A Generic Framework for the Elicitation of Stable and Reliable Gene Expression Signatures

N.K. Chlis, S. Sfakianakis, E.S. Bei, M. Zervakis, *Member, IEEE*

***Abstract*—In the recent years microarray technologies have gained a lot of popularity for their ability to quickly measure the expression of thousands of genes and provide valuable information for linking complex diseases such as cancer to their genetic underpinnings. Nevertheless the large number of parameters to be estimated in relation to the small number of available samples gives rise to an "ill posed" problem where the possible solution is not stable under slight changes either in the dataset or the initial conditions and starting points. In this work we present a generic classification framework that works in an iterative manner and converges to a stable solution that combines good accuracy with biologically meaningful feature selection. The methodology is orthogonal to the specific classification algorithm used. We compare some of the most widely used classifiers based on their average discrimination power and the size of the derived gene signature. According to our proposed model named Stable Bootstrap Validation (SBV), a unified '77 common-gene signature' was selected, which is closely associated with several aspects of breast tumorigenesis and progression, as well as patient-specific molecular and clinical characteristics.**

I. Introduction

Microarray technology has proven to be a valuable tool in measuring the expression levels of different genes of a biological system. By analyzing these expression levels, scientists can search for gene expression profiles for classification of samples among certain classes of interest and provide insight into complex diseases, such as cancer. A variety of classification algorithms have been used for that specific task, having varying performances on different datasets. In particular, selecting the optimal classification method for a given dataset is not trivial.

Classifier performance is severely hindered due to the "curse of dimensionality", since the number of features (genes) used is orders of magnitude larger than the number of available training samples, introducing multi-colinearity in the input data matrix. For this reason, feature subset selection (FSS) methods have been developed in order to reduce the number of features that participate in the classification process. FSS methods aim to select a "small" subset of features that are "important". Different categories of FSS methods use different criteria when deciding the "importance" of features [5],[12].

This research is supported by "ONCOSEED" project funded by the NSRF2007-13 of the Greek Ministry of Development.

N.K. Chlis, E.S. Bei and M. Zervakis are with the Department of Electronic and Computer Engineering, TUC, Chania 73100, Greece (nchlis@isc.tuc.gr, abei@isc.tuc.gr, michalis@display.tuc.gr).

S. Sfakianakis is with the Institute of Computer Science, FORTH, Heraklion 71110, Greece (sgsfak@ gmail.com)

Filter (univariate) methods aim towards selecting the genes with the highest discrimination ability among the classes of interest, acting as a preprocessing step independent of the classifier used. 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. Wrapper (multivariate) methods on the other hand, attempt to derive subsets of genes with optimal prediction properties, according to a specific classifier. The classifier is perceived as a black box, independent of the feature selection method. 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. Embedded methods attempt to merge the benefits of the previous technologies by embedding the search of feature subset within the training of the classifier. However, due to the embedding of feature selection in the training process, they can prove to be harder to implement [14].

Another important aspect of microarray analysis algorithms is the stability of the observed results. In this respect, the observed results, while not random, are severely affected by the selection of the training and testing samples, as well as other algorithmic parameters. Finally, there is a need to determine the statistical significance of results and validate the extent to which they are observed merely by chance, by utilizing significance metrics such as the p-value [2] and comparing the performance of randomly selected features [4], [5].

Our approach aims to formulate a complete algorithmic framework to address and effectively handle the above-mentioned issues. This framework is based on three sequential modules of different utility, aiming to combine the selection criteria, address the issue of algorithmic stability and validate the results on a rigorous statistical scheme. The first module is used for merging the conceptual gap of univariate and multivariate methods for feature selection. A filter method is used for the initial selection of a significant (selective) subset of features, which is refined through wrapper techniques to also account for strong prediction power in the process of dimensionality reduction. Subsequently, we introduce a new method, namely Stable Bootstrap Validation (SBV), for stabilizing the results of feature selection and classification methods and compare it to standard cross validation. Finally, we evaluate the statistical significance of the observed accuracy and compare

978-1-4799-3163-7/13/$31.00 ©2013 IEEE

the performance of the emerged gene signature to random signatures. Within this framework, the performance of different classifiers is assessed on the same dataset through internal and external cross validation, in order to ensure generalization of the selected gene signatures. It is verified that the proposed framework manages to reduce the randomness of results and the variability of the signatures selected by different algorithms. Overall, it manages to derive signatures of not only statistical but also biological significance.

II. METHODOLOGY

A. Method Overview

The aim of this section is to propose a methodology for stable feature selection and classification, while evaluating the statistical significance of the observed results. An abstract representation of the proposed methodology is presented in Figure 1 in modular form. Single sub-units of the algorithmic scheme may have been used before, but they are now being employed in a holistic framework addressing the stability of prediction results. The 2-step feature selection aims to utilize the benefits of filter and wrapper methods, by first selecting a small number of features with the highest discrimination ability (Univariate FSS) and then tuned to the predictive requirements of classification performance through a specific classifier (Bootstrap Validation). The proposed SBV test during the second step of feature selection and classification leads to performance metrics of accuracy and signature size that are stable across independent runs of the evaluation method. At the final stage of our framework, the significance of the observed results is evaluated with the aim of illustrating whether they reflect the underlying biological system to a significant extent, or they are merely observed due to random noise.

B. Two-Step FSS

We propose a two-step approach to feature selection, aiming to utilize the benefits of both univariate and multivariate FSS methods in a sequential scheme. A filter FSS is first employed on the dataset in order to reduce the number of features using computationally efficient methods. The univariate gene selection algorithm used is the Significance Analysis of Microarrays (SAM) [15] that employs a modified t-statistic and repeated permutations of the data to determine if the expression of genes are strongly related to the response. This step preserves "significant" features providing insight into class differences through their differentiation ability. At the second step we employ an embedded FSS on the dataset in order to further refine the remaining features through recursive feature elimination (RFE). In essence, this step aims at preserving features with "significant" prediction ability, which maximize the performance of a certain classifier. The RFE scheme proceeds iteratively by eliminating the least significant feature and reassessing the classifier performance. 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 this study, several categories of classifiers are evaluated in conjunction with RFE. Regularized Least Squares (RLS) methods such as Ridge Regression (RR) [6] and Least Absolute Shrinkage and Selection Operator (LASSO) [7] assign weights to features in order to minimize the mean square error of the predictor, known as the empirical risk. The difference between RR and LASSO lies in the selection of the regularization term. Partial Least Squares (PLS) methods [8] including PLS-VIP and PLS-BETA utilize a metric known as the Variable Importance in Projection (VIP) score in order to set a larger weight to features more important in finding the latent variables during partial least squares regression. Support Vector Machines (SVM) [10][11] on the other hand, aim to construct the maximum margin hyperplane separating the classes. Finally, the k Nearest Neighbor method (kNN) assigns a label to a new sample, based on the labels of the majority of k known samples closest to it, according to a distance metric.

C. Stable Bootstrap Validation

An important issue of feature selection and classification is stability of results, since small alterations in the training data can cause significant variations in the features selected as well as the observed classification accuracy. One common assessment technique is k-fold Cross Validation (CV). A value for the test statistic is estimated for each fold, while all estimated values can be averaged or combined. However, the performance estimation of a given method varies between independent executions of k-fold CV. Alternative methods have emerged in an attempt to better stabilize the estimation results. StabPerf proposed in [1] utilizes random splitting of the dataset, while Min-Max model selection [13] utilizes bootstrap resampling and leave one out CV.

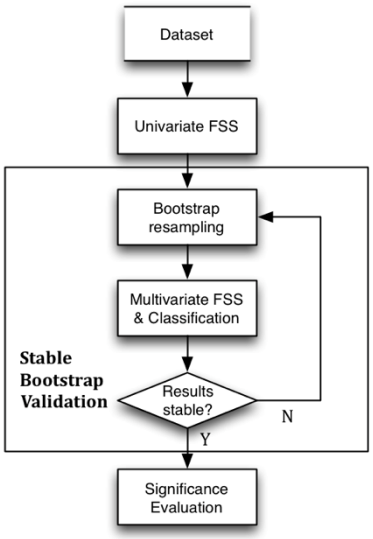

Figure 1. Abstract view of the proposed methodology.

In this paper, a model evaluation method called Stable Bootstrap Validation (SBV) is proposed, which aims to

produce stable estimation values for test statistics of a given method. SBV relies on evaluating a given method based on estimations from a large number of bootstrap datasets. Unlike [13], in SBV each bootstrap dataset is split only once into a training and test set, avoiding the added computational cost of leave-one-out CV. Moreover, instead of assessing the stability of only the observed accuracy, signature size stability is assessed as well. As pointed out in [1], average estimation performance tends to stabilize after a sufficient number of random datasets are evaluated. Since some methods can be computationally expensive, running an arbitrary large number of trials, as in StabPerf, can prove to be impractical. In that manner, we propose a formal criterion of stability for both the number of genes (features) selected by the embedded FSS method and the classifier accuracy.

The SBV criterion proceeds as follows. First, we define B as the "bootstrap window", i.e. a fixed number of bootstrap datasets. A number of 3B Bootstrap datasets of the same size and class ratio as the original dataset are created by random sampling with replacement. The embedded FSS and classification method is then executed 3B times, which provides the mean classification accuracy, $\overline{A}$, and $\overline{G}$, the total number of genes selected for the signature size. Next, the stability of the mean value of the above statistics is assessed in batches of subsequent B trials. We define the mean accuracies as $A_{w,i}, i = 1,2,3$ at the end of the first, second and third bootstrap window, respectively. Then, the maximum difference of mean accuracy between windows 1, 2 and 1,3 is defined as $\Delta A = max\left(\left|A_{w1} - A_{w2}\right|, \left|A_{w1} - A_{w3}\right|\right)$. The use of three windows enables the method to overcome possible risk of local stability. When the mean accuracy between the three last windows has been determined, it is checked for stabilization. The classification accuracy is considered stable if ΔA<acc_thresh, where acc_thresh is a fixed threshold.

Stability in terms of the mean number of genes $G_{w,i}, i = 1,2,3$ in the genomic signature is considered in a similar manner. However, we notice that different FSS methods may lead to signatures whose sizes can differ in orders of magnitude. For this reason the corresponding threshold on the number of genes is normalized by the largest signature size (number of genes) and is defined as $gen\_thresh = \left|G_{wi} - G_{wj}\right| / max\left(G_{wi}, G_{wj}\right)$ where i, j the windows being compared.

The SBV procedure ends when both $\overline{A}$ and $\overline{G}$ are found stable. Else, another set of B datasets is generated and the stability assessment is performed for the 3 windows, which now extend to cover the additional datasets. The above steps are repeated until stability for both statistics is reached.
In each iteration the following formula applies for the mean accuracy, as well as the mean number of selected genes:

$$A^{(n)}_{w_j} = \frac{1}{(n+j-1)B} \sum_{b=1}^{(n+j-1)} acc_b$$

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

where n is the iteration number, j is the window checked (1, 2, or 3), b runs over all the bootstrap datasets, acc is the accuracy result achieved by the specific algorithm in each dataset, and gen is the number of genes selected. Finally, the $\overline{G}$ genes with the highest selection frequency across all bootstrap datasets are selected as the genomic signature of the method. An abstract overview of SBV is displayed in Figure 2.

D. Statistical Significance Evaluation

After SBV has finished, two stable values have been produced: the classification accuracy $\overline{A}$ and the fixed genomic signature of size $\overline{G}$. One question that arises is whether those results are statistically significant and reflect the nature of the underlying model, or are observed merely by chance.

In order to test the significance of the observed accuracy, permutation tests are performed as to calculate the corresponding p-value. More specifically, a fixed number of 1000 bootstrap datasets are generated and the labels of the two classes are permuted. Then, the embedded "FSS & Classification" method is performed on the permuted datasets. Given an observed classification accuracy $\overline{A}$, the p-value of $\overline{A}$ is defined as the number of times that accuracy greater or equal to $\overline{A}$ was observed when a permuted dataset was used for training, divided be the total number of bootstrap datasets. If the p-value calculated is smaller than 0.05, then the observation is considered to be statistically significant.

To test the significance of the genomic signature, a bootstrap resampling approach is followed. However, permutation tests are not performed. Instead, the performance of the observed signature is compared to that of random signatures of the same size. A fixed number of 1000 bootstrap datasets are generated. For each dataset, the embedded FSS is omitted and the classifier is trained on a bootstrap dataset given a random signature of size $\overline{G}$, resulting into two performance metrics. The mean accuracy across all datasets and Prs, the percentage of times that equal or greater accuracy than $\overline{A}$ was observed when the model was trained with a random signature. Please note that Prs is not a p-value since no permutations of labels were performed.

E. PLS-kNN FSS&Classification Method

In our study, PLS-kNN outperformed all other methods in terms of accuracy and execution speed, while it leads to the smallest signature size when the VIP>2 criterion is used instead of the usually selected VIP>1. Thus, as a side effect we emphasize the joint use of PLS-kNN, which utilizes the fast and stable feature subset selection of PLS methods, while the use of kNN classifier proves to be quite efficient when applied on new samples. Two versions of the method

are evaluated, one using PLS-VIP and the other PLS-BETA for feature selection, with the former showing more balanced performance in terms of size and accuracy of the signature.

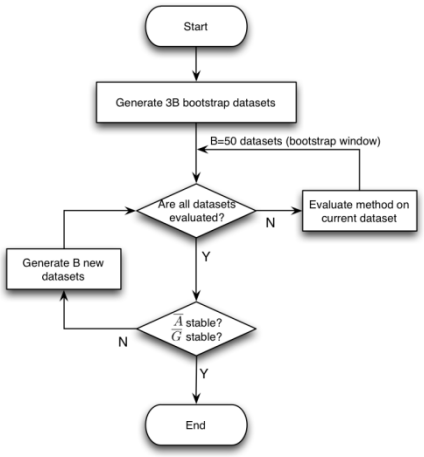

Figure 2. Flowchart of SBV.

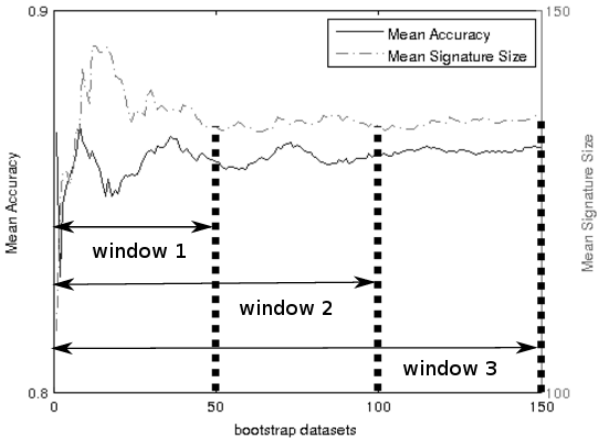

Figure 3. Accuracy and signature size stability assessment windows of SBV

III. RESULTS

The proposed methodology is tested on a breast cancer dataset. Our dataset consists of 529 samples, 104 of which are control and 425 cancer produced by the integration of 5 publicly available datasets (GEO access numbers: 22820, 19783, 31364, 9574, 18672). For each sample there are measurements of all remaining 4174 genes after the first step of univariate feature selection using the SAM algorithm. For the second step of feature selection, recursive feature elimination is embedded within the classification environment. The known classification methods evaluated are RR, LASSO, PLS-VIP, PLS-BETA, SVM, PLS-VIP-kNN, PLS-BETA-kNN. Each bootstrap dataset is split into a training (90%) a test set (10%). Moreover, each of the training and test sets has the same cancer/control ratio as the original dataset (4 to 1). The bootstrap window B of SBV was set to 50 bootstrap datasets, the accuracy threshold acc_thresh was set to 0.01 and the signature size threshold gen_thresh to 0.05. Finally, the threshold for the PLS methods was first set to VIP>1 and then to VIP>2 for a second set of results. SBV was run for each FSS & Classification method, all methods were stable after at most 300 bootstrap datasets.

The results for each method under the proposed evaluation framework are presented in table I. The method with the highest classification accuracy proved to be PLS-VIP-kNN. The exact value of k did not seem to have significant effect on classification accuracy, as the results for both k=3 and k=5 are comparable. PLS methods are also the fastest, with runtime of a few seconds for each bootstrap dataset. Finally, PLS methods are the most selective, when the VIP>2 criterion is utilized instead of the usual VIP>1. Moreover, the VIP>2 criterion leads to a small accuracy loss while the reduction of selected genes is significant. The SVM proved to be second best in classification accuracy, after PLS-kNN methods. Finally, RLS methods like RR and LASSO had acceptable classification accuracies, while LASSO selected only 136 genes, being the second most selective method after the PLS-VIP>2 case. Finally, RLS methods and SVM had comparable runtimes of roughly 2 minutes per bootstrap dataset, significantly larger that the execution time of PLS methods.

To assess the statistical significance of the observed accuracy of SBV, permutation tests were performed. 1000 bootstrap datasets with permuted class labels were generated for each method and the corresponding p-value was calculated. All observed accuracies were statistically significant with p-value 0.001, implying that the achieved accuracies are not obtained merely by chance.

To assess the significance of the observed signatures, a grouping approach was implemented. Instead of examining the signature of each method separately, the 77 common genes of all methods were selected as a unified gene signature for SBV. Only VIP>1 for the PLS methods was considered in extracting these common genes. The performance of this genomic signature of 77 common genes was then compared to the performance of random signatures on 1000 bootstrap datasets. Keeping the number of genes fixed, the simple 3NN and SVM classifiers were used to evaluate the accuracy of both the 77 common genes and the random signatures. As also mentioned in [4], [5], the performance of random signatures might be comparable to that of the 77 selected genes. However, the common-gene signature results in higher classification accuracy on average.

Comparing the results of SBV with those of 10fold CV we observe that in most cases they are comparable. However, 10fold CV systematically results in larger signature size and smaller estimates of classification accuracy than SBV. Moreover, in the case of the RLS classifiers RR and LASSO, the classification accuracy estimate of 10fold CV is off by 60%. That is most likely observed as a result of uneven class ratios among samples in random selection of CV and could possibly be avoided if stratified CV is used instead. Despite the fact that CV selects larger signatures per method, the common genes of all methods are fewer (59) when compared to the 77 common

genes of SBV. That leads to the conclusion that signatures produced by CV include more random noise than those produced by SBV.

For the biological interpretation of the selected unified '77 common-gene signature' and of the '19 gene signature' from PLS-VIP-3NN* method, the Genes-to-Systems Breast Cancer (G2SBC) Database [16] and WebGestalt (WEB-based GEne SeT AnaLysis Toolkit) [17] are used. G2SBC provides literature based evidences that 51.95% of genes of the unified '77 common-gene signature' and 36.84% of genes of the '19 gene signature' are altered in breast cancer cells. The '19 gene signature', which includes six genes from '77 common-gene signature', shares similar attributes with the unified '77 common-gene signature'. Moreover, the WebGestalt functional analysis in terms of gene ontology biological processes and KEGG pathways provides significant ($p<0.05$) enriched processes (cell proliferation, cell development, growth, cell differentiation, cell migration, extracellular matrix organization) and KEGG pathways (Pathways in cancer, ECM-receptor interaction, MAPK signaling pathway, ErbB signaling pathway, Cytokine-cytokine receptor interaction) that are implicated in breast tumorigenesis, progression and metastasis. Furthermore, the disease association analysis verified the significant ($p<0.05$) relation of both signatures to breast cancer disease (referred as breast diseases, breast neoplasms) but also creates gene sets that are related to neoplastic process, pathologic processes, carcinoma *in situ*, disease progression, recurrence, neoplasm invasiveness, neoplasm metastasis, disease susceptibility, genetic predisposition to disease, chromosome aberrations, hypersensitivity, and a variety of diseases including breast cancer comorbidities (hypertension, diabetes mellitus, osteoarthritis, depression, etc.) or treatment side effects (e.g. musculoskeletal diseases). Thus, both signatures are giving insights to breast carcinogenesis, and patients' clinical and molecular profiles in part according to the original studies (GEO access numbers: 22820, 19783, 31364). Gene sets of both signatures that are not yet associated with breast cancer but participate in various significant ($p<0.05$) biological processes, pathways and a broad spectrum of diseases (248 diseases for the '77 common-gene signature' and 103 for the '19 gene signature') should be further explored in order to decoding their implication in breast carcinogenesis and clinical outcome in breast cancer disease. Apart from this fact, both signatures appear to be potential useful as clinical signatures.

IV. Conclusion

In this paper we propose a framework for selection of stable genomic signatures that maximize classifier performance. We introduce a two-step feature selection scheme, aiming to utilize the advantages of both univariate and multivariate FSS methods, leading to small genomic signatures while being computationally efficient. Furthermore, we propose a validation method called SBV that employs bootstrap resampling of the dataset and a convergence metric in order to extract stable estimates of the classification accuracy as well as the genomic signature. Moreover, the statistical significance of the observed results is evaluated. Next, the performance of SBV is compared to that of standard 10fold CV across a set of different types of FSS&Classification methods: LASSO, RR, SVM, PLS and kNN on a breast cancer dataset. It is observed that even though the derived results are comparable, SBV leads to signatures that include less random noise. We also notice that the PLS-kNN method, which utilizes feature selection using PLS and classification with the kNN classifier performed best on our dataset. Finally, the classification performance of the signature consisting of the 77 common genes of all FSS&Classification methods is assessed, using a 3NN classifier as well as an SVM.

The unified '77 common-gene signature' that was selected from all methods according to our proposed model evaluation method SBV and the individual '19 gene signature' that resulted from PLS-VIP-3NN* method that is accompanied with highest classification accuracy, are highly associated with breast cancer disease and patients' clinical features that could be eventually viewed as promising clinical signatures.

	SBV		10fold CV	
FSS&Classification Method	Accuracy	Signature Size	Accuracy	Signature Size
RFE-LASSO	86.4%	136	27.8%	301
RFE-RR	88%	1372	26.9%	2044
RFE-PLS VIP	87.6%	825	87.5%	754
RFE-PLS VIP*	82.2%	18	82.6%	20
RFE-PLS BETA	88.8%	1159	87.3%	1083
RFE-PLS BETA*	81.2%	16	81.1%	15
RFE-SVM	89.9%	640	75.2%	930
RFE-PLS-VIP-3NN	94.7%	793	93.6%	802
RFE-PLS-VIP-3NN*	90.5%	19	86.6%	22
RFE-PLS-VIP-5NN	94.3%	776	92.8%	804
RFE-PLS-VIP-5NN*	90.4%	18	87.1%	21
RFE-PLS-BETA-3NN	94.3%	1194	93.6%	1122
RFE-PLS-BETA-3NN*	90.7%	15	86.8%	15
RFE-PLS-BETA-5NN	94.3%	1142	91.9%	1080
RFE-PLS-BETA-5NN*	89.7%	16	88.5%	14

Table I. Comparison of SBV and 10foldCV results. PLS methods marked with * utilize the criterion VIP>2 instead of VIP>1.

Classifier	Mean accuracy	p-value	Random accuracy	Prs
3NN	95.2%	<0.001	92.7%	25.5%
SVM	81.1%	0.013	67.5%	23.4%

Table II. Genomic signature for 77 common genes evaluation.

References

[1] 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.

[2] 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.

[3] 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.

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

[5] 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.

[6] C. Saunders, A. Gammerman, V. Vovk, "Ridge Regression Learning Algorithm in Dual Variables," Royal Holloway, University of London Proceedings of the 15th International Conference on Machine Learning, ICML '98.

[7] R. Tibshirani "Regression, Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society. Series B (Methodological)* vol. 58, no. 1, pp. 267-288, 1996.

[8] I.G. Chong, C. H. Jun, "Performance of some variable selection methods when multicollinearity is present," Chemometrics and Intelligent Laboratory Systems., vol 78, pp. 103-112, 2005.

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

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

[11] 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.

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

[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] 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

[15] V.G. Tusher, R. Tibshirani, 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

[16] E. Mosca, R. Alfieri, I. Merelli, F. Viti, A. Calabria, L. Milanesi , " A multilevel data integration resource for breast cancer study," *BMC Syst. Biol.*, vol. 4, pp. 76, Jun 2010.

[17] J. Wang, D. Duncan, Z. Shi, B. Zhang, " WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013," *Nucleic. Acids. Res.,* 2013, vol. 41, Web Server issue, pp. W77-83, 2013.