A Three-Stage Method to Select Informative Genes from Gene Expression Data in Classifying Cancer Classes

Article · July 2013

DOI: 10.1109/ISMS.2010.39

CITATIONS
0

READS
40

4 authors, including:

Mohd Saberi Mohamad
Universiti Teknologi Malaysia
175 PUBLICATIONS **401** CITATIONS
SEE PROFILE

Safaai bin deris
Universiti Teknologi Malaysia
205 PUBLICATIONS **683** CITATIONS
SEE PROFILE

A Three-Stage Method to Select Informative Genes from Gene Expression Data in Classifying Cancer Classes

Mohd Saberi Mohamad[1,2], Sigeru Omatu[1], Safaai Deris[2], Michifumi Yoshioka[1]
[1]Department of Computer Science and Intelligent Systems, Osaka Prefecture University,
Sakai, Osaka 599-8531, Japan
mohd.saberi@sig.cs.osakafu-u.ac.jp,{omatu, yoshioka}@cs.osakafu-u.ac.jp
[2]Department of Software Engineering, Universiti Teknologi Malaysia,
81310 Skudai, Johore, Malaysia
safaai@utm.my

***Abstract*—The process of gene selection for the cancer classification faces with a major problem due to the properties of the data such as the small number of samples compared to the huge number of genes, irrelevant genes, and noisy data. Hence, this paper aims to select a near-optimal (small) subset of informative genes that is most relevant for the cancer classification. To achieve the aim, a three-stage method has been proposed. It has three stages: 1) pre-selecting genes using a filter method; 2) optimizing the gene subset using a multi-objective hybrid method; 3) analyzing the frequency of appearance of each gene. By performing experiments on three public gene expression data sets, classification accuracies and the number of selected genes of the proposed method are better than those of other experimented methods and previous works. A list of informative genes in the final gene subsets is also presented for biological usage.**

Keywords-component; cancer classification; genetic algorithm; gene selection; gene expression data; three-stage method;

I. INTRODUCTION

Microarray technology is used to measure the expression levels of thousands of genes simultaneously, and finally produce gene expression data. A comparison between the gene expression levels of cancerous and normal tissues can also be done. This comparison is useful to select those genes that might anticipate the clinical behavior of cancers. Thus, there is a need to select informative genes that contribute to a cancerous state. An informative gene is useful for cancer classification. However, the gene selection process poses a major challenge because of the following characteristics of gene expression data: the huge number of genes compared to the small number of samples (high-dimensional data), irrelevant genes, and noisy data.

To overcome the challenge, a gene selection method is used to select a subset of genes for cancer classification. The gene selection method has several advantages such as maintaining or improving classification accuracy, reducing the dimensionality of data, and removing irrelevant and noisy genes.

There are two types of gene selection methods [1]: if a gene selection method is carried out independently from a classifier, it belongs to the filter approach; otherwise, it is said to follow a hybrid (wrapper) approach. In the early era of microarray analysis, most previous works have used the filter approach to select genes because it is computationally more efficient than the hybrid approach [2-3]. However, the filter approach results in inclusion of irrelevant and noisy genes in a gene subset for the cancer classification. The hybrid approach usually provides greater accuracy than the filter approach. Until now, several hybrid methods, especially a combination between a genetic algorithm (GA) and a support vector machine (SVM) classifier (GASVM), have been implemented to select informative genes [1],[4-8]. The drawbacks of the hybrid methods (GASVM-based methods) in the previous works are [1],[4-8]: 1) intractable to efficiently produce a small subset of informative genes when the total number of genes is too large (high-dimensional data); 2) the high risk of over-fitting problems.

In order to solve the problems derived from gene expression data and overcome the limitations of the hybrid methods in the previous works [1],[4-8], we propose a three-stage method (3-SGS) for gene selection. This method is able to perform well in the high-dimensional data and reduce the high risk of over-fitting problems since it has three stages as follows: stage 1 for producing a subset of genes; stage 2 for resulting near-optimal subsets of genes; stage 3 for yielding a small (final) subset of informative genes based on the frequency of appearance for each gene in the near-optimal subsets. The diagnostic goal is to develop a medical procedure based on the least number of possible genes to detect diseases. Thus, the ultimate goal of this paper is to select a small subset of informative genes (minimize the number of selected genes) for yielding high cancer classification accuracy (maximize the classification accuracy). To achieve the goal, we adopt 3-SGS where 3-SGS is evaluated on three real gene expression data sets of tumor samples.

The outline of this paper is as follows: Sections 2 and 3 discuss previous works and the detail of the proposed 3-SGS, respectively. In Section 4, gene expression data sets, experimental setup, and experimental results are described. The conclusion of this paper is provided in Section 5.

II. PREVIOUS WORKS

Several hybrid methods, i.e., GASVM-based methods have been proposed for genes selection of gene expression

978-0-7695-3973-7/10 $26.00 © 2010 IEEE
DOI 10.1109/ISMS.2010.39

data [1],[4-8]. The hybrid methods usually provide greater accuracy than filter methods since genes are selected by considering relations among genes. Generally, our previous GASVM-based methods performed well in high-dimensional data, e.g., gene expression data since we proposed a modified chromosome representation and a multi-objective approach [4-6]. However, the methods yielded inconsistent results when they were run independently.

The work of Huang and Chang can simultaneously optimize genes and SVM parameter settings by using a GASVM-based method [7]. Next, integrated algorithms based on GASVM have been proposed by the works of Shah and Kusiak [1] to produce a small subset of genes. Peng *et al*. introduced a feature elimination post-processing step after the step of a GASVM-based method in order to reduce the number of selected genes again [8].

Nevertheless, the GASVM-based methods of the previous works are still intractable to efficiently produce a small subset of informative genes from high-dimensional data due to their binary chromosome representation drawback [1],[5-8]. The total number of gene subsets produced by GASVM-based methods is calculated by $M_c = 2^M - 1$ where M_c is the total number of gene subsets, whereas M is the total number of genes. Based on this equation, the GASVM-based methods are almost impossible to evaluate all possible subsets of selected genes if M is too large (high-dimensional data). The work of Peng *et al*. has implemented a pre-processing step to decrease the dimensionality of data, but this step can only reduce a small number of genes, and many genes are still available in the data [8]. The GASVM-based methods also face with the high risk of over-fitting problems. The over-fitting problem of hybrid methods (e.g., GASVM-based methods) was also reported in a review paper in Saeys *et al*. [9].

III. The Proposed Three-stage Method (3-SGS) for Gene Selection

In order to overcome the drawbacks of GASVM-based methods in the related previous works [1],[4-8], we propose 3-SGS that contains three stages for gene selection. 3-SGS in our work differs from the methods in the previous works in one major part. The major difference is that our proposed method involves three stages (using a filter method, a hybrid method, and frequency analysis), whereas the previous works usually used only one stage (using a hybrid method) [1],[4-7] or two stages (using a filter method and a hybrid method) [8]. The difference is necessary in order to produce near-optimal gene subsets from high-dimensional data, reduce the high risk of over-fitting problems, and finally yield a small subset of informative genes. The computational flow of 3-SGS for gene selection is shown in Fig. 1.

A. Stage 1: Pre-selecting Genes Using a Filter Method

A filter method such as gain ratio (GR) or information gain (IG) is used in this stage (stage 1) to pre-select genes and produce a subset of genes. After the pre-select process, the dimensionality of data is also decreased. The filter method calculates and ranks a score for each gene. Genes with the highest scores are selected and put into a gene subset. This subset is then used as an input to the second stage. A GASVM-based method, i.e., a multi-objective GASVM (MOGASVM) that performs poorly in high-dimensional data is implemented in the second stage of 3-SGS. Therefore, the filter method (GR or IG) is firstly used to reduce the high-dimension in order to overcome the drawback of the GASVM-based method. If the subset that produced by the filter method is in small-dimension, the combination of genes is not complex, and then MOGASVM is possible to produce near-optimal genes subsets.

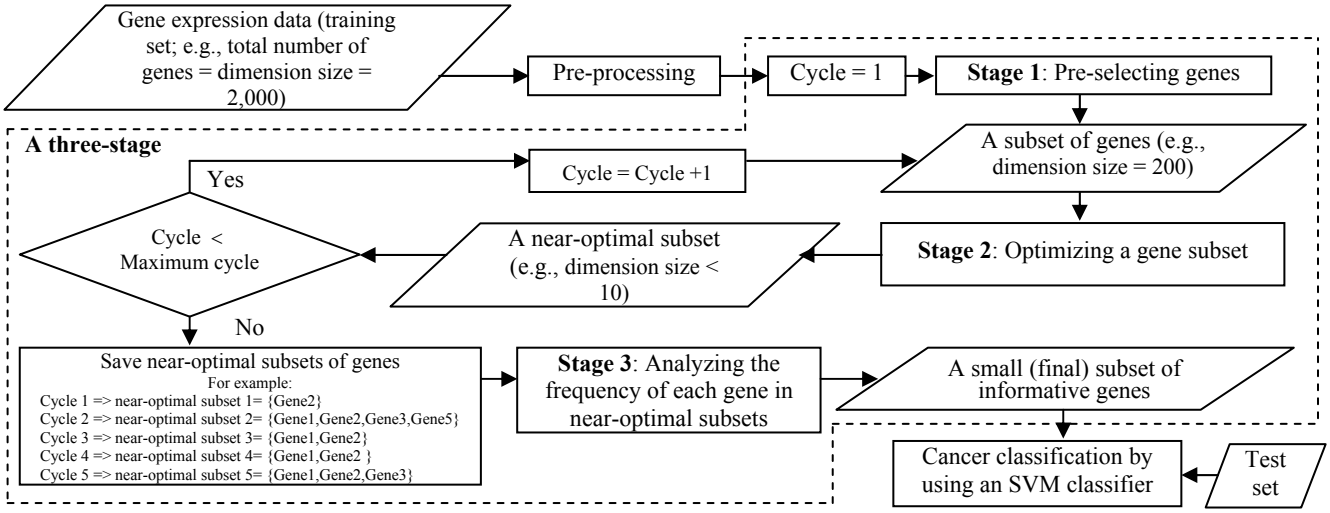


Figure 1. The proposed three-stage method (3-SGS).

B. Stage 2: Optimizing a Gene Subset Using MOGASVM

In this stage, MOGASVM optimizes gene subsets that are produced by the first stage, and finally yields near-optimal subsets of genes. This stage is cycled until the maximum number of cycles is satisfied. The near-optimal subsets are identified by an evaluation function in MOGASVM that uses two criteria: maximization of leave-one-out-cross-validation (LOOCV) accuracy and minimization of the number of selected genes. MOGASVM selects and optimizes genes by considering relations among them in order to remove irrelevant and noisy genes. The near-optimal subsets can be obtained since the dimensionality and complexity of data have been firstly reduced by the first stage. The high risk of over-fitting problems can be also decreased because of the reduction in the first stage. The detail of MOGASVM can be found in the previous work [6].

C. Stage 3: Analyzing the Frequency of Each Gene in Near-optimal Subsets

In this stage, frequency analysis is implemented to identify the most frequently selected genes in near-optimal gene subsets. The frequency of appearance of each gene in each near-optimal gene subset is examined and analyzed to assess the relative importance of genes for cancer classification. The most frequently selected genes in near-optimal gene subsets are presumed to be the most relevant for the classification. Finally, a small (final) subset of informative genes (K genes, K is a number of genes) is produced and used to construct an SVM classifier. This subset contains a small number of informative genes with high classification accuracy. Table 1 shows an example on how to obtain the frequency of each gene and the final subset of informative genes. This paper has produced two methods of 3-SGS obtained from combinations of two different filter methods (GR and IG) and MOGASVM. These methods are 3-SGS-GR and 3-SGS-IG.

TABLE I. AN EXAMPLE TO OBTAIN THE FREQUENCY OF EACH GENES (ASSUME THAT THE MAXIMUM NUMBER OF CYCLES IS FIVE)

Cycle	Near-Optimal Gene Subset			
	Gene 1	Gene 2	Gene 3	Gene 4
1	N	Y	N	N
2	Y	Y	Y	N
3	Y	Y	N	N
4	Y	Y	N	N
5	Y	Y	Y	N
Frequency	4	5	2	0
A Final Subset of Informative Genes (following the most frequently selected genes)	Gene 2;	Gene 1;	Gene 3;	

Note: 'Y' means that the corresponding gene is included in a near-optimal gene subset. Otherwise, 'N' means that the corresponding gene is not included.

IV. EXPERIMENTS

A. Data Sets and Experimental Setup

Three benchmark gene expression data sets that contain binary classes and multi-classes of cancer samples are used to evaluate 3-SGS. They are summarized in Table 2. Table 3 contains parameter values for 3-SGS. These values are chosen based on the results of preliminary runs. Three criteria following their importance are considered to evaluate the performance of 3-SGS: test accuracy on the test set, LOOCV accuracy on the training set, and the number of selected genes. High accuracies and a small number of selected genes are needed to obtain an excellent performance. The top 200 genes are pre-selected by using GR and IG in the first stage of the 3-SGS, and are then used for the second stage.

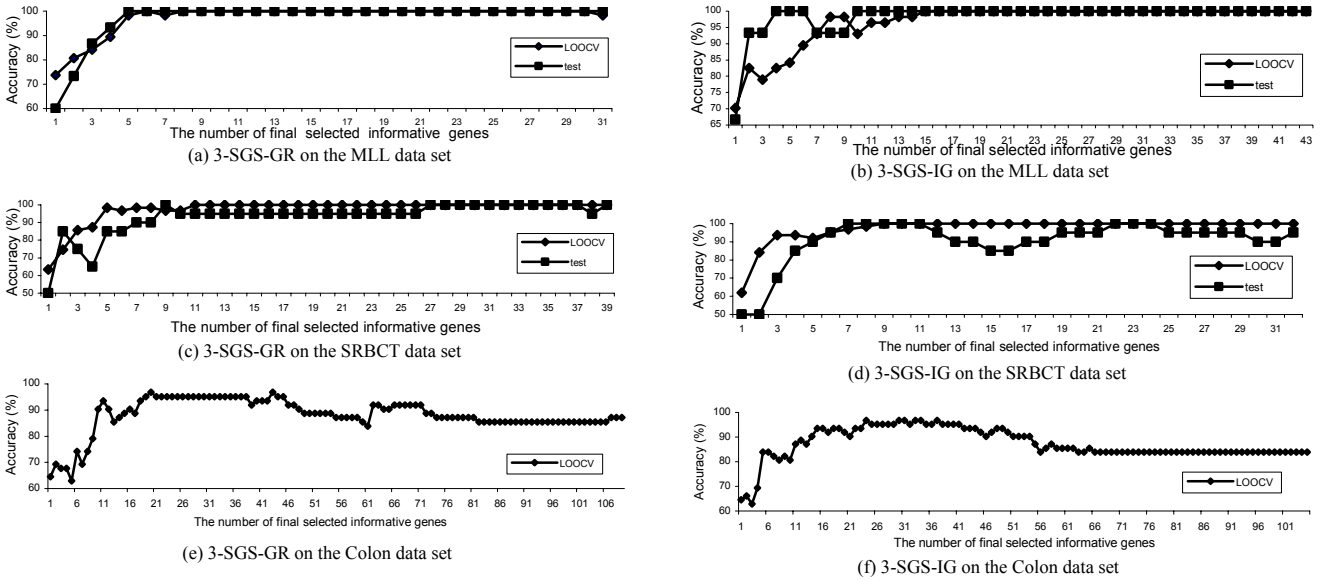
TABLE II. THE SUMMARY OF GENE EXPRESSION DATA SETS

Data set	Number of classes	Number of samples in the training set	Number of samples in the test set	Number of genes	Source
MLL [2]	3 (ALL, MLL, and AML)	57 (20 ALL, 17 MLL, and 20 AML)	15 (4 ALL, 3 MLL, and 8 AML)	12,582	http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi
SRBCT [3]	4 (EWS, RMS, NB, and BL)	63 (23 EWS, 20 RMS, 12 NB, and 8 BL)	20 (6 EWS, 5 RMS, 6 NB, and 3 BL)	2,308	http://research.nhgri.nih.gov/microarray/Supplement/
Colon [8]	2 (Normal and tumor)	62 (22 normal and 40 tumor)	Not available	2,000	http://microarray.princeton.edu/oncology/affydata/index.html

Note:
ALL = acute lymphoblastic leukemia.
MLL = mixed-lineage leukemia.
AML = acute myeloid leukemia.
EWS = ewing family of tumors.
RMS = rhabdomyosarcoma.
NB = neuroblastoma.
BL = burkitt lymphomas.
SRBCT = small round blue cell tumors.

TABLE III. PARAMETER SETTINGS FOR 3-SGS

Parameter	MLL data set	SRBCT data set	Colon data set
Size of population	100	100	100
Number of generation	300	300	300
Crossover rate	0.7	0.7	0.7
Mutation rate	0.01	0.01	0.01
Maximum number of cycles	10	10	10
Cost for an SVM classifier	100	100	100



(a) 3-SGS-GR on the MLL data set

(b) 3-SGS-IG on the MLL data set

(c) 3-SGS-GR on the SRBCT data set

(d) 3-SGS-IG on the SRBCT data set

(e) 3-SGS-GR on the Colon data set

(f) 3-SGS-IG on the Colon data set

Figure 2. A relation between classification accuracies and the number of final selected informative genes (*K* genes) using 3-SGS.

B. Experimental Results

1) Classification Accuracies of Final Informative Genes: As shown in Fig. 2, the best results of the MLL (100% LOOCV and 100% test accuracies), the SRBCT (100% LOOCV and 100 % test accuracies), and the colon data sets (96.77% LOOCV) are obtained by using the only six (using 3-SGS-GR), nine (using 3-SGS-IG), and 20 (using 3-SGS-GR) final selected informative genes (*K* genes), respectively. Many runs have achieved 100% LOOCV accuracy on all the data sets. These results have proved that 3-SGS has efficiently selected and produced a small subset of informative genes from high-dimensional data..

2) A List of Informative Genes for Biological Usage: The informative genes and their rank scores (frequency) of the final subsets as produced by the proposed 3-SGS and reported in Fig. 2 are listed in Table 4. These informative genes among the thousand of genes may be the excellent candidates for clinical and medical investigations. Biologists can save much time since they can directly refer to the genes that have high possibility to be useful for cancer diagnosis and drug target in the future.

3) 3-SGS versus Other Previous Methods: Table 5 displays the benchmark of this work and previous related works. 3-SGS had achieved 100% LOOCV accuracy on the MLL and SRBCT data sets. The accuracy result of the colon data set only has LOOCV accuracy since this data only has the training set.

TABLE IV. THE LIST OF INFORMATIVE GENES IN THE FINAL GENE SUBSETS

Data Set	Rank Score	Gene ID / Gene Card ID
MLL	9	M11722
	7	M13143
	3	U41843
	3	Z83844
	2	L08895
	2	U59878
SRBCT	5	GC16M088332
	4	GC01P149298
	4	GC02M091189
	3	GC02P191818
	3	GC08M042151
	3	GC13M046243
	2	GC07P115952
	2	GC18M023784
	2	GC11M002110

Overall, our proposed 3-SGS has outperformed the previous works (one-stage and two-stage methods) on all the data sets except the colon data set in terms of test accuracy, LOOCV accuracy, and the number of selected genes. This is due to the fact that a filter method in the first stage of 3-SGS reduces the dimensionality of the solution space in order to produce a gene subset. Next, MOGASVM in the second stage of 3-SGS optimizes the subset automatically to yield near-optimal subsets of genes. These subsets are obtained since MOGASVM in 3-SGS considers and optimizes a relation among genes. Finally, the first *K* genes appearing most frequently are selected as the final selected informative genes for cancer classification.

TABLE V. THE BENCHMARK OF 3-SGS WITH PREVIOUS METHODS ON THE MLL, SRBCT, AND COLON DATA SETS

Category	Gene Selection Method	MLL Data Set				SRBCT Data Set				Colon Data Set		
		#Selected Genes	Accuracy (%)		Time Taken (Hour)	#Selected Genes	Accuracy (%)		Time Taken (Hour)	#Selected Genes	CV Accuracy (%)	Time Taken (Hour)
			CV	Test			CV	Test				
Three-stage	3-SGS (our proposed method)	6	100	100	(9.23)	9	100	100	(3.51)	20	96.77	(3.23)
Two-stage	GASVM [8]	-	-	-	-	-	-	-	-	12	93.55	-
	GASVM [7]	(3.5)	(100)	-	-	(6.2)	(98.75)	-	-	-	-	-
	Principal component analysis [2]	100	95	-	-	-	-	-	-	-	-	-
One-stage	Principal component analysis [3]	-	-	-	-	78	100	-	-	-	-	-
	GASVM-II+GASVM [5]	(6.5)	(100)	(92)	(46.48)	-	-	-	-	(11.6)	(99.52)	(11.87)
	GASVM-II [4]	(30)	(100)	(84.67)	(22.64)	(10)	(99.84)	(68)	(12.86)	(30)	(99.03)	(10.24)
	MOGASVM [6]	(4,465.2)	(94.74)	(90)	(260.54)	(444.7)	(100)	(81.5)	(86.56)	(446.3)	(93.23)	(76.46)
	GASVM [4]	(6,298.8)	(94.74)	(87.33)	(534.08)	(1146)	(98.41)	(78.5)	(157.69)	(979.8)	(91.77)	(98.23)

Note: The results of the best subsets shown in shaded cells. '-' means that a result is not reported in the related previous work. A result in '( )' denotes an average result. CV and #Selected Genes represent cross-validation and a number of selected genes, respectively. Methods in *italics* style are experimented in this work..

Generally, filter methods in previous works [2-3] achieved poor performances since they may result in inclusion of irrelevant and noisy genes in a gene subset for the cancer classification. These bad performances occurred because the methods evaluated a gene based on its discriminative power for the target classes without considering its relations with other genes.

GASVM-based methods [1],[4-8] may be unable to produce a small subset of informative genes because they perform poorly in high-dimensional data due to their chromosome representation drawback. GASVM-II method is impractical to be used in real applications because a variety number of selected genes should be tested in order to obtain the near-optimal one [4]. On the contrary, the proposed 3-SGS that pre-selects a number of genes at the first stage can reduce the data dimensionality and produce a gene subset. This subset is then optimized by MOGASVM in the second stage of 3-SGS to yield near-optimal subsets. Finally, the first *K* genes appearing most frequently are selected as the final selected informative genes (a small subset) for cancer classification.

The gap between LOOCV accuracy and test accuracy that resulted by 3-SGS was also lower. This small gap shows that the risk of the over-fitting problem can be reduced. On the other hand, the results of LOOCV accuracy of the related previous works [1],[4-6] were much higher than their test accuracy because they were unable to avoid or reduce the risk of over-fitting problems. Other previous works by GASVM-based methods [7-9] did not provide any test accuracy results and thus, the over-fitting problem could not be investigated in their works. Over-fitting is a major problem on hybrid methods in gene selection and classification of gene expression data when the classification accuracy on training samples, e.g., LOOCV accuracy is much higher than the test accuracy. This was also supported by a review paper in Saeys *et al*. [9] which reported that hybrid methods (e.g., GASVM-based methods) confront with the high risk of over-fitting problems because of the high-dimensional data.

V. CONCLUSION

In this paper, 3-SGS has been proposed and tested for gene selection on three gene expression data sets that contain binary classes and multi-classes of tumor samples. Based on the experimental results, the performance of 3-SGS was superior to other methods in related previous works. This is due to the fact that the filter method in the first stage of the 3-SGS can pre-select genes and reduce dimensionality of data in order to produce a subset of genes. When the dimensionality was reduced, the combination of genes and complexity of solution spaces were automatically decreased. The second stage of 3-SGS can automatically optimize the subset that is yielded by the first stage in order to produce near-optimal gene subsets. Finally, the first *K* genes appearing most frequently are selected as the final selected informative genes (a small subset) for cancer classification. Hence, the gene selection using 3-SGS is needed to produce a small subset of informative genes for better cancer classification of gene expression data. Generally, 3-SGS in this paper also obtains short running time because of the large number of genes are removed by a filter technique in the first stage. However, due to the application of a filter method in the first stage of 3-SGS, the number of pre-selected genes is difficult since it is manually done. Even though 3-SGS has classified tumors with high accuracy, it is still not able to completely avoid the over-fitting problem. Therefore, a combination between a statistical approach and a hybrid method will be proposed to solve the problem.

ACKNOWLEDGMENT

This study was supported and approved by Osaka Prefecture University, Universiti Teknologi Malaysia, and Malaysian Ministry of Higher Education. The authors gratefully thank the referees for the helpful suggestions.

REFERENCES

[1] S. Shah and A. Kusiak, "Cancer Gene Search with Data-Mining and Genetic Algorithms," Computers in Biology & Medicine, vol. 37, no. 2, Feb. 2007, pp. 251-261, doi:10.1016/j.compbiomed.2006.01.007.

[2] T. R. Golub, D. K. Slonim, P. Tomayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander, "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," Science, vol. 286, Oct. 1999, pp. 531-537, doi:10.1016/j.compbiomed.2006.01.007.

[3] J. Khan , J. S. Wei, M. Ringner, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson, P. S. Meltzer, "Classification and Diagnostic Prediction of Cancers Using Gene Expression Profiling and Artificial Neural Networks," Nature Medicine, vol. 7, no. 6, Jun. 2001, pp.673-679, doi:10.1038/89044.

[4] M. S. Mohamad, S. Deris, and R. M. Illias, "A Hybrid of Genetic Algorithm and Support Vector Machine for Features Selection and Classification of Gene Expression Microarray," International Journal of Computational Intelligence and Applications, vol. 5, no. 1, Mar. 2005, pp. 91-107, doi:10.1142/S1469026805001465.

[5] M. S. Mohamad, S. Omatu, S. Deris, M. F. Misman, and M. Yoshioka, "Selecting Informative Genes from Microarray Data by Using Hybrid Methods for Cancer Classification," International Journal of Artificial Life & Robotics, vol. 13, no. 2, Mar. 2009, pp. 417-417, doi:10.1007/s10015-008-0534-4.

[6] M. S. Mohamad, S. Omatu, S. Deris, M. F. Misman, and M. Yoshioka, "A Multi-objective Strategy in Genetic Algorithm for Gene Selection of Gene Expression Data," International Journal of Artificial Life & Robotics, vol. 13, no. 2, Mar. 2009, pp. 410-413, doi: 10.1007/s10015-008-0533-5.

[7] H. L. Huang, F. L. Chang, "ESVM: Evolutionary Support Vector Machine for Automatic Feature Selection and Classification of Microarray Data," Biosystems, vol. 90, Sep. 2007, pp. 516-528, doi:10.1016/j.biosystems.2006.12.003.

[8] S. Peng, Q. Xu, X. B. Ling, X. Peng, W. Du, and L. Chen, "Molecular Classification of Cancer Types from Microarray Data Using the Combination of Genetic Algorithms and Support Vector Machines," FEBS Letters, vol. 555, Dec. 2003pp. 358-362, doi:10.1016/S0014-5793(03)01275-4.

[9] Y. Saeys, I. Inza, and P. Larranaga, "A Review of Feature Selection Techniques in Bioinformatics," Bioinformatics, vol. 23, no. 19, Aug. 2007, pp. 2507-2517, doi: 10.1093/bioinformatics/btm344.