International Journal of Medical, Medicine and Health Sciences
World Academy of Science, Engineering and Technology
ISSN: 2517-9969
Vol:6 2012-11-28
Vol:6, No:11, 2012

Gene Expression Signature for Classification of Metastasis Positive and Negative Oral Cancer in Homosapiens

A. Shukla, A. Tarsauliya, R. Tiwari, and S. Sharma

International Science Index Vol:6, No:11, 2012 waset.org/Publication/283

Abstract**—**Cancer classification to their corresponding cohorts has been key area of research in bioinformatics aiming better prognosis of the disease. High dimensionality of gene data has been makes it a complex task and requires significance data identification technique in order to reducing the dimensionality and identification of significant information. In this paper, we have proposed a novel approach for classification of oral cancer into metastasis positive and negative patients. We have used significance analysis of microarrays (SAM) for identifying significant genes which constitutes gene signature. 3 different gene signatures were identified using SAM from 3 different combination of training datasets and their classification accuracy was calculated on corresponding testing datasets using k-Nearest Neighbour (kNN), Fuzzy C-Means Clustering (FCM), Support Vector Machine (SVM) and Backpropagation Neural Network (BPNN). A final gene signature of only 9 genes was obtained from above 3 individual gene signatures. 9 gene signature's classification capability was compared using same classifiers on same testing datasets. Results obtained from experimentation shows that 9 gene signature classified all samples in testing dataset accurately while individual genes could not classify all accurately.

Keywords**—**Cancer, Gene Signature, SAM, Classification.

I. Introduction

CLASSIFICATION related problems have been key research in the field of medical diagnosis in last few decades. Genes hold the information to build and maintain an organism's cells and pass genetic traits to offspring. All organisms have many genes corresponding to various different biological traits, some of which are immediately visible, such as eye colour or number of limbs, and some of which are not, such as blood type or increased risk for specific diseases, or the thousands of basic biochemical processes that comprise life. As genes posses all the information associated with survival and inheritance. Considering the information associated with genes, their sequences can be used for the purpose of classification of clinically positive and negative patient of oral cancer metastasis [1-6]. During the evolution of cancers through the primary to metastasis stage, genes related to cancer evolution information posses malignant nature. With the help of these identified genes, we can classify them into lymphatic non-metastasis and metastasis stage.

A genome posse's number of genes which makes problem data of a higher dimension. Various techniques such as Significance Analysis of Microarrays (SAM) [7], Empirical Bayes Analysis of Microarrays (EBAM), Limma are used to identify significant genes out of the genome. In this experimentation, SAM has been used for the purpose of identification of significant differential genes. SAM gives as output of set of genes which can be used for isolation between clinically positive and negative tested patients.

Classification models generally follow either supervised learning or unsupervised learning. In case of supervised learning, training data target classes are given to learn their parameters. While in unsupervised learning, data target class is not given. Model in case of unsupervised learning has to identify the class itself based upon the data given. Supervised learning classification models like k-nearest neighbour, artificial neural network, support vector machine are used for the task of classification. While Fuzzy C-Means Clustering algorithm has also been used, incorporates unsupervised learning. Further optimization methods like genetic algorithm and particle swarm optimization are used for optimizing the neural network parameters.

The k-nearest neighbour classifier labels an unknown object O with the label of the majority of the k nearest neighbours. An input is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its k nearest neighbors [13]. Fuzzy C-Means Clustering is a data clustering technique wherein each data point belongs to a cluster to some degree that is specified by a membership grade. It provides a method that shows how to group data points that populate some multidimensional space into a specific number of different clusters [14]. SVM is primarily a statistical method that performs classification tasks by constructing hyperplanes in a multidimensional space that separates cases of different class labels. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other [16]. The success of Artificial Neural Network (ANN) applications can be qualified of their features and powerful pattern recognitions capability. The use of ANN in this field has been growing due to their ability to model complex nonlinear systems on sample data. ANN functions by finding correlations and patterns in the data which you provide [18].

A.Shukla is with the Indian Institute of Information Technology and Management, Gwalior, India. He is with the Department of ICT working as a professor (e-mail: dranupamshukla@gmail.com).

A. Tarsauliya is with the Indian Institute of Information Technology and Management, Gwalior, India (e-mail: anupam8391[@]gmail.com).

R. Tiwari is with the Indian Institute of Information Technology and Management, Gwalior, India. She is with the Department of ICT working as a Assistant professor (e-mail: tiwariritu2@gmail.com).

S. Sharma is with the Indian Institute of Information Technology and Management, Gwalior, India (e-mail: sanjeev.sharma1868@gmail.com).

Lymphatic metastasis is the process of expansion of cancer cells from existing parts of the body to new areas through lymph nodes. Lymph node metastasis could be predicted by gene expression profiles of primary oral cavity squamous cell carcinomas [1]. The combination of alteration observed and earlier and later stage can be used to determine extent of metastasis in lung adenocarcinoma [2]. Oral tongue squamous cell carcinoma progression can be predicted by gene expression analyses of primary tumors [3]. Gene information of an patient can be used in task of prediction or classification of cancer as altered genes set the basis of carrying out the task [1-6, 10-12]. The task of identifying significant genes can be carried out using statistical test hypothesis. SAM has been illustrated for ionizing radiation purpose [7]. In this paper, SAM has been used for identifying significant genes for oral cancer lymphatic metastasis. Quantitative reverse transcription-polymerase chain reaction was used to validate the genes selected by Adaboost algorithm, the model was constructed with eight highly significant genes and then it was evaluated against test case group [9]. DNA microarray analysis on primary breast tumors of 117 young patients were used with supervised classification model. The gene information was recorded at short intervals of metastasis and this was used to predict the clinical outcome of breast cancer [10]. K- Nearest Neigbour was used for gene data analysis and modeling for classification using gene expression analysis [13]. Clustering techniques like K-Nearest Neigbour are considered as hard clustering as they assign strictly a single class. While Fuzzy C-Means is a soft clustering which assigns a score of membership to the different classes. Fuzzy C-means applicability to microarray data clustering and influence of fuzziness parameter is discussed [14-15]. A method of gene selection utilizing Support Vector Machine methods based on Recursive Feature Elimination is addressed to overcome the difficulty of selection of small subset of genes [17].

II. Experiments and Results

A. Research Data

We have oral cavity cancer database obtained from the repository of European Bioinformatics Institute. The database [1] consists of samples of 27 assays aging 42-83 years with clinical history of metastasis and non-metastasis. Assays are both male and female. Affected organism parts under the disease squamous cell carcinoma are floor of mouth, floor of mouth/buccal, floor of mouth/tonsil, gingiva, larynx, lymph node metastasis, mandible and tongue. Genomes corresponding to each subject consist of 22283 genes.

B. Methodology

Classification task involves use of micro-array data of genes which is used for the purpose. Dataset used is of 27 subjects, out of which 22 subjects are of primary tumors and 5 subjects are of lymphatic node metastasis. Out of the 22 primary tumors assays, 14 of them are positive to metastasis and remaining 8 of them are negative to metastasis. A subject is associated with thousands of genes recorded out of which genes affected or showing malignant behavior are identified. These identified genes form a sequence referred as the gene signature. Gene signatures of these subjects are used to carry out the classification and prediction task. Here, dataset is divided into training and testing datasets from primary tumor samples. Lymphatic metastasis samples are used as independent to check the classification accuracy of classifying models along with test set. A 9 genes signature was profiled from three different gene signatures for three different sample set of primary tumors using SAM. Genes which were present in more than 2 of the gene signatures were selected. Training dataset is used to train the prediction models like artificial neural networks, clustering based classification; support vector machines etc. are used. Final results of the systems are calculated using test dataset and lymphatic metastasis dataset.

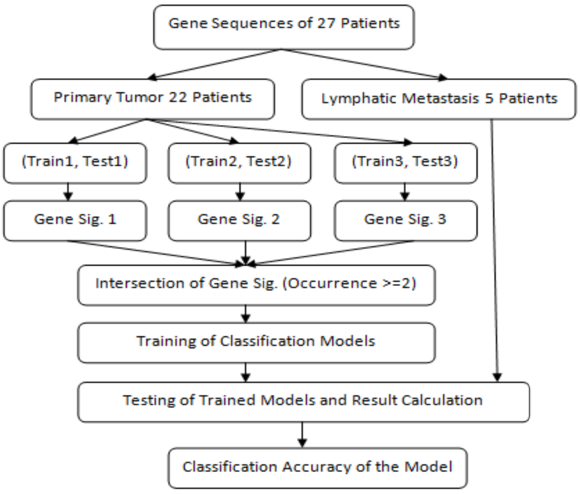
Gene Sequences of 27 Patients
Primary Tumor 22 Patients
Lymphatic Metastasis 5 Patients
(Train1, Test1)
(Train2, Test2)
(Train3, Test3)
Gene Sig. 1
Gene Sig. 2
Gene Sig. 3
Intersection of Gene Sig. (Occurrence >=2)
Training of Classification Models
Testing of Trained Models and Result Calculation
Classification Accuracy of the Model

Fig. 1 Flowchart diagram of Methodology Used

B.1. Gene Sequences

Gene sequences posses all the information about the survival and inheritance of organism. All the adaption, diseases or any changes are reflected into corresponding genes into the gene sequence. In our database, genomes recorded per patient consist of different 22883 genes. Every gene has its significance pertaining to its representative. In same way, there are some genes which affected either by getting upregulated or downregulated during the oral cancer metastasis.

B.2. Train and Test Dataset

B.2.1 E-GEOD-2280

The whole training database consists of 27 assays, out of which 22 subjects are of primary tumors and 5 subjects are of lymphatic node metastasis. Out of the 22 primary tumors assays, 14 of them are positive to metastasis and remaining 8 of them are negative to metastasis. Three different training and testing datasets are formed for carrying out the experimentation as described in Table I.

International Science Index Vol:6, No:11, 2012 waset.org/Publication/283

TABLE I
DESCRIPTION OF TRAINING AND TESTING DATASETS USED

Dataset	N+	N-
Training Set 1	9	6
Test Set 1	5	2
Training Set 2	9	5
Test Set 2	5	3
Training Set 3	10	5
Test Set 3	4	3

B.3. Gene Signature

Not all genes into the gene sequences are affected due to some diseases or changes into organisms. Changes in some of them might be due to cancer diseases while some of them might have change due to diabetes although some affected genes might be common. Thus corresponding to a disease we first find gene signatures for carrying out the further.

B.3.1 Significance Analysis of Microarrays (SAM)

SAM proposed by Tusher et al. (2001) is used for identifying differential genes for making the gene signature. The input to SAM is gene expression measurements from a set of microarray experiments, as well as a response variable from each experiment. SAM computes a statistic d(i) for each gene i, measuring the strength of the relationship between gene expression and the response variable. It uses repeated permutations of the data to determine if the expressions of any genes are significantly related to the response. The cutoff for significance is determined by a tuning parameter delta, chosen by the user based on the false positive rate. One can also choose a fold change parameter, to ensure that called genes change at least a pre-specified amount.

B.3.1.1 Gene Signature 1

Gene signature 1 was identified using SAM from training data set 1 which consists of 9 metastasis positive and 6 metastasis negative patients. SAM plot for the same showing delta and False Discovery Rate (FDR) for the training dataset 1 is shown in Fig. 2. Delta value of 0.6 was taken and 70 significant genes were identified. SAM plot for corresponding delta of 0.6 has been shown in Fig. 2. Clustergram in Fig. 4 uses hierarchical clustering to visualize the classification capability of identified gene signature. It is able to correctly classify all the samples in training dataset 1.

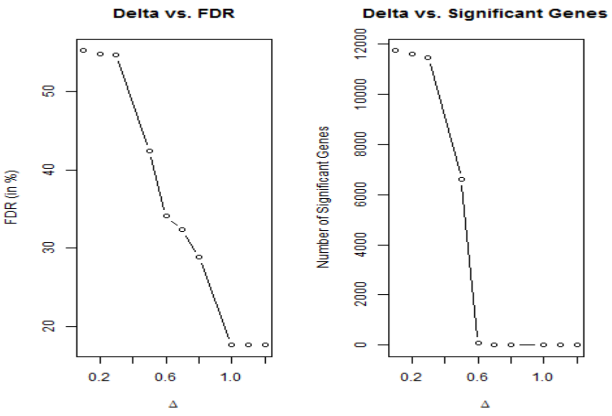

Fig. 2 SAM plot for set1 obtained from siggenes package of R

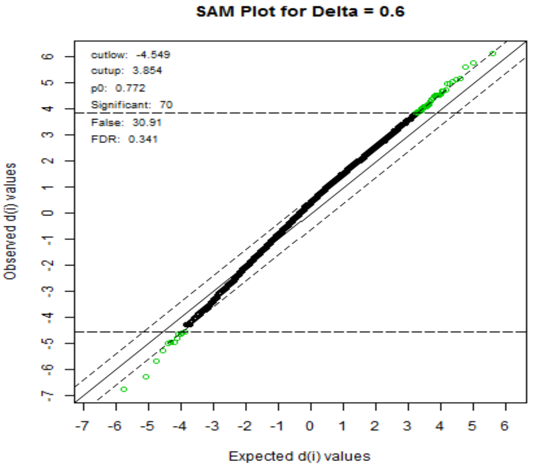

Fig. 3 SAM plot showing significant genes corresponding to delta=0.60

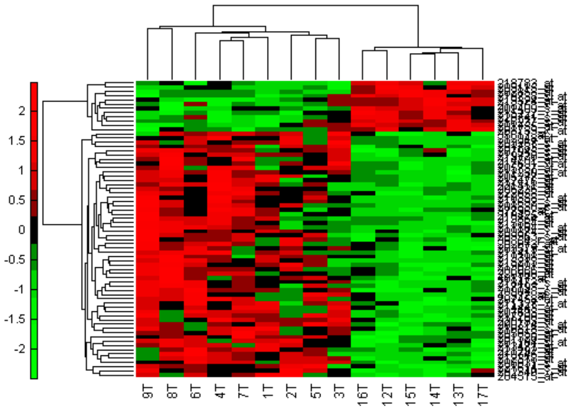

Fig. 4 Clustergram showing discrimination ability of gene signature 1

B.3.1.2 Gene Signature 2

Gene signature 2 was identified using SAM from training data set 2 which consists of 9 metastasis positive and 5 metastasis negative patients. SAM plot for the same showing delta and False Discovery Rate (FDR) for the training dataset 1 is shown in Fig. 5. Delta value of 0.565 was taken and 70

International Science Index Vol:6, No:11, 2012 waset.org/Publication/283

significant genes were identified. SAM plot for corresponding delta of 0.565 has been shown in Fig. 6. Clustergram in Fig. 7 uses hierarchical clustering to visualize the classification capability of identified gene signature. It is able to correctly classify all the samples in training dataset 2.

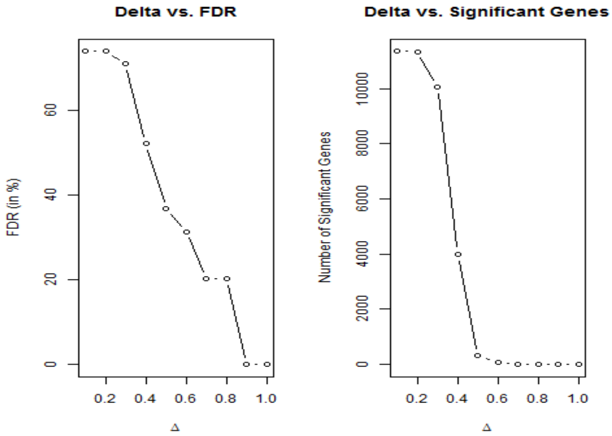

Fig. 5 SAM plot for set2 obtained from siggenes package of R

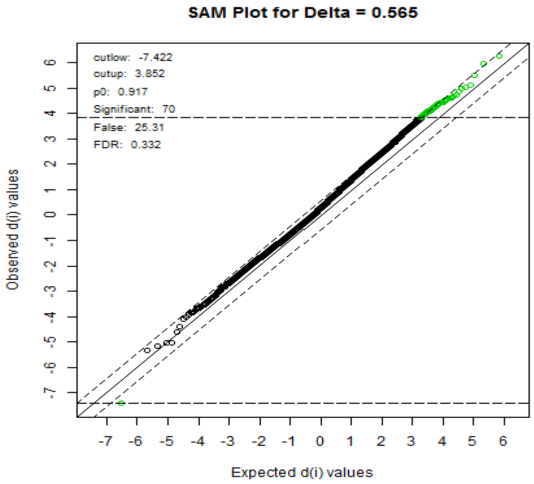

Fig. 6 SAM plot showing significant genes corresponding to delta=0.565.

B.3.1.3 Gene Signature 3

Gene signature 3 was identified using SAM from training data set 3 which consists of 10 metastasis positive and 5 metastasis negative patients. SAM plot for the same showing delta and False Discovery Rate (FDR) for the training dataset 3 is shown in Fig. 8. Delta value of 0.3425 was taken and 73 significant genes were identified. SAM plot for corresponding delta of 0.3425 has been shown in Fig. 9. Clustergram in Fig. 10 uses hierarchical clustering to visualize the classification capability of identified gene signature. It is able to correctly classify all the samples in training dataset 3.

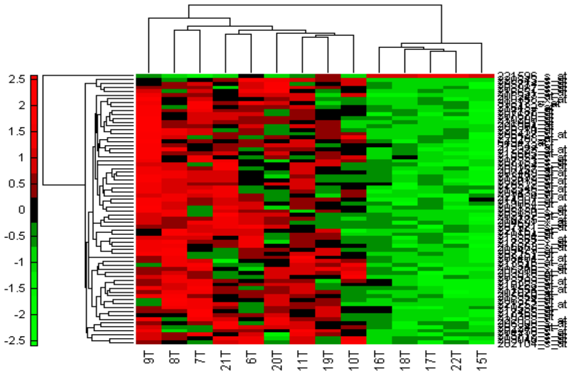

Fig. 7 Clustergram showing discrimination ability of gene signature 2

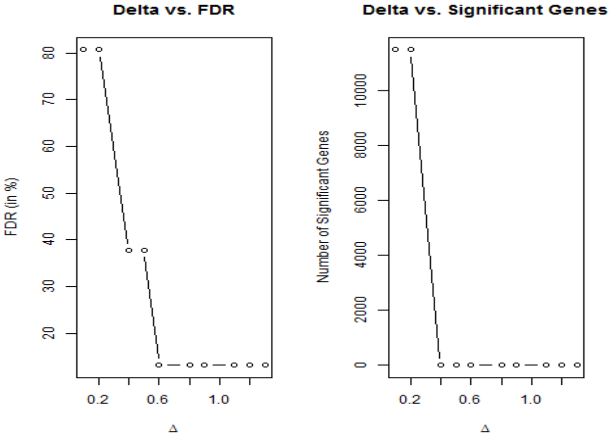

Fig. 8 SAM plot for set 3 obtained from siggenes package of R.

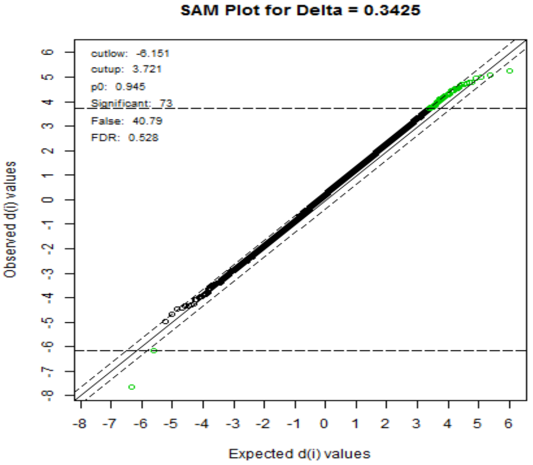

Fig. 9 SAM plot showing significant genes corresponding to delta=0.3425

International Science Index Vol:6, No:11, 2012 waset.org/Publication/283

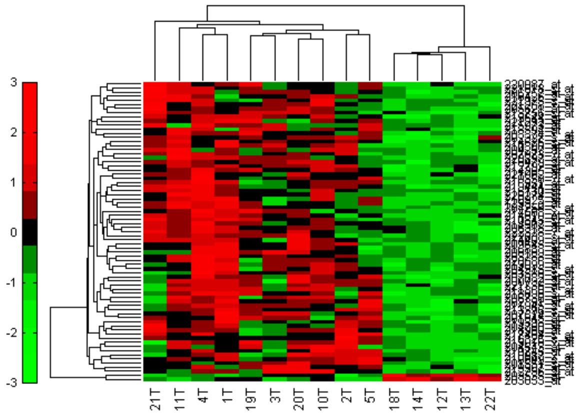

Fig. 10 Clustergram showing discrimination ability of gene signature 3

B.4. Classification Algorithms

For the purpose of classification and prediction of positive patients, models like Neural Network Models, Clustering based classification; Support Vector Machines etc. can be used.

B.4.1 k- Nearest Neighbour Classifier (k-NN)

The k-nearest neighbour classifier labels an unknown object O with the label of the majority of the k nearest neighbours. An input is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its k nearest neighbors. A neighbour is deemed nearest if it has the smallest distance, in the Euclidian sense, in feature space. For k = 1, this is the label of its closest neighbour in the learning set. A disadvantage of this method is its large computing power requirement, since for classifying an object its distance to all the objects in the learning set has to be calculated.

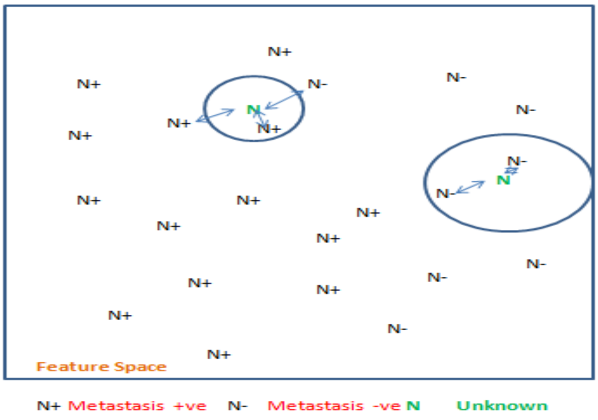

Fig. 11 Figure illustrating working of k-NN in feature space

B.4.1 Fuzzy C-Means Clustering (FCM)

FCM is a data clustering technique wherein each data point belongs to a cluster to some degree that is specified by a membership grade. It provides a method that shows how to group data points that populate some multidimensional space into a specific number of different clusters. It starts with an initial guess for the cluster centers, which are intended to mark the mean location of each cluster. The initial guess for these cluster centers is most likely incorrect. Additionally, it assigns every data point a membership grade for each cluster. By iteratively updating the cluster centers and the membership grades for each data point, it iteratively moves the cluster centers to the right location within a data set. This iteration is based on minimizing an objective function that represents the distance from any given data point to a cluster center weighted by that data point's membership grade.

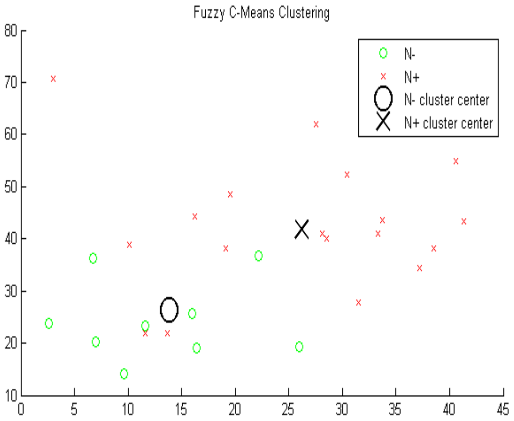

Fig. 12 Figure illustrating working of FCM

B.4.3 Support Vector Machine (SVM)

SVM is primarily a statistical method that performs classification tasks by constructing hyperplanes in a multidimensional space that separates cases of different class labels [13,14]. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other. A SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on. Fig. 4 describes the SVM working; left hand side shows the class 1 and 2 samples in input space which are not linearly separable. SVM using its kernel function maps the input space data to a higher dimensional feature space where two classes are linearly separable. SVM finds the hyperplane which has maximum margin from the class 1 and class 2 to avoid the misclassification.

International Science Index Vol:6, No:11, 2012 waset.org/Publication/283

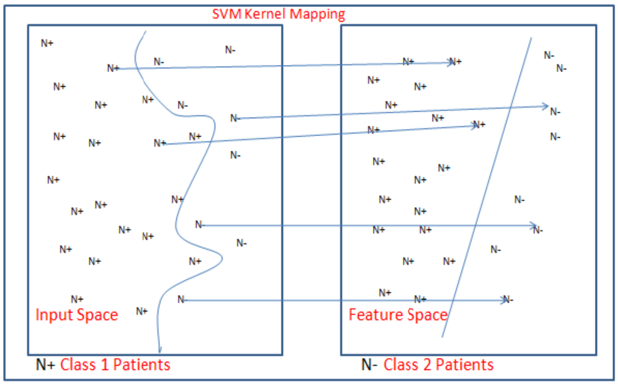

Fig. 13 Figure illustrating SVM kernel mapping from Input to Feature Space

SVM uses method to identify support vectors si, weights αi, and bias b that are used to classify vectors x according to the following equation:

$$class(c) = \sum_i \alpha_i \, k(s_i, x) + b$$

where *k* is a kernel function. We used linear kernel *k* which is the dot product. Depending upon the value of c, we decide which group the sample belong s to.

B.4.4. Backpropagation Neural Network (BPNN)

General backpropagation neural network architecture includes input layer, hidden layer and output layer. Each neuron in input layers are interconnected with neurons in hidden layers with appropriate weights assigned to them. Similarly each neuron of hidden layer in interconnected with output layer neuron with weights assigned to the connection. On providing learning data to the network, the learning values are passed through input to hidden and finally to output layer where response for input data is obtained. For optimizing the error obtained, the error values are back propagated to make changes in weights of input to hidden layer and hidden to output layer. With error back propagation input response are made converged to desired response. A general structure of BPA neural network has been shown below.

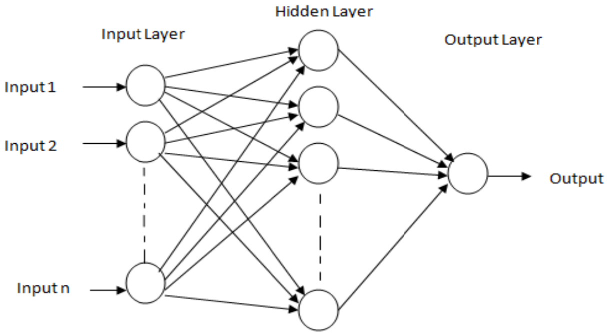

Fig. 14 General architecture of a Backpropagation Neural Network

Each neuron output is found by equation:

$$Y = f(Wij * Xi + b)$$

BPA uses supervised learning in which trainer submits the input-output exemplary patterns and the learner has to adjust the parameters of the system autonomously, so that it can yield the correct output pattern when excited with one of the given input patterns. An acceleration method for training the back propagation network, based on a gradient descent approach, is defined by:

$$Wij(k+1) = Wij(k) - \eta \frac{\delta E}{\delta Wij} + \alpha[Wij(k) - Wij(k-1)]$$

The term $\alpha[Wij(k) - Wij(k-1)]$ is called the momentum term, and α is the momentum rate. Selection of the training parameters, namely training rate η and momentum rate α, is largely a matter of experience; a small difference in these parameters can lead to large differences in training time.

B. 5. Validation and Result Calculation from Trained Models

Test Dataset is used for validating and calculating success rate of models which were trained using the training dataset.

C. Empirical Results

Table II shows models used for classification of cancer patients with their respective parameters used in this study. Table III shows the empirical results obtained on all used classification models for 3 individual gene signatures and 9 gene signature's on all 3 combination of testing dataset and lymphatic metastasis test dataset. It can be observed that performances of classifiers were significantly improved while support vector machine and backpropagation neural network were able to classify all the testing samples accurately to their corresponding class.

TABLE II
MODELS AND CORRESPONDING USED PARAMETERS

Models	**Parameters Used**
k- Nearest Neigbour	k=1, Euclidian Distance
Fuzzy C-Means Clustering	Euclidian Distance
Support Vector Machine	Linear Kernel Function
Backpropagation Neural Network	(Input, Hidden, Target)= (66,22,1) Learning Rate=0.05 Momentum=0.1

TABLE III
MODELS AND CORRESPONDING RESULTS

Model	Gene Sig. 1	Gene Sig. 2	Gene Sig. 3	9 Gene Sig.
k- Nearest	3 / 7	5 / 8	3 / 7	(5/7,6/8,7/7)
Neighbour	4 / 5	4 / 5	4 / 5	(5/5,4/5,5/5)
Fuzzy C-Means	3 / 7	5 / 8	3 / 7	(5/7,6/8,7/7)
Clustering	4 / 5	3 / 5	4 / 5	(4/5,4/5,4/5)
Support Vector	4 / 7	5 / 8	5 / 7	(7/7, 8/8,7/7)
Machine	4 / 5	4 / 5	5 / 5	(5/5,5/5,5/5)
Backpropagation	5 / 7	6 / 8	6 / 7	(7/7,8/8,7/7)
Neural Network	5 / 5	4 / 5	5 / 5	(5/5,5/5,5/5)

III. CONCLUSION

It can be concluded from results that gene signature identified due to difference between N+ and N- type can be used for the purpose of classification. 9 genes signature obtained after intersection of 3 gene signatures shows better discriminating ability than individual gene signatures. SVM and BPNN were able to accurately classify all the samples from test set and lymphatic node metastasis samples into corresponding classes using 9 genes signature. Results obtained concluded that 9 genes signature obtained from primary tumor samples can be used for classification of positive, negative metastasis cancer samples and lymphatic metastasis.

REFERENCES

[1] O'Donnell, K. R.et al. (2005), Gene expression signature predicts lymphatic metastasis in squamous cell carcinoma of the oral cavity, *Oncogene, 24, 1244–1251.*

[2] Ming-Jian Ge et al. (2009), Gene expression signature for lymphatic metastasis of human lung adenocarcinoma . *Chinese Journal of Cancer 28:3, 220-224.*

[3] Zhou X. et al. (2006), Global Expression-Based Classification of Lymph Node Metastasis and Extracapsular Spread of Oral Tongue Squamous Cell Carcinoma. *Neoplasia. Vol. 8, No. 11, pp. 925 – 932.*

[4] Van 't Veer LJ, Dai H, Van de Vijver MJ et al. (2002), Gene expression profiling predicts clinical outcome of breast cancer. Nature 415:530–536

[5] Lu Y. et al. (2006), A Gene Expression Signature Predicts Survival of Patients with Stage I Non-Small Cell Lung Cancer**.** *PLoS Medicine, Vol. 3, Issue 12, e467.*

[6] Roepman P. et al. (2006), Multiple Robust Signatures for Detecting Lymph Node Metastasis in Head and Neck Cancer**.** *Cancer Research; 66:2361-2366.*

[7] Tusher, Tibshirani and Chu (2001); "Significance analysis of microarrays applied to the ionizing radiation response", *PNAS 2001 98: 5116-5121.*

[8] Pramana J. et al. (2007), Gene Expression Profiling to Predict Outcome After Chemoradiation in Head and Neck Cancer," *International Journal of Radiation Oncology * Biology * Physics, Vol. 69, Issue 5, Pages 1544-1552.*

[9] Millenaar FF. et al. (2007), Identification of a predictive gene expression signature of cervical lymph node metastasis in oral squamous cell carcinoma, *Cancer Science, Vol.98, Issue 5, Pages 740-746.*

[10] Bertucci F., Finetti P., Cervera N et al (2006) Gene expression profiling and clinical outcome in breast cancer. *Omics 10:429–443.*

[11] Kondohet N. et al. (2007), Gene expression signatures that can discriminate oral leukoplakia subtypes and squamous cell carcinoma, *Oral Oncology, Volume 43, Issue 5, Pages 455-462.*

[12] Van de Vijver MJ, He YD, Van't Veer LJ et al. (2002), A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med 347:1999–2009.*

[13] Parry M. R. et al. (2010), k- Nearest neighbor models for microarray gene expression analysis and clinical outcome prediction, *The Pharmacogenomics Journal (2010) 10***,** *292–309.*

[14] Dembélé D. and Kastner P. (2003) Fuzzy C-means method for clustering microarray data. *Bioinformatics, 19(8): 973-980.*

[15] Han, L. K., Zeng X. and Yan H. (2008) Fuzzy clustering analysis of microarray data, *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine October 1, 2008 222: 1143-1148.*

[16] Suykens, J (2001), "Support Vector Machines : a nonlinear modelling and control perspective" , *European Journal of Control, Special Issue on fundamental issues in control, 7(2-3):311-327.*

[17] Guyon, I, Weston, J, Barnhill, S, and Vapnik, V (2002)., "Gene selection for cancer classification using support vector machines"*, Machine Learning, 46:389–422.*

[18] Chen L. and Boggess L. (2002), Neural Networks for genome signature analysis, Proceedings of the 2002 International Conference on Neural Information Processing , pp. 1554-1558.

Anupam Shukla is a Professor in the ICT Department of ABV-Indian Institute of Information Technology and Management Gwalior, India. His research interest includes speech processing, artificial intelligence, soft computing and bioinformetics. He has published more than 80 technical papers and 4 books.

A. Tarsauliya is a student of final year of 5-year Integrated Post Graduate Course (BTech + MTech in IT) in Indian Institute of Information Technology and Management Gwalior. His areas of research are artificial neural networks, hybrid system design, artificial intelligence and soft computing and biomedical engineering.

Ritu Tiwari is an Assistant Professor in the ICT Department of ABV-Indian Institute of Information Technology and Management, Gwalior, India. Her field of research includes biometrics, artificial neural networks, signal processing, robotics and soft computing. She has authored or co-authored more than 50 technical papers and 4 books.

S. Sharma is a PhD student of Information Technology at ABV-Indian Institute of Information Technology and Management, Gwalior, India. He received his MTech in Computer Science and Engineering From ABV-IIITM, Gwalior. His Area of research is Artificial intelligence, biometrics, bioinformatics etc.