Scientific Journal of Impact Factor (SJIF): 4.14

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 3, Issue 12, December -2016

Use of Neural Network in the field of Bioinformatics for prediction of cancer

Sangeeta Sharma[1], Kodanda Dhar Sa [2]

[1]Electronics And Telecommunication Engineering, Indira Gandhi Institute of Technology,Sarang
[2]Electronics And Telecommunication Engineering,Indira Gandhi Institute of Technology,Sarang

Abstract — *The purpose of this analysis was to develop a method for classifying cancers to specific diagnostic categories based on their gene expression signatures using artificial neural networks (ANNs). We trained the ANN by using the small, round blue-cell tumors (SRBCTs) as the model. These cancers belong to four distinct diagnostic categories and usually present diagnostic dilemmas in medical study. As their name implies, these cancers are difficult to distinguish by light microscopy, and currently no single test can accurately distinguish these types of cancers. The ANN properly classified the whole samples and identified the genes most relevant to the classification. To test the ability of the trained ANN models to identify SRBCTs, we examined additional blinded samples that were not previously used for the training purpose, and correctly classified them in all cases. This study demonstrates the potential applications of these methods for tumor diagnosis and the identification of candidate targets for therapy.*

Keywords- *Multi-class classification, Principal component Analysis, Artificial Neural Network, Backpropagation Algorithm, cancer classification and diagnostic prediction of cancer.*

I. INTRODUCTION

The small, round blue cell tumors (SRBCTs) of childhood, which include neuroblastoma (NB), rhabdomyosarcoma (RMS), non-Hodgkin lymphoma (NHL) and the Ewing family of tumors (EWS). As their name suggests, these cancers are difficult to categorize by light microscopy, and currently no single test can accurately classify these cancers. Gene-expression profiling using cDNA microarrays permits a simultaneous study of multiple markers, and has been used to classify the cancers into subgroups.

However, despite the many statistical techniques to analyze gene-expression data, none so far has been rigorously tested for their ability to accurately distinguish cancers belonging to several diagnostic categories. Here, we have used Artificial Neural Network (ANN) for the classification and diagnostics prediction of cancer.

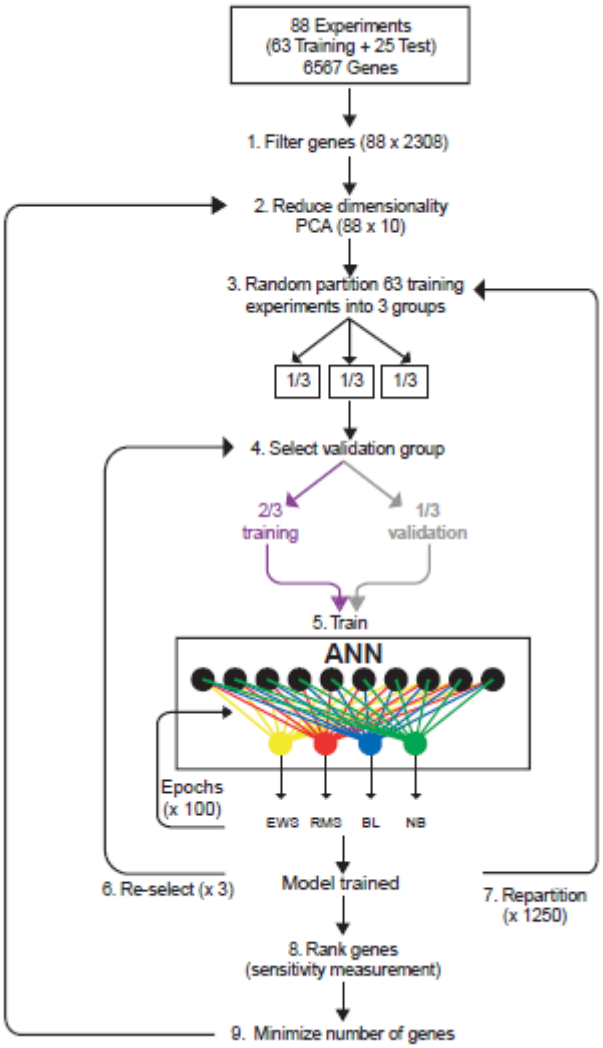
II. ARTIFICIAL NEURAL NETWORK

Artificial neural networks (ANNs) are computer-based algorithms which are modeled on the structure and behavior of neurons in the human brain and can be trained to recognize and categorize complex patterns. Pattern recognition is achieved by adjusting parameters of the ANN by a process of error minimization through learning from experience. They can be calibrated using any type of input data, such as gene-expression levels generated by cDNA microarrays, and the output can be grouped into any given number of categories. ANNs have been recently applied to clinical problems such as diagnosing myocardial infarcts and arrhythmias from electrocardiograms and interpreting radiographs and magnetic resonance images. Here we applied ANNs to decipher gene-expression signatures of SRBCTs and used them for diagnostic classification. The algorithm used here is backpropagation algorithm.

The backpropagation algorithm (Rumelhart and McClelland, 1986) is used in layered feed-forward ANNs. This means that the artificial neurons are organized in layers, and send their signals “forward”, and then the errors are propagated backwards. The network receives inputs by neurons in the input layer, and the output of the network is given by the neurons on an output layer. There may be one or more intermediate hidden layers. The backpropagation algorithm uses supervised learning, which means that we provide the algorithm with examples of the inputs and outputs we want the network to compute, and then the error (difference between actual and expected results) is calculated. The idea of the backpropagation algorithm is to reduce this error, until the ANN learns the training data. The training begins with random weights, and the goal is to adjust them so that the error will be minimal. The back propagation algorithm changes the synaptic weights in an iterative manner so as to minimize the error and bring the network output as close as possible to the target.

III. METHOD : CALIBRATION AND VALIDATION OF THE ANN MODELS:

To calibrate ANN models to recognize cancers in each of the four SRBCT categories, we used gene-expression data from cDNA microarrays containing 6567 genes. The 63 training samples included both tumor biopsy material (13 EWS and 10 RMS) and cell lines (10 EWS, 10 RMS, 12 NB and 8 Burkitt lymphomas (BL; a subset of NHL)).



"Figure.1:Schematic illustration of the analysis process"

The artificial neural network ***a*** is the Schematic illustration of the Analysis Process:
The entire dataset of all 88 experiments was first quality filtered and then the dimensionality was further reduced by principal component analysis (PCA) to 10 PCA projections from the original 6567 expression values. Next, the 25 test experiments were set aside and the 63 training experiment were randomly partitioned into 3 groups. One of these groups was reserved for validation and the remaining 2 groups for calibration. ANN models were then calibrated using for each sample the 10 PCA values as input and the cancer category as output. For each model the calibration was optimized with 100 iterative cycles (epochs). This was repeated using each of the 3 groups for validation and the samples were again randomly partitioned and the entire training process repeated. For each selection of a validation group one model was calibrated, resulting in a total of 3750 trained models. Once the models were calibrated they were used to rank the genes according to their importance for the classification. The entire process was repeated using only top ranked genes and then the 25 test experiments were subsequently classified using all the calibrated models.

IV. RESULT AND DISCUSSION

Table 1 shows the result of the classified test samples. From table 1 we can analyze that the blinded test samples and properly classified and the diagnosis of the cancer type is being done accurately except the sample label 1, according to histological diagnosis sample label 1 belongs to NB type of cancer. Now, we have calculated the Confusion Matrix as per the output generated by the SVM.

Here, based on the classified samples a simple model of the Artificial Neural Network is generated, where all the 63 data are taken as the input of the artificial neuron. Here, Multilayer feedforward network is used .The model makes use of the bias term whose weight is w but with a fixed input of b=1. The transfer function used here is the bipolar Sigmoidal function whose value is given by,

$$\Theta = \tan [\lambda I]$$

With the help of Artificial Neural Network and Support Vector Machine we classified the SRBCTs type of cancer.

DIAGNOSTIC PREDICTION OF CANCERS USING ANN:

"Table 1. ANN Classification and Diagnostic Prediction"

SAMPLE LABEL	EWS	RMS	NB	BL	SVM CLASSIFICATION	SVM DIAGNOSIS	HISTOLOGICAL DIAGNOSIS
1	0.5309	0.6038	0.413	0.2959	RMS	-	NB
2	0.8725	0.4222	-0.1516	0.507	EWS	EWS	EWS
3	0.7941	-0.1337	0.2472	-0.1092	EWS	-	Osteosarcoma-C
4	-0.025	0.9964	-0.5043	0.0888	RMS	RMS	RMS
5	0.6978	0.2978	0.0506	0.301	EWS	-	Sarcoma-C
6	0.9986	-0.1651	0.0593	-0.1795	EWS	EWS	EWS
7	0.5727	0.2042	-0.2934	0.8884	BL	BL	BL
8	-0.194	-0.1921	0.9106	0.061	NB	NB	NB
9	0.5638	0.9339	-0.8337	0.0191	RMS	RMS	Sk.Muscle
10	-0.242	0.84	0.3037	-0.1013	RMS	-	RMS
11	0.8154	0.2891	-0.0897	0.4589	EWS	-	Prostate Ca.-C
12	0.9873	-0.2136	0.3569	-0.0477	EWS	EWS	EWS
13	0.0483	0.8877	-0.496	-0.2887	RMS	RMS	Sk.Muscle
14	0.387	-0.8291	0.9654	-0.4258	NB	NB	NB
15	0.3146	-0.3221	-0.006	0.9296	BL	BL	BL
16	0.3865	-0.8039	0.9715	-0.417	NB	NB	NB
17	0.4465	0.9917	-0.4679	-0.0171	RMS	RMS	RMS
18	-0.054	-0.0882	0.2159	0.8281	BL	-	BL
19	0.9822	-0.4849	0.354	0.0463	EWS	EWS	EWS
20	0.9458	0.3399	-0.238	0.0795	EWS	EWS	EWS
21	0.9059	0.1183	0.2455	0.0946	EWS	EWS	EWS
22	-0.318	0.9193	0.2946	0.022	RMS	RMS	RMS
23	0.2859	-0.7083	0.9342	-0.1841	NB	NB	NB
24	-0.114	0.9355	-0.0671	0.4128	RMS	RMS	RMS
25	0.5596	-0.6861	0.9339	-0.3711	NB	NB	NB

CONFUSION MATRIX:

A confusion Matrix is a table that is often used to describe the performance of a classification model on a set of the test data for which the true values are known. Table 2 represents the confusion matrix of the given test samples.

"***Table 2. Confusion Matrix***"

	EWS	RMS	NB	BL	NON-SRBCTs
EWS	6	0	0	0	0
RMS	0	4	0	0	2
NB	0	0	5	0	0
BL	0	0	0	2	0
NON-SRBCTs	0	0	0	0	3

ACCURACY RATE:

Accuracy Rate gives the value of overall how often the classifier is correct which can be calculated from the confusion matrix.

$$\text{Accuracy Rate} = \frac{6+4+5+2+3}{25} = 0.80$$

Accuracy (%) = 80%

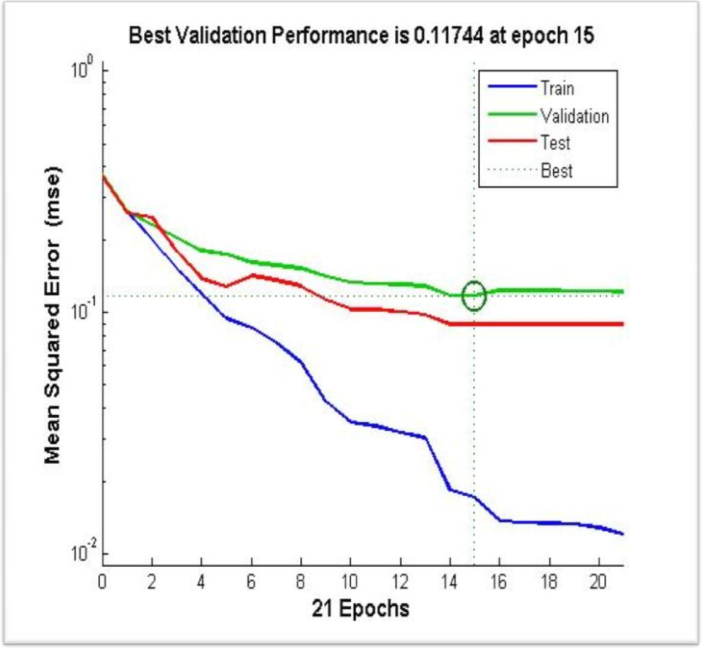
MISCLASSIFICATION RATE:

Misclassification Rate can be calculated by confusion matrix simply by the equation,

Misclassification Rate= (false positive + false negative) ÷ (Total number of test samples)

So,

$$\text{Misclassification Rate} = \frac{2}{25} = 0.08$$



"Figure.2: Epochs vs Mean Square Error"

The above figure shows the relation between the mean square error and the number of epochs. The average classification error is plotted during the training iterations (epochs) for both the training, the validation and testing samples. Here, the blue line is for training samples, green is for validation and red is for testing samples. The decrease in

the classification errors with increasing epochs demonstrates the learning of the models to distinguish these cancers. All the models performed well for training, validation and test samples. We can see from the above graph that as the number of epochs increases the mean square error decreases and the models are more trained. And the curve of the validation and testing samples becomes parallel after sometime i.e. at epochs 15 which show the best validation performance.

DISCUSSION:

Tumors are currently diagnosed by histology and immunohistochemistry based on their morphology and protein expression respectively. However, poorly differentiated cancers can be difficult to diagnose by routine histopathology. In addition, the histological appearance of a tumor cannot reveal the underlying genetic aberrations or biological processes that contribute to the malignant process. Here we developed a method of diagnostic classification of cancers from their gene-expression signatures and identified the genes that contributed to this classification using Artificial Neural Network(ANN).

We calibrated ANN models on the expression profiles of 63 SRBCTs of 4 diagnostic categories. Due to the limited amount of training data and the high performance achieved.

As the main goal of this analysis was to make the most effective classification of these cancers, we used a precise quality filter to use only the genes which shows good measurement results for all the samples. This may remove certain genes that are highly expressed in some cancers, but not expressed in other cancers, or may not appear to be expressed because of an artefact in a particular cDNA spot. However, we found that this quality filtration produce more vigorous prediction models and led to the identification of these types of cancers.

We used the SRBCTs of childhood as a model because these cancers occasionally present diagnostic difficulties. Here we developed a method of diagnostic classification of cancers from their gene expression signatures using ANNs. The genes are ranked and accordingly the Classification and diagnostic prediction of cancers are carried out using Artificial Neural Networks (ANN). In addition, there was no sign of 'over-training' of the models, it rises the summed square error for the validation set with increasing training iterations or 'epochs'. So, the hidden layers should be taken accurately so that there will not be more error and the result will also be of an accurate one.

Although ANN analysis leads to identification of genes specific for a cancer with implications for biology and therapy, a strength of this method is that it does not require genes to be exclusively associated with a single cancer type. This allows for classification based on complex gene-expression patterns.

Future applications of this method will include studies to classify cancers according to their stages and biological behavior in order to predict diagnosis and thereby use the direct therapy.

V. CONCLUSION

Cancer diagnosis is one of the most emerging medical applications of gene expression microarray technology. Here we developed a method of diagnostic classification of cancers from their gene-expression signatures and identified the genes that contributed to this classification using ANNs. We also identified in ranked order the genes that contributed to this classification, and we were able to define a minimal set that can correctly classify our samples into their diagnostic categories. Although we achieved high sensitivity and specificity for diagnostic classification, we believe that with larger arrays and more samples it will be possible to improve on the sensitivity of these models for purposes of diagnosis in clinical practice.

VI. REFERENCES

[1] McManus, A.P., Gusterson, B.A., Pinkerton, C.R. & Shipley, J.M. The molecular pathology of small round-cell tumours—relevance to diagnosis, prognosis, and classification. *J. Pathol.* 178, 116–121 (1996).

[2] Khan, J. *et al.* Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays. *Cancer Res.* **58**, 5009–5013 (1998).

[3] Alizadeh, A.A. *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503–511 (2000).

[4] Khan, J., Wei, J. S., Ringner, M., Saal, L. H., Ladanyi, M., Westermann, F., et al. (2001).Classification and diagnostic prediction of cancers using gene expressing profiling and artificial neural network. Nature Medicine, 7, 673–679.

[5] Bittner, M. *et al.* Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature* **406**, 536–540 (2000).

[6] Lindsay I Smith, A tutorial on Principal Components Analysis.

[7] Aguilar-Ruiz, J. S., & Divina, F. (2006). Biclustering of expression data with evolutionary computation. IEEE Transactions on Knowledge and Data Engineering, 18(5), 590–602.

[8] E. Domany, Cluster analysis of gene expression data, Journal of Statistical Physics 110 (3–6) (2003) 1117–1139.

[9] Cavazzana, A.O.,Miser,J.S., Jefferson, J. &Triche, T.J.Experimental Evidence for a neuralorigin of Ewing's Sarcoma of bone .Am. J. Pathol. 127,507-518(1987).

[10] Paul O'Neill,George D. Magoulas, Member, and Xiaohui Liu "Improved Processing of Microarray Data Using Image Reconstruction Techniques" IEEE Transactions On Nanobioscience, Vol. 2, No. 4, December 2003.

[11] Gordon K. Smyth and Terry Speed "Normalization of the cDNA Microarray Data" IEEE Paper on Normalization April 4,2003.

[12] Yang, Y.H. et al. (2001) Analysis of cDNA microarray images. Brief. Bioinform. 2, 341 – 349.

[13] Leung,Y.F. et al. (2002) Microarray Software Review. In A Practical Approach to Microarray data Analysis,Kluwar academic.

[14] Lee, P.D. (2002) Control genes and variability: absence of ubiquitous reference transcripts in diverse mammalian expression studies. Genome Res .12,292-2979 Nadon, R. and Shoemaker, J. (2002) Statistical issues with microarrays: processing and analysis. Trends Genet.18, 265-271.

