

# Tissue Classification Through Analysis of Gene Expression Data Using A New Family of ART Architectures

Rui Xu<sup>1</sup>, Georgios C. Anagnostopoulos<sup>2</sup> and Donald C. Wunsch II<sup>1</sup>

<sup>1</sup>Applied Computational Intelligence Laboratory  
Dept. of Electrical and Computer Engineering  
University of Missouri - Rolla  
Rolla, MO 65409-0249 USA

<sup>2</sup>School of Electrical Engineering & Computer Science  
University of Central Florida  
Orlando, Florida, USA

rxu@umr.edu, anagnostop@email.com, dwunsch@ece.umr.edu

**Abstract - Correct classification is crucial to cancer diagnosis and treatment. In the paper, we demonstrate that a new family of neural network architectures – Ellipsoid ART and ARTMAP (EA/EAM) can cluster and classify tissues successfully through analysis of gene expression data generated by DNA microarray experiments.**

## I INTRODUCTION

Correct classification is crucial to cancer diagnosis and treatment. Traditional classification methods are mostly dependent on morphological appearance of tumors and their applications are limited by the existing uncertainties [9]. Advance in DNA microarray techniques makes it possible to measure gene expression levels of thousands of genes simultaneously under different cancerous or normal samples [1]. Therefore, it provides a new way for people to understand molecular behaviors in abnormal tissues and make more accurate predictions in cancer diagnosis and treatment [9].

Currently, there are two major microarray technologies based on the nature of the attached DNA (cDNA with length varying from several hundred to thousand bases or oligonucleotides containing 20-30 bases). For both technologies, each DNA microarray consists of a solid substrate to which large amount of DNA molecules are attached according to some certain order. Fluorescently labeled cDNA obtained from RNA samples through the process of reverse transcription is hybridized with the probes on the microarray. Usually, a reference sample with different fluorescent label is needed for the purpose of comparison. Image analysis techniques are then used to measure the fluorescence of each dye and the ratio reflects relative levels of gene expression. Microarray technologies open a door for people to investigate gene activities from the angle of the whole genome. At the same time, they lead to many open issues for computational biologists with large amount of data generated. Baldi describes three levels to represent the complexity of gene expression data analysis [10]. The bottom level studies the activities of single genes under different conditions. The second level focuses on the relations and

interactions between genes. The top level tries to infer the whole genetic network that finally determines all the patterns we observed.

Many unsupervised clustering methods [2-7], supervised learning algorithms [8] and statistical techniques [10-11] have been successfully used in recent years. Eisen et al., apply pairwise average-linkage cluster algorithm to analyze gene expression in the budding yeast *Saccharomyces cerevisiae* [2]. Reference [4] and [5] show other examples of implementation of hierarchical clustering (HC) algorithm. Although HC provides a very informative visualization of the clustered data, it lacks robustness [7] and does not have favorable scalability properties. This is mainly because of its huge memory demands in the case of very large data sets, which are typical of genome expression data clustering problems. Neural networks provide a good solution for gene expression data analysis with their features and capabilities that have already been proven in pattern recognition and many other applications. Brown et al. illustrate the value of Support Vector Machines in classifying genes functionally by using gene expression data [8]. Tamayo et al. construct Self-Organizing Feature Map (SOFM) architectures to cluster gene expression patterns both in yeast cell cycle and in hematopoietic differentiation across four cell lines [7]. While SOFMs [25] enjoy the merits of topological order preservation and input space density approximation [26], convergence-related aspects of its training phase may become serious issues with very large data sets. Additionally, trained SOFMs may be suffering from input space density misrepresentation [26], where areas of low pattern density may be over-represented and areas of high density under-represented.

Recent cancer research based on DNA microarray expression experiments demonstrates the effectiveness of cancer classification by gene expression data [5,9,12-15]. Golub et al. describe cancer classification as two challenges: class discovery and class prediction and use Self-Organizing Map to discriminate two types of human acute leukemias [9]. Alizadeh et al. identify two molecularly distinct forms of

diffuse large B-cell lymphoma by corresponding gene expression profiling [13]. Furthermore, Ross et al. construct a gene expression database to study the relationship between genes and drugs for 60 human cancer cell lines, which provides an important criterion for therapy selection and drug discovery [5].

In the study, we use a new family of neural network architecture – Ellipsoid ART and ARTMAP (EA/EAM) [17] to analyze publicly accessible data sets on cancer research, including AML/ALL leukemia data set and the Colon cancer data set. EA/EAM have the properties of fast, stable and finite learning. They can create nonlinear boundaries by using hyper-ellipsoids to represent the generated categories. In the paper, we demonstrate the potential of EA/EAM in successfully addressing the challenge of processing and then interpreting massive, multidimensional data collections with computational efficiency and satisfying results.

The paper is organized as follows. Section 2 presents a brief introduction to EA/EAM. Section 3 describes the data sets and experimental methods. The results of experiments are presented and discussed in section 4 and section 5 concludes the paper.

## II. ELLIPSOID ART & ARTMAP

Ellipsoid ART (EA) and Ellipsoid ARTMAP (EAM) were first introduced in [17] and are two neural network architectures based on the Adaptive Resonance Theory (ART) developed by S. Grossberg in [19]. While EA utilizes unsupervised learning to cluster unlabeled input patterns, EAM is capable of learning associative maps between an input and an output space. As a special case, when the output space consists of a set of class labels, EAM can be used as a classifier. EA and EAM naturally evolved as a generalization of Hyper-sphere ART (HA) and Hyper-sphere ARTMAP (HAM) presented in [20] and follow the same learning and functional principles of Fuzzy ART (FA) [21] and Fuzzy ARTMAP (FAM) [22].

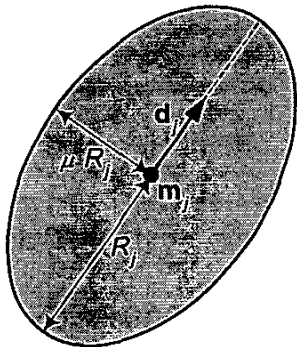


Figure 1: Ellipsoid ART category embedded in a 2-dimensional input space.

FA and FAM accomplish their clustering and mapping tasks by aggregating distributions of data via the use of FA categories. The geometric representations of such categories are hyper-rectangles embedded in the input space. The motivation behind EA/EAM's designs was to develop new types of ART-based neural networks that have similar structure and properties of learning to FA and FAM respectively, while utilizing a different, more effective and efficient geometric shape for the summarization of data.

Both EA and EAM employ EA categories for the task of data aggregation, whose geometric representations are hyper-ellipsoids. Figure 1 shows an example of such a representation for a 2-dimensional input space.

As shown in the figure, each EA category  $j$  is characterized by a collection of descriptive quantities, which are called *template elements*: a location vector  $m_j$ , an orientation vector  $d_j$  and the length  $R_j$  of its major semi-axis (*radius*). During the training phase of EA and EAM, learning is accomplished by creating new categories or by expanding already existing ones. A category's template elements are updated incrementally in the light of new evidence provided by the presentation of input patterns.

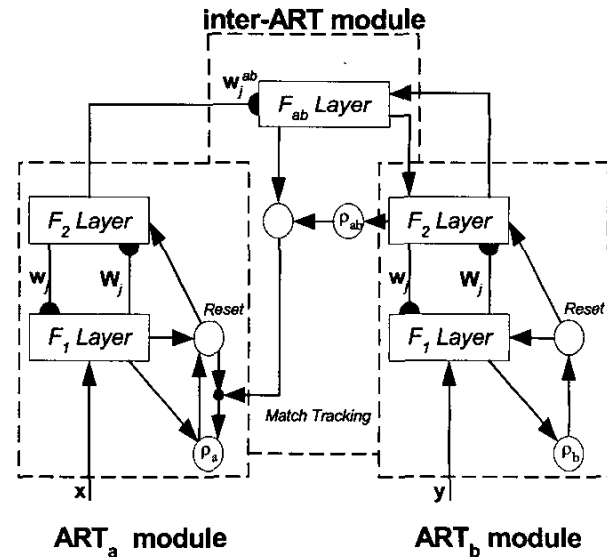


Figure 2: Ellipsoid ARTMAP block diagram

Figure 2 illustrates the block diagram of an EAM network, which resembles the one of FAM. EAM consists of two EA modules ( $ART_a$  and  $ART_b$ ) interconnected via an inter-ART module. The  $ART_a$  module clusters patterns of the input domain and  $ART_b$  the ones of the output domain. The information regarding the input-output associations is stored in the weights  $w_j^{ab}$  of the inter-ART module, while EA category descriptions are contained in the template vectors  $w_j$ . These vectors are the top-down weights of  $F_2$ -layer nodes

in each module. Finally, let us note that an EA network consists of a single, standalone ART module.

Due to their design, EA/EAM inherit all the characteristics and properties of learning of FA/FAM. Among those there are particular features that are very desirable, when it comes to clustering or classification tasks. First, due to their incremental learning nature, EA/EAM are capable of both on-line and off-line (batch) learning. Using *fast learning* [17] in off-line mode both networks stabilize fast in a finite number of epochs. The computational cost of their training phase is relatively low and both networks can cope with large amounts of multidimensional data maintaining the same efficiency. Another important feature of EA/EAM is the capability of detecting atypical patterns during either their training or performance phase. The detection of such patterns is accomplished via the employment of a match-based criterion that decides to which degree a particular pattern matches the characteristics of EA categories that have been formed due to previously experienced inputs. Owing again to their structure, EA and EAM are *transparent* learning machines, meaning that their response to an input pattern can be interpreted and explained in a straightforward manner. This fact contrasts other, *opaque* neural network architectures, for which it is difficult, in general, to explain why an input  $x$  produced a particular output  $y$ . Finally, EA/EAM can be easily implemented as algorithms. The interested reader may find more details regarding EA, EAM and their characteristics in [23].

### III. DATA SETS AND EXPERIMENTS

We use two data sets to test EA/EAM performance in cancer classification. The first data set is the leukemia data set that can be downloaded at [http://www-genome.wi.mit.edu/MPR/data\\_set\\_ALL\\_AML.html](http://www-genome.wi.mit.edu/MPR/data_set_ALL_AML.html). This data set consists of 72 samples, including bone marrow samples, peripheral blood samples and childhood AML cases. These samples are divided as a training set (38 samples) and a test set (34 samples). Among 72 samples, 25 are acute myeloid leukemia (AML) and 47 are acute lymphoblastic leukemia (ALL), which is also composed of two subclasses due to the influences of T-cells and B-cells. The expression levels for 7129 genes (including 312 control genes) were measured across all the samples by high-density oligonucleotide microarrays [9]. The data are finally expressed as the matrix  $E = \{e_{i,j}\}_{72 \times 7129}$ , where  $e_{i,j}$  represents the expression level of gene  $j$  in tissue sample  $i$ . Linear transformation is needed to scale all inputs into the interval  $[0,1]$ , so that EAM can attain better compression. The other data set is available at <http://microarray.princeton.edu/oncology/affydata/index.html>. Gene expressions for more than 6500 genes were measured using oligonucleotide microarrays and 2000 genes with highest minimal intensity were selected [12]. There are 62 colon tissue samples in the data set in which 22 are normal tissues while 40 are tumor ones. The final matrix is in the

form of  $T = \{t_{i,j}\}_{62 \times 2000}$ . We also studied a subset of the colon cancer data set, which includes 18 paired adenocarcinoma samples [15] and available at <http://microarray.princeton.edu/oncology>. 4002 genes were selected with the average expression level equal to or greater than 10.

EAM was utilized to classify two types of leukemia cancer: AML and ALL. Considering that many genes may not contribute to the discrimination of these two types of tumors, we generated another seven subsets by choosing the top 3000, 1000, 500, 100, 50, 10, and 5 genes. Following the criterion used in [9], we can rank genes by calculating the discriminability of each gene as follows:

$$D(i) = \frac{|\mu_{ALL}(i) - \mu_{AML}(i)|}{\sigma_{ALL}(i) + \sigma_{AML}(i)},$$

Where  $\mu_{ALL}(i)$  and  $\mu_{AML}(i)$  are the mean values of gene  $i$  for the samples in class ALL and class AML,  $\sigma_{ALL}(i)$  and  $\sigma_{AML}(i)$  are the standard deviations of gene  $i$  for the samples in class ALL and class AML. Then the value of  $D$  reflects the expression level difference between the two classes for each gene. Gene expresses itself most differently in the classes has the highest score. Top genes were selected according to the standard. Clustering experiments were also performed for the leukemia data set by EA. Top genes were still chosen in this case.

As done in [14], we used the jackknife approach, which is also called leave one out cross validation, to analyze the Colon cancer data set and its subset. For a data set with  $n$  samples, the classifier is trained  $n$  times. Each time, a different single sample is left out as the test point and the other  $n-1$  samples are used to train the classifier. Performance evaluation of the classifier is estimated by considering the average accuracy of the  $n$  cross-validation experiments. In the experiment, performance was measured by examining the number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). True positives are normal samples classified in accordance with their labels. True negatives represent the correct classification for tumor samples. False positives are tumor samples that are misclassified as normal ones and false negatives are the normal samples wrongly placed in the tumor class. Informative genes were also selected as the method described above.

### IV. RESULTS

Table I describes the classification results for the leukemia data set. For the training set, no matter how many genes we selected, EAM can identify all samples in it due to its feature of fast learning. While for the test set, the best result was obtained when we chose the top 100 to top 10 genes. In these

cases, EAM can classify 33 out of 34 test samples correctly. Among all examples, sample AML66 was consistently misclassified as ALL. This shows that AML66 may be an outlier. The same conclusion is derived by other analyses, such as principal component analysis and multi-dimensional scaling algorithm [16].

TABLE I. BEST PERFORMANCE OF EAM FOR CLASSIFICATION OF AML AND ALL

The number of genes	Training accuracy	Test accuracy
7129	38/38	29/34
3000	38/38	30/34
1000	38/38	32/34
500	38/38	32/34
100	38/38	33/34
50	38/38	33/34
10	38/38	33/34
5	38/38	31/34

The results also suggest that it is crucial to select appropriate number of informative genes in the preprocessing phase. Too many or too few genes both will deteriorate the performance of the EAM classifier. Many genes are not related to the ALL/AML classification and including them in the data set will bring noises into the classification system. On the other hand, important information will be wrongly discarded with inadequate genes chosen. Table II lists the top 10 genes with which EAM classifier obtained the highest accuracy.

TABLE II. TOP 10 GENES WITH WHICH EAM CLASSIFIER OBTAINED THE HIGHEST ACCURACY

Gene Accession Number	Gene description
M55150	FAH Fumarylacetoacetate
U50136	Leukotriene C4 synthase (LTC4S) gene
X95735	Zyxin
U22376	C-myb gene extracted from Human (c-myb) gene, complete primary cds, and five complete alternatively spliced cds
M16038	LYN V-yes-1 Yamaguchi sarcoma viral related oncogene homolog
M23197	CD33 CD33 antigen (differentiation antigen)
M84526	DF D component of complement (adipsin)
Y12670	LEPR Leptin receptor
U82759	GB DEF = Homeodomain protein HoxA9 mRNA
D49950	Liver mRNA for interferon-gamma inducing factor (IGIF)

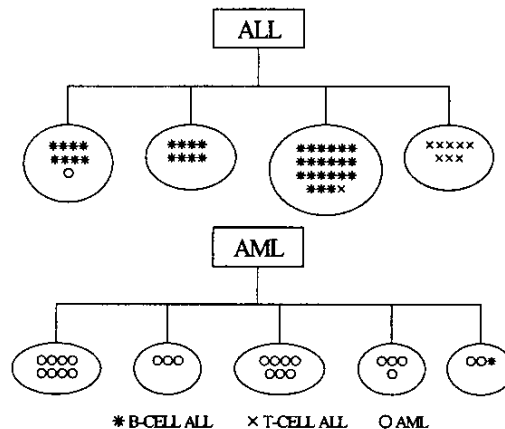


Figure 3: Clusters of ALL and AML by EA

Figure 3 depicts the clusters for the leukemia data set generated by EA with a set of parameters (the number of genes  $n=100$ , the ratio between the length of the major axis and minor axes  $\mu=0.6$ , the learning rate  $\gamma=0.8$ , the vigilance parameter  $\rho=0.6$ , the choice parameter  $\alpha=0.5$ ). In the figure, AML samples are represented as circles and B-cell ALL and T-cell samples are shown as asterisks and crosses, respectively. The ALL samples are grouped into 4 classes while the AML samples form 5 categories. Again, AML66 is misclassified as ALL. Another error occurs when ALL6 is clustered with other two AML samples. From the figure, we also can see that EA can identify two subsets of ALL samples. T-cell samples are all clustered as one category except ALL11 and B-cell samples are distributed in the other three categories.

Results for the Colon cancer data set and the subset are summarized in Table III. In this case, there is just a slight (5%) improvement when we used the reduced-dimensional data set. For the best performance ( $n=50$ ,  $\mu=0.8$ ,  $\gamma=0.14$ ,  $\rho=0.55$ ,  $\alpha=0.9$ ), false positives include tissue samples T30, T33 and T36, and false negatives are normal tissues N8, N34, and N36. This is similar to the clustering results shown in [12], in which three normal samples N8, N12, and N34 and five colon tissues T2, T30, T33, T36 and T37 are placed in the wrong branches of the binary tree. Alon provided an explanation for the observed error through studying the muscle index of each tissue [12]. It is said that normal tissues usually have higher muscle indices than those of tumor tissues. Our results show stronger support for the interpretation. All false positives we obtained have muscle indices (T30: 0.4, T33: 0.7, T36: 0.7) much higher than the average of the rest tumor tissues (0.119), and so for the false negatives (N8: 0.2, N34: 0.2, N36: 0.1, average: 0.626). The results also manifest that EAM can work very well for the colon subset, with 97% classification accuracy for all 4002 genes and 100% for top 50 informative genes. The previous misclassified samples, including T33, N34 and N8 can be identified correctly. This may result from the different

microarray experiment and need biologists to make further analysis.

TABLE III. PERFORMANCE FOR COLON CANCER CLASSIFICATION

Data set	Number of genes	TP	TN	F P	F N	Accuracy
Colon	2000	19	33	7	3	52/62
	50	19	37	3	3	56/62
Subset	4002	17	18	0	1	35/36
	50	18	18	0	0	36/36

## V. CONCLUSIONS

We utilize a new family of neural networks architectures – EA/EAM in tissue classification by analyzing gene expression profiling. The produced experimental results demonstrate that EA/EAM can help to find potential information under these large-scale data sets. These can be very useful for diagnosis of different types of tumors. We also expect that EA/EAM can be used in other tasks like functional classification of genes. Future work will include more experiments with more complex data sets.

## References

[1] M.B. Eisen and P.O. Brown, "DNA Arrays for Analysis of Gene Expression", *Methods Enzymol*, vol. 303, pp. 179-205, 1999.

[2] M.B. Eisen, P.T. Spellman, P.O. Brown and D. Botstein, "Cluster Analysis and Display of Genome-Wide Expression Patterns", *Proc. Natl. Acad. Sci. USA* 95, 14863-14868, 1998.

[3] A. Ben-Dor and Z. Yakhini, "Clustering Gene Expression Patterns", *Proceedings of the Third International Conference on Computational Biology (Recomb 99)*, pp. 33-42, 1999.

[4] P.T. Spellman, G. Sherlock, M.Q. Ma, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein, B. Futcher, "Comprehensive Identification of Cell Cycle-regulated Genes of The Yeast *Saccharomyces Cerevisiae* by Microarray Hybridization", *Mol. Biol. Cell*, 9, pp. 3273-3297, 1998.

[5] Uwe Scherf, Douglas T. Ross, Mark Waltham, Lawrence H. Smith, Jae K. Lee, Lorraine Tanabe, Kurt W. Kohn, William C. Reinhold, Timothy G. Myers, Darren T. Andrews, Dominic A. Scudiero, Michael B. Eisen, Edward A. Sausville, Yves Pommier, David Botstein, Patrick O. Brown, and John N. Weinstein, "A Gene Expression Database for The Molecular Pharmacology of Cancer", *Nature Genetics*, 24(3), pp.236-44, 2000.

[6] S. Tavazoie et al., "Systematic Determination of Genetic Network Architecture", *Nature Genetics*, 22(3): pp. 281-5, 1999.

[7] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E.S. Lander, and T.R. Golub, "Interpreting Patterns of Gene Expression with Self-Organizing Maps: Methods And Application to Hematopoietic Differentiation", *PNAS*, vol.96, pp. 2907-2912, 1999.

[8] M.P. S. Brown, W.N. Grundy, D. Lin, N. Cristianini, C.W. Sugnet, T.S. Furey, M. Ares Jr., and D. Haussler, "Knowledge-based Analysis of Microarray Gene Expression Data by Using Support Vector Machines", *Proc. Natl. Acad. Sci., USA* 97:262-267, 2000.

[9] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M. 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*, 286: 531-537, 1999.

[10] P. Baldi, A. D. Long, "A Bayesian Framework for The Analysis of Microarray Expression Data: Regularized t-test And Statistical Inferences of Gene Changes", *Bioinformatics*, 17:509-519, 2001.

[11] Kerr and Churchill, "Statistical Design And The Analysis of Gene Expression Microarrays", *Genetical Research*, 77:123-128, 2001.

[12] U. Alon, 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", *Proc. Natl. Acad. Sci. USA* 96, pp.6745-6750, 1999.

[13] A. Alizadeh et al., "Distinct Types of Diffuse Large B-cell Lymphoma Identified by Gene Expression Profiling", *Nature*, vol. 403, pp.503-511, 2000.

[14] A. Ben-Dor et al., "Tissue Classification with Gene Expression Profiles", *Proceedings of the Fourth Annual International Conference on Computational Molecular biology*, pp.598-583, 2000.

[15] D.A. Notterman, U. Alon, A. J. Sierk, and A.J. Levine, "Transcriptional Gene Expression Profiles of Colorectal Adenoma, Adenocarcinoma, And Normal Tissue Examined by Oligonucleotide Arrays", *Cancer Research*, 61, pp.3124-3130, 2001.

[16] Xijin Ge et al., "Modified Multi-dimensional Scaling (MDS) Algorithm for Mining Gene Expression Patterns", *CAMDA'00*, 2000.

[17] G. C. Anagnostopoulos and M. Georgiopoulos, "Ellipsoid ART and ARTMAP for Incremental Unsupervised and Supervised Learning", *Proceedings of the IEEE-INNS-ENNS Intl. Joint Conf. on Neural Networks (IJCNN'01)*, Vol. 2, pp.1221-1226, Washington, Washington D. C., 2001.

[18] [http://www.geocities.com/g\\_anagnostop/](http://www.geocities.com/g_anagnostop/)

[19] S. Grossberg, "Adaptive Pattern Recognition and Universal Encoding II: Feedback, Expectation, Olfaction, and Illusions", *Biological Cybernetics*, Vol. 23, pp. 187-202, 1976.

[20] G.C. Anagnostopoulos and M. Georgiopoulos, "Hypersphere ART and ARTMAP for Unsupervised and Supervised Incremental Learning", *Proceedings of the IEEE-INNS-ENNS Intl. Joint Conf. on Neural Networks (IJCNN'00)*, Vol. 6, pp.59-64, Como, Italy, 2000.

[21] G.A. Carpenter, S. Grossberg, and D.B. Rosen, "Fuzzy ART: Fast Stable Learning and Categorization of Analog Patterns by an Adaptive Resonance System", *Neural Networks*, vol.4, pp759-771, 1991.

[22] G.A. Carpenter, S. Grossberg, N. Markuzon, J.H. Reynolds and D.B. Rosen, "Fuzzy ARTMAP: A Neural Network Architecture for Incremental Supervised Learning of Analog Multidimensional Maps", *IEEE Transactions on Neural Networks*, 3(5), pp. 698-713, 1992.

[23] G.C. Anagnostopoulos, "Novel Approaches in Adaptive Resonance Theory for Machine Learning", *Doctoral Dissertation, University of Central Florida*, 2001.

[24] R. O. Duda, P. E. Hart and D. G. Stork, *Pattern Classification*, 2<sup>nd</sup> Ed., Wiley & Sons, New York, 2001.

[25] T. Kohonen, *Self-Organizing Maps*, Springer, Berlin, 1997.

[26] S. Haykin, *Neural networks: A Comprehensive Foundation*, 2<sup>nd</sup> Ed., Prentice Hall, NJ, 1998.