Spectroscopic Detection of Cervical Pre-Cancer through Radial Basis Function Networks

Kagan Tumer
kagan@pine.ece.utexas.edu
Dept. of Electrical and Computer Engr.
The University of Texas at Austin

Nirmala Ramanujam
nimmi@ccwf.cc.utexas.edu
Biomedical Engineering Program
The University of Texas at Austin

Rebecca Richards-Kortum
kortum@mail.utexas.edu
Biomedical Engineering Program
The University of Texas at Austin

Joydeep Ghosh
ghosh@ece.utexas.edu
Dept. of Electrical and Computer Engr.
The University of Texas at Austin

Abstract

The mortality related to cervical cancer can be substantially reduced through early detection and treatment. However, current detection techniques, such as Pap smear and colposcopy, fail to achieve a concurrently high sensitivity and specificity. *In vivo* fluorescence spectroscopy is a technique which quickly, non-invasively and quantitatively probes the biochemical and morphological changes that occur in pre-cancerous tissue. RBF ensemble algorithms based on such spectra provide automated, and near real-time implementation of pre-cancer detection in the hands of non-experts. The results are more reliable, direct and accurate than those achieved by either human experts or multivariate statistical algorithms.

1 Introduction

Cervical carcinoma is the second most common cancer in women worldwide, exceeded only by breast cancer (Ramanujam et al., 1996). The mortality related to cervical cancer can be reduced if this disease is detected at the pre-cancerous state, known as squamous intraepithelial lesion (SIL). Currently, a Pap smear is used to

screen for cervical cancer (Kurman et al., 1994). In a Pap test, a large number of cells obtained by scraping the cervical epithelium are smeared onto a slide which is then fixed and stained for cytologic examination. The Pap smear is unable to achieve a concurrently high sensitivity[1] and high specificity[2] due to both sampling and reading errors (Fahey et al., 1995). Furthermore, reading Pap smears is extremely labor intensive and requires highly trained professionals. A patient with a Pap smear interpreted as indicating the presence of SIL is followed up by a diagnostic procedure called colposcopy. Since this procedure involves biopsy, which requires histologic evaluation, diagnosis is not immediate.

In vivo fluorescence spectroscopy is a technique which has the capability to quickly, non-invasively and quantitatively probe the biochemical and morphological changes that occur as tissue becomes neoplastic. The measured spectral information can be correlated to tissue histo-pathology, the current "gold standard" to develop clinically effective screening and diagnostic algorithms. These mathematical algorithms can be implemented in software thereby, enabling automated, fast, non-invasive and accurate pre-cancer screening and diagnosis in hands of non-experts.

A screening and diagnostic technique for human cervical pre-cancer based on laser induced fluorescence spectroscopy has been developed recently (Ramanujam et al., 1996). Screening and diagnosis was achieved using a multivariate statistical algorithm (MSA) based on principal component analysis and logistic discrimination of tissue spectra acquired *in vivo*. Furthermore, we designed Radial Basis Function (RBF) network ensembles to improve the accuracy of the multivariate statistical algorithm, and to simplify the decision making process. Section 2 presents the data collection/processing techniques. In Section 3, we discuss the MSA, and describe the neural network based methods. Section 4 contains the experimental results and compares the neural network results to both the results of the MSA and to current clinical detection methods. A discussion of the results is given in Section 5.

2 Data Collection and Processing

A portable fluorimeter consisting of two nitrogen pumped-dye lasers, a fiber-optic probe and a polychromator coupled to an intensified diode array controlled by an optical multi-channel analyzer was utilized to measure fluorescence spectra from the cervix *in vivo* at three excitation wavelengths: 337, 380 and 460 nm (Ramanujam et al., 1996). Tissue biopsies were obtained only from abnormal sites identified by colposcopy and subsequently analyzed by the probe to comply with routine patient care procedure. Hemotoxylin and eosin stained sections of each biopsy specimen were evaluated by a panel of four board certified pathologists and a consensus diagnosis was established using the Bethesda classification system. Samples were classified as normal squamous (NS), normal columnar (NC), low grade (LG) SIL and high grade (HG) SIL. Table 1 provides the number of samples in the training (calibration) and test sets. Based on this data set, a clinically useful algorithm needs to discriminate SILs from the normal tissue types.

Figure 1 illustrates average fluorescence spectra per site acquired from cervical sites at 337 nm excitation from a typical patient. Evaluation of the spectra at 337 nm ex-

[1]Sensitivity is the correct classification percentage on the pre-cancerous tissue samples.
[2]Specificity is the correct classification percentage on normal tissue samples.

Table 1: Histo-pathologic classification of samples.

Histo-pathology	Training Set	Test Set
Normal	107 (SN: 94; SC: 13)	108 (SN: 94; SC: 14)
SIL	58 (LG: 23; HG: 35)	59 (LG: 24; HG: 35)

citation highlights one of the classification difficulties, namely that the fluorescence intensity of SILs (LG and HG) is less than that of the corresponding normal squamous tissue and greater than that of the corresponding normal columnar tissue over the entire emission spectrum[3]. Fluorescence spectra at all three excitation wavelengths comprise of a total of 161 excitation-emission wavelengths pairs. However, there is a significant cost penalty for using all 161 values. To alleviate this concern, a more cost-effective fluorescence imaging system was developed, using component loadings calculated from principal component analysis. Thus, the number of required fluorescence excitation-emission wavelength pairs were reduced from 161 to 13 with a minimal drop in classification accuracy (Ramanujam et al., 1996).

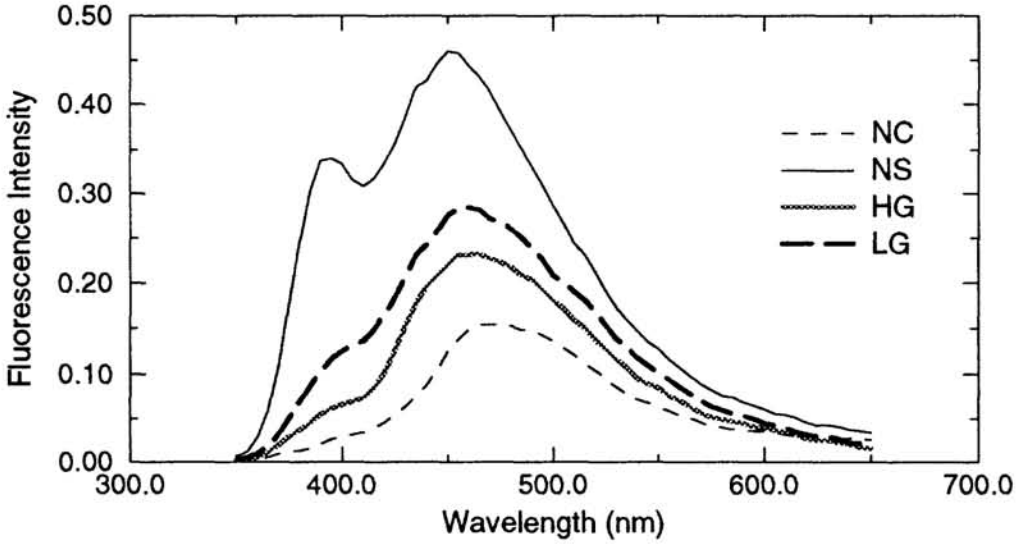

Figure 1: Fluorecsence spectra from a typical patient at 337 nm excitation.

3 Algorithm Development

3.1 Multivariate Statistical Algorithms

The multivariate statistical algorithm development described in (Ramanujam et al., 1996) consists of the following five steps: (1) pre-processing to reduce inter-patient and intra-patient variation of spectra from a tissue type, (2) dimension reduction of the pre-processed tissue spectra using Principal Component Analysis (PCA), (3) selection of diagnostically relevant principal components, (4) development of a classification algorithm based on logistic discrimination, and finally (5) retrospective and prospective evaluation of the algorithm's accuracy on a training (calibration) and test (prediction) set, respectively. Discrimination between SILs and the two normal tissue types could not be achieved effectively using MSA. Therefore two

[3]Spectral features observed in Figure 1 are representative of those measured at 380 nm and 460 nm excitation (not shown here).

constituent algorithms were developed: algorithm (1), to discriminate between SILs and normal squamous tissues, and algorithm (2), to discriminate between SILs and normal columnar tissues (Ramanujam et al., 1996).

3.2 Algorithms based on Neural Networks

The second stage of algorithm development consists of evaluating the applicability of neural networks to this problem. Initially, both Multi-Layered Perceptrons (MLPs) and Radial Basis function (RBF) networks were considered. However, MLPs failed to improve upon the MSA results for both algorithms (1) and (2), and frequently converged to spurious solutions. Therefore, our study focuses on RBF networks and RBF network ensembles.

Radial Basis Function Networks: The first step in applying RBF networks to this problem consisted of retracing the two-step process outlined for the multivariate statistical algorithm. For *constituent* algorithm (1) the kernels were initialized using a k-means clustering algorithm on the training set containing NS tissue samples and SILs. The RBF networks had 10 kernels, whose locations and spreads were adjusted during training. For *constituent* algorithm (2), we selected 10 kernels, half of which were fixed to patterns from the columnar normal class, while the other half were initialized using a k-means algorithm. Neither the kernel locations nor their spreads were adjusted during training. This process was adopted to rectify the large discrepancy between the samples from each category (13 for columnar normal vs. 58 for SILs). For each algorithm, the training time was estimated by maximizing the performance on one validation set. Once the stopping time was established, 20 cases were run for each algorithm[4].

Linear and Order statistics Combiners: There were significant variations among different runs of the RBF networks for all three algorithms. Therefore, selecting the "best" classifier was not the ideal choice. First, the definition of "best" depends on the selection of the validation set, making it difficult to ascertain whether one network will outperform all others given a different test set, as the validation sets are small. Second, selecting only one classifier discards a large amount of potentially relevant information. In order to use all the available data, and to increase both the performance and the reliability of the methods, the outputs of RBF networks were pooled before a classification decision was made.

The concept of combining classifier outputs[5] has been explored in a multitude of articles (Hansen and Salamon, 1990; Wolpert, 1992). In this article we use the median combiner, which belongs to the class order statistics combiners introduced in (Tumer and Ghosh, 1995), and the averaging combiner, which performs an arithmetic average of the corresponding outputs.

4 Results

Two-step algorithm: The ensemble results reported are based on pooling 20 different runs of RBF networks, initialized and trained as described in the previous section. This procedure was repeated 10 times to ascertain the reliability of the

[4]Each run has a different initialization set of kernels/spreads/weights.

[5]An extensive bibliography is available in (Tumer and Ghosh, 1996).

result and to obtain the standard deviations. For an application such as pre-cancer detection, the cost of a misclassification varies greatly from one class to another. Erroneously labeling a healthy tissue as pre-cancerous can be corrected when further tests are performed. Labeling a pre-cancerous tissue as healthy however, can lead to disastrous consequences. Therefore, for algorithm (1), we have increased the cost of a misclassified SIL until the sensitivity[6] reached a satisfactory level. The sensitivity and specificity values for *constituent* algorithm (1) based on both MSA and RBF ensembles are provided in Table 2. Table 3 presents sensitivity and specificity values for *constituent* algorithm (2) obtained from MSA and RBF ensembles[7]. For both algorithms (1) and (2), the RBF based combiners provide higher specificity than the MSA. The median combiner provides results similar to those of the average combiner, except for algorithm (2) where it provides better specificity. In order to obtain the final discrimination between normal tissue and SILs, *constituent* algorithms (1) and (2) are used sequentially, and the results are reported in Table 4.

Table 2: Accuracy of *constituent* algorithm (1) for differentiating SILs and normal squamous tissues, using MSA and RBF ensembles.

Algorithm	Specificity	Sensitivity
MSA	63%	90%
RBF-ave	66% ±1%	90% ±0%
RBF-med	66% ±1%	90% ±1%

Table 3: Accuracy of *constituent* algorithm (2) for differentiating SILs and normal columnar tissues, using MSA and RBF ensembles.

Algorithm	Specificity	Sensitivity
MSA	36%	97%
RBF-ave	37% ±5%	97% ±0%
RBF-med	44% ±7%	97% ±0%

One-step algorithm: The results presented above are based on the multi-step algorithm specifically developed for the MSA, which could not consolidate algorithms (1) and (2) into one step. Since the ultimate goal of these two algorithms is to separate SILs from normal tissue samples, a given pattern has to be processed through both algorithms. In order to simplify this decision process, we designed a one step RBF network to perform this separation. Because the pre-processing for algorithms (1) and (2) is different[8], the input space is now 26-dimensional. We initialized 10 kernels using a k-means algorithm on a trimmed[9] version of the training set. The kernel locations and spreads were not adjusted during training. The cost of a misclassified SIL was set at 2.5 times the cost of a misclassified normal tissue

[6]In this case, the cost of misclassifying a SIL was three times the cost of misclassifying a normal tissue sample.

[7]In this case, there was no need to increase the cost of a misclassified SIL, because of the high prominence of SILs in the training set.

[8]Normalization vs. normalization followed by mean scaling.

[9]The trimmed set has the same number of patterns from each class. Thus, it forces each class to have a similar number of kernels. This set is used *only* for initializing the kernels.

sample, in order to provide the best sensitivity/specificity pair. The average and median combiner results are obtained by pooling 20 RBF networks[10].

Table 4: One step RBF algorithm compared to multi-step MSA and clinical methods for differentiating SILs and normal tissue samples.

Algorithm	Specificity	Sensitivity
2-step MSA	63%	83%
2-step RBF-ave	65% ±2%	87% ±1%
2-step RBF-med	67% ±2%	87% ±1%
RBF-ave	67% ±.75%	91% ±1.5%
RBF-med	65.5% ±.5%	91% ±1%
Pap smear (human expert)	68% ±21%	62% ±23%
Colposcopy (human expert)	48%±23 %	94% ±6%

The results of both the two-step and one-step RBF algorithms and the results of the two-step MSA are compared to the accuracy of Pap smear screening and colposcopy in expert hands in Table 4. A comparison of one-step RBF algorithms to the two-step RBF algorithms indicates that the one-step algorithms have similar specificities, but a moderate improvement in sensitivity relative to the two-step algorithms. Compared to the MSA, the one-step RBF algorithms have a slightly decreased specificity, but a substantially improved sensitivity. In addition to the improved sensitivity, the one step RBF algorithms simplify the decision making process. A comparison between the one step RBF algorithms and Pap smear screening indicates that the RBF algorithms have a nearly 30% improvement in sensitivity with no compromise in specificity; when compared to colposcopy in expert hands, the RBF ensemble algorithms maintain the sensitivity of expert colposcopists, while improving the specificity by almost 20%. Figure 2 shows the trade-off between specificity and sensitivity for clinical methods, MSA and RBF ensembles, obtained by changing the misclassification cost. The RBF ensembles provide better sensitivity and higher reliability than any other method for a given specificity value.

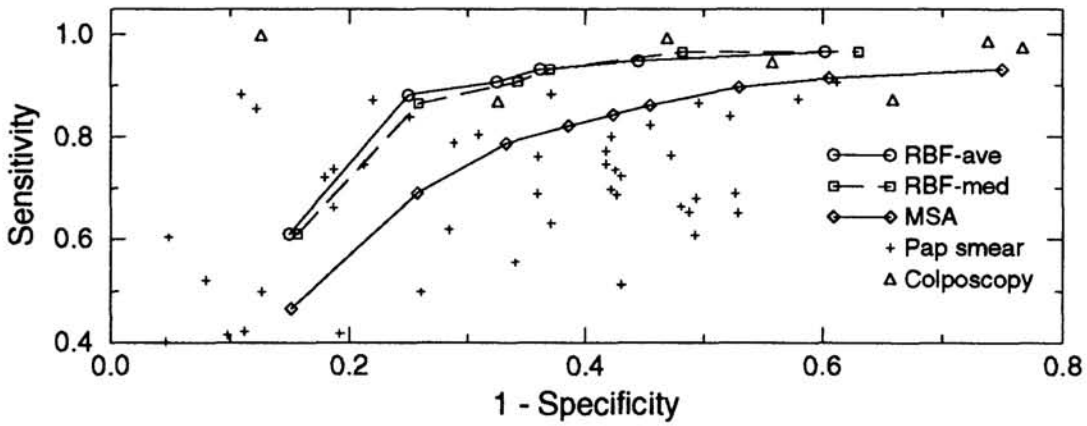

Figure 2: Trade-off between sensitivity and specifity for MSA and RBF ensembles. For reference, Pap smear and colposcopy results from the literature are included (Fahey et al., 1995).

[10]This procedure is repeated 10 times, in order to determine the standard deviation.

5 Discussion

The classification results of both the multivariate statistical algorithms and the radial basis function network ensembles demonstrate that significant improvement in classification accuracy can be achieved over current clinical detection modalities using cervical tissue spectral data obtained from *in vivo* fluorescence spectroscopy. The one-step RBF algorithm has the potential to significantly reduce the number of pre-cancerous cases missed by Pap smear screening and the number of normal tissues misdiagnosed by expert colposcopists.

The qualitative nature of current clinical detection modalities leads to a significant variability in classification accuracy. For example, estimates of the sensitivity and specificity of Pap smear screening have been shown to range from 11-99% and 14-97%, respectively (Fahey et al., 1995). This limitation can be addressed by the RBF network ensembles which demonstrate a significantly smaller variability in classification accuracy therefore enabling more reliable classification. In addition to demonstrating a superior sensitivity, the RBF ensembles simplify the decision making process of the two-step algorithms based on MSA into a single step that discriminates between SILs and normal tissues. We note that for the given data set, both MSA and MLP were unable to provide satisfactory solutions in one step.

The one-step algorithm development process can be readily implemented in software, enabling automated detection of cervical pre-cancer. It provides near real time implementation of pre-cancer detection in the hands of non-experts, and can lead to wide-scale implementation of screening and diagnosis and more effective patient management in the prevention of cervical cancer. The success of this application will represent an important step forward in both medical laser spectroscopy and gynecologic oncology.

Acknowledgements: This research was supported in part by NSF grant ECS 9307632, AFOSR contract F49620-93-1-0307, and Lifespex, Inc.

References

Fahey, M. T., Irwig, L., and Macaskill, P. (1995). Meta-analysis of pap test accuracy. *American Journal of Epidemiology*, 141(7):680–689.

Hansen, L. K. and Salamon, P. (1990). Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(10):993–1000.

Kurman, R. J., Henson, D. E., Herbst, A. L., Noller, K. L., and Schiffman, M. H. (1994). Interim guidelines of management of abnormal cervical cytology. *Journal of American Medical Association*, 271:1866–1869.

Ramanujam, N., Mitchell, M. F., Mahadevan, A., Thomsen, S., Malpica, A., Wright, T., Atkinson, N., and Richards-Kortum, R. R. (1996). Cervical pre-cancer detecion using a multivariate statistical algorithm based on fluorescence spectra at multiple excitation wavelengths. *Photochemistry and Photobiology*, 64(4):720–735.

Tumer, K. and Ghosh, J. (1995). Order statistics combiners for neural classifiers. In *Proceedings of the World Congress on Neural Networks*, pages I:31–34, Washington D.C. INNS Press.

Tumer, K. and Ghosh, J. (1996). Error correlation and error reduction in ensemble classifiers. *Connection Science*. (to appear).

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, 5:241–259.