International Journal of Industrial Engineering, 15(1), 45-52, 2008.

Proteomic Pattern Analysis Using Neural Networks

Rashpal S. Ahluwalia and Sundar Chidambaram

Industrial and Management Systems Engineering Department,
West Virginia University,
Morgantown WV 26506.
E-mail: {Rashpal Ahluwalia; rashpal.ahluwalia@mail.wvu.edu}

Protein profiling of biologic samples by techniques such as surface-enhanced laser desorption/ionization (SELDI) or matrix assisted laser desorption/ionization (MALDI) yields massive amounts of data that require use of automated techniques to detect expression patterns. This paper suggests a neural network based classification and clustering technique for the analysis of proteomic data on serum samples collected from human subjects exposed to diesel exhaust fumes (DEF). Data were collected on samples from 93 subjects exposed to DEF. Proteomic patterns were analyzed using Neuralware Predict® software obtained from Neuralware Inc. The cascade correlation algorithm was used as the classification algorithm and self-organizing maps (SOM) was used as the clustering algorithm. The protein peaks were identified using the Ciphergen Software. The most discriminating peaks were identified by applying a student t-test and using the p-value as the criterion for discrimination. The classification and clustering algorithms were applied to the two data sets. The use of a neural network program for analysis of proteomic patterns from serum samples obtained from human subjects exposed to DEF or not exposed to DEF showed excellent discrimination. Such an approach has potential to play an important role in determining deleterious effects of occupational exposures and discovery of biomarkers.

Keywords: Pattern Analysis, Neural Networks, Classification, Clustering, Proteomics

(Received 29 April 2005; Accepted in revised form 30 January 2007)

1. BACKGROUND

In recent years low molecular weight serum protein profiling to detect disease processes has come into prominence. A great impetus to this field was provided by a recent demonstration of this technique to detect ovarian cancer [1]. The combined use of bioinformatics tools and protein profiling has been suggested to be an effective approach to screen for potential tumor markers [2]. However, the bioinformatics tools for analyzing such data are at a rudimentary stage. A group recently analyzed the ovarian cancer data and discovered non-biologic experimental bias between the cancer and control groups [3]. They caution that such bias may invalidate attempts to use this dataset to find patterns of reproducible diagnostic value. They suggest that in order to minimize false discovery, results using mass spectrometry and data mining algorithms should be carefully reviewed and benchmarked with routine statistical methods.

This paper describes the use of a neural network approach to analyze proteomic data obtained by surface-enhanced laser desorption/ionization (SELDI) time-of-flight mass spectroscopy. Analysis of proteomic patterns obtained by techniques such as MALDI –TOF and SELDI-TOF is proving to be a powerful tool for diagnosing disease states, particularly for early diagnosis of cancer. The analysis of serum samples by these techniques yields massive amounts of data that require sophisticated analytical tools. Currently, three approaches are being used for data analysis, i) Purely statistical techniques, ii) Decision tree techniques and iii) Artificial neural networks (ANN).

1.1 Statistical Approaches

In one study, breast cancer data was analyzed using a multivariate approach, where the different expression levels of the proteins were compared simultaneously. Hierarchical analysis allowed samples that are highly similar to be merged in an agglomerative way, using the complete linkage clustering procedure. The groupings were presented in the form of a dendograms with trees and branches depicting the extent of similarities among the different groups of the samples. Distinct sample clusters were generated using Mann-Whitney statistical analysis between normal breast tissue and benign breast tissue. This served as the training set for a clustering algorithm [4]. A purely statistical approach was also used for analysis of proteomic patterns to detect the tranisitional cell carcinoma of the bladder, but the sensitivity and specificity were poor [5].

Another statistical approach was a pattern matching decision tree algorithm used to analyze prostate cancer data. The data analysis involved three stages; the first stage was peak detection and peak alignment, which were performed using the Ciphergen software and the Peak Miner algorithm respectively. The peaks were selected using the Ciphergen software and

were aligned using the Peak Miner Algorithm, which was based on a statistical criterion (mass error value); the second stage involved the discriminating criterion to select the most significant peaks. The power of each peak to discriminate between Benign Prostate Hyperplasia (BPH), Healthy Men (HM) and Prostate Cancer (PCA) was determined by finding the area under the curve (AUC) of the response operating characteristic curve (ROC). This ranged from 0.5 (no discriminating power) to 1.0 (complete separation); the third stage involved the discrimination of the peaks based on the principle of one rule at a time. The decision is based on the question of the presence or absence of intensity levels of one peak. Splitting process continued until terminal nodes were produced or any further splitting had no gain. For each node a cost function determining the heterogeneity of the node was determined. The mass at which, the intensities are to be compared are determined by the maximum reduction of cost in the two descendants. This method resulted in a sensitivity of 83% and a specificity of 97% [6].

In another study of prostate cancer patients, the difference in peak intensities between the pairs in the group was analyzed using the Wilcoxon test and the peaks were rank ordered (scored) based on their level of significance. A logistic regression model was performed using the most significant peaks [7] with comparable results reported by Vlahou et al. [5].

1.2 Decision Tree Approaches

The prostate cancer data mentioned above was also analyzed using a boosted decision tree algorithm. Boosting is done in classifiers to reduce the error of any "weak" learning algorithm. A decision tree with only one split is called a Decision Stump, which is usually a weak learner. The first two stages that are involved in the analysis of prostate data were as in the previous case, but the thirds stage of classifying the cases involved the use of ADABOOST [8, 9, 10], which can combine weak learners into an accurate classifier. The boosted decision stump feature selection (BSDFS) classifier algorithm is a committee with decision stumps. It has the base classifiers (most discriminating peaks) as its members. The committee makes a decision based on majority vote. The base classifiers are constructed on weighted samples. In the first round of running the ADABOOST algorithm equal weights are assigned to all examples. For the next round the weights were increased for the examples mis-classified in the first case and the weights were decreased for the examples correctly classified in the first stump.

The above procedure is repeated again and again until a desired number of stumps have been created. In the committee each member has its own special training. For example the second member's mistake is to correct the first member's mistake and so on. The classification decision is made by the majority vote by the base classifiers. As the margin increases the confidence becomes greater. The sensitivity and specificity obtained by ADABOOST are 99% and 98.4% respectively. The sensitivity and specificity obtained by the BSDFS algorithm are 91.5% and 94.7% respectively [11].

1.3 Neural Network Approaches

A more powerful tool for analyzing the proteomic patterns has been the use of ANN. ANN can handle high level of noise and can identify the influence of many factors [12, 13], and has proved to be a more generalized model with respect to medical diagnostics [14]. Early reviews of neural networks show their application to be in the area of non-linear data analysis and distributive associative memory. The reason for their application in these areas was due to their noise tolerance, which was a result of their parallel structure and their adaptability to handle any type of data. Medical data generally have low signal to noise ratio, this is a result of early effects of the disease and variations in clinical protocols. The high levels of noise in medical data require a good non-linear model, which can make generalized predictions. Neural network models meet these requirements. Neural network models provide the additional functionality to statistical models, which are linear in the parameters [15].

A study to classify human tumors used parameterization to identify the influences between the inputs in a modeled ANN system. The simplest method of parameterization is to determine the strength of the input parameter's influence based on the analysis of weights of a trained ANN model. This helps in the ranking of the important input parameters and their interactions based on the outputs. A second method is to rank the input parameters and their interactions based on changes in the performance of the ANN model (RMS value) [16]. Determining the important influences on the system being modeled helps in the development of a more generalized model based on the important input parameters and removal of other parameters having the least influence. The data was divided into a number of sub-blocks and the ANN model was run on each of the sub-blocks. The best model was selected based on the RMS criterion and influences between input parameters. A single model was then run, based on the important input parameters selected from the sub-models. A stepwise regression analysis was performed on the final model and the important input parameters and their influences were determined. These input parameters served as training set for the neural network model. The neural network algorithm used was the back propagation algorithm developed using Neuroshell [17].

One study used clustering software, Propeak to obtain clusters for the different stages of breast cancer. This software implements a linear version of the Unified Maximum Separability Analysis (UMSA) algorithm [18, 19]. The key feature of the UMSA algorithm is the incorporation of data distribution information into a structural risk minimization-learning algorithm. Propeak analyzes the data in three stages; the component analysis for cluster formation, boot strap module which performs multiple runs of UMSA algorithm to establish an objective peak selection criterion, and a third component which

applies backward stepwise regression procedure to compute significant score for each peak. A sensitivity of 93% and specificity of 91% was reported using this approach [20].

Another ANN approach was used to differentiate hepatocellular carcinoma from chronic liver disease. The protein peaks were picked by the Biomarker Wizard from the Ciphergen software. The significant peaks were obtained using the Significance Analysis of Microarray (SAM) testing. SAM is a statistical test for biological experiments, which analyzes significant changes in gene expressions of microarrays [21]. The significant peaks generated by SAM were subject to clustering (two-way hierarchical clustering) and classification (feed forward back propagation algorithm). They reported a sensitivity of 90% and specificity of 92% [22]. A similar approach was used to distinguish Barrett's Esophagus (BE) and Esophageal Cancer (EC). The gene filtering process to obtain the most significant peaks was performed using SAM. The ANN model was built using Matlab. The ANN model used a feed forward back propagation algorithm [23].

The ovarian cancer data reported by Petricoin et al. [1] was analyzed using genetic algorithms and cluster analysis. The genetic algorithm starts with a small set of M/Z values and conducts a "fitness test" using cluster analysis to generate the most significant peak that discriminate the cancer and the non-cancer clusters. The unknown cases are tested at these peak values. The ovarian cancer data was analyzed by a proprietary ANN program, not available for general use [1].

We used a commercially available ANN program called Neuralware Predict for the analysis of our diesel data. This program is capable of performing Clustering and Classification functions. The Clustering algorithm failed to yield a segregated pattern between exposed and control subjects. However, the classification algorithm based on the input from peaks selected on the basis p-values showed a sensitivity of 97% and specificity of 95% at optimal data analysis. The classification algorithm used the cascade–correlation principle for generating the model and the learning was based on Adaptive Gradient Descent [24].

Table 1. Summary of sensitivity and specificity obtained for different training and testing set splits

Training Set - Testing Set	High Vs. Low		High Vs. (Low & Medium)		High Vs. Low Vs. Medium	
	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
70 - 30	86.96%	82.98%	65.22%	88.24%	93.48%	78.72%
75 - 25	67.39%	85.11%	93.48%	91.18%	54.35%	70.21%
80 - 20	93.48%	89.36%	95.65%	97.06%	89.13%	87.23%
85 - 15	97.83%	95.74%	91.30%	94.12%	97.83%	91.49%
90 - 10	95.65%	91.49%	95.65%	97.06%	65.22%	87.23%

The ANN programs reported in several studies [17, 21, and 23] used the conventional classification algorithm (back propagation algorithm). Our model uses the cascade–correlation algorithm, which is a variation of the back propagation algorithm. The disadvantages with the back propagation algorithm are that the error convergence rate is slow, and architecture of the network is fixed. In cascade correlation algorithm these are overcome by faster error convergence, which is achieved by training one net at a time and having the other nets frozen. Further cascade correlation starts with a minimal network and builds the network by the addition of one hidden node at a time. This makes the network more efficient [25 and 26]. The optimum number of testing and training set was to use 85% of the data for training and the remaining data for testing. The DEF data had only 93 samples, the classification algorithm require a bigger dataset to be able to generalize the network.

Hence, with the current data, we let the classification algorithm in Neuralware – Predict decide the test and train set. The classification algorithm in Neuralware – Predict uses a "round – robin" approach, where it selects the test set similar to the training set, but with fewer data points (based on the test train split as shown in Table 1).

In summary, we have shown that a commercially available ANN program can be successfully used for data analysis of serum proteomic patterns obtained from individuals exposed to diesel exhaust fumes. Thus, serum proteomic patterns can, not only be used for cancer diagnosis but also for determining exposure to environmental and other toxic agents.

2. METHODS

This paper describes the use of a neural network approach to analyze proteomic data obtained by surface-enhanced laser desorption/ionization (SELDI) time-of-flight mass spectroscopy. The dataset used for this analysis were collected using the WCX2 chip on Cipehrgen ProteinChip® System [Ciphergen Biosystems, Inc., Fremont, CA] on serum samples of human subjects exposed to diesel exhaust fumes [27]. Data for 93 samples exposed to DEF (34, 13 and 46 from low exposure, medium exposure and high exposure subjects, respectively) were obtained from the Ciphergen software.

Table 2. Summary of peaks selected along with their M/Z values and P values

Peak #	M/Z values	P Values
79	8343.81	0.0018
78	8132.995	0.0041
42	3145.116	0.0137
51	4060.915	0.0139
65	6625.625	0.0242
80	8528.873	0.03
62	5600.881	0.0304
104	16705.79	0.0457
72	7558.739	0.0531
103	16368.75	0.0535
121	3869.56	0.0696
49	28074.73	0.0866

The criterion for the classification of low, medium and high diesel exposures are described elsewhere [28, 29]. The Ciphergen software identified 132 peaks for each of those samples. The 132 peaks obtained for each of the samples were fed into the Microsoft Excel spreadsheet for all 93 samples. The most differentiating peaks (12 peaks), which distinguished high diesel exposure from high and low diesel exposure were obtained by performing a t-test using Microsoft Excel and selecting peaks having p-values below 0.10 (Table 2). The intensities from the 93 samples for each of the 12 peaks were fed into the Neural Network model as training data.

The neural network model was run using "Neuralware – Predict" software from Neuralware Inc [24]. Figure 1 illustrates the raw data reduction phase of the analytical approach.

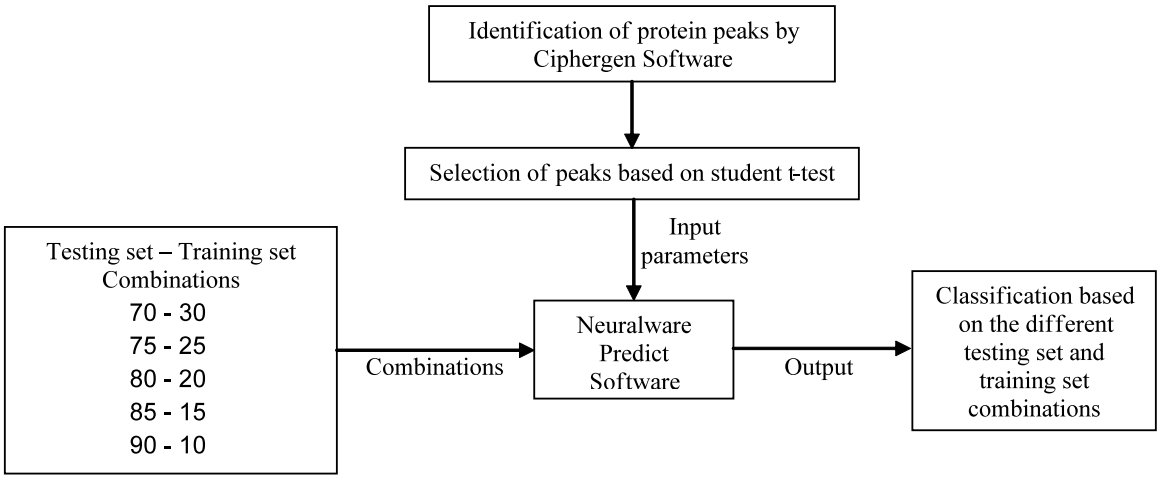

Figure 1. Outline of analytical approach

2.1 Clustering Algorithm

The clustering algorithm used in this study is called the "Winner Take All" algorithm. The weight vector for the clustering unit serves as the representation of the input pattern associated with that cluster. During the self-organization process, the square of the minimum Euclidean Distance (ED) is chosen as the winner.

The weight vectors for a cluster unit serves as an example of the input patterns associated with that cluster. During the self-organization process, the cluster unit whose weight vector matches the input pattern most closely (typically, the square of the minimum Euclidean distance) is chosen as the winner. The winning unit and its neighboring units (in terms of the topology of the cluster units) update their weights. The weight vectors of the neighboring units are not, in general, close to the input pattern [30]. Figure 2 shows an example of three clusters corresponding to the low, medium and high diesel exposures and three unknown input patterns, which are associated to the three clusters by self-organization using Euclidean distance as the criterion.

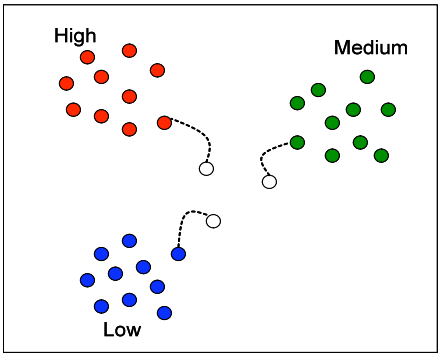

Figure 2. Illustration of clustering through SOM

2.2 Architecture of Clustering Algorithm

The first step of the algorithm is to initialize the weights, i.e. to decide on the number of cluster that is to be formed and assign random weights to each of the elements in the clusters. The step also includes setting topological neighborhood parameters, i.e. to decide on the type of grid to be followed for the clusters. Depending on the type of grid the radius R is set. The step also involves setting the learning rate of the parameters.

The second steps involves the computation of the Euclidean Distance for each element of the input vector x. The Euclidean Distance is computed for the i^{th} element of the input vector x. The input vector with the closest distance ED, which results in the minimum ED, is found.

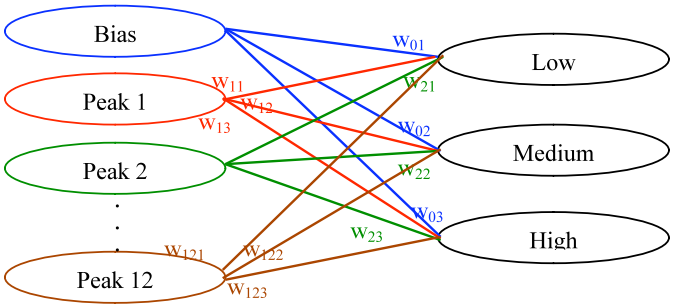

Figure 3. Minimal network of Cascade Correlation algorithm

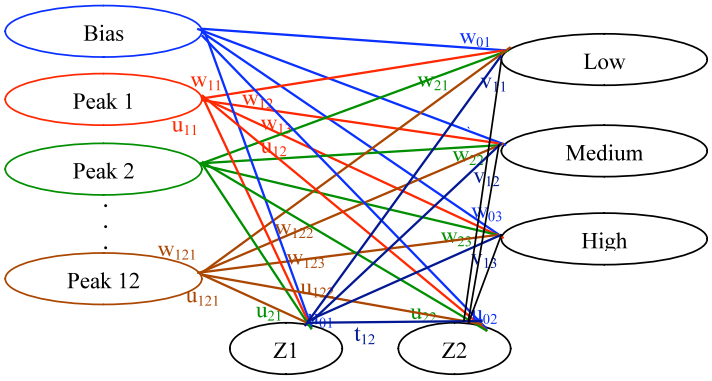

Figure 4. Entire network with all weights for training the network

The third step involves updating the weights. The weights of the winning units are updated. The stopping criterion for the clustering algorithm is fixed number of iterations or when a specified error value is reached. A linear decrease of learning rate α with time is satisfactory for practical computations. The radius of the neighborhood around a cluster unit also decreases as the clustering process progresses [25 and 26].

2.3 Classification Algorithm

Cascade – Correlation starts with a minimal network, consisting of the required input and output units. This net is trained until no further improvement is obtained; the error for the output unit is then computed (summed over all training patterns). Figure 3 shows an example with twelve input units (Peak 1, Peak 2… Peak 12) corresponding to the twelve peak intensities and three outputs corresponding to low, medium, and high diesel exposures. The other parameters in Figure 3 correspond to the weights between the input units and the output units.

One hidden unit is added to the net in a two-step process (hidden units are denoted by z_1, z_2 etc.). In the first step, a candidate unit (hidden unit) is connected to each of the input units, but is not connected to the output units. The weights on the connections from the input units to the candidate unit (hidden unit) are adjusted to maximize the correlation between the candidate's output and the residual error at the output units.

The residual error is the difference between the target and the computed output, multiplied by the output unit's activation function. When training is completed, the weights are frozen and the candidate unit becomes a hidden unit in the net.

The second step in which the new unit is added to the net now commences. The new hidden unit is connected to the output units with the weights on the connection being adjustable. Now all connections to the output units are trained. (The connections from the input units are trained again, and the new connections from the hidden unit are trained for the first time). A second unit is then added using the same process. However, this unit receives an input signal both from the input units and from the previous hidden unit. All weights on these connections are adjusted and then frozen. The connections to the output units from the input nodes and the two hidden nodes are then trained. The process of adding a new weight, training its weights from the input units and previously added hidden units and then freezing the weights, followed by training all connections to the output units, is continued until the error reaches the acceptable level or the maximum number of epochs (or hidden units) is reached. The final network diagram is shown in Figure 4. [25, 26 and 31]

3. RESULTS

3.1 Analysis by Clustering Algorithm

The clustering pattern when all three groups (low, medium and high) were present and the pattern when only low exposure and high groups were present were analyzed. There was no clear segregation pattern in either case. This suggests that the clustering algorithm may not be useful for classification of this set of proteomic data.

3.2 Analysis by Classification Algorithm

Neuralware-Predict provide options for using different number of data sets for training and testing phases. We used a variety of combinations for analyzing the data from diesel exposure subjects and also different combinations of training and testing sets as shown in Table 1. Using this approach we obtained a sensitivity of 98% and a specificity of 96% with diesel data. Further analysis of the different training and testing data splits and the use of different levels of exposures of DEF showed that using 85% of the data as the training set and the remaining 15% of the data as the testing set gave the best results in all cases. When subjects exposed to high levels of DEF were compared with subjects with low exposure DEF, the sensitivity and specificity were 98% and 96% respectively. A comparison of the high DEF exposure group with the two remaining groups together resulted in a sensitivity of 91.3% and a specificity of 94.1%. Finally, analysis of the three groups separately resulted in a sensitivity of 97.8% and a specificity of 91.5%. Thus, the classification algorithm seems effective in identifying the proteomic patterns of subjects exposed to different levels of DEF exposure.

4. CONCLUSION

The application of a neural network program for analysis of proteomic patterns from serum samples obtained from human subjects exposed to DEF or not exposed to DEF showed excellent discrimination. The neural network approach has the potential to play an important role in determining deleterious effects of occupational exposures and discovery of biomarkers.

5. REFERENCES

1. E. F. Petricoin III et.al, Use of proteomic patterns in serum to identify ovarian cancer, LANCET: Mechanism of disease, February 2002, *359*, 572-577.
2. J. Rai et.al., Proteomic approaches to tumor marker discovery: identification of biomarkers from ovarian cancer, Archives of Pathology Lab Medicine, 2002, *126*, 1518-1526.
3. J. M. Sorace et.al., A data review and re-assessment of ovarian cancer serum proteomic profiling, BMC Bioinformatics, 2003, *4*, 24-36.

4. M. V. Dwek et.al., Proteome analysis enables separate clustering of normal breast cancer, benign breast and breast cancer tissues, British Journal of Cancer, 2003, *89*, 305-307.
5. Vlahou et.al., Development of a novel proteomic approach for the detection of transitional cell carcinoma of the bladder in urine, American Journal of Pathology, April 2001, *158*(4), 1491-1502.
6. L. Adam et.al., Serum Protein Fingerprinting coupled with a Pattern-matching Algorithm distinguishes Prostate Cancer from Benign Prostate Hyperplasia and Healthy Men, Cancer Research, July 2002, *62*, 3609-3614.
7. L. H. Cazares et.al., Normal, Benign, Preneoplastic and Malignant Prostate Cells have distinct Protein Expression Profiles resolved by Surface Enhanced Laser Desorption/Ionization Mass Spectrometry, Clinical Cancer Research, August 2002, *8*, 2541-2552.
8. Y. Freund et.al., A decision-theoretical generalization of on-line learning and an application to boosting, Journal of Computer Systems Science, 1997, *55*, 119-139.
9. T. Hastie et.al., The elements of statistical learning, (New York, Springer-Verlag, 2001), 301pp.
10. J. Friedman et.al., Additive logistic regression: a statistical view of boosting, Annals of Statistics, 2000, *28*, 337-407.
11. Y. Qu et.al., Boosted Decision Tree Analysis of Surface-enhanced Laser Desorption/Ionization Mass Spectral Serum Profiles discriminates Prostate Cancer from Noncancer patients, Clinical Chemistry, 2002, *48*(10) 1835-1843.
12. J. T. Wei et.al., Understanding Artificial Neural Networks and exploring their potential applications for the practicing urologist, Urology, 1998, *52*, 161-172.
13. S. C. Kothari et.al., Neural Networks for pattern recognition, *Advances in Computers*, 1993, *37*, 119-166.
14. E. Tafeit et.al, Artificial Neural Networks in laboratory medicine and medical outcome prediction, Clinical Chemical Lab Medicine, 1999, *37*, 845-853.
15. P. J. G. Lisboa, A review of evidence of health benefit from artificial neural networks in medical intervention, Neural Networks, January 2002, *15*(1) 11 – 39.
16. G. R. Balls et.al., Towards unravelling the complex interactions between microclimate, ozone dose and ozone injury in clover, Water, Air Soil Pollution, 1996, *85*, 1467-1472.
17. G. Ball et.al., An integrated approach utilizing artificial neural networks and SELDI mass spectrometry for the classification of human tumors and rapid identification of potential biomarkers, Bioinformatics, 2002, *18*(3), 393-404.
18. Z. Zhang et.al., Applying classification separability analysis to microarray data analysis, Methods of microarray data analysis: papers from CAMDA 2000, Boston: Kluwer Academic Publishers, 2001, 25-26.
19. V. N. Vapnik, Statistical Learning Theory, (New York: John Wiley & Sons, 1998) 401-440pp.
20. J. Li et.al., Proteomics and Bioinformatics Approaches for Identification of Serum Biomarkers to Detect Breast Cancer, Clinical Chemistry, 2002, *48*(8) 1296-1304.
21. V. G. Tusher et.al., Significance Analysis of Microarrays applied to the ionizing radiation response, Proceedings of the National Academy of Sciences, 2001, *98*, 5116-5121.
22. W. T. Poon et.al., Comprehensive Proteomic Profiling Identifies Serum Proteomic Signatures for Detection of Hepatocellular Carcinoma and its Subtypes, Clinical Chemistry, 2003, *49*(2) 752-760.
23. Y. Xu et.al., Artificial Neural Networks and Gene Filtering distinguishes between Global Gene Expression Profiles of Barrett's Esophagus and Esophageal Cancer, Cancer Research, June 2002, *62*, 3493-3497.
24. User Guide – Neuralware Predict, February 2003.
25. L. Fausett, Fundamentals of Neural Networks, Architectures, Applications and Algorithms, 1994.
26. S. Haykin, Neural Networks: A Comprehensive Foundation, (Prentice Hall. 1999).
27. Diesel Human Samples Data, Department of Defence.
28. W. Dockery et.al., An association between air pollution and mortality in six US cities, New England Journal of Medicine, December 1993, *29*, 1753-1759.
29. R. O. McClellan, Health effects of exposure to diesel exhaust particles, Annals Review of Pharmacological Toxicology, 1987, *27*, 279 - 300.
30. T. Kohonen, Self-Organization and Associative Memory, (3rd edition, Berlin, Springer-Verlag, 1989).
31. S. E. Fahlman et.al., The Cascade-Correlation Learning Architecture, *CMU*, August 1991.

BIOGRAPHICAL SKETCH

Rashpal Ahluwalia is professor of Industrial Engineering, West Virginia University, Morgantown, WV. His areas of interest are Information Systems, Computer Integrated Manufacturing, Quality, and Reliability Engineering. Dr. Ahluwalia is a fellow of the American Society for Quality (ASQ) and senior member of the Institute of Electrical and Electronics Engineers (IEEE). He is on the Editorial Board of the International Journal of Industrial Engineering and the Journal of Quality Engineering. He is a registered Professional Engineer (PE). He is certified as Software Quality Engineer by ASQ.

Sundar Chidambaram is currently working as a Quality Engineer at Rolls Royce Energy. He is working on his PhD in Industrial Engineering at West Virginia University. He has a master's degree in Industrial Engineering from University of Southern California and a bachelor's degree in Chemical Engineering from University of Madras. He is a Certified Quality Engineer (CQE) by the Amercian Society for Quality (ASQ).