False-positive reduction technique for detection of masses on digital mammograms: Global and local multiresolution texture analysis

Datong Wei, Heang-Ping Chan,[a)] Nicholas Petrick, Berkman Sahiner, Mark A. Helvie, Dorit D. Adler, and Mitchell M. Goodsitt
Department of Radiology, University of Michigan Hospital, 2910 Taubman Center, Ann Arbor, Michigan 48109-0326

We investigated the application of multiresolution global and local texture features to reduce false-positive detection in a computerized mass detection program. One hundred and sixty-eight digitized mammograms were randomly and equally divided into training and test groups. From these mammograms, two datasets were formed. The first dataset (manual) contained four regions of interest (ROIs) selected manually from each of the mammograms. One of the four ROIs contained a biopsy-proven mass and the other three contained normal parenchyma, including dense, mixed dense/fatty, and fatty tissues. The second dataset (hybrid) contained the manually extracted mass ROIs, along with normal tissue ROIs extracted by an automated Density-Weighted Contrast Enhancement (DWCE) algorithm as false-positive detections. A wavelet transform was used to decompose an ROI into several scales. Global texture features were derived from the low-pass coefficients in the wavelet transformed images. Local texture features were calculated from the suspicious object and the peripheral subregions. Linear discriminant models using effective features selected from the global, local, or combined feature spaces were established to maximize the separation between masses and normal tissue. Receiver Operating Characteristic (ROC) analysis was conducted to evaluate the classifier performance. The classification accuracy using global features were comparable to that using local features. With both global and local features, the average area, A_z, under the test ROC curve, reached 0.92 for the manual dataset and 0.96 for the hybrid dataset, demonstrating statistically significant improvement over those obtained with global or local features alone. The results indicated the effectiveness of the combined global and local features in the classification of masses and normal tissue for false-positive reduction. © *1997 American Association of Physicists in Medicine.* [S0094-2405(97)01406-5]

Key words: mammography, computer-aided diagnosis, wavelet transform, texture analysis, linear discriminant function

I. INTRODUCTION

Breast cancer is the second leading cause of cancer death among American women.[1] Treatment of the cancer while it is still in its early stage is the most promising way to improve the chances of survival for patients with breast cancer.[2] Mammography is presently the most effective method for early detection of breast cancer[3] and is recommended as a routine procedure for the screening of breast cancers. However, among those women who have breast cancers and have undergone mammography, 10%–30% have negative mammograms.[4–7] Approximately two-thirds of these false negative diagnoses were due to missed lesions on the mammograms, which were evident retrospectively.[4,8,9] The low conspicuity of the radiological findings, poor image quality, eye fatigue, or oversight are the common causes of the missed detections.[4] Although double reading by two radiologists may increase sensitivity,[10] it also increases the cost in a mass screening program. As an alternative, CAD may be used to provide a second opinion to reduce the missed detection rate of breast cancer.[11,12]

The detection of masses is more difficult than the detection of microcalcifications due to the similarity between masses and breast tissue. Lai *et al.* used template-matching techniques to detect circumscribed masses after selective median filtering.[13] They found that the number of false positive detection was reduced by analyzing the cross correlation of neighboring pixels. Kegelmeyer *et al.* used the analysis of local oriented edges and a subset of Laws' texture energy features to detect spiculated masses.[12] Yin *et al.* utilized the architectural asymmetry between the right and left breasts to detect masses[14] Laine *et al.* proposed to use the wavelet transform and the ϕ transform for adaptive multiscale processing and contrast enhancement.[15] Brzakovic *et al.* employed a hierarchical region growing with pyramidal multiresolution image representation in the segmentation of microcalcifications and nodules.[16] The reported results of these studies varied. It is difficult to compare the performances of the different algorithms because they depend strongly on the datasets used.

We are developing computerized methods to detect masses on mammograms. In an early study, we used texture features of ROIs manually selected from digitized mammograms and linear discriminant analysis to classify masses and normal tissue.[17] Wei *et al.* used multiresolution texture

analysis[18] and Sahiner *et al.* used a convolution neural network[19,20] with the same dataset and achieved improved results. Adaptive density-weighted contrast-enhancement (DWCE) filtering with Laplacian–Gaussian edge detection was recently developed for the segmentation of low-contrast objects from digitized mammograms.[21–23]

In this study, we introduced the use of local texture features in combination with the global multiresolution texture features for the classification of masses and normal breast tissue. The new feature extraction approach was applied to the same set of manually extracted ROIs in order to compare with our previous results. In addition, false-positive ROIs automatically extracted by the DWCE algorithm were combined with the mass ROIs to examine the effectiveness of these features in reducing the number of false-positive detections. The classification accuracy was evaluated by Receiver Operating Characteristics (ROC) analysis and the improvement in accuracy due to the additional local texture features was examined. We also evaluated the robustness of the classification model by studying its performance with various feature set and training/test set combinations.

II. MATERIALS AND METHODS

A. Dataset

The mammograms used in this study were randomly selected from the patient files in the Department of Radiology at the University of Michigan. The selection criteria were that there was a biopsy-proven mass on the mammogram and there were no visible grid lines. The mammograms were acquired using a Kodak MinR/MRE screen/film system with extended cycle processing. The mammographic systems have a 0.3-mm focal spot, a molybdenum anode, a 0.03-mm-thick molybdenum filter and a 5:1 reciprocating grid. All systems have been certified by the American College of Radiology (ACR) and the image quality is monitored according to the ACRs recommended guidelines.

The films were digitized with a laser film scanner (LUMISYS DIS-1000) at a pixel size of 0.1 mm×0.1 mm with 12-bit gray level resolution. The optical density range of the digitizer was 0–3.5. The light transmission through the mammographic films was amplified logarithmically before digitization. The pixel values were calibrated such that they were linearly proportional to the optical density in the range of 0.1–2.8 optical density units.

The 168 case samples in the dataset contained a mixture of benign ($n = 85$) and malignant ($n = 83$) masses. Forty-five of the malignant masses and six of the benign masses were spiculated. The visibility of the masses was ranked by experienced radiologists on a scale of 1–10 (1=most obvious, 10=most subtle), which corresponded to the range of masses seen on clinical mammograms. The length of the long axis (size) of the masses was also measured by the radiologists and ranged from 5 to 26 mm with a mean size of 12.2 mm. Some of the mammograms were different views obtained from the same patient. A total of 72 different patients was included in the dataset. Different views of the same patients were treated as different case samples in most of the analy-

sis. The mammograms were divided randomly into two subgroups: one for training and one for test. A limited study was performed to evaluate the effect of the possible correlation between films of the same patient on classifier performance.

1. Manually extracted ROIs

We used manually extracted ROIs to study the feasibility of using the extracted features for the classification task and to compare with our previous studies.[18] Four different ROIs, each with 256×256 pixels, were selected by a radiologist experienced in mammography from each mammogram. One ROI contained a true mass, and the other three were normal parenchyma, containing dense tissue, mixed dense/fatty tissue, and fatty tissue, respectively. The dataset was divided randomly and equally into two, with the constraint that ROIs from the same film were grouped into the same subgroup. There were 84 true mass ROIs and 252 nonmass ROIs for each of the subgroups. In the following analysis, we denote the whole set with manually extracted ROIs as M, and the two subgroups as M_1 and M_2.

2. Hybrid dataset

In the manually extracted ROIs, the normal tissue regions were identified by radiologists according to certain criteria for the feasibility studies. The number and tissue type might be different from those extracted by a computer algorithm. To obtain a more realistic evaluation of our false-positive reduction method, we applied it to false positive ROIs extracted automatically by the DWCE procedure. The detailed description of the DWCE method can be found elsewhere.[23] Briefly, each mammogram was processed in two stages. In the first stage, the entire mammogram with reduced spatial resolution was globally filtered with a DWCE adaptive filter to enhance the local contrast of the image based on its local mean pixel values. A Laplacian–Gaussian (LG) edge detection procedure was then used to segment the image into isolated objects. In the second stage, the DWCE filter and the LG edge detector were applied locally to the isolated object regions detected in the first stage. The morphological features of the segmented objects were extracted and used to reduce the number of objects in both stages. The ROIs of the remaining objects were extracted from the full resolution images centered at the centroid locations of the detected objects.

The analysis of dataset containing automatically extracted mass and normal tissue ROIs was more complicated than the manually extracted one due to the possible overlap of the mass and normal tissue ROIs.[23] For our purpose of evaluating the effectiveness of different feature spaces in reducing the number of false positives, we formed a hybrid dataset by using the 168 biopsy-proven masses that were manually extracted (as discussed previously) and 1001 false-positive ROIs containing only non overlapping normal tissue extracted by the DWCE algorithm. In the following analysis, we denote this dataset as H and the two subgroups as H_1 and

H_2. There were 84 true masses in each subgroup, and 503 and 498 non overlapping ROIs with normal tissue in H_1 and H_2, respectively.

3. Adaptive background correction

Masses are superimposed on normal breast structure in a detected ROI. The gray level characteristics of the background structures are basically independent of those of the masses but they will offset some of the extracted properties based on gray level analysis. Therefore, it is necessary to remove the low-frequency background so that different masses can be compared on a common background level. In our previous study,[18] we simply substituted the mean gray level value of each ROI with a constant value that was the same for the entire dataset. An adaptive background correction method was since developed.[17,19] The algorithm estimates the background level based on the image intensity in a band of pixels surrounding the ROI. The first step was the calculation of a moving average of the pixel values along the perimeter of the ROI. A box filter of size 32×16, whose longer side was parallel to the side of the ROI, was used for calculating the moving average. These moving averages provided the estimated background pixel values along the perimeter (four sides) of the ROI. The background level inside the ROI was estimated using a weighted linear combination of these background perimeter pixels. The background level $B(i,j)$ of a pixel (i,j) was calculated as

$$B(i,j) = \frac{p_1/d_1 + p_2/d_2 + p_3/d_3 + p_4/d_4}{1/d_1 + 1/d_2 + 1/d_3 + 1/d_4}, \tag{1}$$

where p_1, p_2, p_3, p_4 are the background perimeter pixel values at the intersection between the four sides of the ROI and the normal directions from pixel (i,j) to each side, and d_1, d_2, d_3, d_4 are the distances between the pixel (i,j) and the intersections. The background corrected image was obtained as the difference between the original ROI and the background image $B(i,j)$. An example of the effects of this background correction method on an ROI can be found in Chan *et al.*[17]

B. Multiresolution texture features

Texture features were calculated from the spatial gray level dependence (SGLD) matrix.[24,25] The (i,j)th element of the SGLD matrix, $p_{d,\theta}(i,j)$, is the joint probability that the gray levels i and j occur in direction θ and at a distance of d pixels apart over the entire ROI. The SGLD matrix is a two-dimensional histogram based on image pixel values. Its size depends on the gray level resolution of the digitized image and the bin width used in determining the histogram. By changing the distance between the pixel pairs in defining the spatial relationship, different SGLD matrices can be constructed.

1. Global texture features

Our previous study[18] demonstrated the feasibility of using eight texture measures calculated from the entire ROI in a multiresolution framework for the classification. In this study, we used 13 texture measures described by Haralick *et al.*[24] The definitions of these texture features are summarized in the Appendix. We have described the methods for the extraction of multiresolution features previously.[18] Briefly, the wavelet transform was employed to decompose an ROI into three levels. Daubechies' filter with four coefficients[26] was used as the wavelet filter. The SGLD matrices were constructed at $d = 1$ for the original image (scale 1) and the subsampled approximation images from the low-pass quadrants in the wavelet coefficients of the next two levels (scales 2 and 4) after the wavelet transform. The wavelet coefficients at scale 8 were obtained with wavelet filtering, but no down-sampling was performed. Additional SGLD matrices were constructed from the approximation image in the wavelet coefficients at scale 8 with d ranging from 2 to 12. These distances corresponded to distances between pixel pairs of 8–48 pixels in the original image. Thirteen texture features were calculated at $\theta=0°$, 45°, 90°, and 135° for each distance. The features at 0° and 90° were averaged, so were those at 45° and 135°. There was a total of 364 features (13 texture measures $\times$14 distances $\times$2 angles) in the global texture feature space. A binwidth of 16 gray levels for the 12-bit image was chosen that corresponded to a reduction to eight-bit gray level resolution, as described previously.[18]

2. Local texture features

Global texture features described above summarize the general textural information of an entire ROI. Since the ROI contains both the suspicious mass object and its peripheral background, the global SGLD matrices formulated above reflect the average properties of the two regions. To further describe the information specific to the mass and the background normal tissue, we calculated local texture features from the local region containing a detected object (object region) and the peripheral regions within each ROI.

For this study, we assumed that the mass or the mass-like object in the false-positive ROI is an object of maximum intensity near the central area of the ROI. The search for this maximum intensity was accomplished by low-pass filtering and by locating the maximum pixel value in the low-pass filtered image. The low-pass filtering was implemented as a two-stage convolution with box filters of different kernel sizes. The kernel size was 60×60 pixels for the first stage and 15×15 pixels for the second. The maximum pixel value in the central 128×128 pixel area in the low-pass filtered ROI was assumed to be the center of the object region (90×90 pixels). The peripheral regions (64×64 pixels) were located at the four corners of the ROI. This segmentation was applied to all ROIs, including mass and normal tissue. Figure 1 illustrates the local regions in four ROIs extracted manually from a mammogram.

Since the size of the object region was small (90×90 pixels), we did not employ the wavelet transform for multiresolution analysis. As found in our previous studies, several SGLD matrices with different distances could be used as an alternative. Therefore, the object region and the peripheral
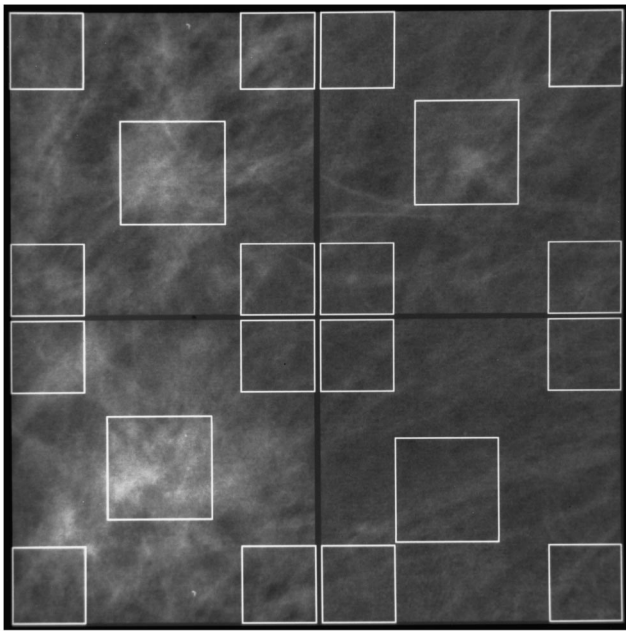
FIG. 1. The segmentation of subregions in the ROIs manually extracted from a mammogram. Upper left: the ROI with a mass; upper right: an ROI with mixed dense/fatty tissue; lower left: an ROI with dense tissue; lower right: an ROI with fatty tissue.

regions in the original images were used for the SGLD matrix formulation. An SGLD matrix of the peripheral region was formed by the accumulation of pixel pair information from all four peripheral subregions. For each region, SGLD matrices were calculated from $d = 1, 2, 4, 8$ and $\theta=0°, 45°, 90°, 135°$. Again, features for $\theta=0°, 90°$ and for $\theta=45°, 135°$ were averaged separately. Different bin widths of the SGLD matrices were evaluated. The local feature space consists of 104 features in the object region (13 texture measures $\times$4 distances$\times$2 angles) and 104 features obtained from the difference of corresponding features in the object and the peripheral regions.

C. Classification method

The classification of mass and normal tissue is a typical statistical problem: there are both between-class and within-class differences. The masses encountered in mammograms vary in shape, size, contrast, and projection, so does the normal tissue. It is extremely difficult, if not impossible, to construct an analytical model to describe the whole class of mass or normal tissue. Therefore, we adopted a linear statistical model to describe the difference between the masses and the normal breast tissues. Linear discriminant analysis[27] is a systematic statistical technique to classify individuals or cases into one of several mutually exclusive classes. For a two-class problem, feature variables are linearly combined to form a canonical discriminant function. The coefficients of the discriminant function are optimized on the basis of feature values of the training group to maximize the separation of the two classes. There are two issues involved in the model building process: the determination of the model components and the optimization of the model coefficients.

The predictor variables used in the discriminant model directly affect the effectiveness and the accuracy of the classification. Each of the global, local, and combined feature spaces was used as a pool of predictor variables. A stepwise feature selection procedure[28] with the maximization of Mahalanobis distance as the optimization criterion was used to select effective predictor variables from each feature pool. We have described the application of the stepwise feature selection procedure to our classification task previously.[17,18] The stepwise feature selection procedure was performed over the entire dataset (M or H) so that the selection would be based on feature distributions that have better statistical properties than those for the subgroups. For model coefficient optimization, the M_1 and M_2 (or H_1 and H_2) subgroups were alternately used as training and test sets. The coefficients of the linear discriminant function were optimized based on the feature values in the training group. The training cases were then classified with the linear discriminant function as a verification of consistency. The other group was used for testing the accuracy of the classifier and the discriminant score of each test case was calculated based on the linear discriminant function.

D. Statistical analysis of classification accuracy

Receiver Operating Characteristic (ROC) analysis[29,30] was used to evaluate the overall performance of the linear discriminant models. The distribution of the discriminant scores of the ROIs in the training or the test group was input into the LABROC1 program,[31] which provided a maximum likelihood estimation of a binormal ROC curve for training or testing, respectively. The area under the fitted ROC curve, A_z, was used as a performance index for the evaluation of the different sets of features selected from the multiresolution feature pools. The CLABROC program was employed to test the statistical significance of the difference between the A_z values of different sets of selected features.[32] The two-tailed p values were reported in the following comparisons. The statistical significance level was chosen at $\alpha=0.05$.

III. RESULTS

A. Manually extracted ROIs

From the manually extracted ROIs, 11 features were selected from the global feature space [Table I(a)]. When training was performed with M_1, the A_z for testing with M_2 was 0.86 ± 0.02 (Table II). When training was performed with M_2, the A_z for testing with M_1 was 0.88 ± 0.02. These results are slightly better than those (0.85 ± 0.03 and 0.86 ± 0.02) in the previous study,[18] where 8 of the 13 texture measures were calculated at several resolutions and/or distances. However, the improvement in A_z is not statistically significant. Table I(a) also shows the 19 features previously selected.[18] It can be seen that some of the additional five texture measures were selected at various distances, indicating that these texture measures contained useful information for the classification.

TABLE I. Texture features selected from the different feature spaces by stepwise linear discriminant analysis. z. axial feature (average of features at 0° and 90°); a: diagonal feature (average of features at 45° and 135°). ●: manual (M) dataset; △: hybrid (H) dataset. ◇: manual (M) dataset in our previous study (Ref. 18). The top eight features are the features used in our previous study and the bottom five are the new features evaluated in this study. F_{object}: object texture feature; $F_{object} - F_{periphery}$: difference in texture feature. (a) Global feature space, (b) local feature space, (c) combined feature space.

(a)

scale	1	2		4		8																
Distance	1	2		4		8		12	16		20		24	28	32		36		40		44	48
θ	z	z	a	z	a	z	a	a	z	a	z	a	z	z	z	a	z	a	z	a	z	a
correlation											•			Δ		Δ	◇				◇	
dif. entropy	◇			◇		◇	Δ				Δ	Δ				•Δ		◇			•◇	
energy							Δ														Δ	
entropy				Δ											◇	•Δ			Δ			◇
inertia				Δ				Δ									◇Δ					
inv. dif. moment				Δ	◇						◇									Δ	◇	Δ
sum average						◇											◇		•	◇	◇	•
sum entropy							Δ							◇		•	Δ				◇	
sum variance																						
dif. variance	Δ	Δ		Δ																		
dif. average							Δ				Δ	Δ									Δ	
info. meas. cor. 1						•Δ																•Δ
info. meas. cor. 2	Δ		Δ				Δ				•			Δ								•

(b)

	F_{object}				$F_{object} - F_{periphery}$									
Distance	1	2	4	8	1		2		4		8			
θ	z	a	a	z	z	a	z	a	z	a	z	a	z	a
correlation				Δ										
dif. entropy			•	Δ		Δ		•						Δ
energy	Δ			Δ		Δ			Δ			•		
entropy				Δ	•		•							
inertia														
inv. dif. moment	•Δ	•			•Δ									
sum average				Δ	•		Δ				•Δ			
sum entropy														
sum variance	•													
dif. variance											•			
dif. average		•Δ			Δ				Δ					
info. meas. cor. 1				Δ	•	Δ					•Δ	•	•Δ	
info. meas. cor. 2						Δ								

(c)

	Global feature space																										Local feature space																
																											F_{object}								$F_{object} - F_{periphery}$							
Distance	1	2	4	8	12	16	20	24	28	32	36	40	44	48													1	2	4	8					1	2	4	8				
θ	z	a	z	a	a	z	a	z	a	z	z	a	z	a	z	a	z	a	z	a	z	a	z	a	z	a	z	a	z	a	z	a	z	a	z	a	z	a	z	a	z	a
correlation	•		•							Δ																			Δ							Δ	Δ				
dif. entropy				Δ	Δ					Δ									•Δ								•		Δ													
energy	•				•		•																		•				•												•
entropy		Δ								Δ	•														•				Δ	•												
inertia														Δ					Δ										•	•Δ	Δ	•	•		Δ							
inv. dif. moment		Δ	•						•	•			•														Δ		Δ		•	•	•									
sum average								Δ						Δ	•						Δ		Δ														Δ	•
sum entropy					Δ		•			•	Δ																		Δ													
sum variance																													Δ													
dif. variance		•					•																								•	Δ			Δ	•		•
dif. average		•			•				Δ	Δ			Δ												•		•		•									
info. meas. cor. 1	Δ	•				•			•	Δ			•Δ	Δ															•Δ			Δ				Δ
info. meas. cor. 2			•		•										Δ																											Δ

TABLE II. Comparison of the area under the ROC curves, A_Z, obtained from different feature spaces for the set of manually extracted ROIs (*M* set).

Feature space		Global	Local	Combined
Number of features		11	17	45
Training set	Test set			
M_1+M_2	M_1+M_2	0.89±0.02	0.89±0.01	0.95±0.01[a,b]
M_1	M_1	0.90±0.02	0.88±0.02	0.96±0.01[a,b]
M_1	M_2	0.86±0.02	0.84±0.03	0.92±0.02[a,b]
M_2	M_2	0.88±0.02	0.88±0.02	0.95±0.02[a,b]
M_2	M_1	0.88±0.02	0.87±0.02	0.91±0.02

[a]The improvement is statistically significant at $\alpha=0.05$, comparing combined to global feature space.
[b]The improvement is statistically significant at $\alpha=0.05$, comparing combined to local feature space.
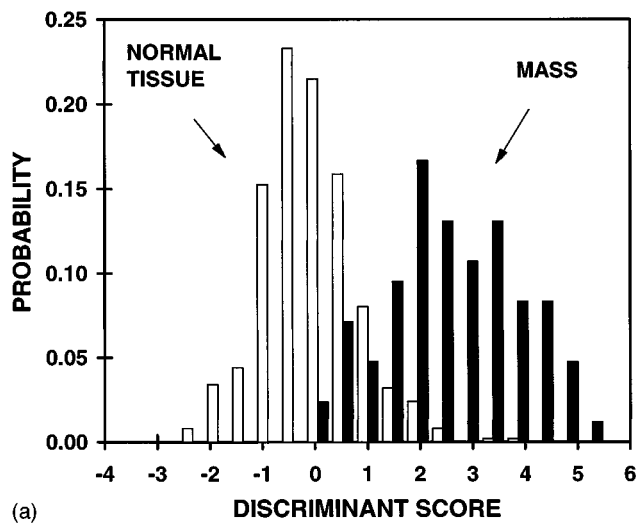
If we used only the local feature space as a pool, 17 features were selected by the stepwise procedure [Table I(b)]. From Table II, the A_z for testing with M_2 was 0.84 ±0.03, while that for testing with M_1 was 0.87±0.02. Statistical tests showed that the differences in A_z values between the local and global spaces were not significant. Therefore the effectiveness of the local feature space was comparable to that of the global feature space.

When we combined the local and global feature spaces as a single pool of 572 features, 45 features were chosen by the stepwise procedure. Table I(c) illustrates the distribution of the selected features in the combined feature space. Table II demonstrates the improvement in A_z values in both the training and the test groups. The A_z for testing with M_2 improved to 0.92±0.02. The improvement was statistically significant compared to that with either the global or local feature space alone. The A_z for testing with M_1 reached 0.91±0.02, although the improvement did not achieve statistical significance for this condition. The improvements in A_z for the training groups were statistically significant, over either the global or local features alone. The effect of bin widths of the SGLD matrices on classification accuracy in the global feature space was evaluated in previous studies.[17] We performed a similar comparison in the local feature space. Table III(a) summarizes the A_z values for different bit depths (varied from seven to ten bits) used in the construction of SGLD matrices, where the local feature space alone was used for classification. Table III(b) lists the A_z values with the combined feature space, where the global texture features were calculated at eight bits while the bit depth of the local texture features was varied from seven to ten bits. Although there was a drop in A_z values at ten bits for the local feature space alone, the dependence of A_z on bit depth for the combined feature space was not statistically significant. Considering these results, we chose nine bits (equivalent to a bin width of eight gray levels) for all local feature calculations in this study.

B. Hybrid dataset

For the manually extracted mass ROIs and automatically extracted normal tissue ROIs, the 32 features selected from the global feature space are shown in Table I(a). The A_z

TABLE III. (a) A_z values with features from local feature space with different bin widths for the manual dataset. (b) A_z values with features from different combined feature spaces (local feature space with different bin widths) for the manual dataset.

(a)

Bin width (pixel value)		32	16	8	4
Equivalent bit depth		7	8	9	10
Number of features		19	18	17	7
Training set	Test set				
M_1+M_2	M_1+M_2	0.87±0.02	0.89±0.01	0.89±0.02	0.85±0.02
M_1	M_1	0.91±0.02	0.91±0.02	0.88±0.02	0.89±0.02
M_1	M_2	0.86±0.02	0.86±0.02	0.84±0.03	0.80±0.03
M_2	M_2	0.88±0.02	0.89±0.02	0.88±0.02	0.81±0.03
M_2	M_1	0.88±0.02	0.88±0.02	0.87±0.02	0.87±0.02

(b)

Bin width (pixel value)		32	16	8	4
Equivalent bit depth		7	8	9	10
Number of features		31	38	45	42
Training set	Test set				
M_1+M_2	M_1+M_2	0.94±0.01	0.95±0.01	0.95±0.01	0.95±0.01
M_1	M_1	0.95±0.01	0.96±0.01	0.96±0.01	0.97±0.01
M_1	M_2	0.92±0.02	0.91±0.02	0.92±0.02	0.90±0.02
M_2	M_2	0.94±0.02	0.95±0.01	0.95±0.02	0.95±0.01
M_2	M_1	0.93±0.02	0.91±0.02	0.91±0.02	0.91±0.02

values for the two test groups were 0.91±0.02 and 0.90 ±0.02, respectively (Table IV). The 22 features selected from the local feature space alone is shown in Table I(b). The A_z values for the two test groups were 0.92±0.02 and 0.95±0.01, respectively (Table IV). When we combined the global and local feature spaces, 41 features were selected [Table I(c)]. The A_z values improved to 0.96±0.01 and 0.97 ±0.01, respectively. The improvements in the A_z values for the test groups with the combined feature space were statistically significant compared to those with either the global or local feature space alone.
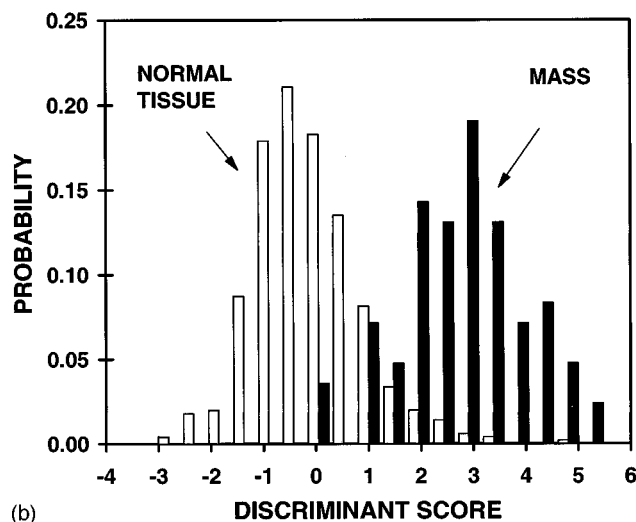
Figure 2 shows the distribution of the discriminant scores for the test groups in the hybrid dataset. Figure 3 shows the ROC curves for the global, local, and the combined feature

TABLE IV. Comparison of the area under the ROC curves, A_z, obtained from different feature spaces for the hybrid dataset.

Feature space		Global	Local	Combined
Number of features		32	22	41
Training set	Test set			
H_1+H_2	H_1+H_2	0.93±0.01	0.96±0.01	0.97±0.01[b,c]
H_1	H_1	0.93±0.01	0.96±0.01	0.98±0.01[a,b]
H_1	H_2	0.91±0.02	0.92±0.02	0.96±0.01[a,b]
H_2	H_2	0.92±0.02	0.93±0.02	0.97±0.01[a,b]
H_2	H_1	0.90±0.02	0.95±0.01	0.97±0.01[a,b]

[a]The improvement is statistically significant at $\alpha=0.05$, comparing combined to global feature space.
[b]The improvement is statistically significant at $\alpha=0.05$, comparing combined to local feature space.
[c]CLABROC did not converge when comparing combined to global feature space.

FIG. 2. Distribution of discriminant scores for the test subgroups. (a) Test with H_2. (b) Test with H_1.

FIG. 3. Comparison of ROC curves for test groups using features from the global, local, and the combined feature spaces. (a) Test with H_2. (b) Test with H_1.
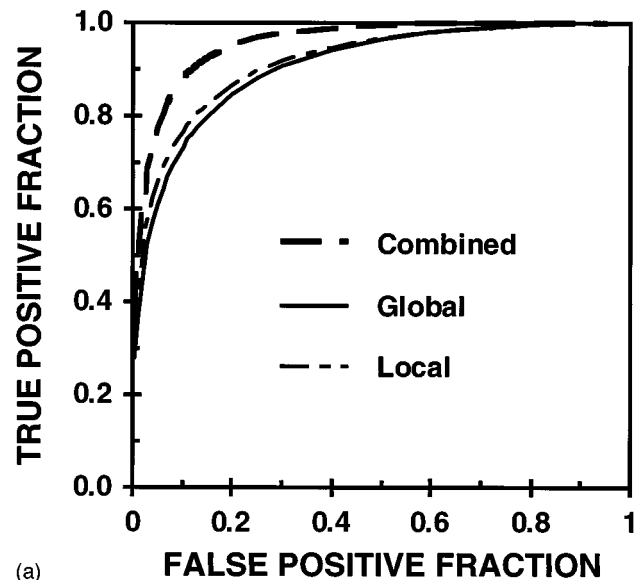
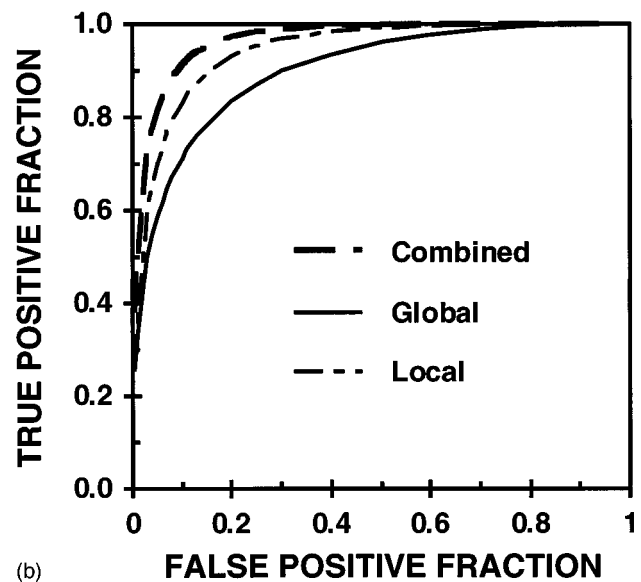spaces, illustrating the improvement in the classification accuracy.

IV. DISCUSSION

We demonstrated previously that texture features in a multiresolution framework could be used to classify masses and normal breast tissue on mammograms. The objectives of this study are to expand the feature pool for the linear discriminant model, and to evaluate the effectiveness of global and local multiresolution texture features for the reduction of false-positive ROIs. In the following, we will discuss the relations and differences among the global and local feature pools, their underlying physical meaning, and the robustness issue of the discriminant model and our approach.

A. Inclusion of more texture measures

Compared with our previous study, we included five more texture measures in the feature pool as potential candidates in the linear discriminant model. Our results demonstrated

that some of these additional texture measures were chosen at various scales and distances by the stepwise feature selection procedure. Under similar conditions to the previous study, the number of features selected from the global feature space decreased from 19 to 11, but the classification results for the test groups were slightly better than those obtained previously [Table I(a)]. These new features therefore appear to include useful information for the differentiation between masses and normal tissue. However, many features in the global space are related to each other. The different texture measures (see the Appendix) essentially describe the shape of the same two-dimensional histogram (or SGLD matrix) from different perspectives. For example, the two information measures of correlation selected in the global feature space [Table I(a)] are nonlinear functions of entropy texture

(see the Appendix). On the other hand, the reason that the multiresolution (essentially multidistance) framework in our previous studies[18] provided statistically better results than the single distance (one SGLD matrix)[17] is the availability of several SGLD matrices that describe the images from different spatial relationships. In order to significantly improve the classification accuracy, it seems to be more important to provide additional and complementary information to the multiresolution feature space than to summarize the information with more texture measures from different perspectives.

B. Difference between global and local features

In this section, we try to examine the difference in the information contained in the global and local features. In the following discussion, ''detail'' is relative and used as a general term that can be defined from different perspectives.

Although the same texture measures and similar multiresolution (or multidistance) method are used, the emphases of the global and local feature spaces are different. The multiresolution analysis with global features summarizes the overall (or ''structural'') information of mass and normal tissue ROIs, since the original image and the low-pass versions of an ROI are the input images to the feature analysis and the extracted global features are directly related to these images. Although the original image contains complete (including overall and detail) information of an ROI, only three features from the original image were selected for the hybrid dataset and none for the manual dataset [Table I(a)]. Therefore, the emphasis of the linear combination of the selected global features is on the overall, rather than detail, information due to the features that are extracted from the low-pass images at larger scales and distances.

It is well known that detailed information plays an important role in the classification of masses by radiologists in mammographic reading. In order to improve the classification accuracy, it is necessary to extract features that can describe image details. The different perspectives and references from which radiologists summarize the detail information, however, remain to be explained. For example, the details revealed from the comparison of mammograms of the same breast obtained at different times are usually different from the details by the asymmetry analysis of the right and left breasts. They are also different from the detailed information contained in the high-pass quadrants of the wavelet coefficients. Although the wavelet transform provides a detailed analysis in a multiresolution framework, we have not found a statistically significant difference in the individual texture features of the high-pass wavelet coefficients between mass and normal tissue ROIs. This might be due to the difficulty in separating the detailed information between the mass and tissue in the high-frequency domain in the presence of image noise. On the other hand, the references from past mammograms or bilateral mammograms are not available for the current analysis of single-view mammograms.

If we examine carefully the detailed analyses by the radiologists and the wavelet transform, we can find that the ref-

erences for the details are different. For radiologists, the references are changes in the tissue structures and appearance of suspicious objects in the same breast; or asymmetry of tissue distribution in the left and right breasts. For the wavelet transform, the references are related to the different frequency bands, where the spatial relationships between pixels are defined. By partitioning an ROI into object and peripheral regions and using these regions as input to the feature analysis, we are changing our reference point and examining the variations of the objects with respect to peripheral regions. Some local features are related to image details, particularly those features extracted from SGLD matrices with $d = 1$. The spatial relationship between pixel pairs at this distance is defined at the highest possible spatial resolution. From Table I(b) we can see that several texture features in the local feature space were selected at $d = 1$. We can also examine the features from the way they are extracted. In global feature extraction, the pixels in the object and its surrounding background of an input ROI are accumulated in the same SGLD matrices, resulting in an averaging of the object and its background information. The smaller the object, the stronger is the influence of the background. For local feature extraction, the averaging effect becomes relatively small since the object region size is 9×9 mm, smaller than the size of most of the masses in our database. The pixels in the object region are used to construct SGLD matrices for the object, thereby providing features that are more specific to the individual objects than the global features. The periphery SGLD matrices are based on the pixels at the four corners of the ROI, so that they provide an estimate of the average properties of the normal tissue background in an ROI. The resulting texture features for the periphery region are less sensitive to the variations in the tissue background than those for each individual subregion. Along with the features for the object subregion, the differences in the corresponding features between the object subregion and its peripheral subregions (i.e., the difference features) are included in the local feature space. These features emphasize the difference between the object and its surrounding tissue background and, therefore, are related to the change in texture in the object region. They also normalize, to some extent, the change in texture by providing the same reference point, i.e., the features of their own background, in the comparison of different objects. Local feature extraction cannot be considered a special case of global feature extraction, since there is no equivalence in the global feature space to the difference features in the local space.

Like ''detail'' or local, the term ''global'' is also relative. The current global feature space is ''global'' with respect to the extracted ROI. If we examine the whole mammogram, this ''global'' feature space becomes ''local.'' Radiologists usually analyze a suspicious object in the context of breast anatomy in their detection process. This anatomical information is more ''global'' than those global features that we extracted. There is no doubt that this kind of structural information at a higher level will provide additional and complementary information in the classification of mass and normal tissue. The inclusion of this kind of information, however, is

beyond the scope of the present study and deserves further investigation in the process of improving the classification accuracy and intelligence of CAD algorithms.

C. Segmentation of suspicious object

With features from the local space alone, we have achieved statistically similar classification accuracy to that with the global features. This implies that the local feature space is as effective as the global space. Evidently, the effectiveness of the local feature space depends on the accurate segmentation of the suspicious object in the ROI. This may become difficult when a subtle mass is adjacent to a structured background tissue. For our feasibility studies, we used a low-pass filter to identify an object based on the mean pixel intensity. This technique proved to be effective for most of the masses in our database, except for one mass that had a piece of dense tissue in the proximity. More sophisticated methods may be necessary to identify subtle masses neighboring dense tissue structure in order to further improve the classification accuracy. It may also be helpful if the size of the object regions is varied according to the size of the individual masses rather than fixed, as used in this study. The DWCE algorithm can provide the location and the size of a bounding box for a suspicious object.[21] We also used this information in our extraction of local texture features. The preliminary results demonstrated a similar level of classification accuracy, suggesting that for the dataset used in this study, our object identification and the feature extraction methods can identify the object location correctly in most cases.

D. Complementary information in the global and local feature spaces

The statistically significant improvement in the classification accuracy with the combined feature space indicates that the global and local features characterize different information in an image. It is a result of the difference in the image regions subjected to SGLD formulation. It indicates that both the structural information associated with the global features based on the entire ROI and the detail information associated with the local features based on subregions in the ROI are necessary to distinguish masses from normal tissue. Although there may be some correlation between the two feature spaces, the stepwise feature selection procedure can select those important features that are complementary to each other. One would also expect that in addition to the global and local feature spaces described here, features that can effectively summarize the detail information embedded in the high-frequency components of the wavelet coefficients, if properly extracted, may be able to contribute to further improvement in the classification accuracy.

E. Hybrid dataset

Our hybrid dataset included manually extracted masses and independent normal tissue ROIs that were detected as false positives in an automated detection program. We did not use the automatically extracted mass ROIs and overlapping normal tissue ROIs[23] in this study. Since our DWCE algorithm at present did not detect all the masses, this would reduce the size of our dataset and the results could not be compared to those for the manual dataset. Our goals in this study were to compare the classification accuracy of the global, local, and combined feature spaces, and to demonstrate that the classifier can distinguish masses from normal tissue and thus can be used to reduce FPs detected by an automated algorithm. The effectiveness of the classifier as an FP reduction technique in an automated mass detection program has been evaluated in a different study.[23]

The difference in A_z values between the manual and hybrid datasets can be explained by the difference in the composition of the FP ROIs. In the H set, the FPs segmented by the DWCE algorithm mainly consisted of dense tissue rather than fatty tissue. For the M set, the dense and dense/fatty ROIs were extracted for their similarity in appearance to a potential mass, and they were generally different from the fatty tissue ROIs.[17,18] Although it was much easier to classify mass and fatty tissue ROIs than to classify mass and dense or dense/fatty ROIs separately, the inclusion of the three different types of FPs together in the normal tissue class increased the within-class variability. Since the FPs in the H set were more homogeneous than those in the M set, it would be easier and more effective for the discriminant model to maximize the ratio of the between-class variation and the within-class variation for the H set than for the M set. This maximization is the essence of the linear discriminant analysis in the separation between mass and normal tissue.

There are two purposes for including the manual dataset: (1) The manual dataset was used in our previous study before the automated segmentation was developed. Therefore, in this paper, it served as a reference point on how the local and combined texture feature spaces improved the classification. (2) The statistical model (including feature selection and model parameter optimization) is data dependent, especially when the number of cases is small. The analysis of two datasets with different FP characteristics provided some information on the adaptivity of our approach. This will be discussed in more detail in the following section.

F. Robustness of the discriminant model and the model building procedure

In robust statistics,[33] one seeks effective prediction methods that are rather insensitive, or robust against, certain types of failures in the model, so that good answers are still obtained, even if some assumptions are only approximately true. The robustness of the classification model is an important issue in computer-aided diagnosis. Whether an optimized model based on a limited dataset can achieve a reasonable accuracy in the general patient population determines the success or failure of that model.

The major issue of the classification model is the selection of model components, i.e., feature selection. To examine empirically the robustness of a selected feature set, we applied the features selected from the manual set to the hybrid set.

The resulting linear model using these feature variables is different from the optimal one. When the coefficients of the classifier were trained on the H_1 subset, the A_z values were 0.97 ± 0.01 for the training set H_1 and 0.95 ± 0.01 for the test set H_2. When the classifier was trained on H_2, the A_z values were 0.96 ± 0.01 for the training set and 0.95 ± 0.01 for the test set. Compared to Table IV, we can see that the A_z values obtained using this suboptimal feature set were only 0.01 to 0.02 lower than the A_z values obtained by optimal feature selection. We also applied the 41 features selected from the hybrid dataset to the manual set. When the classifier was trained on M_1, the training A_z reduced to 0.93 ± 0.02 and the test A_z reduced to 0.87 ± 0.02. When the classifier was trained on M_2, the training A_z reduced to 0.93 ± 0.02 while the test A_z reduced to 0.89 ± 0.02. These were about 0.02–0.05 lower than the corresponding optimal A_z values in Table II. The slightly larger drop in A_z values may be due to the fact that the characteristics of the FPs in the hybrid dataset are a subset of the characteristics of the FPs in the manual set, as discussed in the previous section.

The consistently lower A_z values, compared to those obtained with the optimal feature set observed from these experiments confirm the fact that selection of a feature set from a different population is suboptimal. However, a more important observation is that the suboptimal feature sets can consistently provide very high A_z values under all conditions studied, which were better than (15 out of 16 feature and training/test set combinations) or equal to (1 out of 16) the A_z values obtained from the optimal global or local feature sets alone. This is a strong indication that the feature sets are not very sensitive to the type of false positives in the hybrid or manual dataset.

To explore empirically the robustness of the model with respect to training and test set partitioning, we repartitioned the hybrid dataset into two subsets based on cases. In the new partitioning, all the ROIs belonging to the same patient were grouped into the same subset (either training or test). The training and test procedures were the same as those described in Sec. II C, except for the difference in the data partitioning. The test A_z values were 0.97 ± 0.01 and 0.96 ± 0.01 for the two subsets, respectively. The test A_z values for this by-case partitioning were within 0.01 from those for the by-film partitioning (see last column of Table IV). The differences were not statistically significant ($p = 0.48$), as estimated from the z score. This suggests that, for our dataset, the effect of the possible correlation between the different ROIs from the same patient partitioned into the training and test subsets was less important than the effect of feature selection on the classification accuracy.

It is important to recognize that our empirical analysis on two limited datasets (the M and H sets) does not prove that either the feature sets or the trained classifiers are generalizable to the patient population at large. It will be necessary to retrain the classifier when the characteristics of the false positives change or when large and general datasets become available, as we discussed previously.[17,18] However, the effectiveness of the classification for both the hybrid and manual datasets, which contained false positives of some-what different characteristics, demonstrates the robustness of our model building approach in differentiating mass and normal tissue ROIs.

V. CONCLUSION

This study significantly expanded our previous work by (i) introducing a new local texture feature pool to provide additional discriminatory information to the existing model; (ii) classifying FPs extracted automatically from our computer segmentation program; and (iii) evaluating the robustness of our classification model. Global features from the approximation images in the low-pass wavelet coefficients were used to describe the structural information about the mass and normal tissue, while the local texture features were used to differentiate the specific information of the mass and mass-like normal tissue from their background tissue structure. The classification capability of these two feature spaces were statistically comparable on their own. When the discriminant model included features from the combined feature pool, the improvement in the classification accuracy was, in general, statistically significant for both the manually extracted dataset and the hybrid dataset. This demonstrates the effectiveness of multiresolution feature analysis in the false-positive reduction for automated detection of masses.

APPENDIX: DEFINITION OF SGLD TEXTURE MEASURES

An SGLD matrix element, $p_{\theta,d}(ij)$, is the joint probability of the gray level pairs i and j in a given direction θ separated by a distance of d pixels. For each ROI, 13 texture measures were derived from its SGLD matrix of a given θ and d. The following provides a summary of the mathematical definitions of the texture measures used in our study. A simplified notation $p(i,j)$ will be used to denote the SGLD matrix elements.

(1) Energy.

$$\text{ENERGY} = \sum_{i=0}^{n-1} \sum_{j=0}^{n-1} p^2(i,j),$$

where n is the number of gray levels of the image.

(2) Correlation.

$$\text{CORRELATION} = \frac{\sum_{i=0}^{n-1}\sum_{j=0}^{n-1}(i-\mu_x)(j-\mu_y)p(i,j)}{\sigma_x \sigma_y},$$

where

$$\mu_x = \sum_{i=0}^{n-1} i p_x(i), \quad \sigma_x^2 = \sum_{i=0}^{n-1} (i-\mu_x)^2 p_x(i),$$

$$\mu_y = \sum_{j=0}^{n-1} j p_y(j), \quad \sigma_y^2 = \sum_{j=0}^{n-1} (i-\mu_y)^2 p_y(j),$$

$$p_x(i) = \sum_{j=0}^{n-1} p(i,j), \quad p_y(j) = \sum_{i=0}^{n-1} p(i,j),$$

are the mean and variance of the marginal distributions $p_x(i)$ and $p_y(j)$, respectively.

(3) Entropy.

$$\text{ENTROPY} = -\sum_{i=0}^{n-1} \sum_{j=0}^{n-1} p(i,j) \log_2 p(i,j).$$

(4) Inertia.

$$\text{INERTIA} = \sum_{i=0}^{n-1} \sum_{j=0}^{n-1} (i-j)^2 p(i,j).$$

(5) Inverse difference moment.

INVERSE DIFFERENCE MOMENT

$$= \sum_{i=0}^{n-1} \sum_{j=0}^{n-1} \frac{1}{1+(i-j)^2} p(i,j).$$

(6) Sum average.

$$\text{SUM AVERAGE} = \sum_{k=0}^{2n-2} k p_{x+y}(k),$$

where

$$p_{x+y}(k) = \Sigma_{i=0}^{n-1} \Sigma_{j=0}^{n-1} p(i,j),$$

$$i+j=k, \qquad k=0,\dots,2n-2.$$

(7) Sum entropy.

$$\text{SUM ENTROPY} = -\sum_{k=0}^{2n-2} p_{x+y}(k) \log_2 p_{x+y}(k).$$

(8) Difference entropy.

$$\text{DIFFERENCE ENTROPY} = -\sum_{k=0}^{n-1} p_{x-y}(k) \log_2 P_{x-y}(k),$$

where

$$p_{x-y}(k) = \Sigma_{i=0}^{n-1} \Sigma_{j=0}^{n-1} p(i,j),$$

$$|i-j|=k, \quad k=0,\dots,n-1.$$

(9) Sum variance.

SUM VARIANCE

$$= \sum_{k=0}^{2n-2} (k - \text{SUM AVERAGE})^2 p_{x+y}(k).$$

(10) Difference average.

$$\text{DIFFERENCE AVERAGE} = \sum_{k=0}^{n-1} k p_{x-y}(k).$$

(11) Difference variance. Notice the difference in the definitions of this feature here and in Haralick *et al.*[22]

DIFFERENCE VARIANCE

$$= \sum_{k=0}^{n-1} (k - \text{DIFFERENCE AVERAGE})^2 p_{x-y}(k).$$

(12) Information Measure of Correlation 1.

$$\text{IMC}_1 = \frac{\text{ENTROPY} - H_1}{\max\{H_X, H_Y\}},$$

where

$$H_1 = -\sum_{i=0}^{n-1} \sum_{j=0}^{n-1} p(i,j) \log_2 [p_x(i) p_y(j)],$$

$$H_X = -\sum_{i=0}^{n-1} p_x(i) \log_2 [p_x(i)],$$

$$H_Y = -\sum_{j=0}^{n-1} p_y(j) \log_2 [p_y(j)].$$

(13) Information measure of correlation 2.

$$\text{IMC}_2 = \sqrt{1 - \exp[-2(H_2 - \text{ENTROPY})]},$$

where

$$H_2 = -\sum_{i=0}^{n-1} \sum_{j=0}^{n-1} p_x(i) p_y(j) \log_2 [p_x(i) p_y(j)].$$

[a] Electronic-mail: chanhp@umich.edu

[1] E. Silverberg, C. C. Boring, and T. S. Squires, ''Cancer statistics, 1990,'' Cancer J. Clin. **40**, 9–26 (1990).

[2] S. A. Feig and R. E. Hendrik, ''Risk, benefit, and controversies in mammographic screening,'' in *Syllabus: A Categorical Course in Physics: Technical Aspects of Breast Imaging*, edited by A. G. Haus and M. J. Yaffe (Radiological Society of North America, Inc., Oak Brook, IL, 1993).

[3] H. C. Zuckerman, ''The role of mammography in the diagnosis of breast cancer,'' *Breast Cancer: Diagnosis and Treatment*, edited by I. M. Ariel and J. B. Cleary (McGraw-Hill, New York, 1987), pp. 152–172.

[4] J. E. Martin, M. Moskowitz, and J. R. Milbrath, ''Breast cancer missed by mammography,'' Am. J. Roentgenology **132**, 737–739 (1979).

[5] S. R. Pollei, F. A. Mettler, Jr., S. A. Bartow, G. Moradian, and M. Moskowitz, ''Occult breast cancer: Prevalence and radiographic detectability,'' Radiology, **163**, 459–462 (1987).

[6] C. J. Baines, A. B. Miller, C. Wall *et al.*, ''Sensitivity and specificity of first screen mammography in the Canadian National Breast Screening Study: A preliminary report from five centers,'' Radiology **160**, 295–298 (1986).

[7] L. W. Bassett and R. H. Gold, *Breast Cancer Detection: Mammography and Other Methods in Breast Imaging* (Grune and Stratton, Orlando, 1987).

[8] J. N. Wolfe, A. F. Saftlas, and M. Salane, ''Mammographic parenchymal patterns and quantitative evaluation of mammographic densities: A case-control study,'' Am. J. Roentgenology **148**, 1087–1092 (1987).

[9] J. A. Harvey, L. L. Fajardo, and C. A. Innis, ''Previous mammograms in patients with impalpable breast carcinoma: Retrospective vs blinded interpretation,'' Am. J. Roentgenology **161**, 1167–1172 (1993).

[10] E. L. Thurfjell, K. A. Lernevall, and A. A. S. Taube, ''Benefit of independent double reading in a population-based mammography screening program,'' Radiology, **191**, 241–244 (1994).

[11] H. P. Chan, K. Doi, C. J. Vyborny, R. A. Schmidt, C. E. Metz, K. L. Lam, T. Ogura, Y. Z. Wu, and H. MacMahon, ''Improvement in radiologists' detection of clustered microcalcifications on mammograms: The potential of computer-aided diagnosis,'' Invest. Radiol. **25**, 1102–1110 (1990).

[12] W. P. Kegelmeyer, Jr., J. M. Pruneda, P. D. Bourland, A. Hillis, M. W. Riggs, and M. L. Nipper, ''Computer-aided mammographic screening for spiculated lesions,'' Radiology **191**, 331–337 (1994).

[13] S. M. Lai, X. Li, and W. F. Bischof, ''On techniques for detecting circumscribed masses in mammograms,'' IEEE Trans. Med. Imag. **8**(4), 377–386 (1989).

[14] F. F. Yin, M. L. Giger, C. J. Vyborny, K. Doi, and R. A. Schmidt, ''Comparison of bilateral-subtraction and single-image processing techniques in the computerized detection of mammographic masses,'' Invest. Radiol. **28**(6), 473–481 (1993).

[15] A. Laine, S. Song, J. Fan, W. Huda, and J. Honeyman, ''Adaptive multiscale processing for contrast enhancement,'' Proc. SPIE, **1905**, 521–532 (1993).

[16] D. Brzakovic, P. Brzakovic, and M. Neskovic, ''An approach to automated screening of mammograms,'' Proc. SPIE, **1905**, 690–701 (1993).

[17] H. P. Chan, D. Wei, M. A. Helvie, B. Sahiner, D. D. Adler, M. M. Goodsitt, and N. Petrick, ''Computer-aided classification of mammographic masses and normal tissue: Linear discriminant analysis in texture feature space,'' Phys. Med. Biol. **40**, 857–876 (1995).

[18] D. Wei, H. P. Chan, M. A. Helvie, B. Sahiner, N. Petrick, D. D. Adler, and M. M. Goodsitt, ''Classification of mass and normal breast tissue on digital mammograms: Multiresolution texture analysis,'' Med. Phys. **22**, 1501–1513 (1995).

[19] B. Sahiner, H. P. Chan, N. Petrick, D. Wei, M. A. Helvie, D. D. Adler, and M. M. Goodsitt, ''Classification of mass and normal breast tissue: A convolution neural network classifier with spatial domain and texture images,'' IEEE Trans. Med. Imag. **15**, 598–610 (1996).

[20] D. Wei, B. Sahiner, H. P. Chan, and N. Petrick, ''Detection of masses on mammograms using a convolution neural network,'' *Proceedings of the of IEEE ICASSP-95*, Detroit, MI, 1995, Vol. 5, pp. 3483–3486.

[21] N. Petrick, H. P. Chan, B. Sahiner, and D. Wei, ''Automated detection of breast masses on digital mammograms using adaptive density weighted contrast enhancement filtering,'' IEEE Trans. Med. Imag. **15**, 53–67 (1996).

[22] N. Petrick, H. P. Chan, B. Sahiner, and D. Wei, ''Automated detection of breast masses on digital mammograms using adaptive density-weighted contrast-enhancement filtering,'' *Proceedings of the SPIE, Medical Imaging: Image Processing*, San Diego, CA, 26 February–2 March, 1995, Vol. 2434, pp. 590–597.

[23] N. Petrick, H. P. Chan, D. Wei, B. Sahiner, M. A. Helvie, and D. D. Adler, ''Automated detection of mammographic breast masses using computer segmentation and tissue classification,'' Med. Phys. **23**, 1685–1696 (1996).

[24] R. M. Haralick, K. Shanmugam, and I. Dinstein, ''Texture features for image classification,'' IEEE Trans. Syst. Man. Cybern. **3**, 610–621 (1973).

[25] R. W. Conner, ''Towards a set of statistical features which measure visually perceivable qualities of textures,'' Proc. IEEE Conf. Patt. Recog. Imag. Proc. **X**, 382–390 (1979).

[26] I. Daubechies, ''Orthonormal bases of compactly supported wavelets,'' Comments Pure Appl. Math. **41**, 909–996 (1988).

[27] P. A. Lachenbruch, *Discriminant Analysis* (Hafner, New York, 1975).

[28] M. J. Norusis, *SPSS for Windows Professional Statistics, Release 6.0* (SPSS, Chicago, 1993).

[29] J. A. Swets and R. M. Pickett, *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory* (Academic, New York, 1982).

[30] C. E. Metz, ''ROC methodology in radiologic imaging,'' Invest. Radiol. **21**, 720–733 (1986).

[31] C. E. Metz, J. H. Shen, and B. A. Herman, ''New methods for estimating a binormal ROC curve from continuously distributed test results,'' presented at the *1990 Annual Meeting of the American Statistical Association*, Anaheim, CA, 7 August 1990.

[32] C. E. Metz, P. L. Wang, and H. B. Kronman, ''A new approach for testing the significance of differences between ROC curves measured from correlated data,'' *Information Processing in Medical Imaging: Proceedings of the 8th Conference*, edited by F. Deconinck, Brussels, 29 August–2 September 1983 (Martinus Nijhoff, Boston, 1984), pp. 432–445.

[33] W. Stahel and S. Weisberg, in *Directions in Robust Statistics and Diagnostics, The IMA Volumes in Mathematics and Its Applications* (Springer-Verlag, New York, 1991), Vol. 33.