Computer-aided identification of ovarian cancer in confocal microendoscope images

Saurabh Srivastava
University of Arizona
Department of Electrical & Computer Engineering
360 W. 34th St., Apt. K
New York, New York 10001
E-mail: ss2922@columbia.edu

Jeffrey J. Rodríguez
University of Arizona
Department of Electrical & Computer Engineering
1230 E. Speedway Blvd.
P.O. Box 210104
Tucson, Arizona 85721

Andrew R. Rouse
University of Arizona
Department of Radiology
Radiology Research Laboratory
P.O. 245067
Tucson, Arizona 85724

Molly A. Brewer
University of Arizona
Department of Obstetrics & Gynecology
Arizona Cancer Center
Rm. 1968G
P.O. Box 245024
1515 N. Campbell Avenue
Tucson, Arizona 85724

Arthur F. Gmitro
University of Arizona
Department of Radiology and Optical Sciences
Radiology Research Laboratory
P.O. 245067
Tucson, Arizona 85724

Abstract. The confocal microendoscope is an instrument for imaging the surface of the human ovary. Images taken with this instrument from normal and diseased tissue show significant differences in cellular distribution. A real-time computer-aided system to facilitate the identification of ovarian cancer is introduced. The cellular-level structure present in *ex vivo* confocal microendoscope images is modeled as texture. Features are extracted based on first-order statistics, spatial gray-level-dependence matrices, and spatial-frequency content. Selection of the features is performed using stepwise discriminant analysis, forward sequential search, a nonparametric method, principal component analysis, and a heuristic technique that combines the results of these other methods. The selected features are used for classification, and the performance of various machine classifiers is compared by analyzing areas under their receiver operating characteristic curves. The machine classifiers studied included linear discriminant analysis, quadratic discriminant analysis, and the *k*-nearest-neighbor algorithm. The results suggest it is possible to automatically identify pathology based on texture features extracted from confocal microendoscope images and that the machine performance is superior to that of a human observer. [DOI: 10.1117/1.2907167]

Keywords: pattern recognition; confocal optics; endoscopy; image analysis; image evaluation.

Paper 07245R received Jul. 6, 2007; revised manuscript received Nov. 2, 2007; accepted for publication Nov. 6, 2007; published online May 2, 2008.

1 Introduction

Ovarian cancer is the fifth most common cancer in women. According to statistics from the American Cancer Society, there will be about 22,430 new cases of ovarian cancer in the United States in 2007, and about 15,280 women will die of the disease.[1] If diagnosed early, while still localized, the 5-y survival rate is 93%. However, only 19% of all ovarian cancers are found at this early stage. Clearly, early detection improves the chances that ovarian cancer can be treated successfully. Unfortunately, an effective and routine screening test for women at risk is not available. Noninvasive imaging methods such as computed tomography (CT), magnetic resonance imaging (MRI), and ultrasound can confirm the presence of a pelvic mass, but do not have the resolution to visualize morphological and cellular-level changes typical of early stage cancer and precancer. The only way to diagnose early stage cancer is to remove a tissue sample from a "suspicious" area and examine it under a microscope.[2] Such biopsy procedures are surgically invasive and require significant turnaround time. Moreover, due to the inherent sampling nature of the procedure, diseased areas are often overlooked.

Bench-top confocal microscopes[3] are routinely used to create high-quality optical images of biological samples. A key feature of confocal microscopy is the ability to reject light from out-of-focus planes and provide a clear in-focus image from a thin section of the sample—up to a few hundred micrometers below the surface. Recently, there have been efforts to adapt confocal imaging systems for *in vivo* use to perform optical biopsy.[4–14] Such instruments can be used alone, or inserted through a trocar, catheter, large-bore needle, or the

Address all correspondence to Arthur Gmitro, Department of Radiology, University of Arizona, 1609 N Warren Ave, Tucson, AZ 85724; Tel: 520–626–4720; Fax: 520–626–3893; E-mail: gmitro@radiology.arizona.edu

1083-3668/2008/13(2)/024021/13/$25.00 © 2008 SPIE

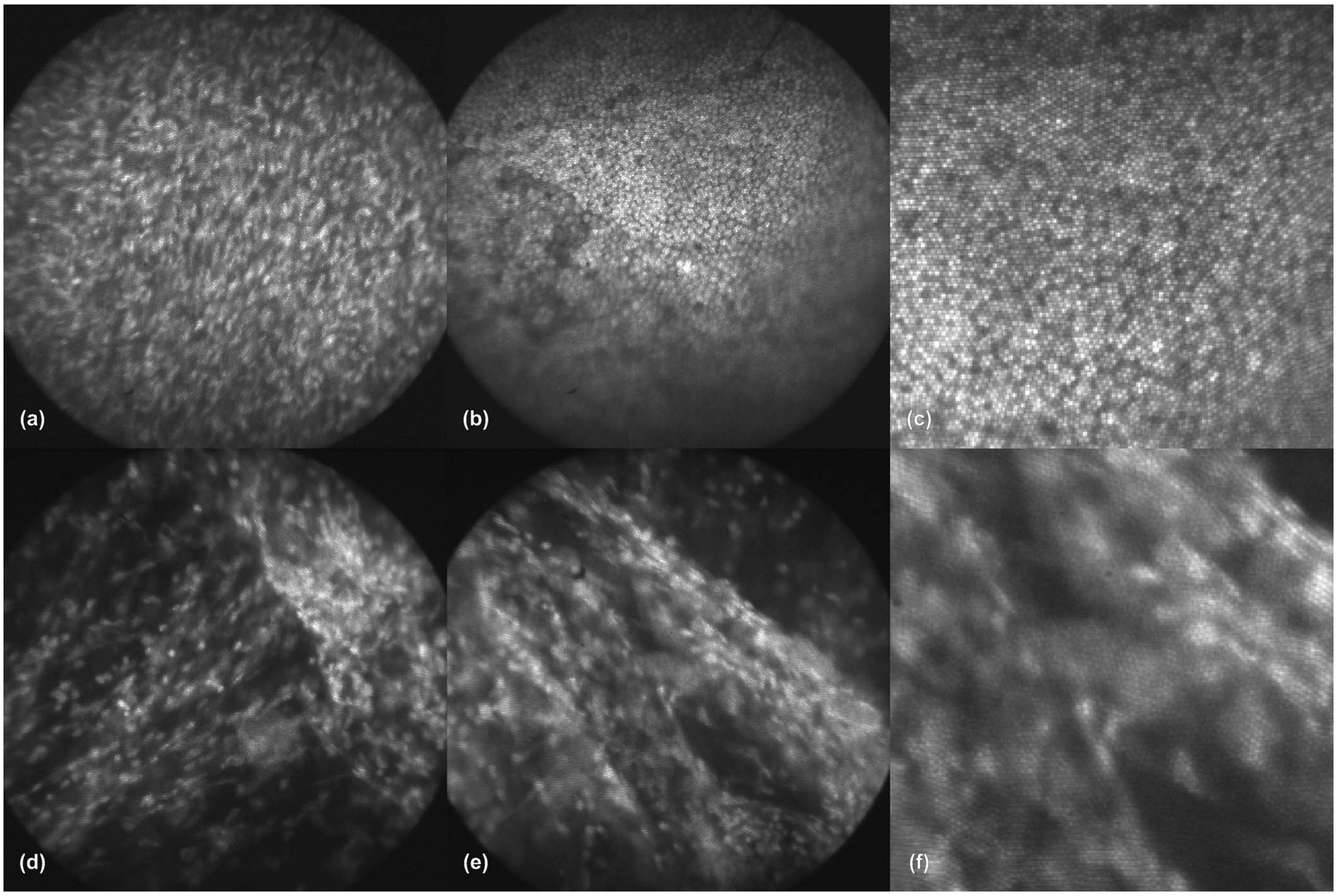

Fig. 1 Representative confocal microendoscope images of the ovary: (a) normal stroma, (b) normal epithelium, (c) close-up of image (b) showing texture of normal epithelium, (d) ovarian cancer, (e) ovarian cancer, and (f) close-up of image (e) showing the texture of cancerous tissue.

instrument channel of a conventional endoscope. We developed and previously reported on such an instrument, which is based on a fluorescence slit-scan confocal microscope coupled to a fiber optic imaging bundle.[11–13,15] Although it is possible to image tissue autofluorescence, higher contrast images are typically obtained using an exogenous fluorescent dye applied to the tissue. This confocal microendoscope is capable of providing physicians with real-time cellular-level images of epithelial tissues, and since about 85% of ovarian cancers are epithelial in nature[1] such a system could potentially be used for ovarian cancer detection.

In a clinical setting, images from the confocal microendoscope are acquired in real time as the probe is placed in contact with and scanned across the tissue surface. A major advantage of the confocal microendoscope is that many areas on the surface of the ovary can be imaged in this way. In such a scenario, it may be difficult for physicians to accurately identify subtle cellular and morphological changes characteristic of pathologies. A real-time computer-aided diagnosis system could potentially be used to provide feedback that would appropriately guide the physician to diseased areas. Furthermore, the system could aid the physician's diagnosis and help determine the best course of action. Similar automated systems have been effective in reducing diagnostic error, cost, and patient suffering associated with unnecessary biopsies.[16]

Confocal microendoscope images of ovary display textural characteristics. Figure 1 shows images of the epithelial surface of *ex vivo* normal and cancerous ovarian tissue acquired using the confocal microendoscope. The tissue was stained topically with the fluorescent dye acridine orange (AO) prior to imaging. Examination of these and similar images reveals that tissue pathologies result in significant differences in cellular distribution patterns, which is an important criteria used by pathologists in making a diagnosis. In general, images from ovarian carcinomas show significantly more heterogeneity than images from healthy ovarian tissue. Figure 2 shows conventional histology images of normal and cancerous ovarian tissue with H&E (hemotoxolin and esosin) staining. Note

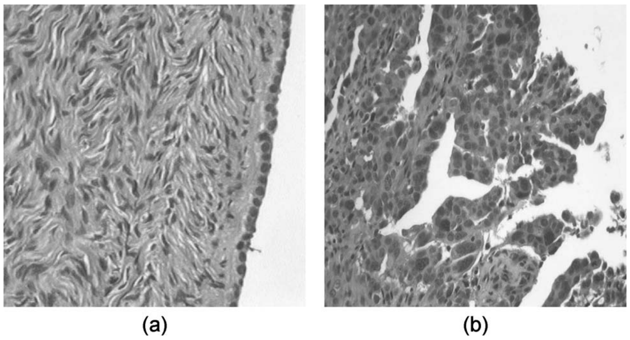

Fig. 2 Conventional histology images from (a) normal and (b) cancerous ovarian tissue.

how the surface epithelium is a relatively uniform single cell layer in Fig. 2(a), whereas in Fig. 2(b) the nuclei at the surface are far more heterogeneous in size and distribution than the nuclei in normal tissue. Note that standard histopathology involves slicing the tissue perpendicular to its surface to produce a cross-sectional view, whereas the confocal microendoscope images are *en face* to the surface. Topical administration of fluorescent dye primarily stains the epithelial surface and although the confocal microendoscope can image at depths up to about 200 μm, it is most often adjusted for best focus at the stained surface layer of cells, typically a depth of a few tens of micrometers.

The analysis of these images for recognition of pathologies can be considered analogous to the problem of texture classification. At the core of this problem is the need to mathematically model the cellular-level structure present in these images as texture. A large number of schemes have been proposed for texture analysis[17] and this remains an active area of research. Esgiar et al.[18] reported on colon cancer detection in microscopic images of colonic mucosa using texture analysis. Other studies have also reported on recognition of pathologies using texture analysis in microscopic images of cervical cancer,[19] prostate cancer,[20] and bladder cancer.[21] In most cases, however, images were obtained from extracted biopsy tissue that had been histologically stained prior to evaluation. Several researchers have published results on ovarian cancer detection by texture analysis.[22–24] However, to the best of our knowledge all of these efforts have been based on nuclear texture analysis rather than texture analysis of a large region of tissue.

The confocal microendoscope is a novel imaging system with the potential to improve ovarian cancer detection. In this paper, we present a technique for recognition of ovarian cancer in confocal microendoscope images. To achieve this goal, we model the cellular-level structure present in these images as texture and extract features based on first-order statistics, spatial gray-level dependence matrices,[25] and spatial-frequency content. For the latter, we present an alternative computational approach to extract texture features from the Fourier power spectrum based on the forward Radon transform[26] and the Fourier projection-slice theorem.[27] The best features for classification and performance of various classifiers are evaluated using a database of images acquired with the confocal microendoscope instrument. Finally, the performance of the best machine classifier is compared to that of expert human observers.

2 Image Acquisition and Preprocessing

2.1 *Tissue Samples*

Ovarian tissue samples from 38 patients (20 normal and 18 cancerous) were obtained from excised human ovaries under protocols approved by the Institutional Review Board of the University of Arizona. Whole ovaries were obtained from patients undergoing oophorectomy. For this study, the surgeon selected several tissue specimens from the ovary and provided a preliminary diagnosis based on macroscopic visualization. Tissue specimens were approximately 4 mm long, 4 mm wide, and 2 mm thick.

2.2 *Imaging*

Tissue specimens were stained with 100 μl of 300 μM AO prior to imaging. AO is a nucleic acid fluorescent dye that is efficiently excited by an argon ion laser at 488 nm and has dual emission spectra at 525 and 650 nm when bound to DNA and RNA, respectively. Images of excised tissue were collected using the confocal microendoscope within 1 h of resection. The catheter of the confocal microendoscope was positioned in contact with the biopsy specimen, and focused on the surface layer of cells. The field of view of the system in tissue is 430 μm with an axial resolution of 25 μm and a lateral resolution of 2 μm. The confocal microendoscope employs a scientific-grade CCD camera from Princeton Instruments to collect gray-scale images of tissue. The camera records 512×512 pixel images with a 12-bit digitizer operating at approximately 4 frames/s. Each image was labeled as "normal" or "cancerous" on the basis of histology results from the corresponding tissue sample. Histology diagnosis was made from the same tissue specimens, with standard processing and H&E staining, immediately following confocal imaging. Standard histology involves slicing tissue perpendicular to the surface, whereas the confocal images are obtained *en face*. Therefore, exact registration between confocal imaging and histology is not possible, but the locations for imaging were estimated to be within 1 mm of each other. In some cases, tissue handling or contact with the confocal imaging probe can denude tissue of the epithelial layer. However, pathologists have stated that using the confocal microendoscope prior to histology preparation does not affect the accuracy of their diagnosis.

2.3 *Image Database and Preprocessing*

A database of exemplar images was developed with 42 images of histology-verified cancerous tissue and 86 images of normal tissue. This set of 128 images was used to study texture classification schemes for ovarian cancer detection.

Imaging systems that utilize fiber optic catheters have pixelation artifacts due to the limited number of fibers in the fiber bundle. To reduce these artifacts, each 512×512-pixel image was smoothed with a 3×3 averaging filter and downsized it by a factor of 2 to 256×256 pixels. All images were normalized using the min max range and requantized from 12 to 8 bits/pixel (i.e., every image covers the full range from 0 to 255). The central 192×192-pixel region was extracted from each image for subsequent evaluation.

3 Texture Features

The first step in the pattern recognition system design cycle is feature extraction. As stated, features were calculated based on first-order statistics, spatial gray-level dependence matrices, and spatial-frequency content.

3.1 *First-Order Statistics*

First-order statistics are measures computed from the normalized histogram of the image. The following five features were calculated:[28] mean, variance, standard deviation, coefficient of variation, skewness, and kurtosis (numbered as features 0 through 5, respectively).

3.2 *Spatial Gray-Level Dependence Matrices*

Spatial gray-level dependence matrices[25] (SGLDMs) are commonly used to extract statistical texture features from images. The normalized SGLDM is an estimation of the second-order joint conditional probability density function $p(i,j:d,\theta)$. The function $p(i,j:d,\theta)$ measures the probability that 2 pixels, which are located with an intersample distance d and direction θ, have gray levels i and j. The matrices are determined empirically by indexing through the image pixels, locating the corresponding pixel at distance d and angle θ away, adding a count to the SGLDM in the appropriate i, j and j, i elements, and then dividing all SGLDM elements by 2 times the number of visited image pixels. The normalization factor is slightly less than the total number of pixels because edge pixels do not have a corresponding pixel at a distance d in the image and are not included in the calculation.

In calculating the SGLDM for texture representation, there are three fundamental parameters that must be defined: the number of bits of resolution B (such that $2^B=L$, where L is the number of quantized gray levels), the displacement value d, and the direction θ. The number of bits of resolution used is an important factor, as it influences computation time and how accurately the texture is represented. It also affects how sparse the SGLDM will be. We set $B=8$ (i.e., $L=256$ gray levels) in all our experiments,[29] but evaluate the final classifier performance as the number of bits is reduced.

In many applications, it is not obvious which value(s) of d will accurately capture the texture. Zucker and Terzopoulos[30] proposed an algorithm for selecting the values of d that best capture the texture. A χ^2 statistic, which was later normalized,[31] is used to compare values at different displacements to determine which are the most significant. We applied the χ^2 test for intersample distances from 1 to 14 pixels on each of the preprocessed confocal microendoscope images. As texture in the confocal microendoscope images is rotation invariant, the value of χ^2 was averaged over four orientations (0, 45, 90, 135 deg) for each image. At 45 and 135 deg, a 1-pixel distance is the adjacent pixel on the diagonal even though the physical distance is greater by a factor $\sqrt{2}$. The analysis indicated that values of $d\leq 6$ capture the most significant aspects of the texture. Based on this result the SGLDM were calculated at six distances, $d=\{1,2,3,4,5,6\}$, and four orientations, $\theta=\{0,45,90,135\text{ deg}\}$. At each distance and orientation we calculated the 14 features[25,32] listed in Table 1.

Since the texture in the confocal microendoscope images is rotation invariant, the features were averaged over the four orientations to produce a total of 14 (features)$\times$6 (distances)$=$84 SGLDM-based features. These features are numbered 6 through 89 with the $d=1$ versions of Table 1 covering features 6 through 19 and so on (i.e., feature 20 is ASM2 and feature 34 is ASM3). Throughout this paper, we refer to the combined set of first-order statistics plus SGLDM-based features as "statistical features."

Table 1 Calculated features.

Feature	*Abbreviation*
Angular second moment	ASM
Contrast	CON
Correlation	COR
Sum of squares	SOS
Inverse difference moment	IDM
Sum average	SA
Sum variance	SV
Sum entropy	SE
Entropy	ENT
Difference variance	DV
Difference entropy	DEN
Information measure of correlation 1	IMC1
Information measure of correlation 2	IMC2
Maximum probability	MP

3.3 *Spatial-Frequency-Based Features*

The quasiperiodic patterns present in confocal microendoscope images suggest the use of texture features based on the Fourier power spectrum of the image. Extracting texture features in the spatial-frequency domain entails the calculation of the square magnitude of the discrete Fourier transform of the image and the definition of features as summations over regions of the spatial-frequency plane.[33] Since rotation invariance is desired in this application, summing energy values between certain frequency ranges in an annular fashion is appropriate. However, the summation of energy values over annular regions is not straightforward as the discrete Fourier transform returns a rectangular array. In this study, we applied the Radon transform[26] in conjunction with the Fourier projection-slice theorem[27] to compute texture features in an annular ring sampling geometry.[34] Specifically, we computed the Radon transform of the preprocessed image, and then performed a 1-D discrete Fourier transform (DFT) operation on the projection at each orientation. The result is a matrix whose columns are radial lines in Fourier space. We call this matrix the projection-slice Fourier transform. This approach simplifies the task of summing energy values in the annular ring sampling geometry. Instead of summing values of Fourier energies on a rectangular grid, we can now accomplish this task by simply summing along rows of the square magnitude of the projection-slice Fourier transform.

The Radon transform of the preprocessed 192$\times$192-pixel confocal microendoscope images was computed with a radial sampling of 1 pixel and an angular sampling of 1 deg. The maximum extent of the projection at 45 deg is 273 pixels, so the 1-D DFT operation produces a projection-slice Fourier transform matrix with dimensions of 273$\times$180 pixels. To extract texture features, nonoverlapping frequency bands of four rows starting from the first row away from the center dc row were added together. The calculated mean and standard deviation of each band represents a feature. Because of Hermetian symmetry in the projection-slice Fourier transform, only half of the $272/4=68$ frequency bands are unique. Thus, the means and standard deviations of 34 frequency bands (plus the mean dc component) generated a set of $(34\times 2)+1=69$ features. These features are numbered 90 through 158 with feature 90 corresponding to the mean dc component and each subsequent pair of features corresponding to the mean and standard deviation, respectively, as the annular rings in frequency space increase from the dc to higher frequencies.

4 Feature Selection

Feature selection is arguably the most crucial step in pattern recognition system design. To design an efficient ovarian tissue classification system, one must select features that capture the salient differences between the texture classes (normal and cancerous). It is well known that a positive bias is introduced in classification accuracy estimates if feature selection is performed on the entire data set and the same data set is used to evaluate classifier performance.[35] To obtain the most reliable estimate of classifier performance, a separate data set should be used to determine the "optimal" feature subset. However, feature selection is unreliable when based on a small amount of data.[36] Due to the limited amount of data available in this study, we utilized the full data set (128 images) for feature selection. The lack of a large number of samples is a common problem in biomedical applications and many researchers have used a similar approach.[37–39] To estimate the true performance of the automated classification system, a larger set of samples will be necessary, and we are currently performing such an evaluation.

4.1 *Techniques for Feature Selection*

Two sets of features, statistical features and spatial-frequency-based features, were already described. To determine the optimal subset of features, a third set was concocted by merging the two sets into a combined set of 159 features (90 statistical features and 69 spatial-frequency features). As the features have different physical units, and thus substantial differences in variances, it is necessary to normalize them prior to feature selection. Therefore, each feature was scaled to have zero mean and unit variance. The normalized features were used in all subsequent analysis.

In this study, the following five approaches were investigated for the selection of the best set of features: stepwise discriminant analysis,[40] forward sequential search,[41] a nonparametric method,[42] principal component analysis,[43] plus a collection of the most popular features from these other four schemes. These techniques were applied to each of the three sets of features (statistical, spatial-frequency, combined). We constrained our experiment to five features. It is generally accepted that using at least 10 times as many training samples per class as the number of features is good practice to follow in classifier design.[44] Adding more features can make matters worse due to the curse of dimensionality. To make a fair comparison between all the feature selection schemes, we forced each algorithm to select exactly five features.

4.1.1 *Stepwise discriminant analysis*

Stepwise discriminant analysis (SDA) was implemented using[45] SPSS. The procedure begins by selecting the individual feature that provides the greatest univariate discrimination. Subsequently, at each step of the procedure, one feature is either added to or removed from this set of features based on the effect of the new feature on the selection criterion. Wilks's lambda, which measures the ratio of the variance in each group to the total variance, was used as the selection criterion.[46] SDA utilizes two threshold values: F_{in} for feature entry and F_{out} for feature removal. The values for F_{in} and F_{out} were chosen in such a way that five features were selected for each experiment.

4.1.2 *Forward sequential search*

Forward sequential search (FSS) is one of the most common search techniques[47] and is often applied to feature selection. This simple procedure adds features one at a time by selecting the next feature that maximizes the criterion function. The procedure terminates when the desired number of features (i.e., 5) is achieved. In this study, we used the parametric Mahalanobis distance criterion for measuring feature set discrimination. The metric is attractive because under Gaussian class-conditional densities, the probability of error is inversely proportional to the Mahalanobis distance.[48]

4.1.3 *Nonparametric method* (*NPM*)

The performance of a non-parametric classifier, k nearest neighbor (k-NN), was used as a criterion for feature subset selection. Classification accuracy of the k-NN classifier was estimated using the leave-one-out error estimation technique.[49] As before, the desired number of features was set to 5. To determine the optimal value for the parameter k of the k-NN algorithm, experiments were conducted using various values ($k=1$, 3, 5, 7, 9) for each feature set. The peak classification accuracy was achieved when $k=7$ for all three feature sets (see Srivastava[50] for additional information).

4.1.4 *Principal component analysis*

Principal component analysis (PCA) is a method to derive a new set of features that are uncorrelated linear combinations of the original variables.[49] In this study, the data were projected into the subspace of the five most significant principal components.

4.1.5 *Popular features*

In an effort to combine the suboptimal feature subsets provided by the already mentioned feature selection schemes (excluding PCA), and to acquire the most stable and consistent features, we defined a new set of features, which consisted of the most commonly selected features of the other approaches. We tabulated the frequency of occurrence of a feature in the feature subsets acquired via SDA, FSS, and NPM, and incorporated the most frequently occurring features in the new "POP" set. If ever there was a tie, it was broken on the basis of a feature's individual discriminatory ability using the Mahalanobis distance criterion.

4.2 *Evaluation of Performance*

Assessment of classification performance in diagnostic systems is often accomplished using receiver operating characteristic (ROC) analysis.[43] The performance of each of the five feature subsets from the preceding was estimated via ROC analysis[51] and bootstrap error estimation[52] using a linear discriminant classifier. The procedure involved random drawing with replacement from the available samples in the data set to create a training set. The remaining samples form an independent test set, which was used with an optimum linear discriminant function designed from the training data. We used the output discriminant score from the linear discriminant as the decision variable in the ROC analysis. Performance of a feature subset was measured by computing the area A_z under the ROC curve using[53] LABROC4. To reduce bias due to case selection, training and testing were repeated many times, each with different training and test data sets. The process was repeated 100 times and the 100 A_z values were averaged to provide a measurement of the effectiveness of that feature subset. If at any time a degenerate data set was encountered and the corresponding A_z could not be accurately computed, the data set was excluded from the analysis and replaced by another iteration of bootstrap sampling, which ensured that 100 A_z values were computed for each five-feature subset.

5 Machine Classifiers

5.1 *Linear and Quadratic Discriminant Functions*

A mathematical representation of the Bayesian classifier is a set of discriminant functions, $g_i(\boldsymbol{x})$, $i=1,\ldots,c$ where c is the number of classes. The discriminant functions classify a pattern $\boldsymbol{x}$ by assigning $\boldsymbol{x}$ to class ω_i if $g_i(\boldsymbol{x})\geq g_j(\boldsymbol{x})$ for all $j\neq i$. In the two-class case, it is equivalent to form a single decision function:

$$g(\boldsymbol{x})=g_1(\boldsymbol{x})-g_2,(\boldsymbol{x}), \qquad (1)$$

and to assign $\boldsymbol{x}$ to class ω_1 if $g(\boldsymbol{x})\geq 0$ (cancerous), and to class ω_2 if $g(\boldsymbol{x})\leq 0$ (normal). Under the assumption that the distribution of the feature vectors $\boldsymbol{x}$ within the i'th class is multivariate Gaussian with mean μ_i and covariance matrix $\boldsymbol{\Sigma}_i$, and that the covariance matrices for both classes are identical, i.e., $\boldsymbol{\Sigma}_1=\boldsymbol{\Sigma}_2=\boldsymbol{\Sigma}$, the resulting decision function attains a linear form, and the classifier results from linear discriminant analysis[54] (LDA). Alternatively, if the covariance matrices are different for each class, the classification rule adopts a quadratic form, and the classifier results from quadratic discriminant analysis (QDA). To evaluate the performance of LDA or QDA classifiers, the output discriminant score $g(\boldsymbol{x})$ was regarded as a decision variable, and the discriminant scores for the test samples were subjected to ROC analysis, as already described.

5.2 *k-NN*

The k-NN algorithm is a simple nonparametric classifier that classifies patterns by assigning them to the class that is most heavily represented in the "votes" of the k nearest samples.[49] To measure classification performance via ROC analysis, it was necessary to assign a confidence measure to each sample in addition to the traditional binary decision (normal/cancerous). To generate such a measure, we used the weighted voting algorithm developed by Arya and Mount[55] that calculates the k-NNs for each sample and assigns a confidence using the following rule:

$$p=\frac{\sum_{i=1}^{k}\exp\left(-d_i^2/\mu^2\right)c_i}{\sum_{i=1}^{k}\exp\left(-d_i^2/\mu^2\right)}, \qquad (2)$$

where $d_1,\ldots,d_k$ are the distances to the k nearest samples from the test sample, μ is the mean of $d_1,\ldots,d_k$, and $c_1,\ldots,c_k$ are the class labels of the k-NNs. For the two-class case, we chose $c=0$ for normal samples and $c=1$ for cancerous samples. The resulting confidence measure is a number ranging from zero (normal) to one (cancerous).

6 Experiments and Results

6.1 *Performance Analysis of Feature Subsets*

Table 2 shows the features selected from the three feature sets (statistical, spatial-frequency, combined) using each of the feature selection schemes already described, as well as the corresponding average A_z values and standard deviations from 100 bootstrap runs. We begin by considering the spatial-frequency-based features. To test whether there are statistically significant differences between the performances of the feature subsets (subsets 6–10 in Table 2), we applied a balanced one-way analysis of variance[56] (ANOVA) in conjunction with a multiple comparison test.[57] Although there was no clear winner; some feature combinations were statistically better than others. For example, subsets 6 and 7 were statistically better than subsets 8 and 9 ($p<0.05$). Feature subset 6 had the highest mean A_z value of the spatial frequency based features.

Features selected from the statistical feature set using SDA (subset 1), FSS (subset 2), and POP (subset 5) were found to be statistically better ($p<0.001$) than the best spatial-frequency features (subset 6). These results indicate that features selected from the statistical feature set are more powerful in discriminating normal from cancerous images. Statistical features selected using the NPM and PCA were found to be statistically poorer than subsets 1, 2, and 5 ($p<0.01$). In looking at the features selected from the combined feature set, it is observed that SDA (subset 11) actually identified the same features as subset 1. This indicates that the combined feature set does not yield any significant improvement in classification accuracy when compared to features selected from statistical features alone. A multiple comparison test revealed that there were no significant differences between subsets 1 (same as 11), 12, and 15. Subsets 13 and 14 were found to be poorer when compared to other feature subsets selected from the combined feature set ($p<0.001$).

We observed a few interesting patterns in looking at the best feature subsets—(1, 2, 5, 12, and 15). There appears to be two basic templates of four features that are commonly selected (16,86,87,73) and (16,87,73,77). These two templates are consistently selected by two different feature selection techniques, namely, FSS and SDA. To investigate the optimum number of features needed for classification, we compared the performance of features selected from the combined feature set using FSS and SDA as we varied the number of features selected from 1 to 7. We again accumulated A_z values from 100 runs of bootstrapped test data. Figure 3 shows the group means and 95% confidence interval for the two feature

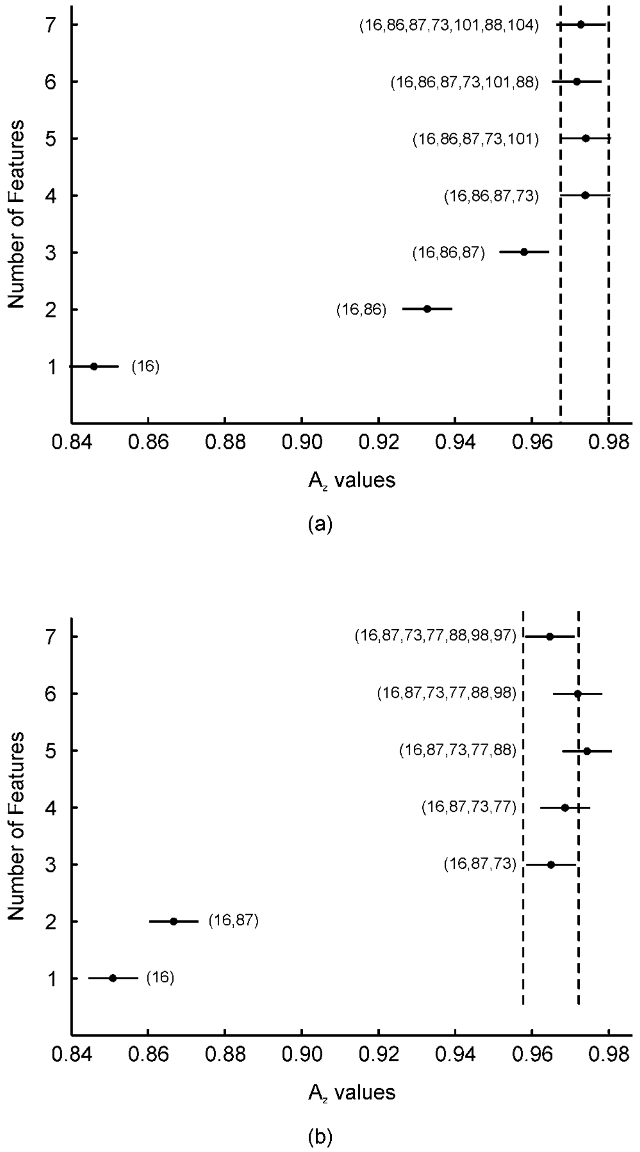

Fig. 3 Performance of increasing number of features selected from the combined feature set using (a) FSS and (b) SDA.

selection methods as the number of features increases from 1 to 7.

According to these results, there are no statistically significant increases in classification accuracy beyond four features using FSS and three features using SDA, as indicated by the overlap in confidence intervals. Also evident from Fig. 3 is the fact that when coupled with feature 16, feature 86 is substantially more effective than feature 87. To make a fair comparison between the two feature selection techniques, we compared the two sets of four features $\{(16,86,87,73), (16,87,73,77)\}$ to test for any statistically significant differences between them. The mean A_z values for the set of four features selected using FSS and SDA were 0.9738 and 0.9684, respectively, and were not significantly different (p=0.0687). The subset $(16,86,87,73)$ selected using FSS was the "best" set of features, as it provided the highest mean A_z value.

Table 2 Comparison of performance of LDA, QDA, and 5-NN algorithm.

Classifier	A_z	*Std. Dev.*	*Sensitivity*	*Specificity*
LDA	0.9738	0.0177	0.98	0.90
QDA	0.9656	0.0227	0.97	0.90
5-NN	0.8742	0.0600	0.67	0.90

6.2 *Performance Analysis of Different Classifiers*

We used the $(16,86,87,73)$ subset of features to compare the performance of the following machine classifiers: LDA, QDA, and k-NN. Specifically, for each classifier we collected the output confidence scores from 100 runs of bootstrapped test data and subjected these scores to ROC analysis using LABROC4. The A_z values generated by each experiment, in addition to the two curve-fitting parameters calculated by LABROC4, were tabulated. To obtain the optimal value for k for the k-NN algorithm, we repeated the experiments for a range of values from k=2 to 9. Although no particular value yielded a statistically significant increase in classification accuracy, we selected k=5, as it provided the highest mean A_z value. Table 2 shows the mean A_z value and standard deviation obtained for each classifier, along with their respective sensitivities at a specificity of 0.90.

The linear and quadratic classifiers were found to be statistically superior to the 5-NN algorithm, whereas no statistical difference was found to exist between the LDA and QDA. As frequently asserted,[58] a simpler classifier design can be expected to perform better on a validation set, we chose the LDA classifier as our preferred approach for this application. The current results pertaining to the LDA classifier indicate that about 2% of patients exhibiting abnormal cell characteristics were incorrectly classified as normal when 10% of patients exhibiting normal characteristics were incorrectly identified as abnormal, the operating point on the ROC curve being (TPF,FPF)=(0.98,0.10).

6.3 *Human Performance Analysis*

To compare performance of the computer-aided system with human classification accuracy, we conducted an observer study using the ROC study protocol. To characterize human observer consistency, we randomly repeated 30% of the data set and randomly rotated these repeated data by 90, 180, or 270 deg. The resulting set of 166 images included 128 original images and 38 repeated images with random orientation.

The images were reviewed by four human observers. Two of the observers (observers 1 and 2) were researchers with experience looking at confocal microendoscope images. The other two (observers 3 and 4) were MDs with experience looking at histopathologic images of ovarian cancer patients. The observers were asked to rate each image using a numerical scale from one to six: (1) definitely cancerous, (2) probably cancerous, (3) possibly cancerous, (4) possibly normal, (5) probably normal, and (6) definitely normal. The images were displayed in random order on a CRT monitor. All observers except observer 1 were blinded to the distribution of

Table 3 Results of human performance study.

Characteristic	*Observer 1*	*Observer 2*	*Observer 3*	*Observer 4*
Accuracy	83.73%	80.72%	75.30%	70.48%
Sensitivity	55.17%	81.03%	68.97%	77.59%
Specificity	99.07%	80.56%	78.70%	66.67%
Positive predictivity	96.97%	69.12%	63.49%	55.56%
Negative predictivity	80.45%	88.78%	82.52%	84.71%
Consistency	60.53%	55.26%	68.42%	55.26%
Classification Consistency	89.47%	89.47%	78.95%	81.58%
A_z value	0.8474	0.8856	0.8425	0.8296

cancer versus normal samples in the data set. The observers were not allowed to change their assigned diagnosis during the test. Prior to reviewing, the observers were "trained" with a set of eight images that were representative of the visual distribution present in each class.

Table 3 shows the results from the study. The total accuracy was computed by calculating the ratio of the number of correct responses to the total number of images. For the table, "consistency" and "classification consistency" differ in the sense that, one is a measure of the consistency with which the observer provided the exact same label $\{1,2,3,4,5,6\}$ when presented with the same image at a different orientation, and the other measures the consistency with which the observer provided the same classification {normal/cancer} when presented with the same image at a different orientation. Ratings of ≤ 3 and ≥ 4 were considered to be classified as cancerous and normal, respectively.

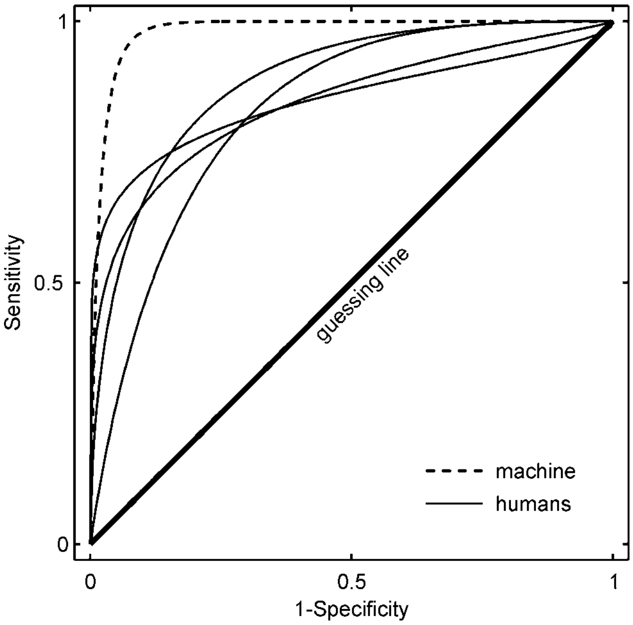

Fig. 4 Comparison of machine performance with human observers.

We also computed the A_z value for each observer. To make a fair comparison with machine performance, we removed the observer ratings from images that had been rotated. The discrete confidence ratings (for 128 images) from each observer were then individually subjected to ROC analysis using LABROC4. The corresponding ROC curves for the human observers are compared to the machine performance (LDA) in Fig. 4. To generate the ROC plot for machine performance, we averaged the ROC curve-fitting parameters previously collected. Clearly, as demonstrated by Tables 2 and 3 and Fig. 4, the performance of the automated linear discriminant classifier is superior to the performance of human observers in this study.

6.4 *Effects of Gray-Level Quantization*

We also investigated how gray-level quantization affects the classification performance. To accomplish this, we reverted back to the original database of images and recomputed the 90 statistical features at various quantization levels from 3 to 8 bits/pixel. Subsequently, we selected a set of four features from each set of 90 features. The FSS strategy was used for selection because of its effectiveness in prior experiments. The performance of each feature subset was estimated using ROC methodology and bootstrap error estimation. Table 4

Table 4 Mean A_z value and standard deviation for each feature subset versus the number of quantization bits.

Bits	*Features*	A_z	*Std.Dev.*
3	16, 80, 38, 17	0.9168	0.0375
4	16, 77, 30, 58	0.9098	0.0413
5	19, 76, 34, 1	0.8558	0.0569
6	6, 80, 22, 66	0.9175	0.0341
7	14, 84, 7, 29	0.9505	0.0272
8	16, 86, 87, 73	0.9738	0.0177

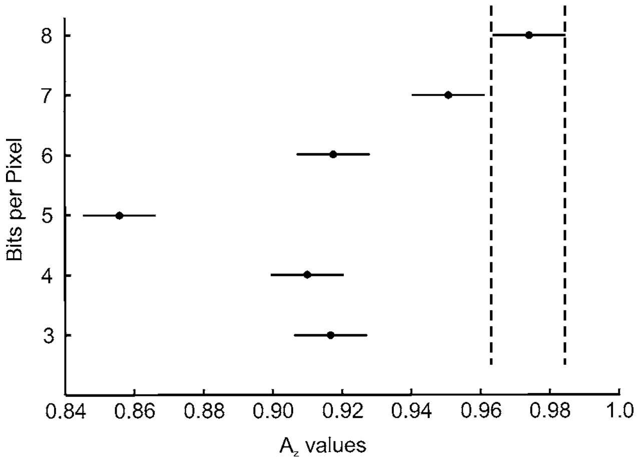

Fig. 5 Performance of a set of four features selected using FSS for different number of quantization bits in the SGLDM.

shows the features selected at each quantization level along with the corresponding mean A_z value and standard deviation obtained from 100 bootstrap runs.

The results show that as the number of bits of resolution decreases, the area under the ROC decreases until reaching a trough at 5 bits/pixel, and then increases slightly. We performed an ANOVA in conjunction with a multiple comparison test to evaluate whether one setting was statistically superior to others. Figure 5 shows the 99.9% confidence intervals for the mean of each group of A_z values derived from the statistical feature subsets as a function of the number of bits of resolution. Clearly, 8 bits/pixel is statistically superior to the other values tested.

7 Discussion

7.1 *Selected Features*

In general, the excellent classification performance indicates that texture is an appropriate image characteristic for distinguishing between cancerous and normal tissue. In terms of spatial-frequency-based texture features, the most commonly selected of the 69 features was 117, which measures the mean Fourier energy within an annular ring with an inside radius of $(54/136)\pi$ and outside radius of $(57/136)\pi$. We denote this feature as $F_{\mu}[53,57)$, where the subscript μ indicates that the mean was computed over the region in the parentheses. The spatial frequency π corresponds to a period of 2 pixels in the image, so $F_{\mu}[53,57)$ is a narrow spatial frequency range centered on a spatial frequency with a period of approximately 5 pixels. Compared to other spatial-frequency-based features, $F_{\mu}[53,57)$ has the highest individual discriminatory ability based on the Mahalanobis distance criterion. Other features with high individual discriminatory ability are $F_{\mu}[45,49)$ and $F_{\mu}[9,13)$. Feature $F_{\mu}[129{,}133)$ has negligible discriminatory information by itself but it can be effective when combined with $F_{\mu}[53,57)$. Note that only three features calculated by computing the standard deviation of the Fourier energy within an annular region were selected, namely, $F_{\sigma}[113{,}117)$, $F_{\sigma}[13,17)$, and $F_{\sigma}[117{,}121)$. Spatial-frequency-based features are effective in discriminating normal from cancerous tissue and an automated classification system based on these features performs significantly better than the human observers. However, spatial-frequency based features are not as effective as statistical features.

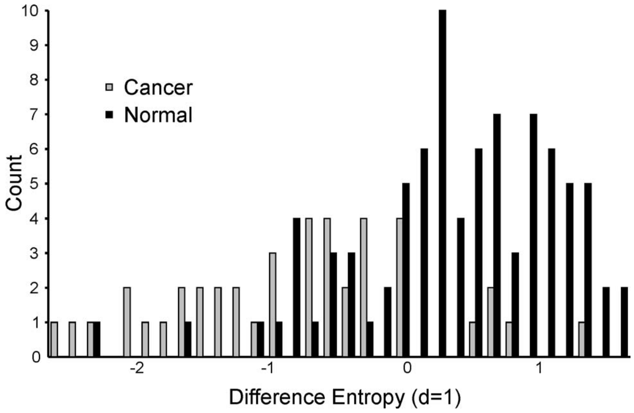

Fig. 6 Class-conditional histogram for feature 16.

In terms of statistical features, none of the first-order features was effective for discrimination. The most commonly selected second-order statistical feature was 16 (difference entropy calculated at intersample distance 1, denoted as DEN_1). It was routinely selected before other features because it provided the highest individual classification accuracy amongst all 159 features. By itself, it accounted for nearly 85% of the total A_z value. Difference entropy is computed from the difference second-order histogram of the image, which represents the probability of occurrence of differences $|i-j|$ in gray-level values for 2 pixels separated by a distance d. As an entropy measure, the feature measures the variation in the distribution of the difference second-order histogram and achieves a maximal value for a uniform distribution of probability. Figure 6 shows a class-conditional histogram of the difference entropy feature. It indicates that on average the difference entropy of cancerous tissue is lower than that of normal tissue, which implies that the gray-level values had larger variation on a local level in normal tissues than in cancerous samples. Intuitively, this makes sense as the regular packing of cells in normal tissue tends to provide a more uniform difference second-order histogram.

According to the results, the "best" set of features is DEN_1, DEN_6, $\mathrm{IMC1}_6$, $\mathrm{IMC1}_5$. The information measure of correlation (IMC1) feature is related to entropy and captures aspects of the correlation in gray values i and j. An interesting observation about this set of features is that they are strongly correlated in pairs. DEN_1 is strongly correlated to DEN_6, and $\mathrm{IMC1}_6$ is strongly correlated to $\mathrm{IMC1}_5$. Yet, when any of the features is removed from the set, we observe a statistically significant decrease in classification performance. This can be attributed to the complementary nature of these features. Guyon and Elisseeff[59] demonstrated that highly correlated features do not necessarily translate to "redundant" information. They suggest that there are two contributions to feature correlation: covariance in the same direction as the "target," which is not more informative, and covariance perpendicular to the "target," which is useful. Methods that prune features based on the linear correlation coefficient without making this distinction are simplistic and ignore the possibility that correlated features may actually be useful.

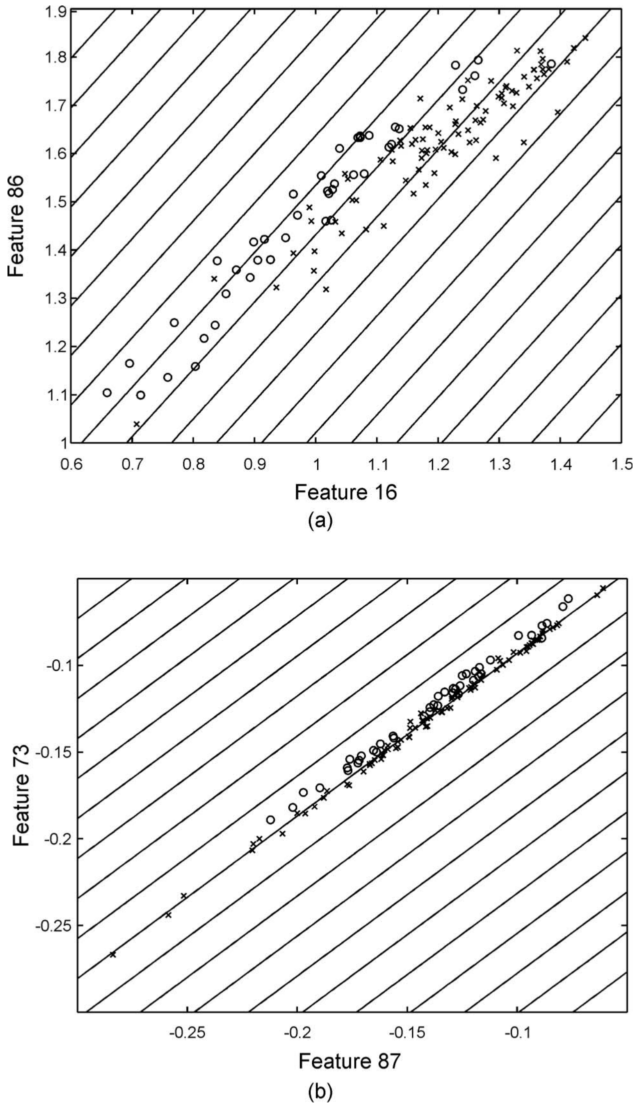

Fig. 7 Scatter plot of (a) features 16 and 86 and (b) features 87 and 73.

Scatter plots of DEN_1 versus DEN_6 and $\mathrm{IMC1}_6$ versus $\mathrm{IMC1}_5$ are shown in Figs. 7(a) and 7(b), respectively. These plots show that even though two features are correlated, they are able to increase classification performance. The profile of these features is unlike that of truly redundant features, where there is often no variation in the direction perpendicular to the class center line (i.e., the covariance is in the same direction as the class center line). In contrast, in our case, the second principal direction of the covariance matrices of these features is perpendicular to the class center line. Therefore, the combined effect of these features leads to greater separation in the higher dimensional feature space.

The results show that no statistically significant improvement in classification accuracy is observed using the combined feature space. In fact, SDA selects the exact same set of features from the combined set as when considering the statistical features alone. Furthermore, in all subsets, statistical features are selected ahead of spatial-frequency-based features, indicating that statistical features are more powerful than spatial-frequency-based features for this application. The results also show that four statistical features are useful for this classification problem, and that classification accuracy is not significantly improved using more than four features.

Features selected using PCA and the NPM performed relatively poorly when compared to features selected using SDA and FSS. One possible explanation for the poor performance of PCA is that PCA does not necessarily capture components that are useful for discrimination. Rather, it finds components that are useful for representing data in a least-squared error sense.[48] In other words, the top five principal components may be useful for accurately capturing the variation in the feature space, but may not be the directions that are necessary to distinguish between the two classes.[58] This inability of PCA to capture the between-class variance could account for the relatively large variance in A_z values (see Table 2). In contrast, FSS and SDA seek features that are efficient for discrimination. The Wilks lambda and the Mahalanobis distance metrics used in FSS and SDA, respectively, naturally capture the property of separating different classes, while at the same time keeping each cluster as tightly packed as possible. For the nonparametric method, we selected features that were optimized for the 7-NN algorithm and subsequently measured performance using a linear discriminant classifier. It is fair to say that the selected features were not designed to be linearly separable as the k-NN algorithm can work equally well for any arbitrary distribution. This argument offers one explanation for the relatively small difference and large variance (see Table 2) in performance for features selected using the nonparametric approach.

An issue of relevance to features derived from the SGLDM is the bit depth of the images. At high bit depth, the SGLDM is sparse. As the bit depth decreases, the sparsity of the SGLDM decreases. The results show that the classification performance decreases as the bit depth decreases from 8 bits, reaching a minimum at a bit depth of 5 bits, and then performance increases with a further decrease in bit depth. Another observation is that the best features selected at each bit depth are different. It is encouraging that the overall performance changes gradually with bit depth and is always better than the performance of the human observers. The significant change of features selected as bit depth changes is an interesting finding and something that should be more carefully investigated to understand the source of this behavior.

7.2 *Classification Performance*

The ideal most-unbiased approach for designing an automated classification system is to have separate large data sets for the three steps of feature selection, classifier design, and performance evaluation. However, as is the case here, it is often difficult or impossible to obtain sufficient data to implement this approach. Leave-one-out has been the recommended[60] technique for evaluating classifier performance on small data sets as it provides almost unbiased estimates of the true error rate. While the leave-one-out estimator is nearly unbiased, its variance is high for small sample sizes. This large variance proves to be problematic and tends to dominate error estimation. In this study we used the bootstrap error estimator, which is reported to be superior to leave-one-out on small data sets.[61] The technique is a low-variance estimator that provides pessimistically biased error estimates.

The test performance of a simple linear classifier is generally better than that of more complex classifiers when the training sample size is small.[36,62] As is often the case in biomedical applications, the data set of images used in this study is relatively limited. That is why we chose to use a linear discriminant function to compare performance of feature subsets. Furthermore, the linear discriminant is an optimal rule for classification if the class-conditional distributions are multivariate normal with a common covariance matrix.

Generally, the performance of the quadratic discriminant is better than that of a linear discriminant as any distribution that can be achieved by a linear discriminant can also be achieved by a quadratic discriminant, provided that the class-conditional statistics are accurately estimated. In this study, we found that the performance of the two techniques was almost identical. Perhaps this indicates the presence of a nearly linearly separable space [see Figs. 7(a) and 7(b)], where a hyperquadric discriminant surface cannot improve classification performance.

The performance of the *k*-NN algorithm is often comparable to the performance of other statistical and neural classifiers.[63] However, in this study, we found evidence to the contrary. We found that the algorithm performed significantly worse than the linear and quadratic classifiers. In addition, the performance did not vary significantly as we varied the *k* parameter. These results are consistent with the work of Weiss,[64] who found that for the *k*-NN algorithm "the estimator [is] overly pessimistic when the true error rate is relatively low." He attributed this characteristic to the repetition of patterns in the training set. Indeed, if a region of the feature space has been neglected due to resampling with replacement, it seems plausible that test patterns from that region will be misclassified.

We found that the success rate with which humans recognized normal/cancerous images varied significantly between 70 to 85% (Table 3). In addition, the performance of the observers varied according to the amount of exposure they had to the database of images. Observers 1 and 2 who had previous experience with confocal microendoscope images performed better than observers 3 and 4 who were more knowledgeable about ovarian pathology but less familiar with these images. This indicates that a learning curve must be overcome for recognizing ovarian pathology in confocal microendoscope images. Overall, there was significant inconsistency in the performance of the human observers. Part of this can be explained by the often confusing texture patterns observed in these images, which can perhaps be discerned effectively only using features such as difference entropy that are able to capture microvariations in these images.

The results of this study suggest that an automated classification system can outperform a human observer in recognizing ovarian cancer in confocal microendoscope images. Due to the small sample size and use of the full data set during feature selection, there is always a concern about overfitting the data and overestimating the performance of the automated classifier. Another study with a larger independent data set will be required to fully validate the findings of this work. Another important issue relates to the performance of the automated classification system when one includes early stage preinvasive lesions, preneoplastic lesions, and the broader spectrum of benign changes occurring in the ovary. A multiclass discriminant approach may be required to distinguish among these various circumstances. It should be remembered, however, that the goal of the automated classifier is to help guide the physician in identifying pathology during an exploratory investigation with a dynamic high frame-rate imaging technology. A high sensitivity to those conditions that warrant more-detailed investigation or intervention is the principal requirement and something that an automated system may be able to achieve without requiring high specificity to the various pathologic and nonpathologic conditions found in ovary.

8 Conclusion

In this study, the efficacy of statistical and spatial-frequency-based features extracted from the confocal microendoscope images for recognition of normal and cancerous ovarian tissue were evaluated. Several feature selection techniques were compared based on performance evaluated using a linear discriminant classifier and ROC analysis. A set of four features selected from the SGLDM-based texture features using forward sequential search provided the highest classification accuracy. The performance of this feature set was also tested using the quadratic discriminant and *k*-NN classifiers. It was found that the linear discriminant classifier using the best set of SGLDM-based features was superior to the other classification methods and significantly outperformed the human observer. Classification performance using spatial-frequency-based features, although not as high as that achieved using SGLDM-based features, also outperformed the human observer. Results of this study indicate that an automated image recognition system may be effective in diagnosing pathologies in a clinical setting and could assist physicians with diagnosis. Although these results were obtained using a relatively small data set, the study demonstrates the potential of computer-aided diagnosis for recognizing ovarian pathologies in confocal microendoscope images.

Acknowledgments

The authors would like to thank William Brands for his help with imaging, preparing tissue, and compiling histology results, Dr. Amy Baruch for participating in the human observer analysis, Dr. Matt Kupinski for his valuable comments and discussion regarding feature selection and ROC methodology, Dr. Elizabeth Krupinski for her help with ROC technique, members of the Kurt Rossman Laboratories at the University of Chicago for providing the LABROC4 software, Dr. Thomas Rauber for providing TOOLDIAG—pattern recognition software. This work was supported by the National Institutes of Health (NIH)—CA94287, the Arizona Disease Control Research Commissions (ADCRC)—7003, and the University Physicians (UPERCC) grants.

References

1. Cancer Facts & Figures 2007, http://www.cancer.org.
2. American Cancer Society, "Detailed guide: ovarian cancer," http://documents.cancer.org/114.00/114.00.pdf.
3. T. Wilson, *Confocal Microscopy*, Academic Press, London (1990).
4. D. L. Dickensheets and G. S. Kino, "Micromachined scanning confocal optical microscope," *Opt. Lett.* **21**(10), 764–766 (1996).
5. R. Juskaitis, T. Wilson, and T. F. Watson, "Real-time white light reflection confocal microscopy using a fibre-optic bundle," *Scanning* **19**(1), 15–19 (1997).

6. K. B. Sung, C. Liang, M. Descour, T. Collier, M. Follen, and R. Richards-Kortum, "Fiber-optic confocal reflectance microscope with miniature objective for in vivo imaging of human tissues," *IEEE Trans. Biomed. Eng.* **49**(10), 1168–1172 (2002).
7. R. H. Webb and F. J. Rogomentich, "Microlaser microscope using self-detection for confocality," *Opt. Lett.* **20**(6), 533–535 (1995).
8. G. J. Tearney, R. H. Webb, and B. E. Bouma, "Spectrally encoded confocal microscopy," *Opt. Lett.* **23**(15), 1152–1154 (1998).
9. W. Mclaren, P. Anikijenko, D. Barkla, T. P. Delaney, and R. King, "Invivo detection of experimental ulcerative colitis in rats using fiberoptic confocal imaging (FOCI)," *Dig. Dis. Sci.* **46**(10), 2263–2276 (2001).
10. J. Knittel, L. Schnieder, G. Buess, B. Messerschmidt, and T. Possner, "Endoscope-compatible confocal microscope using a gradient index-lens system," *Opt. Commun.* **188**(5–6), 267–273 (2001).
11. A. F. Gmitro and D. Aziz, "Confocal microscopy through a fiber-optic imaging bundle," *Opt. Lett.* **18**(8), 565–567 (1993).
12. A. R. Rouse and A. F. Gmitro, "Multispectral imaging with a confocal microendoscope," *Opt. Lett.* **25**(23), 1708–1710 (2000).
13. A. R. Rouse, A. Kano, J. A. Udovich, S. M. Kroto, and A. F. Gmitro, "Design and demonstration of a miniature catheter for a confocal microendoscope," *Appl. Opt.* **43**(31), 5763–5771 (2004).
14. A. A. Tanbakuchi, A. R. Rouse, K. D. Hatch, R. E. Sampliner, J. A. Udovich, and A. F. Gmitro, "Clinical evaluation of a confocal microendoscope system for imaging the ovary," in *Proc. SPIE Int. Soc. for Opt. Eng.*, Vol. **6851**, pp. 685103–10, San Jose, CA (2008).
15. Y. S. Sabharwal, A. R. Rouse, L. Donaldson, M. F. Hopkins, and A. F. Gmitro, "Slit-scanning confocal microendoscope for high-resolution in vivo imaging," *Appl. Opt.* **38**(34), 7133–7144 (1999).
16. Y. Zheng, J. F. Greenleaf, and J. J. Gisvold, "Reduction of breast biopsies with a modified self-organizing map," *IEEE Trans. Neural Netw.* **8**(6), 1386–1396 (1997).
17. M. Tuceryan and A. K. Jain, "Texture analysis," in *The Handbook of Pattern Recognition and Computer Vision*, C. H. Chen, L. F. Pau, and P. S. P. Wang, Eds., pp. 207–248, Springer, New York (1998).
18. A. N. Esgiar, R. N. G. Naguib, B. S. Sharif, M. K. Bennett, and A. Murray, "Microscopic image analysis for quantitative measurement and feature identification of normal and cancerous colonic mucosa," *IEEE Trans. Inf. Technol. Biomed.* **2**, 197–203 (1998).
19. Q. Ji, J. Engel, and E. Craine, "Texture analysis for classification of cervix lesions," *IEEE Trans. Med. Imaging* **19**(11), 1144–1149 (2000).
20. K. Jafari-Khouzani and H. Soltanian-Zadeh, "Multiwavelet grading of pathological images of prostate," *IEEE Trans. Biomed. Eng.* **50**(6), 697–704 (2003).
21. J. T. Arendt, H. S. Levin, E. A. Klein, R. Manoharan, M. S. Feld, and R. M. Cothren, "Investigation of early cancerous changes in bladder tissue by autofluorescence," in *Proc. 19th Annual Int. Conf. IEEE Engineering in Medicine and Biology Society*, Vol. **5**, pp. 2290–2293, (1997).
22. B. Nielsen, F. Albregtsen, and H. E. Danielsen, "Low dimensional adaptive texture feature vectors from class distance and class difference matrices," *IEEE Trans. Med. Imaging* **23**(1), 73–84 (2004).
23. J. P. Geisler, H. E. Geisler, G. A. Miller, M. C. Wiemann, Z. Zhou, and W. Crabtree, "Markov optical texture parameters as prognostic indicators in ovarian carcinoma," *Int. J. Gynecol. Cancer* **9**(4), 317–321 (1999).
24. L. Deligdisch, J. Gil, H. Kerner, H. S. Wu, D. Beck, and R. Gershoni-Baruch, "Ovarian dysplasia in prophylactic oophorectomy specimens: cytogenetic and morphometric correlations," *Cancer* **86**(8), 1544–1550 (1999).
25. R. M. Haralick, K. Shanmugan, and I. H. Dinstein, "Texture features for image classification," *IEEE Trans. Syst. Man Cybern.* **3**, 610–621 (1973).
26. S. Helgason, *The Radon Transform*, Wiley, Boston (1983).
27. L. Ehrenpreis, *The Universality of the Radon Transform*, Clarendon Press, New York (2003).
28. W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes: The Art of Scientific Computing*, Cambridge Univ. Press, Cambrigde, UK (1987).
29. L.-K. Soh and C. Tsatsoulis, "Texture analysis of SAR ice imagery using gray level cooccurrence matrices," *IEEE Trans. Geosci. Remote Sens.* **37**(2), 780–795 (1999).
30. S. W. Zucker and D. Terzopoulos, "Finding structure in co-occurence matrices for texture analysis," *Comput. Graph. Image Process.* **12**, 286–308 (1980).
31. J. Parkkinen and K. Selkainaho, "Detecting texture periodicity from the co-occurrence matrix," *Pattern Recogn.* **11**(8), 43–50 (1990).
32. R. M. Haralick, "Statistical and structural approaches to texture," *Proc. IEEE* **67**, 786–804 (1979).
33. M. F. Augusteijn, L. E. Clemens, and K. A. Shaw, "Performance evaluation of texture measures for ground cover identification in satellite images by means of a neural network classifier," *IEEE Trans. Geosci. Remote Sens.* **33**(3), 616–626 (1995).
34. J. S. Weszka, C. R. Dyer, and A. Rosenfeld, "A comparative study of texture measures for terrain classification," *IEEE Trans. Syst. Man Cybern.* **SMC-6**, 269–285 (1976).
35. B. Sahiner, H.-P. Chan, N. Petrick, R. F. Wagner, and L. Hadjiiski, "Feature selection and classifier performance in computer-aided diagnosis:the effect of finite sample size," *Med. Phys.* **27**(7), 1509–1522 (2000).
36. S. J. Raudys and A. K. Jain, "Small sample size effects in statistical pattern recognition: recommendations for practitioners," *IEEE Trans. Pattern Anal. Mach. Intell.* **13**(3), 252–264 (1991).
37. N. R. Mudigonda, R. M. Rangayyan, and J. E. L. Desautels, "Gradient and texture analysis for the classification of mammographic masses," *IEEE Trans. Med. Imaging* **19**(10), 1032–1043 (2000).
38. C. I. Christodoulou, C. S. Pattichis, M. Pantziaris, and A. Nicolaides, "Texture-based classification of atherosclerotic carotid plaques," *IEEE Trans. Med. Imaging* **22**(7), 902–912 (2003).
39. C.-M. Wu, Y.-C. Chen, and K.-S. Hsieh, "Texture features for classification of ultrasonic liver images," *IEEE Trans. Med. Imaging* **11**, 141–152 (1992).
40. L. Hadjiiski, B. Sahiner, H.-P. Chan, N. Petrick, and M. Helvie, "Classification of malignant and benign masses based on hybrid ART2LDA approach," *IEEE Trans. Med. Imaging* **18**(2), 1178–1187 (1999).
41. M. Gletsos, S. G. Mougiakakou, G. K. Matsopoulos, K. S. Nikita, A. S. Nikita, and D. Kelekis, "A computer-aided diagnostic system to characterize CT focal liver lesions: design and optimization of a neural network classifier," *IEEE Trans. Inf. Technol. Biomed.* **7**(3), 153–162 (2003).
42. H. Ganster, A. Pinz, R. Rohrer, E. Wildling, M. Binder, and H. Kittler, "Automated melanoma recognition," *IEEE Trans. Med. Imaging* **20**(3), 233–239 (2001).
43. K. Chan, T.-W. Lee, P. A. Sample, M. H. Goldbaum, R. N. Weinreb, and T. J. Sejnowski, "Comparison of machine learning and traditional classifiers in glaucoma diagnosis," *IEEE Trans. Biomed. Eng.* **49**(9), 963–974 (2002).
44. A. K. Jain and B. Chandrasekaran, "Dimensionality and sample size considerations in pattern recognition practice," in *Handbook of Statistics*, P. R. Krishnaiah and L. N. Kanal, Eds., pp. 835–855, North-Holland Publishing Co., The Netherlands (1982).
45. SPSS Inc, Chicago, IL.
46. M. J. Norusis, *SPSS Professional Statistics 6.1*, SPSS, Chicago (1993).
47. A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: a review," *IEEE Trans. Pattern Anal. Mach. Intell.* **22**(1), 4–37 (2000).
48. R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*, Wiley, New York (1973).
49. P. A. Devijver and J. Kittler, *Pattern Recognition: A Statistical Approach*, Prentice/Hall International, London (1982).
50. S. Srivastava, *Computer-Aided Identification of Ovarian Cancer in Confocal Microendoscope Images*, University of Arizona, Tucson (2004).
51. J. A. Swets and R. M. Pickett, *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory*, Academic Press, New York (1982).
52. B. Efron and R. Tibshirani, *An Introduction to the Bootstrap*, Chapman & Hall, London (1994).
53. C. E. Metz, B. A. Herman, and J.-H. Shen, "Maximum-likelihood estimation of ROC curves from continuously-distributed data," *Stat. Med.* **17**, 1033 (1998).
54. R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*, Prentice-Hall, Englewood Cliffs, NJ (1992).

55. S. Arya and D. M. Mount, "Approximate nearest neighbor queries in fixed dimensions," in *Proc. 4th ACM-SIAM Symp. Discrete Algorithms*, pp. 271–280 (1993).
56. R. V. Hogg and J. Ledolter, *Engineering Statistics*, MacMillan, New York (1987).
57. Y. Hochberg and A. C. Tamhane, *Multiple Comparison Procedures*, Wiley, New York (1987).
58. R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, Wiley-Interscience, New York (2000).
59. I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.* **3**, 1157–1182 (2003).
60. L. Kanal and B. Chandrasekaran, "On dimensionality and sample size in statistical pattern classification," *Pattern Recogn.* **3**, 225–234 (1971).
61. B. Efron, "Estimating the error rate of a prediction rule," *J. Am. Stat. Assoc.* **78**, 316–333 (1983).
62. H. P. Chan, B. Sahiner, K. L. Lam, N. Petrick, M. A. Helvie, M. M. Goodsitt, and D. D. Adler, "Computerized analysis of mammographic microcalcifications in morphological and texture feature spaces," *Med. Phys.* **25**(10), 2007–2019 (1998).
63. L. Holmstrom, P. Koistinen, and J. Laaksonen, "Neural and statistical classifiers—taxonomy and two case studies," *IEEE Trans. Neural Netw.* **8**(1), 5–17 (1997).
64. S. M. Weiss, "Small sample error rate estimation for k-NN classifiers," *IEEE Trans. Pattern Anal. Mach. Intell.* **13**(3), 285–289 (1991).