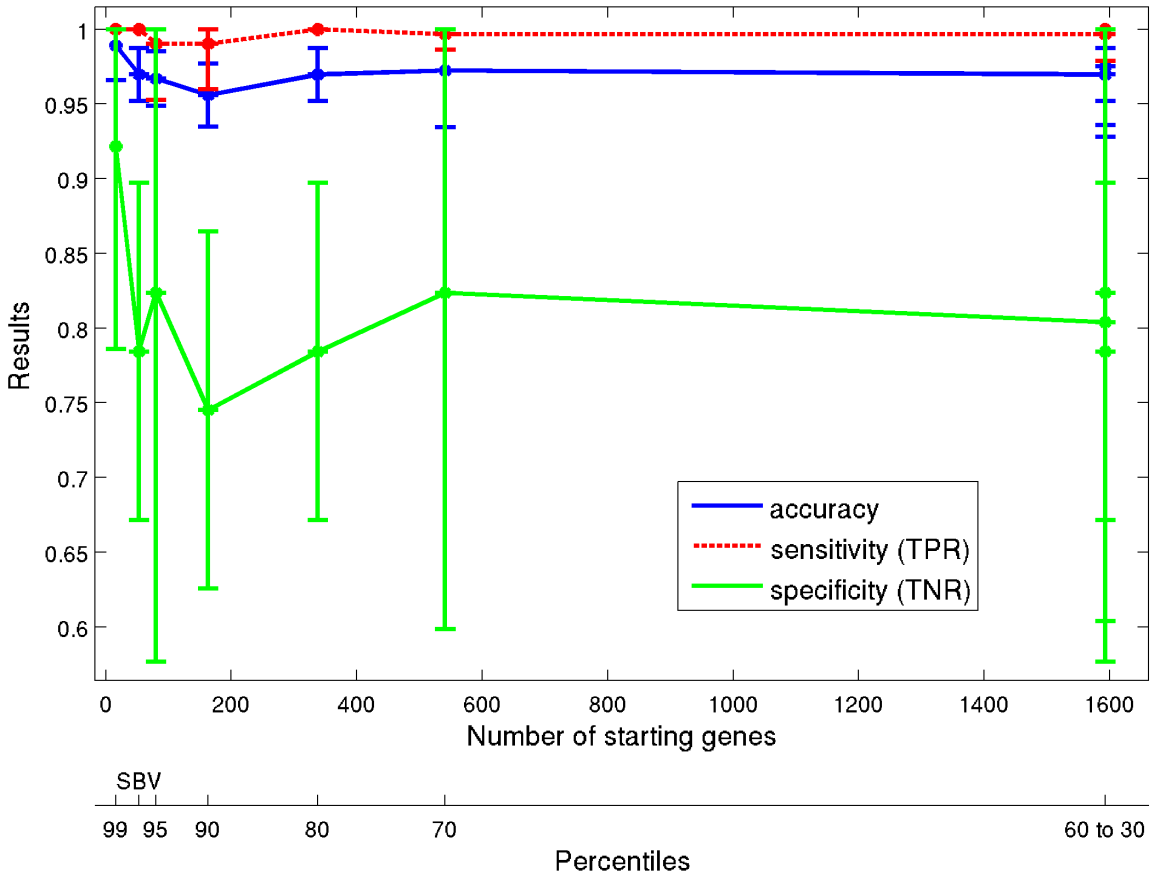

Figure 4.3 RVM classification results for GSE42568 (no sampling).

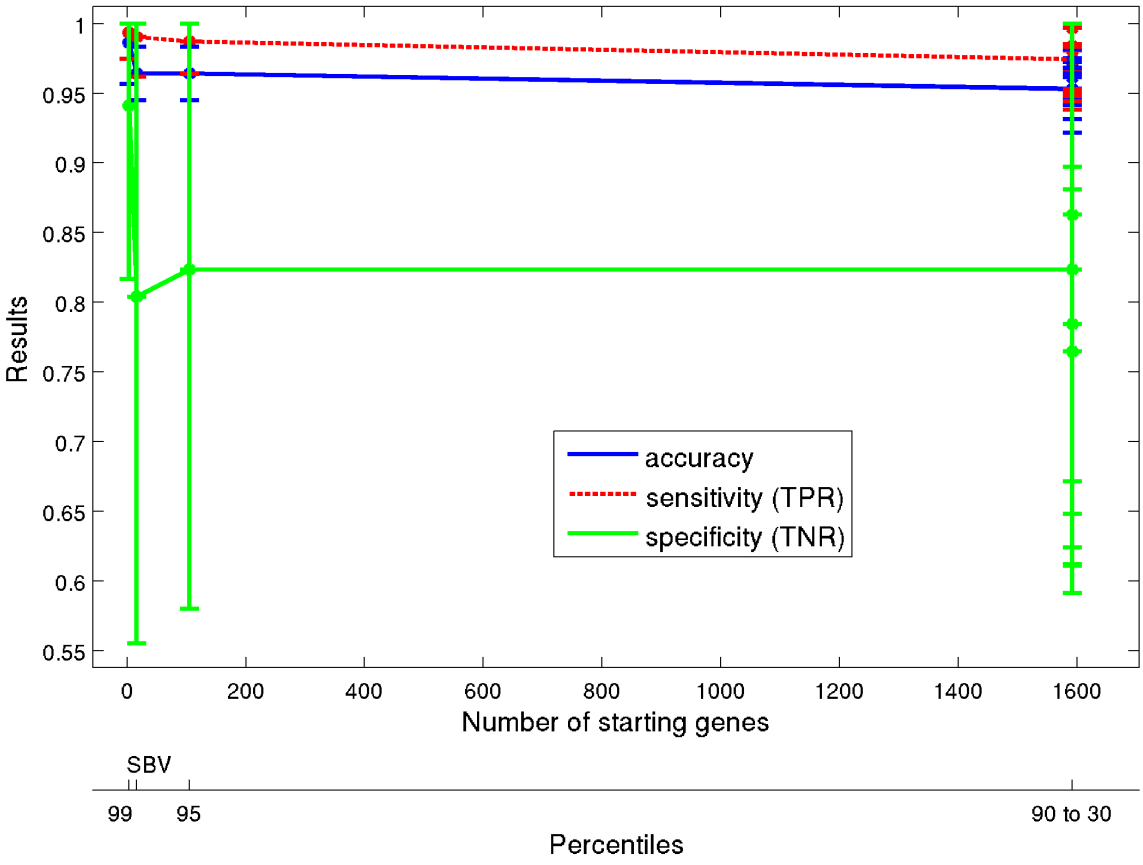

Figure 4.4 RFE-SVM classification results for GSE42568 (no sampling).

4.2.3 GSE35974

In the case of GSE35974, the 127 genes of the 90th percentile are selected by RVM. In a similar fashion, the 132 genes of the corresponding 90th percentile are selected by RFE-SVM. The C parameter of the SVM was set to 10^-2 using cross validation. GSE35974 is similar to an extent to GSE_Merged, since selecting genes based on predictive performance appears to be reasonable method, since different gene sets corresponding to different percentiles yield different predictive performance, unlike the extreme case of the previous dataset (GSE42568). Between the 127 and 132 genes of the two signatures, there is an overlap of 69 common genes (Figure 4.7).

Method	**Genes**	**Accuracy**		**Sensitivity**		**Specificity**	
RVM	127	Normal:	97.9%	Normal:	98.9%	Normal:	96%
	128	Over:	97.7%	Over:	97.5%	Over:	98.0%
	127	Under	97.0%	Under:	96.1%	Under:	98.7%
SVM	132	Normal:	93.1%	Normal:	96.1%	Normal:	87.3%
	125	Over:	91.2%	Over:	93.3%	Over:	87.3%
	128	Under:	91.0%	Under:	89.0%	Under:	94.7%

Table 4.5 Predictive performance of the RVM and RFE-SVM classifiers, including over and undersampling results for GSE35974

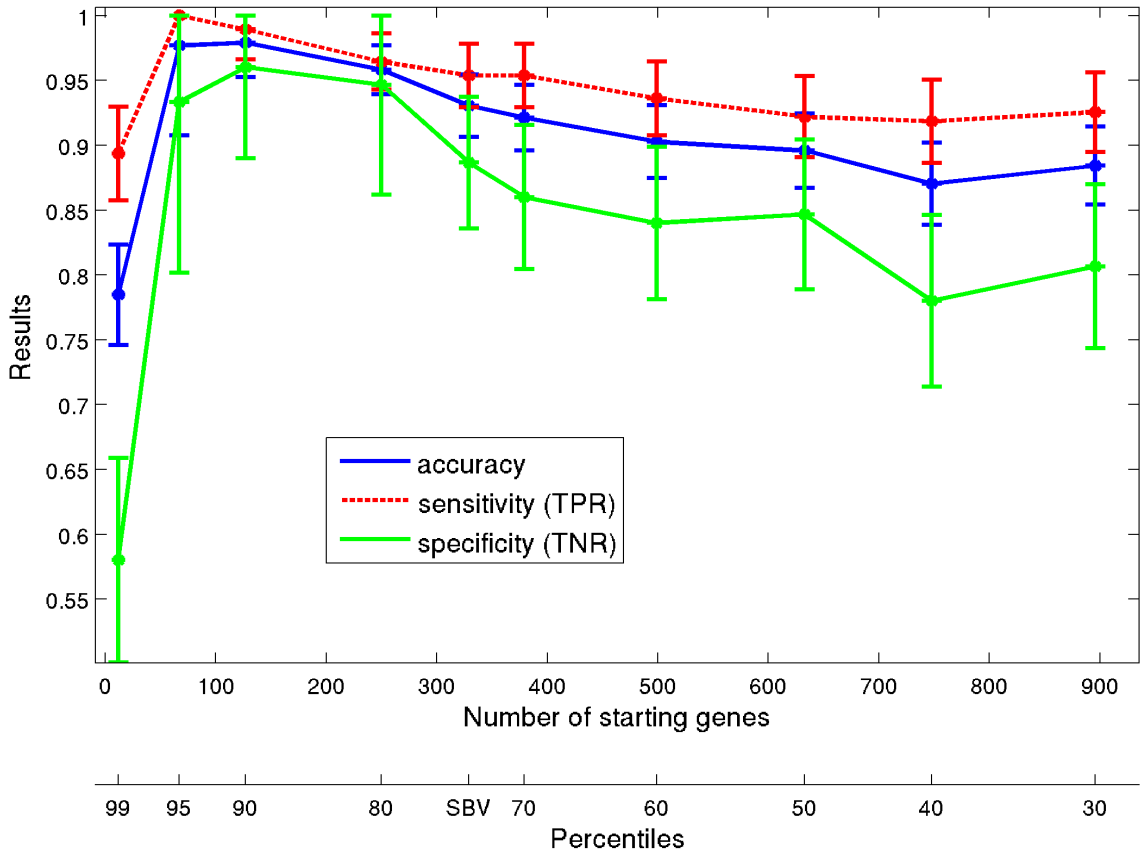

Figure 4.5 RVM classification results for GSE35974 (no sampling).

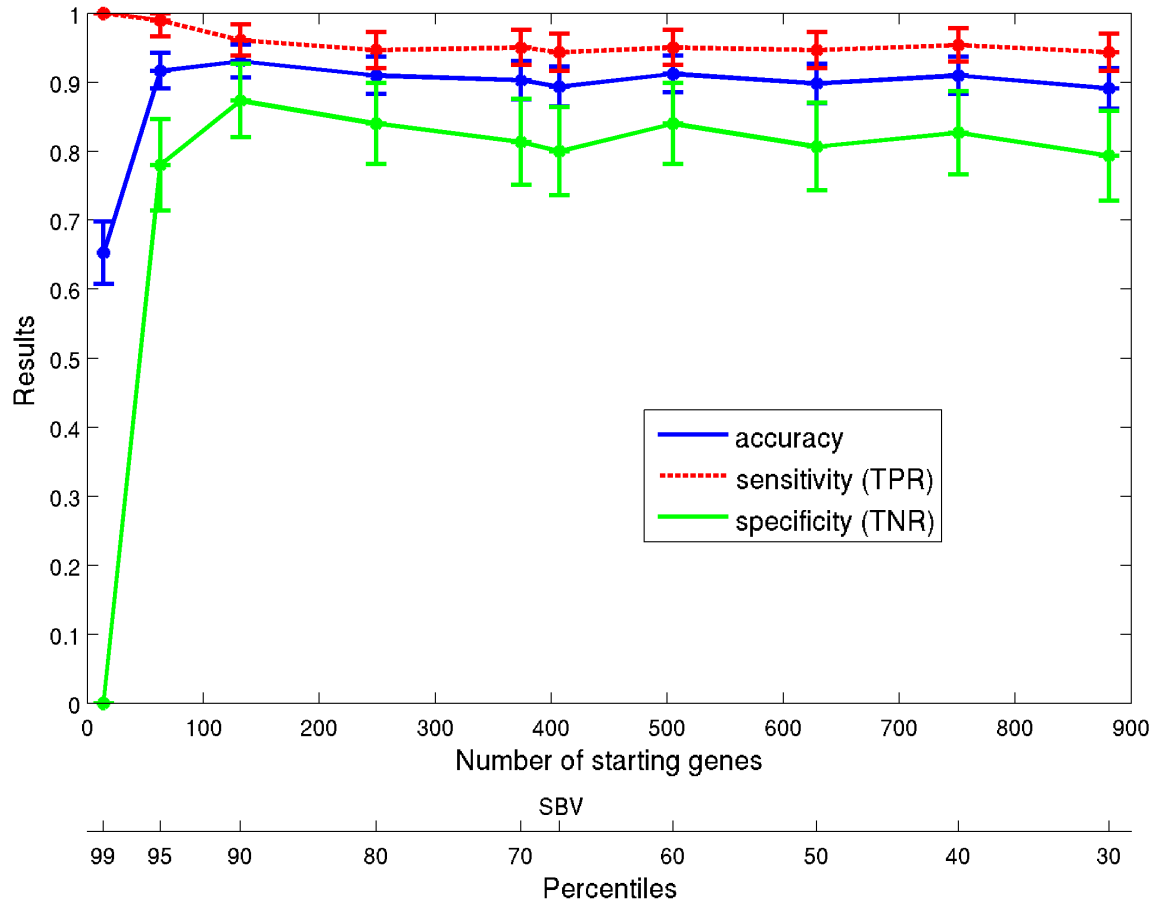

Figure 4.6 RFE-SVM classification results for GSE35974 (no sampling).

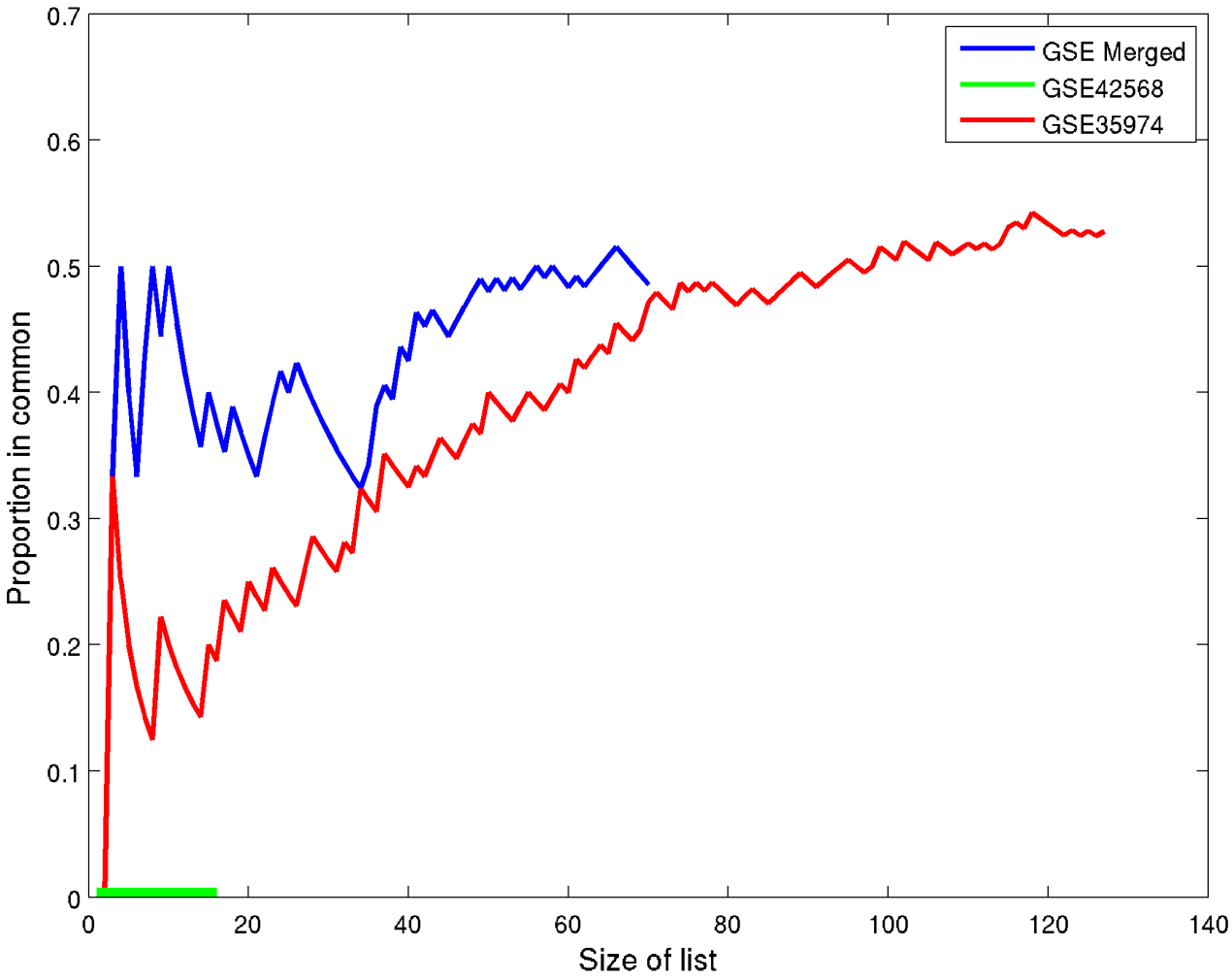

Figure 4.7 CAT plot measuring the degree of "agreement" between the signatures extracted by RVM and RFE-SVM. In two of the three datasets (GSE_Merged, GSE35974) there is considerable overlap/agreement between the two methods. In the case of GSE_Merged the overlap is consistently close 50% as the size of the signature changes when we move from high ranking to low ranking genes in the signatures (left to right). In the case of GSE35974, there is around 30% agreement at the beginning for high ranking genes and the degree of agreement increases towards 50% as more genes are added in the signature. Which means that there is stronger agreement for medium to low ranked genes, than there is for high ranked genes. In the case of GSE42568 there is no overlap between the 16 gene signatures.

4.3 Statistical Significance

According to the first statistical test, the extracted signatures outperform random genes in all datasets in terms of average classification performance on the 10^4 generated bootstrap datasets. However, the predictive performance of the extracted signatures and random genes is comparable, as expected [61] [62]. In the first part of the first test (Test1-A) the empirical probability of a random signature achieving an equal of higher classification accuracy compared to the extracted signature is calculated. Moreover, the average difference in classification accuracy between the extracted and random signatures across all 10^4 bootstrap datasets is also calculated. In the cases of GSE_Merged and GSE35974 the extracted signatures significantly outperform random signatures. In the case of GSE42568, there is very little difference between the extracted and random signatures. However, the result of the second part of the first test (Test1-B) is very

interesting, since all of the observed differences in accuracy are statistically significant and not random, even if they are small, like the case of GSE42568. For the second part of the first test, 10^6 permutations were performed.

According to the Globaltest (Test1-A), all extracted genomic signatures are significantly associated with the response variable, which is to be expected since the genes where selected based on their association with the response variable (phenotype), so the first part of the second test serves mostly as a prerequisite to the second part of the test. It is interesting that once again the signature of GSE42568 scores lower that the signatures of the other two datasets. At the second part of the second test (Test2-B), the association of the extracted signatures to the response is considered statistically significant when compared to random signatures only in the cases of GSE_Merged and GSE35974, according to the permutation test. On the contrary, the signatures extracted for GSE42568 by both RVM and RFE-SVM fail at the permutation test and do not appear to outperform random signatures of the same size to a statistically significant extent. For the second part of the second test, 10^4 permutations were performed.

To summarize, GSE_Merged and GSE35974 where deemed statistically significant according to both tests and as such their predictive performance, as well as their association to the response variable are significantly better when compared to random genomic signatures of the same size. As it was demonstrated earlier, almost any gene set achieved exceptional classification accuracy for GSE42568. As a result, GSE42568 scores low at Test1-A where a random gene set of the same size has a high chance (0.5988 for RVM and 0.6472 for RFE-SVM) of achieving the same or better accuracy. Nonetheless, the 16 gene signature still manages to perform better on average and the difference is statistically significant (Test1-B) even though it is small. Finally, the signatures of GSE42568 do not appear to be associated to the response (phenotype) in a statistically significant manner, when compared to random signatures of the same size (Test2-B). The fact that they succeed at the Globaltest (Test2-A) has no practical value and is to be expected, since they were extracted based on their association to the response by feature selection and classification methods. Due to this, the Globaltest (Test2-A) serves mostly as a prerequisite for (Test2-B) and has no significant value on its own in the context of the proposed methodology.

RVM	**Test1-A**	**Test1-B:**	**Test2-A**	**Test2-B**
GSE_Merged	0.0274 (9.3%)	<10^-6	3.49e-79	0.0009
GSE42568	0.5988 (3.6%)	<10^-6	3.19e-15	0.3106
GSE35974	0.0738 (16.6%)	<10^-6	1.04e-42	<10^-4
SVM				
GSE_Merged	<10^-4 (15.1%)	<10^-6	8.22e-77	0.0089
GSE42568	0.6472 (3%)	<10^-6	0.00425	0.9305
GSE35974	<10^-4 (34.9%)	<10^-6	4.31e-38	<10^-4

Table 4.6 Results of the two statistical significance tests for all datasets. The average difference in classification accuracy between the extracted signature and random signatures of the same size is displayed in parentheses at Test1-A.

4.4 Signature Consistency

To evaluate the consistency of the extracted signatures, 10 independent executions of the proposed methodology were performed, each resulting in genomic signatures. Then, the “degree of agreement” among the 10 signatures was assessed using CAT plots. The maximum value of the CAT plot x-axis is determined by the size of the smallest signatures extracted, as we are interested in the common genes among all 10 signatures. Generally, a more stable method will have a CAT plot that is higher on the y axis than a less stable method, meaning that independent executions of the method will result in more genes in common on average. There was no (significant) observable difference in the consistency of signatures extracted by RVM of RFE-SVM and once again the two classifiers (which also perform feature selection) yield very similar results, possibly with the exception of GSE42568 where the signature size is very small.

In the case of GSE_Merged there is a strong degree of agreement, resulting in 66% and 71% common genes across all iterations for the RVM and RFE-SVM methods, respectively. The signatures of GSE42568 show 75% overlap when extracted using the RVM and 44% overlap when extracted using RFE-SVM. The 44% overlap of RFE-SVM indicates that the heuristic nature of recursive feature elimination does not consistently as the RVM in signatures of very small size (around 16 genes). However, it still is an acceptable degree of overlap considering the small signature size. Finally, in the case of GSE35974, a high degree of overlap is observed: 74% for RVM and 80% for RFE-SVM. The corresponding CAT plots for the three datasets can be found in figures 4.7, 4.8 and 4.9. Beyond assessing the overall consistency of the extracted signatures, the CAT plots also highlight some regions or “spikes” of interest that possibly should be considered for additional biological evaluation, such as a few (3 and 4) genes the beginning of the RVM CAT plot for GSE_Merged, that are common in all 10 independent signatures (100% overlap). In a similar manner, there appears a “spike” of 4 genes at the RFE-SVM CAT plot of GSE42568.

The above results indicate that the genomic signatures extracted by the above methodology are consistent for independent executions of the methodology that lead to variations of the training data. As such, the robustness and stability of the extracted signatures is highlighted. The strong overlap between the signatures extracted by the same method is maintained as the size of the signature increases, which means that the independent signatures “agree” in general, and not only on their top ranked genes. Please note that the CAT plots were used to assess the “agreement” of independent executions of the methodology for the same feature selection and classification method (RVM or RFE-SVM), in order to assess the stability of the extracted signature. The CAT plots in this section were not used in order to test the agreement between the two feature selection and classification methods. The agreement between RVM and RFE-SVM in terms of selected genes is presented in the CAT plot of the previous section (Figure 4.7).

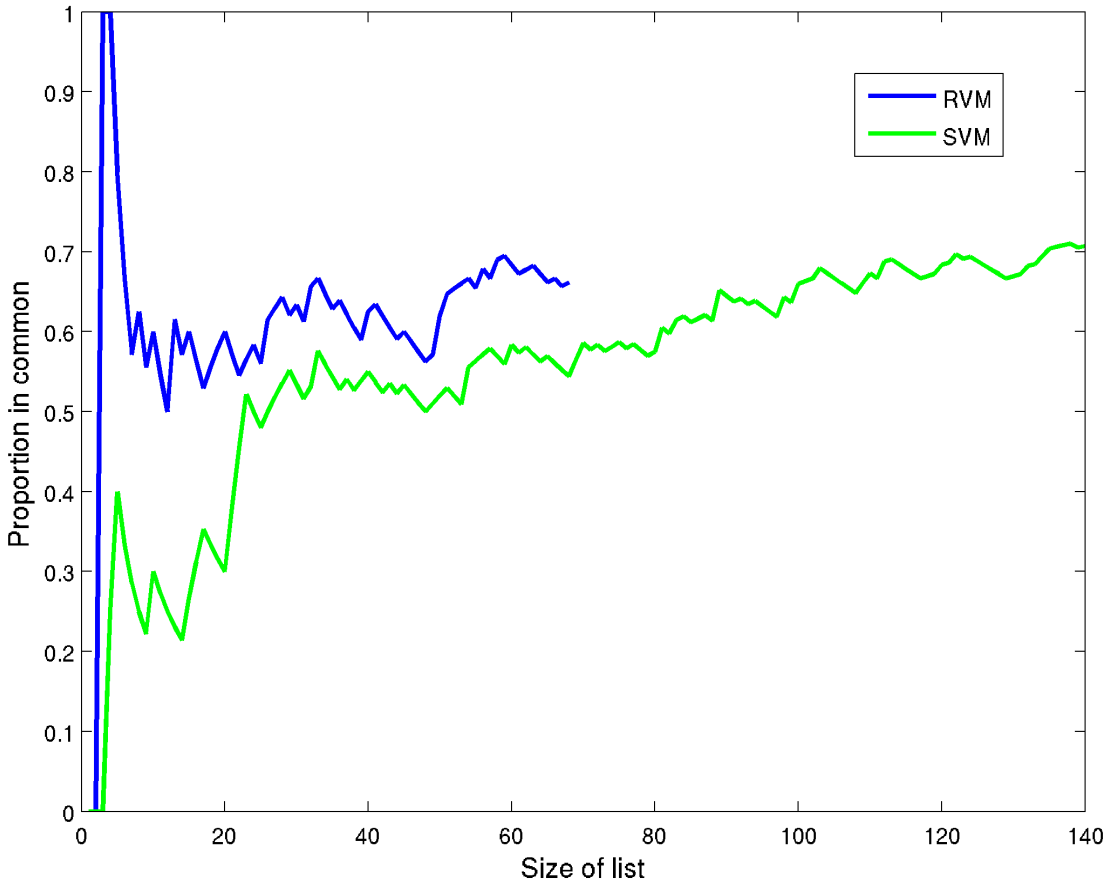

Figure 4.8 Consistency of the GSE_Merged signature.

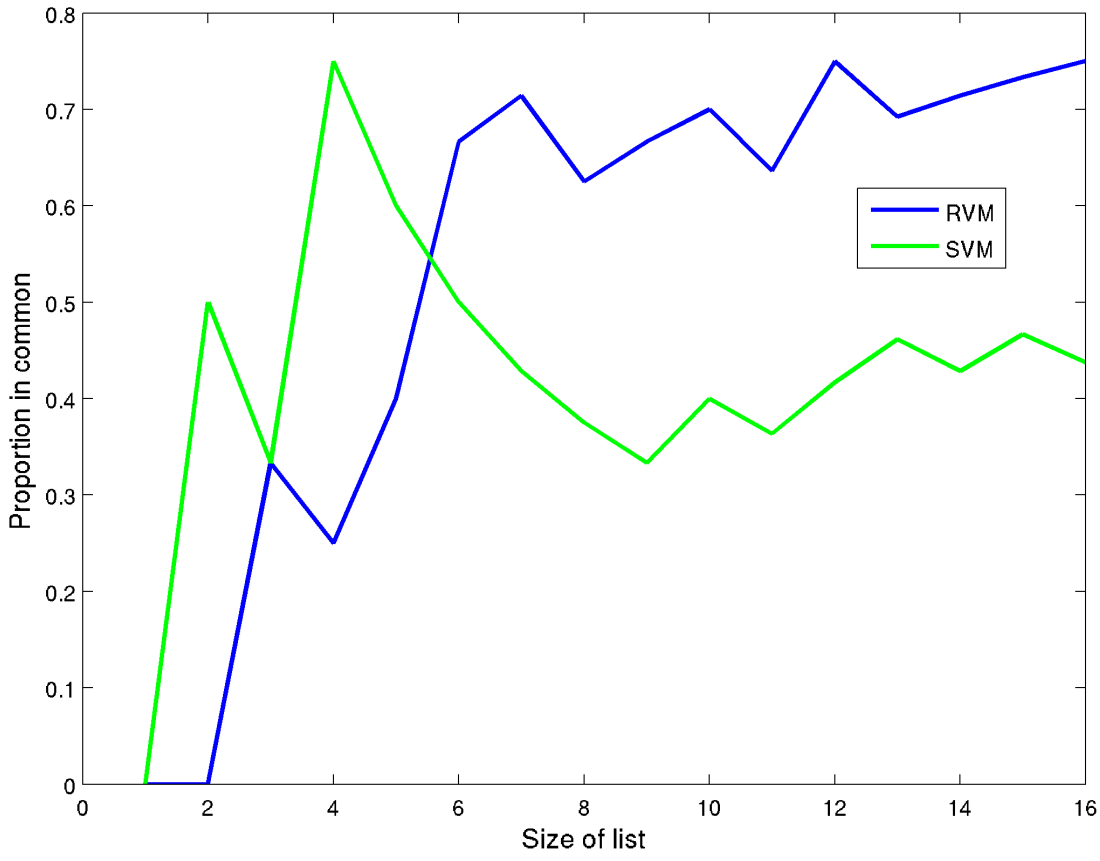

Figure 4.9 Consistency of the GSE42568 signature.

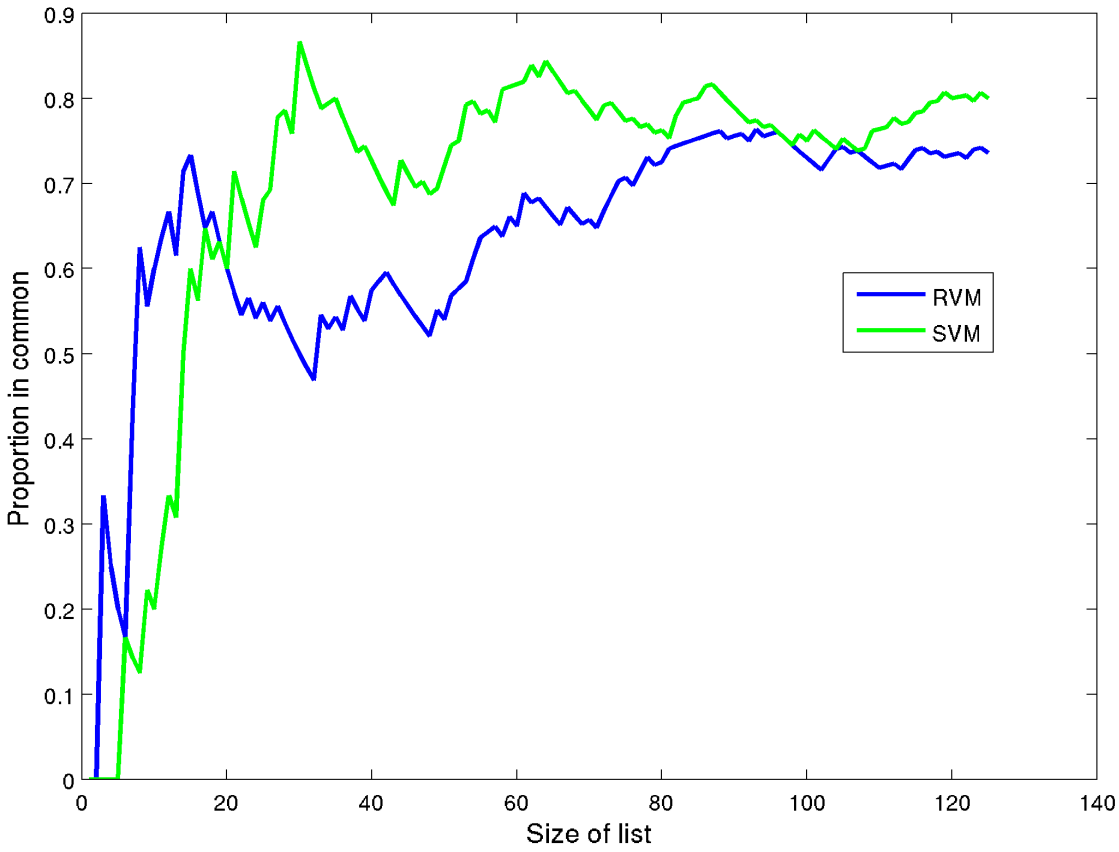

Figure 4.10 Consistency of the GSE35974 signature.

4.5 Biological Evaluation

Measuring the expression levels of different genes in a biological system became possible through high-throughput experimentation techniques such as DNA microarray technology. Genomic analysis by DNA microarrays allows for the identification of “genomic signatures” i.e. set of genes with distinct expression patterns among different classes or with unique features of disease or biological phenotype. An interesting implementation of “genomic signatures” is their usage for the characterization of unknown samples through various classification methods.

In this work, we followed an innovative concept in order to gain gene signatures with statistical significance and high stability, which at the same time reflect the biological or disease phenotype. As mentioned in the methodology section, a number of five datasets (GSE22820, GSE19783, GSE31364, GSE9574, GSE18672) obtained from GEO (Gene Expression Omnibus) repository to generate the GSE_Merged Dataset, while another two free-public high-dimensional GEO Datasets (GSE42568, GSE35974) have been used in order to perform a fair evaluation of the proposed methodology [66]. Two of them are associated to a breast cancer study, i.e. gene expression profiling of primary breast cancer tissues and non cancerous or healthy breast tissues, whereas the third is related to a bipolar disorder study, i.e. transcriptomic levels of postmortem human brain samples, bipolar and control samples. The proposed methodology was carried out on (i) the 11928 **filtered** probesets of GSE_Merged dataset, (ii) the 54676 unfiltered probesets of GSE42568 dataset, and (iii) the 33297 unfiltered probesets of GSE35974 [67], resulted in the selection of stable probesets that were found to be differentially expressed on the corresponding datasets. For each dataset, two different gene signatures were extracted by applying the classification methods RVM and RFE-SVM (Appendix B). The probe identifiers from each gene signature were mapped to unique Entrez Gene Ids, upon which pathway enrichment analysis has been performed.

In particular, to gain insight into the biological significance represented in the gene expression patterns associated with breast cancer or bipolar disorder, we used two approaches. First, we used two annotation tools, namely WebGestalt [68] and GATHER [69], aiming at a functional enrichment of biological pathways associated with the identified gene sets. A functional enrichment, also known as pathway analysis, is recognized as a secondary analysis on large gene sets that resulted from high-throughput genomic methods. Enrichment analysis can yield terms which are statistically over-or under-represented within a genomic signature of interest, providing valuable information concerning the biological functions associated with the signature. Second, we used three Databases, G2SBC [70], BDgene [71], PsyGeNET [72] in order to explore a gene-disease association interpretation within the identified gene signatures.

Briefly, the information about the panel of these bionformatic tools for Pathway Enrichment Analysis and Gene-Disease Association Analysis can be summarized as follows:

- **WebGestalt** “WEB-based GEne SeT AnaLysis Toolkit“ is a “*web-based integrated data mining system that incorporates information from different public resources and supports biologists in exploring large sets of genes generated from genomic, proteomic and large-scale genetic studies”* [68]. Statistic: *P*-value computed using the hypergeometric distribution, Benjamini & Hochberg (1995).
- **GATHER** is a “*tool that integrates various forms of available data to elucidate biological context within molecular signatures produced from high-throughput post-genomic assays”* [69]. Statistic: *P*-value computed using the hypergeometric distribution or chi-square test.
- **G2SBC** “The Genes-to-Systems Breast Cancer” is a “*bioinformatics resource that collects and integrates data about genes, transcripts and proteins which have been reported in literature to be altered in breast cancer cells”*. Also, it is a “*multilevel resource dedicated to the molecular and systems biology of breast cancer, including both the building-blocks level (genes, transcripts and proteins) and the systems level (molecular and cellular systems, cell populations)”* [70]. Statistic: *P*-value computed using the cumulative hypergeometric distribution.

- **BDgene** “Genetic Database for Bipolar Disorder”: The BDgene Database **“***was developed to address the genetic complexity of Bipolar Disorder and its overlap with schizophrenia (SZ) and major depressive disorder (MDD). By profound literature screening, BDgene integrates not only multi-type BD-related genetic factors (e.g. SNP, CNV, gene, pathway), but also overlapping genetic factors between BD and SZ/MDD, which presents a panoramic view of current genetic studies for BD”* [71].
- **PsyGeNET** “Psychiatric disorders Gene association NETwork”: **PsyGeNET** is a *“resource for the exploratory analysis of psychiatric diseases and their associated genes. PsyGeNET database is the result of the integration of information from DisGeNET and data extracted from the literature by text mining, followed by curation by domain experts”* [72].

Moreover, the KEGG (Kyoto Encyclopedia of Genes and Genomes) Database was considered by both GATHER and WebGestalt enrichment analysis, as it provides a reliable source of pathways [73]. Also, it should be noted that KEGG pathway inference from network was also searched by pathway analysis conducted from GATHER, in order to gain further functional relationships of genes that compose the gene signatures. This enabled the extent of annotations, which relies on an analysis of a network of genes in the literature, in order to unfold a broader scope of functional roles that are not immediately obvious in the gene signature.

As mentioned previously, by using statistical methods such as the hypergeometric distribution, it is possible to determine that specific pathways are enriched in a candidate genomic signature, which allows us to assume that these biological paths have important functions in the study undertaken. Indeed, enrichment analysis identified statistically significant KEGG pathways in all six genomic signatures (3 RVM and 3 RFE-SVM), as illustrated in Tables below (Tables 4.7, 4.8, and 4.9). It is important to notice that both algorithms resulted in meaningful genomic signatures, despite the fact that were implemented either on **filtered** probeset of GSE_Merged dataset or unfiltered probesets of GSE42568 and GSE35974 datasets, as well as on different diseases (breast cancer and bipolar disorder).

To facilitate the comparison of the biological pathways associated with the identified gene signatures (RVM or RFE-SVM gene signatures) in each dataset, we provide this information in Table 4.7 for Breast Cancer Dataset GSE_Merged, Table 4.8 for Breast Cancer Dataset GSE42568 and Table 4.9 for Bipolar Disorder Dataset GSE35974. As shown in Tables 4.7, 4.8, and 4.9, in each Dataset the common pathways among RVM or RFE-SVM gene signatures are underlined.

GSE_Merged "Breast Cancer"	
WebGestalt	**Enriched KEGG Pathways (p≤0.05)**
RVM (19) **7 with more than 2 genes**	Pathways in cancer, Cytokine-cytokine receptor interaction, Regulation of actin cytoskeleton, MAPK signaling pathway, Cell cycle, Focal adhesion, Neuroactive ligand-receptor interaction
RFE-SVM (36) **19 with more than 2 genes**	Protein digestion and absorption, ECM-receptor interaction, Focal adhesion, Drug metabolism - cytochrome P450, Cytokine-cytokine receptor interaction, Drug metabolism - other enzymes, Pathways in cancer, Metabolic pathways, Metabolism of xenobiotics by cytochrome P450, ErbB signaling pathway, MAPK signaling pathway, Regulation of actin cytoskeleton, Phagosome, Chemokine signaling pathway, Toll-like receptor signaling pathway, Osteoclast differentiation, Oxidative phosphorylation, Jak-STAT signaling pathway, Endocytosis
GATHER	**Enriched KEGG Pathways (p≤0.05)**
RVM (2)	Cytokine-cytokine receptor interaction, Cell cycle
6 Inferred from Network (1 similar*)	Cytokine-cytokine receptor interaction, Jak-STAT signaling pathway, MAPK signaling pathway, Focal adhesion, Complement and coagulation cascades, Insulin signaling pathway
RFE-SVM (4)	ECM-receptor interaction, Focal adhesion, ATP synthesis, Cytokine-cytokine receptor interaction
8 Inferred from Network (3 similar*)	Cytokine-cytokine receptor interaction, Jak-STAT signaling pathway, MAPK signaling pathway, Toll-like receptor signaling pathway, Pentose and glucuronate interconversions, Focal adhesion, ECM-receptor interaction, Complement and coagulation cascades
G2SBC	**Enriched or Depleted KEGG Pathways (p≤0.05)**
RVM **1 depleted**	Metabolic pathways
RFE-SVM **3 enriched** **2 depleted**	 ECM-receptor interaction, Caffeine metabolism, Drug metabolism - cytochrome P450 Neuroactive ligand-receptor interaction, Metabolic pathways
G2SBC	**Number of genes associated with breast cancer**
RVM	**31 genes representing 44,28%**
RFE-SVM	51 genes representing 36,17%

Table 4.7 Results of Enrichment Analysis and Gene-Breast Cancer Association for GSE_Merged. RVM: 70 gene signature (70 genes mapped); RFE-SVM: 141 gene signature (141 genes mapped); *similar with 1st enrichment by GATHER.

The identified enriched KEGG pathways in GSE_Merged, and GSE42568 Datasets are involved in various processes, such as signal transduction (e.g. MAPK signaling pathway, ErbB signaling pathway, Jak-STAT signaling pathway, TGF-beta signaling pathway), signaling interaction (e.g. Cytokine-cytokine receptor interaction), metabolism (e.g. metabolic pathways), energy metabolism (e.g. oxidative phosphorylation), xenobiotics biodegradation and metabolism (e.g. Drug metabolism - cytochrome P450), transport and catabolism (e.g. Phagosome, Endocytosis), cell growth and death (e.g. cell cycle), cellular community (e.g. Focal Adhesion), development (e.g. osteoclast differentiation), immune system (e.g. Chemokine signaling pathway, Toll-like receptor signaling pathway), digestive system (e.g. protein digestion and absorption), diseases (e.g. pathways in cancer), underlying their importance in both breast cancer studies (GSE_Merged, GSE42568). It is inferred from our study, but also confirmed by numerous studies that these pathways are important in many aspects of breast cancer progression [74-80].

GSE42568 – "Breast Cancer"	
WebGestalt	**Enriched KEGG Pathways (p≤0.05)**
RVM (7)	Cell adhesion molecules (CAMs), Nicotinate and nicotinamide metabolism, Salivary secretion, Lysosome, Pancreatic secretion, ECM-receptor interaction, Osteoclast differentiation
RFE-SVM (10)	ECM-receptor interaction, Focal adhesion, Complement and coagulation cascades, Protein digestion and absorption, TGF-beta signaling pathway, Cell cycle, Ubiquitin mediated proteolysis, Oocyte meiosis, PPAR signaling pathway, Phagosome
GATHER	**Enriched KEGG Pathways (p≤0.05)**
RVM (3)	Nicotinate and nicotinamide metabolism, ECM-receptor interaction, Calcium signaling pathway
2 Inferred from Network (none similar*)	Cytokine-cytokine receptor interaction, Jak-STAT signaling pathway
RFE-SVM (6)	Focal adhesion, Ubiquitin mediated proteolysis, Complement and coagulation cascades, TGF-beta signaling pathway, ECM-receptor interaction, Cell cycle
Inferred from Network (2 similar*)	ECM-receptor interaction, Focal adhesion
G2SBC	**Enriched or Depleted KEGG Pathways (p≤0.05)**
RVM (1)	Cell adhesion molecules (CAMs)
RFE-SVM (1)	ECM-receptor interaction
G2SBC	**Number of genes associated with breast cancer**
RVM	4 genes representing 30,77%
RFE-SVM	**5 genes representing 41.67%**

Table 4.8 Results of Enrichment Analysis and Gene-Breast Cancer Association for GSE42568. RVM: 16 gene signature (13 genes mapped); RFE-SVM: 16 gene signature (12 genes mapped); *similar with 1st enrichment by GATHER.

Similar, as shown in below Table 4.9, the enriched KEGG pathways that are involved in processes such as lipid metabolism (e.g. Steroid hormone biosynthesis), immune system (e.g. Toll-like receptor signaling pathway), carbohydrate metabolism (e.g. Citrate cycle (TCA cycle), metabolism of other amino acids (e.g. Glutathione metabolism), cell growth and death (e.g. apoptosis), nervous system (e.g. Neurotrophin signaling pathway) in bipolar study (GSE35974), underlie their functional role in bipolar disorder according to several research studies [81-86].

GSE35974 – "Bipolar Disorder"	
WebGestalt	**Enriched KEGG Pathways (p≤0.05)**
RVM (7) 1 with more than 2 genes	Pathways in cancer, Neurotrophin signaling pathway, Steroid hormone biosynthesis, Metabolism of xenobiotics by cytochrome P450, Toll-like receptor signaling pathway, Ubiquitin mediated proteolysis, Osteoclast differentiation
RFE-SVM (5) 2 with more than 2 genes	Metabolism of xenobiotics by cytochrome P450, Steroid hormone biosynthesis, Pathways in cancer, Ubiquitin mediated proteolysis, Neuroactive ligand-receptor interaction
GATHER	**Enriched KEGG Pathways (p≤0.05)**
RVM (2) Inferred from Network (2 similar*)	Apoptosis, Toll-like receptor signaling pathway Apoptosis, Toll-like receptor signaling pathway
RFE-SVM (1) Inferred from Network (none similar*)	Citrate cycle (TCA cycle) Toll-like receptor signaling pathway, Apoptosis, Glutathione metabolism, Cytokine-cytokine receptor interaction
BDgene	**Number of genes associated with bipolar disorder**
RVM	2 genes representing 3,08%
RFE-SVM	**5 genes representing 10,42%**
PsyGeNET	**Number of genes associated with bipolar disorder**
RVM	2 genes representing 3,08%
RFE-SVM	**6 genes representing 12,5%**

Table 4.9 Results of Enrichment Analysis and Gene-Breast Cancer Association for GSE35974. Common Genes in RVM & RFE-SVM: 27 Common Protein Coding Genes; 25 common non coding RNA; 2 common Unknown; 15 common Controls); RVM: 127 gene signature (39 Unique Protein Coding Genes; 11 Unique non coding RNA; 5 Unknown; 3 Controls) - RVM: 66 Protein Coding Genes (65 mapped); RFE-SVM: 132 gene signature (21 Unique Protein Coding Genes; 10 Unique non coding RNA; 5 Unknown; 27 Controls) - RFE-SVM: 48 Protein Coding Genes (48 mapped), *similar with 1st enrichment by GATHER.

To facilitate the interpretation of the RVM and RFE-SVM algorithms within the terms of biological pathways, we considered the pathways that converge in both WebGestalt and GATHER databases, as illustrated in the below Table 4.10.

Comparison of Gene Signatures Common KEGG Pathways in WebGestalt & GATHER		
	Common Pathways (p≤0.05)	Common Pathways (p≤0.05)
Datasets	**RVM**	**RFE-SVM**
GSE_Merged	Cytokine-cytokine receptor interaction MAPK signaling pathway Jak-STAT signaling pathway Focal adhesion Insulin signaling pathway Cell cycle	Cytokine-cytokine receptor interaction MAPK signaling pathway Jak-STAT signaling pathway Focal adhesion Toll-like receptor signaling pathway ECM-receptor interaction
GSE42568	ECM-receptor interaction Nicotinate and nicotinamide metabolism	ECM-receptor interaction Focal adhesion Complement and coagulation cascades TGF-beta signaling pathway Cell cycle Ubiquitin mediated proteolysis
GSE35974	Toll-like receptor signaling pathway	None

Table 4.10 Comparison of Gene Signatures in the context of their convergence by applying WebGestalt and GATHER.

The common pathways in RVM and RFE-SVM genomic signatures of GSE_Merged implicate processes such as signaling molecules and interaction, signal transduction, and cellular community processes, whereas the different pathways implicate immune system (RFE-SVM), as well as cell growth/and death, and endocrine system (RVM) processes. However, more different pathways appear in the RVM and RFE-SVM genomic signatures of GSE42568, as shown in Table 4.10. Specifically, metabolism of cofactors and vitamins appear in the RVM genomic signature, while cellular community, immune system, signal transduction, cell growth/and death, and folding, sorting and degradation processes appear in RFE-SVM signature. ECM-receptor interaction (signaling molecules and interaction) is the common pathway in these genomic signatures. Surprisingly, in GSE35974, only one pathway yielded by WebGestalt and GATHER, the Toll-like receptor signaling pathway, and only in RVM genomic signature.

Moreover by comparing the KEGG pathways resulted from WebGestalt and GATHER enrichment analyses in both RVM and RFE-SVM gene signatures in all three datasets, we conclude that:

- RVM and RFE-SVM extracted gene signatures are biologically important
- Both Signatures are governed from biologically meaningful pathways with disease relevance
- RVM and RFE-SVM in GSE_Merged seems to be biologically similar
- RVM and RFE-SVM Signatures in GSE42568 appear to be biologically different
- RVM Signature in GSE35974 seems to provide a pathway consistency between WebGestalt and GATHER compared to RFE-SVM
- Differences between them can be traced back to signature size, or gene information which is unknown at this time (e.g. unmapped genes, non coding RNAs)

Overall, the genomic signatures extracted by RVM and RFE-SVM algorithms provide: (i) enrichment of various KEGG pathways, which are implicated in breast cancer (GSE_Merged, GSE42568) [74-80], or bipolar disorder (GSE35974) [81-86], (ii) enrichment of additional KEGG pathways through network inference, (iii) association of a high number of genes with breast cancer, and (iv) association of a low number of genes with bipolar disorder, which was expected, bearing in mind the very limited information we have in this field. These results support our proposed methodology, which provides a reliable methodology to extract meaningful genomic signatures and was applied to different datasets (filtered, unfiltered) and different phenotypes (breast cancer, bipolar disorder). Finally, these results strengthen the role of RVM algorithm, as an alternative algorithm for genomic signature extraction.

Conclusion and Future Work

Selecting lists of candidate genes according to selection accuracy may produce unstable results, which cause skepticism and hinder the clinical application of findings. In essence, stability is linked to reproducibility and as such, a procedure for selecting candidate genes concerning a biological process should yield stable and repeatable results. To address this issue, an appropriate methodological framework is introduced, which introduces some key innovative points: The extracted genomic signature is stable and robust. The predictive performance of the extracted signature is estimated not only using accuracy, but also sensitivity and specificity which results in a more reliable estimation of its actual predictive capability. Moreover, by using a hybrid method, appropriate and necessary confidence intervals are plotted for the metrics of predictive performance, which are omitted in most studies. Confidence intervals are a necessary tool, since they account for the degree of variability in the observed metrics of predictive performance. This is an important aspect since omitting to account for this variability will probably lead to false assessment of the observed results. Another innovative aspect of the proposed methodology is the introduction of proper statistical tests for the assessment of the quality of the extracted signature in terms of statistical significance. As illustrated in the results section, these statistical tests successfully identify the cases where assessing the importance of a signature based only on classification performance is misleading, while they confirm the statistical significance of actually informative signatures that reflect the underlying biological processes of interest. Finally, the stability of the signatures extracted be the proposed framework is confirmed by assessing the degree of “agreement” among independent executions of the methodology using a special kind of plot called the CAT plot.

The first step of the proposed methodology utilizes a dual feature selection scheme, which aims to combine the advantages of univariate and multivariate feature selection methods, extracting gene sets that yield good differentiation of their expression values among the classes of interest and have maximal predictive performance. This dual feature selection step is more computationally efficient than the direct application of multivariate feature selection, while capturing the associations among different features (genes). During the next step of the proposed methodology, bootstrap resampling is utilized in order to extract candidate gene signatures according to the percentiles of gene selection frequency across all bootstrap datasets. In order to avoid the computational overhead of generating unnecessary additional bootstrap datasets, a stability criterion is introduced, concerning the average number of genes “**G**” selected across all bootstrap datasets by the feature selection and classification method. The top “**G**” genes are also considered as an additional candidate genomic signature. Next, the predictive performance of all candidate signatures is assessed in terms of classification accuracy, sensitivity (true positive rate) and specificity (true negative rate). Assessing all three performance measures is necessary, since taking only classification accuracy into account can lead to misinterpretation of the actual predictive performance, especially in imbalanced datasets which are usually the case in DNA microarray analysis. Then, the signature which yields the best predictive performance among all candidate signatures is extracted as the stable genomic signature of the methodology. Additionally, since the extracted gene signature is selected among many candidate signatures according to predictive performance, it is very important that appropriate confidence intervals are generated. That is, when two or more signatures have similar predictive performance, the signature consisting of fewer genes could be preferred, since it leads to a simpler and sparser model. Moreover, since the proposed methodology also calculates the sensitivity and specificity achieved by each candidate signature, it allows for additional criteria of signature extraction, based on which prognostic aspect of the signature is deemed to be more important. It should be noted however, that in the datasets studied in this thesis sensitivity and specificity appeared to increase or decrease for each candidate signature simultaneously with classification accuracy. In that manner, selecting candidate genes was based primarily on classification accuracy and the other two metrics were considered complementary. Moreover, the predictive performance of all candidate signatures was displayed in an elaborate plot, allowing further

inspection. Assessing whether different candidate signatures, as well as different feature selection and classification methods yield results that are different in a statistically significant manner is another innovative point of the proposed methodology. This is contrary to most microarray studies which ignore the concept of confidence intervals when assessing predictive performance. That is, if two classification methods result in similar classification accuracy, e.g. 90% and 93%, most studies claim that the second method is better since it results in a small increase in classification accuracy. However, this claim is usually wrong since appropriate confidence intervals should be generated that reflect the uncertainty of the observed results, especially in small-N, large-P problems, such as DNA microarray analysis. Moreover, in order to counter the negative effects caused by imbalanced datasets in classification, the effectiveness of random oversampling and random undersampling schemes for balancing the data is investigated. During the next step of the proposed methodology, the statistical significance of the extracted signature is assessed using two distinct statistical tests. The first test assesses whether the extracted signature significantly outperforms random signatures of the same size in terms of classification accuracy, while the second test assesses whether the extracted signature is significantly more associated to the response variable (phenotype/class label), according to a gene set analysis method. Finally, the last step of the proposed methodology utilizes a type of plot called the correspondence at the top (CAT) plot, which measures the degree of “agreement” among signatures extracted by independent executions of the proposed methodology, in order to assess the robustness and stability of the extracted signature.

The proposed methodology is tested on three publicly available datasets, which highlight the advantages of the proposed framework, as well as the limitation of extracting gene signatures according to their predictive performance in terms of classification accuracy. The methodology is run twice on each dataset, once using the RVM and once using the RFE-SVM method for feature selection and classification and both methodologies achieve similar classification performance. However, RFE-SVM is significantly faster, probably since it is set to eliminate half the features during each round of recursive feature elimination. Additionally, the effectiveness of random sampling methods for countering the effect of class imbalance in classification is questionable. While there seems to be a consistent trend where specificity increased at the cost of specificity, the differences are not statistically significant for the available data. For two of the three datasets (GSE_Merged and GSE35974) the methodology yields exceptional results, extracting gene signatures with maximal predictive performance, which are better in a statistically significant manner, in terms of classification accuracy and association to the response, when compared to random signatures of the same size. Moreover, there is a strong overlap of the genes selected by the two different feature selection and classification methods. Contrary to the other two datasets, in the case of GSE42568 the methodology is partially successful. This is a result of GSE42568 being “problematic”, since the classification task seems to be “too easy” for this dataset and almost any gene set can achieve 100% classification accuracy, probably due to the dataset being strongly imbalanced and only having 14% control samples (17 out of 121). While classification performance for the extracted signature is excellent, the signature fails at the first part of the first test of statistical significance. That is, roughly 60% of random signatures achieve equal or better classification accuracy than the extracted signature (59.88% for the RVM and 64.72% for RFE-SVM). This is reflected in the lack of overlap among the signatures extracted by the RVM and RFE-SVM methods for GSE42568. It is interesting however that the signatures extracted for GSE42568 succeed at the second part of the first test and are on average around 3% better than random signatures in terms of classification accuracy (3.6% for the RVM and 3% for RFE-SVM). While 3% might seem like a small difference due to random noise, its statistical significance is confirmed by the second part of the fist test (permutation test). Moreover, both the RVM and RFE-SVM signatures of GSE42568 fail at the second part of the second test. Which means that the GSE42568 signatures are not more associated to the response than random signatures of the same size. This is another indication of the questionable practical significance of the GSE42568 signatures in terms of extracting biological knowledge. The above limitation can be interpreted and expressed as: when any gene set has similar (in this case exceptional) classification performance, then

selecting genes based on accuracy is not a good course of action, since in this case classification accuracy is not a good criterion for gene selection. However, when different gene sets yield different predictive performance, such as the case of GSE_Merged and GSE35974, extracting signatures based on classification accuracy can achieve very good results. This limitation observed in GSE42568 also highlights the importance of utilizing appropriate statistical tests in order to assess the statistical significance of the extracted signatures. Moreover, the signatures extracted by independent executions of the proposed methodology were consistent, according to corresponding the CAT plots. As such, the stability and robustness of the signatures extracted by the proposed methodology is confirmed. Finally, the signatures extracted were biologically meaningful, since they were found to be associated with the underlying biological process of interest in terms of KEGG pathways and disease associated genes.

In terms of future work, more feature selection and classification methods such as random forests and and l1-regularized logistic regression could be used. Additionally, different univariate filtering methods could be used in the preliminary feature elimination step and the difference in the resulting signatures could be further studied. Moreover, the use of CAT plots for measuring the “degree of agreement” among different executions of the same method lead to some interesting “spikes” of common genes. These “spikes” of the consistency step could be further studied by field experts. Additionally, CAT plots could be also employed at the stable gene extraction step in order to assess stability of results (instead of the current criterion), or could also be employed at the next step, the estimation of classification performance, in order to identify additional candidate genomic signatures, whose predictive performance should be assessed. Moreover, since imbalanced datasets are such a frequent case in DNA microarray studies, additional techniques countering this problem should be investigated. Since a trend appeared but it was not statistically significant for the available data, it is suggested that more data should be acquired. However, obtaining additional samples is at the very least impractical and usually impossible. Alternatively, by slightly relaxing the assumption of sample independence in order to plot confidence intervals, the number of folds and repetitions of the repeated K-Fold Cross Validation could be increased, which would lead to more samples being classified and could possibly lead to statistically significant differences in the effects of random sampling on the measures of classification performance. Finally, it could be investigated whether the dataset balancing techniques (random sampling) which where currently used at the estimation of classification performance step would influence the genes selected, if they were also applied to the stable signature extraction step.

References

[1] N. K. Chlis, “Comparison of Statistical Methods for Genomic Signature Extraction”, Diploma Thesis, Technical University of Crete, 2013.

[2] N. K. Chlis, S. Sfakianakis, E. S. Bei and M. Zervakis, “A Generic Framework for the Elicitation of Stable and Reliable Gene Expression Signatures”, in Proc. IEEE BIBE, Chania Greece, 2013. DOI 10.1109/BIBE.2013.6701527.

[3] N. K. Chlis, E. S. Bei, S. Sfakianakis, D. Iliopoulou, D. Kafetzopoulos and M. Zervakis, “Searching for Significant Genes in Cancer Metastasis by Tissue Comparisons”, in Proc. MBEC, 2014.

[4] N.K. Chlis, and M. Zervakis, “Algorithmic Stability Affects the Extraction of Gene Expression Signatures”, proceedings of the 6th Panhellenic Conference on Biomedical Technology, 2015.

[5] N. K. Chlis, E. S. Bei, K. Moirogiorgou and M. Zervakis, “Extracting Reliable Gene Expression Signatures through Stable Bootstrap Validation”, in Proc. IEEE EMBC, Milan, Italy, 2015.

[6] S. Sfakianakis, E. S. Bei, M. Zervakis, D. Vassou and D. Kafetzopoulos, “On the Identification of Circulating Tumor Cells in Breast Cancer,” IEEE Journal of Biomedical and Health Informatics, vol. 18, no. 3, pp. 773-782, 2014. doi: 10.1109/JBHI.2013.2295262.

[7] H. Zengyou and Y. Weichuan, “Stable feature selection for biomarker discovery,” Computational Biology and Chemistry, vol. 34, pp. 215–225, 2010. doi:10.1016/j.compbiolchem.2010.07.002.

[8] L. Yu, Y. Han and M. E. Berens, “Stable Gene Selection from Microarray Data via Sample Weighting,” IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 9, pp. 262-272, 2012.

[9] L. Yu, C. Ding and S. Loscalzo, “Stable Feature Selection via Dense Feature Groups,” in Proc. 14th ACM SIGKDD, pp. 803-811, 2008.

[10] A. L. Boulesteix and M. Slawski, “Stability and aggregation of ranked gene lists,” Briefings in Bioinformatics, vol. 10. no. 5, pp. 556-568, 2009.

[11] C. A. Davis, F. Gerick, V. Hintermair, C. C. Friedel, K. Fundel, R. Küffner, R. Zimmer, "Reliable gene signatures for microarray classification: assessment of stability and performance," Bioinformatics., vol. 22, no. 19, pp. 2356–2363, 2006.

[12] S. Y. Neo, C. K. Leow, V. B. Vega, P. M. Long, A. F.M. Islam, P. B.S. Lai, E. T. Liu, E. C. Ren, "Identification of Discriminators of Hepatoma by Gene Expression Profiling Using a Minimal Dataset Approach," HEPATOLOGY., vol. 39, pp. 944-953, 2004.

[13] I. Suzuki, T. Takenouchi, M. Ohira, S. Oba, S. Ishii, "Robust Model Selection for Classification of Microarrays," Cancer Informatics., vol.7, pp. 141–157, 2009.

[14] A.C. Haury, P. Gestraud, J.P. Vert , "The Influence of Feature Selection Methods on Accuracy, Stability and Interpretability of Molecular Signatures,", PLoS ONE 6(12): e28210. doi:10.1371/journal.pone.0028210, 2011.

[15] R. Armañanzas, I. Inza, P. Larrañaga , "Detecting reliable gene interactions by a hierarchy of Bayesian network classifiers," Computer Methods and Programs in Biomedicine., vol. 91, pp.110-121, 2008.

[16] A. García-Bilbao, R. Armañanzas, Z. Ispizua, B. Calvo, A. Alonso-Varona, I. Inza, P. Larrañaga, G. López-Vivanco, B.Suárez-Merino, M. Betanzos, "Identification of a biomarker panel for colorectar cancer diagnosis," BMC Cancer 2012, 12:43

[17] A. Barrier, P. Boelle, F. Roser, J. Gregg, C. Tse, D. Brault, F. Lacaine, S. Houry, M. Huguier, B. Franc, A. Flahault, A. Lemoine, S. Dudoit, "Stage II Colon Cancer Prognosis Prediction by Tumor Gene Expression Profiling," Journal of Clinical Oncology., vol. 24, no. 29, pp. 4665-4691, 2006.

[18] B.Efron, "Bootstrap Methods: Another Look at the Jackknife," The Annals of Statistics., vol. 7, no. 1, pp. 1-26, 1979.

[19] M. Kathleen Kerr, G. A. Churchill, "Bootstrapping cluster analysis: Assessing the reliability of conclusions from microarray experiments," PNAS., vol. 98, no. 16, pp. 8961-8965, 2001.

[20] N. Friedman, M. Goldszmidt, A. Wyner, "Data Analysis with Bayesian Networks: A Bootstrap Approach," ., Proc. Fifteenth Conf. on Uncertainty in Artificial Intelligence (UAI), 1999., 1999

[21] R. Maglietta, A. D'Addabbo, A. Piepoli, F. Perri, S. Liuni, G. Pesole, N. Ancona, "Selection of relevant genes in cancer diagnosis based on their prediction accuracy," Artificial Intelligence in Medicine., vol. 40, pp. 29-44, 2007.

[22] R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

[23] N. K. Chlis, personal website: http://users.isc.tuc.gr/~nchlis/

[24] D. Stekel, "Microarray Bioinformatics", Cambridge University Press, 2003.

[25] Emmert-Streib, M. Dehmer, "Analysis of Microarray Data", Wiley-VCH, 2008.

[26] A. Zhang, "Advanced Analysis of Gene Expression Microarray Data", World Scientific Publishing, 2006.

[27] D. P. Bernar, W. Dubitzky, M. Granzow, "A Practical Approach to Microarray Data Analysis", Kluwe Academic Publishers, 2003

[28] National Human Genome Research Institute, http://www.genome.gov/

[29] 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.

[30] 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.

[31] F. Chibon, "Cancer gene expression signatures - the rise and fall?," Eur J Cancer, vol. 49, no. 8, pp. 2000-2009, 2013.

[32] T. Hastie, R. Tibshirani, J. Friedman, "The Elements of Statistical Learning: Data Mining, Inference, and Prediction second edition," Springer, 2009.

[33] Christopher M. Bishop, "Pattern Recognition and Machine Learning," Springer, 2006.

[34] Richard O. Duda, Peter E. Hart, David G. Stork , "Pattern Classification, 2nd edition," Wiley, 2000.

[35] S. Theodoridis, K. Koutroumbas, "Pattern Recognition, 4th Edition," Elsevier, 2009.

[36] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in Proc. 14th international joint conference on Artificial intelligence (IJCAI'95), vol. 2, pp. 1137-1143, 1995.

[37] T. Fawcett. "An introduction to ROC analysis", Pattern Recognition Letters, vol. 27, no. 8, pp. 861-874, 2006.

[38] J. Van Hulse, T. M. Khoshgoftaar, A. Napolitano, "Experimental perspectives on learning from imbalanced data", In Proceedings of the 24th international conference on Machine learning, pp. 935-942, 2007.

[39] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, A. Napolitano, "Mining data with rare events: a case stud", Proceedings of IEEE ICTAI, vol. 2, pp. 132-139, 2007.

[40] J. S. Long, “Regression Models for Categorical and Limited Dependent Variables,” SAGE publications, 1997.

[41] C. Cortes, V. Vapnik, "Support-Vector Networks," Machine Learning., vol. 20, pp. 273-297, 1995.

[42] I. Guyon, J. Weston, S. Barnhill and V. Vapnik, “Gene Selection for Cancer Classification using Support Vector Machines,” Machine Learning, vol. 46, pp. 389–422, 2002.

[43] M. E. Tipping, “Sparse Bayesian Learning and the Relevance Vector Machine,” Journal of Machine Learning Research vol. 1, pp. 211-244, 2001.

[44] S. Boyd, L. Vandenberghe, “Convex Optimization”, Cambridge University Press, 2004.

[45] A. Papoulis and S. U. Pillai, Probability, Random Variables and Stochastic Processes. McGraw-Hill Europe, 2002.

[46] D. S. Moore, G. P. McCabe and B.A. Craig, Introduction to the Practice of Statistics, Sixth Edition. W. H. Freeman and Company, 2009.

[47] D. M. Diez, C. D. Barr and M. Çetinkaya-Rundel, OpenIntro Statistics, Second Edition. CreateSpace Independent Publishing Platform, 2012.

[48] "Wolfram MathWorld - Weak Law of Large Numbers", available online at: http://mathworld.wolfram.com/WeakLawofLargeNumbers.html

[49] Y. Saeys, I. Inza, P. Larrañaga, A review of feature selection techniques in bioinformatics," Bioinformatics., vol. 23, no. 19, pp. 2507–2517, 2007. doi:10.1093/bioinformatics/btm344

[50] I. Guyon, A. Elisseeff, "An Introduction to Variable and Feature Selection," Journal of Machine Learning Research., vol. 3, pp. 1157-1182, 2003.

[51] 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.

[52] V. G. Tusher, R. Tibshirani and G. Chu, “Significance analysis of microarrays applied to the ionizing radiation response,” Proceedings of the National Academy of Sciences, vol. 98, no. 9, pp. 5116–5121, 2001. doi:10.1073/pnas.091062498.

[53] R. Tibshirani, G. Chu, B. Narasimhan and J. Li, samr: SAM: Significance Analysis of Microarrays (R package). Available: http://cran.r-project.org/web/packages/samr/.

[54] H. Maciejewski, “Gene set analysis methods: statistical models and methodological differences,” Briefings in Bioinformatics, vol.15, no. 4, pp. 504–518, 2014. doi: 10.1093/bib/bbt002.

[55] J. J. Goeman, S. A. van de Geer, F. de Kort and H. C. van Houwelingen, “A global test for groups of genes: testing association with a clinical outcome,” Bioinformatics, vol. 20 no. 1, pp. 93–99, 2004. DOI: 10.1093/bioinformatics/btg382.

[56] J. J. Goeman1, and Peter Bühlmann, “Analyzing gene expression data in terms of gene sets: methodological issues,” Bioinformatics, vol. 23, no. 8, pp. 980-987, 2007. doi: 10.1093/bioinformatics/btm051

[57] J.Goeman, J.Oosting, L.Finos and A. Solari, “The Global Test and the globaltest R package” (R package). Available: http://www.bioconductor.org/packages/release/bioc/html/globaltest.html

[58] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov, “Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles,” Proceedings of the National Academy of Sciences of the United States of America, vol. 102, no. 43, pp. 15545-15550. doi: 10.1073/pnas.0506580102

[59] B. Efron, and R. Tibshirani. “On testing the significance of sets of genes,” The annals of applied statistics vol. 1, no. 1, pp. 107-129, 2007.

[60] R. A. Irizarry, D. Warren, F. Spencer, I. F. Kim, S. Biswal, B. C. Frank, E. Gabrielson, J. G. Garcia, J. Geoghegan, G. Germino, C. Griffin, S. C. Hilmer, E. Hoffman, A. E. Jedlicka1, E. Kawasaki, F. Martinez-Murillo, L. Morsberger, H. Lee, D. Petersen, J. Quackenbush, A. Scott, M. Wilson, Y. Yang, S. Q. Ye and W. Yu “Multiple-laboratory comparison of microarray platforms,” Nature Methods, vol.2, no. 5, pp. 345-350, 2005. doi:10.1038/nmeth756.

[61] L. Ein-Dor, I. Kela, G. Getz, D. Givol and E. Domany, “Outcome signature genes in breast cancer: is there a unique set?,” Bioinformatics, vol. 21, no. 2, pp. 171–178, 2005.

[62] D. Venet, J. E. Dumont and V. Detours, “Most random gene expression signatures are significantly associated with breast cancer outcome,” PLoS Computational Biology, vol. 7, 2011. doi:10.1371/journal.pcbi.1002240.

[63] M. Kuhn, K. Johnson, “Applied Predictive Modeling”, New York: Springer, 2013.

[64] T. Hothorn, F. Leisch, A. Zeileis, K. Hornik, “The Design and Analysis of Benchmark Experiments.” Journal of Computational and Graphical Statistics, 14(3), 675–699, 2005.

[65] M. Eugster, T. Hothorn, F. Leisch .“Exploratory and Inferential Analysis of Benchmark Experiments.” Ludwig Maximilians University Munich, Department of Statistics, Tech. Rep, 30, 2008.

[66] S. Calza, W. Raffelsberger, A. Ploner, J. Sahel, T. Leveillard, and Y. Pawitan, “Filtering genes to improve sensitivity in oligonucleotide microarray data analysis,” Nucleic Acids Res, vol. 35, no. 16, pp. E1022007, 2007.

[67] 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. Available: http://www.ncbi.nlm.nih.gov/geo/

[68] J. Wang, D. Duncan, Z. Shi, and B. Zhang, “WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013,” Nucleic Acids Res, vol. 41 (Web Server issue), pp. W77-83, 2013. Available: http://bioinfo.vanderbilt.edu/webgestalt/

[69] J. T. Chang, and JR Nevins, “GATHER: a systems approach to interpreting genomic signatures,” Bioinformatics, vol. 22, no. 23, pp. 2926-2933, 2006. Available: http://gather.genome.duke.edu/

[70] E. Mosca, R. Alfieri, I. Merelli, F. Viti, A. Calabria, and L. Milanesi, “A multilevel data integration resource for breast cancer study,” BMC Syst Biol, vol. 4, pp. 76, 2010. Available: http://www.itb.cnr.it/breastcancer/

[71] 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. Available: http://bdgene.psych.ac.cn/

[72] A. Gutiérrez-Sacristán, S. Grosdidier, O. Valverde, M. Torrens, À. Bravo, J. Piñero, F. Sanz, and L. I. Furlong, “PsyGeNET: a knowledge platform on psychiatric disorders and their genes,” Bioinformatics, 2015, pii: btv301, [Epub ahead of print]. Available: http://www.psygenet.org/web/PsyGeNET/menu;jsessionid=1dshdexwnnq84l0rpxmfc7rv6

[73] M. Kanehisa, and S. Goto, “KEGG: Kyoto Encyclopedia of Genes and Genomes,” Nucleic Acids Res, vol. 28, pp. 27-30, 2000. Available: http://www.genome.jp/kegg/kegg1.html

[74] D. L. Dankort, and W. J. Muller, “Signal transduction in mammary tumorigenesis: a transgenic perspective,” Oncogene, vol. 19, no. 8, pp. 1038-1044, 2000.

[75] J. Huan, L. Wang, L. Xing, X. Qin, L. Feng, X. Pan, and L. Zhu, “Insights into significant pathways and gene interaction networks underlying breast cancer cell line MCF-7 treated with 17β-estradiol (E2),” Gene, vol. 533, no. 1, pp. 346-355, 2014.

[76] E. I. Chen, J. Hewel, J. S. Krueger, C. Tiraby, M. R. Weber, A. Kralli, K. Becker, J. R. 3rd Yates, and B. Felding-Habermann, “Adaptation of energy metabolism in breast cancer brain metastases,” Cancer Res, vol. 67, no. 4, pp. 1472-1486, 2007.

[77] V. Tamási, K. Monostory, R. A. Prough, and A. Falus, “Role of xenobiotic metabolism in cancer: involvement of transcriptional and miRNA regulation of P450s,” Cell Mol Life Sci, vol. 68, no. 7, pp. 1131-1146, 2011.

[78] V. Ouellet, K. Tiedemann, A. Mourskaia, J. E. Fong, D. Tran-Thanh, E. Amir, M. Clemons, B. Perbal, S. V. Komarova, and P. M. Siegel, “CCN3 impairs osteoblast and stimulates osteoclast differentiation to favor breast cancer metastasis to bone,” Am J Pathol, vol. 178, no. 5, pp. 2377-2388, 2011.

[79] L. J. Standish, E. S. Sweet, J. Novack, C. A. Wenner, C. Bridge, A. Nelson, M. Martzen, and C. Torkelson, “Breast cancer and the immune system,” J Soc Integr Oncol, vol. 6, no. 4, 158-68, 2008.

[80] P. L. Fernández, P. Jares, M. J. Rey, E. Campo, and A. Cardesa, “Cell cycle regulators and their abnormalities in breast cancer,” Molecular Pathology, vol. 51, no. 6, pp. 305-309, 1998.

[81] R. M. Adibhatla, and J. F. Hatcher, “Altered lipid metabolism in brain injury and disorders,” Subcell Biochem, vol. 49, pp. 241-268, 2008.

[82] E. Uribe, and R. Wix, “Neuronal migration, apoptosis and bipolar disorder,” Rev Psiquiatr Salud Ment, vol. 5, no. 2, pp. 127-133, 2012.

[83] D. P. McKernan, U. Dennison, G. Gaszner, J. F. Cryan, and T. G. Dinan, “Enhanced peripheral toll-like receptor responses in psychosis: further evidence of a pro-inflammatory phenotype,” Translational Psychiatry, vol. 1, no. 8, pp. E36, 2011.

[84] S. S. Valvassori, K. V. Calixto, J. Budni, W. R. Resende, R. B. Varela, K. V. de Freitas, C. L. Gonçalves, E. L. Streck, and J. Quevedo, “Sodium butyrate reverses the inhibition of Krebs cycle enzymes induced by amphetamine in the rat brain,” J Neural Transm, vol. 120, no. 12, pp. 1737-1742, 2013.

[85] O. M. Dean, M. van den Buuse, A. I. Bush, D. L. Copolov, F. Ng, S. Dodd, and M. Berk, “A role for glutathione in the pathophysiology of bipolar disorder and schizophrenia? Animal models and relevance to clinical practice,” Curr Med Chem, vol. 16, no. 23, pp. 2965-2976, 2009.

[86] G. Shaltiel, G. Chen, and H. K. Manji, “Neurotrophic signaling cascades in the pathophysiology and treatment of bipolar disorder,” Curr Opin Pharmacol, vol. 7, no. 1, pp. 22-26, 2007.

SBV Source Code Availability

It is the author's intention to make the R source code of the proposed methodology for Stable Bootstrap Validation which was used in this thesis publicly available online at [23]. Along with a list of necessary packages (dependencies).

Appendix A

Classification Performance of All Candidate Genomic Signatures

In this part of the appendix, the classification performance of all gene signatures is displayed, along with the appropriate confidence intervals of accuracy, sensitivity and specificity. Some confidence intervals may slightly go beyond 100%, which is just a numeric artifact of the bootstrap t-confidence intervals in this case and could be avoided if bootstrap percentile conference intervals are used. The row corresponding to the selected gene signature is highlighted.

Dataset: GSE_Merged

RVM
genes 985.00 acc 0.871+-0.017 sens 0.931+-0.014 spec 0.625+-0.054
genes 985.00 acc 0.875+-0.016 sens 0.939+-0.013 spec 0.615+-0.054
genes 796.00 acc 0.868+-0.017 sens 0.926+-0.014 spec 0.628+-0.054
genes 645.00 acc 0.878+-0.016 sens 0.940+-0.013 spec 0.625+-0.054
genes 431.00 acc 0.871+-0.017 sens 0.930+-0.014 spec 0.628+-0.054
genes 289.00 acc 0.882+-0.016 sens 0.936+-0.013 spec 0.660+-0.053
genes 150.00 acc 0.900+-0.015 sens 0.941+-0.013 spec 0.731+-0.049
genes 71.00 acc 0.931+-0.012 sens 0.965+-0.010 spec 0.788+-0.045
genes 14.00 acc 0.906+-0.014 sens 0.962+-0.010 spec 0.676+-0.052
genes 70.00 acc 0.934+-0.012 sens 0.966+-0.010 spec 0.801+-0.044
Significant position 10: (70 genes)
acc low 0.922 acc_high 0.946
tpr low 0.956 tpr_high 0.976
tnr low 0.757 tnr_high 0.846

RVM OVERSAMPLING
genes 1004.00 acc 0.861+-0.017 sens 0.911+-0.016 spec 0.654+-0.053
genes 1004.00 acc 0.870+-0.017 sens 0.911+-0.016 spec 0.702+-0.051
genes 828.00 acc 0.863+-0.017 sens 0.911+-0.016 spec 0.663+-0.052
genes 602.00 acc 0.859+-0.017 sens 0.900+-0.016 spec 0.689+-0.051
genes 458.00 acc 0.883+-0.016 sens 0.929+-0.014 spec 0.696+-0.051
genes 314.00 acc 0.876+-0.016 sens 0.915+-0.015 spec 0.718+-0.050
genes 140.00 acc 0.904+-0.015 sens 0.931+-0.014 spec 0.792+-0.045
genes 71.00 acc 0.928+-0.013 sens 0.944+-0.013 spec 0.865+-0.038
genes 15.00 acc 0.872+-0.016 sens 0.877+-0.018 spec 0.853+-0.039
genes 75.00 acc 0.919+-0.013 sens 0.937+-0.013 spec 0.846+-0.040
Significant position 10: (75 genes)
acc low 0.906 acc_high 0.933
tpr low 0.924 tpr_high 0.951
tnr low 0.806 tnr_high 0.886

RVM UNDERSAMPLING
genes 1216.00 acc 0.759+-0.021 sens 0.756+-0.024 spec 0.772+-0.047
genes 969.00 acc 0.767+-0.021 sens 0.764+-0.023 spec 0.779+-0.046
genes 742.00 acc 0.777+-0.020 sens 0.776+-0.023 spec 0.779+-0.046
genes 593.00 acc 0.761+-0.021 sens 0.752+-0.024 spec 0.795+-0.045
genes 473.00 acc 0.795+-0.020 sens 0.791+-0.022 spec 0.814+-0.043
genes 289.00 acc 0.802+-0.020 sens 0.802+-0.022 spec 0.804+-0.044
genes 141.00 acc 0.832+-0.018 sens 0.831+-0.021 spec 0.837+-0.041
genes 72.00 acc 0.839+-0.018 sens 0.831+-0.021 spec 0.869+-0.037
genes 14.00 acc 0.842+-0.018 sens 0.836+-0.020 spec 0.865+-0.038
genes 49.00 acc 0.869+-0.017 sens 0.867+-0.019 spec 0.875+-0.037
Significant position 10: (49 genes)
acc low 0.852 acc_high 0.886
tpr low 0.849 tpr_high 0.886
tnr low 0.838 tnr_high 0.912

SVM
genes 997.00 acc 0.928+-0.013 sens 0.984+-0.007 spec 0.699+-0.051
genes 839.00 acc 0.933+-0.012 sens 0.984+-0.007 spec 0.724+-0.050
genes 712.00 acc 0.937+-0.012 sens 0.985+-0.007 spec 0.740+-0.049
genes 563.00 acc 0.937+-0.012 sens 0.986+-0.006 spec 0.737+-0.049
genes 425.00 acc 0.941+-0.012 sens 0.988+-0.006 spec 0.747+-0.048
genes 283.00 acc 0.948+-0.011 sens 0.993+-0.010 spec 0.766+-0.047
genes 141.00 acc 0.936+-0.012 sens 0.991+-0.005 spec 0.715+-0.050
genes 70.00 acc 0.896+-0.015 sens 0.993+-0.010 spec 0.500+-0.055
genes 19.00 acc 0.803+-0.020 sens 1.000+-0.000 spec 0.000+-0.000
genes 865.00 acc 0.926+-0.013 sens 0.980+-0.008 spec 0.705+-0.051
Significant position 7: (141 genes)
acc low 0.924 acc_high 0.948
tpr low 0.985 tpr_high 0.996
tnr low 0.665 tnr_high 0.765

SVM OVERSAMPLING
genes 990.00 acc 0.910+-0.014 sens 0.947+-0.012 spec 0.756+-0.048
genes 848.00 acc 0.919+-0.013 sens 0.954+-0.012 spec 0.779+-0.046
genes 723.00 acc 0.922+-0.013 sens 0.953+-0.012 spec 0.795+-0.045
genes 570.00 acc 0.932+-0.012 sens 0.965+-0.010 spec 0.795+-0.045
genes 423.00 acc 0.944+-0.011 sens 0.971+-0.009 spec 0.833+-0.041
genes 282.00 acc 0.933+-0.012 sens 0.962+-0.011 spec 0.817+-0.043
genes 140.00 acc 0.940+-0.012 sens 0.962+-0.010 spec 0.849+-0.040
genes 73.00 acc 0.934+-0.012 sens 0.950+-0.012 spec 0.869+-0.037
genes 19.00 acc 0.850+-0.018 sens 0.839+-0.020 spec 0.894+-0.034
genes 890.00 acc 0.918+-0.013 sens 0.954+-0.012 spec 0.772+-0.047
Significant position 7: (140 genes)
acc low 0.928 acc_high 0.952
tpr low 0.952 tpr_high 0.973
tnr low 0.810 tnr_high 0.889

SVM UNDERSAMPLING
genes 997.00 acc 0.860+-0.017 sens 0.870+-0.018 spec 0.821+-0.043
genes 839.00 acc 0.876+-0.016 sens 0.887+-0.017 spec 0.833+-0.041
genes 712.00 acc 0.880+-0.016 sens 0.892+-0.017 spec 0.833+-0.041
genes 563.00 acc 0.882+-0.016 sens 0.890+-0.017 spec 0.846+-0.040
genes 425.00 acc 0.902+-0.015 sens 0.915+-0.015 spec 0.853+-0.039
genes 283.00 acc 0.907+-0.014 sens 0.918+-0.015 spec 0.859+-0.039
genes 141.00 acc 0.892+-0.015 sens 0.900+-0.016 spec 0.862+-0.038
genes 70.00 acc 0.890+-0.015 sens 0.889+-0.017 spec 0.891+-0.035
genes 19.00 acc 0.756+-0.021 sens 0.720+-0.025 spec 0.904+-0.033
genes 865.00 acc 0.883+-0.016 sens 0.896+-0.017 spec 0.827+-0.042
Significant position 7: (141 genes)
acc low 0.877 acc_high 0.908
tpr low 0.883 tpr_high 0.916
tnr low 0.824 tnr_high 0.900

Dataset: GSE42568

RVM
genes 1593.00 acc 0.970+-0.018 sens 0.997+-0.011 spec 0.804+-0.234
genes 1593.00 acc 0.970+-0.018 sens 1.000+-0.000 spec 0.784+-0.113
genes 1593.00 acc 0.975+-0.037 sens 1.000+-0.000 spec 0.824+-0.223
genes 1593.00 acc 0.975+-0.032 sens 1.000+-0.000 spec 0.824+-0.235
genes 541.00 acc 0.972+-0.045 sens 0.997+-0.014 spec 0.824+-0.219
genes 338.00 acc 0.970+-0.018 sens 1.000+-0.000 spec 0.784+-0.113
genes 163.00 acc 0.956+-0.021 sens 0.990+-0.020 spec 0.745+-0.120
genes 80.00 acc 0.967+-0.018 sens 0.990+-0.023 spec 0.824+-0.228
genes 16.00 acc 0.989+-0.023 sens 1.000+-0.000 spec 0.922+-0.175
genes 53.00 acc 0.970+-0.018 sens 1.000+-0.000 spec 0.784+-0.113
Significant position 9: (16 genes)
acc low 0.966 acc_high 1.012
tpr low 1.000 tpr_high 1.000
tnr low 0.746 tnr_high 1.097

RVM OVERSAMPLING
genes 1593.00 acc 0.970+-0.018 sens 0.987+-0.035 spec 0.863+-0.208
genes 1593.00 acc 0.959+-0.020 sens 0.981+-0.029 spec 0.824+-0.234
genes 1593.00 acc 0.964+-0.019 sens 0.987+-0.029 spec 0.824+-0.262
genes 661.00 acc 0.964+-0.019 sens 0.984+-0.029 spec 0.843+-0.217
genes 531.00 acc 0.961+-0.020 sens 0.984+-0.038 spec 0.824+-0.223
genes 321.00 acc 0.964+-0.019 sens 0.987+-0.035 spec 0.824+-0.208
genes 162.00 acc 0.964+-0.019 sens 0.990+-0.028 spec 0.804+-0.244
genes 82.00 acc 0.981+-0.028 sens 0.997+-0.012 spec 0.882+-0.179
genes 17.00 acc 0.994+-0.014 sens 1.000+-0.000 spec 0.961+-0.119
genes 63.00 acc 0.981+-0.032 sens 0.997+-0.011 spec 0.882+-0.192
Significant position 9: (17 genes)
acc low 0.981 acc_high 1.008
tpr low 1.000 tpr_high 1.000
tnr low 0.842 tnr_high 1.080

RVM UNDERSAMPLING
genes 1593.00 acc 0.953+-0.022 sens 0.968+-0.020 spec 0.863+-0.199
genes 1593.00 acc 0.953+-0.022 sens 0.974+-0.047 spec 0.824+-0.219
genes 1593.00 acc 0.959+-0.020 sens 0.974+-0.040 spec 0.863+-0.185
genes 663.00 acc 0.964+-0.019 sens 0.984+-0.028 spec 0.843+-0.232
genes 492.00 acc 0.961+-0.020 sens 0.978+-0.047 spec 0.863+-0.200
genes 357.00 acc 0.948+-0.023 sens 0.962+-0.021 spec 0.863+-0.218
genes 172.00 acc 0.970+-0.018 sens 0.987+-0.026 spec 0.863+-0.210
genes 81.00 acc 0.950+-0.022 sens 0.965+-0.020 spec 0.863+-0.192
genes 16.00 acc 0.964+-0.019 sens 0.971+-0.049 spec 0.922+-0.140
genes 44.00 acc 0.967+-0.018 sens 0.978+-0.035 spec 0.902+-0.191
Significant position 9: (16 genes)
acc low 0.945 acc_high 0.983
tpr low 0.923 tpr_high 1.020
tnr low 0.781 tnr_high 1.062

SVM
genes 1592.00 acc 0.953+-0.022 sens 0.974+-0.037 spec 0.824+-0.225
genes 1592.00 acc 0.953+-0.022 sens 0.981+-0.033 spec 0.784+-0.113
genes 1592.00 acc 0.967+-0.018 sens 0.984+-0.029 spec 0.863+-0.209
genes 1592.00 acc 0.961+-0.020 sens 0.984+-0.029 spec 0.824+-0.246
genes 1592.00 acc 0.945+-0.023 sens 0.965+-0.020 spec 0.824+-0.224
genes 1592.00 acc 0.964+-0.019 sens 0.997+-0.011 spec 0.765+-0.116
genes 1592.00 acc 0.972+-0.036 sens 0.997+-0.016 spec 0.824+-0.226
genes 105.00 acc 0.964+-0.019 sens 0.987+-0.026 spec 0.824+-0.226
genes 16.00 acc 0.964+-0.019 sens 0.990+-0.022 spec 0.804+-0.242
genes 3.00 acc 0.986+-0.024 sens 0.994+-0.015 spec 0.941+-0.150
Significant position 9: (16 genes)
acc low 0.945 acc_high 0.983
tpr low 0.969 tpr_high 1.012
tnr low 0.561 tnr_high 1.046

SVM OVERSAMPLING
genes 1592.00 acc 0.967+-0.018 sens 0.994+-0.017 spec 0.804+-0.254
genes 1592.00 acc 0.961+-0.020 sens 0.994+-0.022 spec 0.765+-0.116
genes 1592.00 acc 0.961+-0.020 sens 0.990+-0.025 spec 0.784+-0.113
genes 1592.00 acc 0.961+-0.020 sens 0.984+-0.038 spec 0.824+-0.211
genes 1592.00 acc 0.972+-0.038 sens 0.994+-0.015 spec 0.843+-0.208
genes 1592.00 acc 0.967+-0.018 sens 0.987+-0.025 spec 0.843+-0.217
genes 1592.00 acc 0.959+-0.020 sens 0.981+-0.038 spec 0.824+-0.238
genes 96.00 acc 0.959+-0.020 sens 0.974+-0.037 spec 0.863+-0.213
genes 18.00 acc 0.978+-0.043 sens 0.984+-0.040 spec 0.941+-0.125
genes 3.00 acc 0.983+-0.026 sens 1.000+-0.000 spec 0.882+-0.203
Significant position 9: (18 genes)
acc low 0.935 acc_high 1.021
tpr low 0.944 tpr_high 1.024
tnr low 0.816 tnr_high 1.067

SVM UNDERSAMPLING
genes 1592.00 acc 0.920+-0.028 sens 0.923+-0.030 spec 0.902+-0.173
genes 1592.00 acc 0.890+-0.032 sens 0.894+-0.034 spec 0.863+-0.219
genes 1592.00 acc 0.923+-0.027 sens 0.929+-0.028 spec 0.882+-0.182
genes 1592.00 acc 0.920+-0.028 sens 0.933+-0.028 spec 0.843+-0.228
genes 1592.00 acc 0.898+-0.031 sens 0.904+-0.033 spec 0.863+-0.199
genes 1592.00 acc 0.879+-0.034 sens 0.881+-0.036 spec 0.863+-0.211
genes 1592.00 acc 0.887+-0.033 sens 0.910+-0.032 spec 0.745+-0.120
genes 105.00 acc 0.917+-0.028 sens 0.923+-0.030 spec 0.882+-0.189
genes 17.00 acc 0.901+-0.031 sens 0.910+-0.032 spec 0.843+-0.214
genes 3.00 acc 0.956+-0.021 sens 0.958+-0.022 spec 0.941+-0.140
Significant position 9: (17 genes)
acc low 0.870 acc_high 0.932
tpr low 0.879 tpr_high 0.942
tnr low 0.629 tnr_high 1.057

Dataset: GSE35974

RVM
genes 896.00 acc 0.884+-0.030 sens 0.926+-0.031 spec 0.807+-0.063
genes 748.00 acc 0.870+-0.032 sens 0.918+-0.032 spec 0.780+-0.066
genes 633.00 acc 0.896+-0.029 sens 0.922+-0.031 spec 0.847+-0.058
genes 499.00 acc 0.903+-0.028 sens 0.936+-0.029 spec 0.840+-0.059
genes 379.00 acc 0.921+-0.025 sens 0.954+-0.024 spec 0.860+-0.056
genes 250.00 acc 0.958+-0.019 sens 0.965+-0.022 spec 0.947+-0.074
genes 127.00 acc 0.979+-0.033 sens 0.989+-0.031 spec 0.960+-0.069
genes 67.00 acc 0.977+-0.031 sens 1.000+-0.000 spec 0.933+-0.131
genes 12.00 acc 0.785+-0.039 sens 0.894+-0.036 spec 0.580+-0.079
genes 329.00 acc 0.931+-0.024 sens 0.954+-0.024 spec 0.887+-0.051
Significant position 7: (127 genes)
acc low 0.946 acc_high 1.013
tpr low 0.959 tpr_high 1.020
tnr low 0.891 tnr_high 1.029

RVM OVERSAMPLING
genes 888.00 acc 0.894+-0.029 sens 0.933+-0.029 spec 0.820+-0.061
genes 756.00 acc 0.912+-0.027 sens 0.943+-0.027 spec 0.853+-0.057
genes 628.00 acc 0.907+-0.027 sens 0.936+-0.029 spec 0.853+-0.057
genes 506.00 acc 0.910+-0.027 sens 0.943+-0.027 spec 0.847+-0.058
genes 376.00 acc 0.931+-0.024 sens 0.957+-0.024 spec 0.880+-0.052
genes 252.00 acc 0.961+-0.018 sens 0.968+-0.045 spec 0.947+-0.088
genes 128.00 acc 0.977+-0.032 sens 0.975+-0.038 spec 0.980+-0.040
genes 68.00 acc 0.988+-0.019 sens 0.989+-0.036 spec 0.987+-0.039
genes 12.00 acc 0.778+-0.039 sens 0.784+-0.048 spec 0.767+-0.068
genes 329.00 acc 0.944+-0.022 sens 0.957+-0.024 spec 0.920+-0.043
Significant position 7: (128 genes)
acc low 0.945 acc_high 1.009
tpr low 0.937 tpr_high 1.013
tnr low 0.940 tnr_high 1.020

RVM UNDERSAMPLING
genes 901.00 acc 0.884+-0.030 sens 0.879+-0.038 spec 0.893+-0.049
genes 768.00 acc 0.882+-0.030 sens 0.869+-0.039 spec 0.907+-0.047
genes 637.00 acc 0.894+-0.029 sens 0.904+-0.034 spec 0.873+-0.053
genes 499.00 acc 0.912+-0.027 sens 0.904+-0.034 spec 0.927+-0.042
genes 376.00 acc 0.931+-0.024 sens 0.922+-0.031 spec 0.947+-0.099
genes 250.00 acc 0.954+-0.020 sens 0.957+-0.024 spec 0.947+-0.083
genes 127.00 acc 0.970+-0.016 sens 0.961+-0.023 spec 0.987+-0.033
genes 62.00 acc 0.977+-0.029 sens 0.975+-0.035 spec 0.980+-0.053
genes 12.00 acc 0.775+-0.039 sens 0.770+-0.049 spec 0.787+-0.066
genes 327.00 acc 0.944+-0.022 sens 0.950+-0.025 spec 0.933+-0.091
Significant position 7: (127 genes)
acc low 0.954 acc_high 0.986
tpr low 0.938 tpr_high 0.984
tnr low 0.954 tnr_high 1.019

SVM
genes 881.00 acc 0.891+-0.029 sens 0.943+-0.027 spec 0.793+-0.065
genes 751.00 acc 0.910+-0.027 sens 0.954+-0.024 spec 0.827+-0.061
genes 629.00 acc 0.898+-0.029 sens 0.947+-0.026 spec 0.807+-0.063
genes 505.00 acc 0.912+-0.027 sens 0.950+-0.025 spec 0.840+-0.059
genes 374.00 acc 0.903+-0.028 sens 0.950+-0.025 spec 0.813+-0.062
genes 249.00 acc 0.910+-0.027 sens 0.947+-0.026 spec 0.840+-0.059
genes 132.00 acc 0.931+-0.024 sens 0.961+-0.023 spec 0.873+-0.053
genes 63.00 acc 0.917+-0.026 sens 0.989+-0.023 spec 0.780+-0.066
genes 14.00 acc 0.653+-0.045 sens 1.000+-0.000 spec 0.000+-0.000
genes 407.00 acc 0.894+-0.029 sens 0.943+-0.027 spec 0.800+-0.064
Significant position 7: (132 genes)
acc low 0.907 acc_high 0.955
tpr low 0.938 tpr_high 0.984
tnr low 0.820 tnr_high 0.927

SVM OVERSAMPLING
genes 886.00 acc 0.910+-0.027 sens 0.936+-0.029 spec 0.860+-0.056
genes 753.00 acc 0.898+-0.029 sens 0.922+-0.031 spec 0.853+-0.057
genes 627.00 acc 0.891+-0.029 sens 0.918+-0.032 spec 0.840+-0.059
genes 501.00 acc 0.889+-0.030 sens 0.922+-0.031 spec 0.827+-0.061
genes 378.00 acc 0.889+-0.030 sens 0.901+-0.035 spec 0.867+-0.054
genes 249.00 acc 0.907+-0.027 sens 0.933+-0.029 spec 0.860+-0.056
genes 125.00 acc 0.912+-0.027 sens 0.933+-0.029 spec 0.873+-0.053
genes 62.00 acc 0.954+-0.020 sens 0.957+-0.024 spec 0.947+-0.115
genes 12.00 acc 0.826+-0.036 sens 0.862+-0.040 spec 0.760+-0.068
genes 386.00 acc 0.907+-0.027 sens 0.929+-0.030 spec 0.867+-0.054
Significant position 7: (125 genes)
acc low 0.885 acc_high 0.939
tpr low 0.903 tpr_high 0.962
tnr low 0.820 tnr_high 0.927

SVM UNDERSAMPLING
genes 882.00 acc 0.859+-0.033 sens 0.840+-0.043 spec 0.893+-0.049
genes 766.00 acc 0.882+-0.030 sens 0.876+-0.038 spec 0.893+-0.049
genes 634.00 acc 0.889+-0.030 sens 0.876+-0.038 spec 0.913+-0.045
genes 505.00 acc 0.880+-0.031 sens 0.872+-0.039 spec 0.893+-0.049
genes 378.00 acc 0.891+-0.029 sens 0.883+-0.038 spec 0.907+-0.047
genes 259.00 acc 0.907+-0.027 sens 0.897+-0.035 spec 0.927+-0.042
genes 128.00 acc 0.910+-0.027 sens 0.890+-0.037 spec 0.947+-0.067
genes 62.00 acc 0.940+-0.022 sens 0.922+-0.031 spec 0.973+-0.048
genes 12.00 acc 0.836+-0.035 sens 0.865+-0.040 spec 0.780+-0.066
genes 354.00 acc 0.868+-0.032 sens 0.855+-0.041 spec 0.893+-0.049
Significant position 7: (128 genes)
acc low 0.883 acc_high 0.937
tpr low 0.854 tpr_high 0.927
tnr low 0.879 tnr_high 1.014

Appendix B

Gene Lists of All Candidate Genomic Signatures

In this part of the appendix, all Candidate Genomic Signatures are displayed. In genomic signatures, each protein coding gene is described by its corresponding Entrez Gene ID, Gene Symbol and the Official Full Name. For probes that do not code for proteins (e.g. non coding RNAs), Ensembl ID is provided; for other unmapped probes, Affymetrix transcript-cluster-ID is provided. Briefly, Ensembl provide annotation on genome assemblies that have been deposited into the International Nucleotide Sequence Database Collaboration (INSDC).

Gene-Disease Associations are highlighted. Genes that are associated with breast cancer are colored with red, whereas genes that are associated with bipolar disorder are colored with purple.

'70 RVM Signature' - BREAST CANCER - GSE_Merged (1/2)		
Gene ID	**Gene Symbol**	**Description**
943	TNFRSF8	tumor necrosis factor receptor superfamily, member 8
10150	MBNL2	muscleblind-like 2 (Drosophila)
4948	OCA2	oculocutaneous albinism II
1300	COL10A1	collagen, type X, alpha 1
57604	C8orf79	chromosome 8 open reading frame 79
1975	EIF4B	eukaryotic translation initiation factor 4B
7107	GPR137B	G protein-coupled receptor 137B
3500	IGHG1	immunoglobulin heavy constant gamma 1 (G1m marker)
22800	RRAS2	related RAS viral (r-ras) oncogene homolog 2
259266	ASPM	asp (abnormal spindle) homolog, microcephaly associated (Drosophila)
166647	GPR125	G protein-coupled receptor 125
55013	CCDC109B	coiled-coil domain containing 109B
2940	GSTA3	glutathione S-transferase alpha 3
55713	ZNF334	zinc finger protein 334
4800	NFYA	nuclear transcription factor Y, alpha
84914	ZNF587	zinc finger protein 587
8698	S1PR4	sphingosine-1-phosphate receptor 4
890	CCNA2	cyclin A2
1047	CLGN	calmegin
2259	FGF14	fibroblast growth factor 14
846	CASR	calcium-sensing receptor
6363	CCL19	chemokine (C-C motif) ligand 19
811	CALR	calreticulin
9184	BUB3	budding uninhibited by benzimidazoles 3 homolog (yeast)
9452	ITM2A	integral membrane protein 2A
4680	CEACAM6	carcinoembryonic antigen-related cell adhesion molecule 6 (non-specific cross reacting antigen)
5321	PLA2G4A	phospholipase A2, group IVA (cytosolic, calcium-dependent)
9244	CRLF1	cytokine receptor-like factor 1
23012	STK38L	serine/threonine kinase 38 like
4303	FOXO4	forkhead box O4
140885	SIRPA	signal-regulatory protein alpha
1409	CRYAA	crystallin, alpha A
341	APOC1	apolipoprotein C-I
1440	CSF3	colony stimulating factor 3 (granulocyte)
1950	EGF	epidermal growth factor (beta-urogastrone)
220	ALDH1A3	aldehyde dehydrogenase 1 family, member A3
6932	TCF7	transcription factor 7 (T-cell specific, HMG-box)
54988	ACSM5	acyl-CoA synthetase medium-chain family member 5

'70 RVM Signature' - BREAST CANCER - GSE_Merged (2/2)		
Gene ID	**Gene Symbol**	**Description**
6767	ST13	suppression of tumorigenicity 13 (colon carcinoma) (Hsp70 interacting protein)
22829	NLGN4Y	neuroligin 4, Y-linked
57144	PAK7	p21 protein (Cdc42/Rac)-activated kinase 7
7351	UCP2	uncoupling protein 2 (mitochondrial, proton carrier)
154796	AMOT	angiomotin
26278	SACS	spastic ataxia of Charlevoix-Saguenay (sacsin)
5724	PTAFR	platelet-activating factor receptor
80774	LIMD2	LIM domain containing 2
9021	SOCS3	suppressor of cytokine signaling 3
3667	IRS1	insulin receptor substrate 1
5308	PITX2	paired-like homeodomain 2
898	CCNE1	cyclin E1
9915	ARNT2	aryl-hydrocarbon receptor nuclear translocator 2
7067	THRA	thyroid hormone receptor, alpha (erythroblastic leukemia viral (v-erb-a) oncogene homolog, avian)
7980	TFPI2	tissue factor pathway inhibitor 2
9435	CHST2	carbohydrate (N-acetylglucosamine-6-O) sulfotransferase 2
55765	C1orf106	chromosome 1 open reading frame 106
7136	TNNI2	troponin I type 2 (skeletal, fast)
9510	ADAMTS1	ADAM metallopeptidase with thrombospondin type 1 motif, 1
6362	CCL18	chemokine (C-C motif) ligand 18 (pulmonary and activation-regulated)
9123	SLC16A3	solute carrier family 16, member 3 (monocarboxylic acid transporter 4)
1359	CPA3	carboxypeptidase A3 (mast cell)
2527	FUT5	fucosyltransferase 5 (alpha (1,3) fucosyltransferase)
3833	KIFC1	kinesin family member C1
7080	NKX2-1	NK2 homeobox 1
7545	ZIC1	Zic family member 1 (odd-paired homolog, Drosophila)
79632	FAM184A	family with sequence similarity 184, member A
1288	COL4A6	collagen, type IV, alpha 6
54829	ASPN	asporin
2115	ETV1	ets variant 1
2146	EZH2	enhancer of zeste homolog 2 (Drosophila)
8537	BCAS1	breast carcinoma amplified sequence 1

GENE SYMBOL: associated with breast cancer

GENE SYMBOL: not associated yet with breast cancer

'141 RFE-SVM Signature' - Breast Cancer - GSE_Merged (1/4)		
Gene ID	**Gene Symbol**	**Description**
4948	OCA2	oculocutaneous albinism II
6363	CCL19	chemokine (C-C motif) ligand 19
4680	CEACAM6	carcinoembryonic antigen-related cell adhesion molecule 6 (non-specific cross reacting antigen)
10150	MBNL2	muscleblind-like 2 (Drosophila)
3204	HOXA7	homeobox A7
3500	IGHG1	immunoglobulin heavy constant gamma 1 (G1m marker)
1300	COL10A1	collagen, type X, alpha 1
7980	TFPI2	tissue factor pathway inhibitor 2
9184	BUB3	budding uninhibited by benzimidazoles 3 homolog (yeast)
943	TNFRSF8	tumor necrosis factor receptor superfamily, member 8
9915	ARNT2	aryl-hydrocarbon receptor nuclear translocator 2
2069	EREG	epiregulin
3698	ITIH2	inter-alpha (globulin) inhibitor H2
2676	GFRA3	GDNF family receptor alpha 3
57604	C8orf79	chromosome 8 open reading frame 79
6278	S100A7	S100 calcium binding protein A7
6767	ST13	suppression of tumorigenicity 13 (colon carcinoma) (Hsp70 interacting protein)
84914	ZNF587	zinc finger protein 587
9021	SOCS3	suppressor of cytokine signaling 3
1950	EGF	epidermal growth factor (beta-urogastrone)
220	ALDH1A3	aldehyde dehydrogenase 1 family, member A3
347902	AMIGO2	adhesion molecule with Ig-like domain 2
7107	GPR137B	G protein-coupled receptor 137B
2261	FGFR3	fibroblast growth factor receptor 3
2527	FUT5	fucosyltransferase 5 (alpha (1,3) fucosyltransferase)
4321	MMP12	matrix metallopeptidase 12 (macrophage elastase)
7080	NKX2-1	NK2 homeobox 1
1277	COL1A1	collagen, type I, alpha 1
1288	COL4A6	collagen, type IV, alpha 6
57144	PAK7	p21 protein (Cdc42/Rac)-activated kinase 7
6932	TCF7	transcription factor 7 (T-cell specific, HMG-box)
7545	ZIC1	Zic family member 1 (odd-paired homolog, Drosophila)
1118	CHIT1	chitinase 1 (chitotriosidase)
1592	CYP26A1	cytochrome P450, family 26, subfamily A, polypeptide 1
4232	MEST	mesoderm specific transcript homolog (mouse)

'141 RFE-SVM Signature' - Breast Cancer - GSE_Merged (2/4)		
Gene ID	**Gene Symbol**	**Description**
4800	NFYA	nuclear transcription factor Y, alpha
6696	SPP1	secreted phosphoprotein 1
79645	EFCAB1	EF-hand calcium binding domain 1
811	CALR	calreticulin
23532	PRAME	preferentially expressed antigen in melanoma
2940	GSTA3	glutathione S-transferase alpha 3
4129	MAOB	monoamine oxidase B
4303	FOXO4	forkhead box O4
9	NAT1	N-acetyltransferase 1 (arylamine N-acetyltransferase)
10381	TUBB3	tubulin, beta 3
341	APOC1	apolipoprotein C-I
3667	IRS1	insulin receptor substrate 1
5308	PITX2	paired-like homeodomain 2
6362	CCL18	chemokine (C-C motif) ligand 18 (pulmonary and activation-regulated)
6398	SECTM1	secreted and transmembrane 1
1446	CSN1S1	casein alpha s1
22979	EFR3B	EFR3 homolog B (S. cerevisiae)
50617	ATP6V0A4	ATPase, H+ transporting, lysosomal V0 subunit a4
5744	PTHLH	parathyroid hormone-like hormone
9435	CHST2	carbohydrate (N-acetylglucosamine-6-O) sulfotransferase 2
22800	RRAS2	related RAS viral (r-ras) oncogene homolog 2
25840	METTL7A	methyltransferase like 7A
5650	KLK7	kallikrein-related peptidase 7
1278	COL1A2	collagen, type I, alpha 2
1301	COL11A1	collagen, type XI, alpha 1
3084	NRG1	neuregulin 1
4477	MSMB	microseminoprotein, beta-
63982	ANO3	anoctamin 3
11057	ABHD2	abhydrolase domain containing 2
1975	EIF4B	eukaryotic translation initiation factor 4B
3249	HPN	hepsin
3665	IRF7	interferon regulatory factor 7
54972	TMEM132A	transmembrane protein 132A
7366	UGT2B15	UDP glucuronosyltransferase 2 family, polypeptide B15
9185	REPS2	RALBP1 associated Eps domain containing 2
171586	ABHD3	abhydrolase domain containing 3
1907	EDN2	endothelin 2
4916	NTRK3	neurotrophic tyrosine kinase, receptor, type 3

'141 RFE-SVM Signature' - Breast Cancer - GSE_Merged (3/4)		
Gene ID	**Gene Symbol**	**Description**
5008	OSM	oncostatin M
10	NAT2	N-acetyltransferase 2 (arylamine N-acetyltransferase)
26472	PPP1R14B	protein phosphatase 1, regulatory (inhibitor) subunit 14B
2920	CXCL2	chemokine (C-X-C motif) ligand 2
2938	GSTA1	glutathione S-transferase alpha 1
79919	C2orf54	chromosome 2 open reading frame 54
9636	ISG15	ISG15 ubiquitin-like modifier
1047	CLGN	calmegin
140885	SIRPA	signal-regulatory protein alpha
4778	NFE2	nuclear factor (erythroid-derived 2), 45kDa
5321	PLA2G4A	phospholipase A2, group IVA (cytosolic, calcium-dependent)
55013	CCDC109B	coiled-coil domain containing 109B
9452	ITM2A	integral membrane protein 2A
2115	ETV1	ets variant 1
225	ABCD2	ATP-binding cassette, sub-family D (ALD), member 2
23012	STK38L	serine/threonine kinase 38 like
259266	ASPM	asp (abnormal spindle) homolog, microcephaly associated (Drosophila)
4322	MMP13	matrix metallopeptidase 13 (collagenase 3)
55713	ZNF334	zinc finger protein 334
5608	MAP2K6	mitogen-activated protein kinase kinase 6
6445	SGCG	sarcoglycan, gamma (35kDa dystrophin-associated glycoprotein)
1289	COL5A1	collagen, type V, alpha 1
1359	CPA3	carboxypeptidase A3 (mast cell)
56475	RPRM	reprimo, TP53 dependent G2 arrest mediator candidate
57214	KIAA1199	KIAA1199
10312	TCIRG1	T-cell, immune regulator 1, ATPase, H+ transporting, lysosomal V0 subunit A3
1290	COL5A2	collagen, type V, alpha 2
22829	NLGN4Y	neuroligin 4, Y-linked
29106	SCG3	secretogranin III
64221	ROBO3	roundabout, axon guidance receptor, homolog 3 (Drosophila)
898	CCNE1	cyclin E1
9244	CRLF1	cytokine receptor-like factor 1
11178	LZTS1	leucine zipper, putative tumor suppressor 1
220988	HNRNPA3	heterogeneous nuclear ribonucleoprotein A3
55065	GPR172B	G protein-coupled receptor 172B
8698	S1PR4	sphingosine-1-phosphate receptor 4
1908	EDN3	endothelin 3

'141 RFE-SVM Signature' - Breast Cancer - GSE_Merged (4/4)		
Gene ID	**Gene Symbol**	**Description**
3559	IL2RA	interleukin 2 receptor, alpha
4586	MUC5AC	mucin 5AC, oligomeric mucus/gel-forming
8817	FGF18	fibroblast growth factor 18
1016	CDH18	cadherin 18, type 2
154796	AMOT	angiomotin
2001	ELF5	E74-like factor 5 (ets domain transcription factor)
5307	PITX1	paired-like homeodomain 1
10964	IFI44L	interferon-induced protein 44-like
23057	NMNAT2	nicotinamide nucleotide adenylyltransferase 2
24137	KIF4A	kinesin family member 4A
24141	C20orf103	chromosome 20 open reading frame 103
2897	GRIK1	glutamate receptor, ionotropic, kainate 1
3200	HOXA3	homeobox A3
55286	C4orf19	chromosome 4 open reading frame 19
7067	THRA	thyroid hormone receptor, alpha (erythroblastic leukemia viral (v-erb-a) oncogene homolog, avian)
7351	UCP2	uncoupling protein 2 (mitochondrial, proton carrier)
1230	CCR1	chemokine (C-C motif) receptor 1
166647	GPR125	G protein-coupled receptor 125
3851	KRT4	keratin 4
3852	KRT5	keratin 5
3861	KRT14	keratin 14
5050	PAFAH1B3	platelet-activating factor acetylhydrolase, isoform Ib, subunit 3 (29kDa)
51466	EVL	Enah/Vasp-like
7136	TNNI2	troponin I type 2 (skeletal, fast)
79632	FAM184A	family with sequence similarity 184, member A
2191	FAP	fibroblast activation protein, alpha
2327	FMO2	flavin containing monooxygenase 2 (non-functional)
4314	MMP3	matrix metallopeptidase 3 (stromelysin 1, progelatinase)
514	ATP5E	ATP synthase, H+ transporting, mitochondrial F1 complex, epsilon subunit
7083	TK1	thymidine kinase 1, soluble
79948	PRG2	plasticity-related gene 2

GENE SYMBOL: associated with breast cancer

GENE SYMBOL: not associated yet with breast cancer

'16 RVM Signature' - Breast Cancer - GSE42568		
Gene ID	**Gene Symbol**	**Description**
286133	SCARA5	scavenger receptor class A, member 5 (putative)
1446	CSN1S1	casein alpha s1
1215	CMA1	chymase 1, mast cell
4582	MUC1	mucin 1, cell surface associated
10990	LILRB5	leukocyte immunoglobulin-like receptor, subfamily B (with TM and ITIM domains), member 5
683	BST1	bone marrow stromal cell antigen 1
114800	CCDC85A	coiled-coil domain containing 85A
10053	AP1M2	adaptor-related protein complex 1, mu 2 subunit
257194	NEGR1	neuronal growth regulator 1
6382	SDC1	syndecan 1
7153	TOP2A	topoisomerase (DNA) II alpha 170kDa
83857	TMTC1	transmembrane and tetratricopeptide repeat containing 1
91584	PLXNA4	plexin A4
Unmapped IDs: 234057_at uncharacterized gastric protein ZG33P (not available gene symbol); 1559401_a_at (BI052176); 1557383_a_at (AI925316)		

'16 RFE-SVM Signature' - Breast Cancer - GSE42568		
Gene ID	**Gene Symbol**	**Description**
2167	FABP4	fatty acid binding protein 4, adipocyte
1311	COMP	cartilage oligomeric matrix protein
991	CDC20	cell division cycle 20 homolog (S. cerevisiae)
6659	SOX4	SRY (sex determining region Y)-box 4
3204	HOXA7	homeobox A7
2146	EZH2	enhancer of zeste homolog 2 (Drosophila)
1301	COL11A1	collagen, type XI, alpha 1
84419	C15orf48	chromosome 15 open reading frame 48
84820	POLR2J4	polymerase (RNA) II (DNA directed) polypeptide J4, pseudogene
84163	GTF2IRD2	GTF2I repeat domain containing 2
721	C4B	complement component 4B (Chido blood group)
1300	COL10A1	collagen, type X, alpha 1
Unmapped IDs: 234032_at (AF119847); 235803_at (AA843122); 243329_at (AI074450); 1562235_s_at (AL832146)		

GENE SYMBOL: associated with breast cancer

GENE SYMBOL: not associated yet with breast cancer

'127 RVM Signature' - Bipolar Disorder - GSE35974 (1/3)		
(A) 66 Protein Coding Genes		
Gene ID	**Gene Symbol**	**Description**
84443	FRMPD3	FERM and PDZ domain containing 3
84079	ANKRD27	ankyrin repeat domain 27 (VPS9 domain)
8728	ADAM19	ADAM metallopeptidase domain 19
7189	TRAF6	TNF receptor-associated factor 6, E3 ubiquitin protein ligase
283398	SUCLG2P2	succinate-CoA ligase, GDP-forming, beta subunit pseudogene 2
83850	ESYT3	extended synaptotagmin-like protein 3
1645	AKR1C1	aldo-keto reductase family 1, member C1 (dihydrodiol dehydrogenase 1; 20-alpha (3-alpha)-hydroxysteroid dehydrogenase)
1389	CREBL2	cAMP responsive element binding protein-like 2
25942	SIN3A	SIN3 transcription regulator homolog A (yeast)
5379	PMS2P1	postmeiotic segregation increased 2 pseudogene 1
359845	FAM101B	family with sequence similarity 101, member B
54491	FAM105A	family with sequence similarity 105, member A
100873283	RNA5SP44	RNA, 5S ribosomal pseudogene 44
100873308	RNA5SP74	RNA, 5S ribosomal pseudogene 74
10268	RAMP3	receptor (G protein-coupled) activity modifying protein 3
1646	AKR1C2	aldo-keto reductase family 1, member C2 (dihydrodiol dehydrogenase 2; bile acid binding protein; 3-alpha hydroxysteroid dehydrogenase, type III)
51676	ASB2	ankyrin repeat and SOCS box containing 2
116369	SLC26A8	solute carrier family 26, member 8
134829	CLVS2	clavesin 2
54915	YTHDF1	YTH domain family, member 1
26844	RNU3P2	RNA, U3 small nucleolar pseudogene 2
65110	UPF3A	UPF3 regulator of nonsense transcripts homolog A (yeast)
5984	RFC4	replication factor C (activator 1) 4, 37kDa
132671	SPATA18	spermatogenesis associated 18
7021	**TFAP2B**	**transcription factor AP-2 beta (activating enhancer binding protein 2 beta)**
7428	VHL	von Hippel-Lindau tumor suppressor, E3 ubiquitin protein ligase
610	HCN2	hyperpolarization activated cyclic nucleotide-gated potassium channel 2
406907	MIR124-1	microRNA 124-1
91120	ZNF682	zinc finger protein 682
146760	RTN4RL1	reticulon 4 receptor-like 1

'127 RVM Signature' - Bipolar Disorder - GSE 35974 (2/3)		
(A) 66 Protein Coding Genes		
Gene ID	**Gene Symbol**	**Description**
401491	FLJ35024	uncharacterized LOC401491
5066	PAM	peptidylglycine alpha-amidating monooxygenase
57828	CATSPERG	catsper channel auxiliary subunit gamma
1781	DYNC1I2	dynein, cytoplasmic 1, intermediate chain 2
84959	UBASH3B	ubiquitin associated and SH3 domain containing B
200132	TCTEX1D1	Tctex1 domain containing 1
3152	HMGN2P11	high mobility group nucleosomal binding domain 2 pseudogene 11
574448	MIR202	microRNA 202
100873558	RNA5SP507	RNA, 5S ribosomal pseudogene 507
168374	ZNF92	zinc finger protein 92
402117	VWC2L	von Willebrand factor C domain containing protein 2-like
57654	UVSSA	UV-stimulated scaffold protein A
79780	CCDC82	coiled-coil domain containing 82
10455	ECI2	enoyl-CoA delta isomerase 2
6422	SFRP1	secreted frizzled-related protein 1
285464	CRIPAK	cysteine-rich PAK1 inhibitor
79674	**VEPH1**	**ventricular zone expressed PH domain homolog 1 (zebrafish)**
118980	SFXN2	sideroflexin 2
10201	NME6	NME/NM23 nucleoside diphosphate kinase 6
84767	TRIM51	tripartite motif-containing 51
84125	LRRIQ1	leucine-rich repeats and IQ motif containing 1
65985	AACS	acetoacetyl-CoA synthetase
5295	PIK3R1	phosphoinositide-3-kinase, regulatory subunit 1 (alpha)
340286	FAM183B	acyloxyacyl hydrolase (neutrophil)
3437	IFIT3	interferon-induced protein with tetratricopeptide repeats 3
100302652	GPR75-ASB3	GPR75-ASB3 readthrough
646870	LOC646870	centrosomal protein 57kDa pseudogene
10677	AVIL	advillin
651	BMP3	bone morphogenetic protein 3
392391	OR5C1	olfactory receptor, family 5, subfamily C, member 1
23473	CAPN7	calpain 7
51444	RNF138	ring finger protein 138, E3 ubiquitin protein ligase

'127 RVM Signature' - Bipolar Disorder - GSE 35974 (3/3)		
(A) 66 Protein Coding Genes		
Gene ID	**Gene Symbol**	**Description**
152667	FAM192BP	family with sequence similarity 192, member B pseudogene
285761	DCBLD1	discoidin, CUB and LCCL domain containing 1
441061	MARCH11	membrane-associated ring finger (C3HC4) 11

Gene ID 65110 represented twice (different probes)

In bold: Gene ID-Gene Symbol-Description: associated with bipolar disorder (source: BDgene)

Underlined: Gene ID-Gene Symbol-Description: associated with bipolar disorder (source: PsyGeNET)

In bold and Underlined: Gene ID-Gene Symbol-Description: associated with bipolar disorder (source: BDgene and PsyGeNET)

'127 RVM Signature' - Bipolar Disorder - GSE35974

(B) 61 Unmapped IDs

pos_control (7896112; 7893125; 7893918; **7895038**)

neg_control (7896264; 7893746; 7892741; 7893809; 7896681; 7892798; 7895338; 7894407; 7894891; 7894864; 7895898; 7895887; **7894020; 7893071**)

ncrna_pseudogene:scRNA

(**8017829/RN7SL756P/ENSG00000244610; 8106250/RN7SL814P/ENSG00000244326**)

ncrna:snRNA

(8114579/RNU4-14P/ENSG00000222790; 8016429/ENSG00000207306; 8097062/ENSG00000223225; 8008596/ENSG00000200107; 7993179/ENSG00000199482; 7916743/ENSG00000207190; 8036589/ENSG00000207296; 8078832/ENSG00000206708; 7948306/ENSG00000200817; 8128888/GenBank:AL357515.26/ENSG00000207431; 8156759/ENSG00000212521;
8173823/ENSG00000206826; 8094355/ENSG00000239001; 8112894/ENSG00000206774; 7993112/ENSG00000200869; 8088846/ENSG00000252937; 8007501/ENSG00000252729; 8045846/ENSG00000251980; 7911108/ENSG00000252282; 8089255/ENSG00000201065; **8124162/ENSG00000200957; 8163147/ENSG00000200106; 8003226/ENSG00000252311; 8132692/ENSG00000202350; 7940580/ENSG00000200898; 7944525/ENSG00000199217; 8062693/ENSG00000201021; 7988344/AC091245-AC090888/ENSG00000252117)**

ncrna:snoRNA

(7899988/GenBank: AL160000.15/ENSG00000201542; 8010766/ENSG00000238947)

ncrna:misc_RNA

(8076461/RN7SKP80/ENSG00000202058; 7974253/RN7SKP193/ENSG00000201358;
8124038/ENSG00000207193; **8052370/RN7SKP208/ENSG00000202344**)

Miscellaneous:

ncrna_pseudogene:scRNA (**8031867/ENSG00000240512/not found**)

ncrna_pseudogene:tRNA (**8062962/ENSG00000244543/retired**)

cDNA clone (**8105189/cDNA clone MGC:168815/cDNA clone**)

cdna (**7995580/cdna: Genscan chromosome**)

8066292/Unknown; 7992742/Unknown; 8180246/withdrawn

Abbreviations:

pos_control, normgene->exon, exronic control=probe sets against exon regions of a set of housekeeping genes**; neg_control,** normgene->intron, intronic control=probe sets against intron regions of a set of housekeeping genes**; ncRNA,** non coding RNA**; ncRNA types: (i) tRNA,** transfer RNA**, (ii) scRNA,** small cytoplasmic RNA**, (iii) snRNA,** small nuclear RNA**, (iv) snoRNA,** small nucleolar RNA**, and (v) misc_RNA,** miscellaneous other RNA**; cDNA,** complementary DNA**; Genscan,** a gene finding algorithm. Unique IDs in '127 RVM Signature' are given in "bold" font.

‘132 RFE-SVM Signature’ - Bipolar Disorder - GSE35974 (1/3)		
(A) 48 Protein Coding Genes		
Gene ID	**Gene Symbol**	**Description**
51816	CECR1	cat eye syndrome chromosome region, candidate 1
8728	ADAM19	ADAM metallopeptidase domain 19
7189	TRAF6	TNF receptor-associated factor 6, E3 ubiquitin protein ligase
283398	SUCLG2P2	succinate-CoA ligase, GDP-forming, beta subunit pseudogene 2
8516	ITGA8	integrin, alpha 8
2119	ETV5	ets variant 5
5142	**PDE4B**	**phosphodiesterase 4B, cAMP-specific**
83850	ESYT3	extended synaptotagmin-like protein 3
1645	AKR1C1	aldo-keto reductase family 1, member C1 (dihydrodiol dehydrogenase 1; 20-alpha (3-alpha)-hydroxysteroid dehydrogenase)
9510	ADAMTS1	ADAM metallopeptidase with thrombospondin type 1 motif, 1
5379	PMS2P1	postmeiotic segregation increased 2 pseudogene 1
7783	ZP2	zona pellucida glycoprotein 2 (sperm receptor)
5996	RGS1	regulator of G-protein signaling 1
9734	HDAC9	histone deacetylase 9
168667	BMPER	BMP binding endothelial regulator
100873308	RNA5SP74	RNA, 5S ribosomal pseudogene 74
402117	VWC2L	von Willebrand factor C domain containing protein 2-like
1646	AKR1C2	aldo-keto reductase family 1, member C2 (dihydrodiol dehydrogenase 2; bile acid binding protein; 3-alpha hydroxysteroid dehydrogenase, type III)
6355	CCL8	chemokine (C-C motif) ligand 8
2719	GPC3	glypican 3
5105	PCK1	phosphoenolpyruvate carboxykinase 1 (soluble)
1268	**CNR1**	**cannabinoid receptor 1 (brain)**
57654	UVSSA	UV-stimulated scaffold protein A
79780	CCDC82	coiled-coil domain containing 82
6422	SFRP1	secreted frizzled-related protein 1
116369	SLC26A8	solute carrier family 26, member 8
2944	**GSTM1**	**glutathione S-transferase mu 1**
285464	CRIPAK	cysteine-rich PAK1 inhibitor

‘132 RFE-SVM Signature’ - Bipolar Disorder - GSE35974 (2/3)		
(A) 48 Protein Coding Genes		
Gene ID	**Gene Symbol**	**Description**
5984	RFC4	replication factor C (activator 1) 4, 37kDa
79674	**VEPH1**	**ventricular zone expressed PH domain homolog 1 (zebrafish)**
118980	SFXN2	sideroflexin 2
7704	ZBTB16	zinc finger and BTB domain containing 16
3883	KRT33A	keratin 33A
54885	TBC1D8B	TBC1 domain family, member 8B (with GRAM domain)
84125	LRRIQ1	leucine-rich repeats and IQ motif containing 1
84767	TRIM51	tripartite motif-containing 51
7021	**TFAP2B**	**transcription factor AP-2 beta (activating enhancer binding protein 2 beta)**
7428	VHL	von Hippel-Lindau tumor suppressor, E3 ubiquitin protein ligase
340286	FAM183B	acyloxyacyl hydrolase (neutrophil)
406907	MIR124-1	microRNA 124-1
4986	OPRK1	opioid receptor, kappa 1
7857	SCG2	secretogranin II
5066	PAM	peptidylglycine alpha-amidating monooxygenase
123036	TC2N	tandem C2 domains, nuclear
57828	CATSPERG	catsper channel auxiliary subunit gamma
619383	SCARNA9	small Cajal body-specific RNA 9
441061	MARCH11	membrane-associated ring finger (C3HC4) 11
285761	DCBLD1	discoidin, CUB and LCCL domain containing 1

In bold: Gene ID-Gene Symbol-Description: associated with bipolar disorder (source: BDgene)
Underlined: Gene ID-Gene Symbol-Description: associated with bipolar disorder (source: PsyGeNET)
In bold and Underlined: Gene ID-Gene Symbol-Description: associated with bipolar disorder (source: BDgene and PsyGeNET)

‘132 RFE-SVM Signature’ - Bipolar Disorder - GSE35974 (3/3)

(B) 84 Unmapped IDs

pos_control (7896112; 7893125; 7893918; **7895620; 7895186; 7896558)**

neg_control (7896264; 7893746; 7892741; 7893809; 7896681; 7892798; 7895338; 7894407; 7894891; 7894864; 7895898; 7895887; **7894362; 7893193; 7896630; 7894287; 7893358; 7894731; 7894738; 7896507; 7894651; 7892510; 7896355; 7894507; 7895980; 7896316; 7892508; 7893645; 7896357; 7894142; 7896389; 7892962; 7894678; 7894550; 7896226**)

ncrna:snRNA
(8114579/RNU4-14P/ENSG00000222790; 8016429/ENSG00000207306; 8097062/ENSG00000223225; 8008596/ENSG00000200107; 7993179/ENSG00000199482; 7916743/ENSG00000207190; 8036589/ENSG00000207296; 8078832/ENSG00000206708; 7948306/ENSG00000200817; 8128888/GenBank: AL357515.26/ENSG00000207431; 8156759/ENSG00000212521; 8173823/ENSG00000206826; 8094355/ENSG00000239001; 8112894/ENSG00000206774; 7993112/ENSG00000200869; 8088846/ENSG00000252937; 8007501/ENSG00000252729; 8045846/ENSG00000251980; 7911108/ENSG00000252282; 8089255/ENSG00000201065; **8042460/ENSG00000199460; 8143108/ENSG00000212303; 8016412/ENSG00000206954; 7944716/ENSG00000252556; 8104117/ENSG00000202215; 7983270/ENSG00000201136; 8112801/ENSG00000200331**)

ncrna:snoRNA
(7899988/GenBank: AL160000.15/ENSG00000201542; 8010766/ENSG00000238947; **8142448/ENSG00000202377; 7954690/ENSG00000238661**)

ncrna:misc_RNA
(8076461/RN7SKP80/ENSG00000202058; 7974253/RN7SKP193/ENSG00000201358; 8124038/ENSG00000207193; **7983189/AC068724 AC009852/ENSG00000202211**)

Miscellaneous:
7992742/Unknown; 8180246/withdrawn; **8134429/AK096576/not found; 8088996/ncrna_pseudogene:scRNA/ENSG00000242783/retired; 8150163/ncrna_pseudogene:Mt_tRNA/ENSG00000243610/retired; 8136658/ncrna_pseudogene:Mt_tRNA/ENSG00000240028; 7927031/ncrna:snoRNA/ENSG00000212148; 7896704/other spike**

Abbreviations:
pos_control, normgene->exon, exronic control=probe sets against exon regions of a set of housekeeping genes**; neg_control,** normgene->intron, intronic control=probe sets against intron regions of a set of housekeeping genes**;** other spike, control->affx; **ncRNA,** non coding RNA**; ncRNA types: (i) tRNA,** transfer RNA**, (ii) Mt-tRNA,** transfer RNA located in the mitochondrial genome, **(iii) snRNA,** small nuclear RNA**, (iv) snoRNA,** small nucleolar RNA**, and (v) misc_RNA,** miscellaneous other RNA**; cDNA,** complementary DNA**; Genscan,** a gene finding algorithm. Unique IDs in ‘132 RFE-SVM Signature’ are given in “bold” font.