Extended Principal Orthogonal Decomposition Method for Cancer Screening

Carlyn-Ann B. Lee and Charles H. Lee

***Abstract*—Recent advances in microarray technology offer the ability to study the expression of thousands of genes simultaneously. The DNA data stored on these microarray chips can provide crucial information for early clinical cancer diagnosis. The Principal Orthogonal Decomposition (POD) method has been widely used as an effective feature detection method. In this paper, we present an enhancement to the standard approach of using the POD technique as a disease detection tool. In the standard method, cancer diagnosis of an arbitrary sample is based on its correlation value with the cancerous or normal signature extracted using the POD method from DNA microarray data. In this paper, we extend the POD method by feeding the extracted principal features into Machine Learning algorithms to detect cancer. Particularly, Linear Support Vector Machine, Feed Forward Back Propagation Networks, and Self-Organizing Maps are used on liver cancer, colon cancer, and leukemia data. Sensitivity, specificity, and accuracy are discussed to evaluate predictive abilities of the proposed extended POD methods. Our results indicate overall the proposed methods provide improvements over the standard POD method.**

***Index Terms*—DNA Microarray, principal orthogonal decomposition, machine learning, artificial neural networks, support vector machine, self-organizing map, cancer detection**

I. Introduction

Expressions of thousands of individual genes can be stored in a DNA microarray, which allows one to see genes that are induced or repressed in an experiment. Signatures of a cancer may be encrypted in DNA microarrays, and once found, can be used for diagnoses. The standard Principal Orthogonal Decomposition (POD) method had been used to effectively detect liver and bladder cancers [1]-[2]. In this paper, we propose to extend the standard Principal Orthogonal Decomposition (POD) method to include Machine Learning (ML) algorithms for cancer detection. Namely, we use the POD technique to extract the principal features, both cancerous and normal. We then feed them to ML algorithms such as the Support Vector Machine (SVM), Feed Forward Back Propagation Networks (FFBPN) and Self-Organizing Map (SOM) to train the classifiers for detection of different types of cancers. Results vary depending on a priori information. We include results from varying the number of training data and genes included in the feature selection. Additionally we compare results from classifiers trained when two, four, and six modes are extracted from cancer and normal projections.

Manuscript received March 1, 2012; revised March 31, 2012.

Carlyn-Ann B. Lee is with the Department of Computer Science at California State University, Fullerton, CA 92834, USA (e-mail: cblee@csu.fullerton.edu).

Charles H. Lee is with the Department of Mathematics at California State University, Fullerton, CA 92834, USA (e-mail: CharlesHLee@fullerton.edu).

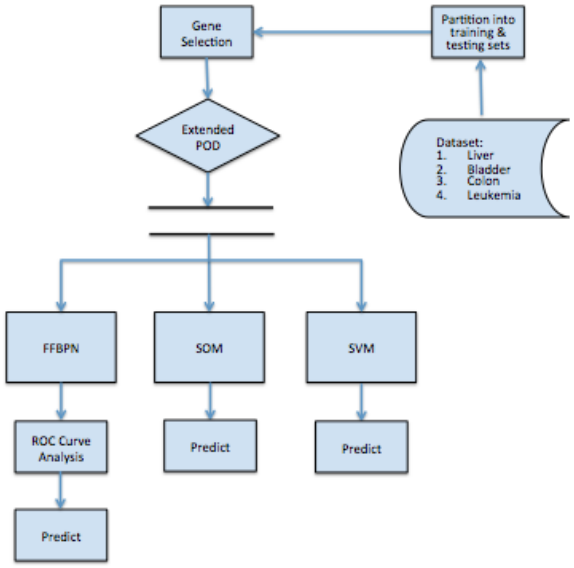


Fig. 1. The cancer screening prediction process.

II. Methodology

A. Dataset

For liver cancer detection, we examined the DNA microarray data from reference [3]. The data, containing both normal and cancerous tissues, are obtained from the Stanford Microarray Database at *genome-www5.stanford.edu.* Only genes with expressions in over 80% of the samples are included. Missing data for a particular gene are imputed with the average of the values for that gene from the other samples. The liver cancer data set contained data from 76 normal tissue samples and 105 primary liver cancer samples, where data for 5520 genes are extracted.

For colon cancer detection, we examined DNA microarray data from reference [4]. Colon cancer data consisted of 40 cancerous samples and 22 normal samples. Samples are taken from epithelial cells of colon cancer patients. The original data contained 6000 gene expression levels. Only 2000 gene expression levels are used based on the confidence in the measured expression levels.

For leukemia detection, we examined DNA microarray data from reference [5]. Leukemia data consisted of 48 samples of Acute Myeloid Leukemia (AML) and 25 samples of Acute Lymphoblastic Leukemia (ALL). The measurements are taken from 63 bone marrow samples and 10 peripheral blood samples. Data for 7129 gene expression levels are extracted.

Mean values for each gene are subtracted off before

selecting the most prominent genes for performing the orthogonal decomposition for all data. Given a cancer training set $\{T_i^C\}_{i=1}^{N_C}$ with N_C samples, and a normal training set $\{T_j^N\}_{j=1}^{N_N}$,, with N_N samples, we define the Signal-to-Noise ratio for each gene g as:

$$SNR(g) = \left| \frac{mean_{1\le i\le N_C}(T_i^C(g))}{std_{1\le i\le N_C}(T_i^C(g))} \right| + \left| \frac{mean_{1\le j\le N_C}(T_j^N(g))}{std_{1\le j\le N_C}(T_j^N(g))} \right|. \quad (1)$$

We sort the SNR values for the genes in descending order and select only the genes with the highest SNR score for our analyses. Figure 3 shows ten genes with the top ten scores using SNR using training data from colon cancer and healthy samples. Cancer samples are plotted in red along x=1-40 and healthy samples are plotted in blue along x=41-62. Expression levels for each sample are plotted against y-axis. Horizontal lines across the chart indicate the means of the cancer and healthy samples for each gene.

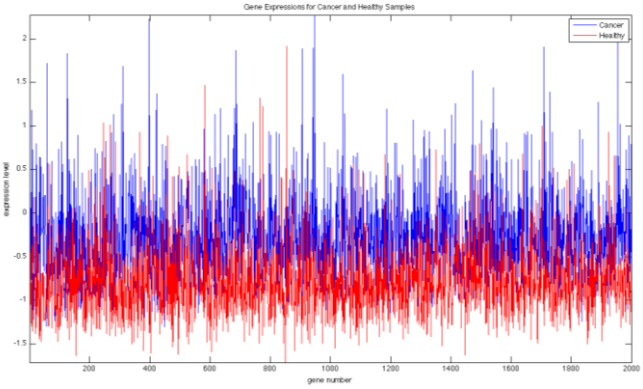


Fig. 2. All gene expressions from a colon cancer data set for arbitrary cancer sample and arbitrary normal sample.

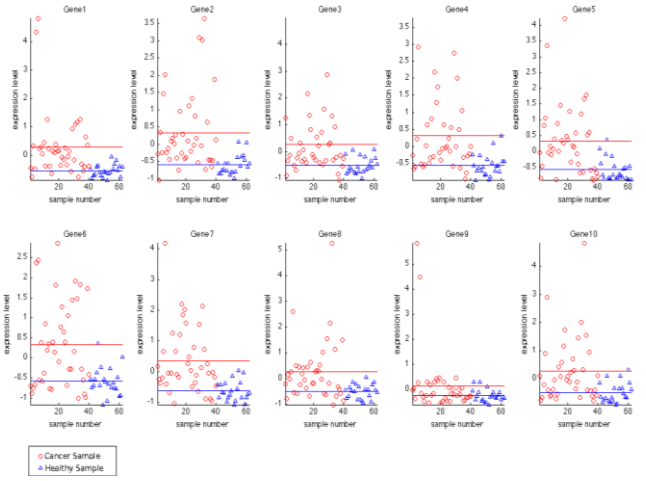


Fig. 3. Ten gene expressions with the highest SNR scores evaluated using the colon cancer testing set.

The top prominent genes with the highest SNR scores are used for analysis. The process was repeated using 10, 40 and 100 genes. Gene selection was based on random training sets. Training sets consist of 80% of cancerous samples and 80% of normal samples. The remaining samples are used for testing. The prediction values from FFBPN were mapped to a numerical range from 0 to 1. Cut-off thresholds are selected to obtain maximum fitness obtained by the training set:

$$f_{fitness}(\tau) = \text{Sensitivity}(\tau) \times \text{Specificity}(\tau). \quad (2)$$

Predictions for the testing set are evaluated using the fitness value defined in equation 2. Training and testing processes are repeated 100 times, with randomly selected training and testing partitions for each of the 100 trials. Averaged sensitivity, specificity and accuracy from these predictions are reported. Results are also reported when the number of training samples was decreased to 50% of the entire dataset and additional modes.

B. Feature Extraction

Given a cancer training set $\{T_i^C\}_{i=1}^{N_C}$ and a normal training set $\{T_j^N\}_{j=1}^{N_N}$, we apply the POD technique to extract the primary dominant features Φ^C and Φ^N, respectively. We use $\{T_k\}_{k=1}^{N_C+N_N}$ to include both training sets, whose elements can be represented as,

$$X^{(k)} = \begin{bmatrix} x_1^{(k)} \\ x_2^{(k)} \\ y^{(k)} \end{bmatrix}, \text{ where } x_1^{(k)} = \langle T_k, \Phi^C \rangle, x_2^{(k)} = \langle T_k, \Phi^N \rangle, \text{ and } y^{(k)} = \begin{cases} 0 & T_k \in \{T_j^N\}_{j=1}^{N_N} \\ 1 & T_k \in \{T_i^C\}_{i=1}^{N_C} \end{cases} \quad (3)$$

The following Machine Learning algorithms are used to construct classifiers F based on the values of $\{X^{(k)}\}_{k=1}^{N_C+N_N}$ of the training sets so that $F(X^{(k)}) = y^{(k)}$ for as many k as possible. We denote by $\{S_m^C\}_{m=1}^{M_C}$, $\{S_n^N\}_{n=1}^{M_N}$, $\{S_l\}_{l=1}^{M_C+M_N}$ the test sets of cancer, normal, and both, respectively. For each member of the testing set, we define the corresponding metric

$$X^{(l)} = \begin{bmatrix} x_1^{(l)} \\ x_2^{(l)} \\ y^{(l)} \end{bmatrix} \text{ where } x_1^{(l)} = \langle S_l, \Phi^C \rangle, x_2^{(l)} = \langle S_l, \Phi^N \rangle, \text{ and } y^{(l)} = \begin{cases} 0 & S_l \in \{S_n^N\}_{n=1}^{M_N} \\ 1 & S_l \in \{S_m^C\}_{m=1}^{M_C} \end{cases} \quad (4)$$

Figure 4 shows the projections of all samples onto the primary tumor and normal modes respectively. Horizontal lines along y=0 are drawn to show separation between projections of tumor and normal samples. The two projections are combined used as $x^{(l)}$, $y^{(l)}$, $x^{(k)}$ and, $y^{(k)}$ to construct $X^{(l)}$ and $X^{(k)}$, and are plotted in the right in figure 4. Typically, projections using primary dominant features resemble normal distribution. However, when the primary dominant features do not provide sufficient accuracy, additional modes contribute significant class information to feature sets.

Additional modes were investigated to add structure to the new features sets. Only the primary dominant modes resemble normal distributions. Figures 5 and 6 demonstrate top ranking normal and tumor modes in order of mean differences. The first charts of each figure demonstrate clear separation between two classes, with the occurrence of a few outliers. Training samples are labeled along x-axis and corresponding projections are plotted along the y-axis. Top ranking modes were used to construct $X^{(l)}$ and $X^{(k)}$. Results using the top 2, 4, and 6 modes are reported.

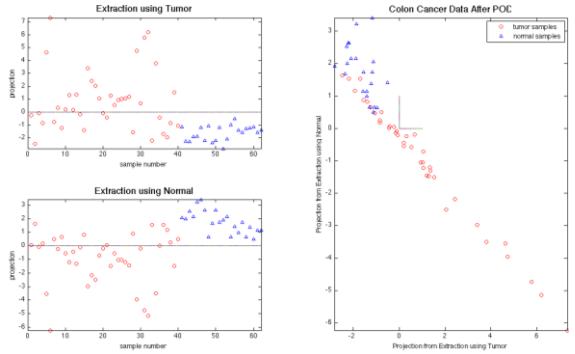


Fig. 4. Projections over primary dominant colon cancer and normal modes for all samples.

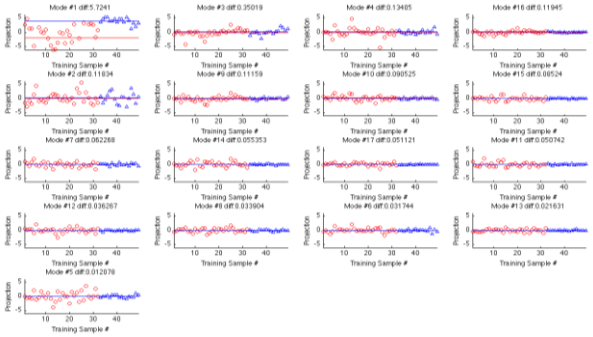


Fig. 5. All normal modes ranked in order by mean differences, with projections from training samples only.

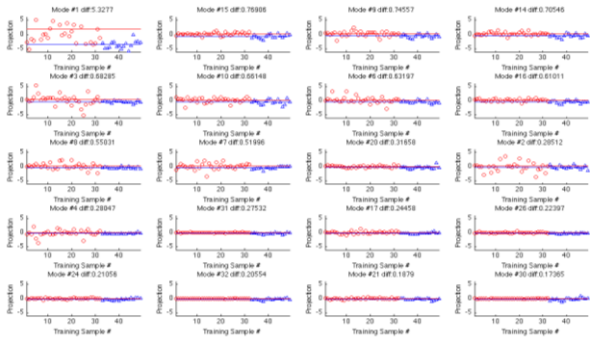


Fig. 6. Top 20 tumor modes ranked in order by mean differences, with projections from training samples only.

III. Machine Learning Models

A. Support Vector Machine

Since we perform our projections onto the dominant cancer and normal POD features, the hyper-plane is two-dimensional and SVM draws a contour between the cancerous and normal classes [6]. For simplicity, we assume that the training data is linearly separable and utilize a linear SVM. The SVM algorithm constructs the line $y = mx + b$ that maximizes the margin between the positive and negative groups. In this case, the classifier is given by

$$F_{SVM}(X^{(l)}) = \begin{cases} 0 & x_2^{(l)} > mx_1^{(l)} + b \\ 1 & \text{otherwise} \end{cases} . \quad (5)$$

B. Feed Forward Back Propagation Network

For the Feed Forward Back Propagation Network (FFBPN), we assume a simple, single layer perceptron with two inputs and one output (see [7] for more details). The FFBPN is constructed using the MATLAB command "newff", where the weights (w_1, w_2) and the bias parameter θ are found based on the training sets. The network architecture is activated by a hyperbolic tangent sigmoid function,

$$F_{FFBPN}(X^{(l)}) = \begin{cases} 0 & G(w_1x_1 + w_2x_2 + \theta) < \tau_{cutoff} \\ 1 & \text{otherwise} \end{cases} \text{ where } G(s) = \frac{e^s - e^{-s}}{e^s + e^{-s}} \quad (6)$$

Note that the cut-off value τ_{cutoff} is determined from the Receiver-Operating-Characteristic (ROC) curve and equation 2.

C. Self Organizing Map

SOM starts out with an initial two-dimensional map and, when introduced to the training set, it updates the map iteratively to fit the distribution of the clusters in the training set. When a testing set is fed into the map, the map classifies it according to its nearest cluster of the training set. We implement the SOM scheme using four neighborhood functions (Bubble, Gaussian, Cut-Gaussian, and Epanechicov) sequentially to exhaust all possible maps. Both the batch and the sequential training algorithms are also explored in this study. SOMs are implemented using the somtoolbox (see [8] for further details).

IV. Results

Sensitivity, specificity, and accuracy are used to determine the performance of classifiers in this study. Sensitivity measures the ability to correctly identify those with the disease, whereas specificity measures the ability to identify those without the disease. Accuracy shows the ratio of true predictions (true positives and true negatives) out of all predictions. For all test set predictions, the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) are determined. Sensitivity, specificity, and accuracy are evaluated to determine the quality of the network:

$$\text{Sensitivity} = \frac{TP}{TP + FN}, \quad \text{Specificity} = \frac{TN}{TN + FP}, \quad \text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}. \quad (7)$$

Results using extended POD with machine learning techniques demonstrate slightly improved predictions when compared to the standard POD method. A study [9] using SVM and SOM without POD on colon and leukemia cancers has been compared to our proposed methods. Our methods produce better results (+90% versus +70%); however, there are different assumptions such as holdout percentages (20% versus 50%) for training and testing. Results reported are obtained using 40 genes selected from 80% of the raw data for training standard and extended POD methods.

A. Liver Cancer Data

POD feature extraction for one trial is plotted in Fig. 7. The horizontal axis is the case number and the vertical axis represents its projection value. Cancerous and normal samples are numbered 1-105 and 106-181, respectively. The line y=0 is drawn horizontally along the plots for each

projection in Fig. 7. A large percentage of projections from cancerous tissue samples exceeded y=0. Similarly, a large percentage of projections from normal tissue samples are less than y=0. In this case, the standard POD method performs rather well as a predictive classifier. The ROC curve for training data using POD cancerous feature (blue), POD normal feature (green), and FFBPN (red) are shown in Figure 8. Points with the largest fitness values are circled and the corresponding cut-off thresholds are used for predicting the test set. In addition, we find from Figure 8 that while the FFBPN obtains a smaller false positive rate than the POD normal feature, it obtains a higher true positive rate than the POD cancer feature. The SOM method, displayed in Fig. 9, indicates distinct cancerous and normal clusters. Labeled neurons show that only a small percentage of the map neurons have predictive capabilities prior to pruning. The SVM hyper-plane, shown in Fig. 10, is constructed using the training set, denoted with red and green. Test data is denoted in magenta and cyan. Average accuracies for five random trials and all classifiers are shown in Fig. 11.

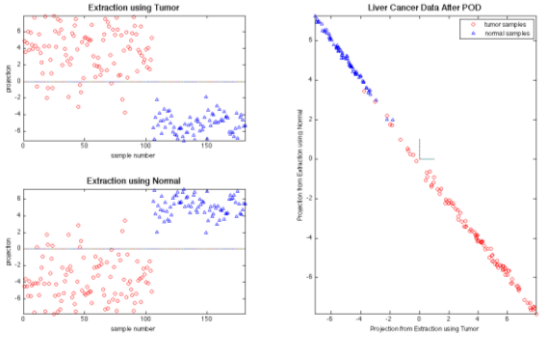


Fig. 7. Projections over primary dominant liver cancer and normal modes for all samples.

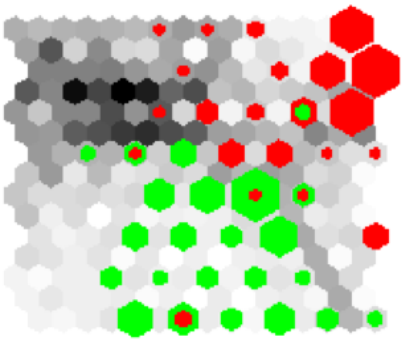


Fig. 8. SOM constructed with primary dominant modes of training liver cancer and normal samples. Weights for mapping of tumor samples are indicated in red, and normal samples are indicated in green.

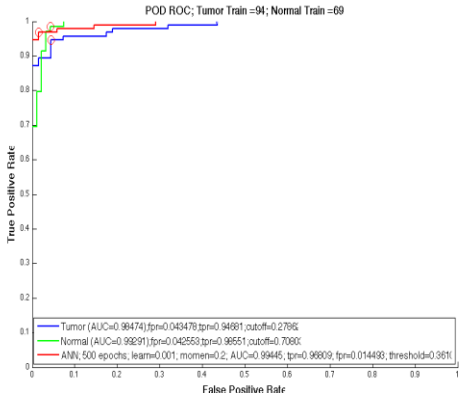


Fig. 9. ROC curves for training sets using standard and extended POD with FFBPN. Cutoffs with max fitness are circled in red.

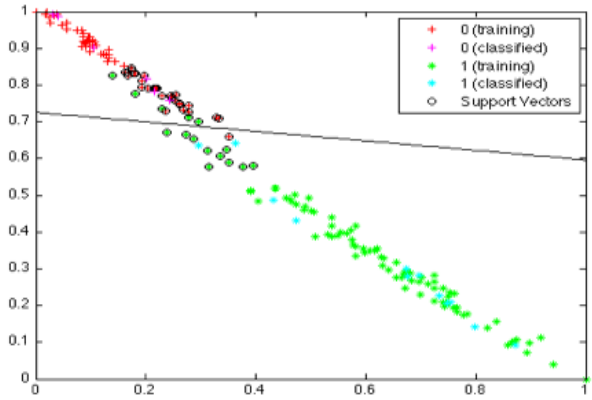
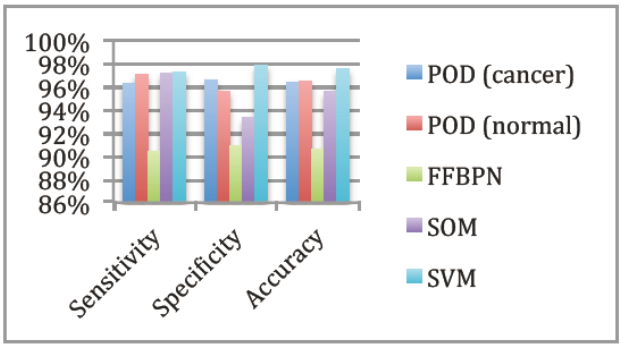


Fig. 10. The SVM hyper-plane, shown in, is constructed using the training set, denoted with red and green.

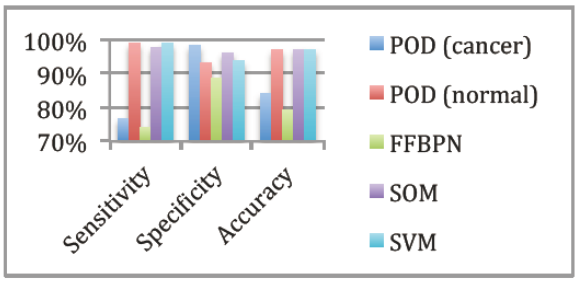


	POD (cancer)	POD (normal)	FFBPN	SOM	SVM
Sensitivity	96.29%	97.10%	90.38%	97.19%	97.29%
Specificity	96.62%	95.63%	90.88%	93.33%	97.88%
Accuracy	96.43%	96.48%	90.59%	95.57%	97.54%

Fig. 11. Average sensitivities, specificities and accuracies for screening liver cancer from 100 random trials.

B. *Leukemia Data*

In this data set, there are no normal samples and the classes are AML and ALL. Here we treat the ALL samples as if they are normal samples. Results from AML and ALL predictions are shown in Figure 12. Accuracy of extended POD predictions improved only slightly using machine-learning techniques. Accuracy for this data using SVM and SOM for extended POD exceed recognition rate for feature selection methods proposed in [9]. Furthermore, using extended POD for prediction of unknowns obtains slightly better results compared to a majority vote ensemble classifier [9] (97.3% versus 97.1%).

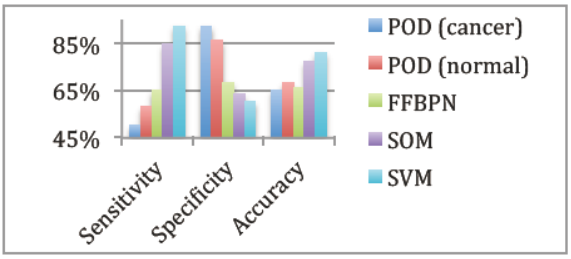


	POD (cancer)	POD (normal)	FFBPN	SOM	SVM
Sensitivity	76.70%	98.98%	74.21%	97.81%	98.98%
Specificity	98.40%	93.40%	88.80%	96.20%	93.80%
Accuracy	84.15%	97.05%	79.20%	97.26%	97.20%

Fig. 12. Average sensitivities, specificities and accuracies for screening AML leukemia from 100 random trials.

C. Colon Cancer Data

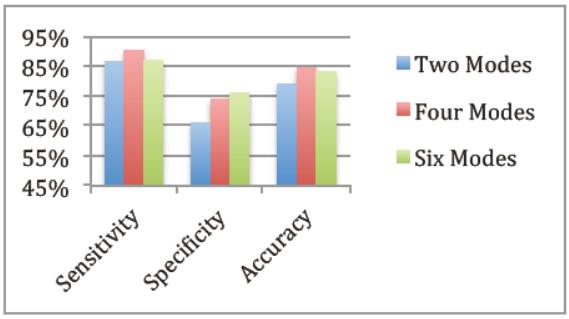
In the results for liver cancer and leukemia, there is little improvement in overall accuracy for extended POD compared to standard POD. Compelling improvements using extended POD were obtained in screening of colon cancer, where standard POD obtained accuracy <70%. Prediction results from screening of colon cancer test data using our proposed methods are summarized in Figure 13. Sensitivity, specificity, and accuracy for extended POD using linear SVM demonstrates improvement to the standard POD with primary dominant features. As shown in figures 5 and 6, projections over the primary dominant modes resemble a normal distribution. Results from extended POD using SVM trained with only primary dominant modes are comparable to accuracy measures obtained using a parametric Gaussian classifier on the primary dominant modes. A Gaussian classifier obtains sensitivity=92.37%, specificity=60.45%, and accuracy=81.08%. Additional modes deviate from the typical normal distribution, so as modes are added to the feature vector, results obtained using linear SVM tend to be exceed results obtained with Gaussian classifier.



	POD (cancer)	POD (normal)	FFBPN	SOM	SVM
Sensitivity	50.38%	58.13%	65.00%	84.50%	86.75%
Specificity	92.30%	86.10%	68.60%	63.65%	66.25%
Accuracy	65.35%	68.10%	66.20%	77.25%	79.27%

Fig. 13. Average sensitivities, specificities and accuracies for screening colon cancer from 100 random trials.

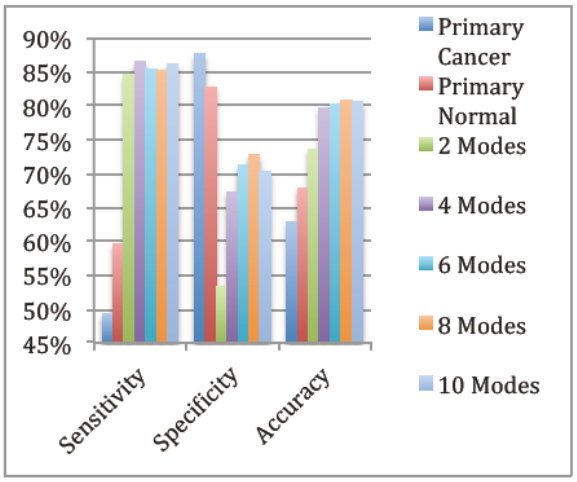
Figure 14 summarizes extended POD results using feature vectors with 2, 4, and 6 total modes respectively. Using 80% of data samples, screening of colon cancer using extended POD trained with two modes from each class provides optimal results in overall accuracy. Additional modes show little to no improvement for the configuration.



	Two Modes	Four Modes	Six Modes
Sensitivity	86.75%	90.50%	87.25%
Specificity	66.25%	74.25%	76.35%
Accuracy	79.27%	84.71%	83.40%

Fig. 14. Average sensitivities, specificities and accuracies for screening colon cancer from 100 random trials using extended POD with feature vectors containing 2, 4, and 6 modes.

Figure 15 shows improvements only 50% of the data for training 100 genes expressions. Predictions are made using extended POD and 1, 2, 4, 6, 8, and 10 modes. Introduction of multiple nodes increases sensitivity and accuracy, at the expense of specificity. Peak accuracy was obtained with 4 normal and 4 cancer modes.



	Primary Cancer	Primary Normal	2	4	6	8	10
Sen	49.35%	59.70%	84.70%	86.60%	85.35%	85.25%	86.25%
Spc	87.64%	82.73%	53.36%	67.36%	71.27%	72.82%	70.36%
Acc	62.94%	67.87%	73.58%	79.77%	80.35%	80.84%	80.61%

Fig. 15. Average sensitivities, specificities and accuracies for screening colon cancer from 100 random trials using extended POD with feature vectors containing 1, 2, 4, and 6, 8, and 10 modes.

V. Summary and Conclusions

The results from 100 random trials for extended and standard POD methods demonstrate that extended POD method improves overall accuracy compared to standard methods. For liver cancer and leukemia, the standard POD alone was capable of extracting linearly separable sets, and was sufficient for obtaining accurate results. In such cases, when standard POD using only the primary dominant normal mode provided sufficient results (accuracy >95%), the extended methods improved specificity.

On the other hand, for colon cancer screening, the standard POD requires the extended method to obtain sufficient results. In such cases, when the standard POD screened poorly (accuracy<70%), extended POD with primary dominant features only improves sensitivity significantly. Although this is at the expense of specificity, overall accuracy using the extended method exceeds that of the standard method.

Further investigation demonstrated that addition of several modes in the feature set contributes more accurate results. Since class projections using additional modes do not exhibit normal distributions, use of parametric learning techniques tend to over fit the data. Machine learning techniques, particularly SVM, demonstrated promising results, obtaining increased accuracy while still generalizing the structure of class information from additional modes.

On average, the extended POD with linear kernel support vector machine outperformed machine-learning algorithms described in this study. Although extended POD with only primary dominant features for colon cancer screening results in high occurrence of false alarms, this method improves

sensitivity enough to improve overall accuracy. Use of additional modes recovers loss in specificity, especially when only a few training samples are known. For predicting colon cancer with few training samples (only 50% of entire dataset), best results were obtained with a feature set consisting of projections from 4 tumor and 4 normal modes.

ACKNOWLEDGMENT

Computation and graphics were generated with Matlab2010. SOM Toolbox was used for constructing all self-organizing maps. The authors would like to thank Professor Spiros Courellis of the department of Computer Science at California State University, Fullerton for his sharing his knowledge of machine learning and inspiring this study.

REFERENCES

[1] D. Peterson and C. H. Lee, "A DNA-based Pattern Recognition Technique for Cancer Detection," in *Proceedings of the 26th Annual Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 2 , pp. 2956-2959, 2004.

[2] C. H. Lee and N. Abbasi, "Feature Extraction Techniques on DNA Microarray Data for Cancer Detection," *2007 World Congress on Bioengineering Proceedings*, Bangkok, Thailand, July 2007

[3] X. Chen, et al. "Gene expression patterns in human liver cancers," *Molecular Biology of the Cel*l. 2002, vol. 13, pp. 1929-1939.

[4] U. Alon, et al., "Broad patterns of gene expression revealed by clustering analysis of cancer and normal colon tissues probed by oligonucleotide arrays," in *Proc. Natl. Acad. Sci*. vol. 96, pp. 6745–6750.

[5] T. R. Golub and D. K. Slonim, *et al*. "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*. 1999, vol. 286, pp. 531-537.

[6] S. Abe, *Support Vector Machines for Pattern Classification*, Springer, *2005*

[7] C. Bishop, *Pattern Recognition and Machine Learning*, Springer-Verlag, 2006

[8] J. Vesanto, J. Himberg, E. Alhoniemi, and J. Parhankangas, SOM Toolbox for Matlab 5, report, Helsinki Univ. of Technol., Helsinki, Finland, 2000.

[9] S. Cho and H. Won. "Machine learning in DNA microarray analysis for cancer classification," in *Proceedings of the First Asia-Pacific bioinformatics conference on Bioinformatics.* APBC '03. 2003.