Phys. Med. Biol. 40 (1995) 857–876. Printed in the UK

Computer-aided classification of mammographic masses and normal tissue: linear discriminant analysis in texture feature space

Heang-Ping Chan, Datong Wei, Mark A Helvie, Berkman Sahiner, Dorit D Adler, Mitchell M Goodsitt and Nicholas Petrick

Department of Radiology, University of Michigan, Ann Arbor, MI, USA

Received 16 August 1994, in final form 3 February 1995

Abstract. We studied the effectiveness of using texture features derived from spatial grey level dependence (SGLD) matrices for classification of masses and normal breast tissue on mammograms. One hundred and sixty-eight regions of interest (ROIs) containing biopsy-proven masses and 504 ROIs containing normal breast tissue were extracted from digitized mammograms for this study. Eight features were calculated for each ROI. The importance of each feature in distinguishing masses from normal tissue was determined by stepwise linear discriminant analysis. Receiver operating characteristic (ROC) methodology was used to evaluate the classification accuracy. We investigated the dependence of classification accuracy on the input features, and on the pixel distance and bit depth in the construction of the SGLD matrices. It was found that five of the texture features were important for the classification. The dependence of classification accuracy on distance and bit depth was weak for distances greater than 12 pixels and bit depths greater than seven bits. By randomly and equally dividing the data set into two groups, the classifier was trained and tested on independent data sets. The classifier achieved an average area under the ROC curve, A_z, of 0.84 during training and 0.82 during testing. The results demonstrate the feasibility of using linear discriminant analysis in the texture feature space for classification of true and false detections of masses on mammograms in a computer-aided diagnosis scheme.

1. Introduction

Mammography is the most efficacious method for detection of early breast cancer. However, retrospective studies have shown that radiologists do not detect all breast cancers that are visible on the mammograms (Wolfe 1966, Martin *et al* 1979, Wallis *et al* 1991, Bird *et al* 1992, Harvey *et al* 1993). Double reading has been suggested to be an effective approach to improve the detection accuracy (Thurfjell *et al* 1994). Double reading is costly because it requires twice as much radiologists' reading time. This cost will be quite problematic considering the ongoing efforts to reduce costs of the health care system. Cost effectiveness is one of the major requirements for a mass screening program to be successful.

In our previous study, we demonstrated that a computer-aided diagnosis (CAD) scheme can serve as a second opinion for radiologists in the film interpretation process (Chan *et al* 1989). A well trained computer program, in effect, can partly assume the role of a second reader and assists radiologists in certain aspects of the detection and decision making processes. Although a computer program may never be able to achieve the level of knowledge and cognitive capability of a radiologist, a trained computer program can perform certain tasks reproducibly and consistently without the interobserver and intraobserver

0031-9155/95/050857+20$19.50

variations that are commonly observed among human observers. The ability of a CAD scheme can therefore be complementary to that of a radiologist. The combination of the two will probably result in improved accuracy in the interpretation of mammograms.

A number of research groups have been developing computer programs for analysis of mammographic abnormalities. For detection of mammographic masses, the methods reported to date mainly utilize morphological features to distinguish a mass from the normal mammographic background (Lai *et al* 1989, Brzakovic *et al* 1990, Yin *et al* 1991, Ng and Bischof 1992). Others have made use of texture or fractal analysis in classification of the four types of normal breast parenchyma according to Wolfe (Magnin *et al* 1986, Caldwell *et al* 1990), in an attempt to predict the risk levels of developing breast cancers. Recently, Kegelmeyer *et al* (1994) detected spiculated masses using local edge orientation and Laws texture features. This latter method, however, is not applicable for detection of non-spiculated masses.

In our previous study, we demonstrated the feasibility of using texture features to discriminate regions containing spiculated or non-spiculated masses from those containing normal breast tissue (Petrosian *et al* 1994). The texture features were derived from a spatial grey level dependence (SGLD) matrix, which characterized the spatial distribution of grey levels in the region of interest (ROI). We classified the ROIs by using a three-layer decision tree. The decision tree classifier was found to be effective in classifying ROIs containing mass and normal breast tissue. Using only three texture features, a true-positive rate of 89% at a false-positive rate of 24% was achieved during training. However, it was also found that training of the decision tree was quite inefficient. The training time increased rapidly when the number of layers and the number of training cases increased. Furthermore, the test results seemed to lag far behind the training results.

In the present study, we evaluated a new approach, which used linear discriminant analysis for classification of the ROIs based on the texture features. The linear discriminant analysis takes full advantage of the combinations of all available features, and the training process is relatively efficient. Using this classifier, we determined the relative importance of each input feature for the classification task based on statistical criteria. The dependence of the classification accuracy on the parameters of the SGLD matrix and on the combination of the texture features was analysed.

2. Materials and methods

2.1. Case samples

The 168 mammograms used in this study were randomly selected from the patient files in the Department of Radiology at the University of Michigan Hospitals by radiologists experienced in mammography. .The only criteria for inclusion were that the mammogram contained a biopsy-proven mass and that no grid lines were visible. The data set therefore included a mixture of benign ($n = 83$) and malignant ($n = 85$) masses. Forty-five of the malignant masses and six of the benign masses were judged as spiculated by the radiologists. The visibility of the masses on the mammograms was ranked by the radiologists on a scale of 1 to 10, with 1 being the most visible (obvious) and 10 the least visible (subtle) relative to the range of masses seen on mammograms. The range would be similar to the case mix encountered in clinical practice because the cases were randomly sampled from clinical films, although some biases would exist due to the small sample size. The size (length of the long axis) of the masses seen on the mammograms was also measured by the radiologists. The histograms of the visibility and the size of the masses are plotted in figure 1(a) and

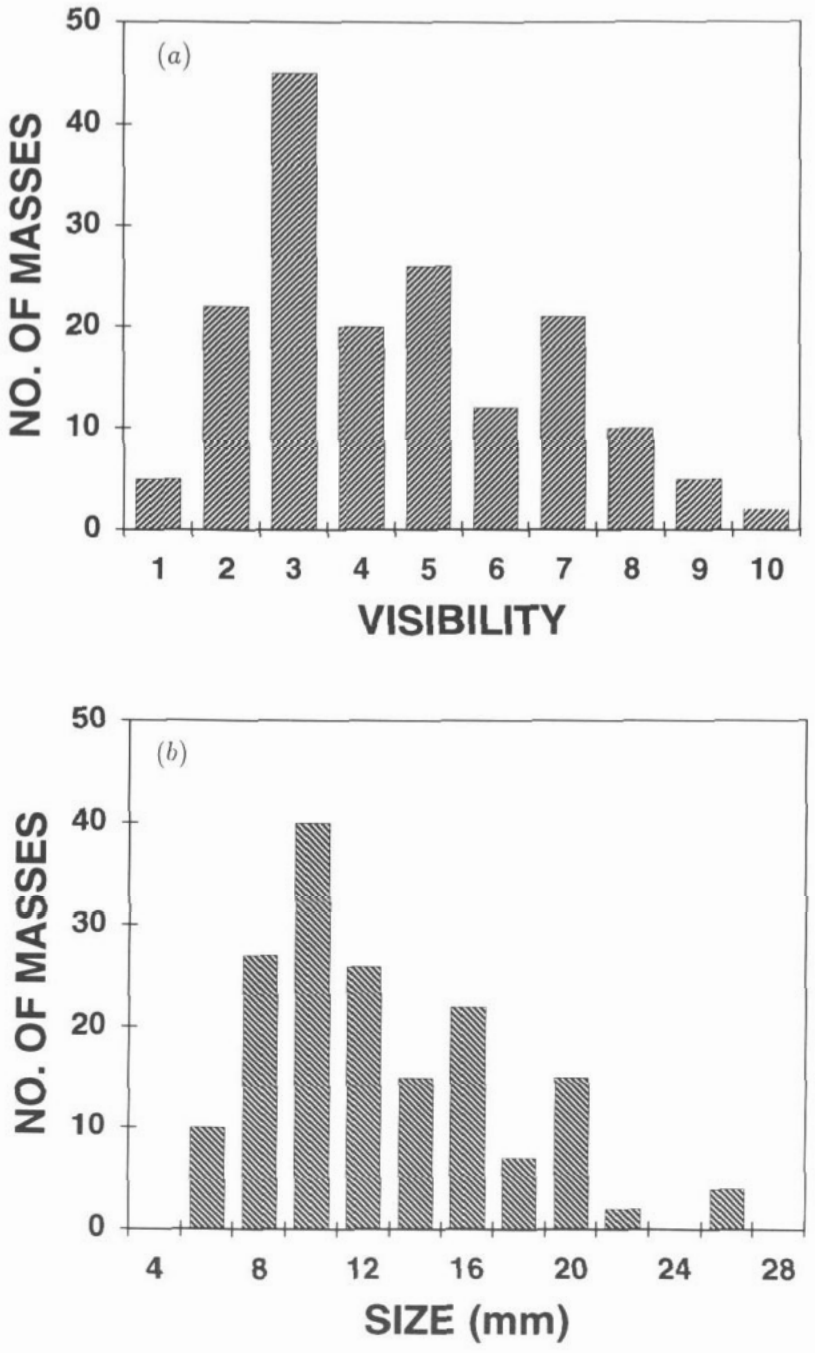
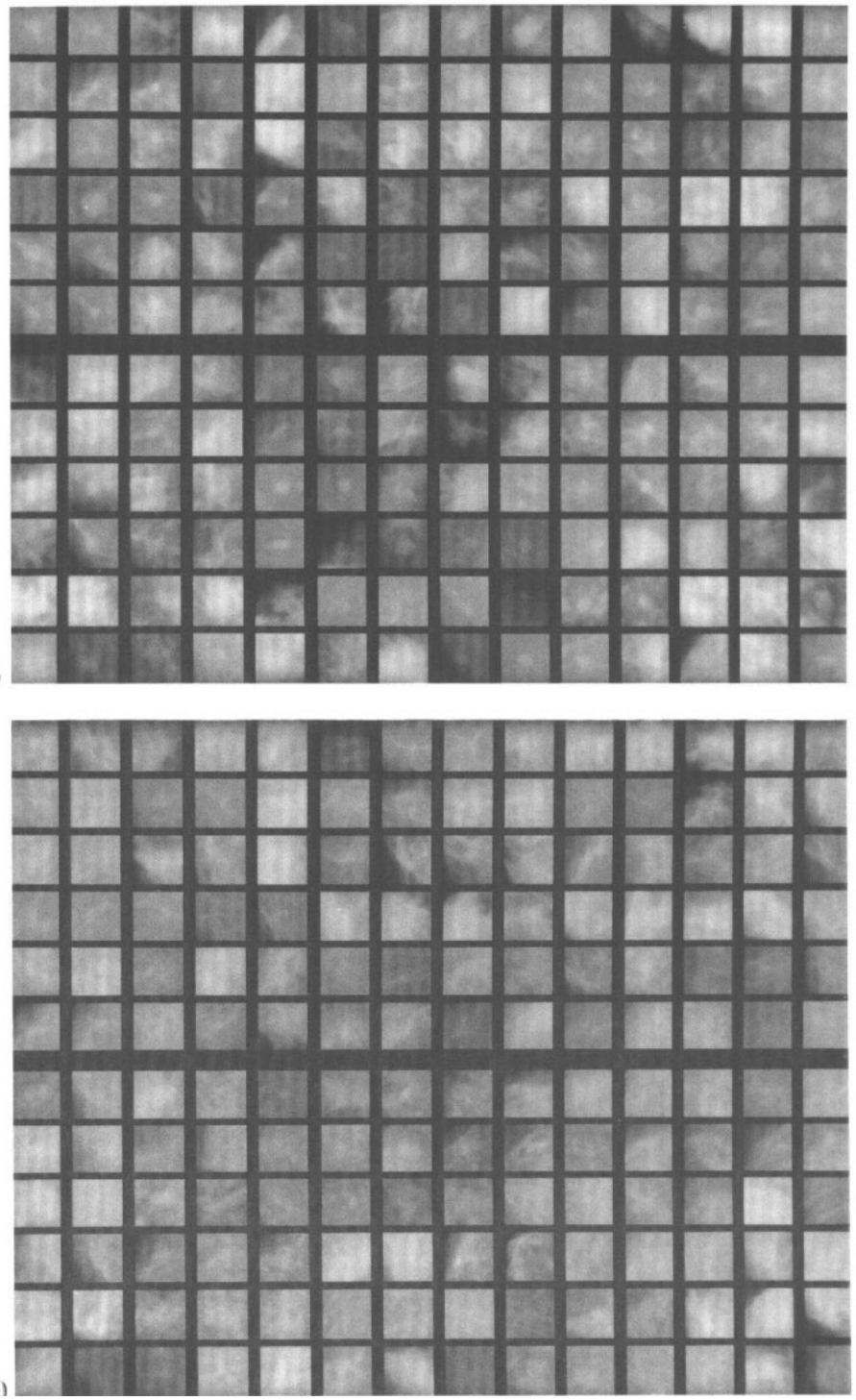


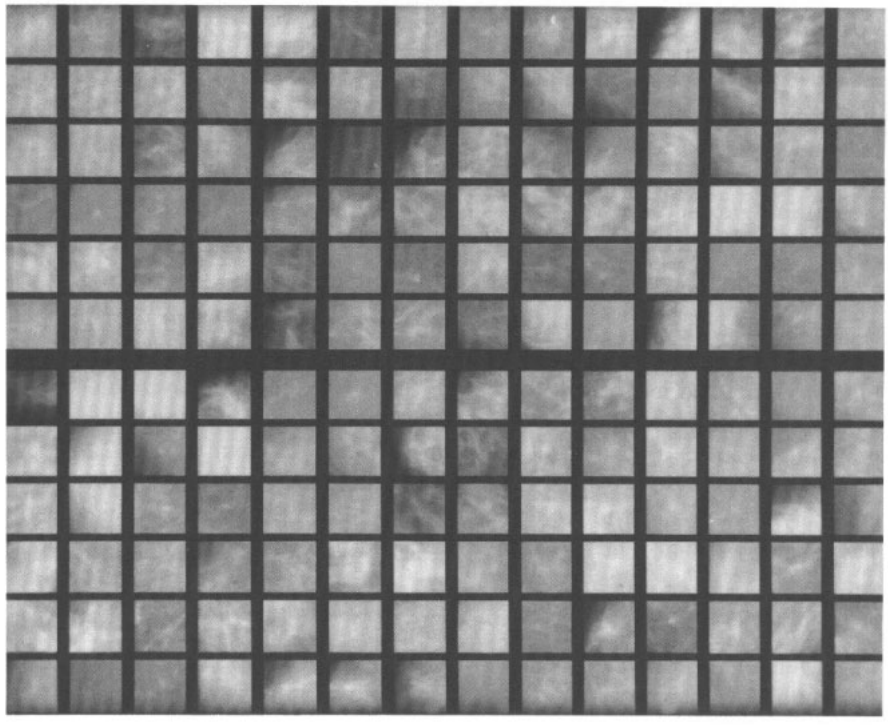
Figure 1. (a) The histogram of the subjective ranking of the visibility of the 168 masses in our data set on mammograms. The ranking ranges from very obvious (1) to very subtle (10) with a mean of 4.5. (b) The histogram of the size (length of the long axis) of the masses. The mean size of the masses is 12.2 mm.



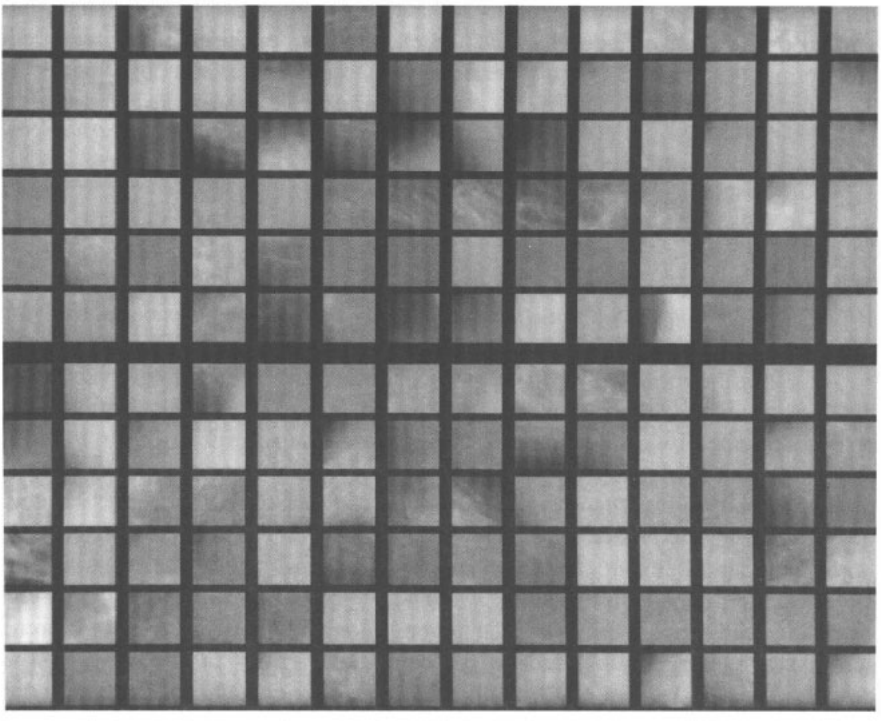
(*a*)

(*b*)

Figure 2. (a) The 168 ROIs containing biopsy-proven masses, (b) the 168 ROIs containing dense breast tissue, (c) the 168 ROIs containing mixed dense/fatty breast tissue, and (d) the 168 ROIs containing fatty breast tissue used in this study.



(*c*)



(*d*)

Figure 2. (Continued)

(b), respectively. All mammograms with grid lines were excluded because the repetitive grid pattern might interfere with the texture features extracted from the images. This will not be a limitation because only systems with reciprocating grids can be used in performing mammography under new regulations. The mammograms were acquired with dedicated mammographic systems with a 0.3 mm focal spot, molybdenum (Mo) anode and 0.03 mm Mo filter, and a 5:1 reciprocating grid. A Kodak Min R/MRE mammographic screen/film system using extended cycle processing was employed as the image receptor.

2.2. *Digitization*

All mammograms were digitized with a laser film scanner (LUMISYS DIS-1000) at a pixel size of 0.1 mm × 0.1 mm and 4096 grey levels (12 bits). The light transmitted through the film was amplified logarithmically before analogue-to-digital conversion. The digitizer had an optical density (OD) range of 0–3.5. It was calibrated so that the OD on film was linearly proportional to the output pixel value in the range of about 0.1 OD–2.8 OD with a slope of 0.001 OD/pixel value. The slope of the calibration curve outside this range decreased gradually. Before input to the detection program, the pixel values were linearly converted such that low optical densities were represented by high pixel values.

2.3. *ROI selection*

On each mammogram, the location of the mass was identified by an experienced radiologist and verified with biopsy reports. Three additional regions of normal breast parenchyma were also chosen by the radiologist from the same mammogram: a region containing the densest tissue, a region of mixed tissue, and a region of fatty tissue. These three types of normal tissue were included because the classifier is developed for differentiation of mass and normal tissue for all breast types. Each of these ROIs was composed of 256 × 256 pixels. The four sets of ROIs, each of 168 samples, are displayed at a reduced spatial resolution in figure 2 to illustrate the masses and normal tissue included. Each ROI corresponds to a 2.56 cm × 2.56 cm area on a mammogram.

2.4. *Texture features*

The texture features used in this study were derived from the SGLD matrix (Haralick *et al* 1973), also known as the concurrence matrix or the co-occurrence matrix, of the ROI as discussed previously (Petrosian *et al* 1994, Cheng *et al* 1994). The SGLD matrix element, $p_{\theta,d}(i, j)$, is the joint probability density of the occurrence of grey levels i and j for two pixels with a defined spatial relationship on an image. The spatial relationship of the pixel pair is described by a selection rule that specifies the relative direction θ and the distance d between the two pixels. Because of the discrete nature of the digital image, the choice of θ is actually limited to 0°, 45°, 90°, and 135° and the distance d is limited to integral multiples of the pixel size.

A number of texture features can be derived from the SGLD matrix (Haralick *et al* 1973, Conners 1979). In this study, we evaluated the discriminant ability of eight features: correlation, entropy, energy (angular second moment), inertia, inverse difference moment, sum average, sum entropy, and difference entropy (Petrosian *et al* 1994). These features describe the shape of the SGLD matrix and generally contain information about the image characteristics, such as homogeneity, contrast, and the presence of organized structures, as well as the complexity and grey level transitions within the image. However, a particular feature cannot be related uniquely to a specific image characteristic (Haralick *et al* 1973).

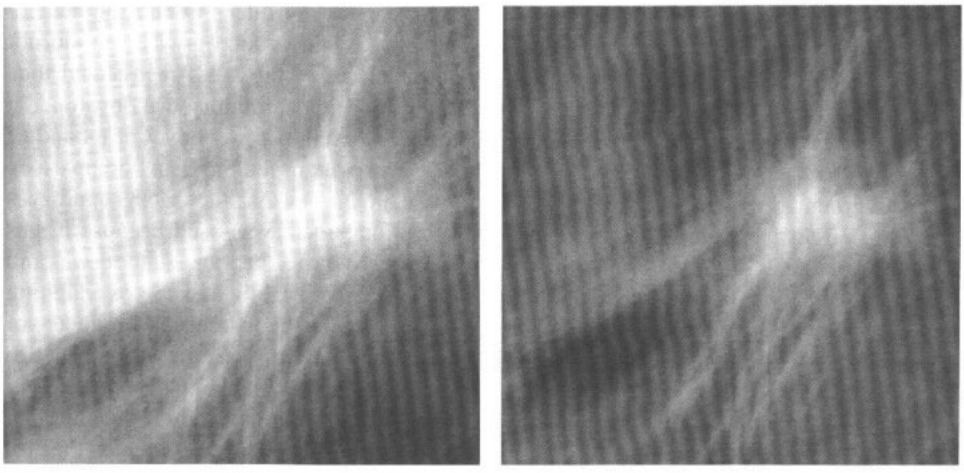
The background grey levels of each ROI depend on the x-ray intensity and the density of the neighbouring and overlapping tissue. In general, the background grey levels do not relate directly to the presence of a mass, but affect the shape of the SGLD matrix and thus the value of some of the texture features. In order to reduce this variability, we developed a technique that estimated the low-frequency background by using the grey level of a band of pixels around the perimeter of the ROI. The method was chosen based on visual comparison of the background-corrected images with the original image so that the background was levelled and no artifact could be seen. This subjective judgement was used because it was impossible to quantitatively determine how good a background correction method was; no 'true background' could be found due to the complexity and variability of the overlapping structures. We first calculated a running average of the pixel values along the perimeter of the ROI using a box filter of a 32×16 kernel, of which the long dimension was parallel to the edge of the ROI. For the perimeter pixels that were within 16 pixels of one of the four corners of the ROI, the long dimension of the box filter kernel was reduced on the side that was limited by the ROI edge. For example, the average pixel value at a corner of the ROI was obtained by a 16×16 box filter, with one apex of the box filter kernel coinciding with the corner pixel. The grey level, $G(i, j)$, of a given pixel (i, j) in the estimated background image of the ROI was then calculated as

$$G(i, j) = \left[\sum_{k=1}^{4} g_k/d_k\right] \Big/ \left[\sum_{k=1}^{4} 1/d_k\right] \tag{1}$$

where g_k is the grey level of the pixel at the intersection between one side of the low-pass-filtered ROI perimeter and the normal from the pixel (i, j) to that side, and d_k is the distance from the pixel (i, j) to the intersection; k ranges from one to four, denoting the four sides of the ROI. More than four pixels around the ROI perimeter might be used in the weighted sum for the background estimation and a low-pass filter might be applied to the interpolated background image to provide further smoothing. We used (1) for the interpolation and applied a low-pass box filter of a 32×32 kernel to the interpolated background image in this study. The background image was then subtracted from the original ROI, thus reducing the background to near zero. The texture features were calculated from the background-corrected ROIs. An example of an original ROI with a malignant mass and the background-corrected ROI is shown in figure 3(a) and (b), respectively. It can be seen that the background of the ROI was flattened while the high-frequency information in the ROI was basically unchanged because the background image only contained low frequencies.

As discussed above, the SGLD matrix depends on the spatial relationship of the pixel pairs. Because it is expected that the texture of a mass may be isotropic whereas the texture of normal breast structures may have a slightly stronger directional dependence as they diverge from the nipple, we first investigated the dependence of the classification accuracy on directional information. For a fixed pixel distance d, we calculated each of the eight texture features from the SGLD matrices at $\theta = 0°$, 45°, 90°, and 135°, resulting in a 32-dimensional feature space. An eight-dimensional feature space was derived by averaging the corresponding features in the four directions, thereby neglecting the directional information of the texture features. The classification accuracy in these two feature spaces was compared. As discussed in section 3 below, the comparison indicated that there was no significant difference between the two approaches. Therefore, the eight features averaged over the four directions were used as the input features to the classifier in the rest of our studies.

The SGLD matrix depends not only on the direction and the distance between the pixel pairs, but also on the bin width (i.e. the grey level interval) used in determining the two-dimensional histogram, which is an estimate of the joint probability density distribution.



(*a*) (*b*)

Figure 3. An example demonstrating the effect of background correction: (a) the original ROI with a malignant mass and (b) the background-corrected ROI. A constant value was added to each pixel of the ROI in (b) to match the mean level of the ROI in (a) for visual comparison purposes. The displayed window width is the same for both images. The sloped background in (a) becomes relatively flat in (b) after background correction.

For a 12-bit image, a 4096×4096 SGLD matrix is obtained with the minimum bin width of one grey level. A matrix size of 256×256 is obtained with a bin width of 16 grey levels, which is equivalent to reducing the grey level resolution (i.e. the bit depth) of the image by eliminating the four least significant bits and using a bin width of one grey level. There is a trade-off between the grey level resolution and the statistics of the estimated distribution. When the bin width is small, the number of counts of pixel pairs in each bin will be small and the statistics of the estimated joint probability density distribution will be poor. The noise in the least significant bits of the image will also affect the distribution. On the other hand, when the bin width is large, the statistics in each bin will improve and the effect of image noise will decrease. However, some characteristic features of the distribution may be lost. In either case, the discriminant power of the texture features derived from the SGLD matrix may be degraded. We therefore investigated the dependence of the classification accuracy of the texture features on grey level resolution in order to determine the best bin width to construct the SGLD matrix. The grey level resolution of the original images was reduced from 12 bit to lower bit depths by eliminating the least significant bits. The texture features of the images of different bit depths were calculated in the same way as described above.

2.5. Linear discriminant analysis

Linear discriminant analysis is a well established statistical technique (Lachenbruch 1975, Tatsuoka 1988). For a two-class problem, one canonical discriminant function can be constructed for classification of the two groups of cases. The discriminant function is formulated by a linear combination of the feature variables:

$$D = a_0 + \sum_{i=1} a_i X_i \tag{2}$$

where n is the number of feature variables, the X_i are the values of the feature variables and the a_i are coefficients (or weights) estimated from the input data during training so that the separation between the distributions of the discriminant scores, D, of the two groups is a maximum. This is accomplished by maximization of the ratio of the between-groups sum of squares to the within-groups sum of squares for the two distributions of the discriminant scores. Geometrically, the linear discriminant function can be considered as a projection of the feature vectors onto an axis in the multidimensional feature space. The component of the feature vector of a given case along this axis corresponds to the discriminant score of that case. An example of the probability density distributions of the discriminant scores for the ROIs with masses and normal tissue used in this study is shown in figure 4. The discriminant scores were obtained with five input features calculated from the SGLD matrix at a distance of 20 pixels and eight bits as discussed below. The goal of the linear discriminant analysis is to find the axis (i.e. the principal axis) that provides the maximum separation between the distributions of the discriminant scores for the two groups. If the population of the feature vectors for each group in the feature space follows a multivariate normal distribution and the population covariance matrices for the two groups are equal, the linear discriminant function provides an optimal classification rule to minimize the probability of misclassification.

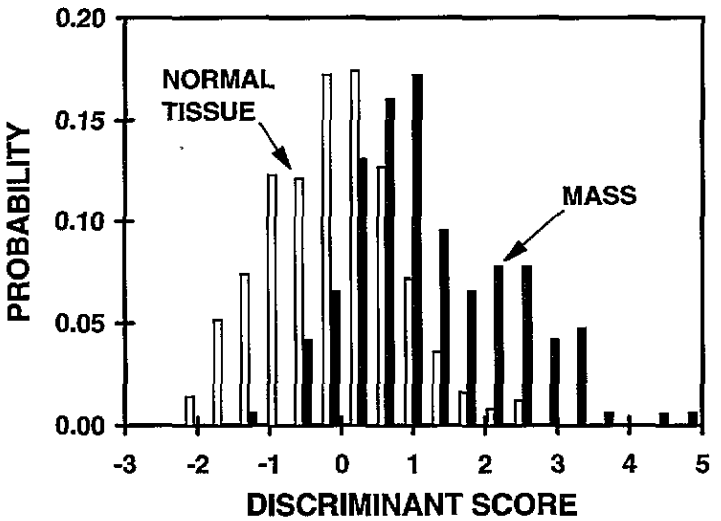


Figure 4. An example of the probability density distributions of the discriminant scores of the normal tissue (white) and masses (black). The discriminant scores were calculated from the discriminant function, which provides a linear combination of five texture features derived from the SGLD matrices constructed at eight bits and a distance of 20 pixels to maximize the separation between the two groups.

The linear discriminant analysis can be performed in a two-stage process (SPSS 1993). First, a stepwise procedure is performed to identify from all available input features the useful feature variables for the formulation of the discriminant function. Second, the selected features of the input cases are used to determine the coefficients of each feature variable in the dscriminant function to achieve maximum separation. In the SPSS implementation, several statistical criteria can be used to choose good feature variables in the stepwise

procedure. These criteria include the maximization of a generalized measure of the distance between two groups (Mahalanobis distance), the minimization of the ratio of the within-group sum of squares to the total sum of squares of the distributions (Wilks' lambda), the Lawley–Hotelling trace (Rao's V), the maximization of the between-groups F statistic value, and the minimization of the sum of unexplained variance. We studied the effect of using the different criteria in the stepwise procedure on feature selection for a feature set from fixed SGLD matrix parameters and found that the same feature variables were selected for our two-class problem when the different criteria were used. The Wilks' lambda criterion was then used for variable selection under the different bit depth and distance conditions in this study. The stepwise procedure using the Wilks' lambda criterion is described briefly as follows. A detailed discussion of the underlying statistical theories is given in the literature (Lachenbruch 1975, Tatsuoka 1988, SPSS 1993).

At the stepwise feature selection stage, the program enters one feature or removes features in alternate steps by analysing their effect on the separation between the two groups based on the Wilks' lambda criterion. The significance of the change in Wilks' lambda when a variable is entered or removed from the model is based on F statistics. Initially, the program calculates the Wilks' lambda values between the two groups when each of the feature variables is used individually. The variable that provides the smallest Wilks' lambda is entered into the model first. In a subsequent feature entry step, each of the variables not yet in the model is assumed to be entered one at a time. The Wilks' lambda and the F value for the change in Wilks' lambda (F-to-enter) are evaluated. The variable that provides the smallest Wilks' lambda will be entered next in the model if the F-to-enter value is larger than the F-to-enter threshold. In the feature removal step, each of the variables already included in the model is assumed to be removed one at a time. The Wilks' lambda and the F value for the change in Wilks' lambda (F-to-remove) are evaluated. If the F-to-remove value is smaller than the F-to-remove threshold, the variable will be actually removed from the model. The stepwise procedure continues until the F-to-enter values for all variables not in the model are smaller than the F-to-enter threshold and the F-to-remove values for all variables in the model are greater than the F-to-remove threshold. At this point, no more variables meet the entry or removal criteria and the procedure terminates.

To evaluate the relative importance of the feature variables for our discrimination task, we first made use of the stepwise procedure to select the important features for each bit depth and each distance of the SGLD matrices. All eight features were entered into the stepwise selection procedure under each condition. We then examined the features selected for all conditions, and ranked the relative importance of the features based on the frequency with which the features were chosen. To evaluate the effect of the combination of the feature variables on the discriminant analysis, linear discriminant functions with one to eight input features were formulated by adding one feature at a time, in the order of importance as found by the stepwise selection procedure. For each feature combination, the dependence of the classification accuracy on bit depth and distance was studied by receiver operating characteristic (ROC) analysis as described below. With this approach, the combinations of features that were effective for the separation of the two classes were systematically evaluated. Although we did not exhaustively study all feature combinations, our approach should find a near-optimal combination under the conditions studied.

After selection of features and optimization of the SGLD matrix parameters, we studied the training and testing of the discriminant classifier. The 168 mass ROIs were randomly and equally divided into two subsets, referred to as group 1 (G1) and group 2 (G2). The three normal ROIs obtained from the same mammogram as the mass ROI were grouped into the same subset as the mass ROI in order to ensure the independence of the training and

test sets. When group 1 was used as training set, group 2 was used as the test set and vice versa. The discriminant classifier was trained with the selected combination of feature variables. The weight of each feature variable in the discriminant function for a given condition was optimized, as described above, by using the feature values in the training set. During testing, the feature values of each case in the test set were entered into the trained discriminant function to calculate the discriminant score for that case.

The accuracy of the discriminant classifier was evaluated by ROC methodology (Swets and Pickett 1982, Metz 1986). The discriminant score was used as the decision variable in the ROC analysis. An ROC curve, which is the relationship between the true positive fraction (TPF) and false positive fraction (FPF), can be generated by setting different decision thresholds. For the distributions shown in figure 4, a decision threshold at a large discriminant score corresponds to a stringent criterion with a small TPF and small FPF; a decision threshold at a small discriminant score corresponds to a lax criterion with a high TPF and high FPF. The LABROC1 program (Metz *et al* 1990), which assumes binormal distributions of the decision variable for the normal and abnormal cases and fits an ROC curve based on maximum-likelihood estimation, was used to estimate the area under the ROC curve, A_z, and the standard deviation (SD) of A_z. We used A_z as an index of classification accuracy. The best combination of features, bit depth, and distance for the classification task was determined by maximization of A_z. To test the statistical significance of the difference in A_z for two conditions, the CLABROC program for correlated data (Metz *et al* 1984) was used.

3. Results

We compared the classification accuracy of using the texture features in four directions separately to that of using the texture features averaged over the four directions. The comparison was performed for features calculated from the SGLD matrix constructed at a bit depth of eight and a pixel distance of 20. These conditions were chosen because they provided near-maximum classification, as discussed below, for the texture features and data set used in this study. For both sets of input features, the stepwise discriminant procedure was used to select the features. We used the entire data set of 672 ROIs as input cases so that the statistical properties of the feature variables could be more reliably determined. For the set of 32 input features, the discriminant analysis selected six features: correlation (45°), correlation (135°), difference entropy (45°), difference entropy (135°), entropy (135°), and inertia (90°). The A_z was 0.842 ± 0.017. For the set of eight input features averaged over the four directions, the discriminant analysis selected four features: correlation, difference entropy, entropy, and inertia. The A_z was 0.834 ± 0.018. Although the directional input features seemed to provide slightly better discriminant power, the difference was less than 0.5 SD. The two-tailed p level obtained from the correlated 'area test' (Metz *et al* 1984) for the difference in the A_z values was greater than 0.05 and did not reach statistical significance. We therefore used the texture features averaged over four directions in the following studies.

For the determination of the relative importance of the features for the classification task, we again used the entire data set of 672 ROIs as input cases in the stepwise feature selection precedure. The threshold values for inclusion or exclusion in the stepwise procedure were kept at the default values in the SPSS program for all conditions (F-to-enter threshold, 3.84; F-to-remove threshold, 2.71). The features selected for grey level resolution of four bits to nine bits and pixel pairs separated by a distance of two to 40 pixels are listed in table 1. The numbers in the table entries indicate the number of bits used in construction of

the SGLD matrices. The correlation feature appeared to be the most discriminative because it was selected for all bit depths and all distances. The difference entropy was the second most important feature. Entropy and inertia were almost equally important. The order of importance of the next three features, energy, inverse difference moment, and sum entropy, was less obvious. Energy was important when the distance was small, sum entropy was more important in the mid-distance range, while inverse difference moment spanned over all distances for relatively low bit depths. The effect of sum average was negligible under almost all conditions. For the following studies, we ordered the features as (i) correlation, (ii) difference entropy, (iii) entropy, (iv) inertia, (v) inverse difference moment, (vi) sum entropy, (vii) energy, and (viii) sum average.

We studied the dependence of classification accuracy on the bit depth and pixel distance parameters of the SGLD matrix when the feature combination was fixed. The entire data set was again used to achieve the best statistical certainty and the study was performed as a training procedure. The dependence of A_z on pixel distance for bit depths from four to nine bits using five feature variables is illustrated in figure 5. The classification accuracy increased with increasing distances initially, reached a broad maximum, and decreased slightly or levelled off at large distances. For distances above 12 pixels, A_z increased as the bit depth increased from four bits to seven or eight bits and fell off slightly at nine bits. The SDs of A_z ranged from 0.017 to 0.018 at distances from 12–40 pixels and six to nine bits; the SDs could be as large as 0.021 at smaller distances or four and five bits. The differences in A_z for distances of 12–40 pixels and six to nine bits were less than one SD. We performed statistical significance tests between some selected pairs of conditions. At a bit depth of eight, the two-tailed p levels were greater than 0.05 for the differences between pairs of A_z at pixel distances larger than 12. The two-tailed p level for the difference between six and eight bits at a pixel distance of 20 was 0.2. These differences therefore were not statistically significant. The difference between five and eight bits at a pixel distance of 20 was statistically significant at a two-tailed p level of 0.04.

The dependence of classification accuracy on feature combination for various bit depth and distance parameters of the SGLD matrix was studied. The features were combined based on their order of importance as found in the stepwise feature selection procedure discussed above. The number of feature variables in the discriminant function was varied from one to eight. The entire data set of 672 ROIs was used as input cases. The effect of the different input features on classification accuracy is plotted in figure 6 for the various bit depths and a fixed distance of 20 pixels used in the SGLD matrices. The value on the horizontal axis indicates the feature combinations. For example, a number of features of four means that the first four features as ordered previously were used to formulate the discriminant function. The results are discrete but the data points are linked with lines to facilitate reading. The accuracy is low when only one or two features are used for classification. It increases significantly when the third feature is added. For five to seven bits, A_z is almost constant for four to eight input features. For eight to nine bits, there was a broad maximum at five or six features. The SDs of the A_z for these conditions were about 0.018. For a fixed grey level resolution of eight bits, the two-tailed p levels of the differences in pairs of A_z between three and eight features were all greater than 0.05. Because the SDs and the differences in A_z between the other conditions from six to nine bits and from three to eight features were of a similar magnitude, their p levels were expected to be in the same range. The difference in A_z between two and three features at eight bits was statistically significant at a two-tailed p level of 0.002.

The dependence of the training and test results on variations in the input data sets is demonstrated in figure 7. The A_z values obtained with five input features and eight bits were

Table 1. Texture features selected by stepwise linear discriminant analysis. Eight features were derived from the SGLD matrices constructed with a given pixel distance and a given bit depth. The numbers in an entry of a given feature and a given distance indicate the bit depths for which the feature was selected.

Distance (pixels)	Correlation	Difference entropy	Entropy	Inertia	Inverse difference moment	Sum entropy	Energy	Sum average
2	4 5 6 7 8 9						5 6 7	
4	4 5 6 7 8 9					4	5 6 7 8 9	
8	4 5 6 7 8 9	4 5 6 7 8 9	7 9	5 6 8	4 6 8	4 5 6 7 8 9	5 6 8	
12	4 5 6 7 8 9	5 6 7 8 9	5 6 7 8 9	7 8 9	4	4 7 8 9	5 7 8 9	
16	4 5 6 7 8 9	5 6 7 8 9	6 7 8 9	6 7 9	4	4 5	7 9	
20	4 5 6 7 8 9	5 6 7 8 9	6 7 8 9	5 6 7 8 9	4	4 5		
24	4 5 6 7 8 9	5 6 7 8 9	6 7 8 9	5 6 7 8 9	4	4 5		
28	4 5 6 7 8 9	5	5	5				4
32	4 5 6 7 8 9							
36	4 5 6 7 8 9	5 6 7 8 9	5 6 7 8 9	5 6 7 8 9	5 6			
40	4 5 6 7 8 9	4 5 6 7 8 9	5 6 7 8 9	5 6 7 8 9	5 6			

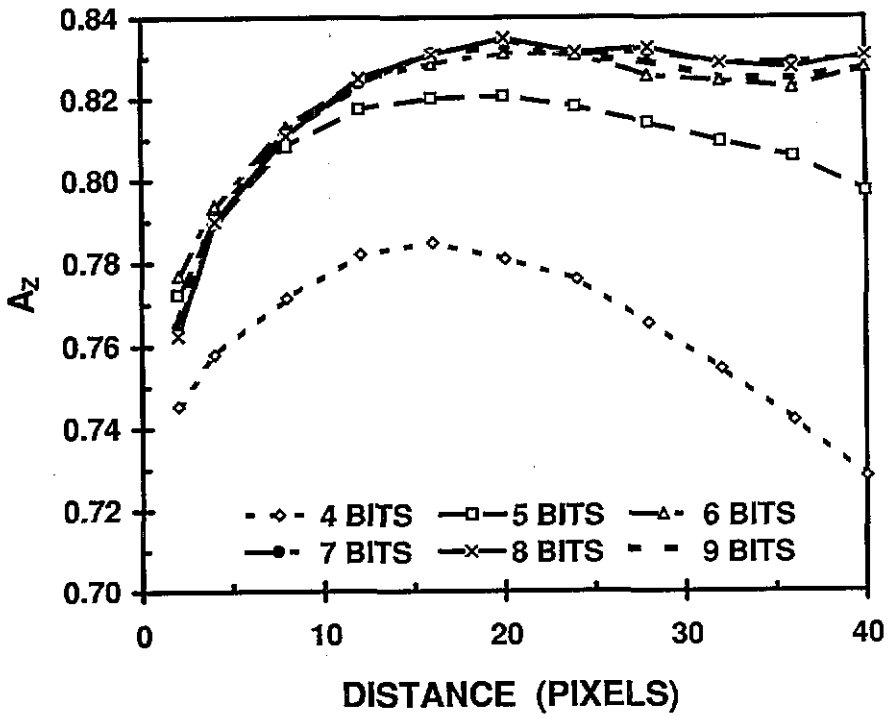


Figure 5. Dependence of classification accuracy, quantified in terms of the area under the ROC curve, A_z, on distance for various bit depths used in the SGLD matrices. The results for five input features were plotted. The two curves for seven bits and eight bits almost overlap. The standard deviations of A_z ranged from 0.017 to 0.018.

plotted as a function of pixel distance of the SGLD matrix. The use of group 1 or group 2 as training set resulted in a difference of about 0.001–0.025 in A_z of training, and a difference of about 0.001–0.023 in A_z of testing except for a difference of 0.051 at $d = 4$. Because of the smaller data sets used to obtain these results, the SDs in A_z ranged from 0.024 to 0.029. The maximum average A_z was 0.840 for training and 0.823 for testing, occurring at a distance of 20 pixels. The average difference between the training A_z and test A_z was 0.015 with a maximum difference of 0.036 at $d = 28$. The ROC curves corresponding to the A_z values at a distance of 20 pixels are plotted in figure 8. The A_z values for the two training curves are 0.831 and 0.850, and the A_z values for the two test curves were 0.829 and 0.817, respectively.

To further estimate the variation in A_z caused by differences in the input data sets, we equally divided the 168 mass ROIs again with a different sequence of random numbers, together with the normal ROIs from the same image, to form group 1′ and group 2′. The analysis described above was repeated using these two groups for training and testing alternately. The differences in A_z at various pixel distances varied from 0.007 to 0.021 for training and from 0.001 to 0.030 for testing, except for a difference of 0.045 at $d = 4$. The SDs in A_z ranged from 0.024 to 0.030. The maximum average A_z for training was 0.839, occurring at a distance of 20 pixels. The corresponding average A_z for testing was 0.810. The average difference between the training A_z and test A_z was 0.027 with a maximum difference of 0.061 at $d = 12$.

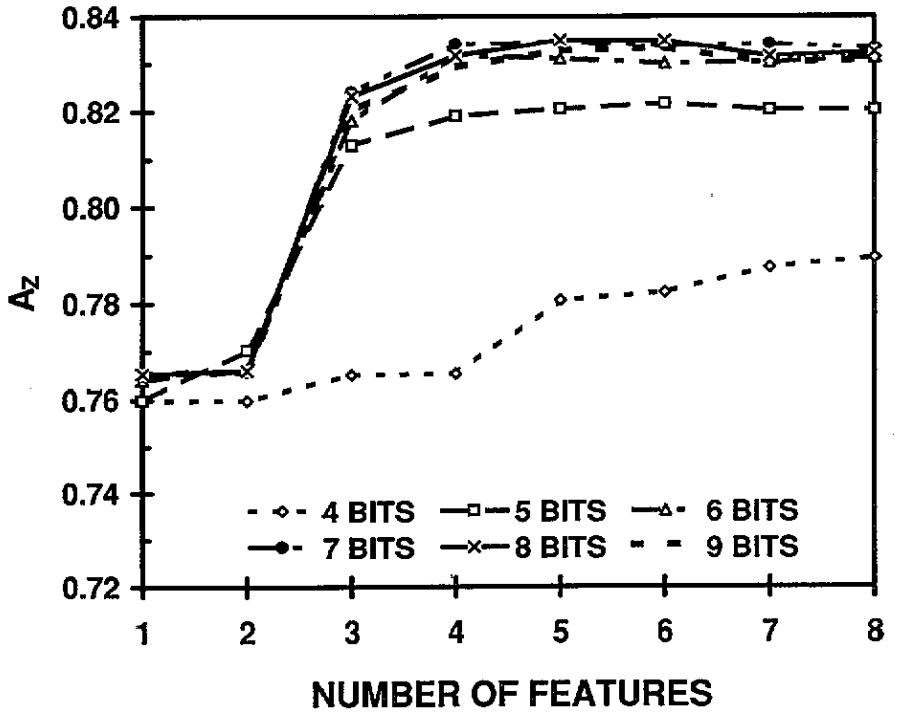


Figure 6. The dependence of classification accuracy on the input features to the linear discriminant classifier. The results for a distance of 20 pixels and four to nine bits are plotted. The order of features being added to the input was (i) correlation, (ii) difference entropy, (iii) entropy, (iv) inertia, (v) inverse difference moment, (vi) sum entropy, (vii) energy, and (viii) sum average.

Since the results from the different grouping were not independent of the first grouping, we did not estimate the variation in A_z from the four training or test groups. However, this study using the four combinations of training and test sets indicates that the variations in the training A_z or test A_z caused by different input data sets were within 1 SD under most of the conditions studied. The differences between the training A_z and the test A_z for each combination ranged from 1 to 2 SD. Similar differences were observed for other input features and bit depths. The linear discriminant classifier therefore appears to be consistent among different training and test data sets and also between training and testing.

4. Discussion

In our previous study, a decision tree was trained to classify masses and normal breast tissue based on texture features extracted from the SGLD matrix (Petrosian *et al* 1994). The data set consisted of 45 mass ROIs and 135 normal ROIs. It was found that the decision tree could provide high accuracy for a training set, a TPF of 89% at an FPF of 24%. However, the test results lagged substantially behind the training results. The trained decision tree provided a TPF of 76% at an FPF of 36% using a leave-one-out test scheme. One possible cause of the discrepancy was that the statistics at each branch of the decision tree in the higher layers were poor. The number of cases reaching each branch was small because the number of

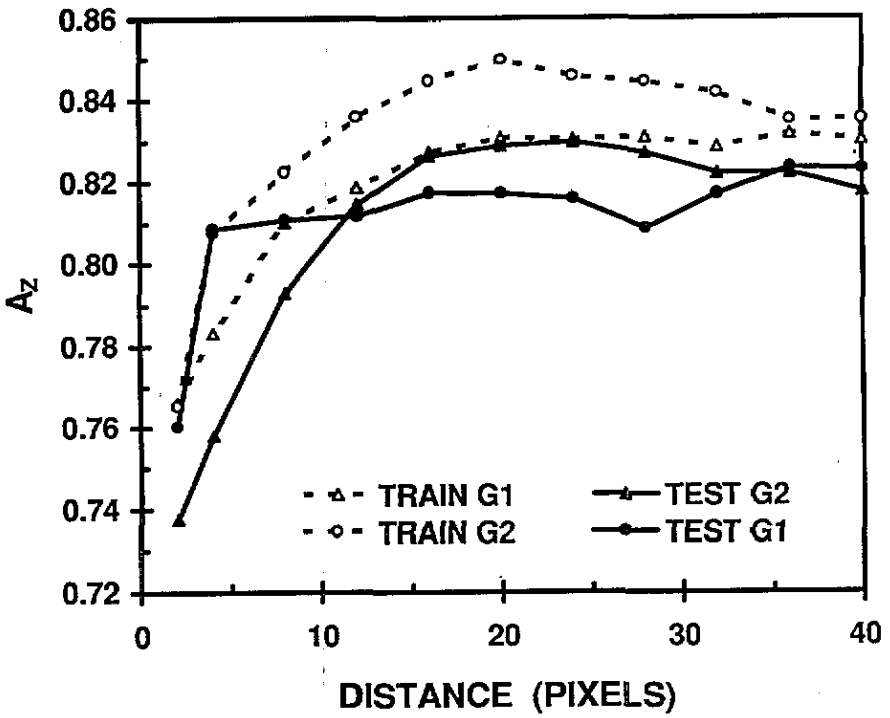


Figure 7. The dependence of classification accuracy on input data sets. The 168 mass ROIs were randomly and equally divided into two groups together with the normal ROIs extracted from the same mammogram. One group (G1) was used as the training set and the other (G2) as the test set and vice versa. The five input texture features were calculated from the SGLD matrices constructed at eight bits.

tree branches increased rapidly with the number of layers. The decision threshold at each branch could not be optimally determined unless the number of input training cases was extremely large so that the population of cases at each branch of the tree was sufficiently representative.

With the linear discriminant classifier for a two-class problem, one single decision threshold was determined based on the probability distributions of linearly combined features of all available training cases. The statistical properties of the test cases were therefore more accurately predicted. The A_z of the test set was generally within a few per cent of the A_z of the training set. This consistency indicates that a linear discriminant classifier could be trained more reliably than a decision tree. One possible limitation of the linear discriminant classifier, however, was that only linear combinations of the features were utilized. The performance of a linear discriminant classifier should be compared with that of a non-linear discriminant classifier or an artificial neural network to determine the effectiveness of non-linear feature combinations in future studies.

Although there is a general trend that the normal tissue pattern on a mammogram diverges from the nipple, how strong this pattern appears depends somewhat on the breast types, i.e. fibroglandular, fatty, or mixed. There is also local branching, as can be seen from a ductogram, and other superimposed structured background, which increases the complexity of the tissue pattern. The general direction of the tissue, if it can be seen at all in an ROI, depends on the location in the breast from which the ROI is extracted. Because our ROIs

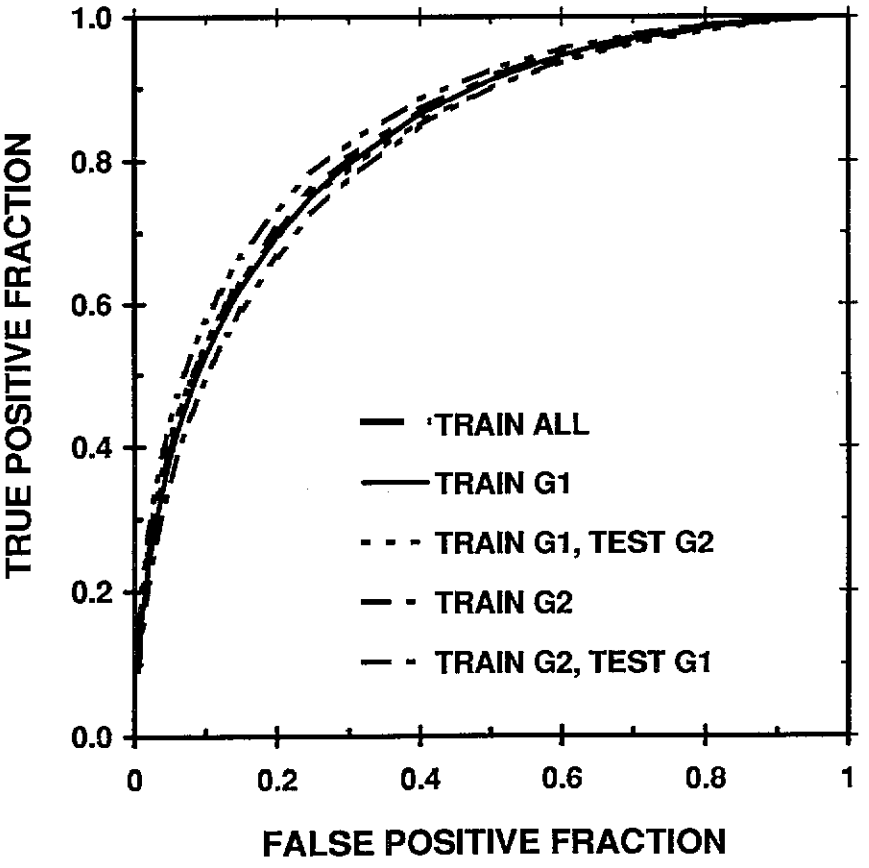


Figure 8. The ROC curves obtained from five texture features calculated at eight bits and a distance of 20 pixels. The A_z values for the two training curves were 0.831 and 0.850, and the A_z values for the two test curves were 0.829 and 0.817, respectively.

were selected based on the tissue of interest instead of the location in the breast or direction of the tissue, the directional information in all normal ROIs in the data set is statistically random as a whole. This is consistent with our observation that the discriminant power of the texture features without being averaged over the four directions is statistically similar to that of the average texture features. When this classifier is used to reduce FP detections, it will have to be applicable to any locations in the breast and breasts of all tissue types. We therefore believe that the directional information of the normal tissue would not be very effective in our classification task and did not pursue this further in this study.

The results of our study indicate that the classification accuracy reached a broad maximum when the input features were obtained from SGLD matrices at about seven or eight bits. In our previous study (Chan *et al* 1994) of the dependence of detection accuracy of microcalcifications on grey level resolution, we found that a nine-bit resolution was required to provide optimal detection, indicating that, for the screen/film system and digitizers used in our study, the three least significant bits contained mainly noise. For the classification of masses with texture features, an additional factor that affected the accuracy was the trade-off between grey level resolution and the statistical uncertainty of the SGLD matrices.

This factor might have further reduced the bit depth at which the classification accuracy of masses was maximized as observed.

The texture features used in this study appear to be promising in distinguishing ROIs containing masses from those containing normal breast parenchyma. Although the classification accuracy at an A_z of 0.823 may not be adequate to be used as a primary tool for detection of masses on mammograms, the classifier at its present stage may be used in combination with a computer-aided detection scheme, which can screen a mammogram for suspicious regions with an automated algorithm, to serve as one of the steps in reducing FP ROIs. For this application, the operating point along the ROC curve may be chosen to be relatively lax, for example, at a TPF of 95% and an FPF of about 60%, so that most of the mass regions are kept while 40% of the normal ROIs are excluded. The selection of an operating point along the ROC curve depends on the application and on cost benefit considerations. The best decision threshold should be determined based upon the specific task to which the classifier is applied.

The results of this study establish the feasibility of using a linear discriminant classifier with texture features for classification of masses and normal breast tissue. Further studies are being conducted in an effort to improve its accuracy. Preprocessing methods are being evaluated for enhancement of the image before the texture features are extracted. Multiresolution analysis using wavelet transform is being investigated as a method that can potentially condense information at each resolution level and make use of any differences in the dependence of the texture features on pixel distance between regions containing masses and normal tissue, as observed in the present study. Additional texture features (Haralick 1986) and morphological features are also being analysed in order to increase the discriminant ability of the classifier. Linearization of the sensitometric response of the screen/film system from optical density to relative exposure before extraction of the texture features will further reduce their variability due to the limited dynamic range of the film. However, it may be difficult to implement the sensitometric conversion in practice because it is difficult to measure the x-ray sensitometric curves routinely at present.

The effectiveness of a classifier for FP reduction in a CAD algorithm is expected to depend on the specific type of FP generated by the detection process, which may be different for different automated detection schemes or human observers. Furthermore, the values of the weights in the discriminant function will depend on the values of the input texture features which may, in turn, depend on the properties of the image acquisition system used. To implement our proposed texture feature classifier in a specific application, the classifier should be trained based on the mass and FP populations obtained from representative case samples in that application, using the optimization procedures proposed in this study. Ideally, the number of training samples should be much larger than that used in this feasibility study in order to provide adequate training for the classifier. The performance of the trained classifier should also be tested for its generalization capability.

5. Conclusion

We studied the effectiveness of using texture features derived from the SGLD matrix for classification of masses and normal breast parenchyma on digital mammograms. With five input features calculated at a distance of 20 pixels and a grey level resolution of eight bits, a linear discriminant classifier provided an average A_z of 0.823 for testing. The feature classifier may be incorporated into a CAD scheme, which automatically scans a digital mammogram for suspicious regions, as one of the steps to differentiate TP and FP

detections. Alternatively, it may be used interactively in mammographic viewing stations to assist radiologists in distinguishing regions containing masses or normal tissue, which are manually localized. Although the accuracy of this method needs to be improved in order to be clinically practical, the results demonstrate the feasibility of using linear discriminant analysis in the texture feature space to differentiate masses and normal breast tissue on mammograms.

Acknowledgments

This work is supported by USPHS grant CA 48129, a faculty research award (FRA-334) from the American Cancer Society, and US Army grant DAMD 17-93-J-3007 (through subgrant GU RX 4300-803UM from Georgetown University). The content of this publication does not necessarily reflect the position of the university or the government and no official endorsement of any equipment and product of any companies mentioned in the publication should be inferred. The authors are grateful to Charles E Metz, PhD, for the LABROC1 programs, and to Diane Williams for secretarial assistance.

References

Bird R E, Wallace T W and Yankaskas B C 1992 Analysis of cancers missed at screening mammography *Radiology* **184** 613–17

Brzakovic D, Luo X M and Brzakovic P 1990 An approcach to automated detecion of tumors in mammograms *IEEE Trans. Med. Imaging* **MI-9** 233–41

Caldwell C B, Stapleton S J, Holdsworth D W, Jong R A, Weiser W J, Cooke G and Yaffe M J 1990 Characterization of mammographic parenchymal pattern by fractal dimension *Phys. Med. Biol.* **35** 235–47

Chan H P, Doi K, Vyborny C J, Schmidt R A, Metz C E, Lam K L, Ogura T, Wu Y and MacMahon H 1989 Improvement of radiologists' detection of clustered microcalcifications on mammograms: the potential of computer-aided diagnosis *Invest. Radiol.* **25** 1102–10

Chan H P, Niklason L T, Ikeda D M, Lam K L and Adler D D 1994 Digitization requirements in mammography: effects on computer-aided detection of microcalcifications *Med. Phys.* **21** 1203–11

Cheng S N C, Chan H P, Helvie M A, Goodsitt M M, Adler D D and St Clair D 1994 Classification of mass and non-mass regions on mammograms using artificial neural network *J. Imaging Sci. Technol.* **38** 598–603

Conners R W 1979 Towards a set of statistical features which measure visually perceivable qualities of textures *Proc. IEEE Conf. on Pattern Recognition and Image Processing* (New York: IEEE) pp 382–90

Haralick R M 1986 Statistical image texture analysis *Handbook of Pattern Recognition and Image Processing* (New York: Academic)

Haralick R M, Shanmugam K and Dinstein I 1973 Texture features for image classification *IEEE Trans. Systems Man Cybernet.* **SMC-3** 610–21

Harvey J A, Fajardo L L and Innis C A 1993 Previous mammograms in patients with impalpable breast carcinomas: retrospective vs blinded interpretation *Am. J. Radiol.* **161** 1167–72

Kegelmeyer W P, Pruneda J M, Bourland P D, Hillis A, Riggs M W and Nipper M L 1994 Computer-aided mammographic screening for spiculated lesions *Radiology* **191** 331–7

Lachenbruch P A 1975 *Discriminant Analysis* (New York: Hafner)

Lai S M, Li X and Bischof W F 1989 On techniques for detecting circumscribed masses in mammograms *IEEE Trans. Med. Imaging* **MI-8** 377–86

Magnin I E, Cluzeau F, Odet C L and Bremond A 1986 Mammographic texture analysis: an evaluation of risk for developing breast cancer *Opt. Eng.* **25** 780–4

Martin J E, Moskowitz M and Milbrath J R 1979 Breast cancer missed by mammography *Am. J. Radiol.* **132** 737–9

Metz C E 1986 ROC methodology in radiologic imaging *Invest. Radiol.* **21** 720–33

Metz C E, Shen J H and Herman B A 1990 New methods for estimating a binormal ROC curve from continuously-distributed test results *1990 Ann. Meeting Am. Stat. Assoc. (Anaheim, CA, 1990)*

Metz C E, Wang P L and Kronman H B 1984 A new approach for testing the significance for difference between ROC curves measured from correlated data *Information Processing in Medical Imaging* ed F Deconinck (The Hague: Martinus Nijhoff)

Ng S L, Bischof W F 1992 Automated detection and classification of breast tumors *Comput. Biomed. Res.* **25** 218–37

Petrosian A, Chan H P, Helvie M A, Goodsitt M M and Adler D D 1994 Computer-aided diagnosis in mammography: classification of masses and normal tissue by texture analysis *Phys. Med. Biol.* **39** 2273–88

1993 *SPSS for Windows* Release 6 Professional Statistics (Chicago, IL: SPSS)

Swets J A and Pickett R M 1982 *Evaluation Of Diagnostic System: Methods From Signal Detection Theory* (New York: Academic)

Tatsuoka M M 1988 *Multivariate Analysis, Techniques for Educational and Psychological Research* 2nd edn (New York: Macmillan)

Thurfjell E L, Lernevall K A and Taube A A S 1994 Benefit of independent double reading in a population-based mammography screening program *Radiology* **191** 241–4

Wallis M G, Walsh M T and Lee J R 1991 A review of false negative mammography in a symptomatic population *Clin. Radiol.* **44** 13–15

Wolfe J N 1966 Mammography: errors in diagnosis. *Radiology* **87** 214–19

Yin F F, Giger M L, Doi K, Metz C E, Vyborny C J and Schmidt R A 1991 Computerized detection of masses in digital mammograms: analysis of bilateral subtraction images *Med. Phys.* **18** 955–63