Computer-aided detection of mammographic microcalcifications: Pattern recognition with an artificial neural network

Heang-Ping Chan, Shih-Chung B. Lo,[a] Berkman Sahiner,
Kwok Leung Lam,[b] and Mark A. Helvie
Department of Radiology, University of Michigan, Ann Arbor, Michigan 48109

(Received 30 December 1994; accepted for publication 14 June 1995)

We are developing a computer program for automated detection of clustered microcalcifications on mammograms. In this study, we investigated the effectiveness of a signal classifier based on a convolution neural network (CNN) approach for improvement of the accuracy of the detection program. Fifty-two mammograms with clustered microcalcifications were selected from patient files. The clusters on the mammograms were ranked by experienced mammographers and divided into an obvious group, an average group, and a subtle group. The average and subtle groups were combined and randomly divided into two sets, each of which was used as training or test set alternately. The obvious group served as an additional independent test set. Regions of interest (ROIs) containing potential individual microcalcifications were first located on each mammogram by the automated detection program. The ROIs from one set of the mammograms were used to train CNNs of different configurations with a back-propagation method. The generalization capability of the trained CNNs was then examined by their accuracy of classifying the ROIs from the other set and from the obvious group. The classification accuracy of the CNNs for the ROIs was evaluated by receiver operating characteristic (ROC) analysis. It was found that CNNs of many different configurations can reach approximately the same performance level, with the area under the ROC curve (A_z) of 0.9. We incorporated a trained CNN into the detection program and evaluated the improvement of the detection accuracy by the CNN using free response ROC analysis. Our results indicated that, over a wide range of true-positive (TP) cluster detection rate, the CNN classifier could reduce the number of false-positive (FP) clusters per image by more than 70%. For the obvious cases, at a TP rate of 100%, the FP rate reduced from 0.35 cluster per image to 0.1 cluster per image. For the average and subtle cases, the detection accuracy improved from a TP rate of 87% at an FP rate of four clusters per image to a TP rate of 90% at an FP rate of 1.5 clusters per image.

Key words: mammography, microcalcification, computer-aided diagnosis, artificial neural network, receiver operating characteristic (ROC) analysis

I. INTRODUCTION

In the United States, breast cancer is the leading cause of death in women between 40 and 55 yr of age.[1] One out of eight women will develop breast cancer in their lifetime.[2] Studies have indicated that early detection and treatment improve the chances of survival for breast cancer patients. At present, mammography is the only proven method that can detect minimal breast cancers.[3–5] However, 10%–30% of the breast cancers that are visible on mammograms in retrospective studies are not detected due to various technical or human factors.[6–9] Double reading can reduce the miss rate on radiographic reading.[10] It has also been shown that computer-aided diagnosis (CAD), in which a computer alerts radiologists to suspicious locations on the images during mammographic reading, can improve the detection accuracy significantly.[11,12] CAD is thus a viable cost-effective alternative to double reading by radiologists.

One of the important indicators of the presence of breast cancers is clustered microcalcifications.[13] Clustered microcalcifications can be seen on mammograms in 30%–50% of breast cancers.[14–17] It is difficult to detect subtle microcalcifications because of the noisy mammographic background. A number of research groups have been developing CAD programs for the detection of microcalcifications. Chan *et al.*[11,18,19] demonstrated that a difference-image technique can effectively detect microcalcifications on digitized mammograms. Fam *et al.*[20] and Davies *et al.*[21] detected microcalcifications using conventional image processing techniques. Qian *et al.*[22] recently devised a tree-structure filter and wavelet transform for enhancement of microcalcifications to facilitate detection. Other groups extracted morphological features such as contrast, size, shape, and edge gradient of microcalcifications, and classified them with various feature classifiers.[23–31] Wu *et al.* scanned for suspected microcalcifications with the difference-image technique[18] then further classified true and false detections by an artificial neural network based on features extracted from their power spectra.[32] Similarly, Zhang *et al.*[33] used a shift-invariant neural network to reduce false-positive microcalcifications. The results reported in all these studies appear to be encouraging for the selected datasets.

In this study, we trained a convolution neural network (CNN) to recognize mammographic microcalcifications. The CNN was first developed for the detection of pulmonary nodules on chest radiographs.[34] This neural network is different from the commonly used back-propagation neural network in that its input is a region of interest (ROI) from the image instead of extracted image features. It is also different from the shift-invariant neural network used by Zhang *et al.*[33] in that the input ROI to the CNN includes an individual microcalcification instead of a cluster, and that the

 0094-2405/95/22(10)/1555/13/$6.00

output of the CNN is a decision score for determination of the presence of a microcalcification instead of a processed image ROI. Therefore, with our approach no further image processing techniques such as thresholding and region growing have to be applied to an output ROI to determine if a microcalcification is present. We have incorporated the trained CNN into our detection program and its effectiveness is evaluated by the improvement in the overall detection accuracy of the CAD program.

II. MATERIALS AND METHODS

A. Case selection

In this study, we used mammograms that contained clustered microcalcifications as case samples. The mammograms were selected from the patient files in the Department of Radiology at the University of Michigan Hospitals by experienced mammographers. The mammograms were acquired with a dedicated mammographic system with a 0.3 mm focal spot, molybdenum (Mo) anode and 0.03 mm Mo filter, and a 5:1 reciprocating grid. Kodak Min R/MRE mammographic screen/film system using extended cycle processing was employed as the image receptor. The presence of the clustered microcalcifications and the histology for each case had been verified by biopsy. The case samples included a mixture of benign and malignant cases. However, in this study, we concentrated on the detection rather than the classification of the malignant/benign nature of the microcalcifications.

Fifty-two mammograms were selected for this study. Each mammogram was ranked by the radiologist regarding the visibility of the cluster of microcalcifications on a rating scale of 1–5 (1=very obvious, 5=very subtle). The scale was established subjectively relative to the cases encountered in clinical practice in our hospitals. After ranking, we divided the 52 mammograms into three groups: the mammograms of ratings 1 and 2 were referred to as the obvious group ($N=14$), the mammograms of rating 3 as the average group ($N=16$), and the mammograms of ratings 4 and 5 as the subtle group ($N=22$). Although this classification was very subjective, it was an attempt to demonstrate the dependence of the performance of the CAD program on the database. We also attempted to describe quantitatively the physical characteristics of the microcalcifications on the digitized image and correlated them with the visual ratings. We extracted digitally, as discussed below, the contrast, the size, and the signal-to-noise ratio (SNR) of the individual microcalcifications. The mean and standard deviation (SD) of these physical characteristics of the microcalcifications in each group were compared.

B. Digitization of mammograms

All mammograms were digitized with a laser film scanner (LUMISYS DIS-1000), with both the sampling distance and the nominal spot size, and thus the pixel size, chosen to be 0.1 mm×0.1 mm.[35] The digitizer has a gray level resolution of 12 bits and an optical density (O.D.) range of 0–3.5. It was calibrated so that the O.D. on film was linearly proportional to output pixel values in the range of about 0.1 O.D. to 2.8 O.D. at 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 density was represented by high pixel values.

To establish a "truth" file with which the computer detection results could be compared, we determined the true locations of the individual microcalcifications on each mammogram manually. The digitized image was displayed on a workstation and the region containing the cluster of microcalcifications was enlarged to full resolution. Each individual microcalcification on the displayed image was identified carefully by comparison with the mammogram on film with a magnifier. The coordinates of the microcalcifications were then determined by a cursor and stored in the "truth" file. The same regional clustering procedure as that used in the detection program described below was applied to the "truth" file to determine the coordinates of the centroid of the clusters. These coordinates were used for scoring the detection of the clusters by the automated procedure.

It may be noted that the "truth" file thus determined may not be the absolute truth because of the difficulties and uncertainties in detecting subtle microcalcifications that are near the human visual threshold. However, this is the best available and practical method. Neither histologic analysis nor specimen radiographs can be used to identify individual microcalcifications seen on mammograms because of the very different geometry and image quality obtained with these techniques. Magnification mammograms are often not available since magnification is not performed for every case or for all views.

C. Extraction of signal characteristics

To describe quantitatively the physical characteristics of the microcalcifications on the digitized image, we have developed a signal extraction program to determine the size, contrast, SNR of the microcalcifications from an unprocessed image based on the coordinate of each individual microcalcification in the "truth" file.[36] In a 51×51 pixel ROI centered at each signal site, the structured background is estimated by polynomial curve fitting in the x and y directions. The fitted pixel values in the x and y directions at the same pixel are averaged. The process may be performed more than one time to reach a well-fitted smooth surface. The central $l \times l$ pixels in the region which contain the signal are excluded from the curve fitting and noise estimation. The size l is chosen to be a constant that is larger than the diameters of the microcalcifications of interest yet much smaller than 51 pixels. After subtraction of the structured background, the local root-mean-square (RMS) noise is calculated. A local threshold gray level is determined as the product of the RMS noise and an input SNR threshold. With a region growing technique, the signal region is then extracted as the connected pixels above the threshold around the manually identified signal location. The size of the microcalcification is estimated as the number of pixels in the signal region. The contrast is defined as the maximum pixel value in the signal region after subtracting the background. The SNR of the microcalcification is the ratio of the contrast to the local RMS

noise. The thresholded image of the microcalcifications superimposed on a background of constant pixel values can also be displayed for visual comparison.

D. Computerized detection of microcalcifications

We have developed a computer program that can automatically detect microcalcifications on mammograms. The program has been described in the literature.[11,18,35] Briefly, there are three major steps in the algorithm: preprocessing, segmentation, and classification. In the preprocessing step, an edge detector detects the breast boundary and divides the image into two regions, one internal and the other external to the breast. Signal detection is applied only to the region within the breast. A signal-enhancement filter (1×1 kernel) is employed to enhance the microcalcifications and a signal-suppression filter (box-rim filter with an 8×8 kernel of constant weights around the rim and a 4×4 central area of zero weights), to remove or suppress the microcalcifications and smooth the noise. Subtracting the two filtered images results in an SNR-enhanced image in which the low-frequency structured background is removed and the high-frequency noise is suppressed. This is also referred to as a difference-image technique.[11,18,19,35] When both the signal-enhancement filter and the signal-suppression filter are linear, as used in this study, the difference-image technique is equivalent to bandpass filtering. In the segmentation step, the program determines the gray level histogram of the preprocessed image within the breast region. A gray level thresholding technique is used to locate potential signal sites above a global threshold. The threshold is changed iteratively until the number of sites obtained falls within the chosen input maximum (4000) and minimum (3000) numbers. At each potential site, a locally adaptive gray level thresholding technique in combination with region growing is performed to determine the number of connected pixels above a local threshold, which is calculated as the product of the local RMS noise and an input SNR threshold. The signal characteristics to be used in the classification step, such as the size, maximum contrast, SNR, and its location, are obtained in this step. This locally adaptive thresholding technique is similar to the signal characteristic extraction technique described above, except that the procedure is performed on the SNR-enhanced image instead of the unprocessed image so that no curve fitting for background correction is necessary.

In the classification step, the previous computer program performs three tests to distinguish signals from noise or artifacts. A lower bound (two pixels) is imposed on the size to exclude signals below a certain size that are likely to be noise and an upper bound (80 pixels) is set to exclude signals greater than a certain size that are likely to be large benign calcifications. A contrast upper bound is also set to exclude potential signals that have a contrast higher than an input number (10) of SDs above the average contrast of all potential signals found with local thresholding. This criterion excludes the very high-contrast signals that are likely to be artifacts and large benign calcifications. A regional clustering procedure is then applied to the remaining signals; a signal is kept if the number of signals found within a neighborhood of a chosen input diameter (1 cm) around that signal is greater than an input minimum number. The remaining signals that are not found to be in the neighborhood of any potential clusters will be considered isolated noise points or calcifications and excluded. This clustering criterion is useful for reducing false positives, because true microcalcifications of clinical interest always appear in clusters on mammograms.[13–17] The specific parameters used in each step have been described previously.[11,19,35]

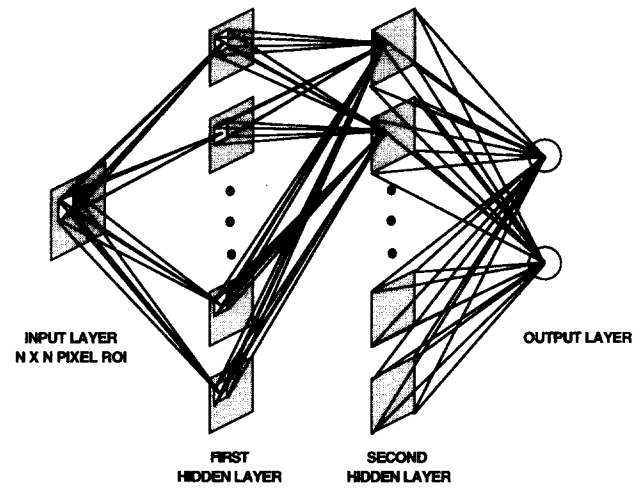
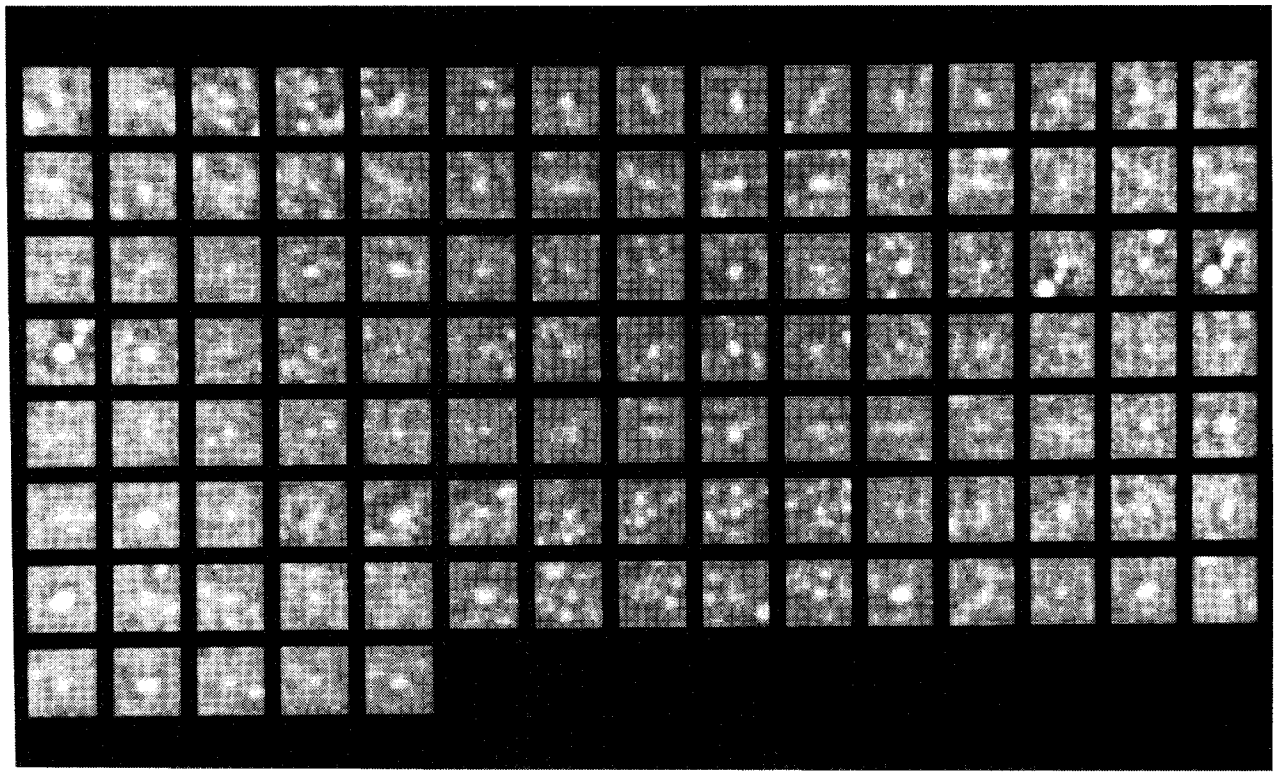


FIG. 1. Schematic diagram of the architecture of a convolution neural network (CNN). The input ROI size, the number of hidden layers, and the number of node groups in each layer are varied in this study.

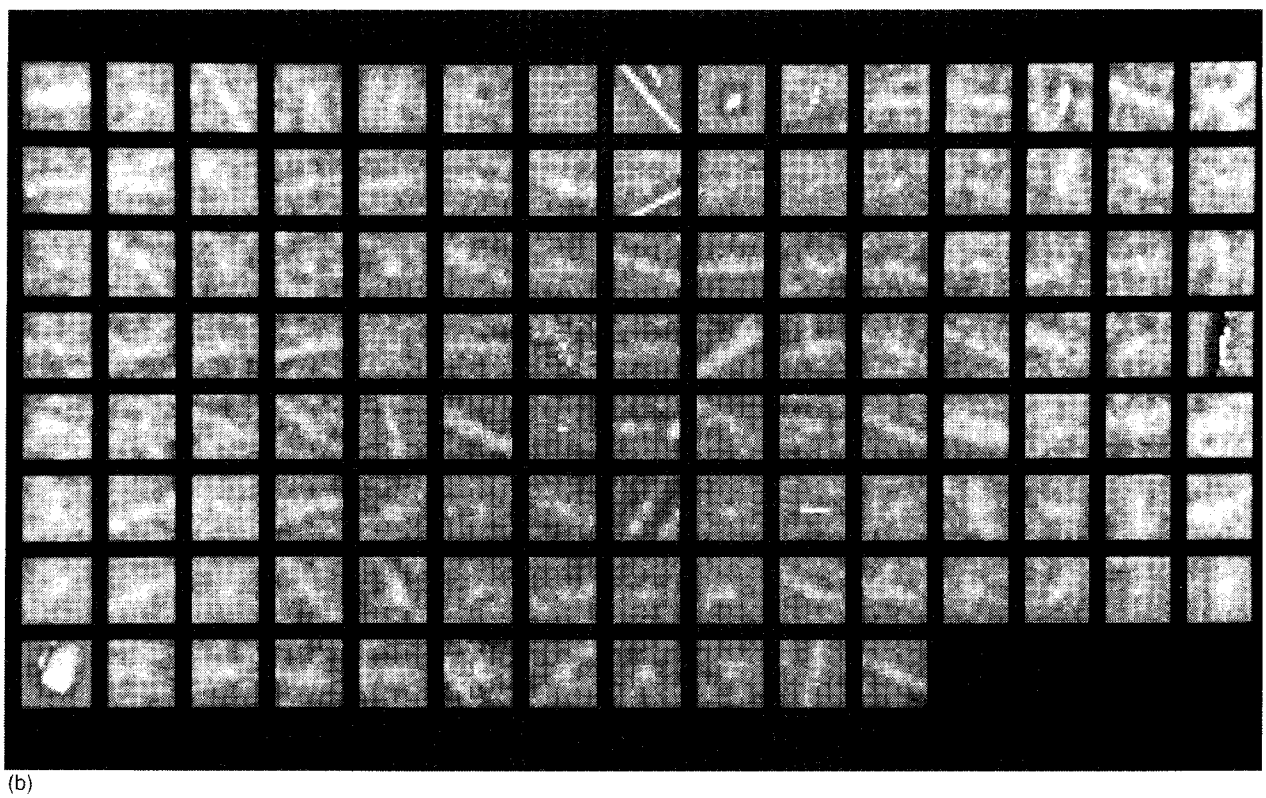
In this study, we investigated the effectiveness of a trained convolution neural network (CNN)[34] in discriminating false signals from true microcalcifications. The chosen CNN classifier was incorporated in the detection program. The potential signals that passed the size and contrast tests in the classification step were further screened by the CNN before being examined by the regional clustering criterion. The overall detection accuracy of microcalcifications with and without the CNN classifier could then be compared.

E. Convolution neural network classifier

The artificial neural network (ANN) used in this application is a convolution-type neural network.[34] The CNN can be considered a simplified version of the neocognitron[37] designed to simulate the human visual system. The general architecture of the CNN used in this study is shown in Fig. 1. It consists of an input layer, one to several hidden layers, and an output layer. The input layer of the CNN contains $N \times N$ input nodes, each of the input nodes is a sensor for an input pixel value in an $N \times N$-pixel ROI containing the normal or abnormal pattern to be recognized. In the hidden layers, the nodes are organized in groups and the groups between adjacent layers are interconnected by weights that are organized in kernels. Learning is constrained such that the kernel of weights connecting the kth group in the $(L-1)$th layer to the nth group in the Lth layer is invariant with nodes in the same groups. Forward signal propagation is thus similar to a spatially invariant convolution operation; the signals from the nodes in the lower layer are convolved with the weight kernel, and the resultant value of the convolution is collected

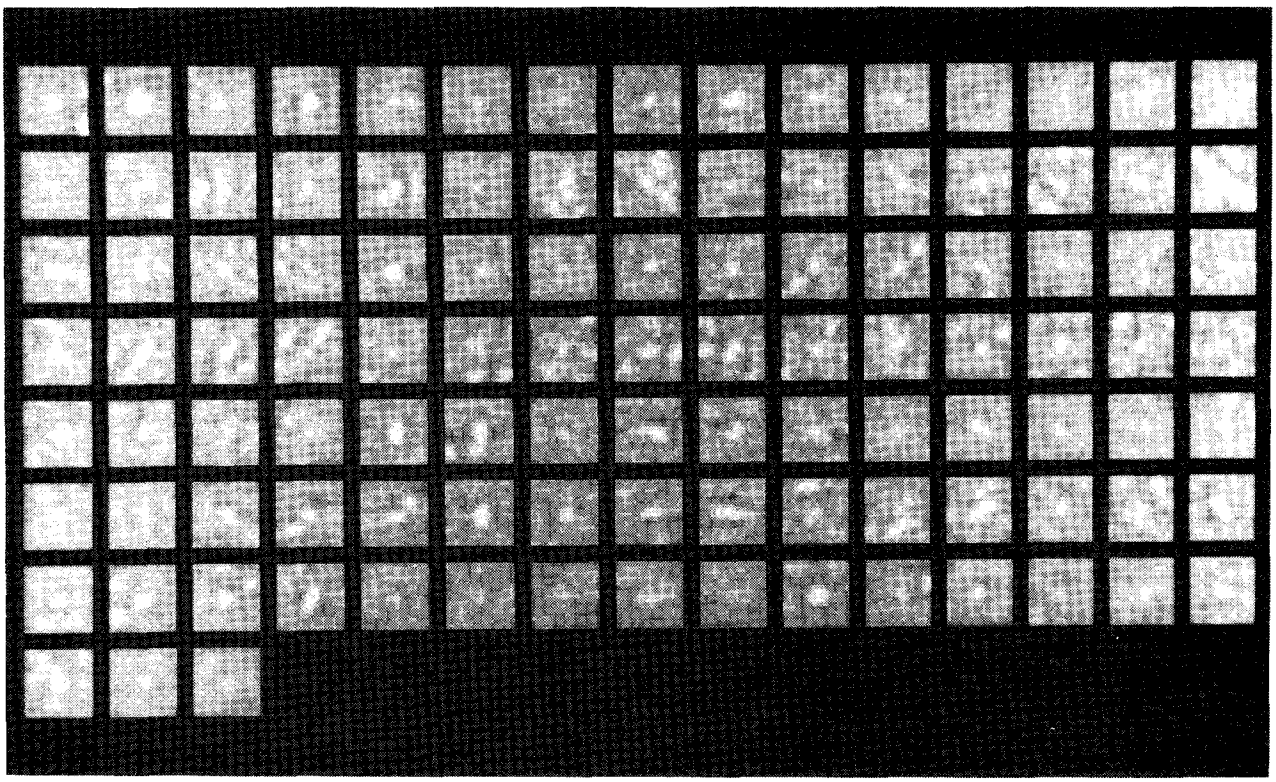


(a)

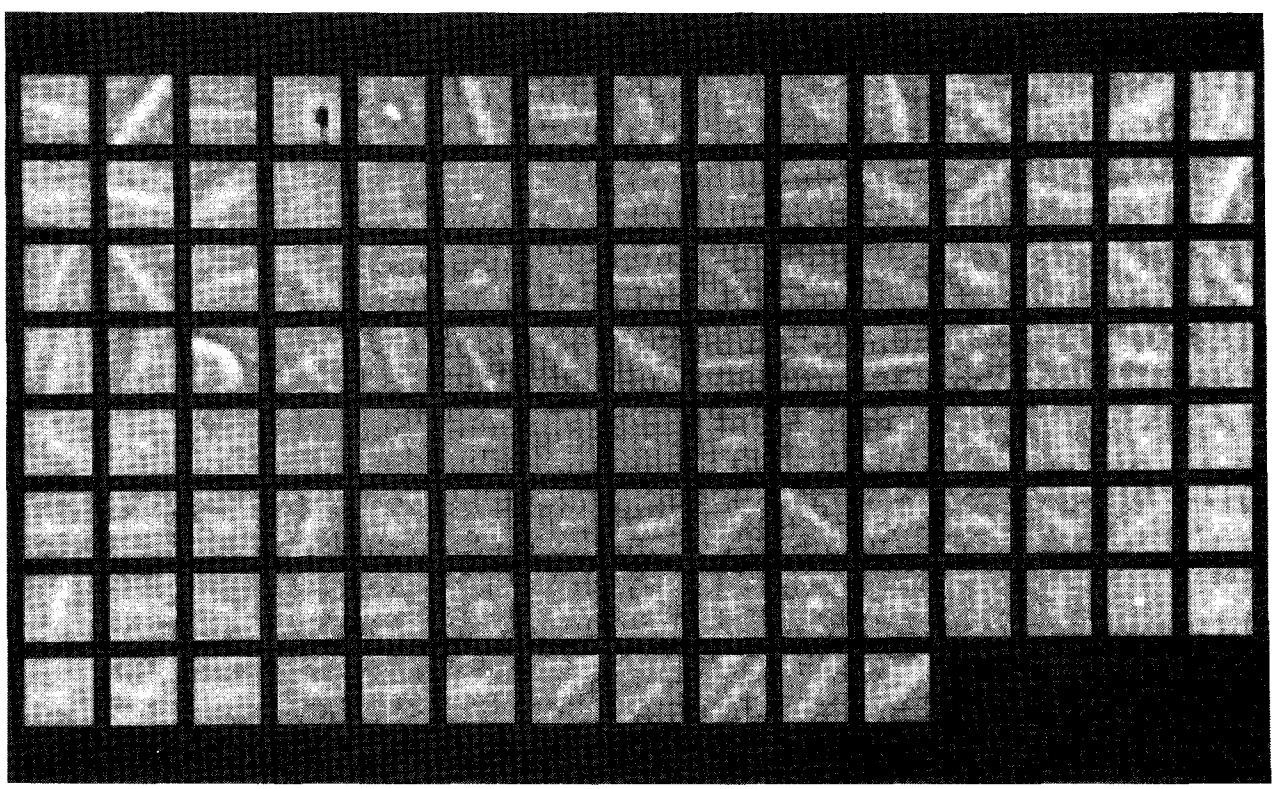


(b)

FIG. 2. The two groups of ROIs with true microcalcifications and false positives used for training of the CNNs in this study. Each of the ROI shown here contains 16×16 pixels (1.6 mm×1.6 mm). (a) ROIs in group 1 with true microcalcifications. (b) ROIs in group 1 with false positives. (c) ROIs in group 2 with true microcalcifications. (d) ROIs in group 2 with false positives.



(c)



(d)

FIG. 2 (*Continued.*)

into the corresponding node in the upper layer. This value is further processed by the node through an activation function and produces an output signal that will, in turn, be forward propagated to the subsequent layer in a similar manner. The convolution kernel incorporates the neighborhood information in the input image pattern and transfers the information to the receiving layers, thus providing the pattern recognition capability of the CNN. The activation function between two

layers is a sigmoidal function, and the signal at the Lth layer is obtained from the signal at the $(L-1)$th layer using the following relationship:

$$S_L((i,j);n) = \frac{1}{1+\exp\{-\Sigma_{\forall k\Rightarrow n}[w_L((i,j);k\Rightarrow n)*S_{L-1}((i,j);k)]\}}, \quad (1)$$

where $S_L((i,j);n)$ denotes the signal at node (i,j) in the nth group and Lth layer, $w_L((i,j);k\Rightarrow n)$ denotes the weight kernel connecting the kth group in the $(L-1)$th layer to the nth group in the Lth layer, * denotes the convolution operation, and the summation is over all groups k that are connected to group n. Note that the weight kernel for a given k and a given n is shift invariant, such that

$$w_L((i',j';i,j);k\Rightarrow n) = w_L((i'-i,j'-j);k\Rightarrow n), \quad (2)$$

where (i',j') denotes the node in the kth group and the $(L-1)$th layer. Because of the convolution operation, the useful matrix size of a node group in the Lth layer, $N_L\times N_L$, is reduced to $(N_{L-1}-K_{L-1}+1)\times(N_{L-1}-K_{L-1}+1)$, where $N_{L-1}\times N_{L-1}$ is the matrix size of a node group in the $(L-1)$th layer and $K_{L-1}\times K_{L-1}$ is the size of a weight kernel between the Lth layer and the $(L-1)$th layer.

In the output layer, there are n_{out} individual output nodes. Each output node is fully connected to all nodes in each group of the preceding hidden layer. The signal at the nth output node is given by Eq. (1), in which the weight matrix size is the same as the group size in the preceding layer and the output group size is 1×1.

F. Back-propagation training

The error back-propagation learning rule is used for supervised training of the CNN. The error function that is to be minimized by training is given by

$$\text{Error} = \frac{1}{2}\sum_{i=1}^{n_{\text{out}}}[S_{\text{in}}(i)-S_{Lo}(i)]^2, \quad (3)$$

where $S_{\text{in}}(i)$ is the input (or desired) value of a given training case at the ith node of the output layer, L_o, $S_{Lo}(i)$ is the network output signal of the case at that node, and n_{out} is the number of nodes in the output layer.

The conventional steepest descent delta rule for back-propagation training of a CNN can be written as

$$w_L((u,v);k\Rightarrow n)[t+1] = w_L((u,v);k\Rightarrow n)[t] + \eta\sum_{i,j}\delta_L((i,j);n) \times S_{L-1}((i+u,j+v);k), \quad (4)$$

where t is the number of iterations, η is the learning rate, and δ_L is the weight-update function given by

$$\delta_L((i,j);n) = S_L((i,j);n)[1-S_L((i,j);n)]Q_L((i,j);n), \quad (5)$$

where

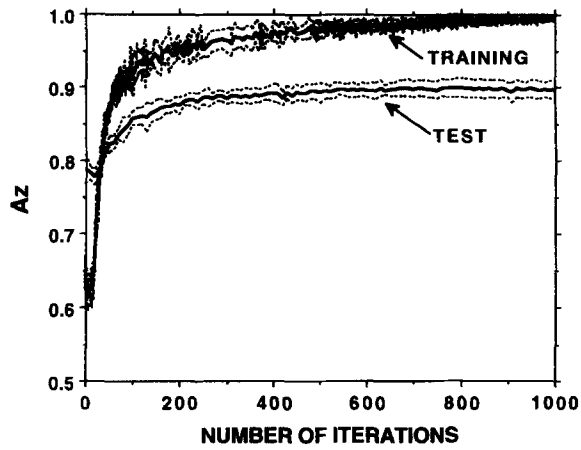


FIG. 3. Dependence of classification accuracy, A_z, on the number of iterations. The solid curves are the average A_z obtained from four repeated runs. The two dotted curves around each solid curve indicate the average±one SD of A_z estimated from the repeated runs. CNN configuration: 16×16 input nodes, first hidden layer: 12 node groups, second hidden layer: 12 node groups, each connected to 8 groups in the first hidden layer, two output nodes, weight kernels between input and first hidden layer: 5×5, weight kernels between first and second hidden layers: 3×3.

$$Q_L((i,j);n) = \sum_{u,v;\forall k\Rightarrow n} w_{L+1}((u,v);k\Rightarrow n) \times \delta_{L+1}((i-u,j-v);k\Rightarrow n). \quad (6)$$

At the output layer, the weight is updated as

$$w_{Lo}((i,j);k\Rightarrow n)[t+1] = w_{Lo}((i,j);k\Rightarrow n)[t] + \eta\delta_{Lo}(k)S_{Lo-1}((i,j);n), \quad (7)$$

where

$$\delta_{Lo}(k) = S_{Lo}(k)[1-S_{Lo}(k)][S_{\text{in}}(k)-S_{Lo}(k)]. \quad (8)$$

Training may be terminated at a selected level of total error, which is the sum of the error for an individual case [Eq. (3)] over all cases in the training set, a selected level of classification accuracy (A_z) as defined below, or a preset number of iterations. In this study, we used the total error as the termination criterion. The total error allowed at termination was chosen to be low enough so that the test A_z could reach a plateau, as demonstrated in Fig. 3.

In our application, all weights in the CNN were initialized to be between −0.5 to +0.5 using a random number generator with a different seed in each training run and normalized by the number of weights in the exponential factor of the sigmoidal activation function [Eq. (1)]. An $N\times N$-pixel region centered at a potential site that passes the size and contrast tests formed the input ROI to the CNN. For a given input SNR threshold, the program would identify a number of potential signals. A low SNR threshold corresponded to a lax criterion with a large number of false-positive (FP) signals. A high SNR threshold corresponded to a stringent criterion with a small number of FP signals and a loss in true-positive (TP) signals. For training the CNN, we arbitrarily divided the 38 mammograms in the average and subtle groups into two subgroups. When the ROIs obtained from one subgroup were used for training, the trained CNN would

TABLE I. Physical characteristics of microcalcifications extracted with an SNR threshold of 2.0 from the three groups of unfiltered mammograms.

Image group	No. of images	No. of μcalc.	Mean no. of μcalc/image	Size (pixels)		Contrast (pixel value)		SNR	
				Mean	Std. dev.	Mean	Std. dev.	Mean	Std. dev.
Ratings 1,2	14	213	15	12.3	11.4	183.4	84.2	5.8	2.6
Rating 3	16	162	10	13.4	12.5	164.6	82.8	5.5	2.6
Ratings 4,5	22	270	12	9.0	9.4	143.9	87.0	4.6	2.2

be applied to the second subgroup for testing, and vice versa. We chose one of the SNR thresholds that yielded a moderate number of FPs and a sufficiently large number of TPs for segmenting the training ROIs. Because the number of FPs were still a few times more than the number of TPs, a subset of FPs with approximately the same number as the TPs were randomly chosen for the training set. It should be noted that the chosen SNR threshold level was not critical as long as the numbers of FP and TP were sufficiently large to provide the variety of ROI patterns for training the CNN. The ROIs obtained by using a high SNR threshold were generally a subset of those obtained by using a low SNR threshold. A chosen ROI input to the CNN was obtained from the SNR-enhanced image. The gray level values of the pixels in the ROI were thus independent of the SNR threshold at which it was chosen, and all ROIs had the same average background pixel value.

The shape of the microcalcifications in the breast parenchyma could be considered randomly oriented if we considered all possible locations of the microcalcifications in the breast and all mammographic views. To increase the variability of the training group, eight input ROIs to the CNN were generated from each ROI by rotating the ROI and its mirror image to 0°, 90°, 180°, and 270°. Each training cycle thus included training of the complete set of training ROIs with the eight orientations. The input order of the training ROIs was randomized with a different random number sequence in each run. A test ROI would be rotated also in the eight orientations, and the average output value of the eight rotated ROIs was taken to be the output value of that test ROI. During training, the desired output of an ROI with microcalcification was set to 1 and that of an ROI without microcalcification was set to 0.

We investigated the dependence of the classification accuracy of positive and negative ROIs on the CNN configurations. Because of the computational requirements in training the CNNs, we did not exhaustively study every possible combination of parameters. The range of parameters that we studied and the corresponding results are tabulated in Table III. CNNs with one and two hidden layers were examined. The number of node groups in the hidden layers was varied from 4 to 12. In most of the two-hidden-layer CNNs, the number of groups was kept the same for both layers. Combinations of 12 groups in the first hidden layer and 4, 8, or 12 groups in the second hidden layer were also studied. All node groups in the two hidden layers are fully connected in these configurations. Additionally, a 12 group–12 group combination in which every 3 of the 12 groups in the second hidden layer were connected to the same 8 selected groups in the first hidden layer was examined.[34] For comparison, a CNN with 8 groups in the first hidden layer and 12 groups in the second hidden layer with full connections was included. We also evaluated the classification accuracy for two combinations of weight kernel sizes; one had a kernel size of 5×5 in the first hidden layer and 3×3 in the second hidden layer and the other had a kernel size of 7×7 in the first hidden layer and 5×5 in the second hidden layer. We did not investigate larger kernel sizes because the sizes of the microcalcifications of interest were generally much smaller than 7×7 pixels and because computation time increased rapidly with kernel size. The input ROI size was adjusted so that the size of the node groups in the last hidden layer was 10×10 for both combinations of kernel sizes. The output nodes were always fully connected to every node group in the last hidden layer with a 10×10 kernel, as shown in Fig. 1.

TABLE II. Number of ROIs with microcalcifications and false positives for training of the CNN.

	Number of ROIs			
	Group 1 (G1)	with rotation	Group 2 (G2)	with rotation
Microcalcifications	110	880	108	864
False positives	116	928	116	928

The classification accuracy of the CNN during training was monitored by receiver operating characteristic (ROC) analysis[38] of the output values from the CNN. After each iteration, or epoch, with the training set was completed, the classification performance with the current weights for all training cases would be determined by inputting the training cases into the CNN as a consistency verification procedure. The distributions of the output values for the positive ROIs and the negative ROIs would be input into the LABROC1 program,[39] which assumes binormal distributions of the decision variable for the normal and abnormal cases and fits an ROC curve based on maximum likelihood estimation. The ROC curve represents the relationship between the true-positive fraction (TPF) and the false-positive fraction (FPF) as the decision threshold varies. The LABROC1 program provides the area under the fitted ROC curve, A_z, and an estimate of the SD of A_z. A_z is used as an index of classification accuracy. The dependence of A_z on the number of iterations was monitored during training. For every ten training iterations, the trained CNN was applied to the other independent set of ROIs to test its generalization capability. The dependence of the test A_z on the number of iterations was also examined.

TABLE III. Dependence of test results, in terms of the area under the ROC curve (A_z), on the configuration of CNN for classification of microcalcifications.

Input ROI size (pixels)		16×16			20×20		
Kernel size (first hidden layer)		5×5			7×7		
Kernel size (second hidden layer)		3×3			5×5		
No. of groups in hidden layer		No. of output nodes			No. of output nodes		
First	Second	1	2	2	1	2	2
		A_z Train: G1 Test: G2	A_z Train: G1 Test: G2	A_z Train: G2 Test: G1	A_z Train: G1 Test: G2	A_z Train: G1 Test: G2	A_z Train: G2 Test: G1
4	4	0.86	0.86	0.86	0.90	0.87	···
6	6	0.88	0.88	0.86	0.89	0.90	0.89
8	8	0.88	0.88	0.88	0.89	0.90	0.89
10	10	0.89	0.89	0.88	0.91	0.90	0.89
12	12	...[a]	0.91	0.89	···	0.91	0.89
12	4	···	0.88	0.86	···	0.90	0.90
12	8	0.89	0.90	0.89	0.90	0.90	0.90
8	12	···	0.89	0.89	···	0.88	0.90
12(8)[b]	12	0.90	0.90	0.90	0.91	0.90	0.89
One hidden layer							
4		0.87	0.85	···	0.83	0.85	···
8		0.85	0.87	···	0.86	0.85	···
12		0.87	0.86	···	0.86	0.86	···

[a]The CNN configuration was not tested if there is no entry.
[b]Eight node groups in the first hidden layer are selectively connected to the 12 node groups in the second hidden layer. The A_z values for this CNN are the averages of four runs shown in Table IV.

G. Analysis of detection accuracy

After passing the size and contrast criteria, being screened by the trained CNN, and passing the regional clustering criterion, the detected individual microcalcifications and clusters would be compared with the "truth" file of the input image. The number of TP and FP microcalcifications and the number of TP and FP clusters were scored. A detected signal was scored as a TP microcalcification if it was within 0.5 mm from a true microcalcification in the "truth" file. A detected cluster was scored as a TP if its centroid coordinate was within a cluster radius (5 mm) from the centroid of a true cluster and at least two of its member microcalcifications were scored as TP. Once a true microcalcification or cluster was matched to a detected microcalcification or cluster, it would be eliminated from further matching. Any detected microcalcifications or clusters that did not match to a true microcalcification or cluster were scored as FPs. The tradeoff between the TP and FP detection rates by the computer program was evaluated by the free-response receiver operating characteristic (FROC) analysis[40] by varying the input SNR threshold. A low SNR threshold corresponded to a lax criterion with a large number of FP clusters. A high SNR threshold corresponded to a stringent criterion with a small number of FP clusters and a loss in TP clusters. The detection accuracy of the computer program with and without the CNN classifier could then be assessed by comparison of the FROC curves.

III. RESULTS

Using the signal extraction program described in Sec. II, the size, contrast, and SNR of the true microcalcifications as indicated in the "truth" file for each of the three groups of mammograms were determined at several SNR thresholds. We examined the extracted signals in the thresholded images and compared visually the extracted signals with those in the original images. When the SNR threshold was too low, the signals merged with one another or with noise in the background. The extracted signals did not represent the true signal size or shape. When the SNR threshold was too high, many subtle microcalcifications were not extracted. The extracted signals appeared to be smaller than those in the original images because only a few pixels in a microcalcification were higher than the threshold. It was determined subjectively that an SNR threshold of 2.0 was a compromise with which the extracted signals were similar in size and shape to those in the original images. At this SNR threshold, an average of about 85% of the microcalcifications were extracted. The other 15% of the microcalcifications could not be extracted at this threshold because their pixel values were lower than the local gray level threshold.

Table I shows the mean and SD of the contrast, size, SNR of the microcalcifications extracted at an SNR threshold of 2.0 for each of the three groups. Note that the "size" of an extracted microcalcification depends on the SNR threshold used because it may merge with an adjacent noise or signal pixels, as discussed previously. The contrast is relatively independent of the SNR threshold, since it depends only on the maximum pixel value in the signal region. We have plotted the histograms of the contrast, size, and SNR of the extracted microcalcifications and found a large overlap in the physical characteristics of the microcalcifications in the three groups of mammograms. As can be seen in Table I, the visual rank-

ing generally correlates with the mean contrast and mean SNR of the microcalcifications. However, the mean number of microcalcifications in the subtle group is larger than that of the average group. These observations indicate that the visibility of a microcalcification cluster is more strongly affected by the contrast and SNR than by the number of microcalcifications in the cluster. This is consistent with the experience of radiologists in visual detection of microcalcifications. The data in Table I should provide more objective information than the visibility ratings in the description of the degree of subtlety for each group of microcalcifications. The quantitative characterization can facilitate comparison of the performance of CAD algorithms in different datasets if a similar signal extraction method and criteria are used in calculation of the data.

Table II shows the number of ROIs with true microcalcifications and false signals used for training of the CNN. Each group of ROIs was detected with the automated algorithm at an SNR threshold of 3.4 from 19 SNR-enhanced images. At this threshold, the average TP rate was 94% at an average FP rate of 7.5 clusters per image. This point was outside the range of the FP rates plotted in Fig. 7. There were no overlapping cases in the two groups. The extracted ROIs of 16×16 pixels are displayed in Figs. 2(a)–2(d). It can be seen that a large number of the FPs extracted by the CAD program was caused by high-frequency structures such as fibrous strands, film artifacts, and noise. Only one-fifth of the FP ROIs were included in the training groups in order to match approximately the number of ROIs with true microcalcifications. With the rotation method, over 800 positive and over 900 negative ROIs were generated in each training set. When one group was used for training, the displayed ROIs in the other group, together with the other four-fifths of the FP ROIs from the same set of images, were used as the test set. The signal of interest was centered at the ROI. The average background gray level was the same for all ROIs after the SNR-enhancement filtering.

The dependence of the classification accuracy on CNN configuration and training set is shown in Table III. The SDs of the A_z as determined by the LABROC1 program ranged from 0.01 to 0.02. The classification accuracy during training generally reached an A_z of 0.99 or greater under all conditions studied. The test results exhibited some variations, as can be seen from Table III. The CNNs with one hidden layer are inferior to the CNNs with two hidden layers. The performance of the CNNs with two hidden layers does not depend strongly on the configuration when the total number of weights in the CNN is large. There is a slight trend, with some minor variations, that the A_z increases as the number of node groups increases. This trend is more systematic for the CNNs with small weight kernels. There is also a trend that, for the same CNN configuration, the test A_z is larger when G1 is used for training than when G2 is used. The difference in the test A_z values between the two training/test group combinations, averaged over all two-hidden-layer, two-output-node CNNs studied, is only about 0.01. This difference, however, is statistically significant at a two-tailed p level of 0.005.

We also compared the difference in performance between CNN with one- and two-output nodes. The two output values from the two-output CNNs were found to be complementary to each other, i.e., for a given case, if the output of one node was x, the output from the other node was very close to $(1-x)$. This outcome is expected because the desired output values for the true and false signals were set to be 1 and 0, respectively, as described above. Therefore, the output from one node was sufficient for the classification task, and the ROC curve could be constructed from either of the output nodes. The difference in the A_z values between the one- and two-output configurations, averaged over the two-hidden-layer CNNs, is 0.001. The difference is not statistically significant ($p=0.68$).

TABLE IV. Reproducibility of test results for two CNNs. The A_z shown is the maximum value reached for a given run.

Input ROI size (pixels)	16×16		20×20	
Kernel size				
First hidden layer:	5×5		7×7	
Second hidden layer:	3×3		5×5	
No. of groups				
First hidden layer:	12(8)		8	
Second hidden layer:	12		8	
Repeated run	Train: G1 Test: G2	Train: G2 Test: G1	Train: G1 Test: G2	Train: G2 Test: G1
	A_z		A_z	
1	0.91	0.91	0.91	0.90
2	0.90	0.88	0.90	0.88
3	0.89	0.89	0.90	0.89
4	0.90	0.91	0.90	0.88
Mean	0.90	0.90	0.90	0.89
std. dev.	0.01	0.01	0.01	0.01

To study the variability in the classification accuracy due to the initialization condition of the weights and training, the training and testing of two selected CNN configurations were repeated four times for each of the two training/test group combinations. The CNN configurations and the results are listed in Table IV. The SD of A_z is estimated from the repeated runs to be 0.01 in each case, and the maximum difference in A_z for the four runs is 0.03. For a given CNN configuration, the mean A_z values for the two training/test combinations agree within 0.01. Figure 3 shows the dependence of A_z for training and testing, averaged over four runs, on the number of iterations for one of the CNNs. The SDs of A_z estimated from the repeated runs are also plotted for the training and the test curves. Both the mean A_z values and SDs stabilize after some large fluctuations in the initial iterations. The shapes of the A_z curves are typical of the conditions included in this study, although the rate of convergence varies with CNN configurations. The curves increase rapidly initially then plateau off and gradually approach its maximum level. The convergence of the CNN training can also be observed from the dependence of the total error on the number of iterations, as shown in Fig. 4. The magnitude of the error depends on the number of output nodes and the number of input cases. However, the trend of the curve is typical among the CNNs studied. It shows a steep descent initially

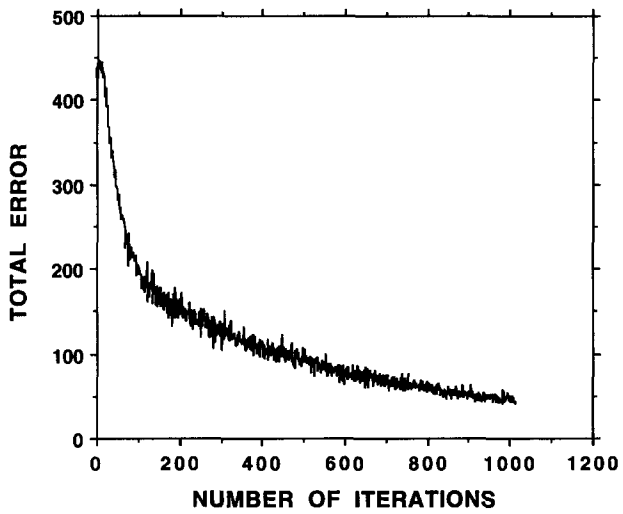


FIG. 4. Dependence of total error of the CNN output on the number of iterations. The CNN configuration is the same as that in Fig. 3. Training group G2 was used.

then gradually levels off at large number of iterations.

The training of each CNN with each group of training cases produces a set of weights at each iteration. Many of the CNN configurations reach approximately the same level of performance (Table III) and may be used as a classifier in the microcalcification detection program. We selected one of the trained CNNs shown in Table IV (2 hidden layers, each with 12 node groups, every 3 node groups in the second hidden layer selectively connected to 8 node groups in the first hidden layer, weight kernels sizes of 5×5 and 3×3) to demonstrate the effect of the CNN classifier on detection accuracy. A weight set trained with the G1 group was used to test the classification accuracy for the G2 group and another set trained with the G2 group was used to test the classification accuracy for the G1 group. The weights were obtained from one of the iterations when the plateau of A_z was reached. The ROC curves for classification of the test groups of ROIs

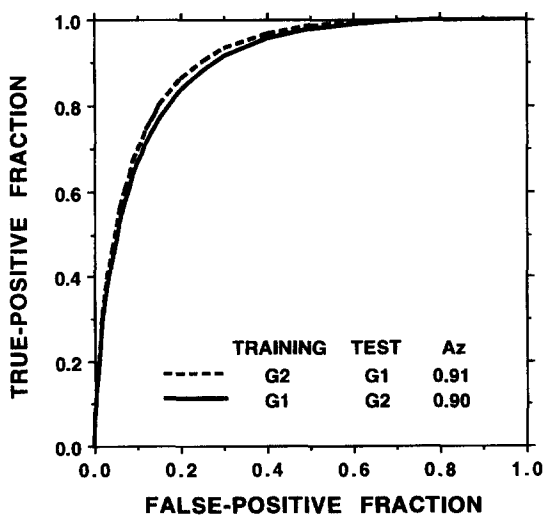


FIG. 5. The ROC curves obtained with the test ROI groups. The CNN configuration is the same as that in Fig. 3.

using the trained CNNs are shown in Fig. 5. The A_z values of the curves are 0.91 and 0.90, which correspond to the best performance obtained with the CNNs tabulated in Table III.

We incorporated the trained CNN into our microcalcification detection program as described previously, and the overall improvement in the detection accuracy was evaluated. For any SNR threshold, each extracted signal that passed the size and contrast criteria was input into the CNN. The set of weights obtained from training with G1 was used for the 19 mammograms from which the G2 ROIs were extracted, and vice versa. For the group of obvious mammograms, either set of weights could be used because the obvious cases were not used for training or testing. The performance of the trained CNNs on the obvious cases was thus an additional independent test for the classifiers. In this application, a constant decision threshold was set for the CNN output value of any input ROI to determine if the ROI was normal or abnormal. To select the appropriate decision threshold for the output value from the CNN, the dependence of the FROC curve on the decision threshold was evaluated. This corresponded to varying the operating point along the ROC curve (Fig. 5) of the classifier. The FROC curves for the three sets of mammograms were plotted in Figs. 6(a)–6(c). The data points along each FROC curve were obtained by varying the SNR thresholds from 3.0 to 5.2. Some of the data points were not plotted if they were outside the range of the graph. A curve without the CNN (decision threshold=0), and two curves with CNN at decision thresholds of 0.5 and 0.8, respectively, were plotted. For a given TP rate, the number of FP clusters decreased as the CNN threshold increased from 0.1 to 0.8. When the CNN threshold was further increased to 0.9, we observed a decrease in the TP rate for a given FP rate for subtle cases, indicating that many of the ROIs with subtle microcalcifications were misclassified with the high CNN threshold. At a CNN threshold of 0.8, the TP rate was 100% at an FP rate of less than 0.1 cluster per image for the obvious cases. For the cases that were ranked average subtle by radiologists, the TP rate was about 93% at an FP rate of one cluster per image. For the subtle cases, the TP rate was 87% at an FP rate of 1.5 clusters per image.

The degree of subtlety of the clustered microcalcifications in the cases ranked from 3 to 5 is similar to that of the cases used in our previous observer performance study.[11] The likelihood that these microcalcifications may be missed is not negligible, and thus it is of particular interest for CAD applications. The average improvement in the detection accuracy for these cases is estimated by comparison of the FROC curves without and with the CNN classifier for all cases ranked 3–5. The FROC curves are shown in Fig. 7. The TP rate improves from about 87% at an FP rate of 4 clusters per image without the CNN classifier to 90% at an FP rate of about 1.5 clusters per image, with the CNN classifier at a decision threshold of 0.8.

IV. DISCUSSION

The computational cost for training a CNN is high. The computational cost per iteration increases as the numbers of nodes and weights increase. However, it was observed that the rate of convergence increased as the number of nodes

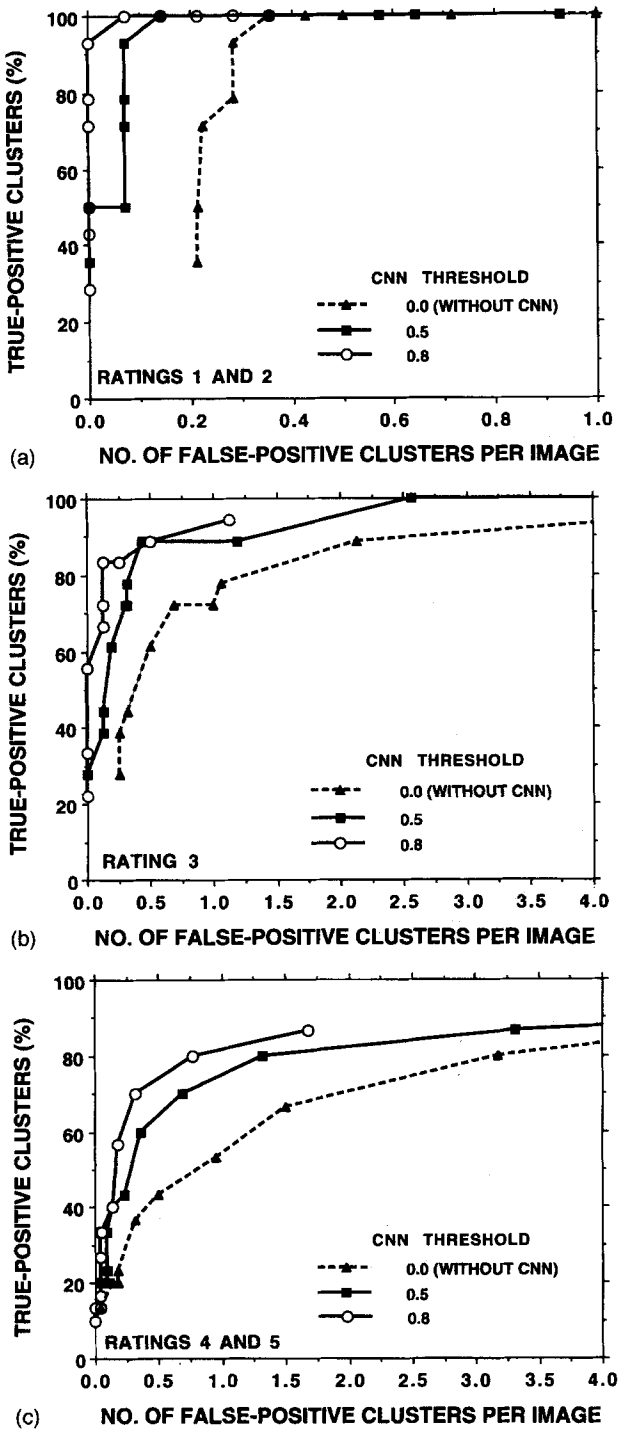


FIG. 6. Comparison of FROC curves for detection of clustered microcalcifications without and with the CNN classifier. The curve without CNN is equivalent to that with the decision threshold of the CNN set to 0. The FROC curves with the decision threshold of the CNN set to 0.5 and 0.8 are plotted for comparison. (a) Mammograms with obvious microcalcifications. (b) Mammograms with average subtle microcalcifications. (c) Mammograms with subtle microcalcifications. The CNN configuration is the same as that in Fig. 3.

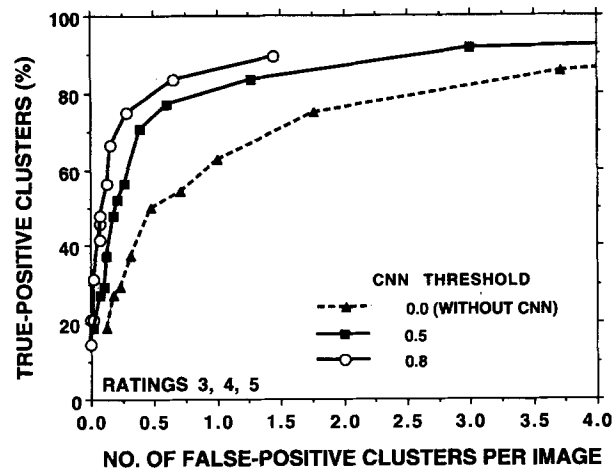


FIG. 7. Comparison of FROC curves for detection of clustered microcalcifications without and with the CNN classifier. The overall detection accuracy for the average and subtle groups of microcalcifications are compared. The CNN configuration is the same as that in Fig. 3.

increased. For example, for CNNs of the same configuration, except for a difference in the number of output nodes, the two-output-node CNN reached the maximum A_z with a smaller number of iterations than the corresponding one-output-node CNN. This trend is more obvious for the CNNs with fewer node groups in the hidden layers. Similarly, the convergence rate increases until the number of node groups in the hidden layers increases to about 10 for CNNs with two-output nodes.

The convergence rate saturates sooner for the CNNs with a larger kernel size. Therefore, the overall training cost of CNNs with complicated configurations may not be higher than those with simpler configurations. We could not perform an exact comparison of the computation time for different CNN configurations because we had to make use of all available workstations that had different CPU speeds and different memory capacities to train the CNNs. It may be noted that the computational cost with a complicated CNN configuration is higher than that of a simple one when it is incorporated in the microcalcification detection program for classification of test cases.

We have attempted to apply the FROCFIT curve fitting program[41] to the FROC curves in this study, but failed to obtain well-fitted curves. This may be caused by the fact that the FROCFIT was developed on the basis of several assumptions, which may not be satisfied for our detection task.[35] We therefore could not arrive at a single value such as the A_1[41] as the performance index for comparison of the different conditions. The generalization capability of the CNN can be observed from the effectiveness of the trained CNN in reducing FPs in the additional independent test group of mammograms of ratings 1 and 2 [Fig. 6(a)]. At a CNN threshold of 0.8, the FP clusters were reduced to zero for almost all TP rates below 100%. Because there is no established method to test the statistical significance of the difference in two FROC curves, we performed t tests on the image-specific paired FP values between the without-CNN and with-CNN (threshold $=0.8$) results at corresponding TP rates, in an effort to estimate the significance of their differences. The p values ranged from 0.04 to 0.08. Although the improvement in the FP rates was very consistent over the entire range of TP rates, as shown in Fig. 6(a), the level of significance for

individual TP rates was not high, probably because the FP rates without CNN were already very low.

The performance of the trained CNN can also be observed from the effective reduction of FPs at different SNR thresholds in the test group of mammograms of ratings 3–5. As shown in Fig. 7, for a given TP rate, the CNN reduced the FP clusters by more than 70% with a CNN threshold of 0.8. We again performed t tests on the image-specific paired FP values at corresponding TP rates. For TP rates between about 20%–75% the p values of the differences between the FP rates without CNN and with CNN (threshold=0.8) ranged from 0.06 to 0.0002. We also performed t tests on the paired TP rates at corresponding FP rates. For FP rates between about 0.1 to about 0.7 clusters per image, all p values of the differences between the TP rates without CNN and with CNN (threshold=0.8) were less than 0.001.

The FROC curves presented here were obtained by varying the SNR threshold in the local gray level thresholding process. The CNN classifier was implemented so that a constant decision threshold for its output value was used to classify ROIs with and without microcalcifications obtained at any SNR threshold. Alternatively, we can select a relatively low SNR threshold that produces a large number of FPs and vary the decision threshold for the output of the CNN classifier, thereby generating pairs of TP and corresponding FP values along an FROC curve. We have studied this approach by using SNR thresholds from 3.0 to 5.2, from each of which an FROC curve was generated by varying the CNN threshold from 0.1 to 0.9. It was observed that the FROC curves obtained with this alternative method were lower than the FROC curve with CNN (threshold=0.8) plotted in Fig. 7. On each of these alternative FROC curves, the data point at a CNN threshold of 0.8 coincided with the data point on the FROC curve with CNN (threshold=0.8) shown in Fig. 7, because they are the data points with the same SNR and CNN thresholds. Other data points on the alternative FROC curves are either comparable to or lower than the FROC curve in Fig. 7, with a few exceptions in the range of very low TP and FP rates.

The goal of this study is to evaluate the feasibility of training a CNN to distinguish FP signals from true microcalcifications obtained from our automated detection program. Although a small dataset was used and sample biases may exist, the effectiveness of the method as one of the steps in the classification process was demonstrated by the relative improvement in the detection accuracy. In the field of CAD, it is known that different detection algorithms or even different human observers may generate FPs of different characteristics. Before the CNN classifier is to be incorporated into a CAD program for clinical implementation, it is important to train the classifier using true and false microcalcifications obtained from the specific application. The training dataset should also be large enough to ensure that the patient population is adequately represented and that the performance of the classifier can be generalized.

V. CONCLUSION

We have developed a computer program for automated detection of clustered microcalcifications on mammograms for CAD applications. In this study, we investigated the effectiveness of a new signal classifier based on artificial neural network methodology for improvement of the detection accuracy of the CAD program. The CNN classifier was trained to recognize individual microcalcifications and incorporated as one of the signal classification steps. It was found that the CNN classifier can achieve a classification accuracy, expressed in terms of the A_z index with ROC analysis, of 0.9. It reduced the average FP rates by more than 70% at all TP rates on mammograms with subtle to obvious microcalcifications. Although the number of cases used in this study is limited, the improvement is consistent and statistically significant. This study demonstrates that a CNN can be trained to recognize mammographic microcalcifications and is effective in reducing FP detections in CAD applications.

ACKNOWLEDGMENTS

This work is supported by USPHS Grant No. CA 48129 and U.S. Army Grant No. DAMD 17-93-J-3007 (through subgrant No. GU RX 4300-803UM from Georgetown University). The content of this publication does not necessarily reflect the position of Georgetown 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, Ph.D., for the LABROC1 programs, to Dev Chakraborty, Ph.D., for the FROCFIT program, and to Diane Williams for secretarial assistance.

[a]Radiology Department, Imaging Physics Division, Georgetown University, Washington, DC 20007.
[b]Department of Radiation Oncology, University of Michigan.

[1]*National Center for Health Statistics. Vital Statistics of the United States, 1987. Vol. 2. Mortality. Part A*, DHHS Publication No. (PHS) 90-1101 (Government Printing Office, Washington, DC, 1990).
[2]J. R. Harris, M. E. Lippman, U. Veronesi, and W. Willett, "Breast cancer," N. Engl. J. Med. **327**, 319–328 (1992).
[3]M. Moskowitz, "Breast cancer: Age-specific growth rates and screening strategies," Radiology **161**, 37–41 (1986).
[4]M. Moskowitz, "Benefit and risk," in: *Breast Cancer Detection: Mammography and Other Methods in Breast Imaging*, edited by L. W. Bassett and R. H. Gold, 2nd ed. (Grune and Stratton, New York, 1987).
[5]H. Seidman, S. K. Gelb, E. Silverberg, N. LaVerda, and J. A. Lubera, "Survival experience in the breast cancer detection demonstration project," Cancer J. Clin. **37**, 258–290 (1987).
[6]J. E. Martin, M. Moskowitz, and J. R. Milbrath, "Breast cancer missed by mammography," Am. J. Roentgenol. **132**, 737–739 (1979).
[7]M. G. Wallis, M. T. Walsh, and J. R. Lee, "A review of false negative mammography in a symptomatic population," Clin. Radiol. **44**, 13–15 (1991).
[8]R. E. Bird, T. W. Wallace, and B. C. Yankaskas, "Analysis of cancers missed at screening mammography," Radiology **184**, 613–617 (1992).
[9]J. A. Harvey, L. L. Fajardo, and C. A. Innis, "Previous mammograms in patients with impalpable breast carcinomas: Retrospective vs blinded interpretation," Am. J. Roentgenol. **161**, 1167–1172 (1993).
[10]E. L. Thurfjell, K. A. Lernevall, and A. A. S. Taube, "Benefit of independent double reading in a population-based mammography screening program," Radiology **191**, 241–244 (1994).
[11]H. P. Chan, K. Doi, C. J. Vyborny, R. A. Schmidt, C. E. Metz, K. L. Lam, T. Ogura, Y. Wu, and H. MacMahon, "Improvement in radiologists' detection of clustered microcalcifications on mammograms. The potential of computer-aided diagnosis," Invest. Radiol. **25**, 1102–1110 (1990).
[12]W. P. Kegelmeyer, J. M. Pruneda, P. D. Bourland, A. Hillis, M. W. Riggs, and M. L. Nipper, "Computer-aided mammographic screening for speculated lesions," Radiology **191**, 331–337 (1994).

[13]L. Tabar and P. B. Dean, *Teaching Atlas of Mammography* (Thieme, New York, 1985).
[14]J. N. Wolfe, "Analysis of 462 breast carcinomas," Am. J. Roentgenol. **121**, 846–853 (1974).
[15]W. A. Murphy and K. DeSchryver-Kecskemeti, "Isolated clustered microcalcification in the breast: Radiologic–pathologic correlation," Radiology **127**, 335–341 (1978).
[16]R. R. Millis, R. Davis, and A. J. Stacey, "The detection and significance of calcifications in the breast: A radiological and pathological study," Br. J. Radiol. **49**, 12–26 (1976).
[17]E. A. Sickles, "Mammographic features of 300 consecutive nonpalpable breast cancers," Am. J. Roentgenol. **146**, 661–663 (1986).
[18]H. P. Chan, K. Doi, S. Galhotra, C. J. Vyborny, H. MacMahon, and P. M. Jokich, "Image feature analysis and computer-aided diagnosis in digital radiography. 1. Automated detection of microcalcifications in mammography," Med. Phys. **14**, 538–548 (1987).
[19]H. P. Chan, K. Doi, C. J. Vyborny, K. L. Lam, and R. A. Schmidt, "Computer-aided detection of microcalcifications in mammograms: Methodology and preliminary clinical study," Invest. Radiol. **23**, 664–671 (1988).
[20]B. W. Fam, S. L. Olson, P. F. Winter, and F. J. Scholz, "Algorithm for the detection of fine clustered calcifications on film mammograms," Radiology **169**, 333–337 (1988).
[21]D. H. Davies and D. R. Dance, "Automatic computer detection of clustered calcifications in digital mammograms," Phys. Med. Biol. **35**, 1111–1118 (1990).
[22]W. Qian, L. P. Clarke, M. Kallergi, H. D. Li, R. Velthuizen, R. A. Clark, and M. L. Silbiger, "Tree-structured nonlinear filter and wavelet transform for microcalcification segmentation in mammography," Proc. SPIE **1905**, 509–520 (1993).
[23]L. N. Mascio, J. M. Hernandez, and C. M. Logan, "Automated analysis for microcalcifications in high resolution digital mammograms," Proc. SPIE **1898**, 472–479 (1993).
[24]R. M. Nishikawa, M. L. Giger, K. Doi, C. J. Vyborny, R. A. Schmidt, C. E. Metz, Y. Wu, F. F. Yin, Y. Jiang, Z. Huo, P. Lu, W. Zhang, T. Ema, U. Bick, J. Papaioannou, and R. H. Nagel, "Computer-aided detection and diagnosis of masses and clustered microcalcifications from digital mammograms," Proc. SPIE **1905**, 422–431 (1993).
[25]D. Brzakovic, P. Brzakovic, and M. Neskovic, "Approach to automated screening of mammograms," Proc. SPIE **1905**, 690–701 (1993).
[26]S. Astley, I. Hutt, S. Adamson, P. Miller, P. Rose, C. Boggis, C. Taylor, T. Valentine, J. Davies, and J. Armstrong, "Automation in mammography: Computer vision and human perception," Proc. SPIE **1905**, 716–730 (1993).
[27]I. N. Bankman, W. A. Christens-Barry, D. W. Kim, I. N. Weinberg, O. B. Gatewood, and W. R. Brody, "Automated recognition of microcalcification clusters in mammograms," Proc. SPIE **1905**, 731–739 (1993).
[28]N. Karssemeijer, "Recognition of clustered microcalcifications using a random field model," Proc. SPIE **1905**, 776–786 (1993).
[29]L. Shen, R. M. Rangayyan, and J. E. L. Desautels, "Automatic detection and classification system for calcifications in mammograms," Proc. SPIE **1905**, 799–805 (1993).
[30]A. P. Dhawan, Y. S. Chitre, and M. Moskowitz, "Artificial-neural-network-based classification of mammographic microcalcifications using image structure features," Proc. SPIE **1905**, 820–831 (1993).
[31]K. S. Woods, J. L. Solka, C. E. Priebe, C. C. Doss, K. W. Bowyer, and L. P. Clarke, "Comparative evaluation of pattern recognition techniques for detection of microcalcifications," Proc. SPIE **1905**, 841–852 (1993).
[32]Y. Wu, K. Doi, M. L. Giger, and R. M. Nishikawa, "Computerized detection of clustered microcalcifications in digital mammograms: Applications of artificial neural network," Med. Phys. **19**, 555–560 (1992).
[33]W. Zhang, K. Doi, M. L. Giger, Y. Wu, R. M. Nishikawa, and R. A. Schmidt, "Computerized detection of clustered microcalcifications in digital mammograms using a shift-invariant artificial neural network," Med. Phys. **21**, 517–524 (1994).
[34]S. B. Lo, M. T. Freedman, J. Lin, and S. K. Mun, "Automatic lung nodule detection using profile matching and back-propagation neural network techniques," J. Dig. Imag. **6**, 48–54 (1993).
[35]H. P. Chan, L. T. Niklason, D. M. Ikeda, K. L. Lam, and D. D. Adler, "Digitization requirements in mammography: Effects on computer-aided detection of microcalcifications," Med. Phys. **21**, 1203–1211 (1994).
[36]H. P. Chan, L. T. Niklason, D. M. Ikeda, and D. D. Adler, "Computer-aided diagnosis in mammography: Detection and characterization of microcalcifications," Med. Phys. **19**, 831 (1992).
[37]K. Fukushima, S. Miyake, and T. Ito, "Neocognitron: A neural network model for a mechanism of visual pattern recognition," IEEE Trans. Syst. Man Cyb. **SME-13**, 826–834 (1983).
[38]J. A. Swets and R. M. Pickett, *Evaluation of Diagnostic System: Methods From Signal Detection Theory* (Academic Press, New York, 1982).
[39]C. E. Metz, J. H. Shen, and B. A. Herman, "New methods for estimating a binormal ROC curve from continuously-distributed test results," presented at the 1990 Annual Meeting of the American Statistical Association, Anaheim, CA, August 7, 1990.
[40]P. C. Bunch, J. F. Hamilton, G. K. Sanderson, and A. H. Simmons, "A free response approach to the measurement and characterization of radiographic observer performance," Proc. SPIE **127**, 124–135 (1977).
[41]D. P. Chakraborty, "Maximum likelihood analysis of free-response receiver operating characteristic (FROC) data," Med. Phys. **16**, 561–568 (1989).