Optimal Neural Network Architecture Selection:
Improvement in Computerized Detection of Microcalcifications[1]

Metin N. Gurcan, PhD, Heang-Ping Chan, PhD, Berkman Sahiner, PhD
Lubomir Hadjiiski, PhD, Nicholas Petrick, PhD, Mark A. Helvie, MD

Rationale and Objectives. The authors evaluated the effect of optimal neural network architecture selection on the performance of a computer-aided diagnostic system designed to detect microcalcification clusters on digitized mammograms.

Materials and Methods. The authors developed a computer program to detect microcalcification clusters automatically on digitized mammograms. Previously, they found that a properly selected and trained convolution neural network (CNN) could reduce false-positive (FP) findings and therefore improve the accuracy of microcalcification detection. In the current study, they evaluated the effectiveness of the CNN optimized with an automated optimization technique in improving the accuracy of the microcalcification detection program, comparing it with the manually selected CNN. An independent test data set was used, which included 472 mammograms selected from the University of South Florida public database and contained 253 biopsy-proved malignant clusters.

Results. At an FP rate of 0.7 cluster per image, the film-based sensitivity was 84.6% for the optimized CNN, compared with 77.2% for the manually selected CNN. For clusters imaged on both craniocaudal and mediolateral oblique views, a cluster could be considered detected when it was detected on one or both views. For this case-based approach, at an FP rate of 0.7 per image, the sensitivity was 93.3% for the optimized and 87.0% for the manually selected CNN.

Conclusion. The classification of true and false signals is an important step in the microcalcification detection program. An optimized CNN can effectively reduce FP findings and improve the accuracy of the computer-aided detection system.

Key Words. Breast, calcification; breast radiography; computers, diagnostic aid; computers, neural network.

Although the 5-year survival rate for breast cancer has improved over the years, possibly due to screening programs, breast cancer remains one of the most common cancers among women in the Western world (1). When breast cancer is detected in its localized stage, the 5-year survival rate is 97%; that rate drops to about 20% if the cancer has metastasized (2). Screening mammography is currently the best tool available for the early detection of breast cancer (3). Although its sensitivity is relatively high compared with that of other breast imaging modalities, its false-negative rate is still as high as 15%–30% (4). Double reading has been shown to improve sensitivity (5), but it is not cost-effective in a clinical setting. Computer-aided diagnosis (CAD) can provide a second opinion and can improve the detection accuracy significantly (6–9).

Several research groups have developed CAD programs for the detection of microcalcifications. The programs use different approaches, employing a number of parameters usually determined during development of the program. Ex-

Acad Radiol 2002; 9:420–429

[1] From the Department of Radiology, University of Michigan Hospitals, 1500 E Medical Center Dr, UH B1F510B, Ann Arbor, MI 48109-0030. Received October 9, 2001; revision requested October 23; revision received and accepted November 5. Supported by U.S. Public Health Service grant CA 48129 and U.S. Army Medical Research and Materiel Command grant DAMD 17-96-1-6254. **Address correspondence to** H.P.C.

amples include the neighborhood size for the normalization of local contrast in reference 10 and the signal-to-noise ratio (SNR) to determine the locally adaptive threshold in reference 11. Generally, these parameters are chosen by experimenting with their values manually until a satisfactory performance is achieved, but there is no guarantee that optimal values will be found by trial and error.

To set the parameters of the CAD systems automatically in an optimal manner, several approaches have been proposed. Anastasio et al (12) used a genetic algorithm–based optimization method to select the values of 10 parameters in a rule-based microcalcification detection system. A genetic algorithm searches a parameter space by using an ad hoc cost function as a guide. By performing training and resubstitution on a data set with 89 images, Anastasio et al observed that the optimization increased the sensitivity of the CAD system from 80% to 87% at a false-positive (FP) rate of 1.0 cluster per image.

Many CAD systems are composed of several independent yet interrelated parts, and some optimization studies have involved optimizing one part of the CAD system. For example, Sahiner et al (13) used a genetic algorithm and a specially designed cost function to select features that could enhance the performance in the high–sensitivity region of a classifier for distinguishing malignant and benign masses. Chan et al (14) used a genetic algorithm to optimize features for differentiating malignant from benign microcalcifications. Leichter et al (15) used feature selection to optimize the characterization of microcalcifications. Yoshida et al (16) optimized the wavelet transform for microcalcification detection based on supervised learning. Tsai et al (17) used a genetic algorithm to determine the optimal set of fuzzy membership functions to classify myocardial heart disease with ultrasound images. Recently, we proposed and compared several automated techniques for selecting optimal neural network architecture for CAD (18–21). In the present study, we evaluated the effect of convolution neural network (CNN) architecture selected with the automated optimization technique on microcalcification detection, in comparison with manually selected architecture. For this comparison we used a publicly available, relatively large, and completely independent data set of digitized mammograms.

MATERIALS AND METHODS

Data Set

The data set of 108 mammograms used for the optimization and training of the CNN architecture was part of our own database collected with Institutional Review Board approval at the University of Michigan, Ann Arbor. The mammograms included both malignant and benign clusters. For validation purposes we used another data set of 152 mammograms, which was also part of our own database but different from the set used for training. The mammograms in our database were randomly selected from the files of patients who had undergone biopsy at the University of Michigan after screening or diagnostic mammography, so they included microcalcifications with a wide range of characteristics, similar to those encountered in clinical practice that radiologists consider to warrant biopsy. The training and validation data sets were digitized with a Lumisys 85 laser scanner (Lumisys, Sunnyvale, Calif) in our laboratory. The optical density (OD) range of the scanner was 0–4.0. The digitizer was calibrated so that the gray values were linearly and inversely proportional to the OD, with a slope of −0.001 OD unit/pixel value.

For test purposes, an independent data set was used. This data set included 472 digitized mammograms, selected from the University of South Florida (USF) digitized mammogram database, which is publicly available over the Internet (22). From all the available cases in this database, only malignant cases digitized with the Lumisys 200 laser scanner were selected (volumes: cancer_01, cancer_02, cancer_05, cancer_09, and cancer_15). The OD range of the scanner for the USF database was 0–3.6. The digitizer was calibrated so that the gray values were linearly and inversely proportional to the OD, with a slope of −0.001 OD unit/pixel value. Details of the case collection method are described at the USF Web site (22). All mammograms in the training, validation, and test sets were digitized at a pixel resolution of 0.05 × 0.05 mm with 4,096 gray levels. We converted these images to 0.1 × 0.1-mm resolution by averaging adjacent 2 × 2 pixels and subsampling. The detection was carried out on these 0.1-mm-resolution images.

Types of the microcalcifications in the selected cases included punctate, amorphous, pleomorphic, round and regular, fine linear branching, round, dystrophic. The distributions of the calcifications were clustered, linear, segmental, and regional. The lesion types, the assessment, the subtlety, and the pathologic findings were provided with the database. The cluster locations were marked on each image as an overlay file. There were a total of 272 microcalcification clusters; 253 proved to be malignant at biopsy, and the rest were benign. The benign clusters were all additional clusters found in the malignant cases.

Figure 1 shows the distribution of the assessment codes for the malignant clusters in our test data set. The assessments are provided in the USF database and follow the American College of Radiology Breast Imaging Reporting and Data System (BI-RADS) categories. This distribution shows that most of the clusters are in the "actionable lesion" category, defined as BI-RADS assessment scores of 3, 4, or 5. These scores require short–interval follow–up or biopsy. Figure 2 shows the breast density information. A majority of the clusters come from breasts with density of 2 or 3. Figure 3 shows the distribution of the subtlety ratings for the malignant clusters provided with the USF cases. There is no BI-RADS standard for the subtlety rating. The ratings were assessed subjectively by expert radiologists who collected the cases for the USF database. In their scale, a subtlety rating of 1 indicated the most obvious clusters, while a rating of 5 indicated the most subtle. This distribution indicates that most of the clusters in this test set could be classified as subtle.

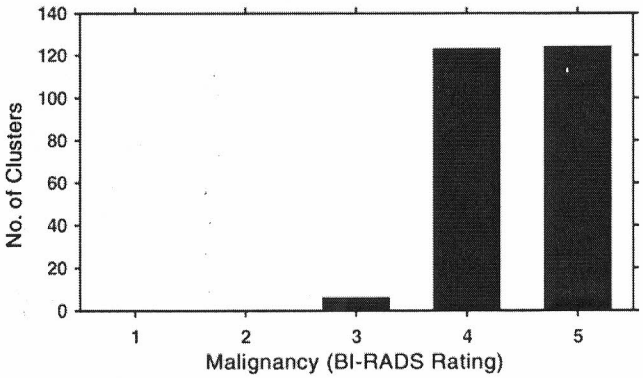

Figure 1. Distribution of assessment ratings for clusters used in our test data set. The assessment follows the American College of Radiology BI-RADS standard and was provided with the USF database. Because only biopsy-proved malignant clusters were included in this test set, the clusters have BI-RADS ratings of 3 (probably benign finding, short-interval follow-up suggested), 4 (suspicious abnormality, biopsy should be considered), or 5 (highly suggestive of malignancy, appropriate action should be taken).

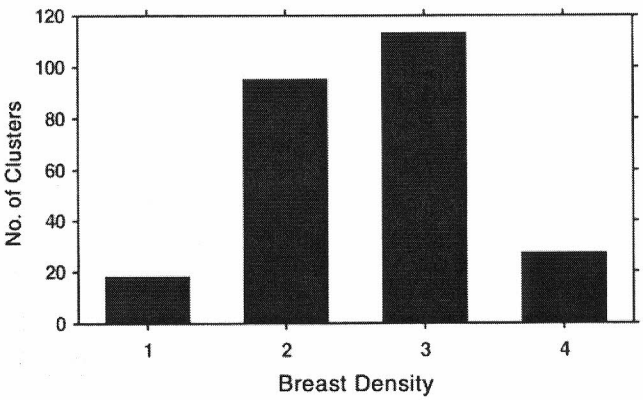

Figure 2. Breast density information for mammograms included in the test data set. The breast density information follows the BI-RADS standard and was provided with the USF database: 1 = almost entirely fat, 2 = scattered fibroglandular densities, 3 = heterogeneously dense, 4 = extremely dense.

Microcalcification Detection Program

We have developed a computer program to automatically detect microcalcification clusters on digitized mammograms (11). The program has three major steps. The first step is preprocessing, in which the breast boundary is automatically determined and the breast region is processed with a bandpass filter to obtain an SNR-enhanced image. The second step is segmentation. In this step, potential microcalcification locations are determined with global and locally adaptive thresholding methods. The local threshold is calculated as the product of the local root-mean-square noise and an input SNR threshold. The microcalcification size, contrast, and SNR are also calculated.

In the third step, the extracted signals are classified as either a true-positive (TP) microcalcification or FP signal. This FP reduction step has three stages. The first is a rule-based classification that uses the size, contrast, and SNR information to generate decision rules. The second-stage classification uses a trained CNN classifier to recognize the abnormal patterns. Finally, regional clustering is used to identify clusters of signals. If a TP signal is within a neighborhood of other TP signals, they are combined to form a cluster. Previously, we have found that the CNN could effectively reduce the number of FP findings and therefore improve the accuracy of the microcalcification detection program (11).

CNN

The CNN is based on the neocognitron structure of Fukushima (23). It was previously used for detection of lung nodules on chest radiographs, detection of microcalcifications on mammograms, and classification of masses and normal breast tissue on mammograms (11,24,25). Figure 4 shows a schematic representation of the CNN structure. The input to the CNN is a region of interest (ROI) image, extracted for each of the detected signals. The nodes in the hidden layers are arranged in groups; each group functions as a filter kernel. The CNN classifies the input ROI as TP or FP. The output node value is

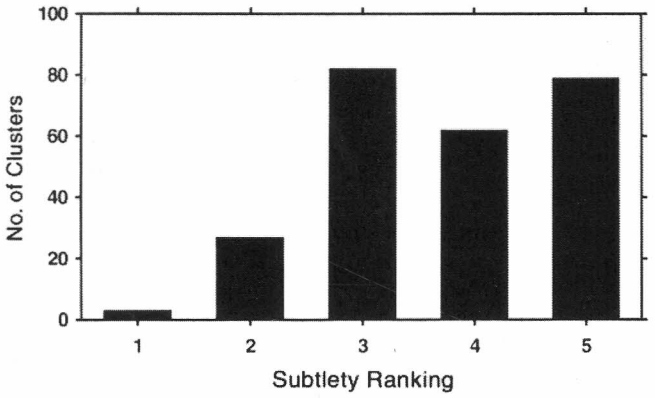

Figure 3. Subtlety rankings (1 = obvious, 5 = subtle) of the 253 clusters provided with the USF data set.

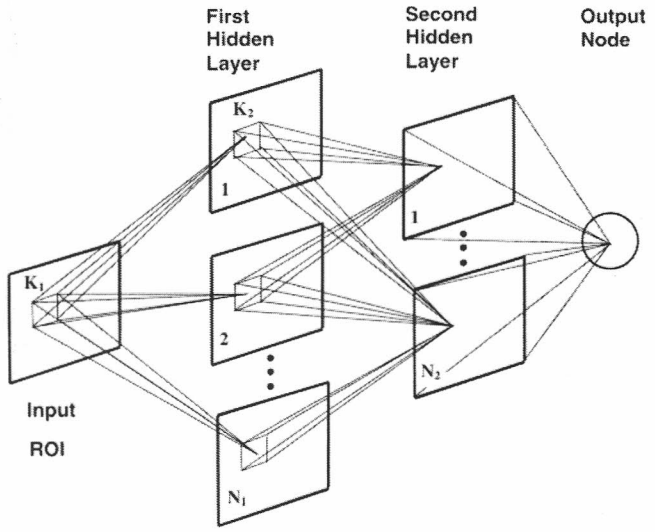

Figure 4. Schematic diagram of the architecture of a CNN. The input to the CNN is an ROI image extracted for each of the detected signals. The output is a scalar that is the relative rating by the CNN representing the likelihood that the input ROI contains a TP microcalcification or an FP signal.

close to 1 for true microcalcifications and close to 0 for FP signals. In this work, the CNN had one input node, two hidden layers, and one output node. All node groups in the two hidden layers were fully connected. The images in each layer were convolved with the filter kernels to obtain the pixel values in the images to be transferred to the following layer. There were N_1 node groups in the first layer and N_2 node groups in the second hidden layer. The kernel sizes of the first group of filters between the input node and the first hidden layer were $K_1 \times K_1$, and those of the second group of filters between the first and second hidden layer were $K_2 \times K_2$. Sigmoidal activation functions were used, and the CNN was trained by means of the error back-propagation rule.

Neural Network Architecture Selection

The CNN architecture used in our earlier studies was selected with a manual optimization technique (11). We recently evaluated the use of automated optimization methods for selecting an optimal CNN architecture. Details of the automated architecture selection study have been described in the literature (21). Briefly, we compared three automated methods: the steepest descent, the simulated annealing, and the genetic algorithm methods. Four main parameters of the CNN architecture, N_1, N_2, K_1, and K_2, were considered for optimization. The area under the receiver operating characteristic (ROC) curve, A_z, was used to design a cost function. The simulated annealing experiments were conducted with four different annealing schedules. Three different parent selection methods were compared for the genetic algorithm experiments.

Our training data set consisted of ROI images extracted from 108 mammograms, described above. The locations of individual microcalcifications in these images were manually identified and saved in a truth file. After the prescreening steps of the microcalcification detection program (11), the detected signals were labeled as TP or FP automatically through comparison with the truth file. A 16 × 16-pixel ROI was then extracted for each of the detected signals, and these ROI images were used for training and testing the CNN. Either a true or a false microcalcification was located at the center of the ROI. The microcalcification detection program detected more FP ROIs than TP ROIs at the prescreening stage. To achieve approximately equal numbers of TP and FP ROIs, we used only a randomly selected subset of FP ROI images.

The selected ROIs were divided into two separate groups. For the first part of the experiments, group 1 was used for training the CNN and group 2 for testing the trained CNN. For the second part of the experiment, the roles of the two groups were switched. Group 1 consisted of 533 ROIs with microcalcifications (240 from malignant clusters and 293 from benign clusters) and 553 FP ROIs. Group 2 had 547 ROIs with microcalcifications (252 from malignant clusters and 295 from benign clusters) and 570 FP ROIs. Therefore, group 1 contained 1,086 ROIs, and group 2 contained 1,117 ROIs. The optimal architecture (N_1-N_2-K_1-K_2) was determined to be 14-4-5-5 when the architecture was trained with group 1 and tested with group 2 and 14-10-5-7 when the training and the test sets

were switched. In our previous study (11), the optimal architecture was determined to be 12-8-5-3 with a manual search technique.

RESULTS

In addition to the 108 mammograms for the training set, we used a data set of 152 mammograms to validate the selected CNN architectures. This data set included 62 mammograms with at least one malignant microcalcification cluster and 90 normal images that were free of clustered microcalcifications. The first two steps of the microcalcification detection program were run on these images. The outputs of these steps provided the potential microcalcification locations. For the last step, classification was run three times with different CNN architectures. In the first run, the manually optimized architecture 12-8-5-3 and its neural network weights were used. In the second and third runs, the two automatically optimized architectures, 14-4-5-5 and 14-10-5-7, and their corresponding weights were used, respectively. For each run, the detection outputs were calculated for three different SNR thresholds: 2.8, 2.9, 3.0. The sensitivity was calculated from the 62 abnormal mammograms, and the FP rates were estimated from the detection output for the 90 normal images. The outputs from these three runs were used to determine the free–response ROC (FROC) curves compared in Figure 5. The comparison indicates that the first optimal architecture (14-4-5-5) generally results in much lower FP rates, but it also reduces the number of TP clusters, thus reducing the sensitivity. The second optimal architecture (14-10-5-7) presents a substantial improvement in terms of both higher sensitivity and lower FP rate. For instance, the sensitivity increases from 78.7% to 84.2% at an FP rate of 0.7 cluster per image. Therefore, these validation results indicate that the best CNN architecture is the second optimal architecture. We tested the performance of this architecture on the independent USF data set described above.

To test the performance of the selected optimal architecture, we ran the detection program at seven SNR threshold values, varying between 2.6 and 3.2 at increments of 0.1. Figure 6 shows the FROC curves of the microcalcification detection program for both the manually optimized and the automatically optimized CNN architectures. The FP rate was estimated from the computer marks on the 184 normal mammograms that were free of microcalcifications in the USF data set. The automatically optimized architecture again outperformed the manually optimized architecture. At an FP rate of 0.7 cluster per image, the film-based sensitivity is 84.6% for the optimized CNN, compared with 77.2% for the manually selected CNN. Figure 7 shows the FROC curves for the microcalcification detection programs when clusters having images in both craniocaudal and mediolateral oblique views are analyzed and a cluster is considered to be detected if detected on one

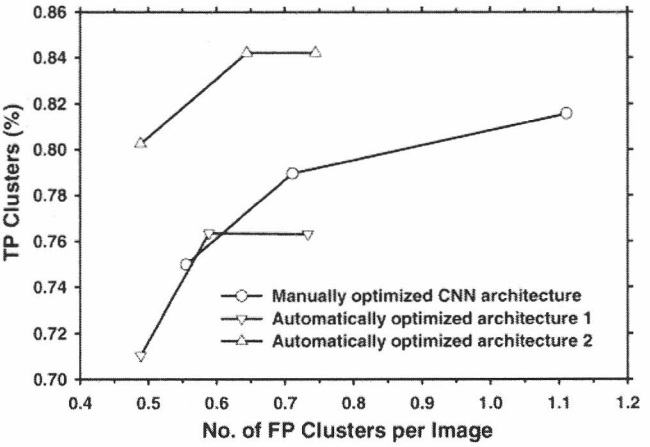

Figure 5. Comparison of validation FROC curves for detection of clustered microcalcifications with different CNN architectures: manually optimized architecture (12-8-5-3), automatically optimized architecture 1 (14-4-5-5), and automatically optimized architecture 2 (14-10-5-7). The evaluation was performed with the 152-image validation data set and three SNR thresholds (2.8, 2.9, and 3.0).

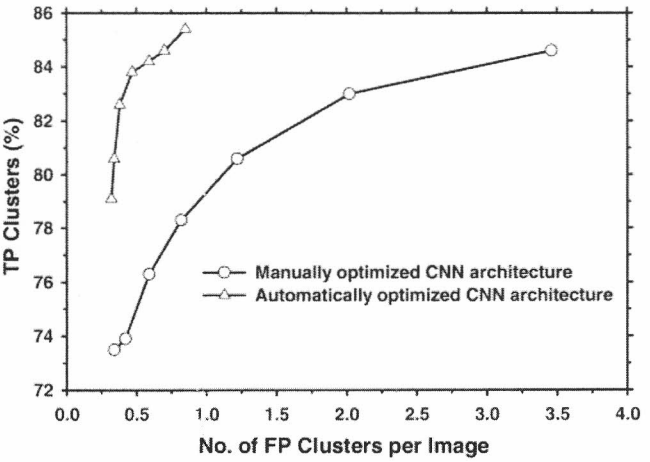

Figure 6. Comparison of test FROC curves for detection of clustered microcalcifications with manually and automatically optimized CNN architectures for film-based (single view) scoring. The automatically optimized architecture is 14-10-5-7. The evaluation was performed with the 472-image test data set and at seven SNR thresholds (between 2.6 and 3.2, varying at increments of 0.1).

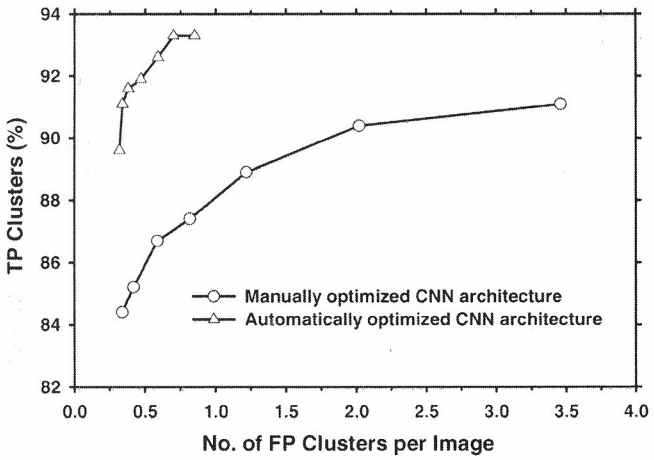

Figure 7. Comparison of test FROC curves for detection of clustered microcalcifications with manually and with automatically optimized CNN architectures for case-based scoring. In case-based scoring, if clusters imaged on both craniocaudal and mediolateral oblique views are analyzed, a cluster is considered to be detected when it is detected on one or both views. The automatically optimized architecture is 14-10-5-7. Evaluation was performed with the 472-image test data set (236 two-view mammograms) and at seven SNR thresholds (between 2.6 and 3.2, varying at increments of 0.1).

or both views. This "case-based" scoring has been adopted for the evaluation of some CAD systems (9). The rationale is that if the CAD system can draw the radiologist's attention to the lesion on one of the views, the lesion is unlikely to be missed. For case-based scoring the sensitivity at an FP rate of 0.7 cluster per image is 93.3% for the automatically optimized CNN and 87.0% for the manually selected CNN.

DISCUSSION

Classification of true and false signals is an important step in the microcalcification detection program. An optimized CNN can effectively reduce FP findings and improve the detection accuracy of the CAD system. Manually searching for the optimal CNN architecture often results in a local optimum, because it is difficult to explore adequately a high-dimensional parameter space with manual experimentation. We have demonstrated previously that an automated optimization algorithm such as simulated annealing can find the global optimum efficiently (18–21).

Our optimization is currently limited to one stage of the detection program, FP reduction with the CNN. Our cost function was based on the A_z of the CNN classifier for its performance in differentiating the TP and FP signals. Ideally, one would prefer to optimize all parameters in the detection program together. In such a case, it is necessary to optimize performance in terms of the FROC curve. To take advantage of some well-established automated optimization methods, such as the genetic algorithm or simulated annealing method, one must define a scalar cost function, but there is no widely accepted form of a scalar cost tunction for comparing FROC curves obtained with different detection methods. In an alternative form of FROC analysis, known as AFROC analysis, a scalar A_1 is calculated, which can be considered a form of cost function, but AFROC analysis requires a special experimental setting (26).

Anastasio et al (12) proposed an ad hoc cost function, $C(f,s)$, in which they incorporated their preferences about their sensitivity-specificity tradeoff into a discrete grid of numbers on the sensitivity-specificity plane; the values between these grid values were determined by means of bilinear interpolation. The fitness of each solution during their genetic algorithm evolution process was assigned by evaluating the cost function for the solution. Since the cost function optimized the FROC curve only at an individual operating point that corresponded to a sensitivity-specificity pair, it did not provide sufficient information to compare two different FROC curves. Moreover, the choice of the preference values is subjective.

For our optimization study we used the ROC analysis, a commonly accepted method of comparing overall classifier performance; therefore, the cost definition was based on the area under the ROC curve, A_z. To extend this definition for FROC curves, we propose the following cost function:

$$C = 100(u - l) - \int_l^u s(f)df, \qquad (1)$$

where l and u are the lower and upper limits of the FP range of interest, respectively; f is the number of FP findings per image and $s(f)$ is the sensitivity at an FP rate of f. This cost function will compare two FROC curves in a chosen range of FP rates.

A similar function was proposed by te Brake et al (27) to measure the quality of a feature for the discrimination of malignant masses from normal tissue on digitized mammograms. In their definition, the area under the logarithmically plotted FROC curve between 0.05

and 4.0 FP findings per image was used as a quality measure:

$$A_f = \int_{0.05}^{4.0} s(f)d\ \ln(f) = \int_{0.05}^{4.0} s(f)\frac{1}{f}df, \qquad (2)$$

where A_f is the area under the FROC curve between the chosen FP range; f and $s(f)$ are defined in Equation (1). As shown in Figure 8, the cost function in Equation (1) calculates the area above the FROC curve and below the 100% sensitivity line. In this cost function, only the operating range of the CAD system needs to be defined in terms of the FP range. For a given FROC curve, the knowledge of $s(f)$ is sufficient for calculating the total cost function. Thus, this cost function is directly related to the performance of the CAD system rather than subjective preferences of the user. Additionally, the cost definition in Equation (1) is flexible in that one can choose the range of FP rates $[l,u]$ along the FROC curve for which the CAD system is to be optimized. Further studies are needed to evaluate the effectiveness of using the cost function defined in Equation (1) to optimize CAD systems.

Of all the available images in the USF database, we used only those scanned with the Lumisys scanner, because it was similar to the scanner we used to acquire digitized mammograms for developing our CAD programs and setting their parameters. It is not uncommon to see drastic decreases in performance if different types of scanners are used for the development and testing of a CAD system. For instance, Velthuizen et al (28) developed a microcalcification detection program that used mammogram images digitized with a DBA ImageClear R3000 scanner (DBA Systems, Melbourne, Fla) and achieved 94% sensitivity at an FP rate of 1.23 clusters per image with a database of 26 images. When they scanned the same images with a Lumiscan 50 scanner (Lumisys) and evaluated the detection performance, the sensitivity dropped to 28% and the FP rate increased to 2.19 clusters per image. In this study, since we were interested in evaluating the performance change due to CNN architecture selection, we limited ourselves to images in the USF database that were obtained with a similar scanner, thereby minimizing the effects of other factors on the performance change. The dependence of our detection program on data set acquired with different film scanners will be investigated in the future.

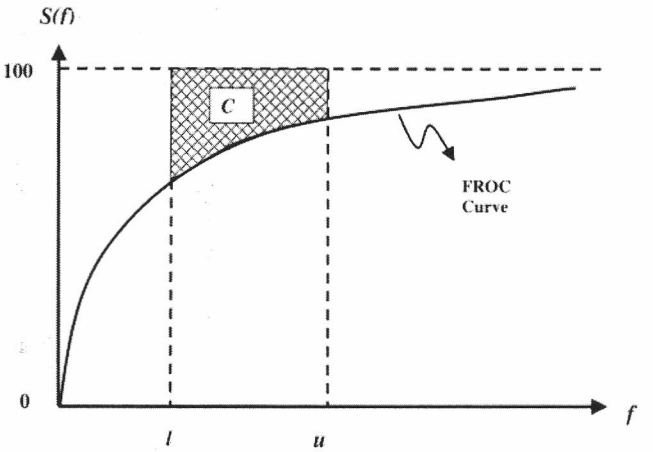

Figure 8. Definition of a scalar cost function for optimization of CAD system, where l and u are the lower and upper limits of the range of FP findings per image on the FROC curve, respectively, f is the number of FP findings per image, and $s(f)$ is the sensitivity at an FP rate of f. The cost, C, is determined as the area above the FROC curve and below the 100% sensitivity line. This area is shaded.

Summary of Data Sets Used in the Training, Validation, and Test Stages

Data Set	Source	No. of Images	No. of Malignant Microcalcification Clusters
Training	University of Michigan	108	29
Validation	University of Michigan	152	76
Test	University of South Florida	472	253

Note.—These three data sets are mutually exclusive, with no overlap of images.

For this optimization study, we followed a three-stage (training, validation, test) CAD development and evaluation method. This method requires separate data sets for each stage. The Table summarizes the information about the images in these data sets. The images in the first two data sets came from the patient files at the University of Michigan. These two data sets were mutually exclusive, however, with no common images. The data set for training was used to find the parameters of the optimal neural network architecture and neural network weights. The images in the validation set were used to evaluate the performance of the selected architectures and identify the best-performing architecture for an independent data set. Once the architecture was selected with the validation set, the parameters of the detection program were fixed, and no further changes were made either to the program or to the CNN architecture and its weights.

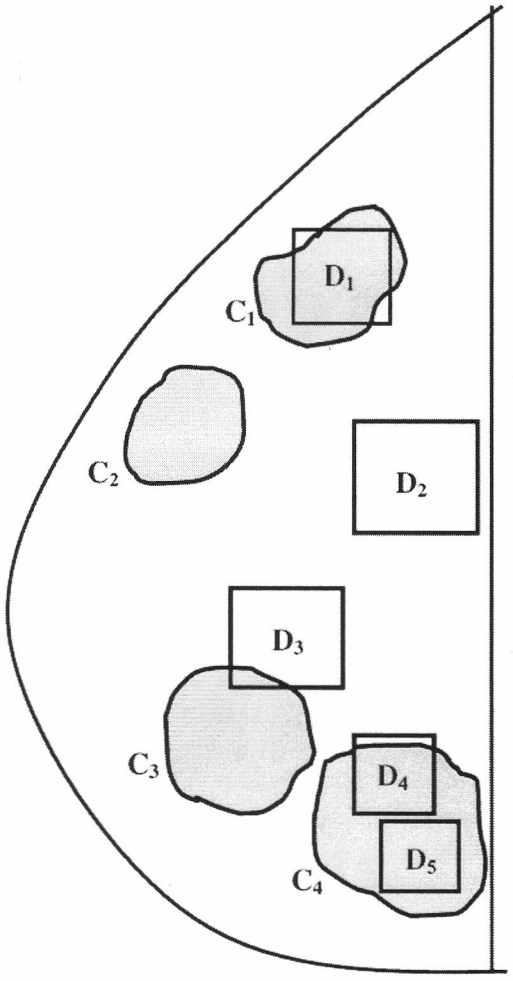

Figure 9. In this schematic mammogram, there are four microcalcification clusters, (C_1, C_2, C_3, C_4), the extents of which are drawn by radiologists. The microcalcification detection program detects five clusters (D_1, D_2, D_3, D_4, D_5). D_1 is a TP detection. D_2 and D_3 are FP detections, because D_2 does not intersect with any cluster and the intersection of D_3 with C_3 is less than 40%, which was chosen as the threshold for detection during training and validation of the automatic scoring criteria. D_4 and D_5 are considered to be detecting the same cluster, C_4. Therefore, for this example, there are two TP findings (C_1, C_4), two false-negative findings (C_2, C_3), and two FP findings (D_2 and D_3).

With this CAD program, microcalcification detection was performed in a completely independent and publicly available test data set. The images in this set were used only to assess the performance of the fully specified optimal architecture. If only a small training set and an "independent" test set are used and the detection performance on the test set is used as a guide to adjust the parameters of the detection program, there is always a bias because the CAD system is fine tuned to this particular "test" data set, which is essentially a validation set. The results achieved with that test set may not be generalizable to other data sets. This consideration is especially important for CAD system development. Before a CAD system can be considered for clinical implementation, it is advisable to follow the three-stage method described here and evaluate the system with an independent random test set containing a large number of mammograms with a wide spectrum of characteristics. Otherwise, the test results may not truly reflect the actual performance of the CAD program in the unknown patient population.

The range of the SNR thresholds (2.8–3.0) for the detection in the validation set was determined by our previous experience with the microcalcification detection program. This range has shown to produce detection results within an acceptable range of FP rates. The range of the SNR thresholds for detection in the test set was wider than that for the validation set in order to compare a wider section of the FROC curve. A smaller SNR threshold will generally result in more potential signals to be considered for detection. Thus, the sensitivity is usually higher but the number of FP clusters also increases. On the other hand, a larger SNR threshold generally reduces the number of FP clusters but this usually comes with a decrease in the sensitivity. Although the SNR threshold can assume any positive value, very small values may not always extend the FROC curve much further beyond its current limits, because at very low thresholds the potential signals are merged with the background, and the noisy background also merges into large patches (11). At very high thresholds, even obvious microcalcifications may be missed, and the sensitivity will drop rapidly.

The scoring of the microcalcification detection program was performed automatically. Figure 9 demonstrates how our automatic scoring scheme was designed. There are two sets of inputs to the automatic scoring program. The first consists of the overlay files, in which the extent of each microcalcification cluster is drawn by an expert radiologist as a polygon. The second consists of outputs of the automated microcalcification detection program, which are the smallest rectangular bounding boxes enclosing the detected microcalcification clusters. The scoring program automatically calculates the intersection of the areas enclosed by these rectangles and the polygons. If the ratio of the intersection area to either the rectangle or the polygon area is more than 40%, then the cluster enclosed by the polygon is considered to be detected. If a polygon area is detected with more than one rectangular region, only one TP finding is recorded. The sensitivity for the film-based FROC curve was based on the number

of malignant clusters detected relative to the total number of malignant clusters present in the data set, with different views of the same cluster considered independent. For case-based scoring, the corresponding clusters in the two views are used to determine whether the same cluster is detected by the CAD system on at least one view. Detection of the same cluster on one or both views are scored as one TP finding, and the sensitivity is normalized to the total number of different malignant clusters in the data set.

At present, there is no established statistical test for comparing the significance in the differences between two FROC curves. Therefore, we cannot evaluate the statistical significance of the improvement in the FROC curves with the optimized CNN. Since the increase in sensitivity is substantial, however (from 77.2% to 84.6% at an FP rate of 0.7 cluster per image), and consistent over the range of FP rates studied, the effectiveness of the CNN is evident. Furthermore, as the improvement is observed for a relatively large independent test set and is consistent with the performance observed for the validation set, it is unlikely that the improvement is biased to the specific data set.

CONCLUSION

We have developed a CAD system to detect microcalcification clusters on digitized mammograms. In this study, we evaluated the effectiveness of an optimal neural network architecture selected by an automated simulated annealing optimization technique for improving the performance of the CAD system. At an FP rate of 0.7 cluster per image, the film-based sensitivity is 84.6% for the optimized CNN, compared with 77.2% for a manually selected CNN. If clusters having images in both craniocaudal and mediolateral oblique views are analyzed and a cluster is considered detected when detected on one or both views, at an FP rate of 0.7 per image, the sensitivity is 93.3% with the optimized CNN and 87.0% with the manually selected CNN. This study demonstrates that classification of true and false signals is an important step in the microcalcification detection program and that an optimized CNN can effectively reduce FP findings and improve the detection accuracy of the CAD system.

ACKNOWLEDGMENT

The authors are grateful to Charles E. Metz, PhD, for providing the LABROC program.

REFERENCES

1. Vogel V. Breast cancer prevention: a review of current evidence. CA Cancer J Clin 2000; 50:156–170.
2. National Cancer Institute. Surveillance, Epidemiology, and End Results (SEER) program public-use CD-ROM (1973–1997). Bethesda, Md: Cancer Surveillance Research Program, Cancer Statistics Branch, 2000.
3. Tabar L, Fagerberg G, Chen HH, et al. Efficacy of breast-cancer screening by age: new results from the Swedish 2-county trial. Cancer 1995; 75:2507–2517.
4. Yankaskas BC, Schell MJ, Bird RE, Desrochers DA. Reassessment of breast cancers missed during routine screening mammography: a community-based study. AJR Am J Roentgenol 2001; 177:535–541.
5. Thurfjell EL, Lernevall KA, Taube AAS. Benefit of independent double reading in a population-based mammography screening program. Radiology 1994; 191:241–244.
6. Chan HP, Doi K, Vyborny CJ, et al. Improvement in radiologists' detection of clustered microcalcifications on mammograms: the potential of computer-aided diagnosis. Invest Radiol 1990; 25:1102–1110.
7. Kegelmeyer WP, Pruneda JM, Bourland PD, Hillis A, Riggs MW, Nipper ML. Computer-aided mammographic screening for spiculated lesions. Radiology 1994; 191:331–337.
8. Freer TW, Ulissey MJ. Screening mammography with computer-aided detection: prospective study of 12,860 patients in a community breast center. Radiology 2001; 220:781–786.
9. Warren Burhenne LJ, Wood SA, D'Orsi CJ, et al. Potential contribution of computer-aided detection to the sensitivity of screening mammography. Radiology 2000; 215:554–562.
10. Veldkamp WJH, Karssemeijer N. An improved method for detection of microcalcification clusters in digital mammograms. Proc SPIE 1999; 3661:512–522.
11. Chan HP, Lo SCB, Sahiner B, Lam KL, Helvie MA. Computer-aided detection of mammographic microcalcifications: pattern recognition with an artificial neural network. Med Phys 1995; 22:1555–1567.
12. Anastasio MA, Yoshida H, Nagel R, Nishikawa RM, Doi K. A genetic algorithm-based method for optimizing the performance of a computer-aided diagnosis scheme for detection of clustered microcalcifications in mammograms. Med Phys 1998; 25:1613–1620.
13. Sahiner B, Chan HP, Petrick N, Helvie MA, Goodsitt MM. Design of a high-sensitivity classifier based on a genetic algorithm: application to computer-aided diagnosis. Phys Med Biol 1998; 43:2853–2871.
14. Chan HP, Sahiner B, Lam KL, et al. Computerized analysis of mammographic microcalcifications in morphological and texture feature space. Med Phys 1998; 25:2007–2019.
15. Leichter I, Lederman R, Buchbinder S, Bamberger P, Novak B, Fields S. Optimizing parameters for computer-aided diagnosis of microcalcifications at mammography. Acad Radiol 2000; 7:406–412.
16. Yoshida H, Zhang W, Cai W, Doi K, Nishikawa RM, Giger ML. Optimizing wavelet transform based on supervised learning for detection of microcalcifications in digital mammograms. Proc IEEE Int Conf on Image Processing, Washington, DC, 1995; 3:152–155.
17. Tsai DY, Watanabe S. A method for optimization of fuzzy reasoning by genetic algorithms and its application to discrimination of myocardial heart disease. IEEE Trans Nucl Sci 1999; 46:2239–2246.
18. Gurcan MN, Sahiner B, Chan HP, Hadjiiski LM, Petrick N. Optimal selection of neural network architecture for CAD using simulated annealing. Proc 22nd Annual International Conference of IEEE Engineering in Medicine and Biology Society, Chicago, IL, 2000; 4:3052–3055.
19. Gurcan MN, Sahiner B, Chan HP, Hadjiiski LM, Petrick N. Selection of an optimal neural network architecture for computer-aided diagnosis: comparison of automated optimization techniques (abstr). Radiology 2000; 217(P):436.
20. Gurcan MN, Chan HP, Sahiner B, Hadjiiski LM, Petrick N, Helvie MA. Improvement of computerized detection of microcalcifications using a convolution neural network architecture selected by an automated optimization algorithm. Presented at the Medical Image Perception Conference IX, Warrenton, Va, September 20–23, 2001.
21. Gurcan MN, Sahiner B, Chan HP, Hadjiiski LM, Petrick N. Selection of an optimal neural network architecture for computer-aided detection

of microcalcifications: comparison of automated optimization techniques. Med Phys 2001; 28:1937–1948.
22. Heath M, Bowyer K, Kopans D, et al. Current status of the digital database for screening mammography. In: Karssemeijer N, Thijssen M, Hendriks J, van Erning L, eds. Digital mammography. Dordrecht, the Netherlands: Kluwer, 1998; 457–460.
23. Fukushima K, Miyake S, Ito T. Neocognitron: a neural network model for a mechanism of visual pattern recognition. IEEE Trans Systems Man Cybernetics 1983; SMC-13:826–834.
24. Lo SCB, Chan HP, Lin JS, Li H, Freedman M, Mun SK. Artificial convolution neural network for medical image pattern recognition. Neural Networks 1995; 8:1201–1214.
25. Sahiner B, Chan HP, Petrick N, et al. Classification of mass and normal breast tissue: a convolution neural network classifier with spatial domain and texture images. IEEE Trans Med Imaging 1996; 15:598–610.
26. Chakraborty DP, Winter LHL. Free-response methodology: alternate analysis and a new observer-performance experiment. Radiology 1990; 174:873–881.
27. te Brake GM, Karssemeijer N, Hendriks JHCL. An automatic method to discriminate malignant masses from normal tissue in digital mammograms. Phys Med Biol 2000; 45:2843–2857.
28. Velthuizen RP, Clarke LP. Image standardization for digital mammography. In: Karssemeijer N, Thijssen M, Hendriks J, van Erning L, eds. Digital mammography. Dordrecht, the Netherlands: Kluwer, 1998; 461–464.