A Study of Cancer Microarray Gene Expression Profile: Objectives and Approaches

Hala M. Alshamlan, Ghada H. Badr, and Yousef Alohali

***Abstract*—Cancer is one of the dreadful diseases, which causes a considerable death rate in humans. Cancer is featured by an irregular, unmanageable growth that may demolish and attack neighboring healthy body tissues or somewhere else in the body. Microarray based gene expression profiling has been emerged as an efficient technique for cancer classification, as well as for diagnosis, prognosis, and treatment purposes. In recent years, DNA microarray technique has gained more attraction in both scientific and in industrial fields. It showed great importance in determining the informative genes that can cause the cancer. This led to improvements in early cancer diagnosis and in giving effective chemotherapy treatment. Studding cancer microarray gene expression data is a challenging task because microarray is high dimensional-low sample dataset with a lot of noisy or irrelevant genes and missing data. In this paper, we conduct a comprehensive study that focuses on exploring the main objectives and approaches that have been applied using cancer microarray gene expression profile. We proceed by making a classification for all approaches, and then conclude by investigating the most efficient approaches that can be used in this field.**

***Index Terms*—Cancer classification, Clustering approaches, Gene expression, Gene selection, Microarray.**

I. Introduction

THE gene expression profiles that are obtained from particular microarray experiments have been widely used for cancer classification to build an effective model. This model can differentiate normal or different cancerous states by using selected informative genes [1]. However, studying microarray dataset according to their gene expression profiles represents a challenging task. The complexity of the problem rises from the huge number of features that contribute to a profile as compared to the very low number of samples normally available in microarray analysis. Another challenge is the presence of noise (biological or technical) in the dataset, which further affects the accuracy of the experimental results.

Microarrays, known as DNA chips or some time called gene chips, are chips that are hybridized to a labeled unknown molecular extracted from a particular tissue of interest. This makes it possible to measure simultaneously the expression level in a cell or tissue sample for each gene represented on the chip [2][3]. DNA microarrays can be used to determine which genes are being expressed in a given cell type at a particular time and under particular conditions. This allows us to compare the gene expression in two different cell types or tissue samples, where we can determine the more informative genes that are responsible for causing a specific disease or cancer [4].

Recently, microarray technologies have opened up many windows of opportunity to investigate cancer diseases using gene expressions. The primary task of a microarray data analysis is to determine a computational model from the given microarray data that can predict the class of the given unknown samples. The accuracy, quality, and robustness are important elements of microarray analysis. The accuracy of microarray dataset analysis depends on both the quality of the provided microarray data and the utilized analysis approach or objective. However, the curse of dimensionality, the small number of samples, and the level of irrelevant and noise genes make the classification task of a test sample more challenging [5][6]. Those irrelevant genes not only introduce some unnecessary noise to gene expression data analysis, but also increase the dimensionality of the gene expression matrix. This results in the increase of the computational complexity in various consequent research objectives such as classification and clustering [7].

Therefore, in our study, we concentrate on the main objectives and approaches that have been applied on cancer microarray gene expression profile. We proceed by investigating the most efficient approaches in this field. The rest of this paper organized as follow: Section 2 gives the reader some background material about microarray gene expression profile. Then, Section 3 illustrates and classifies the main approaches that have been used recently for cancer microarray gene expression profile. Section 4, presents discussion and analysis about the most efficient approaches that are presented through out the paper. Finally, Section 5 concludes the paper.

II. MICROARRAY GENE EXPRESSION PROFILE

All living organisms consist of cells. For instance, Humans have trillions of cells and each cell contains a complete copy of the genome (the program for making the organism), which is encoded in DNA. A gene is a segment of DNA that specifies how to make a protein. For example, Human DNA has about 30-35,000 genes. Gene Expression is the process by which the information encoded in a gene is converted into an observable phenotype (most commonly production of a protein). Therefore, Gene Expression is the degree to which a gene is active in a certain tissue of the body, measured by the amount of mRNA in the tissue. Individual genes can be switched on (exert their effects) or switched off according to the needs and circumstances of the cell at a particular time. It is worth mentioning that in cellular organisms, expression of the right genes in the right order at the right time is particularly crucial during embryonic development and cell differentiation. Thus, abnormalities of gene expression may

Manuscript received March 17, 2013; revised April 1, 2013.

H. M. Alshamlan, Computer Science Department, King Saud University, Saudi Arabia (phone: +96612309954; e-mail: halshamlan@ksu.edu.sa).

G. H. Badr, Computer Science Department, King Saud University, Saudi Arabia and IRI - The City of Scientific Research and Technological Applications, Alex, Egypt (e-mail: badrghada@hotmail.com).

Y. A. Alohali, Computer Science Department, King Saud University, Saudi Arabia (e-mail: yousef@ksu.edu.sa).

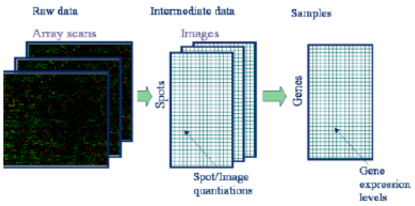
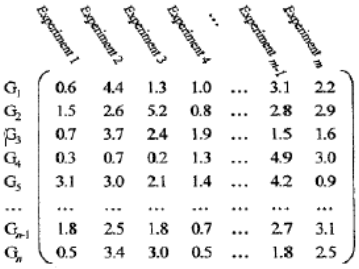


Fig. 1. Generating Matrix from Microarray Experiments [8]



	Experiment 1	Experiment 2	Experiment 3	Experiment 4	...	Experiment m-1	Experiment m
G_1	0.6	4.4	1.3	1.0	...	3.1	2.2
G_2	1.5	2.6	5.2	0.8	...	2.8	2.9
G_3	0.7	3.7	2.4	1.9	...	1.5	1.6
G_4	0.3	0.7	0.2	1.3	...	4.9	3.0
G_5	3.1	3.0	2.1	1.4	...	4.2	0.9
...	...	...	...	...	...	...	...
G_{n-1}	1.8	2.5	1.8	0.7	...	2.7	3.1
G_n	0.5	3.4	3.0	0.5	...	1.8	2.5

Fig. 2. Example of microarray gene expression matrix [8]

result in the death of cells, or their uncontrolled growth, such as in cancer [8].

A Microarray consists of a solid surface onto which known DNA molecules have been chemically bonded at special locations in array. Moreover, each array location is typically known as a probe and contains many replicates of the same molecule. Each probe represents the measurement for a single gene, and an array represents measurements for many genes (the molecules). This means that each array location is carefully chosen so as to hybridize only with the mRNA molecules that corresponds to a single gene [5][8].

Fig. 1 shows how a gene expression matrix is generated. In the gene expression matrix, rows represent genes (as opposed to features/spots in the array) and columns represent measurements from different experimental conditions measured on individual arrays. In case of cancer diagnoses, columns represent different sample tissue (cancerous tissue, or normal tissue) taken from different patients.

Generally, when multiple experiments are conducted, gene expression matrix can be viewed as a two dimensional array, indexed by an integer i identifying a known gene Gi and an integer j identifying a particular experiment trial Ej. Then Aij is the relative amount of hybridization (Gene expression level) for each gene Gi in experiment Ej. In Fig. 2, An example of the hybridization as we explained before gene expression matrix (A) for n genes assayed by m microarray experiments. Each entry represents the relative amount of hybridization for each gene in each experiment. Typically, microarray matrixes contain thousands of rows (Genes) and dozens of columns (Experiments) [8].

III. OBJECTIVES AND APPROACHES

Effective microarray experiments require careful planning that is based on clear objectives [9]. The objectives of many studies using DNA microarrays can be divided into three main groups: Gene selection, class discovery, and classification. Gene finding or gene selection is the process of selecting the smallest subset of informative genes that are most predictive to its related class. This helps in maximizing the classifiers ability to classify samples more accurately. Class discovery concerns with representing a new cancer or disease as a new class. Class prediction (classification) predicts the class of a new specimen based on its expression profile [8][10]. In this section, we will define these three objectives and illustrates the most efficient approaches that are used in order to achieve them in more details.

A. Gene Finding (Gene Selection)

The gene finding studies are very important in microarray study because it is aimed to reduce the dimensionality of microarray dataset by selecting the most informative genes. Moreover, gene-finding methods typically perform class comparison to determining the genes whose expression is correlated to a quantitative measurement or a survival time [48]. Gene selection is a process of selecting the smallest subset of informative genes that are most predictive to its related class for classification and that maximizes the classifiers ability to classify samples accurately. The optimal feature selection problem has been shown to be NP hard [11].

Notably, there are several advantages for gene selection method. For diagnoses, it is much cheaper to focus on the expression of only a few genes rather than on thousands of genes. This leads to a reduction in the cost of clinical diagnosis. In addition, the feature selection reduces the dimensionality problem, and this leads to a reduction in computational cost. Furthermore, feature set selection often gives rise to a much smaller and a more compact gene set [2].

Some gene selection methods do not assume any specific distribution model on the gene expression data and they are referred to as model-free gene selection methods or usually called Filter method. While other gene selection methods assuming certain models are referred to as model-based gene selection methods or may called Wrapper method [12]. In other hand, some researcher applied Filter method and Wrapper method, this method called Hybrid Method. Moreover, hybrid gene selection methods search for an optimal subset of features is built into the classifier construction, and can be seen as a search in the combined space of feature subsets and hypotheses [13].

B. Class Discovery (Clustering)

Class discovery is different form gene finding or class prediction because it does not involve any predefined classes. Class discovery involves analyzing a given set of gene expression profiles with the goal of discovering subgroups that share common features. It involves grouping together specimens that are based on the similarity of their expression profiles with regard to the genes represented on the array [14]. Cluster analysis or clustering is often used for class discovery.

The objective of clustering expression profiles of tumors is to determine new disease (cancer) classifications. Clustering aims at dividing the data points (genes or samples) into groups (clusters) using measures of similarity, such as correlation or Euclidean distance. Discovery of a new class is usually achieved by an unsupervised machine learning method with the help of a clustering technique such as hierarchical clustering, k-means clustering and self organizing maps (SOM) [10] [6]. It is called unsupervised because the grouping is not driven by any phenotype external to the expression profiles, such as tissue type, stage, grade or response to treatment [14] [9].

C. Class prediction (Classification)

Class prediction or Classification (including the assignment of labels to samples based on their expression patterns) is typically based on statistical or supervised machine learning methods [10][15]. It usually requires finding which genes are informative for distinguishing the pre-defined classes, estimating the parameters of the mathematical function that is used, and estimating the accuracy of the predictor. Class prediction is a very useful and helpful data mining method for medical problems of diagnostic classification, prognostic prediction and treatment selection. Also, most cancer studies in microarray expression profiling have class comparison or class prediction objectives [8] [10] [16].

Dougherty et al. (1995) in [17] indicated that supervised methods are better than unsupervised methods. In supervised method we need to train the classifier before we start in classifying process, while unsupervised method we start in classifying process without any training. Moreover, supervised methods are usually more effective in cancer classification researches. And they are used for cancer prediction as follows: A classifier is trained with a part of the samples in the cancer microarray dataset. Then, the trained classifier is used to predict the samples in the rest of the dataset to evaluate the effectiveness of the classifier [18].

Microarray based gene expression profiling has become an important and promising dataset for cancer classification that are used for diagnosis and prognosis purposes. The most important motivation for using microarray datasets is to classify unknown tissue samples according to their expression profiles. For example, it can be used in classifying cancerous or normal samples, or to discriminate different types or subtypes of cancer [5]. Classification tasks are widely used in real-world applications, some of them involves only binary classifies and many of them involve more than two classes, the so-called multi-class classification problem. Moreover, since different subtypes of a cancer respond differently to the same therapy, it is important to diagnose the cancer type of a patient correctly, and then customize the treatment for that patient. It is worth mentioning, that DNA microarrays have been recently receiving big attention in bi- and multi-cancer classification [10]. In the past decade, a number of feature selection and classification methods have been proposed for bi-class and multiclass cancer classification. In order to demonstrate the differences between the binary class classification approaches and multi class cancer classification approaches , in the following subsections we summarize these approaches.

1) Binary Cancer Classification: Classification tasks are widely used in real-world applications, some of them involves only binary classifies and many of them involve more than two classes, called multi class classification problem. Their application domain is diverse; for instance, in the field of bioinformatics, and, in the cancer classification of microarrays. In the literature, binary cancer classification problems have been more extensively studied such as for leukemia, and colon cancer [2][18][19][20]. It is worth mentioning, that there are many benchmarks for two-class cancer microarray that are available online. In Table 1, we summarize the most useful two-class cancer microarray datasets.

TABLE I
BINARY CLASS CANCER MICROARRAY DATASETS

Cancer Microarray	No. Of Classes	No. Of Samples	No. Of Genes
Leukemia [21]	2	72	7129
Lung [22]	2	181	12533
Colon [23]	2	62	2000
Prostate [24]	2	136	12600
Ovarian [25]	2	253	15154
Breast [26]	2	38	7129
Lymphoma [27]	2	96	4026

Most of the proposed binary class classification methods in literature achieved accurate result. There are several techniques that have been applied for classifying two-class cancer microarray dataset including statistical methods, Data mining methods, SVM (Support Vector Machine), k-NN (k- Nearest Neighbour), ANN (Artificial Neural Network), GA (Genetic Algorithms), Practice swam optimization (PSO), Naive Bayes (NB), Decision Trees (DTs). The Support Vector Machine (SVM) algorithm has proven to be one of the most powerful supervised learning algorithms in biological data analysis including microarray-based expression analysis [28] [29] [30]. Also, SVM method utilized as binary categorical classifiers and it has been shown to consistently outperform other classification approaches including weighted voting and k-nearest neighbors [31].

2) Multi Class Cancer Classification: Recently, microarray technology has been considered as a significant approach to classify multi categories (types) for cancer for early diagnosis and chemotherapy treatment purposes. As we noted, a number of systematic methods have been developed and studied to classify cancer types using gene expression data [31][32][21]. However, most of these studies were confined towards binary gene selection problems and only a very few considered multi class gene selection and classification [31][33][34][35]. This is because multi class gene selection and classification is significantly harder than the binary problems [36]. In Table 2, more useful benchmark multi-class cancer-related human gene expression datasets that are gleaned from the literature are described. We have chosen from all multi-class cancer microarray dataset that are used, five datasets, Lung Cancer, Brain Tumor, CNS, NCI60, and GCM. These datasets have less classification accuracy result, when compared with other dataset like leukemia and SRBCT dataset. So, the research in these multi class cancer microarray datasets is challenge and open. In 1990, the National Cancer Institute 60 (NCI60) platform included 60 human tumor cell lines that represented 9 cancer types. Table 3, presents descriptions about NCI60 dataset. GCM is a more complicated microarray dataset that includes 14 types of cancers [16]. This data set contains expression data of 16,306 genes with the total of 198 samples has been already divided into two parts, i.e., 144 for training and the other 54 for testing. Table 4, gives the general information of the GCM dataset.

Notably, multi-class cancer classifiers that are based on support vector machines are the most useful and effective classifiers in performing accurate cancer diagnosis from microarray gene expression data. The first generation of SVMs could only be applied to binary classification tasks. However, most real life diagnostic tasks, especially cancer diagnostic are not binary. Therefore, several algorithms have emerged during the last few years that allow multi-class classification with SVMs, such as DAGSVM [37], a method by Weston and Watkins (WW) [38], and method by Crammer and Singer (CS) [15][32][39]. Furthermore, there are some novel methods in literature that aim to improve the performance of SVM by combining with Evolutionary algorithms, such as ESVM [40], and GASVM [41], or with Fuzzy algorithms like FSVM [34].

However, there are many other multi class classifier that are proposed in the literature such as, Statistical approaches, Evolutionary Algorithm, K-nearest neighbors (KNN), naive Bayes (NB), neural networks (NN), and decision tree (DT). Furthermore, Artificial neural network (ANN) methods provide an attractive alternative to the above approach for a direct multi-class classification problem [42]. Neural networks can map the input data into different classes directly with one network. However, conventional neural networks usually produce lower classification accuracy than SVM [34]. There are several multi-class cancer classification algorithms that based on neural network, such as FNN [16], ELM [42], WNN [43], PNN [44], and SANN [45]. Also, efficient multi-class cancer classification methods that are based on statistical techniques, such as, maximum likelihood classification (MLHD) method in [15][46][26][47].

IV. Analyses and Discussion

As mentioned before, the objectives of many studies using DNA microarrays can be classified into three major groups: gene finding, class discovery, and class prediction. In Table 5, we illustrate each objective with approaches and aim that has been used for studying cancer microarray gene expression profile.

Based on our study, we conclude that cancer classification is a significant field of research for cancer microarray gene expression profile. Also, most cancer studies in microarray expression profiling really have class comparison or class prediction objectives [10] [16] [44]. Moreover, microarray is considered an efficient technique for cancer classification, as well as for diagnosis, prognosis, and treatment purposes. In recent years, DNA microarray technique has gained more attraction in both scientific and in industrial fields, and it is important to determine the informative genes that cause the cancer to improve early cancer diagnosis and to give effective chemotherapy treatment. Classifying cancer microarray gene expression data is a challenging task because microarray is high dimensional-low sample dataset with lots of noisy or irrelevant genes and missing data. The inherent presence of a large number of irrelevant genes increases the difficulty of the classification task influencing the discrimination power of relevant features[9]. Those irrelevant genes do not only introduce some unnecessary noise to gene expression data analysis, but also increase the dimensionality of the gene expression matrix. This results in the increase of the computational complexity in various consequent researches such as classification and clustering. Therefore, finding an accurate gene selection method that reduce the dimensionality and selecting informative genes are very challenging issues in cancer classification.

TABLE II
Multi Class Cancer Microarray Datasets.

Cancer Microarray	No.Of Classes	No.Of Samples	No.Of Genes	Description
Lung [31]	5	203	12600	Four lung cancer types and normal tissues
Brain [48]	4	50	10367	Four malignant glioma types)
CNS [49]	5	90	7129	Central Nervous System Embryonal Tumor CNS consists of 5 subclasses: medulloblastoma (MED), malignant glioma (MG), atypical teratoid/rhabdoid tumors (AT/RT), normal cerebellum (NC) and primitive neuroectodermal (PNET))
NCI60 [50]	9	60	57725	Nine various human tumor types
GCM [31]	14	198	16306	Fourteen various human tumor types
SRBCT [51]	4	63	2304	Small round blue cell tumors (SRBCT) of childhood are hard to classify by current clinical techniques.
Leukemia [52]	3	72	12582	AML , ALL , and Mixed lineage leukemia (MLL)

TABLE III
NCI60 Dataset Description

Type Number	Cancer	Number of cell lines
1	Leukemia	6 lines
2	Melanoma	8 lines
3	Lung	9lines
4	Colon	7 lines
5	Brain	6 lines
6	Ovarian	7 lines
7	Breast	6 lines
8	Prostate	2 lines
9	Kidney	8lines

It worth mentioning, that experimental studies indicate that direct multi class classifications are much more difficult than

TABLE IV
GCM DATASET DESCRIPTION

Type Number	Cancer	Number of Sample
1	Breast	11
2	Prostate	10
3	Lung	11
4	Colorectal	11
5	Lymphoma	22
6	Bladder	11
7	Melanoma	10
8	Uterus	10
9	Leukemia	30
10	Rental	11
11	Pancreas	11
12	Ovarian	11
13	Mesothelioma	11
14	Brain	20

TABLE V
MAIN OBJECTIVES OF STUDDING MICROARRAY GENE EXPRESSION PROFILE

Objective	Approach	Aim
Gene Finding	Feature Selection	To reduce the dimensionality of microarray dataset by selecting the most informative genes
Class Discovery	Clustering	To determine new disease or cancer.
Class Prediction	Classification	To classifying samples (cancerous or normal) or to discriminate different types or subtypes of cancer

binary classifications and that the classification accuracy may drop dramatically when the number of classes increases [5]. Therefore, Instead of directly dealing with multi-class problems, many classification methods for multi class problems use some combination of binary classifiers on a One-Versus-All (OVA) or a One-Versus-One (OVO) comparison Basis [32] [44]. However, this way of implementation results in combining many binary classifiers and thus increases system complexities. It also causes a greater computational burden and longer training time [42]. For example, the support vector machine (SVM) as a binary classifier tries to map the data from a lower-dimensional input space to a higher-dimensional feature space so that to make the data linearly separable into two classes [6] [7].

In literature there are several approaches for gene selection, but we observed that a multi-class cancer classification that is combined with a gene selection method, has not been investigated intensively. Thus, we conclude that we need to use gene selection process as a mandatory step before we start cancer classification on microarray dataset. Also, we notice that ensemble classifiers are also applied for multi-class cancer classification, such as [5], but it does not generally improve the classification performance like SVM based classifier, or non-SVM based classification methods.

V. CONCLUSION

Microarray based gene expression profiling has become an important and promising approach that can be used for cancer classification. This is an important step for diagnosis and prognosis purposes. The most important objectives of microarray dataset is to classify unknown tissue samples according to their expression profiles. Microarray data suffers from the curse of dimensionality, the small number of samples, and the level of irrelevant and noise genes. These make the classification task of a test sample a very challenging problem. As a consequence, it is important to eliminate those irrelevant genes and identify the informative genes that are why a feature selection problem is crucial in gene expression data analysis. Therefore, the first step in processing the expression data is to identify a small subset of genes that are primarily responsible for the cancer. It is required to use a gene selection process as a mandatory step before we start any cancer classification approach on a microarray dataset. Thus, we can conclude that the main objectives of many studies using DNA microarrays can be classified into three major groups: gene finding, class discovery, and class prediction.

REFERENCES

[1] L.-Y. Chuang, C.-H. Yang, K.-C. Wu, and C.-H. Yang, “A hybrid feature selection method for dna microarray data,” *Computers in Biology and Medicine*, vol. 41, no. 4, pp. 228–237, 2011.

[2] H. Yu and S. Xu, “Simple rule-based ensemble classifiers for cancer dna microarray data classification,” in *Computer Science and Service System (CSSS), 2011 International Conference on*, 2011, pp. 2555–2558.

[3] C. FENG and W. LIPO, “Applications of support vector machines to cancer classification with microarray data,” *International Journal of Neural Systems*, vol. 15, no. 06, pp. 475–484, 2005.

[4] A. E, G.-N. J, J. L., and T. E., “Gene selection in cancer classification using pso/svm and ga/svm hybrid algorithms,” in *Evolutionary Computation, 2007. CEC 2007. IEEE Congress on*, 2007, pp. 284–290.

[5] S. Ghorai, A. Mukherjee, S. Sengupta, and P. Dutta, “Multicategory cancer classification from gene expression data by multiclass nppc ensemble,” in *Systems in Medicine and Biology (ICSMB), 2010 International Conference on*, 2010, pp. 4–48.

[6] G. Sheng-Bo, L. M. R., and T.-M. Lok, “Gene selection based on mutual information for the classification of multi-class cancer,” in *Proceedings of the 2006 international conference on Computational Intelligence and Bioinformatics - Volume Part III*, ser. ICIC’06. Springer-Verlag, 2006, pp. 454–463.

[7] L. M. Fu and C. S. Fu-Liu, “Multi-class cancer subtype classification based on gene expression signatures with reliability analysis,” *FEBS Letters*, vol. 561, no. 13, pp. 186 –190, 2004.

[8] R. Simon, “Analysis of dna microarray expression data,” *Best practice and research Clinical haematology*, vol. 22, no. 2, pp. 271–282, 2009.

[9] B. Tjaden1 and J. Cohen, “A survey of computational methods used in microarray data interpretation,” *Applied Mycology and Biotechnology, Bioinformatics*, vol. 6, pp. 7–18, 2006.

[10] Y. T. Young, “Efficient multi-class cancer diagnosis algorithm, using a global similarity pattern,” *Comput. Stat. Data Anal.*, vol. 53, no. 3, pp. 756–765, Jan. 2009. [Online]. Available: http://dx.doi.org/10.1016/j.csda.2008.08.028

[11] N. Patrenahalli and F. K, “A branch and bound algorithm for feature subset selection,” *Computers, IEEE Transactions on*, vol. 26, no. 9, pp. 917–922, 1977.

[12] Y. Kun, C. Zhipeng, L. Jianzhong, and L. Guohui, “A stable gene selection in microarray data analysis,” *BMC Bioinformatics*, vol. 7, no. 1, pp. 1–16, 2006.

[13] S. Yvan, I. aki, and L. Pedro, “A review of feature selection techniques in bioinformatics,” *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, Sep. 2007.

[14] A. L. Tarca, R. Romero, and S. Draghici, “Analysis of microarray experiments of gene expression profiling,” *American journal of obstetrics and gynecology*, vol. 195, no. 2, pp. 373–388, 2006.

[15] J. J. Liu, G. Cutler, W. Li, Z. Pan, S. Peng, T. Hoey, L. Chen, and X. B. Ling, “Multiclass cancer classification and biomarker discovery using ga-based algorithms,” *Bioinformatics*, vol. 21, no. 11, pp. 2691–2697, 2005.

[16] L. Wang, F. Chu, and W. Xie, “Accurate cancer classification using expressions of very few genes,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 4, no. 1, pp. 40–53, 2007.

[17] J. Dougherty, R. Kohavi, and M. Sahami, "Supervised and unsupervised discretization of continuous features," in *MACHINE LEARNING: PROCEEDINGS OF THE TWELFTH INTERNATIONAL CONFERENCE*. Morgan Kaufmann, 1995, pp. 194–202.

[18] H. Jorng-Tzong, W. Li-Cheng, L. Baw-Juine, K. Jun-Li, K. Wen-Horng, and Z. Jin-Jian, "An expert system to classify microarray gene expression data using gene selection by decision tree," *Expert Syst. Appl.*, vol. 36, no. 5, pp. 9072–9081, Jul. 2009.

[19] P. A. Mundra and J. C. Rajapakse, "Gene and sample selection for cancer classification with support vectors based t-statistic," *Neurocomputing*, vol. 73, no. 1315, pp. 2353 – 2362, 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925231210002432

[20] A. Kulkarni, B. N. Kumar, V. Ravi, and U. S. Murthy, "Colon cancer prediction with genetics profiles using evolutionary techniques," *Expert Systems with Applications*, vol. 38, no. 3, pp. 2752 – 2757, 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417410008614

[21] T. Golub, D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, L. Coller, J. Downing, M. Caligiuri, C. Bloomfield, and E. Lander, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–537, 1999.

[22] G. J. Gordon, R. V. Jensen, L. li Hsiao, S. R. Gullans, J. E. Blumenstock, S. Ramaswamy, W. G. Richards, D. J. Sugarbaker, and R. Bueno, "Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma," *Cancer Res*, vol. 62, pp. 4963–4967, 2002.

[23] UAlon, N. Barkai, D. Notterman, K. Gish, S. Ybarra, D. Mack, and A. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proceedings of the National Academy of Sciences*, vol. 96, no. 12, pp. 6745–6750, 1999.

[24] D. Singh, P. G. Febbo, K. Ross, D. Jackson, J. Manola, C. Ladd, P. Tamayo, A. A. Renshaw, A. V. D'Amico, J. P. Richie, E. S. Lander, M. Loda, P. W. Kantoff, T. R. Golub, and W. R. Sellers, "Gene expression correlates of clinical prostate cancer behavior," *Cancer Cell*, vol. 1, no. 2, pp. 203 – 209, 2002.

[25] E. Petricoin, A. Ardekani, B. Hitt, P. Levine, V. Fusaro, S. Steinberg, G. Mills, C. Simone, D. Fishman, E. Kohn, and L. Liotta, "Use of proteomic patterns in serum to identify ovarian cancer," *The Lancet*, vol. 359, no. 9306, pp. 572 – 577, 2002.

[26] A. Su, W. John, S. Lisa, K. Suzanne, D. Petre, L. Hilmar, S. Peter, P. Steven, C. Moskaluk, F. Henry, and H. Garret, "Molecular classification of human carcinomas by use of gene expression signatures," *Cancer Research*, vol. 61, no. 20, pp. 7388–7393, 2001.

[27] A. Alizadeh, M. Eisen, M. Davis, A. Rosenwald, J. Boldrick, T. Sabet, Y. Powell, L. Yang, G. Marti, T. Moore, J. Hudson, L. Lu, D. Lewis, R. Tibshirani, G. Sherlock, W. Chan, T. Greiner, D. Weisenburger, J. Armitage, R. Warnke, R. Levy, W. Wilson, M. Grever, J. Byrd, D. Botstein, P. Brown, and L. Staudt, "Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling," *Nature*, vol. 403, no. 6769, pp. 503–511, 2000.

[28] B. Michael, G. Noble, L. David, C. Nello, S. Walsh, F. Terrence, A. Manuel, and H. David, "Knowledge-based analysis of microarray gene expression data by using support vector machines," *Proceedings of the National Academy of Sciences*, vol. 97, no. 1, pp. 262–267, 2000.

[29] V. N. Vapnik, *Statistical learning theory*. Wiley, 1998.

[30] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler, "Support vector machine classification and validation of cancer tissue samples using microarray expression data," *Bioinformatics*, vol. 16, no. 10, pp. 906–914, 2000.

[31] S. Ramaswamy, T. Pablo, R. Ryan, M. Sayan, Y. Chen-Hsiang, A. Michael, L. Christine, R. Michael, L. Eva, J. P. Mesirov, P. Tomaso, G. William, L. Massimo, L. E. S., and T. R. Golub, "Multiclass cancer diagnosis using tumor gene expression signatures," *Proceedings of the National Academy of Sciences*, vol. 98, no. 26, pp. 15 149–15 154, 2001.

[32] S. Alexander, A. Constantin, T. Ioannis, H. Douglas, and L. Shawn, "A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis," *Bioinformatics*, vol. 21, no. 5, pp. 63–643, Mar. 2005.

[33] H. Chai and C. Domeniconi, "An evaluation of gene selection methods for multi-class microarray data classification," in *Proceedings of the Second European Workshop on Data Mining and Text Mining in Bioinformatics*, 2004, pp. 3–10.

[34] Y. Mao, X. Zhou, D. Pi, Y. Sun, and S. T. C. Wong, "Multiclass cancer classification by using fuzzy support vector machine and binary decision tree with gene selection," *Journal of Biomedicine and Biotechnology*, vol. 2, no. 8, pp. 160–171, 2005.

[35] M. Galar, A. Fernández, E. Barrenechea, H. Bustince, and F. Herrera, "An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes," *Pattern Recogn.*, vol. 44, no. 8, pp. 1761–1776, Aug. 2011.

[36] S. Mukherjee, "Chapter 9. classifying microarray data using support vector machines," in *of scientists from the University of Pennsylvania School of Medicine and the School of Engineering and Applied Science*. Kluwer Academic Publishers, 2003.

[37] J. C. Platt, N. Cristianini, and J. Shawe-taylor, "Large margin dags for multiclass classification," in *Advances in Neural Information Processing Systems*. MIT Press, 2000, pp. 547–553.

[38] J. Weston and C. Watkins, "Support vector machines for multi-class pattern recognition," 1999.

[39] C. Koby and S. Yoram, "On the learnability and design of output codes for multiclass problems," in *Proceedings of the Thirteenth Annual Conference on Computational Learning Theory*, ser. COLT '00. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2000, pp. 35–46.

[40] H.-L. Huang and F.-L. Chang, "Esvm: Evolutionary support vector machine for automatic feature selection and classification of microarray data," *Biosystems*, vol. 90, no. 2, pp. 516 – 528, 2007.

[41] A. El Akadi, A. Amine, A. El Ouardighi, and D. Aboutajdine, "A new gene selection approach based on minimum redundancy-maximum relevance (mrmr) and genetic algorithm (ga)," in *Computer Systems and Applications, 2009. AICCSA 2009. IEEE/ACS International Conference on*, 2009, pp. 69–75.

[42] R. Zhang, G.-B. Huang, N. Sundararajan, and P. Saratchandran, "Multicategory classification using an extreme learning machine for microarray gene expression cancer diagnosis," *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, vol. 4, no. 3, pp. 485–495, 2007.

[43] Z. Zainuddin and P. O, "Improved wavelet neural network for early diagnosis of cancer patients using microarray gene expression data," in *Neural Networks, 2009. IJCNN 2009. International Joint Conference on*, 2009, pp. 3485–3492.

[44] D. Berrar, S. Downes, and W. Dubitzky, "Multiclass cancer classification using gene expression profiling and probabilistic neural networks," in *Pacific Symposium on Biocomputing*, vol. 8, 2003, pp. 5–16.

[45] L. Roland, D. Dawn, S. Holger, T. Dirk, R. Klaus, P. Siegfried, and W. Mathias, "The subsequent artificial neural network (sann) approach might bring more classificatory power to ann-based dna microarray analyses," *Bioinformatics*, vol. 20, no. 18, pp. 3544–3552, 2004.

[46] C. H. Ooi and P. Tan, "Genetic algorithms applied to multi-class prediction for the analysis of gene expression data," *Bioinformatics*, vol. 19, no. 1, pp. 37–44, 2003.

[47] H.-L. Huang, C.-C. Lee, and S.-Y. Ho, "Selecting a minimal number of relevant genes from microarray data to design accurate tissue classifiers," *Biosystems*, vol. 90, no. 1, pp. 78–86, 2007.

[48] N. Catherine, M. D, B. Rebecca, T. Pablo, G. Cairncross, L. Christine, P. Ute, H. Christian, M. Margaret, B. Tracy, B. Peter, von Andreas, P. Scott, G. Todd, and L. David, "Gene expression-based classification of malignant gliomas correlates better with survival than histological classification," *Cancer Research*, vol. 63, no. 7, pp. 1602–1607, 2003.

[49] S. Pomeroy and P. Tamayo, "Prediction of central nervous system embryonal tumour outcome based on gene expression," *Nature*, vol. 415, no. 6870, pp. 436–442, 2002.

[50] X. Ross, U. S. M. Eisen, C. Perou, C. R. P. S. Vishwanath, I. S. Jeffrey, M. V. de Rijn Mark Waltham, A. P. Jeffrey, L. D. L. D. S. Timothy, M. J. W. D. Botstein, and P. O. Brown, "Systematic variation in gene expression patterns in human cancer cell lines," *Nature Genetics*, vol. 24, pp. 227–235, 2000.

[51] J. Khan, J. Wei, and M. Ringner, "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," *Nature Medicine*, vol. 7, no. 6, pp. 673–679, 2001.

[52] S. A. Armstrong, J. E. Staunton, L. B. Silverman, R. Pieters, M. L. den Boer, M. D. Minden, S. E. Sallan, E. S. Lander, T. R. Golub, and S. J. Korsmeyer, "Mll translocations specify a distinct gene expression profile that distinguishes a unique leukemia," *Nature Genetics*, vol. 30, no. 1, pp. 41–47, 2001. [Online]. Available: http://dx.doi.org/10.1038/ng765