Computerized classification of malignant and benign microcalcifications on mammograms: texture analysis using an artificial neural network

Heang-Ping Chan[†], Berkman Sahiner, Nicholas Petrick, Mark A Helvie, Kwok Leung Lam, Dorit D Adler and Mitchell M Goodsitt

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

Received 22 July 1996

Abstract. We investigated the feasibility of using texture features extracted from mammograms to predict whether the presence of microcalcifications is associated with malignant or benign pathology. Eighty-six mammograms from 54 cases (26 benign and 28 malignant) were used as case samples. All lesions had been recommended for surgical biopsy by specialists in breast imaging. A region of interest (ROI) containing the microcalcifications was first corrected for the low-frequency background density variation. Spatial grey level dependence (SGLD) matrices at ten different pixel distances in both the axial and diagonal directions were constructed from the background-corrected ROI. Thirteen texture measures were extracted from each SGLD matrix. Using a stepwise feature selection technique, which maximized the separation of the two class distributions, subsets of texture features were selected from the multi-dimensional feature space. A backpropagation artificial neural network (ANN) classifier was trained and tested with a leave-one-case-out method to recognize the malignant or benign microcalcification clusters. The performance of the ANN was analysed with receiver operating characteristic (ROC) methodology. It was found that a subset of six texture features provided the highest classification accuracy among the feature sets studied. The ANN classifier achieved an area under the ROC curve of 0.88. By setting an appropriate decision threshold, 11 of the 28 benign cases were correctly identified (39% specificity) without missing any malignant cases (100% sensitivity) for patients who had undergone biopsy. This preliminary result indicates that computerized texture analysis can extract mammographic information that is not apparent by visual inspection. The computer-extracted texture information may be used to assist in mammographic interpretation, with the potential to reduce biopsies of benign cases and improve the positive predictive value of mammography.

1. Introduction

Mammography is the most sensitive method for detection of early breast cancer. However, the specificity for classification of malignant and benign lesions from mammographic images is quite low. In the United States, the positive predictive value, i.e., the ratio of the number of breast cancers found to the total number of biopsies, of mammography is typically between 15 and 30% (Kopans 1991, Adler and Helvie 1992). An improvement in the positive predictive value would reduce health care costs and eliminate the anxiety and morbidity of patients who would have to undergo unnecessary biopsy otherwise. One

[†] Address for correspondence: Heang-Ping Chan, PhD, Department of Radiology, University of Michigan Hospital, 1500 E Medical Center Drive, 2910 Taubman Center, Ann Arbor, MI 48109-0326, USA. E-mail address: chanhp@umich.edu

0031-9155/97/030549+19$19.50

of the potential approaches to improving the specificity of mammography is the use of computerized feature extraction techniques to extract information that may not be readily perceived by human readers. The computer-extracted features may complement the visual characteristics of the mammographic abnormalities and provide additional information to the radiologists in distinguishing malignant and benign lesions. The computer-extracted features, alone or in combination with human-perceived features, may also be input to a trained classifier to estimate the likelihood of malignancy of a mammographic lesion, thereby assisting radiologists in making diagnostic decisions.

A number of researchers have attempted to develop feature extraction and classification techniques for masses (Ackerman and Gose 1972; Kilday *et al* 1993, Huo *et al* 1995, Sahiner *et al* 1996a) or microcalcifications (Wee *et al* 1975, Fox *et al* 1980, Chan *et al* 1992, Chitre *et al* 1993, Chan *et al* 1994b, Shen *et al* 1994, Chan *et al* 1995a, c, d, Wu *et al* 1995, Jiang *et al* 1996, Thiele *et al* 1996). Other researchers used radiologists' ratings of mammographic features or encoded the radiologists' readings with numerical values as input to classifiers (Ackerman *et al* 1973, Gale *et al* 1987, Getty *et al* 1988, D'Orsi *et al* 1992, Wu *et al* 1993, Baker *et al* 1996). While the accuracy of lesion characterization in these studies varied, they demonstrated that computer-aided classification has the potential to improve the malignant and benign diagnosis of breast lesions. We have been developing computerized feature-extraction techniques for classification of masses or microcalcifications (Chan *et al* 1992, 1994b, 1995a, c, d, Sahiner *et al* 1996a). The extracted features are analysed by linear or non-linear classifiers which are trained for a specific classification task. We have found that texture features are effective for differentiation of masses and normal tissues (Chan *et al* 1995b, Wei *et al* 1995b), and that morphological features can be used to distinguish malignant and benign clustered microcalcifications (Chan *et al* 1995c). Because the tissue texture in regions containing microcalcifications associated with a malignant process may be different from that associated with a benign process, in the present study we analysed texture features from a region of interest (ROI) containing clustered microcalcifications (Chan *et al* 1995d). The effectiveness of these texture features, in combination with a backpropagation neural network classifier (Freeman and Skapura 1991), for the differentiation of malignant and benign microcalcifications was evaluated. The performance of the neural network was analysed with receiver operating characteristic (ROC) methodology (Swets and Pickett 1982, Metz *et al* 1990).

2. Materials and methods

2.1. Case selection and digitization

In this study, 86 mammograms with clustered microcalcifications were selected from patient files in the Department of Radiology at the University of Michigan. The mammograms were acquired with dedicated mammographic systems with a 0.3 mm focal spot, molybdenum (Mo) anode and 0.03 mm Mo filter. A Kodak Min R/MRE mammographic screen–film system using extended cycle processing was employed as the image receptor. The selection criteria were that the mammogram contained a cluster of microcalcifications, that about half of the case samples were malignant and half were benign, and that no grid lines were visible on the mammogram. The data set included 86 films, some of which were films of different views from the same patient. A total of 54 different patients were included in the data set. There were 41 malignant (26 patients) and 45 benign (28 patients) clusters. The malignant and benign pathology of the microcalcifications had been proven by open surgical biopsy and histologic analysis. The visibility of the microcalcification clusters was ranked by experienced radiologists on a scale of 1–5 (1, very obvious; 5, very subtle) relative to the range of cases seen in clinical practice. The histogram of the visibility for the 86 clusters is shown in figure 1.

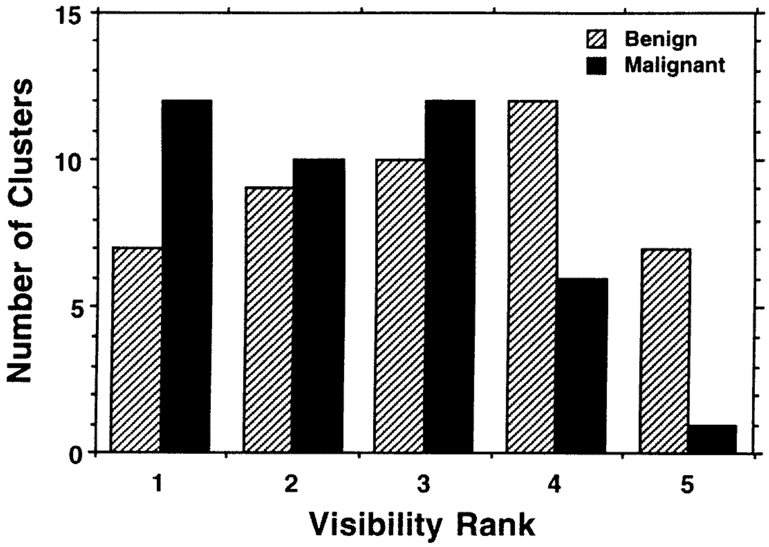


Figure 1. A histogram of the subjective ranking of the visibility of the 86 microcalcification clusters on the mammograms. The clusters were ranked on a five-point scale relative to the range of visibility of clusters found in clinical practice (1, very obvious; 5, very subtle).

All mammograms were digitized with a laser film scanner (Lumisys DIS-1000) at a pixel size of 35 μm × 35 μm and with a 12-bit grey level. The light transmitted through the film was amplified logarithmically before analogue-to-digital conversion. The digitizer had an optical density range of 0–3.5. It was calibrated so that the optical density (O.D.) on film was linearly proportional to the output pixel value in the range of about 0.1–2.8 O.D. with a slope of 0.001 O.D./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.

In this study, the locations of the microcalcification cluster on each mammogram were identified by radiologists so that only true microcalcification clusters were analysed. An ROI of 1024 × 1024 pixels (corresponding to 3.58 cm × 3.58 cm on the film), with the cluster approximately at its centre was extracted for analysis. This ROI size could enclose the majority of the clusters in the data set. A few of the obvious clusters scattered over a larger area, but the main area of the clusters was covered within the ROI.

The low-frequency background grey levels of each ROI depend mainly on the density of the overlapping breast tissue and the x-ray exposure conditions. The background levels therefore do not relate directly to the presence of the microcalcifications, but they bias the numerical values of the texture features. In order to eliminate the variability in the texture feature distributions caused by these factors that are not related to malignancy, we applied a background correction technique to the ROI before texture feature extraction. This technique has been described in detail previously (Chan *et al* 1995b). Briefly, the grey level at a given pixel of the low-frequency background was estimated as the average of the distance-weighted grey levels of four pixels at the intersections of the normals from the given pixel to the four edges of the ROI. An example of an original ROI with a malignant cluster, its estimated background image, and the background-corrected ROI is shown in figure 2(a)–(c), respectively. It can be seen that the sloped background grey level of the ROI was removed by the correction. The high-frequency information in the ROI was basically unchanged because the background image only contained low spatial frequencies.

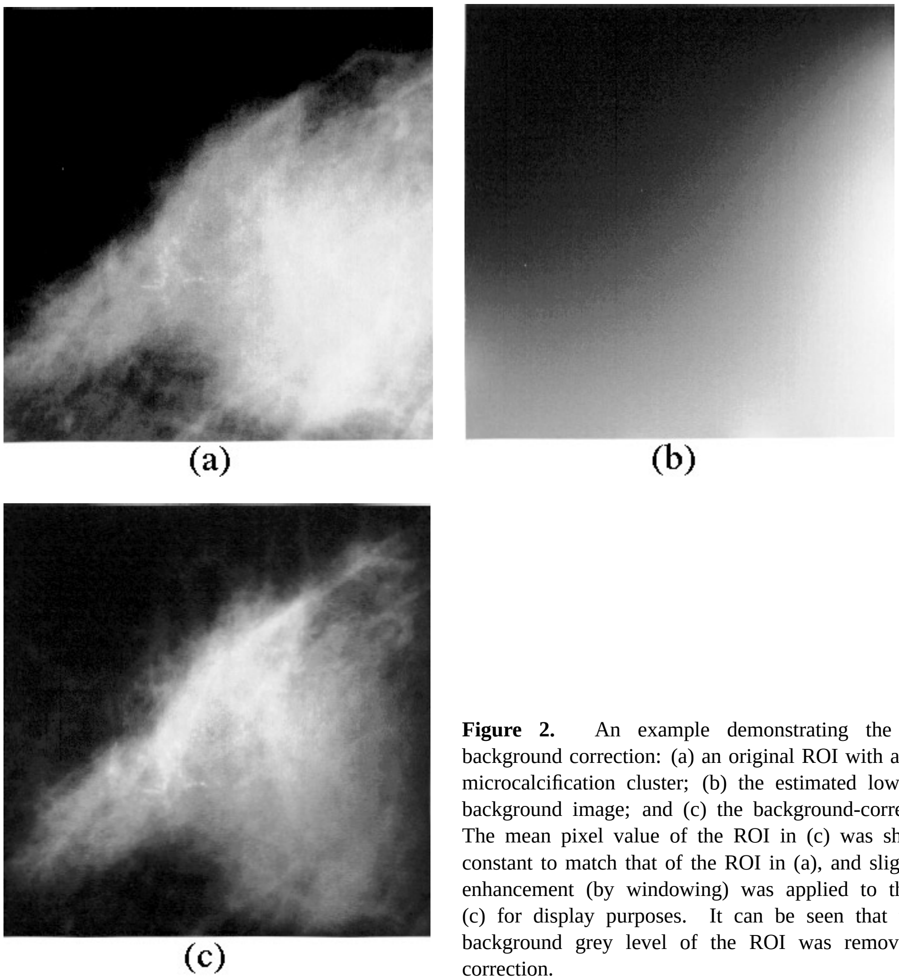


Figure 2. An example demonstrating the effect of background correction: (a) an original ROI with a malignant microcalcification cluster; (b) the estimated low-frequency background image; and (c) the background-corrected ROI. The mean pixel value of the ROI in (c) was shifted by a constant to match that of the ROI in (a), and slight contrast enhancement (by windowing) was applied to the ROI in (c) for display purposes. It can be seen that the sloped background grey level of the ROI was removed by the correction.

2.2. Texture features

Our previous studies indicated that the texture features derived from the spatial grey level dependence matrix (SGLD) (Haralick *et al* 1973), also known as the concurrence matrix or the co-occurrence matrix, of the ROI were useful in classification of masses and normal breast tissue (Cheng *et al* 1994, Petrosian *et al* 1994, Chan *et al* 1995b). We further expanded the texture feature space to include multi-distance features and obtained improved results (Wei *et al* 1995a). In this study, we applied texture analysis to the evaluation of textural changes in the breast tissue due to a developing malignancy. The SGLD matrix element, $p_{\theta,d}(i, j)$, is the joint probability of the occurrence of grey levels i and j for pixel pairs which are separated by a distance d and at a direction θ. Because of the discrete nature of the digital image, the distance d is limited to integral multiples of the pixel size, and the value of θ is limited to 0, 45, 90, and 135° at $d = 1$, and to these and other discrete angles as d increases. We constructed SGLD matrices from pixel pairs in a sub-region of 512×512 pixels centred approximately at the cluster in the background-corrected ROI. Four SGLD matrices, one at each of the four directions, 0, 45, 90, and 135°, were constructed for a given pixel pair distance. The pixel pair distance was varied from four to 40 pixels in increments of four pixels. Therefore, a total of 40 SGLD matrices were derived from each ROI.

The SGLD matrix depends on the bin width (or grey level interval) used in accumulating the histogram. We found in our previous mass classification study (Chan *et al* 1995b) that a bin width of 16 grey levels was a reasonable compromise between grey level resolution and statistical noise. In this study, the ROIs had two times more pixels in width and in height than those in our previous studies, resulting in four times as many pixels in each ROI. Thus, we could use a smaller bin width to obtain approximately the same statistics in the SGLD matrices. Furthermore, our previous study on the digitization requirements of mammograms (Chan *et al* 1994a) indicated that at least nine-bit grey level resolution was required for detection of subtle microcalcifications. We therefore chose a bin width of four grey levels for all SGLD matrices in this study. This is equivalent to reducing the grey level resolution (or bit depth) of the 12-bit image to ten bits by eliminating the two least significant bits.

A number of texture features can be derived from an SGLD matrix (Haralick *et al* 1973, Conners 1979). In our previous studies for mass and normal tissue classification (Chan *et al* 1995b, Wei *et al* 1995a), we evaluated eight texture measures: correlation, entropy, energy (angular second moment), inertia, inverse difference moment, sum average, sum entropy, and difference entropy. In this study, we included five additional texture features: difference average, sum variance, difference variance, information measure of correlation 1, and information measure of correlation 2. The mathematical expressions of these 13 texture features are given in the appendix. 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 (Haralick *et al* 1973).

As discussed in our previous study (Chan *et al* 1995b), we did not find a significant dependence of the discriminatory power of the texture features on the direction of the pixel pairs for mammographic textures. However, since the actual distance between the pixel pair in the diagonal direction was a factor of $\sqrt{2}$ of that in the axial direction, we averaged the feature values at the axial directions (0 and 90°) and also at the diagonal directions (45 and 135°) separately for each texture measure derived from the SGLD matrix at a given pixel pair distance. The average texture features at the ten pixel pair distances therefore formed a 260-dimensional feature space for the classification task.

2.3. Feature selection

The dimension of the texture feature space derived from the SGLD matrices at different pixel distances and directions is very large. It is well known that the presence of ineffective features often degrades classifier performance, especially when the training data set is small (Raudys and Pikelis 1980, Fukunaga and Hayes 1989). Investigators in CAD research have employed different methods for feature selection. Goldberg *et al* (1992) selected features for classifying malignant and benign masses on ultrasound images by evaluation of the discriminatory ability of the individual features. Wu *et al* (1993) selected features based on the difference in the average values of the individual features between the two classes. Lo *et al* (1995) ranked the importance of each feature based on its effect on the classification accuracy, and then eliminated the features, one at a time, from the least important to the most important, to determine the smallest set of features that provided the highest classification accuracy in their data set.

The stepwise procedure in linear discriminant analysis is an established method for selection of useful features for a classification task (Norusis 1993). In our previous studies, we have employed stepwise feature selection and successfully selected a small number of effective features from very large feature spaces (Chan *et al* 1995b, Wei *et al* 1995a). A detailed description of this procedure can be found in the literature. Briefly, one feature is added to or removed from the selected feature set in alternate steps. The effect of the feature on the separation of the two groups is analysed using the Wilks lambda criterion (minimization of the ratio of the within-group sum of squares to the total sum of squares of the two class distributions). The significance of the change in the Wilks lambda when a feature is added to or removed from the model is estimated by F statistics. The user can choose the values of two parameters, the F-to-enter threshold (F_{in}) and the F-to-remove threshold (F_{out}), to control the number of features to be selected. In the feature entry step, each of the features not yet in the model is entered one at a time. The feature variable that causes the most significant change in the Wilks lambda will be included in the feature set if the F value is greater than the F_{in} threshold. In the feature removal step, each of the features already in the model is removed one at a time. The feature variable that causes the least significant change in the Wilks lambda will be excluded from the feature set if the F value is below the F_{out} threshold. The stepwise procedure terminates when the F values for all features not in the model are smaller than the F_{in} threshold and the F values for all features in the model are greater than the F_{out} threshold. Therefore, the number of selected features will decrease if either the F_{in} threshold or the F_{out} threshold is increased. Since the optimal values of the two F thresholds are not known *a priori*, we varied these two thresholds over a wide range to obtain feature sets containing different number of features. The classification accuracies of the different feature sets were then evaluated as described below.

2.4. The artificial neural network (ANN)

We used a feed-forward backpropagation ANN for feature classification in the texture feature space. In this ANN, the nodes are organized in an input layer, an output layer, and one or more hidden layers as shown in figure 3. The nodes are interconnected by weights and information propagates from one layer to the next through a sigmoidal activation function. The learning of the ANN is a supervised process in which known training cases are input to the ANN and the weights are adjusted with an iterative backpropagation procedure in order to achieve a desired input–output relationship. Detailed description of the backpropagation algorithm can be found in the literature (Freeman and Skapura 1991).

To improve the convergence rate and the stability of training, we implemented batch processing in which the weight changes obtained from each training case were accumulated and the weights were updated after the entire set of training cases was evaluated. The batch processing method improves the stability with a tradeoff in the convergence rate. To improve the convergence rate, we included a momentum term and used the delta-bar-delta rule for updating the weights (Sahiner *et al* 1996b). The updated weight is given by

$$w_i(t+1) = w_i(t) - \eta_i(t)\Delta w_i(t)$$

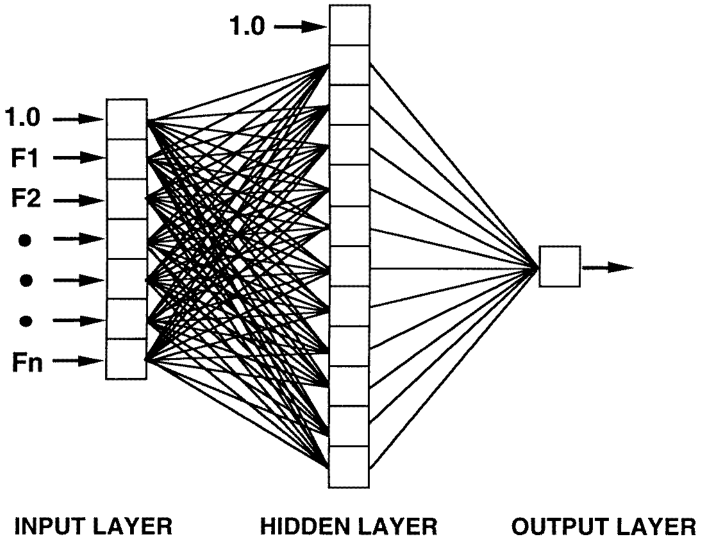


Figure 3. A schematic diagram of the backpropagation neural network classifier used in this study. The number of input nodes was equal to the number of input features. The number of hidden nodes could be varied to obtain the best performance. One output node was used in all ANNs. An ANN with I input nodes, H hidden nodes, and one output node will be denoted as I–H–1.

where $\eta_i(t)$ is the learning rate, $w_i(t)$ is the weight and $\Delta w_i(t)$ is the weight increment for the ith node at training epoch t. When $\eta_i(t)$ is small, the learning is slow but stable. When $\eta_i(t)$ is large, learning is fast but can be unstable. In the delta-bar-delta rule, $\eta_i(t)$ is adjusted adaptively based on the weight increments in two consecutive epochs.

If $\Delta w_i(t-1)\Delta w_i(t) > 0$, $\eta_i(t-1)$ is too small and can be increased:

$$\eta_i(t) = \eta_i(t-1) + \varepsilon \qquad \varepsilon > 0.$$

If $\Delta w_i(t-1)\Delta w_i(t) < 0$, $\eta_i(t-1)$ is too large and should be reduced by a factor r:

$$\eta_i(t) = \eta_i(t-1)r \qquad 0 < r < 1.$$

In this study, we applied a leave-one-out method to training and testing of the ANN classifier. If a data set with N samples is available for training and testing, $(N-1)$ samples will be used for training the classifier and the trained classifier will be evaluated with the left-out test sample. The procedure is repeated N times, each time with a different left-out sample. The test results of the N samples are accumulated to form a distribution of test scores. In the present study, all images of the same patient were left out as test samples in each training cycle and the images from the other $(N-1)$ patients were used for training. The results of all test images from the N training cycles were accumulated to form a distribution of test scores.

Another commonly used method for training and testing a classifier with a small data set is a cross-validation method (Weiss and Kulilowski 1991). In this method, the data set is randomly partitioned into a training set and a test set with a specified training-to-test-case ratio. The training and testing of the classifier are then performed with the partitioned training and test sets, respectively. To reduce the dependence on the training and test cases, the procedure is repeated many times with different partitioning. The results are then averaged over the many partitions to obtain an estimate of the classifier performance. We performed a limited study using the cross-validation method and compared the results with the leave-one-out method. To ensure independence of the training and test sets in the cross-validation method, the case partitioning was performed with the constraint that images of the same patient were always grouped into the same set.

The performance of the ANN classifier was evaluated by ROC methodology (Swets and Pickett 1982, Metz 1986). The output value of the ANN 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), could be generated by setting different decision thresholds on the output values of the ANN. In this study, we used the LABROC 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, to estimate the area under the ROC curve (A_z) and the standard deviation (SD) of A_z. A_z was used as an index of classification accuracy. For the leave-one-out method, the test A_z was obtained from analysis of the accumulated test score distribution from all N cycles. For the cross-validation method, the average performance of the ANN was estimated as the average of the 50 test A_z values obtained from training and testing with 50 different partitions of the data sets.

3. Results

Some representative subsets of features selected by the stepwise procedure from the 260-dimensional texture feature space are listed in table 1. The number of features was varied by changing the F_{in} and F_{out} thresholds as shown in the table. The number of features selected usually remained constant over a range of F_{in} and F_{out} thresholds. For example, there were six selected features when (F_{in}, F_{out}) were reduced from about (2.65, 2.55) to (2.1, 2.0), and seven selected features when reduced from about (1.9, 1.8) to (0.56, 0.55). When the F_{in} and F_{out} were reduced slightly further, the number of selected features increased abruptly to 19.

We evaluated each feature subset by using the feature subset as input to the ANN and estimating the classification accuracy A_z. For a given feature set containing I features, an ANN with I input nodes, one to ten hidden nodes, and one output node was trained with the leave-one-case-out method as described above. For each training cycle with $(N - 1)$ training cases, the ANN was trained up to 30 000 epochs. The test result for the left-out case was obtained at fixed intervals of epochs (e.g., every 1000 epochs). After the N training cycles were completed, the test results of the entire dataset would have been accumulated at the fixed intervals of epochs. Therefore, an ROC curve could be fitted to the output of the test cases and the A_z estimated at the fixed intervals of epochs. Figure 4 shows the typical convergence trend of the test A_z results as the training epochs increased. The test A_z generally increased rapidly for the first few thousand epochs and then levelled off gradually. In this example, the test A_z remained at a constant level of about 0.88 when the ANN was trained for more than 8000 epochs. In some cases, the test A_z decreased if the ANN was over-trained. The test A_z values reported in the following discussion were obtained at the maximum plateau region.

The dependence of the classification accuracy, A_z, on the ANN architecture is shown in figure 5 for the different feature subsets. For convenience of comparison, an ANN without a hidden layer was plotted as an ANN with zero hidden nodes. The number of hidden nodes for a three-layer ANN was varied from one to ten. The standard deviation (SD) of the A_z, estimated by the LABROC1 program, ranged from 0.035 to 0.045. For a given feature set,

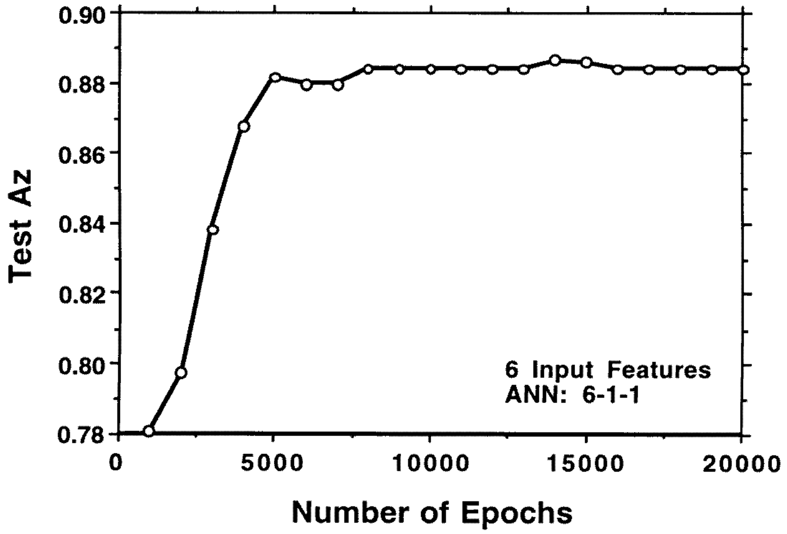


Figure 4. An example demonstrating the dependence of test A_z on the number of training epochs. The test A_z generally increased rapidly during the first 5000 epochs and then gradually reached a plateau or a broad maximum.

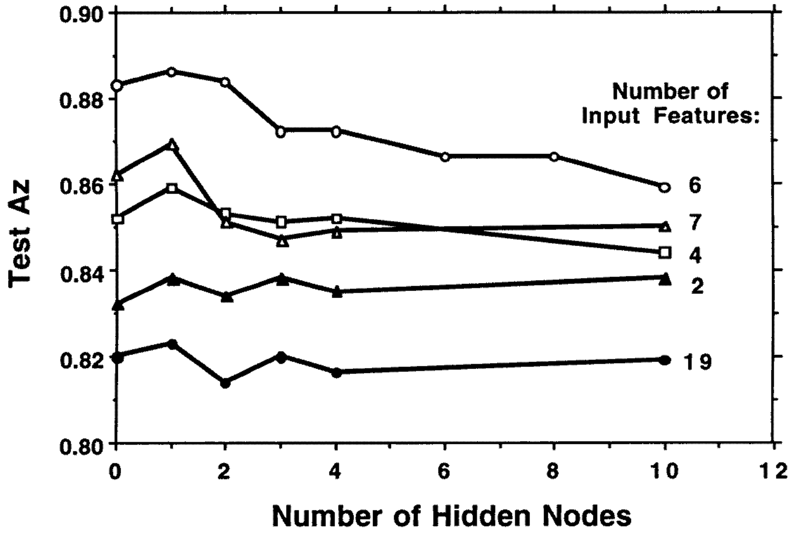


Figure 5. The dependence of the classification accuracy, A_z, on the number of hidden nodes in the ANN classifier. To facilitate comparison, the results for a two-layer ANN that had no hidden layer were plotted as data points with zero hidden nodes. The ANNs with one hidden node consistently provided higher accuracy than the other ANNs for all input feature sets. The input feature set with six selected features was the most effective in classifying malignant and benign microcalcifications among the selected feature sets.

Table 1. Texture features selected by stepwise feature selection procedure for different F-to-enter (F_{in}) and F-to-remove (F_{out}) thresholds.

$F_{in} = 3.84$ $F_{out} = 2.71$	$F_{in} = 2.7$ $F_{out} = 2.6$	$F_{in} = 2.5$ $F_{out} = 2.4$	$F_{in} = 1.7$ $F_{out} = 1.5$	$F_{in} = 0.55$ $F_{out} = 0.45$
Diff. entropy ($d = 8$) Inv. diff. moment ($d = 4$)	Correlation ($d = 40$, diagonal) Diff. entropy ($d = 8$) Inertia ($d = 40$) Inv. diff. moment ($d = 4$)	Diff. average ($d = 4$) Diff. entropy ($d = 8$) Diff. entropy ($d = 32$, diagonal) Inertia ($d = 4$) Inertia ($d = 40$) Inv. diff. moment ($d = 12$)	Correlation ($d = 40$, diagonal) Diff. average ($d = 4$) Diff. entropy ($d = 8$) Diff. entropy ($d = 32$, diagonal) Inertia ($d = 40$) Inv. diff. moment ($d = 12$) Inv. diff. moment ($d = 4$)	Correlation ($d = 8$) Diff. average ($d = 32$) Diff. average ($d = 4$) Diff. average ($d = 40$, diagonal) Diff. entropy ($d = 8$) Diff. entropy ($d = 32$, diagonal) Diff. variance ($d = 40$, diagonal) Energy ($d = 24$, diagonal) Information measure of correlation 1 ($d = 36$) Information measure of correlation 1 ($d = 40$, diagonal) Information measure of correlation 2 ($d = 24$) Information measure of correlation 2 ($d = 36$) Information measure of correlation 2 ($d = 4$, diagonal) Inertia ($d = 4$) Inertia ($d = 40$) Inv. diff. moment ($d = 12$) Inv. diff. moment ($d = 8$) Inv. diff. moment ($d = 4$, diagonal) Inv. diff. moment ($d = 8$, diagonal)

the variation of the A_z values with the number of ANN hidden nodes was within one SD. However, the maximum A_z consistently occurred at the ANN with one hidden node for all feature sets. The feature set with six features provided the highest A_z over the entire range of hidden nodes studied. The maximum A_z of 0.88 was obtained with an ANN of six input nodes, one hidden node, and one output node. The ROC curves that had the two highest A_z values obtained with six and seven input features and one hidden node are plotted in figure 6.

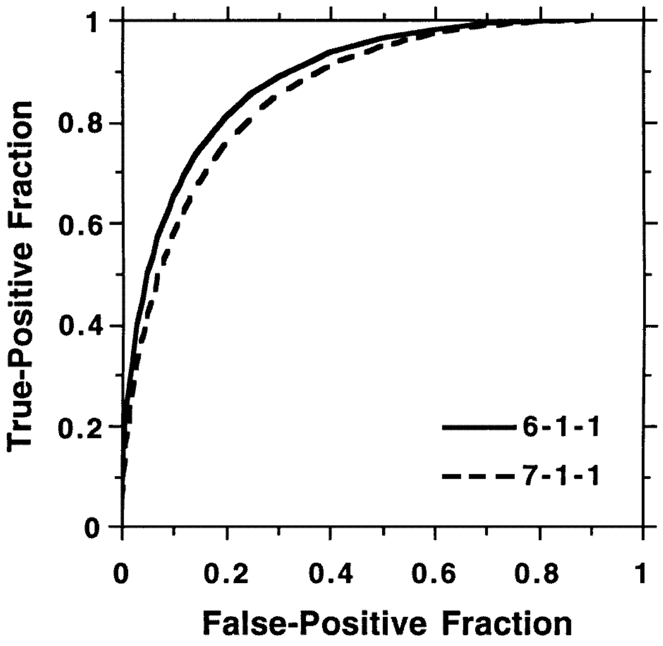


Figure 6. ROC curves that had the two highest A_z values obtained with the six-feature and seven-feature sets and one hidden node shown in figure 5.

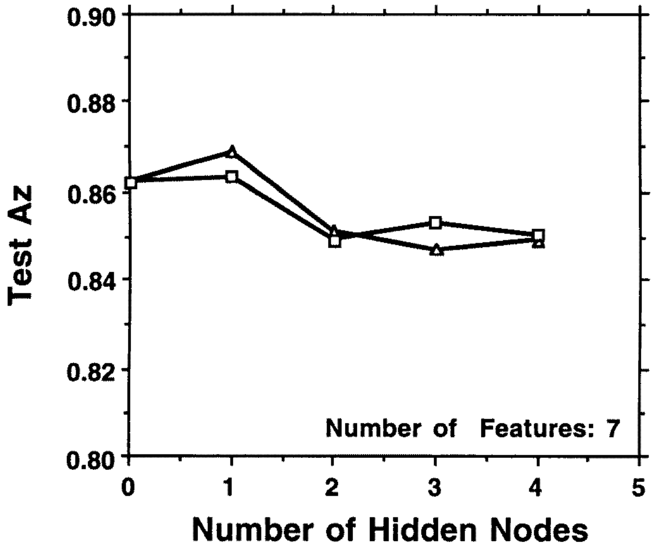


Figure 7. The effect of the initialization of the weights in the ANN on classifier performance. An ANN with seven input nodes, zero to four hidden nodes, and one output node was studied. The two data points at each ANN configuration represent the two different initializations of its weights. The difference in the initial weights appears to have very small effect on the convergence of the ANN.

To evaluate the variation of the classification accuracy on the initialization of the ANN, we used two different random number seeds to generate the initial weights for the ANNs with seven input features. The A_z values are plotted in figure 7 for the ANNs with different numbers of hidden nodes. The differences in A_z were within 0.01 for the different ANNs, indicating that the initial weights do not have a strong effect on the convergence of the ANNs.

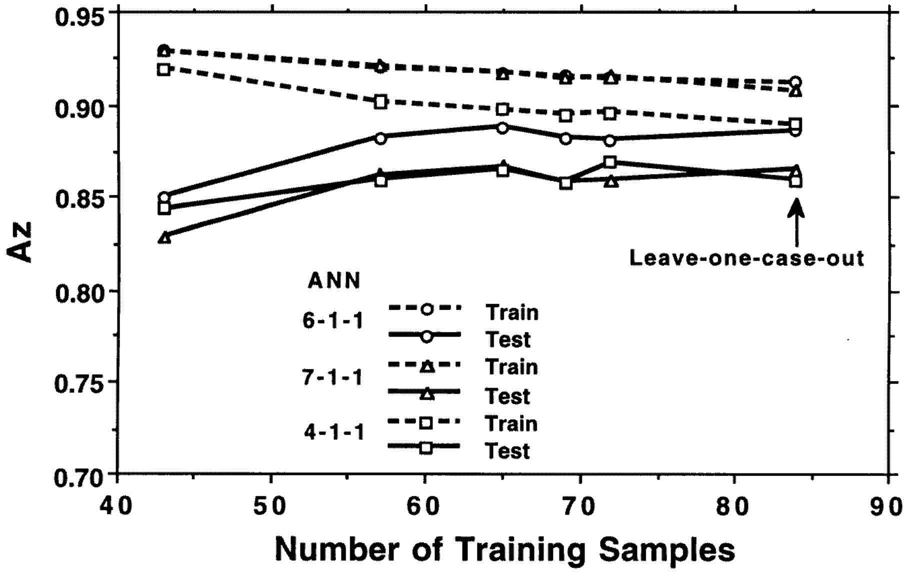


Figure 8. The dependence of A_z on the number of training cases obtained from a cross-validation method. The number of training cases was varied by randomly partitioning the data set into a training set and a test set with training-to-test-sample ratios of one to five. For a given training-to-test-sample ratio, the training (or test) A_z plotted was the average of the 50 A_z values obtained from the 50 random partitions of the data set. For comparison, the A_z values obtained with the leave-one-case-out training and test method were also plotted as the data points with 84 training samples.

Figure 8 shows the performance of the ANN classifiers which were trained and tested with a cross-validation method. The number of input nodes of the ANNs corresponded to the number of input features; the numbers of hidden nodes and output nodes were both set to be one. The training-to-test sample ratio was varied from one to five. The data set was randomly partitioned 50 times at each ratio and the mean training and test A_z values from the 50 partitions were plotted against the number of training samples. Because of the constraint that films of the same patient were always grouped into the same set, the number of training (or test) samples in each of the 50 partitions might not be equal: the expected number of training samples calculated as the nearest integer of $[86R/(R+1)]$, where R is the training-to-test-sample ratio, was plotted as the abscissae. As the training-to-test-sample ratios increased from one to five, the expected number of training samples increased from 43 to 72. To facilitate comparison, the A_z for the corresponding ANN classifiers trained with the leave-one-case-out method was plotted as the data point having an expected number of training samples of 84.

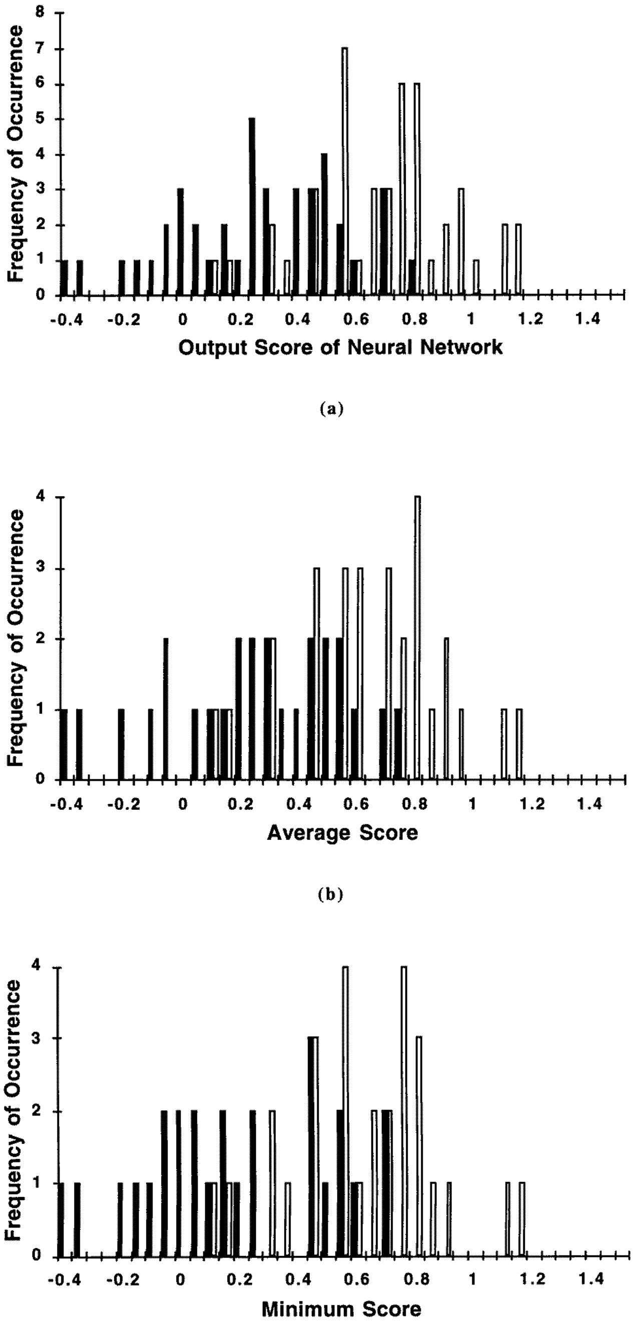
It can be seen that the training A_z decreased slowly as the number of training samples increased. The test A_z, on the other hand, increased as the number of training samples increased. Because the number of test samples was small when the ratio was large, the SD for each test A_z ranged from 0.06 to 0.12 when the ratio increased from one to five. However, the SDs of the mean test A_z values from the 50 partitions varied from 0.01 to 0.02. It can be seen that the fluctuations of the data points were within one SD of the mean. The trend of the curves generally agrees with the expectations that small training sets over-estimate the classifier performance and the trained classifiers perform poorly on test sets, and that both the training and test results will approach the 'true' performance as the number of training samples approaches infinity (Raudys and Pikelis 1980, Fukunaga and Hayes 1989).

The output scores of the ANN with six input features, one hidden node, and one output node for the 86 test samples obtained with the leave-one-case-out method are plotted in figure 9(a). The output scores of the ANN have been scaled linearly for the purpose of plotting the graph. The linear transformation simply expands the horizontal scale without any effect on the relative distribution of the scores. It can be seen that there was good separation between the malignant and benign clusters. If the decision threshold was set at 0.85, 11 of the 45 benign samples were correctly classified without any false negatives (a sensitivity of 100% at a specificity of 24%). At a decision threshold of 0.75, 23 of the benign samples were correctly classified but one malignant sample was missed (a sensitivity of 98% at a specificity of 51%). When the ANN output scores were analysed with the LABROC1 program, the area under the fitted ROC curve was 0.88.

Because some of the samples are films from the same patient, it will be reasonable to make the malignant or benign decision on a case-by-case basis. Two approaches were investigated: one used the average score from all films of the same patient and the other used the minimum score from all films of the same patient for decision making. The latter was a more conservative approach because a lower score corresponded to higher likelihood of malignancy in our analysis. The distributions of the average scores and the minimum scores for the 54 cases are shown in figure 9(b) and (c), respectively. If a decision threshold were set at an average score of 0.80, ten of the 28 benign cases would be correctly classified without any false negative (a sensitivity of 100% at a specificity of 36%). Alternatively, if a decision threshold was set at a minimum score of 0.75, 11 of the 28 benign cases would be correctly classified without missing any malignant cases (a sensitivity of 100% at a specificity of 39%).

4. Discussion

We have investigated the usefulness of texture analysis in predicting the malignant and benign nature of abnormal breast tissue containing clustered microcalcifications. All case samples used in this study had been surgically biopsied, indicating that definitive diagnosis could not be made by the mammographic appearance of the benign clusters. Our results show that there are changes in the texture of the breast tissue in which a malignancy is developing, and that these changes can be distinguished from the benign tissue texture by computerized analysis although their differences are not visually apparent on mammograms. Based on the results of texture analysis and ANN classification, a significant fraction of benign cases can be correctly identified. This information may be used to reduce the number of biopsies, thereby improving the positive predictive value of mammography. Our preliminary study therefore demonstrates that computerized classification may be a useful aid in mammographic interpretation. Further investigation to determine if this approach can be generalized to large data sets is warranted.



(c)

Figure 9. The distributions of the discriminant scores for the malignant (black) and benign (white) microcalcification clusters. The test results for the ANN with six input features, one hidden node, and one output node trained with the leave-one-case-out method are shown. The output scores of the ANN have been scaled linearly for the purpose of plotting. (a) Distribution of the output scores from the ANN classifier for 86 test samples; (b) distribution of the average scores for 54 cases; (c) distribution of the minimum scores for 54 cases.

We used an ANN as a feature classifier for this classification task. By varying the structure of the ANN, both linear and non-linear classifiers could be studied. An analysis of the dependence of the classification accuracy on ANN architecture (figure 5) indicated that ANNs with one hidden node provided the best performance for all feature sets. Because an ANN with one hidden node is equivalent to a linear classifier, the results appear to indicate that a linear classifier may be the optimal choice for this classification task. However, it should be cautioned that the performance of a classifier depends on the number of training samples relative to the number of parameters to be trained in the classifier (Raudys and Pikelis 1980, Fukunaga and Hayes 1989). Since the data set in this study was small and the number of weights to be trained in an ANN increased rapidly with the number of hidden nodes, the observed reduction in classification accuracy with the increasing number of hidden nodes could be caused by insufficient training samples. The optimal choice of a feature classifier for this classification task will have to be investigated further when a large data set is available.

Thiele *et al* (1996) recently studied the classification of the tissue texture surrounding calcification clusters to predict malignant or benign outcomes. They used texture measures calculated from the SGLD matrices and fractal geometry as input to a linear discriminant classifier or a logistic discriminant classifier. Their results also demonstrated that texture analysis showed significant discriminatory power between benign and malignant tissue. In a data set of 54 cases (36 benign, 18 malignant), they obtained a sensitivity of 89% at a specificity of 83%. In their calculation of the SGLD matrices in the tissue region, they included subtle microcalcifications but excluded the pixels containing large and bright calcifications by manually identifying the calcification areas with grey level thresholding. In our SGLD matrix calculation, all pixels in the 512×512 ROI containing the microcalcification cluster were included. Because of the many differences between the two studies and the difference in the data set, it is not known which approach will provide more effective texture features. However, the advantage of our approach is that no manual identification of individual microcalcifications is needed and the analysis can be much more efficient. Minimal operator intervention will be a practical consideration if the computerized classification technique is to be implemented in clinical settings.

In this study, we performed background correction in a 1024×1024 ROI but calculated texture features in a subregion of 512×512 pixels centred approximately at the cluster of microcalcifications. The use of a subregion smaller than the original 1024×1024 ROI would avoid any potential edge effects caused by background correction. Furthermore, because many of the clusters in our data set could be enclosed by a 512×512 region, calculation of texture features in the original ROI would average the texture features in the cluster region with those in a large region of possibly normal tissue. The choice of the subregion size was subjective in this study, taking into consideration the tradeoff between the averaging effect and the statistics needed in the SGLD matrix formation. Whether a different choice of the region size, or use of variable size according to the cluster diameter, would improve the effectiveness of the texture features remains to be studied.

In this study, we did not perform a systematic optimization of the parameters for texture extraction. Many of the parameters were chosen based on our experience in other applications. The goal of this study is to demonstrate the feasibility of using computerized texture analysis for classification of malignant and benign microcalcifications. Our results indicate that the SGLD texture features are useful in such an application although the techniques have not been optimized. In future studies, both the feature extraction techniques and the classifier should be improved by optimization of the various parameters using a large data set.

5. Conclusion

We have developed a computerized method for classification of malignant and benign microcalcification clusters on mammograms. The computer extracts texture features from an ROI containing the microcalcification cluster and predicts its pathology using a trained neural network classifier. The effectiveness of our approach has been demonstrated with a small data set. The classifier could correctly identify a significant fraction of benign cases, which had been recommended for surgical biopsy under current clinical criteria, without missing any malignant cases. The computerized texture analysis may therefore provide useful information for reducing the number of negative biopsies. Further investigation will be conducted with a larger data set to determine the generalizability of these results. The combination of this texture classification method with other morphological features or patient information will be investigated. The optimization of the classifier design will also be examined.

Acknowledgments

This work is supported by USPHS grant CA 48129. The authors are grateful to Datong Wei, PhD, for his assistance in the initial stage of this study, and Charles E Metz, PhD, for the LABROC1 program.

Appendix. The spatial grey level dependence (SGLD) matrix and texture features

The (i, j)th element of the SGLD matrix, $p_{\theta,d}(i, j)$, is the joint probability that the grey levels i and j occur in a direction of angle θ and at a distance of d pixels apart over the entire ROI. The joint probability $p_{\theta,d}(i, j)$ is normalized by the number of grey level pairs obtained from the ROI with a pixel distance of d. For each ROI, thirteen texture measures were derived from its SGLD matrix as described below. Most of the expressions can be found in the literature (Haralick *et al* 1973). Some differences in the expressions may be noted. A simplified notation $p(i, j)$ will be used to denote the SGLD matrix elements in the following equations.

$$\text{Energy} = \sum_{i=0}^{n-1}\sum_{j=0}^{n-1}[p(i, j)]^2 \tag{A1}$$

where n is the number of grey levels in the image.

$$\text{Correlation} = \left(\sum_{i=0}^{n-1}\sum_{j=0}^{n-1}(i-\mu_x)(j-\mu_y)p(i, j)\right)\bigg/(\sigma_x\sigma_y) \tag{A2}$$

where

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

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

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

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

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

$$\text{Inertia} = \sum_{i=0}^{n-1} \sum_{j=0}^{n-1} (i - j)^2 p(i, j) \tag{A3}$$

$$\text{Entropy} = -\sum_{i=0}^{n-1} \sum_{j=0}^{n-1} p(i, j) \log_2 p(i, j) \tag{A4}$$

$$\text{Inverse difference moment} = \sum_{i=0}^{n-1} \sum_{j=0}^{n-1} \frac{1}{1 + (i - j)^2} p(i, j) \tag{A5}$$

$$\text{Sum average} = \sum_{k=0}^{2n-2} k p_{x+y}(k) \tag{A6}$$

where

$$p_{x+y}(k) = \sum_{i=0}^{n-1} \sum_{j=0}^{n-1} p(i, j) \qquad i + j = k \qquad k = 0, \ldots, 2n - 2.$$

$$\text{Sum variance} = \sum_{k=0}^{2n-2} (k - \text{sum average})^2 p_{x+y}(k) \tag{A7}$$

$$\text{Sum entropy} = -\sum_{k=0}^{2n-2} p_{x+y}(k) \log_2 p_{x+y}(k) \tag{A8}$$

$$\text{Difference average} = \sum_{k=0}^{n-1} k p_{x-y}(k) \tag{A9}$$

where

$$p_{x-y}(k) = \sum_{i=0}^{n-1} \sum_{j=0}^{n-1} p(i, j) \qquad |i - j| = k \qquad k = 0, \ldots, n - 1.$$

$$\text{Difference variance} = \sum_{k=0}^{n-1} (k - \text{difference average})^2 p_{x-y}(k) \tag{A10}$$

$$\text{Difference entropy} = -\sum_{k=0}^{n-1} p_{x-y}(k) \log_2 p_{x-y}(k) \tag{A11}$$

$$\text{Information measure of correlation 1} = (\text{entropy} - H_1)/\max\{H_x, H_y\} \tag{A12}$$

where

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

$$H_x = -\sum_{i=0}^{n-1} p_x(i) \log_2 p_x(i)$$

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

$$\text{Information measure of correlation } 2 = \sqrt{1 - \exp[-2(H_2 - \text{entropy})]} \qquad \text{(A13)}$$

where

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

References

Ackerman L V and Gose E E 1972 Breast lesion classification by computer and xeroradiograph *Cancer* **30** 1025–35

Ackerman L V, Mucciardi A N, Gose E E and Alcorn F S 1973 Classification of benign and malignant breast tumors on the basis of 36 radiographic properties *Cancer* **31** 342–52

Adler D D and Helvie M A 1992 Mammographic biopsy recommendations *Current Opinion Radiol.* **4** 123–9

Baker J A, Kornguth P J, Lo J Y and Floyd C E 1996 Artificial neural network: improving the quality of breast biopsy recommendations *Radiology* **198** 131–5

Chan H-P, Niklason L T, Ikeda D M and Adler D D 1992 Computer-aided diagnosis in mammography: detection and characterization of microcalcifications *Med. Phys.* **19** 831

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

Chan H-P, Sahiner B, Lam K L, Wei D, Helvie M A and Adler D D 1995a Classification of malignant and benign microcalcifications on mammograms using an artificial neural network *Proc. World Congress on Neural Networks (Washington, DC, 1995)* vol 2 (Mahwah, NJ: INNS) pp 889–92

Chan H-P, Wei D, Helvie M A, Sahiner B, Adler D D, Goodsitt M M and Petrick N 1995b Computer-aided classification of mammographic masses and normal tissue: linear discriminant analysis in texture feature space *Phys. Med. Biol.* **40** 857–76

Chan H-P, Wei D, Lam K L, Lo S-C B, Sahiner B, Helvie M A and Adler D D 1995c Computerized detection and classification of microcalcifications on mammograms *Proc. SPIE* **2434** 612–20

Chan H P, Wei D, Lam K L, Sahiner B, Helvie M A, Adler D D and Goodsitt M M 1995d Classification of malignant and benign microcalcifications by texture analysis *Med. Phys.* **22** 938

Chan H-P, Wei D, Niklason L T, Helvie M A, Lam K L, Goodsitt M M and Adler D D 1994b Computer-aided classification of malignant/benign microcalcifications in mammography *Med. Phys.* **21** 875

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

Chitre Y, Dhawan A P and Moskowitz M 1993 Artificial neural network based classification of mammographic microcalcifications using image structure features *Int. J. Pattern Recognition Artificial Intell.* **7** 1377–401

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

D'Orsi C J, Getty D J, Swets J A, Pickett R M, Seltzer S E and McNeil B J 1992 Reading and decision aids for improved accuracy and standardization of mammographic diagnosis *Radiology* **184** 619–22

Fox S H, Pujare U M, Wee W G, Moskowitz M and Hutter R V P 1980 A computer analysis of mammographic microcalcifications: global approach *Proc. IEEE 5th Int. Conf. on Pattern Recognition* (New York: IEEE) pp 624–31

Freeman J A and Skapura D M 1991 *Neural Networks—Algorithms, Applications, and Programming Techniques* (Reading, MA: Addison-Wesley)

Fukunaga K and Hayes R R 1989 Effects of sample size on classifier design *IEEE Trans. Pattern Anal. Machine Intell.* **11** 873–85

Gale A G, Roebuck E J, Riley P and Worthington B S 1987 Computer aids to mammographic diagnosis *Br. J. Radiol.* **60** 887–91

Getty D J, Pickett R M, D'Orsi C J and Swets J A 1988 Enhanced interpretation of diagnostic images *Invest. Radiol.* **23** 240–52

Goldberg V, Manduca A, Ewert D L, Gisvold J J and Greenleaf J F 1992 Improvement in specificity of ultrasonography for diagnosis of breast tumors by means of artificial intelligence *Med. Phys.* **19** 1475–81

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

Huo Z, Giger M L, Vyborny C J, Bick U, Lu P, Wolverton D E and Schmidt R A 1995 Analysis of spiculation in the computerized classification of mammographic masses *Med. Phys.* **22** 1569–79

Jiang Y, Nishikawa R M, Wolverton D E, Metz C E, Giger M L, Schmidt R A, Vyborny C J and Doi K 1996 Malignant and benign clustered microcalcifications: automated feature analysis and classification *Radiology* **198** 671–78

Kilday J, Palmieri F and Fox M D 1993 Classifying mammographic lesions using computerized image analysis *IEEE Trans. Med. Imaging* **12** 664–9

Kopans D B 1991 The positive predictive value of mammography *Am. J. Roentgenology* **158** 521–6

Lo J Y, Baker J A, Kornguth P J and Floyd C E 1995 Computer-aided diagnosis of breast cancer: artificial neural network approach for optimized merging of mammographic features *Acad. Radiol.* **2** 841–50

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 *Annu. Meeting Am. Stat. Assoc. (Anaheim, CA, 1990)*

Norusis M J 1993 *SPSS for Windows Professional Statistics* release 6.0. (Chicago, IL: SPSS)

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

Raudys S and Pikelis V 1980 On dimensionality, sample size, classification error, and complexity of classification algorithm in pattern recognition *IEEE Trans. Pattern Anal. Machine Intell.* **2** 242–52

Sahiner B, Chan H P, Petrick N, Helvie M A, Adler D D and Goodsitt M M 1996a Classification of masses on mammograms using rubber-band straightening transform and feature analysis *Proc. SPIE* **2710** 44–50

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

Shen L, Rangayyan R M and Desautels J E L 1994 Application of shape analysis to mammographic calcifications *IEEE Trans. Med. Imaging* **13** 263–74

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

Thiele D L, Kimme-Smith C, Johnson T D, McCombs M and Bassett L W 1996 Using tissue texture surrounding calcification clusters to predict benign vs malignant outcomes *Med. Phys.* **23** 549–55

Wee W G, Moskowitz M, Chang N-C, Ting Y-C and Pemmeraju S 1975 Evaluation of mammographic calcifications using a computer program. *Radiology* **116** 717–20

Wei D, Chan H P, Helvie M A, Sahiner B, Petrick N, Adler D D and Goodsitt M M 1995a Classification of mass and normal breast tissue on digital mammograms: multiresolution texture analysis *Med. Phys.* **22** 1501–13

——1995b Multiresolution texture analysis for classification of mass and normal breast tissue on digital mammograms *Proc. SPIE* **2434** 606–11

Weiss S M and Kulilowski C A 1991 *Computer Systems that Learn* (San Mateo, CA: Morgan Kaufmann)

Wu Y, Freedman M T, Hasegawa A, Zuurbier R A, Lo S C B and Mun S K 1995 Classification of microcalcifications in radiographs of pathologic specimens for the diagnosis of breast cancer *Acad. Radiol.* **2** 199–204

Wu Y, Giger M L, Doi K, Vyborny C J, Schmidt R A and Metz C E 1993 Artificial neural networks in mammography: application to decision making in the diagnosis of breast cancer *Radiology* **187** 81–7