CANCER CLASSIFICATION BASED ON MICROARRAY GENE EXPRESSION DATA USING DCT AND ANN

AHMAD M. SARHAN
Assoc. Prof., Department of Electrical Engineering, King Faisal University, Al-Ahsa, Saudi Arabia.
E-mail:asarhan@hotmail.com

ABSTRACT

In this paper, a stomach cancer detection system based on Artificial Neural Network (ANN), and the Discrete Cosine Transform (DCT), is developed. The proposed system extracts classification features from stomach microarrays using the DCT. The features extracted from the DCT coefficients are then applied to an ANN for classification (tumor or non—tumor). The microarray images used in this study were obtained from the Stanford Medical Database (SMD). Simulation results showed that the proposed system produces a very high success rate.

Keywords: *Stomach cancer, Microarrays, SMD database, cosine transform, Artificial Neural Network, Feature extraction*

1. INTRODUCTION

Conventional methods of monitoring and diagnosing cancer rely on a human observer to detect certain features. Cancer diagnosis is usually achieved using an imaging system, such as x—ray, Magnetic Resonance Imaging (MRI), Computed Tomography (CT), and ultrasonography. Other conventional methods used in cancer assessment include the classic morphological and clinical methods.

Microarray technology is a new tool that can automate the diagnostic task and improve the accuracy of the traditional diagnostic techniques. With microarrays, it is possible to examine the expression of thousands of genes at once. Testing for elevated expression of certain genes can assist in predicting cancer.

The difficulty in microarray analysis, however, is the ultra—high dimensionality of gene expression data (microarray image). The high dimensionality of microarrays makes processing them a very difficult task with a high time and space complexity. Therefore, to make processing microarrays feasible, it is important to reduce the dimensionality of the microarray before further processing.

In this paper, we employ the two—dimensional (2—D) Discrete Cosine transform (DCT) to compress the microarray image/matrix and obtain distinctive features from the image. Classification of the microarray image is then achieved by applying an Artificial Neural Network (ANN) to the coefficients (features) extracted from the DCT (frequency) matrix.

2. MICROARRAY TECHNOLOGY AND ANALYSIS

DNA Microarrays are glass microscope slides onto which genes are attached at fixed and ordered locations. Each gene sequence is identified by a location of a spot in the array. Using a Microarray printer (Fig.1), the DNA is spotted directly onto the slide. With microarrays, it is possible to examine a gene expression within a single sample or to compare gene expressions within two tissue samples, such as in tumor and non—tumor tissues.

Fig. 1: Microarray Printer [1]

A. Microarray Technology Review

In order to prepare a microarray for cancer diagnosis, a biopsy is taken from a suspect tumor or tissue. The messenger RNA (mRNA) is then extracted from the cells. Tiny droplets containing gene fragments are placed by robotics on specific locations on a glass slide.

The complimentary DNA (cDNA) is then labeled with fluorescent dyes, and added to the gene fragments on the slide. The labeled DNA of an active gene will bind to the gene fragment, and produce a brighter color. The intensity of the spots is then measured by a scanning microscope and the results are recorded on a graph.

The resulting microarray image is an orderly arrangement of known or unknown DNA samples attached to a solid support. Each DNA spot (probe) on the microarray is often less than 20μm in diameter. An entire array typically contains thousands of spots, producing a microarray image in tiff format [2].

Two—channel microarrays or two—color microarrays are typically hybridized with cDNA prepared from two samples to be compared (e.g. healthy and diseased tissues). The samples are labeled with two different fluorophores [3]. Fluorescent dyes commonly used for cDNA labeling include Cy3 and Cy5. Relative intensities of each fluorophore may then be used in ratio—based analysis to identify up—regulated and down—regulated genes [4]. One—channel detection can also be used. In this paper, both channels and their relative intensities are used.

B. Microarray Clustering

Clustering microarray features helps investigate their biological significance. Thus, a major use of microarray data is to classify genes with similar expression profiles into groups. A wide variety of clustering algorithms have been employed in the literature for these purposes, which include the EM algorithm [5], the LBG algorithm [6], the Self—organizing maps [7] and the K—means algorithm [8].

C. SMD Database

One of the common difficulties in doing research in microarray analysis is that microarray images are not easily obtained due to costly equipments and the need of professional labs. Researchers, therefore, usually resort to microarray databases. One of the most famous microarray databases is the SMD database [9]. The SMD database is available freely online, and is the most commonly used database by researchers in this field.

The SMD stores raw and normalized data from microarray experiments, and provides web access for researchers to retrieve, analyze and visualize microarray data. The SMD also serves as a storage site for microarray data from ongoing research at Stanford University, and facilitate the public access and use of that data [10].

The SMD database contains microarray images of many types of cancers. For each experiment, the SMD records the name of the researcher, a category and subcategory that describe the biological nature of the experiment, and the organism that served as the source of the DNA. A query can be based on any combination of these criteria.

3. THE STATE OF THE ART OF MICROARRAY PROCESSING

ANNs have been at the heart of the major portion of microarray—based cancer diagnostic systems. Mostly, multi—layered feed—forward neural networks (FFNN) have been used.

Literature review includes the following contributions. Yang worked on multi—class cancer diagnosis algorithm using a global similarity pattern where for each cancer subtype, genes were ranked to determine a characteristic pattern [11]. Hang *et al.* worked on the classification problem by expressing each testing sample as a linear combination of all the training samples, a method they called "sparse representations of test samples" [12]. They showed that the performance of their method is comparable with that of Support Vector Machine (SVM).

Rattikorn *et al.* used Multi—Dimensional Ranker (MDR) to analyze microarray data of 11 different types and subtypes of cancer [13]. Huynh *et al.* used the compact single hidden layer feed—forward neural networks (C—SLFNs) trained by an improved extreme learning machine (ELM) algorithm to classify microarray data for cancer diagnosis [14]. They showed that the simple structure of the (C—SLFNs) is faster than other algorithms such as the SVM and Fisher Discriminate Analysis (FDA).

Rao *et al.* worked on ANNs and statistical techniques to identify prostate cancers and classify them using metrics call values [15]. Ziaei *et al.* presented a system for lymphoma cancer classification where genes were ranked based on their signal to noise (S/N) ratios. They used PCA

for more dimensionality reduction. Selected genes were applied to Perceptron neural network for classification [16]. The study was based on 40 patients and 4026 genes.

Wang *et al.* worked on the classification of genomic data using two—layer multilayer perceptions (MLP) [17]. Markus *et al.* provided a survey of the application of machine learning algorithms to classification and diagnosis of cancer based on expressions profiles [18].

Dudoit *et al.* compared the performance of different discrimination methods for the classification of tumors based on gene expression data [19]. The methods included the nearest—neighbor classier, linear discriminant analysis, and classification trees.

Khayat *et al.* proposed a Genetic Algorithm (GA) approach combined with Multilayer Perceptron using Back Propagation (BP) algorithm [20]. The approach was associated with a fuzzy logic —based pre—filtering technique. Lee *et al.* tackled the problem using wavelets for feature extraction and ANNs for classification [21].

Soares *et al.* worked on General Regression Neural Networks (GRNN), in conjunction with a particle Swarm Optimizer to perform microarray classification [22]. Liang presented a method named (X-AI) for accurate cancer classification and the acquisition of knowledge from DNA microarray data [23]. Statnikov *et al.* used 22 diagnostic and prognostic datasets and showed that SVM outperforms random forests in microarray classification [24]. Al Timemy used Self Organizing Map (SOM) for Kidney cancer classification [25].

4. MATERIALS AND METHODS

The dataset used in the training and testing phases was obtained from the SMD database. It consisted of a total of 60 microarrays (30 tumors and 30 non—tumors). Each microarray consisted of a channel 1 image and a channel 2 image. Each image was represented by a matrix of 5524 x 1956 pixels, with 256 grey levels per pixel.

The system was trained with 50 microarrays (25 tumors and 25 non—tumors). The system was then tested with 30 microarrays; 15 microarrays from the training data and 15 microarrays that were not present in the training.

A bloc k diagram of the proposed system is shown in Fig.2

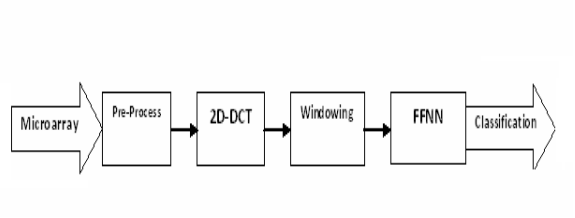


Fig. 2: Block diagram of the proposed system.

As shown in Fig.2, the microarray input image is first preprocessed. Specifically, the microarray channel 1 and channel 2 images are normalized, according to [19] to produce a matrix X, given by the following equation:

$$X = \log_2 \frac{channel\_2}{channel\_1}$$

The 2—D DCT is then applied to the matrix X to produce the matrix C which contains the DCT coefficients. Next, the lower frequency coefficients located in the upper—left corner of the matrix C are extracted using a windowing method.

The extracted coefficients/features are then applied to a FFNN for classification. The flow chart of the system is shown in Fig. 3.

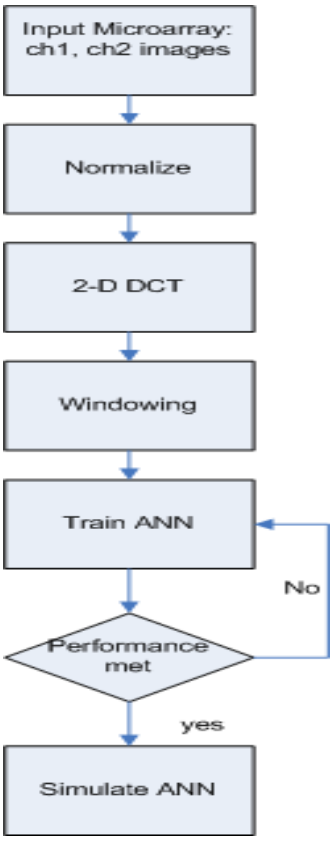


Fig. 3: Proposed System Flow Chart

A. Dimensionality Reduction and the DCT

Some of the methods that have been used in microarray data projection and dimensionality reduction include: the Principal Component Analysis (PCA) [26], the Singular Value Decomposition (SVD) [27] and the Independent Component Analysis (ICA) [28].

The strong capability of the DCT to compress energy makes the DCT a good candidate for pattern recognition applications. Coupled with classification techniques such as Vector Quantization (VQ) and ANN, the DCT can constitute an integral part of a successful pattern recognition system [29, 30].

The DCT has been used in many practical applications, especially in signal compression. For example, the compression achieved in the famous JPEG image format is based on the DCT.

The discrete cosine transform of an N×N image, f(x,y) is defined by

$$F(u,v)=\sum_{x=0}^{(N-1)}\sum_{y=0}^{(N-1)}C(u)C(v)f(x,y)\cos\frac{(2x+1)u\pi}{2N}\cos\frac{(2}{}$$

and the inverse transform is defined by

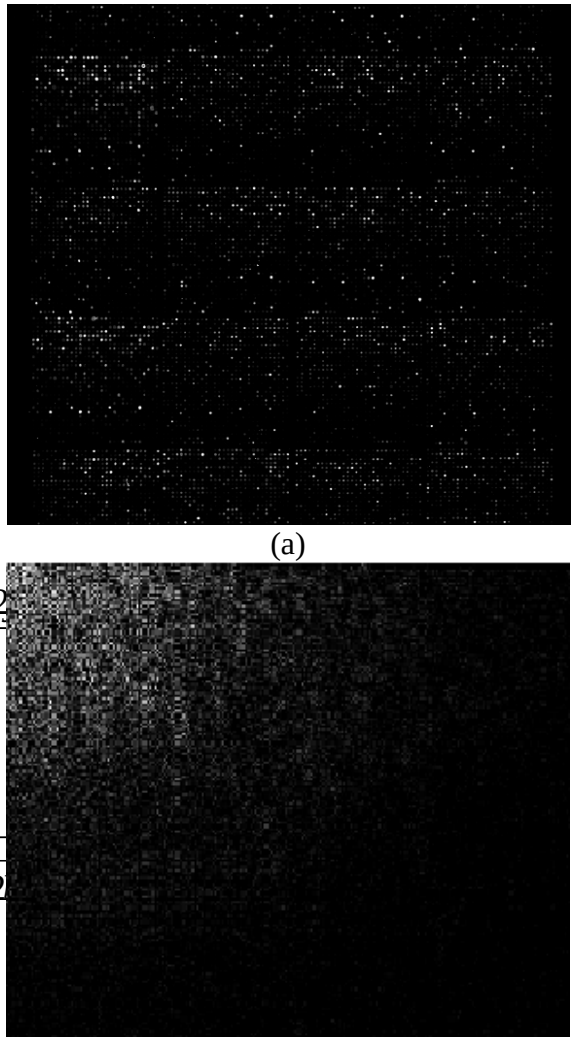
$$f(x,y)=\sum_{u=0}^{(N-1)}\sum_{v=0}^{(N-1)}C(u)C(v)F(u,v)\cos\frac{(2x+1)u\pi}{2N}\cos\frac{(2y-}{2}$$

$$\text{where } C(u)=C(v)=\frac{1}{\sqrt{N}}, \text{for } u=v=0,$$

$$\text{and } C(u)=C(v)=\sqrt{\frac{2}{N}}, \text{for } u,v\neq 0.$$

The DCT decomposes a signal into its elementary frequency components. When applied to an M X N image, the 2—D DCT compresses all the information of the image and concentrates it in a few coefficients located in the upper—left corner of the resulting real—valued M X N DCT matrix.

The energy compactness property of the DDCT is illustrated in Fig. 4 which shows a microarray image, Fig 4(a), and its DCT transform, Fig. 4(b). Note that the transform image has zeros or low—level intensities except at the top left corner where the intensities are very high. These low—frequency, high—intensity coefficients, are therefore, the most important coefficients in the frequency matrix and carry most of the information about the original image.



(a) (b)

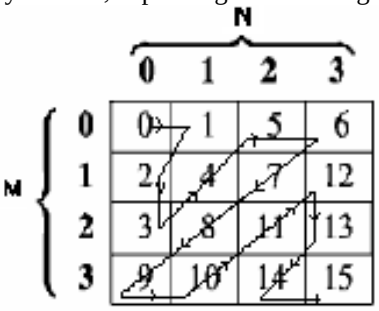
Fig. 4: (a) Microarray image and, and (b) Its DCT transform.

Two methods were followed in this paper to extract features from these low—frequency DCT coefficients. The first method is a square—windowing method that extracts the $k \times k = k^2$ lowest—frequency coefficients in the upper—left corner of the DCT matrix, as shown in Fig.5(b). This windowing method makes use of the fact that the DCT pushes most of the energy/information of the signal in the dc component and the lower frequency components. The dc coefficient contains the highest value or most of the energy. The second

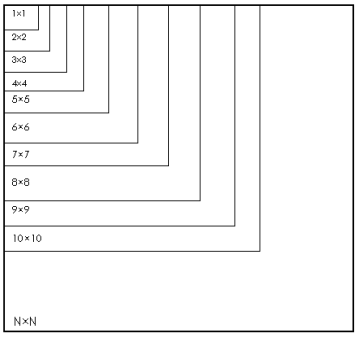
harmonic often has the second highest value, and so on.

To illustrate the scanning scheme of the square window, let c_{mn} designate the coefficient in the DCT matrix located in the m^{th} row and n^{th} column. Then a 1x1 window, generates the 1—element vector $W_{1\times1} = [c_{11}]$. Similarly, a 2x2 window generates the vector $W_{2\times2} = [c_{11}\ c_{12}\ c_{21}\ c_{22}]$ and a 3x3 window produces the vector $W_{3\times3} = [c_{11}\ c_{12}\ c_{13}\ c_{21}\ c_{22}\ c_{23}\ c_{31}\ c_{32}\ c_{33}]$.

The second method is a ziq—zaq method [31], as depicted in Fig. 5(a). Here the coefficients are more selectively scanned, depending on their magnitudes.



(a)



(b)

Fig.5: DCT coefficient extraction techniques: (a) zig—zag method, and (b) Square—window method.

B. ARTIFICIAL NEURAL NETWORKS

ANNs were introduced by McCulloch and Pitts in 1943 [32]. ANNs are trainable algorithms that can learn to solve complex problems from training data that consists of a set of pairs of inputs and desired outputs (targets). They can be trained to perform tasks such as prediction (regression), and classification. ANNs have been applied successfully in many fields including pattern recognition, image processing, and adaptive control.

An ANN consists of interconnected processing elements called neurons that work together to produce an output. Neural networks learning techniques are divided into two categories: supervised learning and unsupervised learning [33]. Supervised learning requires a set of input examples and their corresponding desired outputs. During the training phase, the network is presented with sets of pairs (input and desired output). The network is iteratively updated to reach a desired Mean—Squared—Error (MSE) and optimally provide generalization for test inputs.

One of the most popular ANNs is the FFNN. One of the most popular learning algorithms for the FFNN is the Backpropagation algorithm.

All the ANNs examined in this study had the following specifications in all the simulations:

- FFNN structure.
- Backpropagation as the learning algorithm.
- Logsigmoid functions as the transfer functions for the output layer.
- The output layer always contained a constant number of 2 neurons, which corresponds to the number of classes (tumor and non—tumor).

5. RESULTS

Here we explore the optimum number of DCT coefficients, and optimum ANN structure (number of layers and number of neurons in each layer).

First, we investigate the optimum number of DCT coefficients to be used as features. This is achieved by using a simple 2—layer structure and varying the number of coefficients used. Specifically, the input layer had 5 neurons and the output layer had 2 neurons. Fig. 6 shows the success rate as a function of the number of DCT coefficients used.

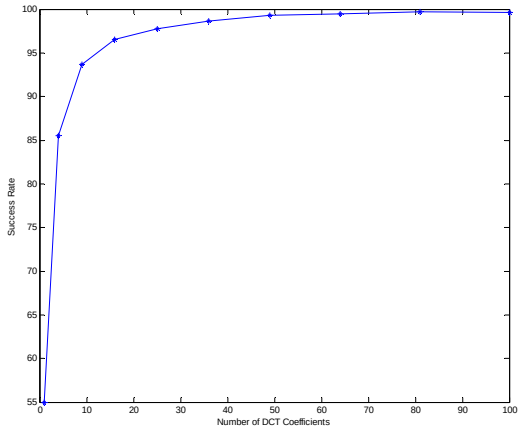


Fig. 6: Percentage Success Rate vs. number of coefficients for a 2—layer ANN.

It is clear from Fig. 6 that as the number of DCT coefficients/features used increases, the success rate increases or levels off. The maximum success rate of 99.7 % occurs, however, when the number of coefficients used is 100, which corresponds to the maximum number of coefficients used, and is obtained using a square window of 10 X 10 coefficients. This finding is consistent with that found by Sarhan for the optimum number of DCT coefficients used for iris recognition [34]. Sarhan showed that the maximum number of DCT coefficients produces the lowest error rates.

The performance of the proposed algorithm was evaluated by computing the percentages of Sensitivity (SE), Specificity (SP) and Accuracy (AC). As given in [25], the respective definitions are as follows:

- Sensitivity: is the fraction of real events that are correctly detected among all real events.
- Specificity: is the fraction of nonevents that has been correctly rejected.

Sensitivity, specificity and accuracy of prediction have been calculated according to the following formulas:

$$SE = \frac{TP \times 100}{(TP+FN)}$$

$$SP = \frac{TN \times 100}{(TN+FP)}$$

$$AC = \frac{(TP+TN) \times 100}{(TN+TP+FN+FP)}$$

where,

FP: Predicts non-tumor as tumor.
TP: Predicts tumor as tumor.
FN: Predicts tumor as non-tumor.
TN: Predicts non-tumor as non-tumor.

The calculated SE, SP and AC are given in Table 1.

No. of cases	SE	SP	AC
60	99.2%	100 %	99.6 %

Table-1: The performance metrics of the proposed method.

From table 1, the obtained accuracy shows that there were no misidentifications, indicating that the proposed system is robust and very reliable.

6. DISCUSSION

The results show that the DCT is capable of mapping the high-dimensionality and complex microarray problem into a simple, manageable-size problem that has low-processing complexity and complete accuracy.

Moreover, the study shows that the DCT coupled with an ANN classifier that has a simple 2-layer structure with a small number of neurons in the first and second layers, was able to produce about a 100% success rate.

7. CONCLUSION

In this paper, a robust system for stomach cancer detection using microarrays is presented. The system consists of a feature—extraction stage followed by an ANN classification stage. The feature extraction stage uses the 2—D DCT to compress the input microarray. Low—frequency components of the DCT array constitute most of the energy/information of the input microarray. These

components were, thus, used as distinctive features and were extracted using a windowing technique.

The paper also investigates through simulations, optimal parameters such as the optimal number of DCT coefficients/features and the optimal ANN structure for the recognition of stomach cancer.

The proposed method produces a success rate of 99.7%. The sensitivity, specificity, and accuracy of the system were found to be equal to 99.2%, 100%, and 99.66% respectively.

Experimental tests on the SMD Database achieved 99.7% of recognition accuracy using only100 DCT coefficients, with a simple 2-layer ANN structure and low computational cost.

REFRENCES:

[1] Cancer Definition http://www.nlm.nih.gov/medlineplus/medlineplus.html.

[2] Sherlock G, Hernandez-Boussard T, Kasarskis A, Binkley G, Matese JC, Dwight SS, Kaloper M, Weng S, Jin H, Ball CA, Eisen MB, Spellman PT, Brown PO, Botstein D, Cherry JM. The Stanford Microarray Database. Nucleic Acids Res 2001 Jan 1; 29(1):152-5.

[3] Kulesh DA, Clive DR, Zarlenga DS, Greene JJ (1987). "Identification of interferon-modulated proliferation-related cDNA sequences".Proc Natl Acad Sci U S A., 84(23): 8453–8457

[4] Schena M, Shalon D, Davis RW, Brown PO (1995). "Quantitative monitoring of gene expression patterns with a complementary DNA microarray,". Science; 270(5235):467-70.

[5] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977): Maximum likelihood from incomplete data via the EM algorithm. J. Royal Stat. Soc. series B, 39: 1-38.

[6] Linde, Y., Buzo, A., and Gray, R. M. (1980): An algorithm for vector quantiser design. IEEE Trans. Commun., COM-28(1):84-95.

[7] Kohonen, T. (1984): Self-organization and associative memory. Berlin, Springer-Verlag.

[8] Gersho, A., and Gray, R. M. (1993): Vector Quantization and Signal Compression. Boston, Kluwer Academic Publishers.

[9] Stanford Microarray Database, http://smd.stanford.edu.

[10] Demeter J, Beauheim C, Gollub J, Hernandez-Boussard T, Jin H, Maier D, Matese JC, Nitzberg M, Wymore F, Zachariah ZK, Brown PO, Sherlock G, Ball CA. The Stanford Microarray Database: implementation of new analysis tools and open source release of software. Nucleic Acids Res 2007 Jan 1;35(Database Issue):D766-770.

[11] T. Y. Yang, "Efficient multi-class cancer diagnosis algorithm, using a global similarity pattern," Computational Statistics & Data Analysis, vol. 53, pp. 756-765, january 2009.

[12] Hang, F. Wu, "Sparse Representation for Classification of Tumors Using Gene Expression Data," Journal of Biomedicine and Biotechnology, 15 March 2009.

[13] H. Rattikorn, K. Phongphun, "Tumor classification ranking from microarray data," BMC genomics journal, vol. 9, pp s21, September 2008.

[14] H. T. Huynh, J. Kim, Y. Won, "DNA Microarray Classification with Compact Single Hidden-Layer FeedForward Neural Networks," Frontiers in the Convergence of Bioscience and Information Technologies (fbit), pp.193-198, 2007.

[15] K.V. G. Rao, P. P. Chand, M.V.R. Murthy, "A neural Network Approach in Medical Decision Systems" Journal of Theoretical and Applied Information Technology, vol. 3 No. 4, 2007.

[16] L. Ziaei, A. R. Mehri, M. Salehi, " Application of Artificial Neural Networks in Cancer Classification and Diagnosis Prediction of a Subtype of Lymphoma Based on Gene Expression Profile," Journal of Research in Medical Sciences, vol. 11, No. 1; Jan. & Feb. 2006.

[17] Zuyi Wang , Yue Wang , Jianhua Xuan , Yibin Dong , Marina Bakay , Yuanjian Feng , Robert Clarke , and Eric P. Hoffman, Optimized multilayer perceptrons for molecular classification and diagnosis using genomic data, Bioinformatics, 22(6): 755-761, 2006.

[18] R. Markus, P. Carsten, "Microarray-based cancer diagnosis with artificial neural

networks," BioTechniques Journal, pp 30-35, 27 Feb 2003.

[19] S. Dudoit, J f Fridlyand, T. P Speed, "Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data," Journal of the American Statistical Association, vol. 97, No. 457, pp. 77-87, 1 march 2002.

[20] O. Khayat, H. R. Shahdoosti, A. J. Motlagh, "A hybrid GA & back propagation approach for gene selection and classification of microarray data," World Scientific and engineering Academy And Society (WSEAS), ED-7, pp. 56-61, 2008.

[21] J. Lee, B. Zee, "Application of wavelet-based neural network on DNA microarray data," Biomedical Informatics Publishing Group, ED-5, pp. 223-229, December 31, 2008.

[22] C. Soares, L. Montgomery, K. Rouse, J. E. Gilbert, "Automating Microarray Classification Using General Regression Neural Network," Seventh International Conference on Machine Learning and Applications (ICMLA), pp. 508-513, 2008.

[23] Liang-Tsung Huang, "An integrated method for cancer classification and rule extraction from microarray data," Journal of Biomedical Science 2009.

[24] Statnikov, Alexander (A); Wang, Lily (L); Aliferis, Constantin F (CF),"A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification,"Journal BMC bioinformatics,2008.

[25] A. H. A. Al Timemy, "Self-Organization Maps for Prediction of Kidney Dysfunction", In Proc. 16th Telecommunications Forum TELFOR, Belgrade, Serbia, 2008.

[26] Misra, J., Schmitt, W., Hwang, D., Hsiao, L., Gullans, S., Stephanopoulos, G., and Stephanopoulos, G. (2002): Interactive Exploration of Microarray Gene Expression Patterns in a Reduced Dimensional Space. Genome Res, 12(7):1112-1120.

[27] Wall, M. E., Rechtsteiner, A., Rocha, L. M. (2003) Singular Value Decomposition and Principal Component Analysis. In .A Practical Approach to Microarray Data Analysis. 91-109. Berrar, D.P., Dubitzky, W., Granzow, M., (eds). Norwell, MA , Kluwer.

[28] Liao, X., Dasgupta, N., Lin, S. M., and Carin, L., ICA and PLS modelling for functional analysis and drug sensitivity for DNA microarray signals. Proc. Workshop on Genomic Signal Processing and Statistics, CP 1-11, 2002

[29] Ahmad M. Sarhan, "A Comparison of Vector Quantization and Artificial Neural Network Techniques in Typed Arabic Character Recognition," International Journal of Applied Engineering Research (IJAER), 4(5):805-817, May, 2009.

[30] A. M. Sarhan, and O. I. Al-Helalat "A Novel Approach to Arabic Characters Recognition Using A Minimum Distance Classifier," In Proceedings of the World Congress on Engineering, London, U.K, July 2007.

[31] Gonzales, R. C., Woods, R. E., 1993. "Digital Image Processing". Addison-Wesley, Reading, Massachusetts.

[32] W. S. McCulloch, and W. H. Pitts, A logical calculus of the ideas immanent in nervous activity, Bulletin of Mathematical Biophysics, 5:115-133 1943.

[33] Haykin, "Neural networks: a comprehensive foundation". New York: Macmillan, 1994.

[34] Ahmad M. Sarhan, "Iris Recognition Using the Discrete Cosine Transform and Artificial Neural Networks," *Journal of Computer Science (JCS)*, 5(4): 369-373, 2009.

Ahmad M. Sarhan, Ph.D. was born in Jordan in 1970. He received a B.Sc. in biomedical engineering form Syracuse University, NY, USA in 1991 and MS. in electrical engineering from Syracuse University in 1992, and a Ph.D. in electrical engineering from University of Dayton, Ohio, USA in 1996. .He is currently an associate professor of Engineering at King Faisal University in Al-Ahsa, Saudi Arabia. He has worked as a faculty member at several universities including Yamouk University, Jordan University, Ajman University. He also worked in the industry as an executive manager of a medical company in the UAE. He received several rewards including, The Best Student Paper Award for the paper "Partition-Based Filters," the Most Outstanding Paper Published in the NAECON Conference, Dayton, Ohio, May 1995. Dr. Sarhan received a Tuition scholarship from Syracuse University and a Teaching and Research Assistantship from the University of Dayton. His main research interests include adaptive signal processing and pattern recognition.