

# Diagnostic Classification of Cancer Using DNA Microarrays and Artificial Intelligence

BRADEN T. GREER AND JAVED KHAN

*Advanced Technology Center, National Cancer Institute, National Institutes of Health, Gaithersburg, Maryland 20877, USA*

**ABSTRACT:** The application of artificial intelligence (AI) to microarray data has been receiving much attention in recent years because of the possibility of automated diagnosis in the near future. Studies have been published predicting tumor type, estrogen receptor status, and prognosis using a variety of AI algorithms. The performance of intelligent computing decisions based on gene expression signatures is in some cases comparable to or better than the current clinical decision schemas. The goal of these tools is not to make clinicians obsolete, but rather to give clinicians one more tool in their armamentarium to accurately diagnose and hence better treat cancer patients. Several such applications are summarized in this chapter, and some of the common pitfalls are noted.

**KEYWORDS:** artificial neural networks; support vector machines; artificial intelligence; microarray

## INTRODUCTION

The widespread applications of cDNA and oligonucleotide microarrays, with whole-genome expression scanning at >40,000 clones in a single experiment, have heralded a new era of molecular genomics and, at the same time, are generating vast amounts of data. There is promise of accurate diagnosis and prognosis, identification of therapeutic targets, characterization of yet unknown genes, and personalized chemotherapy guided by molecular expression signatures. Classification based on gene expression signatures continues to be challenging in part due to the varied microarray platforms, labeling methods, scanners, image analysis tools, as well as classification algorithms currently available. Huge strides have been made over the past few years to increase the quantity of clones printed on an array, and the quality of RNA extraction, amplification, hybridization, and cDNA printing. Additionally, an ever-increasing array of algorithms has been reported to analyze the data produced from these high-quality, dense arrays.

Artificial intelligence (AI) is the term used to describe the ability of a computer or machine to perform activities or make decisions that normally require human

Address for correspondence: Javed Khan, Advanced Technology Center, National Cancer Institute, National Institutes of Health, Room 134E, 8717 Grovemont Circle, Gaithersburg, MD 20877. Voice: 301-435-2937; fax: 301-480-0314.  
khanjav@mail.nih.gov

**Ann. N.Y. Acad. Sci. 1020: 49–66 (2004). © 2004 New York Academy of Sciences.  
doi: 10.1196/annals.1310.007**

**TABLE 1. Examples of unsupervised, supervised, and supervised learning data analysis methods**

Unsupervised	Supervised	Supervised learning
<i>k</i> -means	<i>t</i> test	ANN
Self-organizing maps	ANOVA	SVM
Hierarchical clustering	Golub	
KNN	WGA	
PCA	Wilcoxon	
MDS	Kruskal-Wallis	
Reshuffling	TNoM	

intelligence. AI has been successfully applied to problems ranging from distinguishing four types of small round blue-cell tumors<sup>1</sup> to predicting estrogen receptor status in breast cancer<sup>2</sup> and a host of other applications,<sup>3-8</sup> several of which will be detailed later in this chapter. AI affords many significant benefits over its simplistic clustering counterparts such as hierarchical clustering, *k*-means clustering, and traditional statistical methods.

### CLUSTERING VERSUS MACHINE LEARNING

Simple unsupervised clustering methods such as hierarchical clustering, principal component analysis (PCA), and multidimensional scaling (MDS) allow the visualization of data for the purpose of class discovery, or finding hitherto unknown relationships between samples or genes. On the other hand, supervised methods (see TABLE 1) allow for class prediction and are useful tools that can cluster data sets into meaningful groups and identify genes using a priori knowledge of the data such as stage, diagnosis, tissue type, etc. (see refs. 9 and 10 for good reviews). AI methods such as artificial neural networks (ANNs) or support vector machines (SVMs) are highly specialized forms of supervised clustering. Generally, simple clustering methods weight each input feature (or gene) the same, while ANNs and SVMs have the ability to weight input features according to their relevance to the classification scheme as determined through the learning process. The simpler methods cluster samples based on the summation of all of the inputs, while ANNs and SVMs cluster samples based on the collective effect of all of the input features. In addition, while most of the simpler methods are linear, ANNs and SVMs can learn nonlinear features of the input data. In most of the simpler clustering methods, either a sample belongs to a cluster or it doesn't; in contrast, in ANNs and SVMs, a continuous variable (i.e., an average vote) predicts whether or not a sample belongs to a particular cluster. This gives the ANN or SVM the freedom to conclude that a sample does not belong to any of the known classes if its vote lies too close to the decision boundary. This is advantageous, for example, if in a set of blinded test samples there exist samples that belong to none of the known categories and have been added to determine the specificity of the network.<sup>1</sup> Another limitation of most of the nonlearning methods is the way they treat more than two classification groups: in a one-verse-all

manner. ANNs (but not SVMs) can simultaneously classify a sample set with more than two a priori groups directly without having to resort to a one-verse-all method. For the rest of the chapter, we will turn our attention to the technical aspects and application of two particular forms of machine learning: feed-forward ANNs and feed-forward SVMs.

### ARTIFICIAL NEURAL NETWORKS

ANNs are computer algorithms that model mammalian systems of decision making using the mammalian neuron as a fundamental unit. If one thinks of “the brain [as] a complex, nonlinear, and parallel computer (information processing system)”,<sup>11</sup> then it is logical to want to mimic the brain’s capacity to learn from experience and make decisions. A neuron’s ability to respond to experiences and hard-wire itself accordingly, thus adapting to its local environment, is commonly referred to as the *plasticity* of the brain.

A neural network is a massively parallel distributed processor made up of simple processing units, which has a natural propensity for storing experiential knowledge and making it available for use. It resembles the brain in two respects:

- (1) knowledge is acquired by the network from its environment through a learning process;
- (2) interneuron connection strengths, known as synaptic weights, are used to store the acquired knowledge.<sup>11</sup>

Neural networks are powerful because of their massively parallel, distributed structure and ability to learn and generalize. *Generalization* is the ability of a neural network to predict an input that was not used for training.<sup>11</sup> Generalization is vital to the utility of a neural network. If generalization is not possible, then AI has done little more than put known groups of samples into their corresponding groups. In this way, it offers little more than simple hierarchical or *k*-means clustering. Several properties of neural networks make them particularly powerful for the study of microarray data. The nonlinearity property of a neural network allows it to learn and adapt to nonlinear signals. This is important if the underlying signal is nonlinear, that is, the effect of one gene on another may be in a nonlinear fashion due to a positive feedback loop via a particular protein or transcriptional regulator. Input-output mapping refers to a neural network’s ability to modify its synaptic weights through observing a set of training samples to minimize the error between the input classes and the predicted output classes. The concepts of *adaptivity* and *error-minimization* are what make this input-output mapping possible.

A special form of feed-forward machine learning called support vector machines (SVMs) has some interesting differences from feed-forward ANNs. In essence, the goal of an SVM is to calculate the hyperplane in *n*-dimensional feature space that optimally separates two groups of samples. Feature space is a mapping of the original data to a higher dimensional space. A sample is classified based on where it lies in relation to the hyperplane. When more than two classes are being examined, SVMs employ a one-verse-all strategy in conjunction with the hyperplane. In addition, the robustness of classification can be determined, in part, by the distance between each sample and the hyperplane in *n*-dimensional space. It has been said

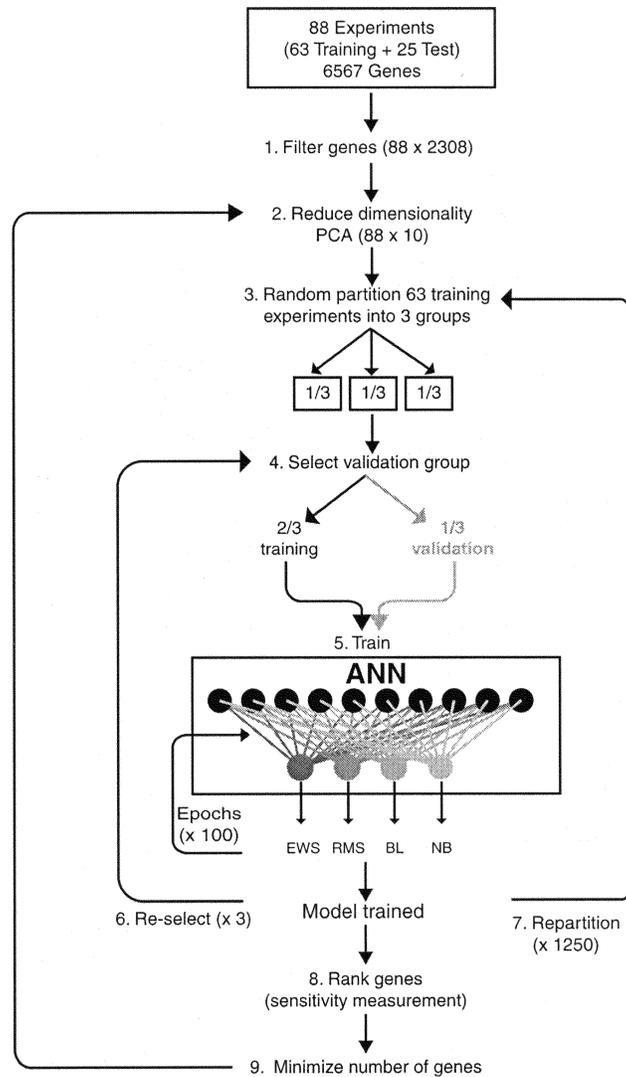
that SVMs can overcome the “curse of dimensionality”. This is the unfavorable situation where the number of tunable parameters is much larger than the number of training samples and can lead to overfitting of the data and thus a nongeneralizable classifier. Both SVMs and ANNs have been used in various contexts for pattern recognition, including imaging, ECG, and biometric pattern recognition (e.g., voice, retina, and palm).

## APPLICATIONS

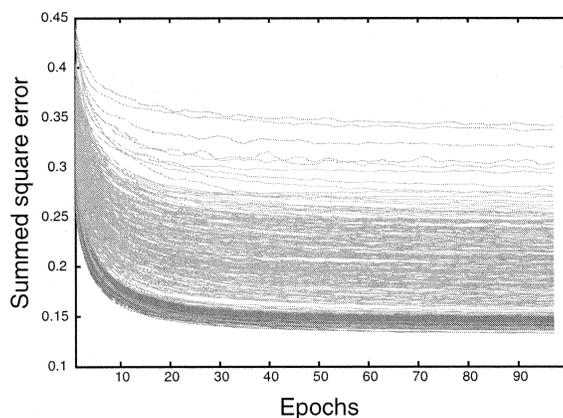
### *Classification of Small Round Blue-Cell Tumors*

Khan *et al.*<sup>1</sup> performed the proof-of-principle study for predicting unknown tumor samples using ANNs on gene expression data from a set of four small round blue-cell tumors (SRBCTs). SRBCTs are a group of tumors that are difficult to diagnose by routine histology and thus pose a significant challenge to the clinician for making an accurate diagnosis and subsequent recommendation for therapy. Accurate diagnosis is crucial because treatment, responses, and prognosis vary greatly depending on the diagnosis. In this study, the training set for the ANNs consisted of SRBCT tumors and cell lines across four tumor types: neuroblastoma (NB), rhabdomyosarcoma (RMS), a subset of non-Hodgkin lymphoma (NHL), and the Ewing’s family of sarcomas (EWS).

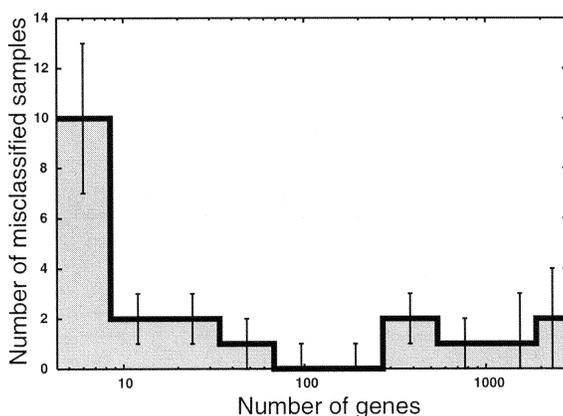
In this paper, they suggest a schema for the application of ANN for diagnostic classification (FIG. 1a). Genes are first filtered for quality based on a quality metric calculated by the image analysis software used to extract data from the scanned microarray images. To avoid the “curse of dimensionality”, PCA is employed and the first  $n$ -components can be used to train the network (in the manuscript, 10 components were used). To further ensure that the data are not overfitted, a cross-validation scheme is employed as follows. The training samples are randomly partitioned into groups, and one of these is used for validation and the rest for training. By this way, several hundred network models are trained (FIG. 1b) and an average vote can be calculated. Because a DNA microarray will typically monitor thousands of genes, and a particular disease will generally affect the expression of on the order of tens or hundreds of genes, the majority of the genes on a chip will not experience significant change in expression. This means that as much as 90% of the measurements on a chip represent noise rather than meaningful biological signal. Additionally, the decision-making process of ANNs is often considered as a “black box” with little information available as to how models determine a particular “expression signature”. For these reasons, they developed a method to identify the genes that contribute most to a particular classification. After training the network as described above, the sensitivity for each gene was calculated as the derivative of the output with respect to the gene expression input. Genes with high sensitivity to the classification scheme received a high rank and genes with low sensitivity to the classification received a low rank. This can be understood functionally as the effect that a perturbation of a particular gene’s expression ratio will have on the classification of each sample. If the perturbation of a particular gene affects the classification result significantly, then that gene receives a high rank. In contrast, if the perturbation of a particular



**FIGURE 1a.** Example of application of ANN (with permission from ref. 1). The 88 experiments were quality filtered (1) and the dimension of the data set further reduced from 2308 to 10 by PCA (2). Next, the 63 training samples were randomly partitioned into 3 groups (3) and 1 of these groups was selected for validation (4). The network was trained for 100 epochs using the 2 remaining groups (5). The samples in the validation group were tested and a different group was selected for validation (6). This process (steps 4–6) was repeated until each group was used for validation exactly one time. Then, the data were repartitioned into 3 new random groups (3) and steps 4–6 repeated again. In total, the data were repartitioned 1250 times (7), generating 3750 trained models. After this procedure, the genes were ranked using the sensitivity measurement (8), increasing numbers of the top-ranking genes were used for training (steps 2–6), and the gene set that produced the minimal number of errors (9) was used to calibrate the ANNs for testing the 25 blinded samples.

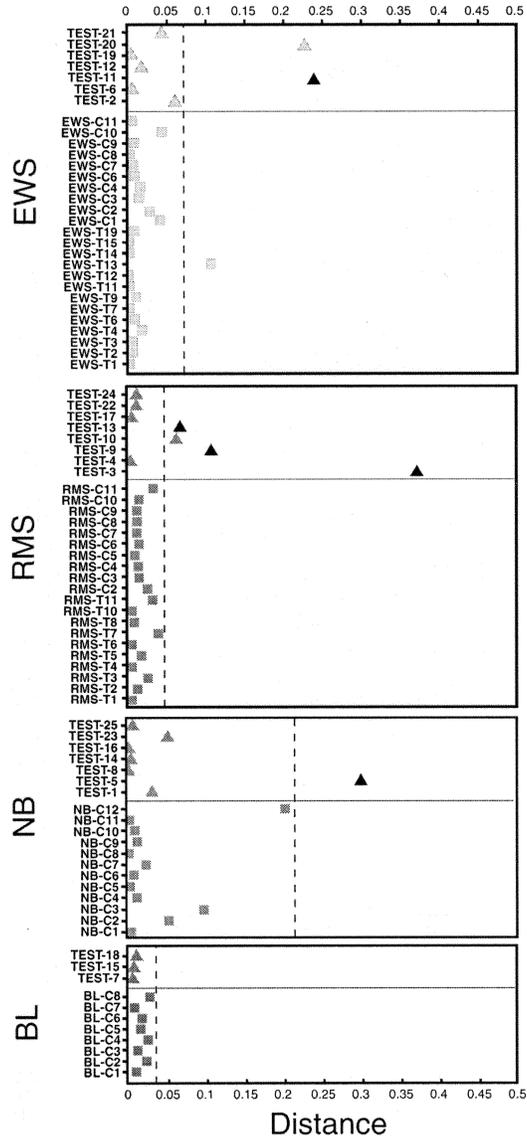


**FIGURE 1b.** Training error results from step 5 of FIGURE 1a. A plot of the classification error with increasing training epochs. The *light gray lines* represent the error of the validation samples and the *darker lines* represent the classification error of the training samples. The consistent decrease in error over increasing epochs implies that overfitting of the data did not occur.

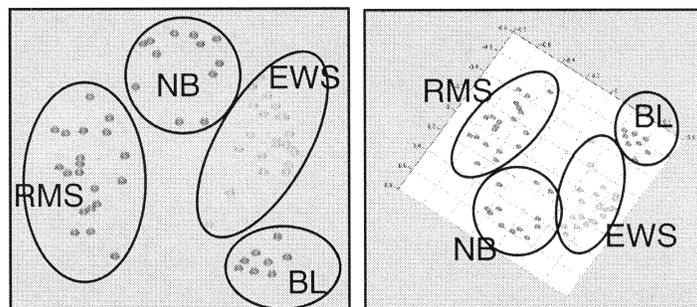


**FIGURE 1c.** Gene minimization results from step 9 of FIGURE 1a. This is a plot of the average number of misclassifications when increasing numbers of genes were used: the number of misclassifications minimized at 96 and 192 genes. The top-ranking 96 genes were used to calibrate the neural networks for subsequent training and testing of the 25 blinded test samples.

gene has a minimal effect on the classification result, then it receives a low rank. Then, in order to determine the best number of high-ranking genes, a gene minimization strategy was employed in which the network was trained with increasing numbers of the top genes (i.e., 6, 12, 24, 48, 96, 192, 384, 768, 1536, and 2308 genes) and the performance of the network determined using the a priori information about the training samples. When the network was trained using the top 96 genes, the average number of misclassifications was near zero ( $<0.5$ ) (FIG. 1c). They clas-



**FIGURE 2.** Classification and diagnosis of the SRBCT samples (with permission from ref. 1). The  $x$ -axis is the Euclidean distance between an ideal ANN output vote and the observed average vote. The vertical dashed line represents the empirical 95 percentile boundary beyond which diagnosis is not confident. Testing samples are represented by triangles, and training samples are shown as squares. Black triangles are the non-SRBCT samples not associated with any of the diagnostic categories. Two testing samples are correctly diagnosed, but lie outside the 95 percentile boundary (Test20-EWS and Test10-RMS). Only one training sample (EWS-T13) lies outside the 95 percentile boundary. All 5 non-SRBCT samples lie outside the 95 percentile boundary as they should.



**FIGURE 3a.** MDS using the top 96 discriminating genes (with permission from ref. 1). Two views of the MDS results depicting the difference in gene expression of the four SRBCT classes.

sified all the training samples correctly with >98% accuracy and correctly diagnosed >90% of the blind test samples, which included 5 non-SRBCT samples (see FIG. 2). Based on all training and testing samples, the sensitivity of the network was 93% for EWS, 96% for RMS, and 100% for both NB and BL. The specificity was 100% for all four SRBCT categories. Many of these 96 genes were not previously known to be associated to the four cancers examined. MDS using these 96 genes is shown in FIGURE 3a. Hierarchical clustering of the samples and the top 96 genes is shown in FIGURES 3b and 3c. One gene, *FGFR4*, a tyrosine kinase receptor, is highly expressed in RMS and could be useful as a therapeutic target. Thus, they demonstrated that a diagnostically sound classifier of cancer could be achieved using microarray technology and ANNs and that meaningful biology could be extracted through ANN gene ranking. Since this study was published, it is worth noting that the microarray technology used, from the RNA preparation to the cDNA printing to the image scanning, has been significantly improved upon. cDNA chips, for example, have gone from parallel monitoring of ~6k clones to ~40k–100k clones. In addition, the improvement of microarray scanner technology has brought much greater consistency to the scanning process. This study marks the first successful attempt at using ANNs to classify cancers according to their gene expression profiles.

#### *Estrogen Receptor Status in Breast Cancer*

In another study, a very similar ANN approach (see FIG. 4a) was used to classify breast cancers based on their estrogen receptor status.<sup>2</sup> In this study, Gruvberger *et al.* analyzed the gene expression profiles of 58 node-negative breast carcinomas using a 6.7k cDNA microarray: 47 samples were used to train the network and 11 samples were used as test samples to verify the universality of the classifier. They were able to predict all 47 training and 11 testing samples with 100% accuracy using the top 100 genes (FIG. 4b and TABLE 2). When they repeated the classification procedure excluding the ER gene, 1 of the 47 training samples was incorrectly classified, but all of the 11 test samples remained correctly classified. This led them to look at how far down the list of discriminating genes they could go and still obtain accurate sample prediction. When they used the genes that ranked between 301 and 400, the

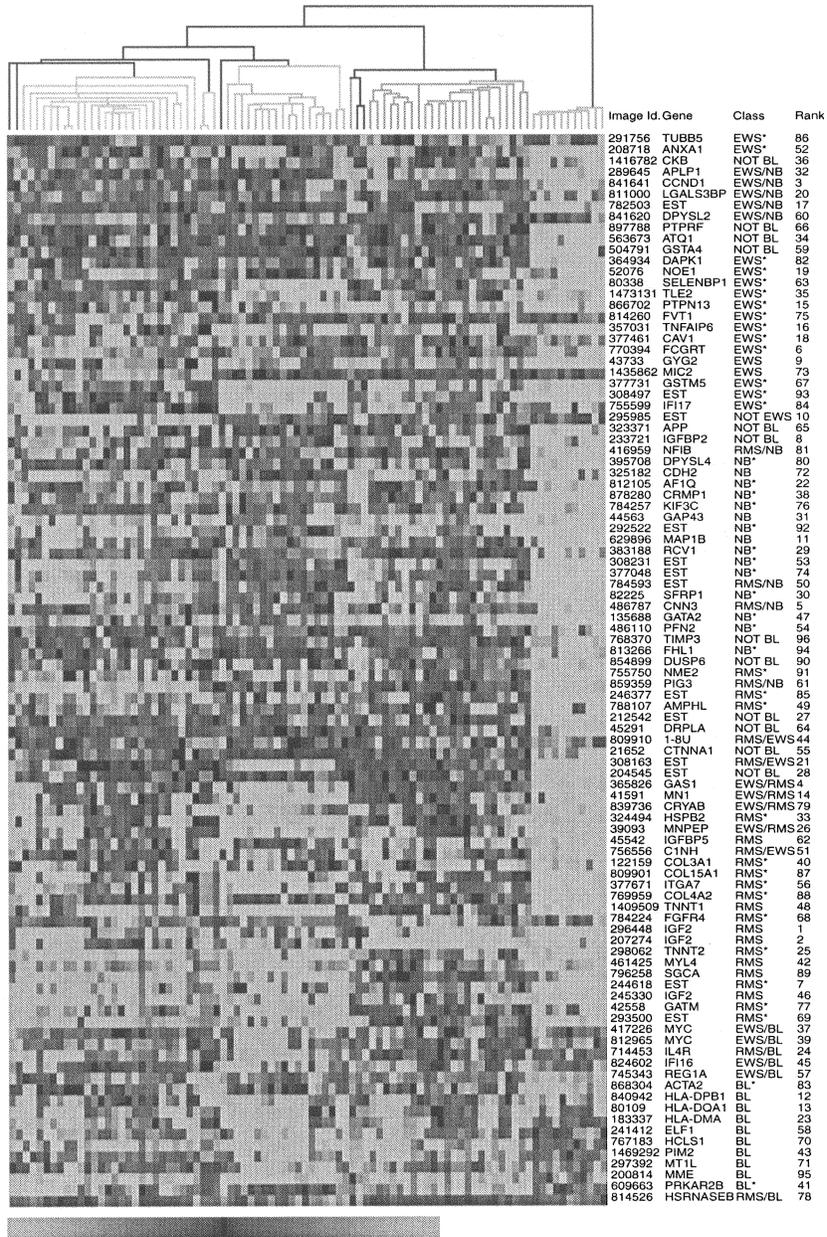
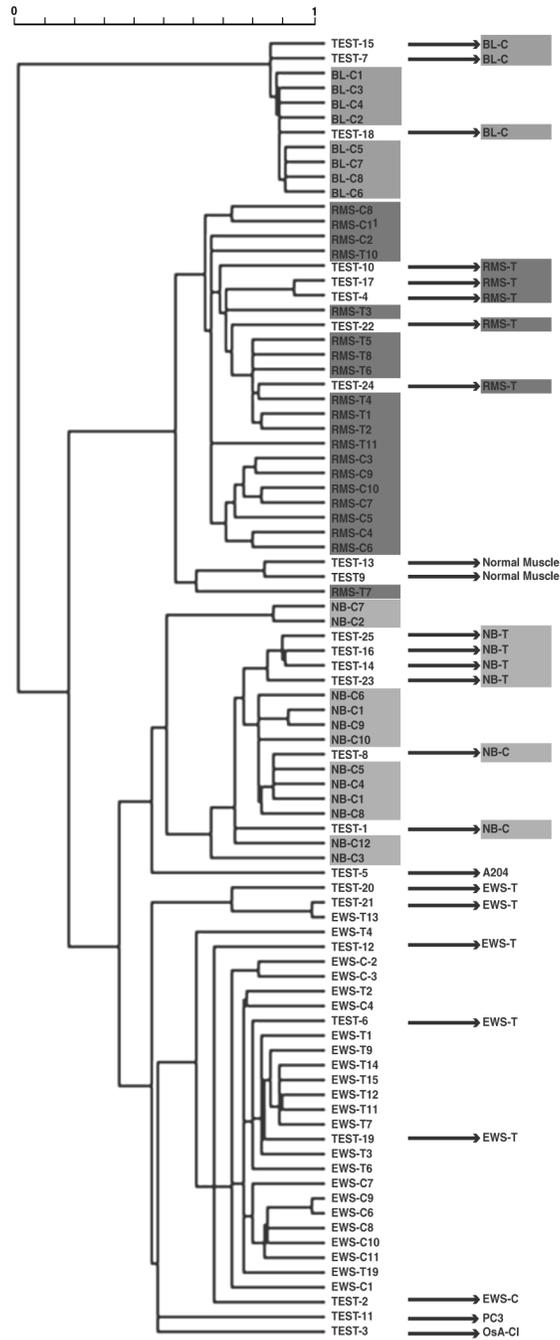
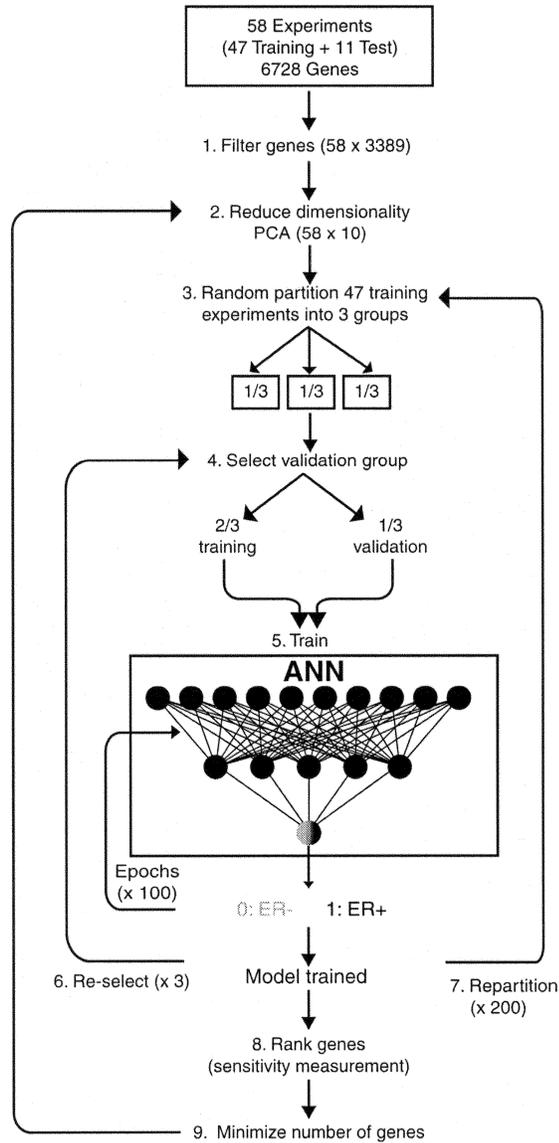


FIGURE 3b. Hierarchical clustering using the top 96 discriminating genes (with permission from ref. 1). Hierarchical clustering and heatmap of genes and samples with dendrogram colored according to clinical diagnosis.

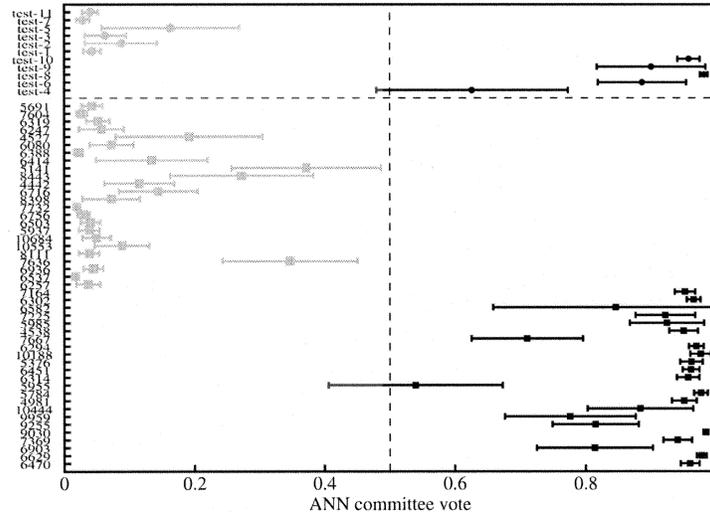


**FIGURE 3c.** Hierarchical clustering using the top 96 discriminating genes (with permission from ref. 1). Enlargement of the sample dendrogram in FIGURE 3b. All 63 training samples were correctly clustered within their diagnostic categories.

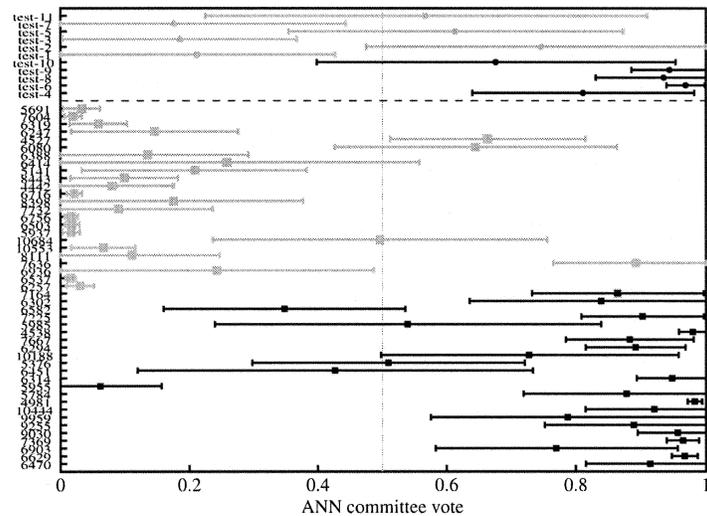


**FIGURE 4a.** ER status ANN work flow (with permission from ref. 2): 6728 genes from 58 experiments were filtered for quality, reducing the number of genes to 3389 (1). Dimensionality was reduced by using the top 10 components from PCA (2). The 47 training samples were randomly partitioned into 3 groups (3). Two groups were used for training, while the third was held aside for validation (4). The multilayer perceptron ANN was trained with 5 nodes in the hidden layer for 100 epochs (5). A different third (see step 4) was selected for validation, and training (5) was performed again. Steps 4–6 were performed one more time such that each third was chosen for validation once. After this, steps 3–6 were repeated 199 more times, making a total of 200 random partitions (7). The genes were ranked for their sensitivity to the classification scheme (8). Next, the number of top-ranking genes required to classify accurately was determined through a gene minimization process using increasing numbers of the top-ranking genes (9). Finally, these selected genes were taken through steps 2–7.

partitioned into 3 groups (3). Two groups were used for training, while the third was held aside for validation (4). The multilayer perceptron ANN was trained with 5 nodes in the hidden layer for 100 epochs (5). A different third (see step 4) was selected for validation, and training (5) was performed again. Steps 4–6 were performed one more time such that each third was chosen for validation once. After this, steps 3–6 were repeated 199 more times, making a total of 200 random partitions (7). The genes were ranked for their sensitivity to the classification scheme (8). Next, the number of top-ranking genes required to classify accurately was determined through a gene minimization process using increasing numbers of the top-ranking genes (9). Finally, these selected genes were taken through steps 2–7.



**FIGURE 4b.** ER status ANN results (with permission from ref. 2): ANN average voting results and standard deviations (*solid lines*) using the top 100 genes. An average vote (*x-axis*) of 1 represents the ideal vote for ER+ (*black*) and an average vote of 0 represents the ideal vote for ER- (*gray*). The decision threshold is 0.5. The 47 training samples are below the *dashed black horizontal line*, and the 11 test samples are located above this line.



**FIGURE 4c.** ER status ANN results (with permission from ref. 2): ANN average voting results and standard deviations (*solid lines*) using the top 301–400 genes. An average vote (*x-axis*) of 1 represents the ideal vote for ER+ (*black*) and an average vote of 0 represents the ideal vote for ER- (*gray*). The decision threshold is 0.5. The 47 training samples are below the *dashed black horizontal line*, and the 11 test samples are located above this line.

TABLE 2. Prediction of ER status

Genes	Validation ( $n = 47$ )		Test ( $n = 11$ )	
	Correct	ROC area	Correct	ROC area
Top 100	47	100.00%	11	100.00%
51–150	43	97.80%	9	100.00%
101–200	45	99.30%	11	100.00%
151–250	44	97.50%	9	100.00%
201–300	41	93.70%	11	100.00%
251–350	39	95.30%	9	93.30%
301–400	41	93.10%	8	96.70%
Random	$38.8 \pm 0.2$	$91.8 \pm 0.2\%$	$5.5 \pm 0.2$	$53.0 \pm 2.6\%$

NOTE: Adapted with permission from ref. 2.

classification still had ROC areas of 93.7% for the training set and 96.7% for the test set (FIG. 4c and TABLE 2). They did report, however, that the ANN committee votes when using this list of genes were closer to the decision boundary and thus should be invested with less confidence. By this experiment, they demonstrated that some of the information content was found in the 301–400 range of discriminators. Using the weighted-gene analysis (WGA) method,<sup>12</sup> the authors discovered a set of 113 genes that were able to classify the samples with 96% accuracy using hierarchical clustering (FIG. 5b). The MDS of the samples using the 113 genes is also shown in FIGURE 5a. Using ANNs, this study showed that ER+ and ER– tumors have very different gene expression profiles with the ability to accurately predict even when the ER gene and the top-ranking genes are removed from the analysis.

#### *Other Applications of Machine Learning to Cancer Classification*

West *et al.*<sup>6</sup> used Bayesian regression modeling to develop a classifier to predict both the estrogen receptor status and the categorized lymph node status of primary breast tumors. Bayesian modeling does not assign a sample to a particular class, but rather assigns a probability that each sample belongs to each output class. Lymph node status is the single most important prognostic indicator for breast cancer,<sup>13</sup> and estrogen receptor status has received attention as a factor in breast cancer development and progression.<sup>4,6,14,15</sup> This study was published only one month after Gruberger's ER paper<sup>2</sup> and represents further proof that both ER and clinical status can be predicted using gene expression profiling.

Van't Veer *et al.*<sup>5</sup> used a leave-one-out supervised correlation-based method to predict clinical outcome of breast cancer. They defined "poor prognosis" ( $n = 34$ ) as patients who were lymph node–negative, but developed distant metastases within 5 years (mean time to metastasis was 2.5 years), and "good prognosis" ( $n = 44$ ) as patients that were lymph node–negative and did not develop distant metastases within 5 years (mean follow-up time of 8.7 years). The authors first selected ~5000 genes from the 25,000 genes measured using an unsupervised selection method (i.e., two-fold regulation and  $P$  value of less than .01 in more than 5 tumors). In one of their

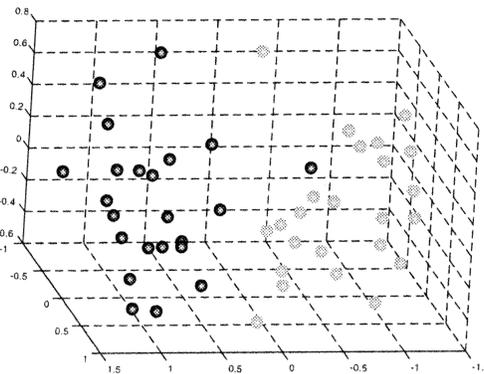
**TABLE 3. Breast cancer patients eligible for adjuvant systemic therapy**

Consensus	Patient group		
	Total patient group ( <i>n</i> = 78)	Metastatic disease at 5 years ( <i>n</i> = 34)	Disease-free at 5 years ( <i>n</i> = 44)
St. Gallen	64/78 (82%)	33/34 (97%)	31/44 (70%)
NIH	72/78 (92%)	32/34 (94%)	40/44 (91%)
Prognosis profile <sup>a</sup>	43/78 (55%)	31/34 (91%)	12/44 (27%) [18/44 (41%) <sup>b</sup> ]

NOTE: The conventional consensus criteria are as follows: tumor  $\geq 2$  cm, ER-, grade 2–3, patient < 35 years (either one of these criteria; St. Gallen consensus); tumor > 1 cm (NIH consensus). Table adapted with permission from ref. 5.

<sup>a</sup>Number of tumors having a poor prognosis signature using microarray profile, defined by the optimized sensitivity threshold in the 70-gene classifier.

<sup>b</sup>Number of tumors with a poor prognosis signature in the group of disease-free patients when the cross-validated classifier is applied.



**FIGURE 5a.** ER+ (*black*) and ER- (*gray*) clustering using the 113 genes selected by weighted-gene analysis (WGA) (with permission from ref. 2): MDS plot of the 47 training samples. The distance between each of the samples represents their approximate degree of correlation.

**FIGURE 5b.** ER+ (*black*) and ER- (*gray*) clustering using the 113 genes selected by weighted-gene analysis (WGA) (with permission from ref. 2): Hierarchical clustering of the samples (*in columns*) and genes (*in rows*). The cluster of genes denoted “ER Cluster” are those genes that clustered with the ER gene (*ESR1*).

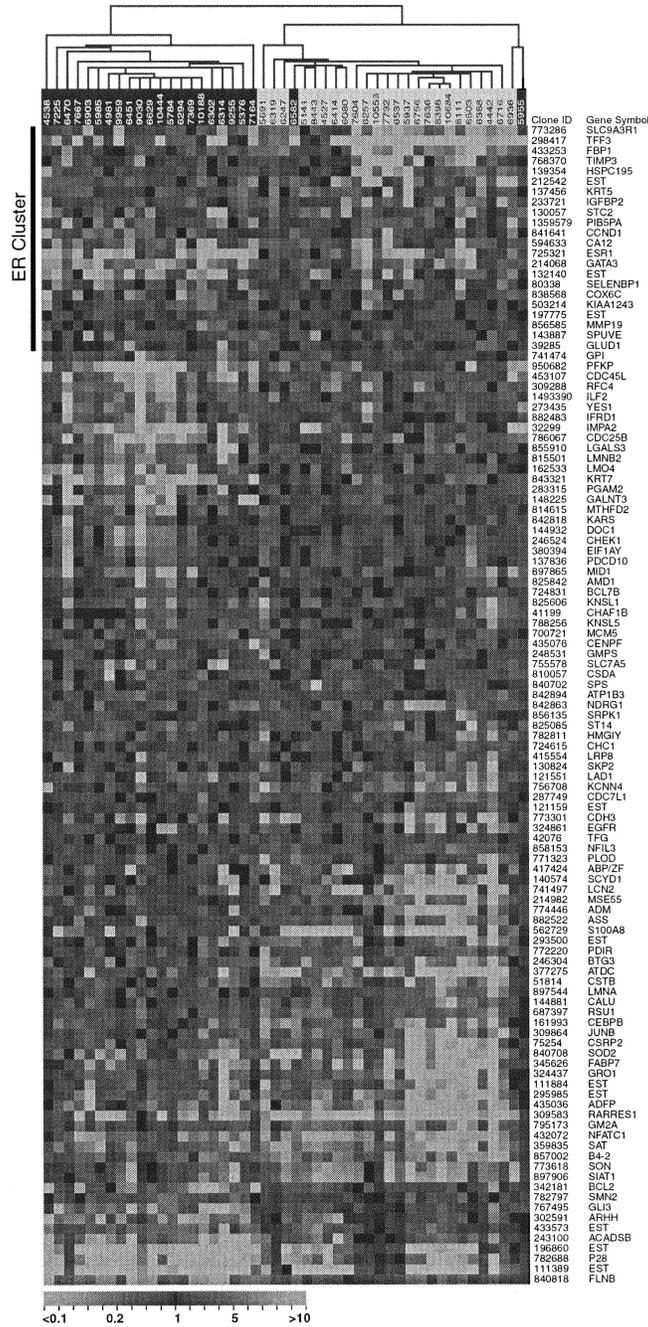


FIGURE 5b. See previous page for legend.

analyses, they performed 78 leave-one-out classifications, each time using the 77 training samples to select the discriminatory gene list and predict the left out 78th sample. With this validation scheme, they successfully predicted the left out samples between 56 and 68 times out of 78 depending on the correlation threshold used. In addition, they correctly predicted 17 out of 19 tumors from an independent test set using a 70-gene classifier. Van't Veer *et al.* compared their classifier to the St. Gallen and NIH consensus conference guidelines for eligibility for adjuvant chemotherapy. These guidelines are based on histological and clinical observations. The comparison is detailed in TABLE 3. For patients who developed metastatic disease and should receive adjuvant systemic therapy, van't Veer's classifier was slightly less sensitive than both the St. Gallen and NIH consensuses. Of the patients who remained disease-free, however, van't Veer's classifier was significantly more specific than either of the two consensuses. The findings of this and similar subsequent studies may significantly change the way a breast cancer patient is determined to be eligible for adjuvant chemotherapy. Disease-free patients may be more accurately diagnosed and suffer much less in the way of harmful side effects.

Furey *et al.*<sup>4</sup> used SVMs and the expression of 97,802 clones to discriminate between 16 ovarian cancers and 15 normal tissues (mixed ovarian tissue and other normal tissue). With various optimization attempts of the SVM parameters and the number of genes used as input, they achieved between 71% and 84% accuracy using leave-one-out cross-validation. Interestingly, their analysis misclassified a normal ovarian tissue sample as ovarian cancer and, upon further investigation, the sample was discovered to be mislabeled. This study demonstrates the power of SVMs to (1) predict whether a tissue sample is cancerous and (2) validate sample information based on gene expression data.

Other studies include that of Xu *et al.*,<sup>3</sup> who used ANNs to distinguish between Barrett's esophagus (BA) ( $n = 14$ ) and esophageal cancer (CA) (3 squamous cell carcinomas and 5 adenocarcinomas). They selected the 160 most relevant genes using SAM (Statistical Analysis of Microarray).<sup>16</sup> After training the network with 12 samples (8 BAs and 4 CAs), it correctly predicted 10 test samples (6 BAs and 4 CAs).

Shipp *et al.*<sup>7</sup> used a supervised learning approach to predict the outcome of diffuse large B cell lymphoma (DLBCL) using oligonucleotide arrays with 6817 probes and tumor samples from 58 DLBCL patients. The 58 patients were separated into two groups: cured disease ( $n = 32$ ) and fatal or refractory disease ( $n = 26$ ). The 5-year overall survival (OS) rate of the 58 patients was 54%. Their supervised learning algorithm employed a leave-one-out cross-validation procedure that generated 58 sets of 13 discriminatory genes. Seven of the 13 genes were common to all of the 58 sets of 13. Using these 58 sets of 13 genes and testing on the single sample that was left out, they generated a prognosis prediction for each patient. Recalculating the 5-year OS for the two prediction classes produced the following: predicted to be cured, 5-year OS = 70%; predicted to have fatal/refractory disease, 5-year OS = 12% (nominal log rank  $P = .000004$ ). Thus, using a supervised learning approach, Shipp *et al.* demonstrated that there is likely to be a gene expression profile at time of diagnosis that can help clinicians predict the outcome of DLBCL patients and proceed accordingly.

Another application of ANNs to DLBCL was performed by O'Neill and Song,<sup>8</sup> who achieved improved classification accuracy using ANNs on the data of Alizadeh *et al.*<sup>17</sup> Alizadeh *et al.* performed cluster analysis and achieved 93% diagnostic

accuracy and were not able to successfully predict prognosis with their methods. O'Neill and Song, using the same data set, were able to achieve 99% diagnostic accuracy and 100% prognostic accuracy. Thus, they demonstrate the significant superiority of ANNs over cluster analysis.

### OUTLOOK

It is clear that AI is bringing to reality some of the promises made by proponents of microarray. AI can accurately predict tumor subtype, metastatic state, and estrogen receptor status. Studies are starting to emerge that demonstrate the power of ANNs to predict clinical outcome. One of the keys to unlocking the power of AI to predict clinical outcome is identifying the misregulated genes in each cancer type. Using this information, we can develop custom arrays containing on the order of 10 to 100 unique genes (and replicates of these genes) that have been implicated in a particular cancer's prognosis profile. We predict that, once a set of diagnostic/prognostic specific genes are identified for all human diseases, handheld computers with an embedded or hard-coded, trained neural network could scan this gene chip, combine the acquired transcriptional data with known clinical parameters such as age, sex, date of presentation, etc., and output a diagnosis or prognosis prediction and a suggested course of treatment. Personalized medicine such as this may sound futuristic, but the fundamental building blocks are already a reality.

### REFERENCES

1. KHAN, J. *et al.* 2001. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat. Med.* **7**: 673–679.
2. GRUVBERGER, S. *et al.* 2001. Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns. *Cancer Res.* **61**: 5979–5984.
3. XU, Y. *et al.* 2002. Artificial neural networks and gene filtering distinguish between global gene expression profiles of Barrett's esophagus and esophageal cancer. *Cancer Res.* **62**: 3493–3497.
4. FUREY, T.S. *et al.* 2000. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* **16**: 906–914.
5. VAN'T VEER, L.J. *et al.* 2002. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**: 530–536.
6. WEST, M. *et al.* 2001. Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc. Natl. Acad. Sci. USA* **98**: 11462–11467.
7. SHIPP, M.A. *et al.* 2002. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat. Med.* **8**: 68–74.
8. O'NEILL, M.C. & L. SONG. 2003. Neural network analysis of lymphoma microarray data: prognosis and diagnosis near-perfect. *BMC Bioinformatics* **4**: 13.
9. RINGNER, M., C. PETERSON & J. KHAN. 2002. Analyzing array data using supervised methods. *Pharmacogenomics* **3**: 403–415.
10. QUACKENBUSH, J. 2001. Computational analysis of microarray data. *Nat. Rev. Genet.* **2**: 418–427.
11. HAYKIN, S. 1999. *Neural Networks: A Comprehensive Foundation*. Prentice-Hall. Upper Saddle River, NJ.
12. BITTNER, M. *et al.* 2000. Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature* **406**: 536–540.
13. SHEK, L.L. & W. GODOLPHIN. 1988. Model for breast cancer survival: relative prognostic roles of axillary nodal status, TNM stage, estrogen receptor concentration, and tumor necrosis. *Cancer Res.* **48**: 5565–5569.

14. OSBORNE, C.K. 1998. Steroid hormone receptors in breast cancer management. *Breast Cancer Res. Treat.* **51**: 227–238.
15. PARL, F.F. 2000. *Estrogens, Estrogen Receptor, and Breast Cancer*. IOS Press. Amsterdam.
16. TUSHER, V.G., R. TIBSHIRANI & G. CHU. 2001. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA* **98**: 5116–5121.
17. ALIZADEH, A.A. *et al.* 2000. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**: 503–511.