AD________________

Award Number: DAMD17-96-1-6058

TITLE: Advanced Methods for the Computer-Aided Diagnosis of Lesions in Digital Mammograms

PRINCIPAL INVESTIGATOR: Maryellen Giger, Ph.D.

CONTRACTING ORGANIZATION: The University of Chicago
Chicago, Illinois 60637

REPORT DATE: July 2000

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE | OMB No. 074-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503

1. AGENCY USE ONLY (Leave blank)	**2. REPORT DATE** July 2000	**3. REPORT TYPE AND DATES COVERED** Final (7 Jun 96 - 6 Jun 00)

4. TITLE AND SUBTITLE
Advanced Methods for the Computer-Aided Diagnosis of Lesions in Digital Mammograms

5. FUNDING NUMBERS
DAMD17-96-1-6058

6. AUTHOR(S)
Maryellen Giger, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
The University of Chicago
Chicago, Illinois 60637

E-MAIL:
m-giger@uchicago.edu

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 Words)*

The objective of the proposed research is to develop computer-aided diagnosis methods for use in mammography in order to increase the diagnostic accuracy of radiologists. Specifically we have developed advanced computerized schemes for the detection spiculated lesions and architectural distortions based on the calculation of the Hough spectrum and for the detection of small, low-contrast early cancers based on gradient and circularity filters. Also, computerized classification schemes for masses using artificial neural networks, rule-based methods, and hybrid systems have been developed. We have investigated the effect of database on feature selection and classifier training. We also investigated a computerized method for including temporal change between mammographic examinations. We have also developed an intelleigent search workstation for aiding radiologist in making diagnostic decisions by providing them with both CAD output and images of known cases that are "similar" to the case in question. The efficacy and efficiency of the CAD methods for mammography are being evaluated on a clinical workstation. The potential significance of this research project lies in the fact that if the detectability of cancers can be increased by employing a computer to aid the radiologist's diagnosis, then the treatment of patients with cancer can be initiated earlier and their chance of survival improved.

14. SUBJECT TERMS
Breast Cancer, Research

15. NUMBER OF PAGES
69

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT Unclassified	**18. SECURITY CLASSIFICATION OF THIS PAGE** Unclassified	**19. SECURITY CLASSIFICATION OF ABSTRACT** Unclassified	**20. LIMITATION OF ABSTRACT** Unlimited

FOREWORD

Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the U.S. Army.

___ Where copyrighted material is quoted, permission has been obtained to use such material.

___ Where material from documents designated for limited distribution is quoted, permission has been obtained to use the material.

___ Citations of commercial organizations and trade names in this report do not constitute an official Department of Army endorsement or approval of the products or services of these organizations.

N/A In conducting research using animals, the investigator(s) adhered to the "Guide for the Care and Use of Laboratory Animals," prepared by the Committee on Care and use of Laboratory Animals of the Institute of Laboratory Resources, national Research Council (NIH Publication No. 86-23, Revised 1985).

X For the protection of human subjects, the investigator(s) adhered to policies of applicable Federal Law 45 CFR 46.

N/A In conducting research utilizing recombinant DNA technology, the investigator(s) adhered to current guidelines promulgated by the National Institutes of Health.

N/A In the conduct of research utilizing recombinant DNA, the investigator(s) adhered to the NIH Guidelines for Research Involving Recombinant DNA Molecules.

N/A In the conduct of research involving hazardous organisms, the investigator(s) adhered to the CDC-NIH Guide for Biosafety in Microbiological and Biomedical Laboratories.

Mayelle Sizer 8/8/00

Table of Contents

INTRODUCTION

Our first-year report was accepted as an excellent report as submitted. The review of our first year report indicated that the report was well-written with extensive background material and meticulous description of the theoretical basis for the algorithms, which are not necessary in future annual reports and could be referred to with appropriate citations. Thus, in this final report, we have substantially shortened the background sections and refer the reviewer to our first-year report.

Although mammography is currently the best method for the detection of breast cancer, between 10-30% of women who have breast cancer and undergo mammography have negative mammograms. In approximately two-thirds of these false-negative mammograms, the radiologist failed to detect the cancer that was evident retrospectively. Low conspicuity of the lesion, eye fatigue and inattentiveness are possible causes for these misses. We believe that the effectiveness (early detection) and efficiency (rapid diagnosis) of screening procedures could be increased substantially by use of a computer system that successfully aids the radiologist by indicating locations of suspicious abnormalities in mammograms. In addition, many breast cancers are detected and referred for surgical biopsy on the basis of a radiographically detected mass lesion or cluster of microcalcifications. Although general rules for the differentiation between benign and malignant breast lesions exist, considerable misclassification of lesions occurs with the current methods. On average, only 10-30% of masses referred for surgical breast biopsy are actually malignant. Surgical biopsy is an invasive technique that is an expensive and traumatic experience for the patient and leaves physical scars that may hinder later diagnoses (to the extent of requiring repeat biopsies for a radiographic tumor-simulating scar). A computerized method capable of detecting and analyzing the characteristics of benign and malignant masses, in an objective manner, should aid radiologists by reducing the numbers of false-positive diagnoses of malignancies, thereby decreasing patient morbidity as well as the number of surgical biopsies performed and their associated complications.

Purpose of the present work

The main hypothesis to be tested is that given dedicated computer-vision programs for the computer-assisted interpretation of mammograms, the diagnostic accuracy for mammographic interpretation will be improved, yielding earlier detection of breast cancer (i.e., a reduction in the number of missed lesions) and a reduction in the number of benign cases sent to biopsy. Computer-aided diagnosis (CAD) can be defined as a diagnosis made by a radiologist who takes into consideration the results of a computerized analysis of radiographic images and uses them as a "second opinion" in detecting lesions and in making diagnostic decisions. The final diagnosis would be made by the radiologist.

Methods of approach

The objective of the proposed research is to develop computer-aided diagnosis methods for use in mammography in order to increase the diagnostic decision accuracy of radiologists and to aid in mammographic screening programs. The CAD methods will include a parallel method for the detection of a range of mass types and for the incorporation of information from multiple views (i.e., CC and MLO, and prior mammograms).

The **specific objectives** of the research to be addressed are:
(1) Development of advanced computerized schemes for the detection and classification of masses in digital mammograms.

(a) Development of a computerized detection scheme for spiculated lesions and architectural distortions based on the calculation of the Hough spectrum.

(b) Development of a computerized detection scheme for small, low-contrast early cancers based on gradient and circularity filters.

(c) Incorporation of the two new methods with a previously-developed bilateral-subtraction method along with feature analyses into a system for lesion detection.

(d) Further development of computerized classification schemes for masses.

(2) Development of computerized methods based on multiple views for enhanced mammographic interpretation.

(a) Development of computerized methods for the incorporation of image information from the CC and MLO views of mammographic examinations.

(b) Development of computerized methods for analysis of temporal change between mammographic examinations.

(3) Incorporation of the computer-vision methods with an Mammo/Icon mammographic review system for enhanced diagnosis.

(a) Expansion of the Mammo/Icon database descriptors to include CAD derived parameters.

(b) Calculation of the computer extracted features of images in the Mammo/Icon database.

(c) Development of hardware and software interfaces for CAD and Mammo/Icon.

(4) Evaluation of the CAD methods for mammography.

BODY (Results to date)

Development of advanced computerized schemes for the detection of masses in digital mammograms.

With the single-image method for detection of small invasive breast cancers localized density peaks on mammograms are identified using a gradient/circularity filter. Lesion contours were generated by matching a deformable template onto a second derivative edge map. In our study (without further feature analyses to reduce false positives) using 45 non-palpable invasive breast cancers, all with a size less than 1 cm (median size of 7 mm), 82% of the cancers were detected with an average false-positive rate of 2.8 per image.

In the Hough spectrum geometric texture analysis technique, the mammogram is analyzed ROI by ROI. Each ROI is transformed into its Hough spectrum and then thresholding is performed with its threshold level based on the statistical properties of the spectrum. ROIs with strong signals of spiculation are then screened out as regions of potential lesions. In a preliminary study, 32 images containing spiculated lesions/architectural distortions (biopsy confirmed) were analyzed using information extracted from the Hough spectrum. Our studies, using only the Hough spectrum based technique without further feature analyses to reduce false positives, yielded sensitivities of 81% for spiculated masses and 67% for architectural distortions at false positives rates of 0.97 and 2.2 per image, respectively. We also converted the method into an AVS based program to expedite the development and optimization of the parameters such as ROI size.

Output from the bilateral subtraction method and that of the gradient/circularity filtering were combined and analyzed. Many masses were detected by both preprocessing methods. For a database of 20 cancer cases, the bilateral yielded a sensitivity of 75% (at 1.8 false-positives detections per image) and the gradient/circularity filter yielded a sensitivity of 70% at the same false postivie rate. Upon comparison, the gradient/circularity filter found masses that the bilateral did not, thus allowing the sensitivity to increase to 80%.

Since November 8, 1994, all screening mammograms taken at the University of Chicago Hospitals have been analyzed on our clinical prototype mammography worstation, except during downtimes. Over 25,000 screening cases have been digitized. Downtime has been minimal, less than 20 days in total, which includes a 3-week period when the mammography section moved to a new outpatient center. During that move, networking problems in the new facility contributed to computer system difficulties. For cases in which a cancer was detected, we also retrospectively reviewed any previous mammograms that were in our study cohort. Two radiologists independently reviewed the cases and stated whether the cancer was visible in a previous exam and whether, knowing that the lesion was present, that they would call the patient back for a diagnostic exam based on the findings in the previous exam. In this way, the number of cancers detected by the computer that were initially missed by the radiologists was determined. With follow-up on the first 10,000 cases, 61 patients have been diagnosed with breast cancer. In 12 of these cases, the screening mammogram(s) were negative even in retrospect. For the mammographically visible cases (n=49), the sensitivity of the two schemes was 68% (34/49). Clinically, 96% of the cancers were detected (47/49). More important than the absolute sensitivity of the workstation is its ability to detect breast cancers that may be missed by a radiologist. In 30 of the 61 cancers, the patient had a screening exam that was read as negative and was included in our study. That is, a screening mammogram that was read as normal, which preceded the cancer being diagnosed. In 14 of these cases, no lesion could be seen in retrospect, i.e., mammographically negative. In 9 of 16 cases, the computer was able to identify the region on the negative-read (cancer visible in retrospect) screening mammogram that corresponded to where the cancer was subsequently detected. Overall, the computer was able to identify the cancer approximately one year before it was diagnosed in approximately 15% (9/61) of all cancer cases and in 56% (9/16) of cases were the cancer was visible in retrospect on a negative-read screening mammogram. The false-positive rate was approximately 1.3 false clusters per image and 2.1 false masses per image. The types of false-positive detections found by the computer in mass detection and clustered microcalcification detection were investigated for 1296 cases. Of the false positives that were indicated by the computer, over 80% of the mass false positives were due to nodular densities on the film.

In order to determine the effect of false-positive detections on mammographic interpretation, we calculated the call-back rate in one-year time periods before and after implementation of the workstation in the clinical area. The callback rate is the fraction of screening mammograms read as abnormal. Before introduction of CAD, 13.2% of screeners were called back for further workup and after the introduction of CAD, 12.6% of screeners were called back for further workup. Thus, the false-positive output from the computer did not increase the number of women called back.

A new development, which has been implemented into the detection scheme for mammographic masses, is a new region growing algorithm. The segmentation of lesions from surrounding background is a vital step in many computerized mass detection schemes. We have developed two novel lesion segmentation techniques -- one based on a single feature called the radial gradient index (similar feature to that described above) and one based on a simple probabilistic model to segment mass lesions from surrounding background. In both methods a series of image partitions is created using gray-level information as well as prior knowledge of the shape of typical mass lesions. With the former method the partition that maximizes the radial gradient index is selected. In the latter method, probability distributions for gray-levels inside and outside the partitions are estimated, and subsequently used to determine the probability that the image occurred for each given partition. The partition that maximizes this probability is selected as the final lesion partition (contour). We tested these methods against our previous region-growing algorithm using a database of biopsy-proven, malignant lesions and found that the new lesion segmentation algorithms more closely match radiologists' outlines of these lesions. At an overlap threshold of 0.30, gray level region growing correctly delineates 62% of the lesions in our database while the radial gradient index (RGI) algorithm and the probabilistic segmentation algorithm correctly segment 92% and 96% of the lesions, respectively. With these new segmentation results we hope to find and extract new features that will help differential between actual lesions and false-positive detections, thus improving the overall performance of computerized mass detection.

In order to improve the classifier performance in the detection method for distinguishing between actual lesions and false-positive detections, we investigated feature selection with limited datasets and the use of probabilistic artificial neural networks. In many computerized schemes, numerous features can be extracted to describe suspect image regions. A subset of these features is then employed in a data classifier to determine whether the suspect region is abnormal or normal. Different subsets of features will, in general, result in different classification performances. A feature selection method is often used to determine an ``optimal" subset of features to use with a particular classifier. A classifier performance measure (such as the area under the receiver operating characteristic (ROC) curve) must be incorporated into this feature selection process. With limited datasets, however, there is a distribution in the classifier performance measure for a given classifier and subset of features. We investigated the variation in the selected subset of ``optimal" features as compared with the true optimal subset of features caused by this distribution of classifier performance. We considered examples in which the probability that the optimal subset of features is selected can be analytically computed. We showed the dependence of this probability on the dataset sample size, the total number of features from which to select, the number of features selected, and the performance of the true optimal subset. Once a subset of features has been selected, the parameters of the data classifier must be determined. We showed that, with limited datasets and/or a large number of features from which to choose, bias is introduced if the classifier parameters are determined using the same data that were employed to select the ``optimal" subset of features.

It is well understood that the optimal classification decision variable is the likelihood ratio or any monotonic transformation of the likelihood ratio. An automated classifier which maps from an input space to one of the likelihood ratio family of decision variables is an optimal classifier or an ideal observer. Artificial neural networks (ANNs) are frequently used as classifiers for many problems. In the limit of large sample sizes, an ANN approximates a mapping function which is a monotonic transformation of the likelihood ratio, i.e., it estimates an ideal observer decision variable. The disadvantages of conventional ANNs include the potential over-parameterization of the mapping function which results in a poor approximation of an optimal mapping function for smaller sample sizes. Recently, Bayesian methods have been applied to ANNs in order to regularize training to improve the robustness of the classifier. A Bayesian ANN should thus better approximate the optimal decision variable given small sample sizes. We have evaluated the accuracy of Bayesian ANN models of ideal observer decision variables as a function of the number of hidden units used, the signal-to-noise ratio of the data, and the number of features or dimensionality of the data. We showed that when enough training data are present, excess hidden units do not substantially degrade the accuracy of Bayesian ANNs. The minimum number of hidden units required to best model the optimal mapping function, however, varies with the complexity of the data.

A new extension of the region growing method was developed to perform as a filter. The radial gradient index (RGI) region growing method is now being implemented at the very first stage of the mass detection algorithm in order to increase the sensitivity for mass detection. Thus, this RGI algorithm replaces the bilateral subtraction methodology in the overall computerized mass detection method. The benefit of this change is that cases with unilateral mammograms can be analyzed by the computer method. In addition, the sensitivity of the detection algorithm increased by 15%.

In addition to the RGI filtering method, we have developed a method that uses a Bayesian neural network to merge multiple feature images (RGI being one of them) pixel by pixel. This method, which represents only the preprocessing stage of the algorithm, reaches a similar performance level of the current mass detection method. We are now incorporating this new preprocessing stage with the feature-extraction stage for further improvements.

Development of advanced computerized schemes for the classification of masses in digital mammograms.

We are investigating the potential usefulness of computer-aided diagnosis as an aid to radiologists in the characterization and classification of mass lesions in mammography. Ninety-five mammograms containing masses from 65 patients were digitized. Various features related to the margin, shape and density of each mass were extracted automatically from the neighborhoods of the computer-identified mass regions. Selected features were merged into an estimated likelihood of malignancy using three different automated classifiers. The performance of the three classifiers in distinguishing between benign and malignant masses was evaluated by receiver operating characteristic (ROC) analysis, and compared with those of an experienced mammographer and of five less experienced mammographers. Our computer classification scheme yielded an A_Z value of 0.94, similar to that of an experienced mammographer (A_Z=0.91) and statistically significantly higher than the average performance of the radiologists with less mammographic experience (A_Z=0.80). With the database we used, the computer scheme achieved, at 100% sensitivity, a positive predictive value of 83%, which was 12% higher than that of the experienced mammographer and 21% higher than that of the average performance of the less experienced mammographers at a *p*-value of less than 0.001. Thus, automated computerized classification schemes may be useful in helping radiologists distinguish between benign and malignant masses.

We have also investigated the effect of dominant features on neural network performance in the task of classification of mammographic lesions. Two different classifiers, an artificial neural network (ANN) and a hybrid system (one step rule-based method followed by an artificial neural network) were investigated to merge computer-extracted features in the classification of malignant and benign masses. Four computer-extracted features were used in the study: spiculation, margin sharpness and two density-related measures. ROC analysis showed that the hybrid system performed significantly better than the ANN method at the high sensitivity levels, yielding an A_z of 0.94 with a specificity of 69% at 100% sensitivity, whereas, the ANN method yielded an A_z of 0.90 with a specificity of 19% at 100% sensitivity. To understand the difference between the two classifiers in their performance, we investigated their learning and decision-making processes by studying the relationships between the outputs and input features. The correlation study showed that the outputs from the ANN alone method strongly correlated with one of the input features (spiculation measure), yielding a correlation coefficient of 0.91 while the correlation coefficients (absolute value) for the other features range from 0.19 to 0.40. The strong correlation between the ANN output and spiculation measure indicates the learning and decision-making processes of the ANN alone method was dominated by the spiculation measure. A series of three-dimensional plots of the computer output as functions of the input features demonstrate that the ANN method did not learn as effectively as the hybrid system from the other three features in differentiating subtle (non-spiculated) malignant masses from benign masses, thus, resulting in the inferior performance at the high sensitive levels. We found that with a limited database, it is detrimental for an ANN to learn the significance of other features in the presence of a dominant feature. The hybrid system, which initially applied a rule on the spiculation measure prior to an ANN, prevents the over-learning from the dominant feature and performed better than the ANN alone method in merging the computer-extracted features into a correct diagnosis on the malignancy of the masses.

Currently in mammography, the digital image on which CAD analysis is performed is obtained by digitizing a screen-film mammogram. Since the image is sampled when digitized, the digitization of a image using two different scanners will not produce exactly the same digital image (because of different designs, sampling aperture, sampling distance and internal electronic noise, etc. of the laser scanners and the different calibration curves for the transformation of the optical density (OD) to pixel value). Thus, the contrast, noise and resolution of the two images may differ. Thus, as long as CAD analysis relies upon digitized screen-film images, a CAD system (film digitization and computer analysis) may suffer from the variability in the digital formats of a image, which may lead to variations in the performance of the CAD scheme. Two different databases and three different digitizers were involved in this study. One database consisted of 95 mammograms collected from 65 cases: 39 biopsy-confirmed malignant cases, 25 biopsy-confirmed benign cases and one benign case which was determined through more than five years of follow-up. These mammograms were digitized using an optical drum scanner (FIP II, Fuji Film, Tokyo, Japan) at a sampling distance of 0.1 mm and 10-bit quantization. Another database consisted of 110 new cases which were collected from the University of Chicago Radiology files. Of these, 50 cases are biopsy-confirmed malignant, 50 cases are biopsy-confirmed benign diseases and 10 cases are aspiration-confirmed cysts. For each case, two standard views of the affected breast were chosen from a single screening exam. Of the 110 cases, 8 cases had a mass appearing on one view only. Each mammogram this second database was digitized twice using two different laser scanners -- a Konica digitizer (LD 4500; Konica Medical, Wayne, NJ) at 0.1-mm pixel size and 10-bit quantization and a Lumisys laser scanner (Lumiscan 100, Lumisys, Sunnyvale, CA) at a 0.1-mm pixel size and 12-bit quantization. In the evaluation of our classification scheme, both A_z and ${}_{0.90}A_z'$ are important indices. The A_z value was used to evaluate the overall performance, while the partial area index (${}_{0.90}A_z'$) was designed to evaluate the performance of a scheme at a preselected high sensitivity level for those who are interested in knowing the performance at the high sensitivity. In addition, the difference in the partial area index ${}_{0.90}A_z'$ quantitatively evaluates, to some degree, the difference in the shape of the two ROC curves. The differences in A_z between the two digital formats were the same for both the ANN-alone and hybrid classifiers. Two-tailed *p* values obtained from the CLABROC programs showed that the difference in the performance of the classification scheme, due to the difference between the two digitization techniques, using both the ANN and the hybrid classifier were not statistically significant at the level of 0.05 in terms of the A_z and ${}_{0.90}A_z'$.

In order to observe the effect of the computer aid on radiologists' performance in the task of distinguishing between malignant and benign lesions, we performed two observer studies. The first observer study was at RSNA '98. The mass classification method was run on both the MLO and CC views and the magnification views. We showed that the average performance of 128 radiologists (who participated in the study) increased significantly from an A_z of 0.89 to an A_z of 0.94 ($p < 0.05$) when the computer aid was used in distinguishing 20 mass lesions cases. In addition, at RSNA 99, we presented results of an observer study in which the radiologist-observers interpreted 110 mammographic mass lesion cases without and with the computer output of the likelihood of malignancy. Six general radiologists (certified in reading mammograms) and five mammographers participated. The average of performance of the radiologists in term of A_z showed a statistically significant increase when the computer aid was used ($p < 0.05$) for all eleven observers, for just the general radiologists, and for just the mammographers.

At RSNA 99, we presented preliminary results from our investigation of the potential of the computerized mass classification method to digital mammography (106). We retrospectively obtained 96 mass cases imaged on a LORAD small/medium-field digital mammography system at Northwestern University. We first tested the computerized method that has been trained on the screen/film database which yielded an A_z of 0.72 on the digital mammography data. However, after retraining the computerized method, but using the same computer-extracted features, we obtained an A_z of 0.91 in a round-robin analysis. This was similar to the subjective rating given by the radiologists during their clinical workup of the cases (0.92). We concluded that the CAD screen/film method could be ported over for use with digital mammography after recalibration of the parameters in the computerized method. We are continuing to investigate the differences as well as examine the porting of the screen/film computerized classification method to FFDM.

Development of computerized methods based on multiple views for enhanced mammographic interpretation

We have evaluated the potential benefit of incorporating a temporal subtraction scheme with our bilateral subtraction technique for improving the sensitivity of mass detection. A database of 79 cases was used, each of which contained a lesion in at least the current exam. Two methods for image registration of the temporal images were investigated: one used translation and rotation based on computer-determined skin lines and the other used a warping technique based on the cross-correlation of regions of interest located throughout the parenchyma. The characteristics of the false-positive detections resulting from the bilateral subtraction and from the temporal subtraction were analysed. The distribution of the true positives and false positives were similar despite the fact that many of the false positives resulting from the two schemes were in different locations in the breast parenchyma. At a false-positive rate of four per image, the combined (Logical OR) scheme detected 85% of the masses, which was 8% greater than the bilateral subtraction technique alone. The combined use of bilateral and temporal subtraction methods shows potential for an improvement in sensitivity in the detection of masses.

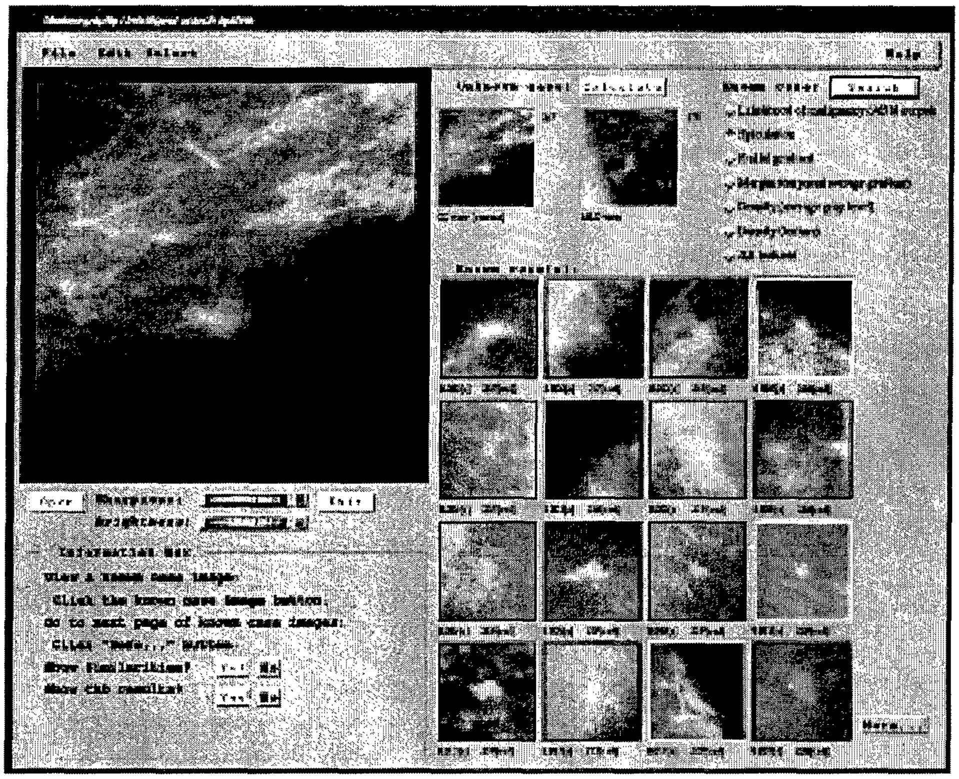
We are investigating how the lesion features as calculated from the CC, MLO, and magnification views vary. To date we have collected approximately 150 cancer cases from digitized films. Spiculation has been shown to be a dominate feature and is influenced by linear-shaped parenchymal structures that transverse the lesion on the 2-D projection image. This is one of the reasons radiologists prefer to have a computer rating given per view as opposed to per case – since the projected view of a lesion and a linear parenchymal pattern could lead to an erroneous increase in the degree of spiculation as calculated by the computer method. It is important that radiologists understand what feature the computer is "looking" at and understand is the computer under-or over calls a lesion. This evaluation is presented later in this report.

Incorporation of the computer-vision methods with an Mammo/Icon mammographic review system for enhanced diagnosis.

Dr. Swetts at Yale has left academics and gone into private practice in Seattle. No grant funds have been transferred to him. Instead researchers on the team at the University of Chicago are creating an "Mammo/Icon-like" system. The features (as well as the merged values from the artificial neural network) from the malignant and benign cases are tabulated and retained in the computer.

We have developed an intelligent search display into which we have incorporated our computerized mass classification method. Upon viewing an unknown mammographic case, the display shows both the computer classification output as well as images of lesions having both known diagnoses (e.g., malignant vs. benign) and similar computer-extracted features. The similarity index used in the search can be selected by the radiologist and can be based on a single feature, multiple features, or on the computer estimate of the likelihood of malignancy. Note that the output of a computer-aided diagnostic scheme can take a variety of forms such as the estimated likelihood that a lesion is malignant either in terms of probabilities or along a standardized rating scale. This information is then available for use by the radiologist as he or she sees fit when making decisions regarding patient management. An alternative approach, which is provided for with the intelligent search workstation, is for the computer to display a variety of lesions that have characteristics similar to the one at hand and for which the diagnosis is known, thereby providing a visual aid for the radiologist in decision making.

Our intelligent workstation recalls lesions in the known database based either on a single feature, multiple features, or a computer-estimate of the likelihood of malignancy. The computer workstation displays similar malignant and benign known cases by use of a color-coding scheme allowing instant visual feedback to the radiologist. (Figure 1) The probability distributions of the malignant and benign cases in the known database are shown by images along with the "location" of the unknown case relative to these two distributions. Features calculated include a spiculation measure, a radial gradient index, margin sharpness, and two density measures. For the example shown, based on the degree of spiculation, similar images from the known database of 175 cases (301 images) are displayed. Each of the known images are enclosed in either a white box corresponding to a benign case or in a black box corresponding to a malignant case. In addition, the user has the option to have the computer-extracted feature value and the "distance in feature value" from the unknown case displayed adjacent to each of the known images. The distribution of malignant and benign cases in the known database of the workstation will affect which cases are displayed when an unknown case is examined. The search for similar images can also be performed using multiple features, the output from an artificial neural network [2] or the likelihood of malignancy, as criteria.



View a known case image:
Click the known case image button.
Go to next page of known case images:
Click "More..." button.
Show Similarities? Yes No
Show CAD results? Yes No

Figure 1. Display interface pf the intelligent search workstation showing a malignant case. Black borders around known cases indicate that they are malignant. White border indicate the lesions are benign.

The intelligent search workstation combines the benefit of computer-aided diagnosis with prior knowledge obtained from confirmed clinical cases. It is expected that the display of known lesions with similar features will aid the radiologist in his/her workup of a suspect lesion, especially when the radiologist's assessment of the lesion differs from the computer output. An observer study involving the workstation is ongoing.

Evaluation of the CAD methods for mammography

Databases are continuously being collected. For mass detection, we have approximately 175 clinical cases of malignant masses. New data for the classification database includes the 150 malignant cases as well as 100 benign cases, which were used in the robustness evaluation, and an additional 100 cases collected for the intelligent search workstation evaluation.

Our computerized mass classification method was independently evaluated on a 110-case database of digitized screen/film mammograms containing 50 malignant and 60 benign mass cases [3]. The effects of variations in both case mix and in film digitization technique were assessed. The method achieved an A_Z value (area under the ROC curve) of 0.90 on the prior training database for categorization of lesions as malignant or benign (Fuji scanner digitization) in a round-robin evaluation, and A_Z values of 0.82 and 0.81 for the independent database with Konica and Lumisys digitization formats, respectively [3]. However, we failed to show a statistical significant difference between the performance on the training database and that on the independent validation database (p-values > 0.10). Thus, our computer-based method for the classification of lesions on mammograms would seem robust to variations in case mix and film digitization technique [3,4].

We evaluated the mass classification method at RSNA 98. The mass classification method was run on both the MLO and CC views and the magnification views. We showed that the average performance of 128 radiologists (who participated in the study) increased significantly from an A_Z of 0.89 to an A_Z of 0.94 ($p < 0.05$) when the computer aid was used in distinguishing 20 mass lesions cases.

We also evaluated the mass classification method on mammographic benign and malignant mass lesions based on the analysis of mammograms of special views. The scheme was trained on a database consisting of 65 cases (95 mammograms of standard views - CC and MLO) and yielded an Az of 0.90 in a round-robin analysis. With no retraining of the scheme, the computerized method was validated on an independent database consisting of 71 cases, each case having mammograms of the two standard views and a special view (a magnified or spot compression view. The classification scheme achieved Az values of 0.78, 0.75 and 0.95 for CC, MLO and special views, respectively, in differentiating between benign and malignant masses. The performance based on the analysis of the special views is significantly better (p=0.0055, 0.0050) than that based on the analysis of the standard mammographic views. Our computerized classification scheme performed well on the independent database. Computerized analysis of special mammographic views is important for the diagnosis of a mammographic mass lesion.

We are also investigating the potential usefulness of computer-aided diagnosis as an aid to radiologists in the characterization of mass lesions on digital mammography. We evaluated our computerized classification method, initially developed using digitized screen/film mammograms, on a database of digital mammograms. We retrospectively collected 212 consecutive digital mammograms from 110 patients obtained with a LORAD stereotactic imaging system. These images had initially been performed for needle localization or core biopsy of a suspect mass lesion. The database consisted of 44 malignant cases and 66 benign cases. The computer classification method, as described above for the digitized screen/film mammogram study, includes automated segmentation of the mass lesions from the breast parenchyma, automated extraction of lesion features, and automated classification of the suspect lesion into an estimate of the likelihood of malignancy. It should be noted, however, that in this study, the automatic classification was performed by a Bayesian neural network (BANN). The BANN was used to merge the four features of spiculation, margin sharpness, average gray level, and texture. The BANN uses regularization to prevent overtraining of the network [5,6]. The computerized classification method , which incorporated the BANN, was trained on our screen/film database and yielded an A_z of 0.90 in the training. This trained computer method yielded an A_z of 0.79 on the independent digitized S/F database and an A_z of 0.71 on the independent digital mammography (LORAD) database. Note here that the computer method/network weights and parameters were set using the digitized screen/film database. The BANN was then re-trained using the digital mammographic images from the LORAD digital system. Thus, the structure, weights, and parameters of the network changed although the same features were automatically extracted from each mass lesion. After re-training of the BANN, the A_z reached a value of 0.89 for the digital mammographic images. Radiologists' ratings of suspicion, from the clinical interpretation, of the same lesions on the basis of prior screen/film images, the LORAD digital images, and clinical data achieved a similar A_z (0.92). Further investigation of the features in the study showed that the spiculation feature performed better on the screen/film database, whereas the texture feature performed better on the digital mammography database. In summary, we have extended our computerized method for the characterization of mass lesions on mammography to the analysis of mammographic images obtained directly from a digital system. Results indicate that the computer-extracted features are robust and can be used to classify lesion on digital mammography. However, retraining of the classifier, which uses the extracted features as input, may be require "calibration." That is, due to differences in the physical characteristics of the two image acquisition systems, however, the classifier may have to be retrained with images obtained using the same modality to optimize performance.

KEY RESEARCH ACCOMPLISHMENTS

1. Improvements in the computerized detection of mass lesions on mammograms
   - Incorporation of temporal image data
   - Development of a new lesion extraction (region growing) method
   - Development of new single image detection method instead of bilateral subtraction
   - Development of a Bayesian neural network technique for merging features as well as feature images
   - Investigation into feature selection and feature merging with limited datasets
   - Development of a new pixel-based multiple-feature image filtering technique

2. Improvements in the computerized classification of mass lesions on mammograms
   - Investigation of a hybrid (rule-based plus ANN) system for classification
   - Validation on an independent database showing robustness of the method
   - Incorporation of special views, beyond just MLO and CC views, in the computer analysis

3. Incorporation of the CAD methods into an intelligent search workstation
   - Expansion of known database descriptors to include CAD derived parameters
   - Development of a similarity index for extracting similar cases from the known database
   - Development of the intelligent search workstation

4. Evaluation of the CAD methods for mammography.
   - Evaluation of the computerized mass detection method on 10,000 consecutive screening mammography cases
   - Evaluation of the computerized mass classification method on an independent database of cases of cancers and benign cases
   - Evaluation of the computerized mass classification method (developed on digitized screen/film mammograms) on digital mammograms (from a LORAD unit)
   - Evaluation (by way of observer studies) of the computerized method as an aid to radiologists in the task of distinguishing between malignant and benign lesions and in recommending biopsy.

REPORTABLE OUTCOMES

PAPERS

- Huo Z, Giger ML, Vybomy CJ, Wolverton DE, Schmidt RA, Doi K: Automated computerized classification of malignant and benign mass lesions on digitized mammograms. Academic Radiology 5: 155-168, 1998.

- Kupinski MA, Giger ML: Automated seeded lesion segmentation on digital mammograms. IEEE Trans on Medical Imaging, 17: 510-517, 1998.

- Doi K, Giger ML, Nishikawa RM, Schmidt RA: Computer vision and artificial intelligence in mammography, Proc. of the International Symposium on Diagnosis and Therapy of Breast Cancer, Germany, [In] Diagnostik und Therapie des Mammakarzinoms: State ofthe Art, editors M. Untch, G. Konecny, H. Sittek, M. KeBler, M. Reiser, and H. Hepp, Published by W. Zuckschwerdt Verlag pgs. 11-16, 1998.

- Huo Z, Giger ML, Metz CE: Effect of dominant features on neural network performance in the classification of mammographic lesions. PMB 44: 2579-2595, 1999.

- Kupinski MA, Giger ML: Feature selection with limited datasets. Medical Physics 26: 2176-2182, 1999.

- Giger ML: Current issues in CAD for mammography. Digital Mammography '96. Proceedings of the 3rd International Workshop of Digital Mammography, Elsevier, New York, pp. 53-60, 1996.

- Nishikawa RM, Schmidt RA, Papaioannou J, Osnis RB, Haldemann RA, Giger ML, et al.: Performance of a prototype clinical "intelligent" mammography workstation. Digital Mammography '96. Proceedings of the 3rd International Workshop of Digital Mammography, Elsevier, New York, pp. 93-96, 1996.

- Bick U, Giger ML, Schmidt RA, Nishikawa RM, Wolverton DE, Doi K: Computer-aided breast cancer detection in screening mammography. Digital Mammography '96. Proceedings of the 3rd International Workshop of Digital Mammography, Elsevier, New York, pp. 97-104, 1996.

- Schmidt RA, Nishikawa RM, Osnis RB, Schreibman KL, Giger ML, et al.: Computerized detection of lesions missed by mammography. Digital Mammography '96. Proceedings of the 3rd International Workshop of Digital Mammography, Elsevier, New York, pp. 105-110, 1996.

- Huo Z, Giger ML, Vybomy CJ, Wolverton DE, Schmidt RA, Doi K: Computer-aided diagnosis: Automated classification of mammographic mass lesions. Digital Mammography '96. Proceedings of the 3rd International Workshop of Digital Mammography, Elsevier, New York, pp. 207-212, 1996.

- Zhang M, Giger ML, Vybomy CJ, Doi K: Mammographic texture analysis for the detectio of spiculated lesions. Digital Mammography '96. Proceedings of the 3rd International Workshop of Digital Mammography, Elsevier, New York, pp. 347-350, 1996.

- Kupinski M, Giger ML, Doi K: Optimization of neural network inputs with genetic algorithms. Digital Mammography '96. Proceedings of the 3rd International Workshop of Digital Mammography, Elsevier, New York, pp. 401-404, 1996.

- Zouras WK, Giger ML, Lu P, Wolverton DE, Vybomy CJ, Doi K: Investigation of a temporal subtraction scheme for computerized detection of breast masses in mammograms. Digital Mammography '96. Proceedings of the 3rd International Workshop of Digital Mammography, Elsevier, New York, pp. 411-416, 1996.

- Huo Z. Giger ML: Integrating rules and artificial neural networks in the classification of mass lesions in digital mammograms. Proc. World Congress on Neural Networks '96, vol 1, pgs. 1166-1169, 1996.

- Kupinski MA, Giger ML: Feature selection and classifiers for the computerized detection of mass lesions in digital mammography. Proc. IEEE International Congress on Neural Networks, (ICNN '97), pp. 1336-13339, IEEE/ICNN, 1997.

- Giger ML, Nishikawa RM, Kupinski MA, Bick U, Zhang M, Schmidt RA, et al.: Computerized detection of breast lesions in digitized mammograms and results with a clinically-implemented intelligent workstation. CAR'97 pgs. 325-330, 1997, 1997.

- Nishikawa RM, Giger ML, Jiang Y, Huo Z, et al.: Automated classification of breast lesions on digital mammograms. CAR'97 pgs. 347-351, 1997.

- Kupinski M, Giger ML: Investigation of regularized neural networks for the computerized detection of mass lesions in digital mammograms. Proc. EMBS'97 , pp. 1336-1339, 1997.

- Giger ML, Nishikawa RM, Vybomy CJ, Schmidt RA, Wolverton DE, Comstock C, Metz CE, Doi K: Development of methods for computerassisted interpretations of digital mammograms for early breast cancer detection. Proc. Era of Hope, Department of Defense Breast Cancer Research Program Meeting, Vol. I, pgs. 83-84, 1997.

- Nishikawa RM, Giger ML, Wolverton DE, Schmidt RA, Comstock CE, Papaioannou J, Collins SA, Doi K: Prospective testing of a clinical mammography workstation for CAD: Analysis of the first 10,000 cases. Digital Mammography 1998, Eds.: Karssemeijer N, Thijssen M, Hendriks J, van Erning L. Kluwer Academic Publishers, Dordrecht, The Netherlands, pp. 401-406, 1998.

- Giger ML, Huo Z, Wolverton DE, Vybomy C, Moran C, Schmidt RA, Al-Hallaq H, Nishikawa R, Doi K: Computer-aided diagnosis of digital mammographic and ultrasound images of breast mass lesions. Digital Mammography 1998, Eds.: Karssemeijer N, Thijssen M, Hendriks J, van Erning L. Kluwer Academic Publishers, Dordrecht, The Netherlands, pp. 143-147, 1998.

- Anastasio MA, Kupinski MA, Nishikawa RM, and Giger ML: A multiobjective approach to optimizing computer aided diagnosis schemes, Proceedings of the 1998 IEEE Nucl. Sci. Symposium and Med. Imaging Conference, Toronto, Canada, 1998.

- Giger ML, Huo Z: Artificial neural netwroks in breast cancer diagnosis: Merging of computer-extracted features from breast images. Proc. Of Conference on Evolutionary Computing (CEC'99), pp. 1768-1769, 1999.

- Huo Z, Giger ML: Robustness of a computerized scheme for the classification of malignant and benign masses on digitized mammograms. In: K Doi, H MacMahon, ML Giger, KR Hoffmann, eds. Computer-Aided Diagnosis in Medical Imaging: Proceedings of the 1st International Workshop on Computer-Aided Diagnosis, Elsevier, Amsterdam, 277-280, 1999.

- Nishikawa RM, Giger ML, Schmidt RA, Wolverton DE, Doi K: Prospective testing of a clinical CAD workstation for the detection of breast lesions on mammograms. In: K Doi, H MacMahon, ML Giger, KR Hoffmann, eds. Computer-Aided Diagnosis in Medical Imaging: Proceedings of the 1st International Workshop on Computer-Aided Diagnosis, Elsevier, Amsterdam, 209-214, 1999.

- Kupinski MA, Giger ML: Multiobjective genetic optimization of diagnostic classifiers used in the computerized detection of mass lesions in mammography, SPIE Medical Imaging Conference , 2000, (in press).

- Huo Z, Giger ML: Incorporation of clinical data into a computerized method for the assessment of mammographic breast lesions. SPIE Medical Imaging Conference Proc. SPIE, 2000 (in press).

- Huo Z, Giger ML: Evaluation of an automated segmentation method based on performances of an automated classification method. Proc. SPIE, 2000 (in press).

- Kupinski MA, Giger ML: A comparison of Bayesian ANN and multiobjective training using limited datasets. Proc. CARS 2000 (in press).

- Giger ML, Huo Z, Lan L, Vyborny CJ: Intelligent search workstation for computer-aided diagnosis. Proc. CARS 2000 (in press).

- ABSTRACTS

- Kupinski MA, Giger ML, Doi K: Use of genetic algorithms in the computerized detection of masses in digital mammograms. Med Phys 23:1133, 1996.

- Giger ML: Update course on technical aspects of breast imaging: Computer aided diagnosis in mammography (refresher course). Radiology 205: 118, 1997.

- Huo Z, Giger ML, Vyborny CJ, Wolverton DE, Schmidt RA, Doi K: Automated computerized classification of malignant and benign mass lesions on digitized mammograms (scientific presentation). Radiology 205: 275, 1997.

- Comstock CE, Giger ML, Nishikawa RM, Wolverton DE, Schmidt RA, Doi K, et al.: Two-year clinical evaluation of computer aided diagnosis (CAD) in the detection of breast cancer (scientific presentation). Radiology 205: 275, 1997.

- Nishikawa RM, Giger ML, Comstock CE, Papaioannou J, Urbas AM, Doi K, et al.: Performance of a prototype clinical mammography workstation for computer-aided diagnosis (CAD) (scientific presentation). Radiology 205: 217, 1997.

- Doi K, Giger ML, Nishikawa RM, Hoffmann KR, Schmidt RA, MacMahon H, et al.: Computer-aided diagnostic schemes in mammography, chest radiography, angiography, and computed tomography (scientific exhibit). Radiology 205: 624, 1997.

- Nishikawa RM, Giger ML, Jiang Y, Yoshida H, Schmidt RA, Doi K, et al.: Computer-aided diagnosis for the detection and classification of breast lesions (infoRad exhibit). Radiology 205: 740, 1997.

- Kupinski M, Giger ML: Probabilistic Lesion segmentation in digital mammography. Med Phys 25: A111, 1998.

- Giger ML: Computer-aided diagnosis in breast imaging. Radiology 209:130, 1998.

- Kupinski MA, Giger ML: Computer-aided diagnosis: Feature selection with limited datasets. Radiology 209: 163, 1998.

- Nishikawa RM, Giger ML, Schmidt RA, Wolverton DE, Collins SA, Doi K, et al.: Computer-aided diagnosis in screening mammography: Detection of missed cancers. Radiology 209: 353, 1998.

- Jiang Y, Nishikawa RM, Giger ML, Huo Z, Schmidt RA, Wolverton DE: Computer-aided diagnosis (CAD) of breast lesions: An interactive demonstration. Radiology 209: 520, 1998.

- Doi K, Giger ML, Nishikawa RM, Hoffman KR, MacMahon H, Schmidt RA, et al.: Computer-aided diagnosis: From lab to practice. Radiology 209, 584, 1998.

- Giger ML, Nishikawa RM, Huo Z, Jiang Y, Wolverton DE, Doi K, et al.: Computer-aided diagnosis in breast imaging. Radiology 209: 673, 1998.

- Huo Z, Giger ML, Vyborny C, Metz C, Wolverton DE: Validation of a computerized method for the diagnosis of mammographic lesions. Med Phys 26: 1064, 1999.

- Kupinski M, Edwards D, Giger ML, Baehr A: Bayesian artificial neural networks in the computerized detection of mass lesions. Med Phys 26: 1081, 1999.

- Huo Z, Giger ML, Vyborny C, Jiang Y, Nishikawa R, Engelmann R: Effectiveness of computer aid for radiologist's classification of mammographic mass lesions. Radiology 213: 200, 1999.

- Nishikawa R, Giger ML, Yarusso L, Kupinski M, Baehr A, Venta L,: Computer-aided diagnosis (CAD) of images obtained on full-field digital mammography. Radiology 213: 229,1999.

- Kupinski M, Giger ML, Baeher A: Computerized detection of mass lesions in digital mammography using radial gradient index filtering. Radiology 213: 229, 1999.

- Maloney M, Huo Z, Giger ML, Venta L, Vyborny C: Computerized classification of mass lesions on images from a medium-field digital mammography unit. Radiology 213: 230, 1999.

- Giger ML, Nishikawa R, Huo Z, Jiang Y, Venta L, Doi K: Computer-aided diagnosis (CAD) in breast imaging. Radiology 213: 507, 1999.

Personnel who received pay from this research effort:

Maryellen Giger
Kunio Doi
Charles Metz
Robert Schmidt
Zhimin Huo
Kenneth Gilhuijs
Samuel Armato
Alexandra Baehr
Young-Jin Kim
Catherine Moran
Matthew Maloney
Chun-Wai Chan
Augustine Urbas
Wendy Zouras
Jordan Samuels
Michael Carlin

CONCLUSIONS

The research supported by this grant has produced improvements for the computerized detection and classification of lesions on mammography. The research has also provided investigators with new methods for feature selection, feature merging (i.e., classifiers), and methods for evaluation. In addition, observer studies from this research have shown at a statistically significant level that use of the computer aid does improve the diagnostic decision making of radiologists. This work is ready to be translated to clinical environment.

Phys. Med. Biol. 44 (1999) 2579–2595. Printed in the UK PII: S0031-9155(99)03913-5

Effect of dominant features on neural network performance in the classification of mammographic lesions

Zhimin Huo, Maryellen L Giger† and Charles E Metz

Kurt Rossmann Laboratories for Radiologic Image Research, Department of Radiology,
5841 South Maryland Avenue, The University of Chicago, Chicago, IL 60637, USA

Received 4 May 1999, in final form 17 June 1999

Abstract. Two different classifiers, an artificial neural network (ANN) and a hybrid system (one step rule-based method followed by an artificial neural network) have been investigated to merge computer-extracted features in the task of differentiating between malignant and benign masses. A database consisting of 65 cases (38 malignant and 26 benign) was used in the study. A total of four computer-extracted features—spiculation, margin sharpness and two density-related measures—was used to characterize these masses. Results from our previous study showed that the hybrid system performed better than the ANN classifier. In our current study, to understand the difference between the two classifiers, we investigated their learning and decision-making processes by studying the relationships between the input features and the outputs. A correlation study showed that the outputs from the ANN-alone method correlated strongly with one of the input features (spiculation), yielding a correlation coefficient of 0.91, whereas the correlation coefficients (absolute value) for the other features ranged from 0.19 to 0.40. This strong correlation between the ANN output and spiculation measure indicates that the learning and decision-making processes of the ANN-alone method were dominated by the spiculation measure. Three-dimensional plots of the computer output as functions of the input features demonstrate that the ANN-alone method did not learn as effectively as the hybrid system in differentiating non-spiculated malignant masses from benign masses, thus resulting in an inferior performance at the high sensitivity levels. We found that with a limited database it is detrimental for an ANN to learn the significance of other features in the presence of a dominant feature. The hybrid system, which initially applied a rule concerning the value of the spiculation measure prior to employing an ANN, prevents over-learning from the dominant feature and performed better than the ANN-alone method in merging the computer-extracted features into a correct diagnosis regarding the malignancy of the masses.

1. Introduction

Various classifiers (Gale *et al* 1987, Getty *et al* 1988, Cook and Fox 1989, Swets *et al* 1991, Wu *et al* 1993) are being investigated for use in merging computer-extracted image features and in medical decision making, such as in the classification of mass lesions in mammography. Mammographic classification of mass lesions is a difficult task, because mass lesions vary in appearance and similar attributes are shared by some benign and malignant masses (Tabar and Dean 1985, Sickles 1991, D'Orsi and Kopans 1993, Knutzen and Grisvold 1993). In addition, it has been shown that general radiologists can extract individual features from radiographs at a level similar to that of experienced mammographers (Getty *et al* 1988). However, general radiologists differ from experienced radiologists in their ability to merge extracted information into a correct diagnostic decision. Researchers have shown that computer-based systems reached a correct diagnosis more often than did general radiologists (Gale *et al* 1987, Getty *et al*

† Author to whom correspondence should be addressed.

1988, Swets *et al* 1991, Wu *et al* 1993). Studies have shown also that computer classifiers were able to merge features extracted by a computer into a correct diagnosis at a level similar to that of expert radiologists (Jiang *et al* 1996, Huo *et al* 1998).

Computer-based systems, however, have limitations in their learning abilities and decision-making processes when they are used alone. For example, while a rule-based expert system can adopt pre-existing knowledge and, thus, can employ prior information directly, there are limitations in the availability of knowledge and knowledge usage. Also, the translation of particular 'intuitive' knowledge into rules may be difficult and even detrimental if expressed incorrectly. Artificial neural networks, on the other hand, are able to learn ill-defined relationships from noisy examples and, in this way, can acquire their own knowledge that can be used to classify new cases if the ANNs have the proper architecture and are 'taught' with a sufficiently large number of training data (Haykin 1994). However, artificial neural networks may not provide users with explanations about the internal decisions and may not be able to incorporate well-established prior knowledge. Moreover, it is uncertain in some situations as to whether a final learning goal of an ANN is achieved, because overtraining or undertraining of an ANN may occur when only a limited sample of training data is available (Haykin 1994). To emulate humans in their learning and decision-making for particular practical problems, two or more types of computer classifiers may need to be integrated into a hybrid system in order to overcome the limitations of each kind of system and, thereby, to improve the learning and decision-making processes.

Rule-based approaches have be used as classification tools in making diagnostic decisions (Cook and Fox 1989). Also, artificial neural networks evolved into one of the major alternative methods in medical decision-making as the ability of ANNs to learn and generalize became recognized in the field (Wu *et al* 1993, Jiang *et al* 1996, Lo *et al* 1997). Levels of decision performance obtained with different classifiers, such as rule-based methods and artificial neural networks, have been compared (Nagel *et al* 1995, Katsuragawa *et al* 1997, Huo *et al* 1998). Results showed that combined methods yielded better performance than either the rule-based methods or the artificial neural networks (Nagel *et al* 1995, Katsuragawa *et al* 1997, Huo *et al* 1998); however, the differences in the levels of performance were not investigated.

We have developed a computerized scheme that differentiates malignant masses from benign masses on the basis of information in digitized mammograms (Huo *et al* 1995, 1998). The scheme automatically extracts four radiographic features, similar to those used by radiologists in the classification of masses, from a mass and merges these features into an estimated likelihood of malignancy (Huo *et al* 1998).

We report here a study in which we investigated the effect of having a dominant feature on the training and performance of classifiers used in the classification of mass lesions. We studied the advantages of a combined rule-based and ANN system (i.e. a hybrid system) over a method that employs an ANN alone in merging the four computer-extracted features into a correct diagnosis. We present a detailed discussion about the advantages and the limitations of the two classifiers in their ability to learn the significance of the four features in classifying masses in a database. In addition, to understand the difference between the two classifiers, we studied the relationships between the input features and the classifier output.

2. Materials and methods

2.1. Database

Ninety-five mass-containing mammograms were collected from 65 patients: 39 with breast cancer and 26 with benign breast disease. Sixty-four of the 65 patients were biopsied for the suspicion of breast cancer; the remaining one was deemed benign by more than 5 years follow-up. The mammograms were digitized with an optical drum scanner (FIP II, Fuji Film, Tokyo, Japan) at a sampling distance of 0.1 mm and 10-bit quantization. Two expert mammographers characterized the 95 mammographic masses in terms of their margin spiculation, shape, density and size (Huo *et al* 1998). According to the subjective opinion of two expert mammographers, the database represents a typical clinical distribution of mammographic masses in terms of margin, shape, density and size, though not in terms of the ratio of benign to malignant cases.

2.2. *Computerized classification method*

Our computerized classification method has been presented and evaluated elsewhere (Huo *et al* 1998). We summarize its components and status here. We used this classification method in our study of the effect of dominant features on performance. The following two sections review the computer-extracted features and classifiers used in the classification methods.

2.2.1. Computer-extracted features. The four computer-extracted features used in the study are spiculation, margin sharpness, mass density and texture measure, and have been discussed in detail elsewhere (Huo *et al* 1998). The lesion is first automatically extracted from the parenchymal background in the mammographic image. The features are then automatically extracted from the neighbourhood of each mammographic mass.

Degree of spiculation of a mass is defined as the average angle (in degrees) by which the direction of the maximum gradient along the margin of a mass deviates from the radial direction (Huo *et al* 1995). The margin sharpness measure is defined as the average magnitude of the maximum gradients along the margin of the mass and is used to characterize the margin of a mass as well-defined, partially well-defined or ill-defined (Huo *et al* 1995, 1998). The density of a mass is quantified by both the average grey level (opacity) within a mass and a texture measure, which is the standard deviation of the gradients within a mass (Huo *et al* 1998). In clinical practice, spiculation is the major diagnostic feature for malignancy that is used by radiologists, with a spiculated mass having a greater than 95% probability of being a cancer (Kopans 1989). An ill-defined mass is associated with a higher probability of malignancy than a well-defined mass. A mass with higher radiographic density is associated with a higher probability of being a malignant mass than one with lower radiographic density.

It should be noted that we used two features to quantify the density of a mass because it is difficult to quantitatively assess the density of a mass radiographically. A mass is a three-dimensional object. Factors that include overlying tissue and x-ray exposure conditions affect measures of mass density that are based upon the absolute value of grey level. Therefore, the texture feature was employed to quantify the density of a mass from a different perspective, by characterizing patterns that arise from veins, trabeculae and other structures which may be visible through a low-density mass, but not through a high-density mass. A mass of low radiographic density should have a low value of average grey level and a high value of the texture feature, whereas a mass of high radiographic density should have a high value of average grey level and a low value of the texture feature. Since the average grey level within a mass depends on the x-ray exposure condition, the average grey-level measure may not be as robust as the other three features, which are gradient-based measures and depend only on the relative values of absolute grey values. However, variations in digitization may affect the performances of these features. In our study, the mammograms in the database were digitized using the same digitizer under the same condition. In a separate study using an independent database (110 cases), we digitized the database twice using two different digitizers. Results from that

Table 1. Performance of the four individual computer-extracted features in differentiating between malignant masses and benign masses in terms of A_z based on ROC analysis of the 95 mammographic mass images and of the 36 mammographic mass images after the spiculation cutoff.

Features	A_z (entire database) $n = 95$	A_z (non-spiculated masses) $n = 36$
Spiculation	0.88	0.53
Sharpness	0.56	0.68
Average grey level	0.65	0.66
Texture measure	0.54	0.71

independent study showed that the classification method is robust to variations in digitization technique (Huo 1998).

Our previous studies demonstrated that these computer-extracted features agree well with radiologists' visual impressions in characterizing benign and malignant mass lesions (Huo *et al* 1995, 1998). The individual abilities of these four features to differentiate malignant from benign masses for the 95 mammographic mass images in our database were evaluated with ROC analysis (Metz 1986, 1989). The area under a fitted ROC curve, A_z, was generated from the ROC analysis as an index to evaluate the performance of the individual features. The spiculation measure is more significant than the other three features in distinguishing between benign and malignant masses, yielding an A_z value of 0.88 compared with A_z values ranging from 0.54 to 0.65 for the other three features, as shown in table 1.

It is interesting to note that spiculation is not only a clinically important feature used by radiologists in the classification of mass lesions, but also the dominant feature for our computer scheme. In our previous study, we compared the performance of the spiculation feature alone with that of an experienced mammographer's spiculation ratings. Our analysis showed that the computer-extracted spiculation feature performed at a level ($A_z = 0.88$) similar to that of an expert mammographer's spiculation rating ($A_z = 0.85$) in terms of ability to distinguish between benign and malignant masses, and correlated well with the experienced mammographer's spiculation ratings ($r = 0.63$; $p < 0.0001$) (Huo *et al* 1995).

2.2.2. Automated classifiers. Figure 1 illustrates the structures of the artificial neural network (ANN) and the hybrid system that were used as classifiers in our study. In the ANN-alone method, the ANN had four input units (each corresponding to a computer-extracted feature), two hidden units and one output unit. All four features were input to the four-input ANN used in the ANN-alone method. In the hybrid system, the spiculation measure was input to the first part of the hybrid system (i.e. the rule-based component), which initially classifies the masses into 'spiculated' and 'non-spiculated' categories. Masses classified as 'non-spiculated' in this way were analysed subsequently by an artificial neural network that had three input units (corresponding to the three computer-extracted features other than spiculation measure), two hidden units and one output unit. Both artificial neural networks were trained using an error back-propagation algorithm with a sigmoid activation function (Haykin 1994).

The four-input ANN was trained on the entire database. The round-robin (i.e. leave-one-out) method (Gong 1986) was used to test the generalization ability of the ANN architecture for this data set. In the round-robin method, all cases but one were used to train the neural network, with the single left out case used to test the neural network. To avoid bias for cases with images having two views (medio-lateral and cranio-caudal views) of the breast, both

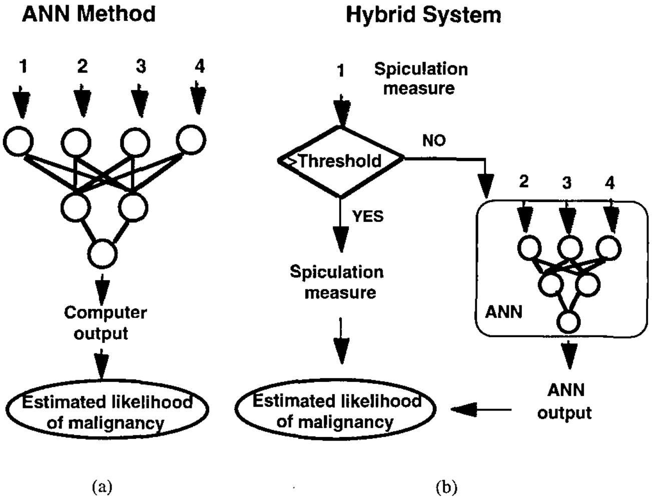


Figure 1. The structures of (a) the ANN-alone method and (b) the hybrid classifier used in the study.

images of the pair were left out in the round-robin training, and the higher of the two values so obtained was reported as the estimated likelihood of malignancy for that case. This procedure was repeated for all the cases so that every case in the database serves once as a cross-validation case. The ANN used in the hybrid system was trained and tested on the non-spiculated masses only, again using the round-robin technique.

ROC analysis (Metz 1986, 1989, Metz *et al* 1998) was employed to evaluate the performance of the two classifiers in their ability to merge the four computer-extracted features into a correct diagnosis regarding malignancy. The statistical significance of differences in the area under fitted binormal ROC curves, A_z, and of differences in a partial-area index, ${}_{\mathrm{TPF}}A'_z$, was tested by using a modified version of our CLABROC (Metz *et al* 1984, 1998, Jiang *et al* 1996) algorithm. The partial area index ${}_{\mathrm{TPF}}A'_z$ is the portion of the area under the ROC curve that lies above a preselected sensitivity threshold (TPF > selected threshold) in a conventional ROC graph divided by the constant (1-TPF) (Jiang *et al* 1996). It should be noted that the ROC curve for the hybrid system was determined using the spiculation values for the lesions that 'passed' the 'spiculation rule' and using the output of the ANN for the lesions that did not pass the rule.

Results from our previous study showed that the overall levels of classification performance in terms of A_z value are 0.94 and 0.90 for the hybrid classifier and the ANN-alone method respectively in the task of distinguishing between benign and malignant masses (Huo *et al* 1998). The ROC curves from the round-robin ANN outputs of the two classifiers are shown in figure 2. As can be seen from the ROC curves (figure 2), the difference in the performance of the two classifiers is mainly due to the difference in the upper parts of the ROC curves. Although the difference in A_z for the two classifiers is not statistically significant ($p = 0.2$), the differences in partial areas (Jiang *et al* 1996) at the high-sensitivity levels (TPF > 0.90 and TPF > 0.80) are statistically significant ($p = 0.007$ and $p = 0.036$ respectively), as shown in table 2. The performance at the high-sensitivity levels, particularly above 90%, is clinically relevant, because high sensitivity is demanded for mammography in order to maximally reduce the mortality of breast cancer. Therefore, the statistically significant difference between the performance of two classifiers at high sensitivity is important.

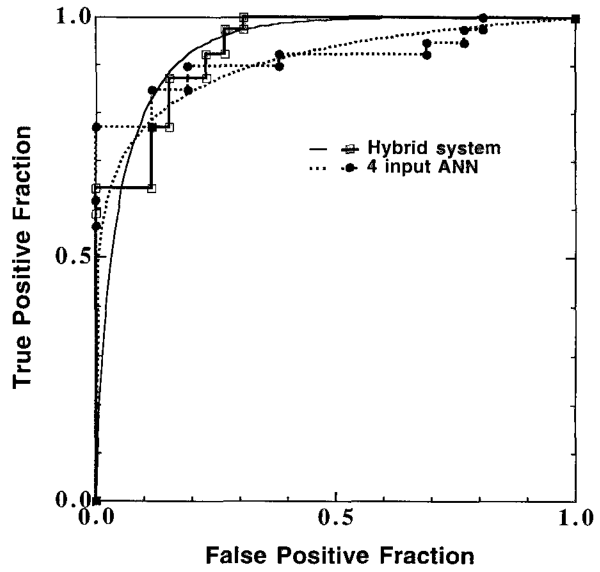


Figure 2. The ROC curves of the ANN-alone and hybrid classifiers in distinguishing malignant masses from benign masses. Individual symbols indicate the empirical (TPF, FPF) points prior to ROC curve fitting.

Table 2. Performances of the ANN and hybrid classifiers in distinguishing between malignant masses and benign masses in terms of A_z, ${}_{0.90}A_z'$ (TPF > 0.90) and ${}_{0.80}A_z'$ (TPF > 0.80) based on ROC analysis of the 65 cases.

	A_z	${}_{0.90}A_z'$ (TPF > 0.90)	${}_{0.80}A_z'$ (TPF > 0.80)
ANN-alone method	0.90	0.40	0.58
Hybrid system	0.94	0.73	0.80
Two-tailed p-value	0.2	0.007	0.036

It should be noted that the two ROC curves (figure 2) are quite similar in the range of sensitivities below 0.80, which indicates that the abilities of the two classifiers are similar in distinguishing obvious malignant masses from benign masses. On the other hand, the difference in the upper parts of the ROC curves indicates that the two classifiers differ in distinguishing subtle malignant masses from benign masses.

3. Results

The hybrid system was introduced in this study to improve the learning process of the classifier. In clinical practice, many radiologists look for spiculation first in their determination of the likelihood of malignancy of a mass due to the high specificity of this measure. Thus, in our study, we first applied a threshold on the spiculation measure to classify masses as spiculated or non-spiculated. The masses classified as spiculated were deemed highly suspicious for malignancy and were not subjected to further computer analysis. The likelihood of malignancy for a spiculated mass was determined solely on the basis of its degree of spiculation. The other features were used in the classification of the non-spiculated (i.e. remaining) masses, because these features are important only in the clinical evaluation of non-spiculated masses

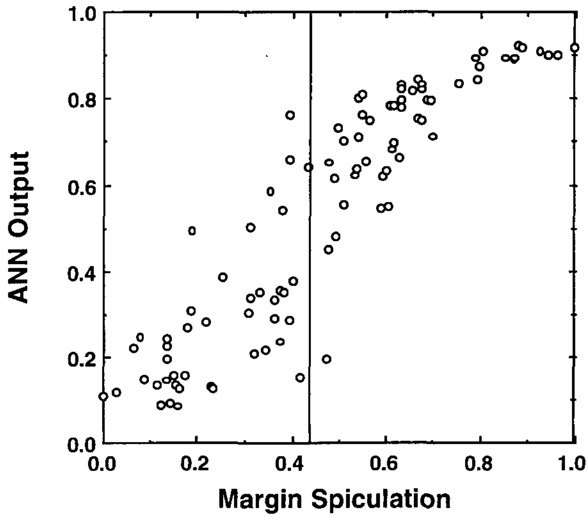


Figure 3. The relationship between outputs from the four-input ANN and the spiculation measures from the 95 mammographic mass images. The vertical line corresponds to the spiculation rule cutoff (critical value) used in the hybrid classifier.

(Tabar *et al* 1985). Our feature analysis showed that these three features were more useful in the evaluation of non-spiculated masses in the sense that the levels of performance of these features, in terms of A_z, improved when they were used to classify only the non-spiculated masses (table 1). These feature are less specific and are interrelated in the determination of malignancy of non-spiculated masses, and thus the merging of these three features is well suited to the use of an ANN.

Intuitively, one might expect that with a sufficient number of hidden units in the ANNs and with sufficient training, the two (i.e. hybrid and ANN-alone) classifiers would be able to perform at a similar level in differentiating both the spiculated and non-spiculated malignant masses from the benign masses. In other words, instead of the sequential learning process in the hybrid system (rule-based on spiculation followed by ANN to acquire decision rules on the other three features), a four-input ANN trained on the entire database should be able to learn the simple rule applied on the spiculation measure to differentiate the spiculated malignant masses from benign masses *and at the same time* learn the rules on the other three features to differentiate the non-spiculated malignant masses from the benign masses. However, as shown from the ROC analysis, the four-input ANN performed statistically significantly worse than did the hybrid system in differentiating non-spiculated malignant masses from benign masses.

3.1. Correlation between the ANN output and individual input features

To understand how the two classifiers determine the likelihood of malignancy of a mass on the basis of input features, we first studied the correlation between the output from the four-input ANN (ANN-alone method) and each of its four input features for the 95 masses, as well as the correlation between the output from the three-input ANN and each of its three input features for the 36 'non-spiculated' masses. Figure 3 shows the correlation ($r = 0.91$; $p < 0.0001$) between the output from the four-input ANN and its input spiculation measure for the 95 masses. Compared with other features (table 3), the spiculation measure shows a strong correlation with the ANN output, apparently causing the decision-making process to be

Table 3. Correlation coefficients between the outputs from the four-input ANN trained on the entire database and its four input features, and between the outputs from the three-input ANN trained on the non-spiculated masses and its three input features.

Classifier	Margin spiculation	Margin sharpness	Average grey level	Texture measure
Four-input ANN ($N = 95$)	0.91	−0.19	0.36	−0.34
Three-input ANN ($N = 36$)	—	−0.40	0.38	−0.33

dominated by the spiculation measure. This strong correlation also indicates that the four-input ANN trained on the entire database was able to learn the simple spiculation rule and determine the likelihood of malignancy of spiculated malignant masses based on their spiculation measures, thus obtaining a performance similar to that of the hybrid system in differentiating the obvious (spiculated) malignant masses from the benign masses as demonstrated in the lower parts of the ROC curves (figure 2). The spiculation rule cut-off (critical value) used in the hybrid classifier is indicated in figure 3 by a line.

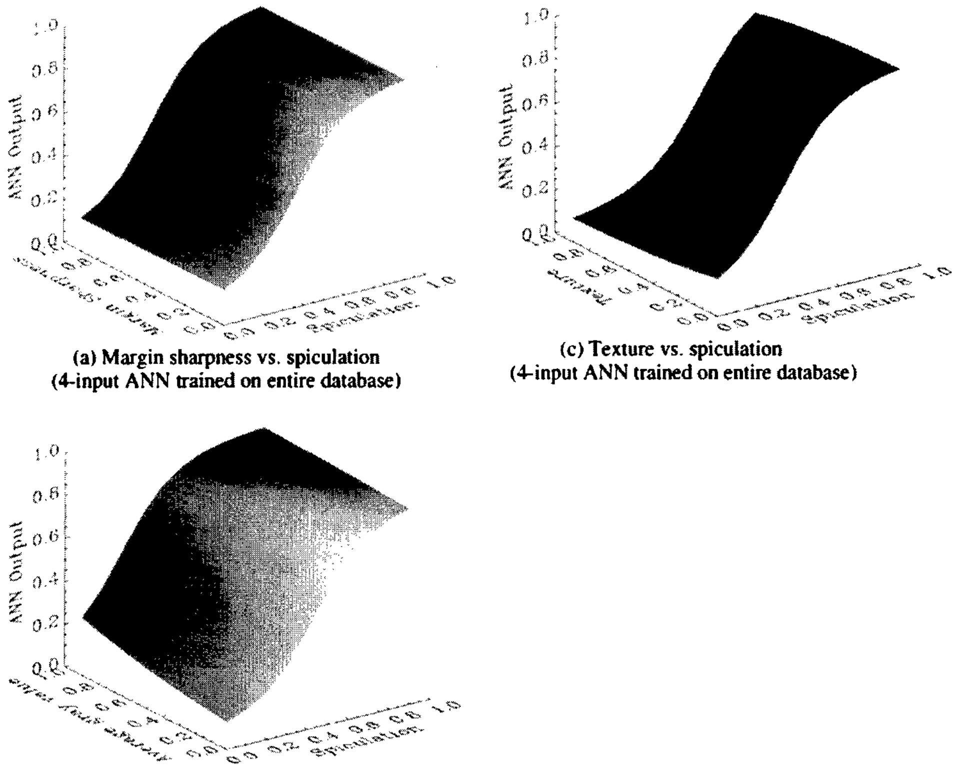
This strong correlation was expected, because clinical experience suggests that the likelihood of malignancy of a mass is determined largely by spiculation and because our computer-extracted spiculation measure performed at the level similar to a radiologist's spiculation ratings. The clinical utility of the spiculation measure was brought into the hybrid system through a one-step rule-based method.

3.2. Input–output mapping in two-dimensional feature space

The relationships between the outputs of the ANNs and the input features were analysed also in terms of a series of three-dimensional surface plots, which illustrated the relationships of the ANN output as a function of two input features, with each input feature ranging from zero to one in increments of 0.1. The input features were normalized relative to the minimum and maximum values of each individual feature for the 95 masses by assigning zero to the minimum and unity to the maximum. These plots represent hypothetical masses having the range of values of the given two input features. The features not shown in these plots were held constant.

Figures 4(a)–(c) are three-dimensional plots of outputs from the four-input ANN as functions of the spiculation measure and one of the other three features. The spiculation measure can be seen to dominate the decision-making process for the benign and malignant masses, as indicated by the fast-descending curvature of the ANN outputs along the spiculation measure as compared with the slow change along the direction of the other features in the surface plots. It should be noted that the critical value of the spiculation rule (normalized value of 0.43) used in the hybrid system lies in the range of the steepest slope of these surface plots.

Figures 5(a)–(c) show the relationships between outputs from the four-input ANN with two of the other three features (excluding the spiculation measure). Figures 5(d)–(f) show the relationships of outputs from the three-input ANN with two of its three input features. In order to compare the performance difference between the four-input ANN and the three-input ANN in analysing the non-spiculated masses, the spiculation measure in figures 5(a)–(c) was set at 0.35, which is below the threshold value (0.43) we used on the spiculation measure in the hybrid system. Thus, the ANN outputs in figures 5(a)–(c) represent masses in the non-spiculated category (spiculation measure < threshold). Features not shown in figure 5 were held constant in the two ANNs.



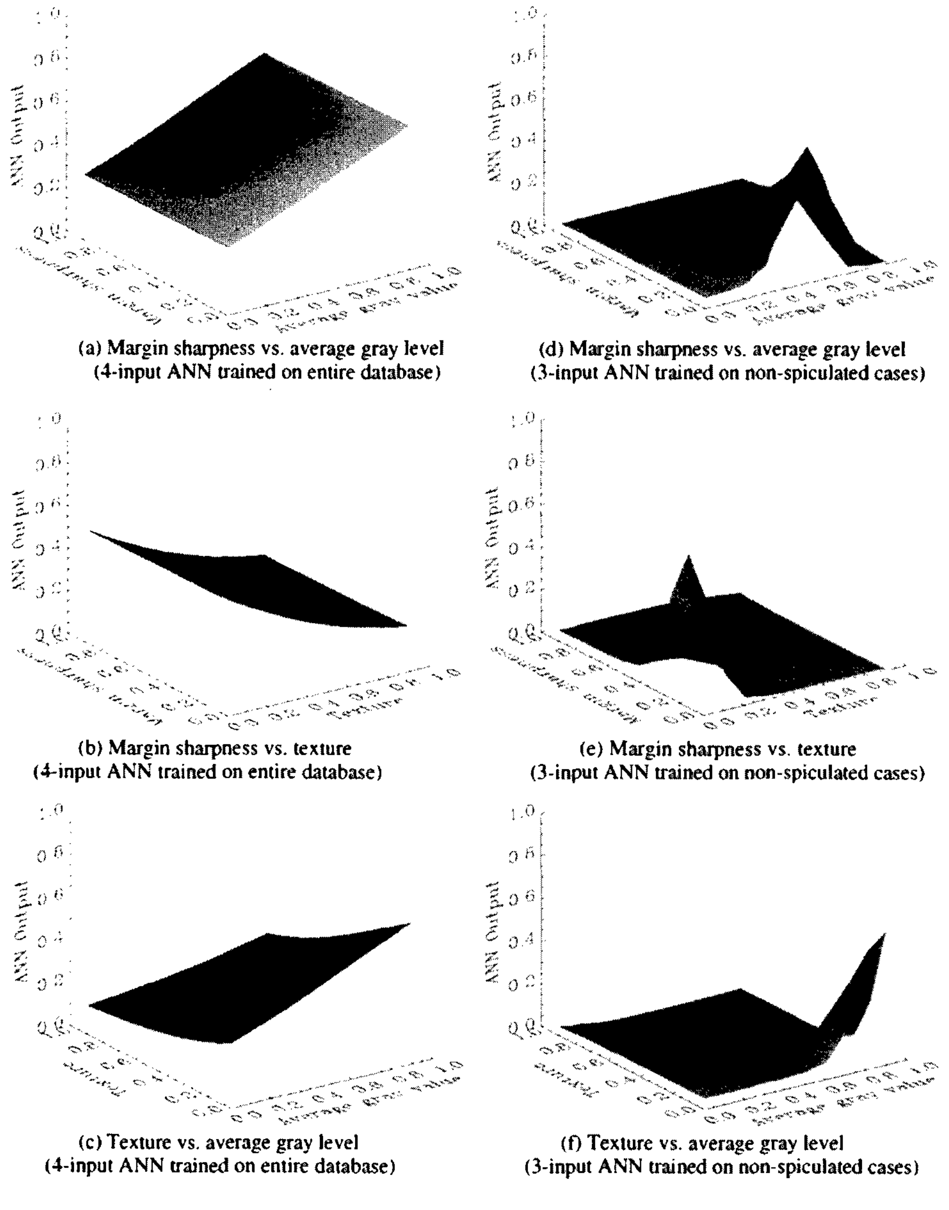
(a) Margin sharpness vs. spiculation
(4-input ANN trained on entire database)

(c) Texture vs. spiculation
(4-input ANN trained on entire database)

(b) Average gray level vs. spiculation
(4-input ANN trained on entire database)

Figure 4. Three-dimensional surface plots showing outputs from the four-input ANN (trained on the entire database) as functions of: (a) margin sharpness and spiculation; (b) average grey value and spiculation; and (c) texture and spiculation.

The difference between the three- and four-input ANNs in deciding the malignancy of masses in the non-spiculated category can be understood from figure 5. As shown in these plots, the general trend of the ANN outputs in terms of the directionality of the maximum ANN output values is similar for the four-input ANN (figures 5(a)–(c)) and three-input ANN (figures 5(d)–(f)). For example, masses with low margin sharpness and high average grey-value measures (figure 5(a) and 5(d)), masses with low margin sharpness and low texture measures (figures 5(b) and 5(e)), and masses with low texture and high average grey-value measures (figures 5(c) and 5(f)) yielded high output values from both classifiers. However, the change of the ANN output in the three-input ANN is more dramatic, whereas the change of the ANN output in the four-input ANN is more gradual. In clinical practice, as mentioned earlier, the likelihood of malignancy increases as margin sharpness decreases from well-defined to ill-defined, and the likelihood of malignancy increases with density; therefore, a dense mass with an ill-defined margin should be associated with an even higher likelihood of malignancy. As shown in figure 5(d) and (e), the sharp increase in output at the corners of low margin sharpness and high density for the three-input ANN emphasize the increasing likelihood of malignancy for these masses. Unfortunately, the four-input ANN (figures 5(a) and (b)) did not emphasize this as much as the three-input ANN trained with the non-spiculated cases only. Moreover, it seems that the significance of the margin characteristics (margin sharpness measure) was suppressed



(a) Margin sharpness vs. average gray level (4-input ANN trained on entire database)

(d) Margin sharpness vs. average gray level (3-input ANN trained on non-spiculated cases)

(b) Margin sharpness vs. texture (4-input ANN trained on entire database)

(e) Margin sharpness vs. texture (3-input ANN trained on non-spiculated cases)

(c) Texture vs. average gray level (4-input ANN trained on entire database)

(f) Texture vs. average gray level (3-input ANN trained on non-spiculated cases)

Figure 5. Outputs from the four-input ANN (trained on the entire database) as functions of: (a) margin sharpness and average grey value; (b) margin sharpness and texture; and (c) texture and average grey value. Outputs from the three-input ANN (trained only on the non-spiculated cases) are shown as functions of: (d) margin sharpness and average grey value; (e) margin sharpness and texture; and (f) texture and average grey value.

by that of the density characteristics (the average grey value and texture measures) in the learning of the four-input ANN, because the output from the four-input ANN varies less with the margin sharpness measure than with the average grey level and texture measures. As shown in figures 5(a) and (b), masses with similar densities but different margin sharpness values produce approximately the same outputs from the four-input ANN. The three-input ANN (trained on

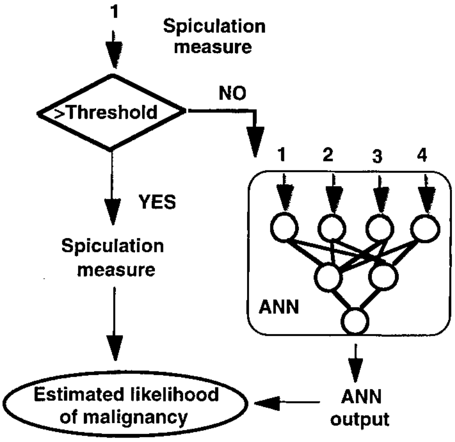
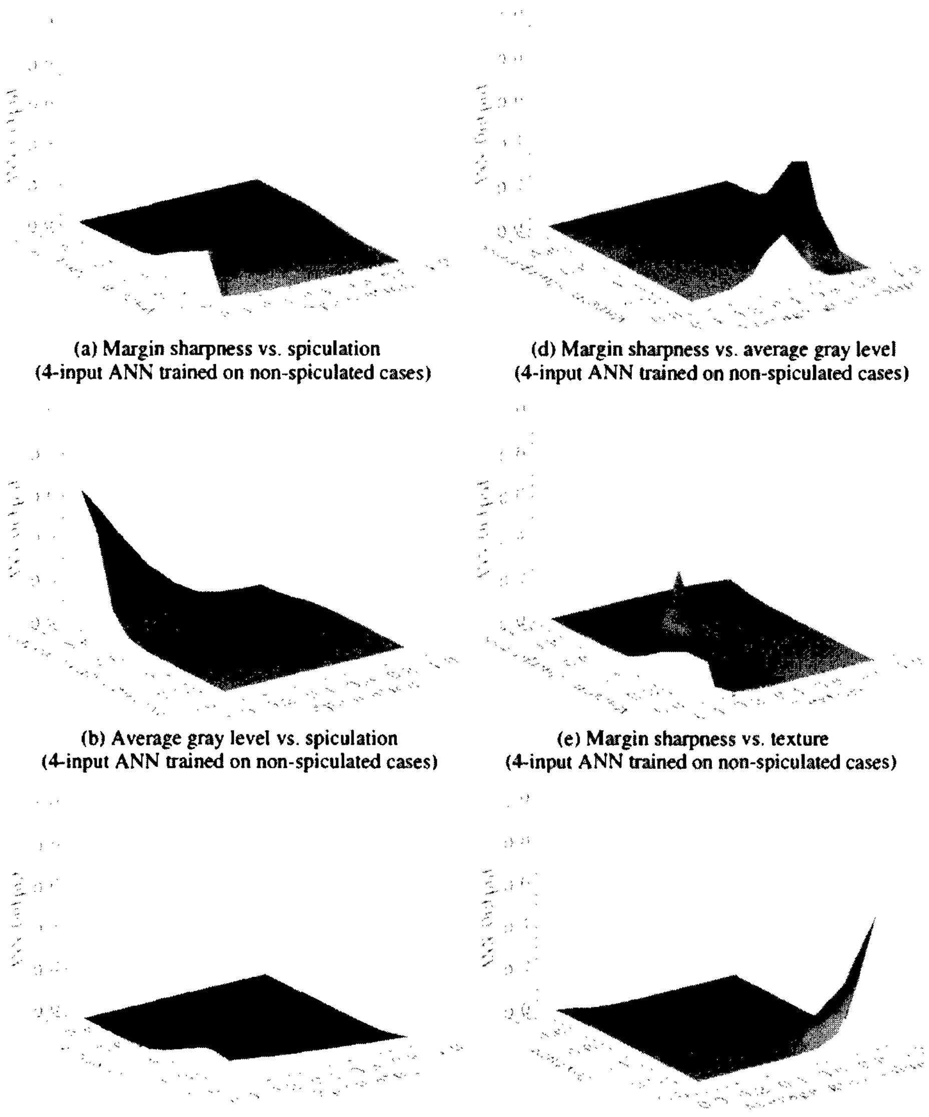


Figure 6. The structure of the alternative hybrid system, which consists of a rule-based method followed by a four-input ANN.

the non-spiculated masses) was able to learn better the significance of the margin sharpness and density measures in determining the likelihood of malignancy for a non-spiculated mass than did the four-input ANN (trained on the entire database). This difference in their learning resulted in a significant difference in their performance levels at high sensitivity.

The dip, instead of a peak, in figure 5(d) at very low margin sharpness and very high average grey value is probably due to over-learning from the small database, because such masses are not included in the clinical database we used. This does not necessarily mean that an ill-defined, high-density mass will have a lower estimate of the probability of malignancy. Since we also used the texture measure to quantify the density of a mass, an ill-defined, high-density mass can be correctly characterized on the basis of its margin sharpness and texture, as shown in figure 5(e). As mentioned earlier, the density of a mass cannot be evaluated exclusively on the basis of absolute grey value.

As shown in the graphs (figure 4), the spiculation measure dominated the learning and decision-making processes of the four-input ANN in determining the malignancy of a mass when spiculated masses are included. The learning of the other three features to classify non-spiculated masses is substantially limited in that situation. To show that it was the dominant nature ($A_z = 0.88$) of the spiculation measure and not the ANN's structure that introduced the difference in the learning of the two ANNs, we replaced the three-input ANN in the hybrid system with a four-input ANN, thereby including the spiculation measure as an additional input feature (figure 6). Note that the spiculation measure is not a prominent feature ($A_z = 0.53$) in classifying malignant and benign masses in the non-spiculated category (table 1). Three-dimensional surface plots were generated for the four-input ANN trained only on the non-spiculated masses. As shown in figures 7(a)–(f), the spiculation measure is not a dominant feature in the learning and decision-making processes of the four-input ANN when trained only on the non-spiculated masses as it was for the four-input ANN when trained on the entire database. This can be seen from the fact that relationships between the ANN outputs from the four-input ANN (trained on the non-spiculated masses) and the three input features as shown in figures 7(d)–(f) are similar to the relationships between the ANN outputs from the three-input ANN (trained on the non-spiculated masses) and the same three input features as shown in figures 5(d)–(f). Note that the features not shown in figure 7 were kept constant at the same values as those used to produce figures 4 and 5.



(a) Margin sharpness vs. spiculation
(4-input ANN trained on non-spiculated cases)

(b) Average gray level vs. spiculation
(4-input ANN trained on non-spiculated cases)

(c) Texture vs. spiculation
(4-input ANN trained on non-spiculated cases)

(d) Margin sharpness vs. average gray level
(4-input ANN trained on non-spiculated cases)

(e) Margin sharpness vs. texture
(4-input ANN trained on non-spiculated cases)

(f) Texture vs. average gray level
(4-input ANN trained on non-spiculated cases)

Figure 7. Three-dimensional surface plots showing outputs from the four-input ANN trained only on the non-spiculated cases as functions of: (a) margin sharpness and spiculation; (b) average grey value and spiculation; (c) texture and spiculation; (d) margin sharpness and average grey value; (e) margin sharpness and texture; and (f) texture and average grey value.

4. Discussion

We found that the two classifiers were actually similar in distinguishing spiculated (obvious) malignant masses from benign masses but differed significantly in distinguishing non-spiculated (subtle) malignant masses from benign masses. Our studies demonstrate that the ANN-alone method (the four-input ANN), when trained on all cases, learned to rely heavily upon the spiculation measure in classifying masses as malignant or benign (figure 3).

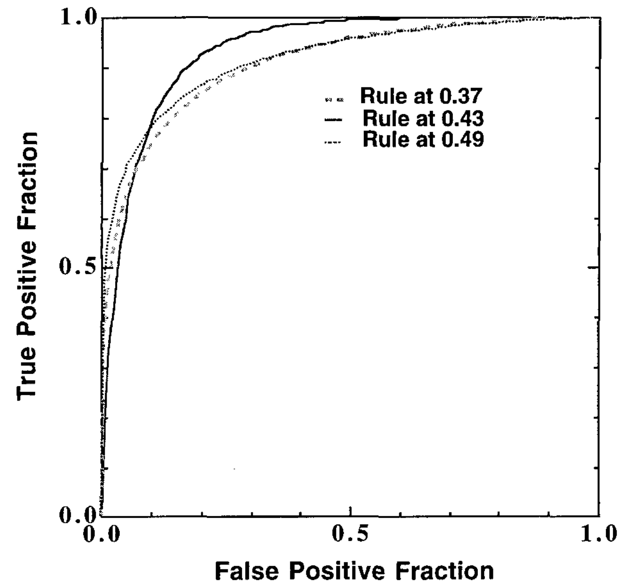


Figure 8. Comparison of ROC curves of the hybrid system when the critical value for the initial rule on the spiculation measure was varied by ±10°.

Thus, similar levels of performance were achieved by the ANN-alone method and by the hybrid classifier in determining the likelihood of malignancy for spiculated masses (corresponding to the lower portion of the ROC curves in figure 2). However, the ANN-alone method did not perform as well as the hybrid classifier in distinguishing between malignant masses and benign masses at the high sensitivity levels. The difference in their performance (figure 2) is due mainly to the difference in their ability to distinguish non-spiculated malignant masses from benign masses, as shown in figure 5. The dominant nature of the spiculation measure prevented the four-input ANN from learning the significance of the other three features in differentiating non-spiculated malignant masses from benign masses. Although slight over-learning occurred in the three-input ANN, probably because it was trained with a small number of non-spiculated cases (figure 5(d)), it seems that *only* when the ANN was employed after the spiculation criterion did the ANN learn effectively to interpret the complicated interrelationships among the remaining three features in determining the likelihood of malignancy of the non-spiculated cases. Therefore, it may be advantageous to employ a rule-based method when a single computer-extracted feature provides a strong separation between two classes, particularly when the computer-extracted feature (e.g. the spiculation measure here) correlates well with that used by humans in the decision-making task.

A hybrid system of the kind we employed can be optimized by varying the initial rule's critical value and then retraining the three-input ANN for each setting of the critical value. The result of such variation is demonstrated in figure 8, which shows results obtained when the critical value for the initial rule on the spiculation feature was varied by ±10°. Such variation resulted in a reduction in A_z and partial A_z. This result further indicates the need to accurately categorize the lesions as spiculated or non-spiculated prior to the introduction of the other three features and the training of the ANN.

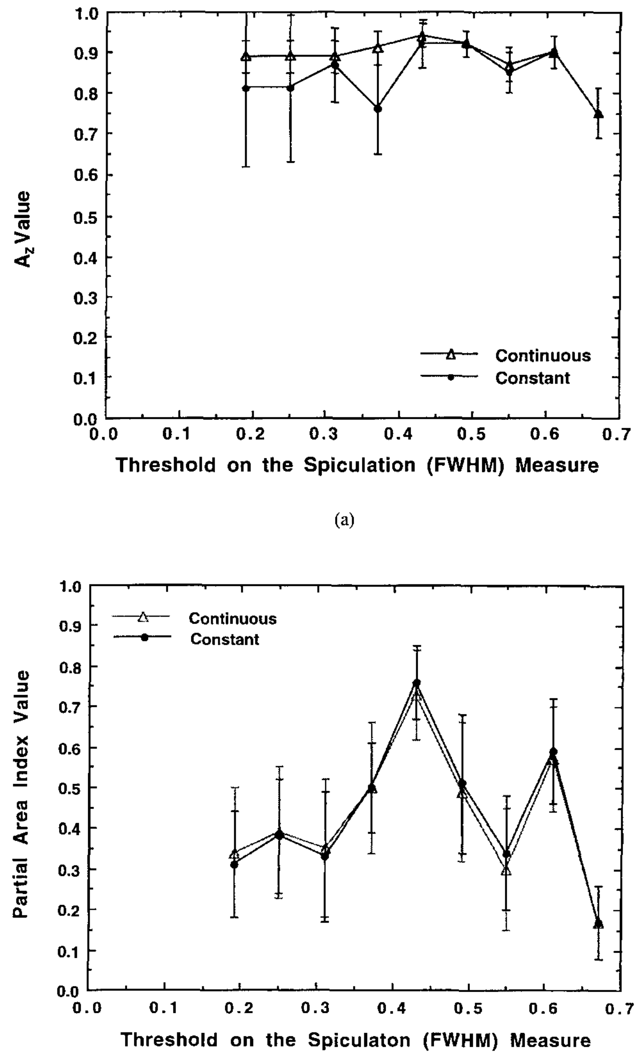
The rule in our hybrid system was used as a classifier in which lesions were categorized as either spiculated or non-spiculated. For the purpose of ROC analysis, values of the

Table 4. Variation in thc performance of thc hybrid system with changes in the critical value of thc spiculation rule as well as with changes in the conversion of the spiculation feature value after implementation of thc rule. The values after '±' are the standard deviations of the corresponding A_z values obtained from LABROC4 programs (Metz *et al* 1998).

	Use of continuous spiculation value or a constant value of one if lesions pass the spiculation rule			
	Continuous		Constant	
Hybrid system critical valuc (rule)	A_z	$_{0.90}A'_z$ (TPF > 0.90)	A_z	$_{0.90}A'_z$ (TPF > 0.90)
0.19	0.89 ± 0.04	0.34 ± 0.16	0.81 ± 0.19	0.31 ± 0.13
0.25	0.89 ± 0.04	0.39 ± 0.16	0.82 ± 0.18	0.38 ± 0.14
0.31	0.89 ± 0.04	0.35 ± 0.17	0.87 ± 0.09	0.33 ± 0.16
0.37	0.91 ± 0.04	0.50 ± 0.16	0.76 ± 0.11	0.50 ± 0.11
0.43	0.94 ± 0.03	0.73 ± 0.11	0.92 ± 0.06	0.76 ± 0.09
0.49	0.92 ± 0.03	0.49 ± 0.17	0.92 ± 0.03	0.51 ± 0.17
0.55	0.87 ± 0.04	0.30 ± 0.15	0.85 ± 0.05	0.34 ± 0.14
0.61	0.90 ± 0.04	0.57 ± 0.13	0.90 ± 0.04	0.59 ± 0.13
0.67	0.75 ± 0.06	0.17 ± 0.10	0.75 ± 0.06	0.17 ± 0.10

spiculation feature that are categorized as 'spiculated' can be either used directly as a decision variable (perhaps after monotonic transformation and/or renormalization), as in our analysis, or converted to a constant value that is strongly indicative of malignancy. However, the resulting ROC curve may depend upon the treatment of the feature after implementation of the rule. In fact, the lower-left part of the fitted ROC curve loses meaning if a constant value is assigned to those lesions that 'pass' the rule (i.e. are classified as 'spiculated'). Table 4 and figure 9 compare the calculated A_z and $_{\mathrm{TPF}}A'_z$ values for the alternative analyses in which (a) the lesions that passed the spiculation rule are assigned a continuous value, or (b) the lesions that passed the spiculation rule are assigned a constant value of 1.0 prior to ROC analysis. (Note that the output of the ANN varies from 0 to 1.) It should be noted that the A_z values can vary substantially depending on whether the feature is assigned a continuous or constant value, whereas the partial A_z values show relatively little variation (figure 9). This is expected, because the method of assignment for the feature after passing the rule only affects the lower part of the ROC curve.

Our rationale for integrating the rule-based method and the artificial neural network is to take advantage of the benefits of both approaches and to assign to each the tasks that best match their inherent abilities. Such an integration may provide us with maximum discriminant power and flexibility for the classification task, especially when the database and learning resources are limited. It is not that an artificial neural network is inherently unable to learn by itself a strategy to deal with this particular problem: artificial neural networks are good at learning ill-defined relationships from noisy examples and therefore can acquire their own knowledge for complex problems. However, there is great concern regarding proper training and performance evaluation of ANNs when training sample sizes are small (Tourassi and Floyd 1997). It is desirable to obtain good generalization even with few training data, because it is impossible to guarantee sufficient appropriate data for real-world problems. Other investigators have studied specialized networks with algorithms that reduce the network complexity by putting restrictions on synaptic weights to improve the ability to generalize from few training data (Mozer and Smolensky 1989, Nowlan and Hinton 1992, Fukushima 1993, Reed 1993). Over-learning of a particular important feature may occur in neural networks because a large number of neurons



(a)

(b)

Figure 9. Plots of (a) the A_z and (b) $_{0.90}A_z'$ values, along with the standard deviations, indicating the performance level of the hybrid system as functions of the cut-off value (for the initial rule on the spiculation measure) for the situation when the lesions that passed the cut-off value are assigned a continuous value and for when they are assigned a constant value of 1.0.

may be involved in the learning of the significance of the feature (Haykin 1994). Our hybrid system can be thought of as a way to prevent over-learning of the important feature—spiculation measure, in our situation—and thus to 'free' the ANN (or a majority of neurons thereof) to learn the significance of the other three features in the classification of non-spiculated masses. Further, it is apparent that more effort will be required to build a specialized network that can incorporate prior information concerning the spiculation than is required to develop a one-step rule-based method. In other words, one can take advantage of existing rules and different computer classifiers to intelligently tailor them into a hybrid system. The hybrid system can be used as an alternative way to optimize the learning process of the system for a particular problem, thus reaching the final learning goal. However, one must be cautious in determining which features are important. When the database is small, a feature that yields good separation between two classes in the database may result purely from fortuitous case sampling. In our study, the rule on the spiculation measure was not only determined on the basis of the separation seen in our database, but was also consistent with a feature used visually by radiologists.

Finally, it is important to note that with a sufficiently large database, the ANN-alone method (with a sufficient number of hidden units) would be expected to function as well as the hybrid system in the classification of mass lesions, i.e. it would be able to learn the significance of the dominant feature as well as the significance of the other features in classifying mass lesions when given enough samples. Nevertheless, it appears that use of a hybrid system is more efficient to bring well-known knowledge directly into a system in order to avoid lengthy training times, uncertainty concerning whether the final learning goal (well-known rules) is achieved, and the need for a large database. In another study, we did perform an independent validation of our computerized classification method on a 110-case database and showed that the ANN-alone and the hybrid classifiers were robust to case mix and digitization technique (Huo 1998).

Acknowledgments

The authors are grateful to Kunio Doi, PhD, Carl J Vyborny, MD, PhD, Robert M Nishikawa, PhD, Dulcy E Wolverton, MD and Robert A Schmidt, MD for their useful discussions. This work was supported in parts by the American Cancer Society (FRA-390), USPHS grants CA48985, US Army Medical Research and Development Command grant DAMD17-96-1-6058, and the National Digital Mammography Development Group NDMDG.

Maryellen L Giger is shareholder of R2 Technology, Inc., Los Altos, CA. It is the policy of the University of Chicago that investigators disclose publicly actual or potential significant financial interests that may appear to be affected by their research activities.

References

Cook H K and Fox M D 1989 Application of expert systems to mammographic image analysis *Am. J. Physiol. Image* **4** 16–22

D'Orsi C J and Kopans D B 1993 Mammographic feature analysis *Semin. Roentgenol.* **28** 204–30

Fukushima K 1993 Improved generalization ability using constrained neural network architecture *Proc. Int. Joint Conference on Neural Networks (IJCNN'93) (Nagoya, Japan)* pp 2049–54

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

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

Gong G 1986 Cross-validation, the jackknife, and the bootstrap: excess error estimation in forward logistic regression *J. Am. Statist. Assoc.* **81** 108–13

Haykin S 1994 *Neural Networks: A Comprehensive Foundation* (New York: Macmillan) pp 113–38

Huo Z 1998 Computerized methods for classification of masses and analysis of parenchymal patterns on digitized mammograms *PhD Thesis* University of Chicago pp 79–112

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

Huo Z, Giger M L, Vyborny C J, Wolverton D E, Schmidt R A and Doi K 1998 Automated computerized classification of malignant and benign masses on digitized mammograms *Acad. Radiol.* **5** 155–68

Jiang Y, Metz C E and Nishikawa R M 1996 A receiver operating characteristics partial area index for highly sensitive diagnostic tests *Radiology* **201** 745–50

Jiang Y, Nishikawa R M, Wolverton D E, Metz C E, Giger M L, Schimidt R A, Vyborny C J and Doi K 1996 Malignant and benign clustered microcalcifications: automated feature analysis and classification *Radiology* **198** 671–8

Katsuragawa S, Doi K, MacMahon H, Monnier-Cholley L, Ishida T and Kabayashi T 1997 Classification of normal and abnormal lungs with interstitial disease by rule-based method and artificial neural networks *J. Digit. Imaging* **10** 108–14

Knutzen A M and Grisvold J J 1993 Likelihood of malignant disease for various categories of mammographically detected, nonpalpable breast lesions *Mayo Clin. Proc.* **68** 454–60

Kopans D B 1989 *Breast Imaging* (Philadelphia: Lipincott) pp 351–3

Lo J Y, Baker J A, Kornguth P J, Iglehart J D and Floyd C E 1997 Predicting breast cancer invasion with artificial neural networks on the basis of mammographic features *Radiology* **203** 159–63

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

——1989 Some practical issues of experimental design and data analysis in radiological ROC studies *Invest. Radiol.* **24** 234–45

Metz C E, Herman B A and Shen J 1998 Maximum-likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data *Statist. Med.* **17** 1033–53

Metz C E, Wang P L and Kronman H B 1984 A new approach for testing the significance of differences between ROC curves measured from correlated data *Information Processing in Medical Imaging* ed F Deconick (Boston: Martinus Nijhoff) pp 432–45

Mozer M C and Smolensky P 1989 Skeletonation: a technique for trimming the fat from a network via relevance assessment *Advances in Neural Information Processing Systems* ed D Touretzky (San Mateo, CA: Morgan Kaufmann) pp 107–15

Nagel R H, Nishikawa R M, Papaioannou J, Giger M L and Doi K 1995 Comparison of rule-based and artificial neural network approaches for improving the automated detection of clustered microcalcifications in mammograms *Proc. SPIE* **2622** 775–9

Nowlan S J and Hinton G E 1992 Simplifying neural networks by soft weight sharing *Neural Comput.* **4** 473–93

Reed R 1993 Pruning algorithms—a survey *IEEE Trans.* **4** 740–7

Sickles E A 1991 Periodic mammographic follow-up of probably benign lesions: results in 3184 consecutive cases *Radiology* **179** 463–8

Swets J A, Getty D J, Pickett R M, D'Orsi C J, Seltzer S E and McNeil B J 1991 Enhancing and evaluating diagnostic accuracy *Med. Decision Making* **11** 9–18

Tabar L and Dean P B 1985 *Teaching Atlas of Mammography* (New York: Georg Thieme) pp 17–86

Tourassi G D and Floyd C E 1997 The effect of data sampling on the performance evaluation of artificial neural networks in medical diagnosis *Med. Decision Making* **17** 186–92

Wu Y, Giger M L, Doi K, Vyborny C J, Schmidt R A and Metz C E 1993 Artificial neural networks in mammography: application to decision making in the diagnosis of breast cancer *Radiology* **87** 81–7

Automated Computerized Classification of Malignant and Benign Masses on Digitized Mammograms[1]

Zhimin Huo, MSc, Maryellen L. Giger, PhD, Carl J. Vyborny, MD, PhD
Dulcy E. Wolverton, MD, Robert A. Schmidt, MD, Kunio Doi, PhD

Rationale and Objectives. To develop a method for differentiating malignant from benign masses in which a computer automatically extracts lesion features and merges them into an estimated likelihood of malignancy.

Materials and Methods. Ninety-five mammograms depicting masses in 65 patients were digitized. Various features related to the margin and density of each mass were extracted automatically from the neighborhoods of the computer-identified mass regions. Selected features were merged into an estimated likelihood of malignancy by using three different automated classifiers. The performance of the three classifiers in distinguishing between benign and malignant masses was evaluated by receiver operating characteristic analysis and compared with the performance of an experienced mammographer and that of five less experienced mammographers.

Results. Our computer classification scheme yielded an area under the receiver operating characteristic curve (A_z) value of 0.94, which was similar to that for an experienced mammographer (A_z = 0.91) and was statistically significantly higher than the average performance of the radiologists with less mammographic experience (A_z = 0.81) (P = .013). With the database used, the computer scheme achieved, at 100% sensitivity, a positive predictive value of 83%, which was 12% higher than that for the performance of the experienced mammographer and 21% higher than that for the average performance of the less experienced mammographers (P < .0001).

Conclusion. Automated computerized classification schemes may be useful in helping radiologists distinguish between benign and malignant masses and thus reducing the number of unnecessary biopsies.

Key Words. Breast, biopsy; breast neoplasms, diagnosis; computers, diagnostic aid; computers, neural network.

The present widespread use of mammography for early detection of breast cancer in asymptomatic women increases the importance of radiologists recognizing the mammographic features that distinguish carcinomas from benign abnormalities. Despite improvements in the criteria used to differentiate benign from malignant lesions of the breast (1–6), considerable misclassification of lesions occurs in everyday clinical practice. At many centers, only 15%–30% of mammographically detected lesions analyzed by means of surgical breast biopsy are actually malignant (7,8). There also is great variation (7%–40%) in positive biopsy rates among individual radiologists (9).

Computer-aided diagnosis in mammography can be defined as a diagnosis made by a radiologist who takes into account the output from a computer analysis of a mammogram. Many investigators have studied the use of computer analysis as an aid in the early detection of breast cancer (10–13). The development of computer aids that help in the classification portion of a mammographic work-up also has been studied. An objective computer classification scheme capable of differentiating between benign and malignant masses at a level similar to that of

Acad Radiol 1998; 5:155–168

[1] From Kurt Rossmann Laboratories for Radiologic Image Research, Department of Radiology, MC2026, University of Chicago, 5841 S Maryland Ave, Chicago, IL 60637. Received July 16, 1997; revision requested September 2; revision received September 25; accepted October 8. Supported in part by the American Cancer Society (FRA-390), U.S. Public Health Service grants CA48985, CA24806, U.S. Army grant DAMD 17-93-J-3021 and Medical Research and Development Command grant 92153010, and the National Digital Mammography Development Group (NDMDG). **Address reprint requests to** M.L.G.

Maryellen L. Giger, Carl J. Vyborny, Robert A. Schmidt, and Kunio Doi are shareholders of R2 Technology, Los Altos, Calif. It is the policy of the University of Chicago that investigators disclose publicly actual or potential significant financial interests that may appear to be affected by the research activities.

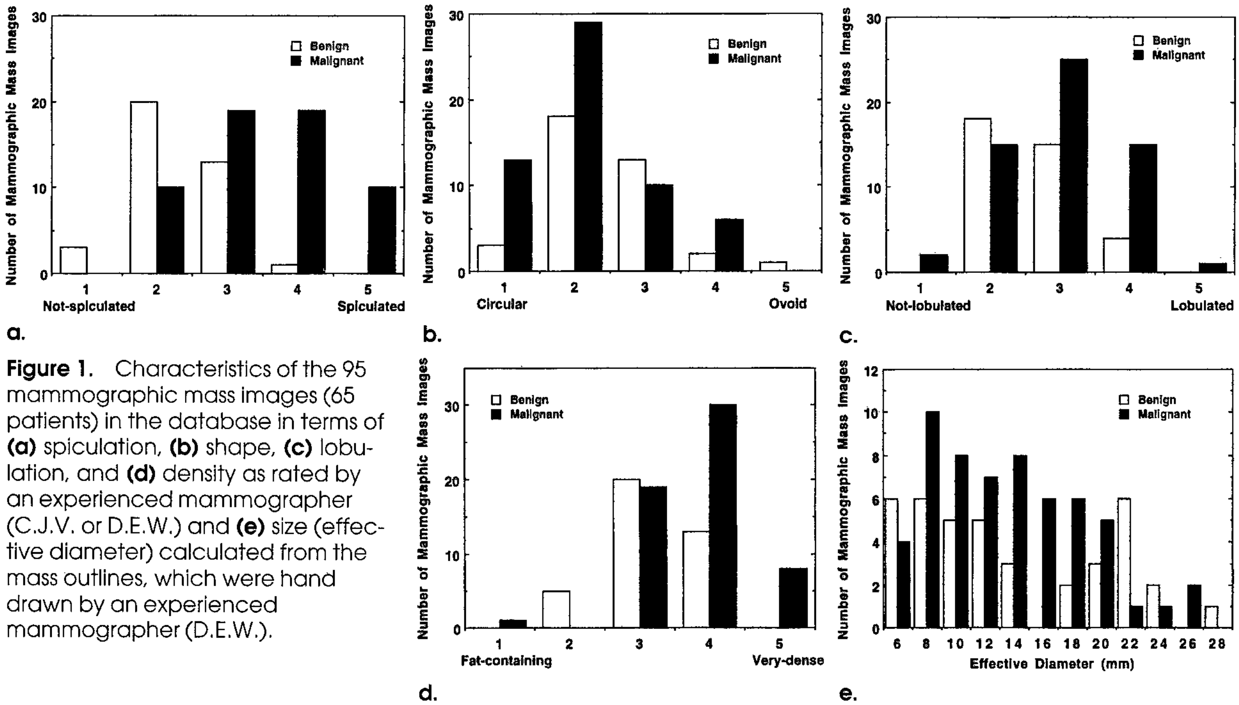


Figure 1. Characteristics of the 95 mammographic mass images (65 patients) in the database in terms of **(a)** spiculation, **(b)** shape, **(c)** lobulation, and **(d)** density as rated by an experienced mammographer (C.J.V. or D.E.W.) and **(e)** size (effective diameter) calculated from the mass outlines, which were hand drawn by an experienced mammographer (D.E.W.).

experienced mammographers would help radiologists improve accuracy, decrease variability, and reduce the number of unnecessary biopsies.

In the classification of lesions, investigators have taken advantage of the ability of radiologists to extract features related to the margin and density of mammographic abnormalities and have used computers to merge these (human-extracted) features into diagnoses (14–18). Use of computer-based decision systems such as rule-based methods, discriminant analysis, and artificial neural networks (ANNs) to merge the information extracted by either human observers or computers has been investigated (14–19). In addition, computerized techniques can be used to automatically extract individual image features such as spiculation (20–22), margin sharpness (23), irregularity (24), and density (25). Some investigators have attempted to use multiple computer-extracted features to classify masses (24,26).

In this study, we address the classification task in mammographic work-up and introduce a set of morphologic features similar to the ones used by practicing radiologists to characterize margin and density of a mass. We then merge these features with a spiculation measure into an estimated likelihood of malignancy for individual lesions. It should be noted that our fully automated computerized method includes automated lesion segmentation, automated feature extraction, and automated classification. The effectiveness of each individual feature and the role of each feature in classification of masses were studied.

To process the computer-extracted features more effectively, a two-step rule-based method and an ANN were used to merge these features. To overcome the limitations of these two individual types of classifiers for this particular task, integration of a rule-based method and an ANN was introduced as a hybrid information-processing approach. The hybrid system provides more power as a computer-based classifier by allowing emulation of humans in their information-processing and decision-making capabilities. The ability of the three classifiers to merge the computer-extracted features into a correct diagnosis was evaluated in 65 patients by using receiver operating characteristic (ROC) analysis (27,28). The performance of the computer was compared with that of an experienced mammographer and five radiologists with less mammographic experience.

MATERIALS AND METHODS

The database used in this study consisted of 95 clinical mammographic images (Min-R screen/OM-1 film; Eastman Kodak, Rochester, NY), each of which contained a mass. Thirty-eight of the images showed benign lesions, and 57 showed malignant lesions. The 95 mammograms were collected from examinations of 65 patients and represented an entire database gathered in our laboratory from December 1985 to October 1989. Twenty-six of the 65 patients had benign breast abnormalities, and 39 patients had breast cancer. Both mediolateral oblique and craniocaudal views were available for 30 of the patients (12 of 26 patients with benign lesions and 18 of 39 patients with malignant lesions). According to the original selection criteria, patients were chosen who had masses that were difficult to classify and who had undergone open biopsy or long-term mammographic follow-up. All but one patient underwent biopsy for the suspicion of breast cancer, and in the remaining one patient the disease was deemed benign on the basis of follow-up of more than 5 years. The screen-film mammograms were digitized with an optical drum scanner (FIP II; Fuji Film, Tokyo, Japan) at a sampling distance of 0.1 mm and 10-bit quantization.

To characterize the database, two experienced mammographers (C.J.V., D.E.W.) rated each mass with respect to spiculation, lobulation, shape, and density by using a five-point scale in which 1 corresponded to not spiculated, not lobulated, circular, or fat containing and 5 corresponded to definitely spiculated, lobulated, ovoid, or very dense, respectively. These distributions are shown in Figure 1a–1d. The size of each mass in terms of effective diameter was also estimated based on the region outlined on the computer by an experienced mammographer (D.E.W.). The effective diameter of a mass is defined as the diameter of the equivalent circle (whose area is the same as the area of the grown region) of the identified mass region (29). The distribution of size in terms of effective diameter for the masses depicted on the 95 images is shown in Figure 1e; the average size was approximately 1.3 cm.

Our current classification scheme consists of three stages: (*a*) automated segmentation of mammographic masses from surrounding parenchyma, (*b*) automated feature extraction, and (*c*) automated classification, which yields an estimation of malignancy of a mass by means of one of three classifiers—a rule-based method, an ANN, or a hybrid system (ie, a combination of a one-step rule-based method and an ANN).

The area under the ROC curve (A_z) was used to evaluate the ability of our computer classification scheme to utilize the three different classifiers to differentiate benign from malignant masses. Clinically, the specificities at high sensitivity levels are most relevant because the "cost" of missing a cancer is greater than the cost of performing a biopsy to assess a benign lesion. Thus, the average performances in a high sensitivity range (true-positive fraction [TPF_0] above 0.90) were evaluated for both our classification schemes and the observers by using a partial area index, $_{TPF}A_z'$, from 0 to 1, which is the portion of the A_z that lies above a preselected sensitivity threshold (TPF_0) in a conventional ROC graph divided by the constant ($1 - TPF_0$) (30). These performances, in terms of specificity at a given sensitivity level, were also evaluated. In this study, we chose to calculate specificity at a sensitivity level of 100% because the aim of creating the computer output in our research was to aid radiologists in reducing the number of unnecessary biopsies performed without misclassifying any cancers.

Segmentation

Segmentation of a mass from the background parenchyma was accomplished by using a multiple-transition-point, gray-level, region-growing technique (22). Segmentation begins within a 512 × 512-pixel region of interest manually centered about the abnormality in question, as illustrated in Figure 2a and 2b. In clinical practice, the location of the mass could be identified either by a radiologist or with a computer-detection scheme (31) and then fed into the classification scheme for an output in regard to the likelihood of malignancy. To correct for the nonuniformity of the background distribution and to enhance image contrast for better segmentation of masses, background trend–correction and histogram-equalization techniques were applied to the 512 × 512-pixel region of interest (22). The corresponding enhanced images of the malignant and benign masses are shown in Figure 2c and 2d, respectively. The computer-identified margins of the malignant and benign masses are superimposed on the images of the original masses in Figure 2e and 2f. For comparison, margins of the same images hand drawn by an experienced mammographer are shown in Figure 2g and 2h.

Computer-extracted Radiographic Features: Margin and Density

The margin, shape, and density of a mass are three major characteristics used by radiologists in classifying masses. Different characteristics of these features are associated with different levels of probability of malignancy (4,6,32). To determine the likelihood of malignancy associated with different margin and density characteristics, we developed algorithms that extract two features that characterize the margin of a mass (spiculation, sharpness) and three features that characterize the density of a mass (average gray level, contrast, texture).

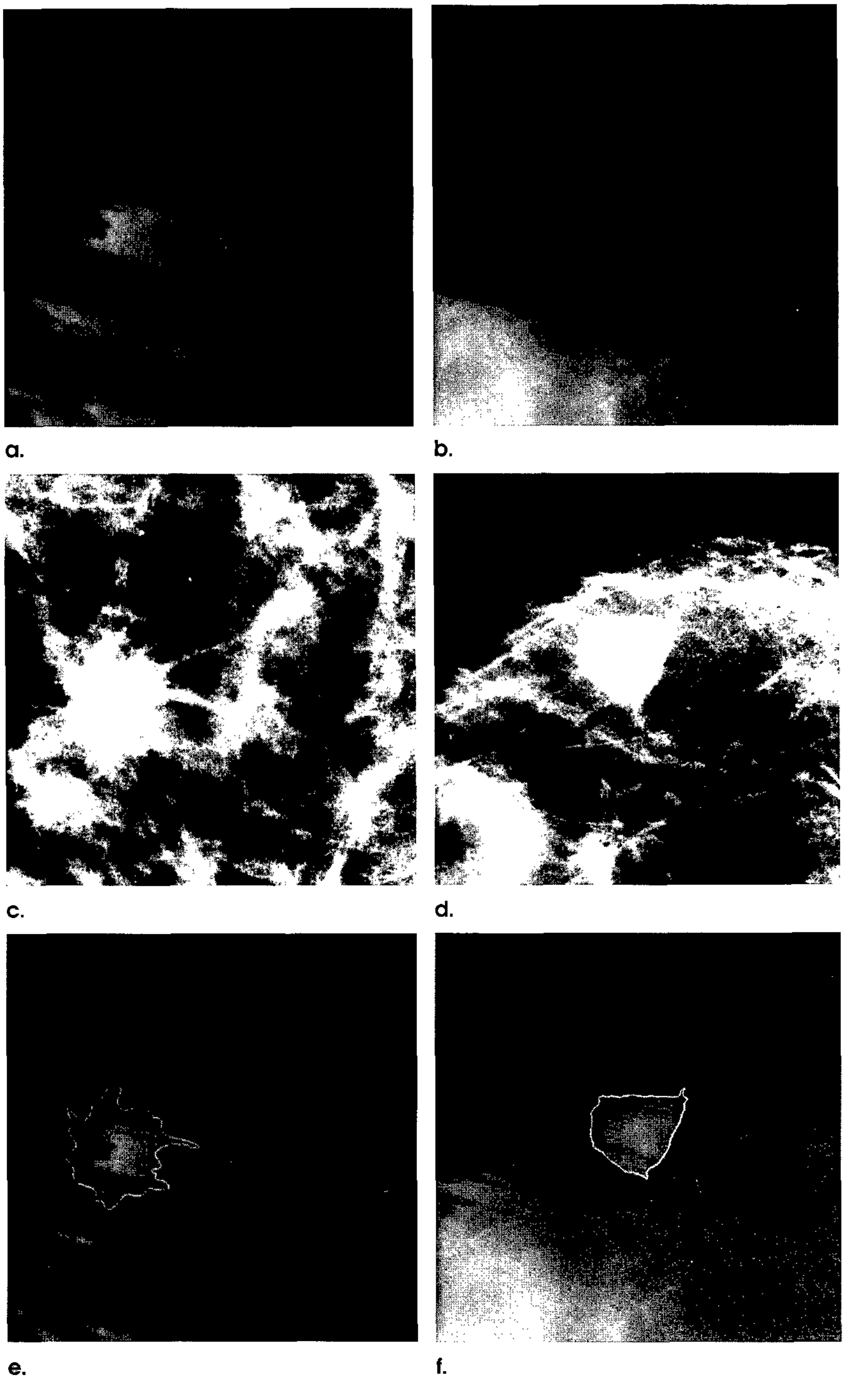


Figure 2. Mammographic images of **(a)** a malignant mass and **(b)** a benign mass in a 512 × 512-pixel region of interest; enhanced images of the **(c)** malignant and **(d)** benign masses after image processing; and computer-extracted margins superimposed on the **(e)** malignant and **(f)** benign masses (*Fig 2 continues*).

We did not explicitly devise a specific measure to characterize the shape of a mass for the purpose of classification, but measures related to shape are embedded within the other measures.

Margin.—Margin characteristics are very important in differentiating between benign and malignant masses. To determine the likelihood of malignancy of a mass based on its margin, two major margin characteristics—spiculation and sharpness—were measured. Margin spiculation is the most important indicator for malignancy, with spiculated lesions having a greater than 90% probability of malignancy (6). Margin sharpness is also very important in determining whether a mass is benign or malignant; an ill-defined margin indicates possible malignancy, and a well-defined margin indicates likely benignity. Only about 2% of well-defined masses are malignant (2).

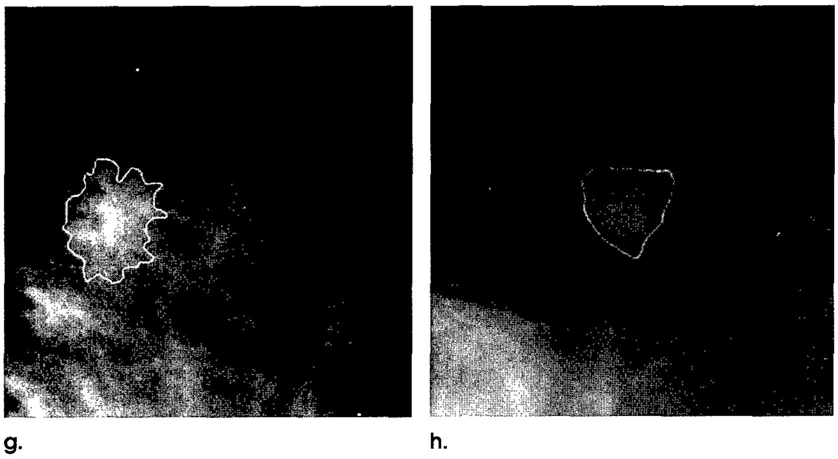


Figure 2 (*continued*). An experienced mammographer's hand-drawn margins of the **(g)** malignant and **(h)** benign masses.

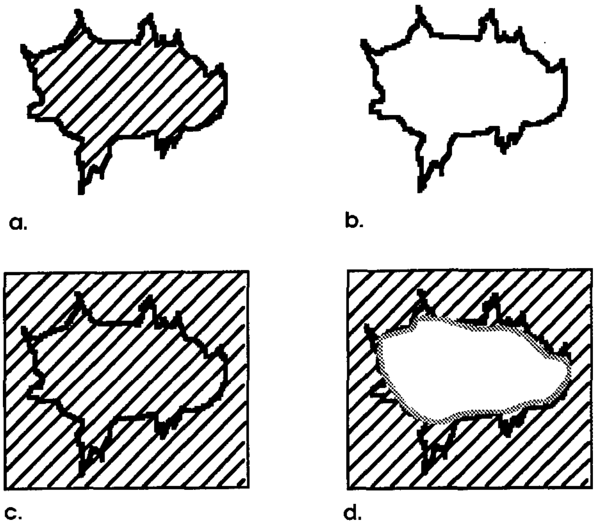


Figure 3. Illustration of the four neighborhoods used for feature extraction: **(a)** grown region, **(b)** margin, **(c)** encompassing region, and **(d)** surrounding periphery (cross-hatched region).

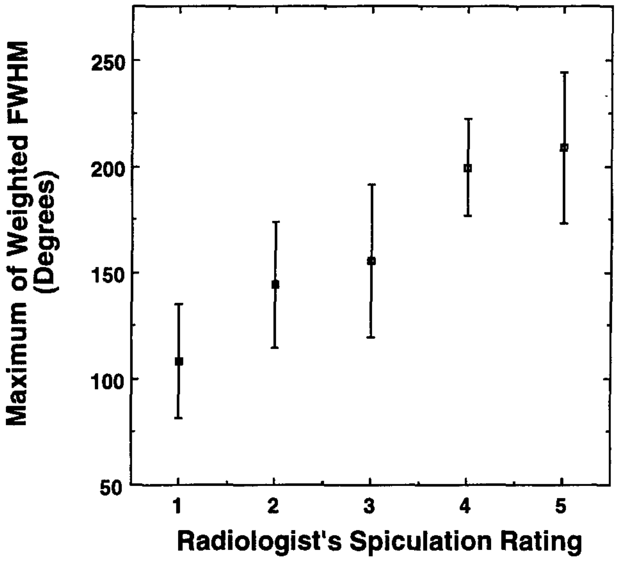


Figure 4. Correlation of the spiculation measure (weighted FWHM, in degrees) with the spiculation ratings (Fig 1a) of an experienced mammographer for a database of 95 mass images. The error bars indicate the variation in the spiculation measure for each spiculation rating given by the radiologist.

The spiculation measure is determined from an analysis of radial edge gradients (22). The spiculation measure evaluates the average angle (in degrees) by which the direction of the maximum gradient at each point along the margin of a mass deviates from the radial direction, the direction pointing from the geometric center of the mass to the point on the margin. The actual measure is the full width at half maximum (FWHM) of the normalized edge-gradient distribution calculated for a neighborhood of the grown region of the mass with respect to the radial direction (22). This measure is able to quantify the degree of spiculation of a mass primarily because the direction of maximum gradient along the margin of a spiculated mass varies greatly from its radial direction, whereas the direction of the maximum gradient along the margin of a smooth mass is similar to its radial direction. The spiculation measure was extracted not only along and within the margin of a mass (Fig 3a, 3b) but also in enlarged neighborhoods of the computer-identified mass region as shown in Figure 3c and 3d. In this way, potentially more subtle spicules that are difficult to delineate by region growing could be better extracted. The two enlarged neighborhoods included 20 additional pixels around the computer-identified mass region. A neighborhood of this size is large enough to accommodate thin or short spicules radiating from the margin of a mass.

To maximize the sensitivity of the spiculation measure, all the possible signs of spiculation identified from the four neighborhoods were considered; the greatest value of the FWHM measures from the four neighborhoods was used to indicate the spiculation of a mass. However, because of differences in the ability of FWHM measures from the four neighborhoods to capture spiculation information (22), the FWHM measures were weighted differently. The weighting factor used for the two enlarged neighborhoods was 1.0, and that used for the other two neighborhoods was 0.85. This weighted spiculation measure correlates well with an experienced mammographer's spiculation rating ($r = .64$; $P < .0001$) (Fig 4). In addition, the level of performance of the spiculation measure ($A_z = 0.88$) was similar to that of the experienced mammographer's spiculation ratings ($A_z = 0.85$) in terms of the ability to distinguish between benign and malignant masses based solely on spiculation (22).

The sharpness of the margin of a mass can be described as well defined, partially ill defined, or ill defined. The average margin sharpness can be quantified by calculating the magnitude of the average gradient along the

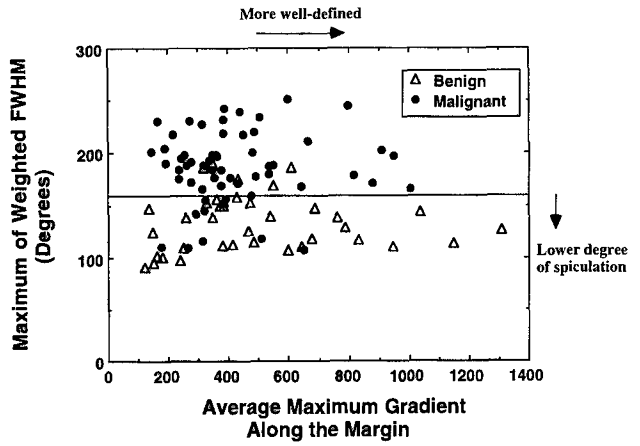


Figure 5. Cluster plot of the spiculation measure (weighted FWHM) versus the margin sharpness measure (average gradient) along the margin for 95 mass images. The horizontally drawn line indicates the cutoff on the FWHM measure chosen to distinguish between spiculated and nonspiculated masses.

Figure 6. Examples of masses with **(a)** a spiculated margin, **(b)** an ill-defined margin, **(c)** a partially ill-defined margin, and **(d)** a well-defined margin shown by mammography.

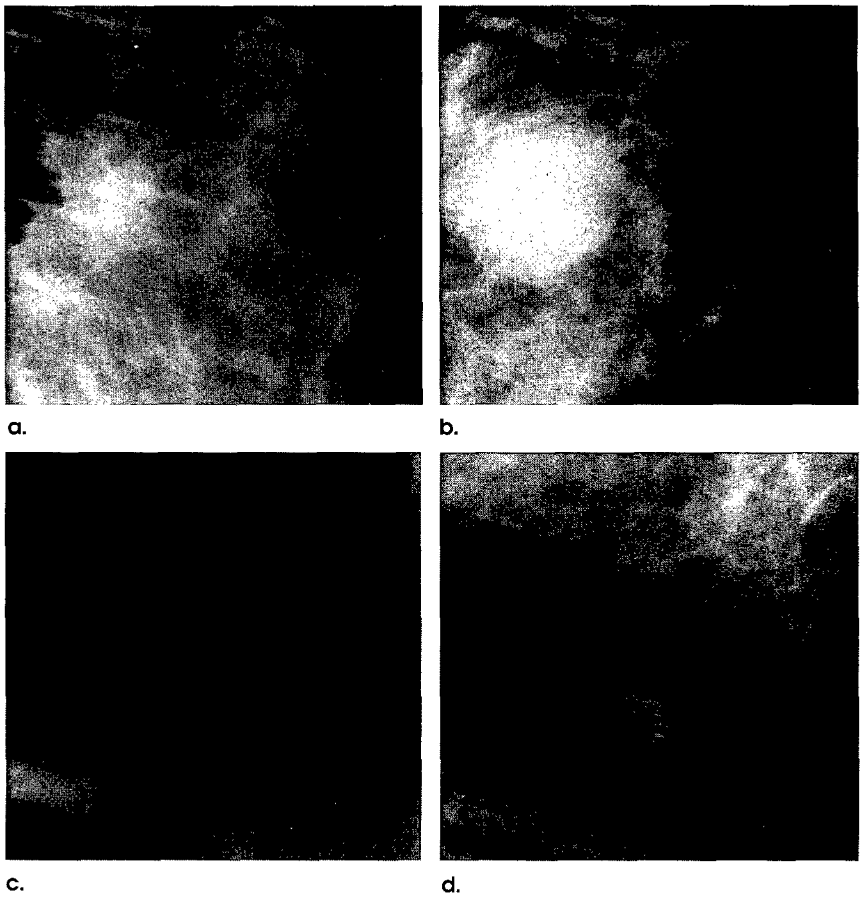


Table 1
Examples of Spiculation and Margin-Sharpness Measures for Four Selected Masses

Mass Type	Radiologist's Spiculation Rating	FWHM Measure	Average Gradient Along the Margin
Spiculated	5	240	513
Ill defined/obscured	3	118	318
Partially ill defined/obscured	3	111	962
Well defined	2	111	1,315
Maximum value in the database	5	242	1,315
Minimum value in the database	1	88	127

margin of the mass. A well-defined margin has a large value for the average margin sharpness measure, whereas an ill-defined margin has a small value.

Figure 5 shows the relationship between the two margin measures for the 95 mass images. The horizontally drawn line indicates a cutoff on the FWHM measure used to categorize spiculated masses and nonspiculated masses. It should be noted that there is much more overlap between benign and malignant masses in terms of margin sharpness than in terms of margin spiculation. With the threshold of 160° for the spiculation measure, most of the malignant masses were in the spiculated category (FWHM > 160°). At this threshold, five of 39 malignant masses and 22 of 26 benign masses were classified as nonspiculated. In addition, in the nonspiculated category, masses with a higher value for the margin-sharpness measure tended to be benign. This finding is in agreement with radiologists' visual perception in determining the benign versus malignant nature of masses. Thus, to determine the likelihood of malignancy of a mass based on the two margin characteristics described, it is more effective to use first the spiculation measure to identify spiculated masses (which are very likely to be malignant) and to determine their likelihood of malignancy based on their degree of spiculation. The margin-sharpness measure can then be used further to determine the likelihood of malignancy of the remaining (ie, nonspiculated) masses.

Figure 6 shows examples of masses with spiculated, ill-defined, partially ill-defined, and well-defined margins. The calculated spiculation and margin-sharpness measures for these four masses are listed in Table 1. The spiculated mass (radiologist's spiculation rating of 5) had a FWHM measure of 240° in a database with a maximum degree of spiculation of 242° and a minimum of 88°. This mass was correctly identified as highly spiculated and thus was not further analyzed with the margin-sharpness measure. The three smoother masses, each with spiculation ratings by the radiologist of 2 or 3, had similar spiculation measures that ranged from 111° to 118°. They were classified as nonspiculated masses (FWHM < 160°) and were further evaluated with the margin-sharpness measure. The margin-sharpness measures of the three masses were well separated, with the well-defined margin having the highest value (sharpness of 1,315), the partially ill-defined margin having the second highest value (sharpness of 962), and the ill-defined margin having the lowest value (sharpness of 318) in a database with margin-sharpness measures that ranged from 127 to 1,315. This illustrates the usefulness of the margin-sharpness measure in further discriminating between masses in the nonspiculated category.

Density.—Although the radiographic density of a mass may not by itself be as powerful a predictor as the margin features in distinguishing between benign and malignant masses, taken with these features density assessment can be extremely useful (4). The evaluation of the density of a mass is of particular importance in diagnosing circumscribed, lobulated, indistinct, or obscured masses (4) that are not spiculated.

To assess the density of a mass radiographically, we introduced three density-related measures (average gray level, contrast, texture) that characterize different aspects of the density of a mass. These measures are similar to those used intuitively by radiologists. Average gray level is obtained by averaging the gray-level values of each point within the grown region of a mass. Contrast is the difference between the average gray level of the grown mass and the average gray level of the surrounding fatty areas (areas with gray-level values in the lower 20% of the histogram for the total surrounding area). Texture is defined here as the standard deviation of the average gradient within a mass, and it is used to quantify patterns that arise from veins, trabeculae, and other structures that may be visible through a low-density mass but not through a high-density mass. A mass of low radiographic

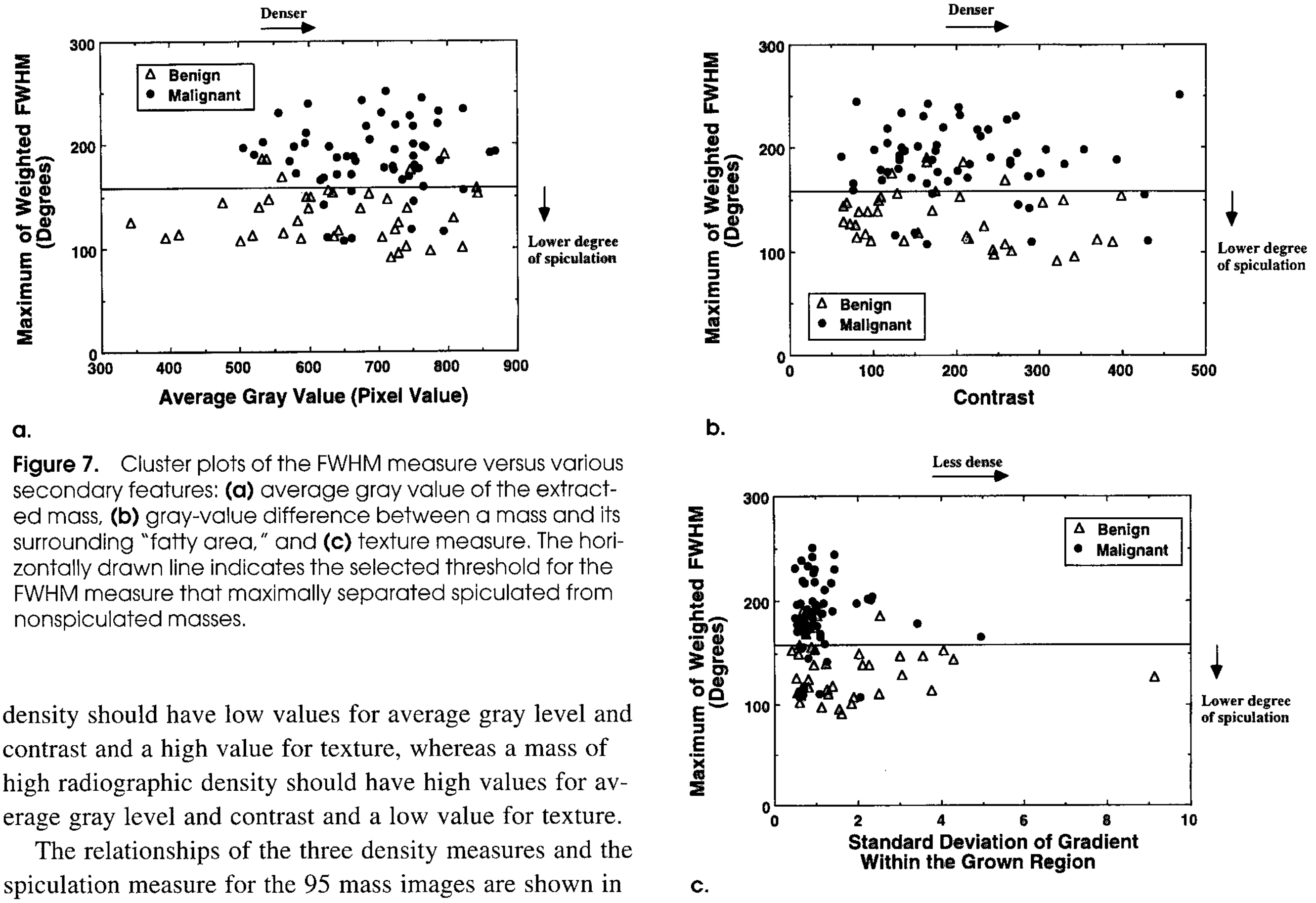


Figure 7. Cluster plots of the FWHM measure versus various secondary features: **(a)** average gray value of the extracted mass, **(b)** gray-value difference between a mass and its surrounding "fatty area," and **(c)** texture measure. The horizontally drawn line indicates the selected threshold for the FWHM measure that maximally separated spiculated from nonspiculated masses.

density should have low values for average gray level and contrast and a high value for texture, whereas a mass of high radiographic density should have high values for average gray level and contrast and a low value for texture.

The relationships of the three density measures and the spiculation measure for the 95 mass images are shown in Figure 7. The drawn line indicates a cutoff, based on the spiculation measure (FWHM of 160°), that categorizes spiculated and nonspiculated masses. As can be seen from these cluster plots, the distribution of benign and malignant nonspiculated masses in terms of the density agrees with radiologists' general perception; namely, the benign masses in the nonspiculated category tend to have low image density, whereas the malignant masses in the nonspiculated category tend to have high image density for all three density measures. The ability to separate low-density benign masses from high-density malignant masses only in the nonspiculated category stresses the importance of using the density measures to differentiate between benign and malignant masses only after having excluded the spiculated masses.

Classification of Masses

The ability of each individual computer-extracted feature to aid in the differentiation between benign and malignant mass images was evaluated for the entire database with ROC analysis. The calculated A_z values are listed in Table 2. The ROC analysis shows that the spiculation measure outperformed the other features in distinguishing between benign and malignant masses; the Az value was 0.88 for spiculation compared with 0.54–0.65 for the other four features. We have therefore found that margin spiculation is as important a feature for the computerized method as it is for radiologists.

After the rule based on the spiculation measure (FWHM of 160°) was applied, the ability of these features to further distinguish between benign and malignant mass images in the remaining database (nonspiculated) was also studied with ROC analysis. The calculated A_z values for these features are listed in Table 2. As can be seen in Table 2, the spiculation measure is no longer a dominant feature in discriminating between benign and malignant masses in the nonspiculated category. The other four features perform better, however, in differentiating malignant from benign masses in the nonspiculated category than in the complete database (ie, both spiculated and nonspiculated). This finding indicates the importance of using these features to differentiate between benign and malignant masses only after the spiculated masses have been excluded.

Table 2
Performances of the Five Computer-extracted Features in Distinguishing between Benign and Malignant Mass Images

Feature	A_z for All 95 Mass Images	A_z for 36 Mass Images in the Nonspiculated Category
Margin		
Spiculation (FWHM)*	0.88	0.53
Sharpness*	0.56	0.68
Density		
Average gray level*	0.65	0.66
Contrast	0.59	0.70
Texture measure*	0.54	0.71

*These features were used as inputs in the ANN and the combined rule-based ANN classifiers.

Three automated classifiers were investigated for the task of merging the computer-extracted features into an estimate of the likelihood of malignancy: a rule-based method, an ANN, and a hybrid system. In determining the likelihood of malignancy for the cases that had both the mediolateral oblique and the craniocaudal views, the measurements obtained from both views were considered, and the view that the computer estimated had a higher likelihood of malignancy was used in the evaluation. For example, a mass would be classified as malignant if either one of the two views showed suspect signs (ie, either one of the FWHM measures from its two views satisfied the cutoff on the FWHM measure).

Rule-based method.—A rule-based method adapts knowledge from experts into a set of simple rules. Certain criteria for differentiating between benign and malignant masses have been established by expert mammographers (4,6,32). The rules used in our approach for measures of spiculation, margin sharpness, and density were based on these criteria.

A two-step rule-based method was studied for this database. Because of its clinical diagnostic importance, the spiculation measure was applied first in our rule-based method. After the spiculation measure (FWHM) was applied to identify spiculated masses (including some irregular masses) and categorize them as malignant first, a second feature was applied to characterize further the masses in the nonspiculated category as discussed in the previous section. To investigate the potential discriminant ability of the spiculation measure along with all the possible secondary features, we applied separately each of the remaining four features—the margin-sharpness measure and the three density measures—after the spiculation measure. The threshold of the spiculation measure (FWHM = 160°) was determined based on the entire database. The thresholds of the other four features were determined based on the remaining database only.

ANN.—The ANN approach is very different from the rule-based method. Instead of using predetermined empirical algorithms based on prior knowledge, ANNs are able to learn from examples and therefore can acquire their own knowledge through learning. Also, neural networks are capable of processing large amounts of information simultaneously. Neural networks do not, however, provide the user with explanations for their decisions and may not be able to bring preexisting knowledge into the network.

Here we used a conventional three-layer, feed-forward neural network with a back-propagation algorithm, which has been used in medical imaging and medical decision making (33,34). The structure of the neural network included four input units (each of which corresponded to a computer-extracted feature), two hidden units, and one output unit. The four features used as inputs to the ANN were the FWHM measure, the margin-sharpness measure, and two density measures (indicated by asterisks in Table 2). Similar performances were obtained when all three density measures were used. Because limiting the number of input features is critical in reducing the number of training samples needed, we kept the number of inputs to the ANN to a minimum; thus, only two density measures were used.

To determine the ability of our neural network to generalize from the training cases and make diagnoses for cases that had not been included in the database, we used a round-robin method, also known as the leave-one-out method. In this method, all but one of the cases were used to train the neural network. The single case that was left out was used to test the neural network. For the cases that had both mediolateral oblique and craniocaudal views, both images were left out in the round-robin training. The higher value of the two from the round-robin test was reported as the estimated likelihood of malignancy. This procedure was repeated for all the cases.

Hybrid system.—Each classifier has its advantages and limitations. With rule-based methods, one could adopt preexisting knowledge as rules. There are limitations, however, in the availability of knowledge and knowledge translation. Even the experts find it difficult to articulate particular types of "intuitive" knowledge, and the process of translating particular knowledge into rules is limited by this expressive power. ANNs are capable of learning from examples and therefore can acquire their own knowledge. It may be most advantageous to use ANNs when intuitive knowledge cannot be explicitly expressed or is difficult to translate. The ANN needs a sufficiently large database, however, to learn effectively. Also, with an ANN there may be uncertainty as to whether the final learning goal is achieved in some situations.

To take advantage of both rule-based systems and ANNs in the task of classifying masses, we integrated a rule-based method and an ANN into a hybrid system. In the hybrid system, we initially applied a rule on the spiculation measure because both spiculated and irregular masses are highly suspect for malignancy. We then applied an ANN to the remaining masses. Basically, this method frees the ANN from having to "learn" the importance of spiculation to the detriment of learning the importance of the other features.

The threshold of the spiculation measure for the hybrid system was the same as the one used in the rule-based method. The ANN applied in the hybrid system was a three-layer, feed-forward neural network with a back-propagation algorithm that had a structure of three input units (corresponding to the three remaining features used in the ANN method), two hidden units, and one output unit. The same round-robin method was applied to test the generalization ability of this neural network to differentiate between benign and malignant masses in the nonspiculated category.

Observer Study

One experienced radiologist who specializes in mammography (D.E.W.) and five other radiologists with some experience in mammography participated in the observer study. The experienced radiologist characterized some features of the database after the observer study was completed. Three of the five other radiologists were general radiologists from Europe participating in visiting fellowships. At the time of the observer study, they had a total of 6 months to 2 years of experience in mammography. One of the remaining two was a fellow in mammography in his 3rd month of training. The fifth radiologist had 3 months of training in mammography (beyond the standard 2 months of training in residency) in a combination fellowship. In the study, each observer was asked to estimate the probability of malignancy for each of the 65 cases by using a 100-point scale based on the mammograms available. The performance of each observer in distinguishing between benign and malignant masses was evaluated by using ROC analysis. An ROC curve was generated for the five less experienced radiologists as a group by averaging the two binormal parameters of their individual ROC curves (27,28).

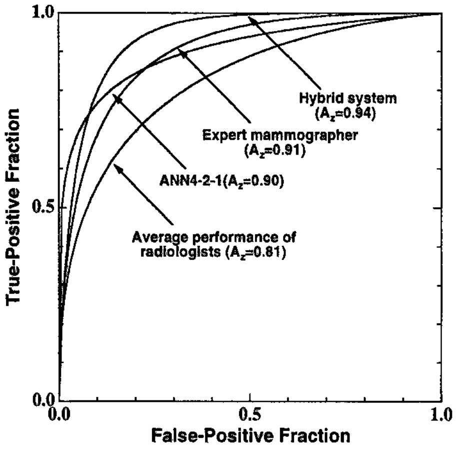


Figure 8. ROC curves for the performance of an experienced mammographer, five radiologists, the computerized scheme with ANN alone, and the computerized scheme with the hybrid system. *ANN4-2-1* = ANN with four input units, two hidden units, and one output unit.

RESULTS

Because the computer outputs from each individual classifier were monotonically correlated with the estimated probabilities of malignancy, we were able to evaluate the ability of each classifier to merge computer-extracted features into a correct estimated probability of malignancy based on the computer output with ROC analysis. The ROC curves of the ANN and hybrid system are shown in Figure 8. The A_z and the partial area index $_{0.90}A_z'$ values of the three classifiers are listed in Table 3. Among the three classifiers, the hybrid system yielded the highest A_z and $_{0.90}A_z'$ values. The specificities and positive predictive values of the three classifiers at 100% sensitivity were calculated. As shown in Table 3, the hybrid system yielded the highest specificity (69.2%), the two-step rule-based method the second highest (42.3%, 34.6%, 30.8%, 30.8%), and the ANN the third highest (19.2%).

The performance of the hybrid system was compared with that of the other two types of classifiers. No statistically significant difference ($P > .05$) was found for the A_z values based on the evaluation from the CLABROC program (35,36). Statistically significant differences for the

Table 3
Performances of the Three Classifiers in Distinguishing between the 26 Benign and the 39 Malignant Masses

Classifier	Specificity*	Positive Predictive Value*	Full Curve A_z	Partial Area Index	
				$_{0.90}A_z'$	P†
Two-step rule-based system (1st rule on FWHM)					
Margin sharpness	34.6%	69.6%	0.92	0.59	0.014
Average gray level	30.8%	68.4%	0.90	0.45	0.001
Contrast	30.8%	68.4%	0.92	0.58	0.021
Texture measure	42.3%	72.2%	0.92	0.63	0.015
ANN (4-2-1)	19.2%	65.0%	0.90	0.40	0.008
Hybrid system (rule-based and ANN 3-2-1)	69.2%	83.0%	0.94	0.73	. . .

Note.—ANN 4-2-1 = ANN with four input units, two hidden units, and one output unit; ANN 3-2-1 = ANN with three input units, two hidden units, and one output unit.

*Sensitivity was 100%.

†The P values were calculated for the difference in the $_{0.90}A_z$ between the hybrid system and the other two classifiers.

Table 4
Performances of the Human Observers in Distinguishing between the 26 Benign and the 39 Malignant Masses

Observer	Specificity*	Positive Predictive Value*	Full Curve A_z	Partial Area Index $_{0.90}A_z'$
A	3.8%	60.9%	0.85	0.29
B	11.5%	62.9%	0.86	0.37
C	11.5%	62.9%	0.85	0.40
D	0%	60.0%	0.70	0.07
E	3.8%	60.9%	0.80	0.27
Average performance of A–E	6.1%	61.5%	0.81	0.28
Experienced mammographer	38.5%	70.9%	0.91	0.58

*Sensitivity was 100%

$_{0.90}A_z'$ values were found, however, at the levels of the two-tailed P values as listed in Table 3. Differences in positive predictive value and specificity at 100% sensitivity between the hybrid system and the two-step rule-based method on average were 13% and 34%, respectively. Differences in positive predictive value and specificity at 100% sensitivity between the hybrid system and the ANN were 18% and 50%, respectively.

The ability of each radiologist to distinguish between benign and malignant masses was determined based on the radiologist's subjective ratings of the probability of malignancy for the 65 cases. The ROC curves of the experienced mammographer and of the average performance of the five radiologists are shown in Figure 8. Table 4 lists their individual performances in terms of A_z, $_{0.90}A_z'$, positive predictive value, and specificity at 100% sensitivity. The average performance for the five radiologists was calculated (Table 4). The experienced mammographer had an A_z of 0.91, whereas the average of the five radiologists yielded an A_z of 0.81. The partial area index $_{0.90}A_z'$ for the experienced mammographer was 0.58, whereas the partial area index $_{0.90}A_z'$ for the five radiologists was 0.28. Student t test for paired data was employed to evaluate the statistical significance of these differences (16). Results showed the differences in A_z and $_{0.90}A_z'$ to be statistically significant (two-tailed P values of .032 and .006).

The ability of the observers to distinguish malignant from benign masses was compared with that of the computerized method using the hybrid system. The differences in A_z and $_{0.90}A_z'$ between the experienced mammographer and the computerized method were found to be not statistically significant (two-tailed P values of 0.38 and 0.30) based on the evaluation from the CLABROC program (35,36) and the modified version of the CLABROC program (30). Results of the Student t test for paired data showed that the differences in A_z and ${}_{0.90}A_z'$ between the average performance of the five radiologists and the computerized method were both statistically significant (two-tailed *P* values of .0131 and .0015). Furthermore, statistically significant differences were found between the two in terms of the positive predictive value and the specificity at the 100% sensitivity level (two-tailed *P* values < .0001).

The differences in positive predictive value and specificity at 100% sensitivity between the average performance of the five radiologists and the performance of the experienced mammographer were 9% and 32%, respectively; these differences were also found to be statistically significant. The differences in positive predictive value and specificity at 100% sensitivity between the average performance of the five radiologists and that of the computerized scheme were even larger, 21% ($P < .0001$) and 63% ($P = .0001$), respectively. In other words, with the database we used, at a 100% sensitivity level (ie, no loss of malignant cases), the average radiologists misclassified or essentially overcalled 24 of the 26 benign cases, whereas the computer scheme misclassified only eight of the 26 benign cases.

DISCUSSION

We have developed a computer scheme that automatically extracts features of masses that are similar to those perceived by radiologists. Feature analysis has indicated that our computer-extracted features correlate well with the major features perceived by radiologists, as shown in Figures 4, 5, and 7. We have shown that spiculation (FWHM measure) is a dominant feature for analysis by both radiologists and computerized methods.

The shape of a mass can be described as regular or irregular, lobulated or not lobulated, circular or ovoid. Generally, shape is not as important as margin characteristics in the determination of the benign versus malignant status of a mass. An irregular shape can, however, be a useful sign of malignancy. We did not use a single measure to directly characterize the shape of a mass in our scheme. However, one can correctly identify irregular masses as suspicious for malignancy based on the spiculation measure, because the direction of the maximum gradient along the margin of an irregularly shaped mass can vary as greatly as that of a spiculated mass. A lobulated mass also has a higher spiculation value than a smooth circular or ovoid mass because the direction of the maximum gradient relative to the radial direction along the margin of a lobulated mass varies more than that along the margin of a smooth mass. Thus, a smooth lobulated mass will be ranked as more suspect for malignancy than a smooth circular or ovoid mass, similar to the rank ordering radiologists would give. Some lobulated masses might be classified into the spiculated category if they were very highly lobulated.

We have studied three types of classifiers with which to merge the computer-extracted features. The three classifiers mimic three possible ways that radiologists might merge the information that they perceive in the task of classification of masses. The combination of the rule-based method by using the spiculation measure with the ANN is probably the one that serves this task best for several reasons. First, introduction of the well-known importance of spiculation, which was also shown here, into our system with a one-step rule-based method allows the ANN to "concentrate" on acquiring its own knowledge for the more difficult features for which considerable overlap in the appearance of benign and malignant masses occurs. Second, the good correlation between the computer spiculation measure and an expert mammographer's spiculation ratings, as well as the similar performance of the two in distinguishing between benign and malignant masses, allows a reliable translation of the "intuitive" knowledge into a simple rule in the hybrid system. Third, in clinical practice, radiologists are more likely to process the information they perceive in the same way that is used in our hybrid system, namely, examining for spiculation, the only truly diagnostic feature for malignancy, first and then analyzing all the possible secondary features to determine the likelihood of malignancy.

We evaluated the classifiers by using self-consistency and round-robin methods. The consistency method yielded A_z values of 0.92, 0.93, and 0.98 for the two-step rule-based method, the four-input ANN, and the hybrid system, respectively. The round-robin evaluation was performed to test the generalization ability of the classifiers. The two-step rule-based method was investigated only to understand the features, and so we did not proceed with round-robin testing of this method. The round-robin evaluation of the four-input ANN yielded an A_z value of 0.90. The evaluation of the hybrid system with round-robin analysis on the three-input ANN yielded an A_z of 0.94. The rule was set on the spiculation measure in the hybrid system because spiculation is the major feature used intuitively by radiologists in predicting malignancy, and this rule did not undergo round-robin analysis. In addition, the aim of the spiculation measure was not to separate malignant masses completely from benign masses but to identify only those masses that were very likely to be malignant.

Generalization of a trained network is influenced by three factors: the size and efficiency of the training set, the architecture of the network, and the physical complexity of the problem. Round-robin method is one of the ways to validate a trained network on a data set different from the one used in the training. A valid generalization, however, can be guaranteed only when the training set size is sufficiently large relative to the architecture of the network (37). We are aware of the inadequacies of using a finite database, which is usually what is available in the medical field, and thus we provided details about the characteristics of the clinical database (Fig 1) used in the study.

Of the three types of classifiers, the hybrid system performed the best in differentiating malignant from benign masses. One could expect that the performance of the ANN would be similar to that of the hybrid system because they both use the same four features as the inputs. Of the three classifiers, however, the ANN performed the worst in terms of A_z, $_{0.90}A_z'$, and the performance at 100% sensitivity. Although the difference in A_z between the hybrid system and the ANN was not statistically significant (two-tailed *P* value of .2), the difference in performance between the two classifiers at the high sensitivities ($_{0.90}A_z'$) was found to be statistically significant (two-tailed *P* value of .008). This difference resulted from the dominant nature of the spiculation measure, which kept the ANN-alone method from learning the importance of the other three features in differentiating subtle malignant masses from benign masses (38). It seems that only when the ANN was used after the spiculation criterion did it learn to effectively interpret the complicated interrelationship among the remaining features in determining the benign or malignant status of the subtle cases. With an unlimited database, the ANN-alone method might learn as well as the combined rule-based ANN method in distinguishing between benign and malignant masses (both spiculated and nonspiculated). Nevertheless, it was still more efficient to bring well-known knowledge directly into a classifier to avoid lengthy training times, the need for larger databases, and uncertainty in whether the final learning goal (well-known rules) would be achieved.

The performances of the expert mammographer and our computerized classification scheme were significantly better than the average performance of radiologists with less mammographic experience, and this difference was even greater at the high sensitivity levels (*P* values ranging from .013 to <.001). Variability in radiologists' interpretations of mammograms is due to the differences in their knowledge and experience and has been demonstrated in our observer study and in other's work (39). The superior performance of the computerized classification scheme in distinguishing malignant masses from benign masses, especially at high sensitivity levels, emulates the performance of an expert mammographer. This finding underscores the potential usefulness of a computer-aided diagnosis classification scheme as an aid to improving the performance of less experienced mammographers and thus reducing the variability among radiologists in their mammographic interpretation and reducing the number of biopsies performed for benign masses.

ACKNOWLEDGMENTS

The authors are grateful to Charles E. Metz, PhD, and Robert M. Nishikawa, PhD, for their useful discussions; to Wei Zhang, PhD, and Ping Lu, MS, for discussions on artificial neural networks; and to Ulrich Bick, MD, Regina C. Haldemann-Heusler, MD, Laurence Monnier-Cholley, MD, Mitchell Sussman, MD, and Jon Lee, MD, for participation in the observer study.

REFERENCES

1. Sickles EA. Mammographic features of 300 consecutive nonpalpable breast cancers. AJR 1986; 146:661-663.
2. Sickles EA. Periodic mammographic follow-up of probably benign lesion: result in 3,184 consecutive cases. Radiology 1991; 179:463-468.
3. D'Orsi CJ, Swets JA, Pickett RM, Seltzer SE, McNeil BJ. Reading and decision aids for improved accuracy and standardization of mammographic diagnosis. Radiology 1992; 184:619-622.
4. D'Orsi CJ, Kopans DB. Mammographic feature analysis. Semin Roentgenol 1993; 28:204-230.
5. Knutzen AM, Grisvold JJ. Likelihood of malignant disease for various categories of mammographically detected, nonpalpable breast lesions. Mayo Clin Proc 1993; 68:454-460.
6. Kopans DB. Breast imaging. Philadelphia, Pa: Lippincott, 1989.
7. Hall FM, Storella JM, Silverstone DZ, Wyshak G. Nonpalpable breast lesions: recommendations for biopsy based on suspicion of carcinoma at mammography. Radiology 1988; 167:353-358.
8. Kopans DB. The positive predictive value of mammography. AJR 1992; 158:521-526.
9. Mckenna RJ. The abnormal mammogram: radiographic findings, diagnostic options, pathology, and stage of cancer diagnosis. Cancer 1994; 74:244-255.
10. Chan HP, Doi K, Vyborny CJ, et al. Improvement in radiologists' detection of clustered microcalcification on mammograms: the potential of computer-aided diagnosis. Invest Radiol 1990; 25:1102-1110.
11. Giger ML. Computer-aided diagnosis. In: Haus AG, Yaffe MJ, eds. Syllabus: RSNA categorical course on technical aspects of breast imaging. Oak Brook, Ill: Radiological Society of North America, 1994; 287-302.
12. Kegelmeyer WP, Pruneda JM, Bourland PD, Hillis A, Riggs MW,

Nipper ML. Computer-aided mammographic screening for spiculated lesions. Radiology 1994; 191:331-337.
13. Vyborny CJ, Giger ML. Computer vision and artificial intelligence in mammography. AJR 1994; 162:699-708.
14. Gale AG, Roebuck EJ, Riley P, Worthington BS. Computer aids to mammographic diagnosis. Br J Radiol 1987; 60:887-891.
15. Getty DJ, Pickett CJ, D'Orsi CJ, Swets JA. Enhanced interpretation of diagnostic images. Invest Radiol 1988; 23:240-252.
16. Wu Y, Giger ML, Doi K, Vyborny CJ, Schmidt RA, Metz CE. Artificial neural networks in mammography: application to decision making in the diagnosis of breast cancer. Radiology 1993; 87:81-87.
17. Swets JA, Getty DJ, Pickett RM, D'Orsi CJ, Seltzer SE, McNeil BJ. Enhancing and evaluating diagnostic accuracy. Med Decis Making 1991; 11:9-18.
18. Cook HK, Fox MD. Application of expert systems to mammographic image analysis. Am J Physiol Image 1989; 4:16-22.
19. Jiang Y, Nishikawa RM, Wolverton DE, et al. Malignant and benign clustered microcalcifications: automated feature analysis and classification. Radiology 1996; 198:671-678.
20. Kegelmeyer WP. Computer detection of stellate lesions in mammograms. Proc SPIE 1992; 1660:446-454.
21. Giger ML, Vyborny CJ, Schmidt RA. Computerized characterization of mammographic masses: analysis of spiculation. Cancer Lett 1994; 77:201-211.
22. Huo Z, Giger ML, Vyborny CJ, et al. Analysis of spiculation in the computerized classification of mammographic masses. Med Phys 1995; 22:1569-1579.
23. Claridge E, Richter JH, Stonelake P. Measuring edge blur in mammographic lesions. In: Lemke HV, Rhodes ML, Jaffe CC, Felix R, eds. Computer Assisted Radiology (CAR '93). New York, NY: Springer-Verlag, 1993; 612-617.
24. Kilday J, Palmieri F, Fox MD. Classifying mammographic lesions using computerized image analysis. IEEE Trans Med Imaging 1993; 12:664-669.
25. Burdett CJ, Longbotham HG, Desai M, Richardson WB, Stoll JF. Nonlinear indicator of malignancy. Proc SPIE 1993; 1905:853-860.
26. Ackerman LV, Gose EE. Breast lesion classification by computer and xeroradiograph. Cancer 1972; 30:1025-1035.
27. Metz CE. ROC methodology in radiologic imaging. Invest Radiol 1986; 21:720-733.
28. Metz CE. Some practical issues of experimental design and data analysis in radiological ROC studies. Invest Radiol 1989; 24:234-245.
29. Giger ML, Doi K, MacMahon H, Metz CE, Yin FF. Pulmonary nodules: computer-aided detection in digital chest images. RadioGraphics 1990; 10:41-51.
30. Jiang Y, Metz CE, Nishikawa RM. A receiver operating characteristics partial area index for highly sensitive diagnostic tests. Radiology 1996; 201:745-750.
31. Giger ML, Doi K, MacMahon H. An intelligent workstation for computer-aided diagnosis. RadioGraphics 1993; 13:647-656.
32. Tabar L, Dean PB. Teaching atlas of mammography. New York, NY: Thieme, 1985.
33. Boone JM, Gross GW, Greco-Hunt V. Neural networks in radiologic diagnosis. I. Introduction and illustration. Invest Radiol 1990; 25:1012-1016.
34. Gross GW, Boone JM, Greco-Hunt V, Breenberg B. Neural networks in radiologic diagnosis. II. Interpretation of neonatal chest radiographs. Invest Radiol 1990; 25:1017-1023.
35. Metz CE, Shen JH, Herman BA. 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, Calif, August 1990.
36. Metz CE, Wang PL, Kronman HB. A new approach for testing the significance of differences between ROC curves measured from correlated data. In: Information Processing in Medical Imaging: Proceedings of the 8th Conference. Boston, Mass: Martinus Nijhoff, 1984; 432-445.
37. Baum EB, Haussler D. What set net gives valid generalization? Neural Computation 1989; 1:151-160.
38. Huo Z, Giger ML. Integrating rules and artificial neural networks in the classification of mass lesions in digital mammograms. Presented at the 1996 International Neural Network Society Annual Meeting, San Diego, Calif, 1996.
39. Elmore JG, Wells CK, Lee CH, Howard DH, Feinstein AR. Variability in radiologists' interpretations of mammograms. N Engl J Med 1994; 331:1493-1499.

Automated Seeded Lesion Segmentation on Digital Mammograms

Matthew A. Kupinski* and Maryellen L. Giger, *Member, IEEE*

Abstract—**Segmenting lesions is a vital step in many computerized mass-detection schemes for digital (or digitized) mammograms. We have developed two novel lesion segmentation techniques—one based on a single feature called the radial gradient index (*RGI*) and one based on simple probabilistic models to segment mass lesions, or other similar nodular structures, from surrounding background. In both methods a series of image partitions is created using gray-level information as well as prior knowledge of the shape of typical mass lesions. With the former method the partition that maximizes the *RGI* is selected. In the latter method, probability distributions for gray-levels inside and outside the partitions are estimated, and subsequently used to determine the probability that the image occurred for each given partition. The partition that maximizes this probability is selected as the final lesion partition (contour). We tested these methods against a conventional region growing algorithm using a database of biopsy-proven, malignant lesions and found that the new lesion segmentation algorithms more closely match radiologists' outlines of these lesions. At an overlap threshold of 0.30, gray level region growing correctly delineates 62% of the lesions in our database while the *RGI* and probabilistic segmentation algorithms correctly segment 92% and 96% of the lesions, respectively.**

Index Terms—**Computer-aided diagnosis, digital mammography, lesion segmentation, mass detection.**

I. Introduction

The University of Chicago is currently developing computerized schemes to detect mass lesions in digital (or digitized) mammograms [1]–[3]. Many computerized schemes initially return a number of locations called "potential lesion" sites. These are regions that the computer deems suspicious and require a closer examination. A lesion segmentation algorithm is then employed to extract the lesion or potential lesion from its surrounding tissues. Features can then be calculated using the segmentation information and classification can be accomplished using these features [4].

Numerous techniques have been developed to segment lesions from surrounding tissues in digital mammograms. Petrick *et al.* [5] employed density-weighted contrast enhancement (DWCE) segmentation algorithm to extract lesions and potential lesions from their surrounding tissues. Comer *et al.* [6] and Li *et al.* [7] used Markov random fields to classify the different regions in a mammogram based on texture. A lesion segmentation algorithm was developed by Sameti *et al.* [8] used fuzzy sets to partition the mammographic image data. Despite the difficulty and importance of this step in many computerized mass-detection schemes, few have attempted to analyze the performance of these segmentation algorithms alone, choosing instead to collectively analyze all components of a scheme.

In this paper, we present two methods for segmenting lesions in digital or digitized mammograms: a radial gradient index (*RGI*)-based algorithm and a probabilistic algorithm. These techniques are seeded segmentation algorithms; they begin with a point, called the seed point, which is defined to be within the suspect lesion. Many current computerized mass-detection schemes first employ an initial detection algorithm which returns locations that are used as seed points for the segmentation algorithm. In our previous research [4], a region growing algorithm [9], [10] was performed to extract the lesion from its surrounding tissues. Region-growing is a local thresholding process which utilizes only the gray-level information around the seed point. A series of partitions containing the seed point is created by thresholding, and rules (relating to size and circularity, for example) determine which partition best segments the suspect lesion. Potential problems with such methods are that the rules devised to choose the suspect lesion's partition are heuristic and often based on the first or second derivatives of noisy data. The new methods discussed in this paper attempt to solve the problems associated with conventional region growing by utilizing shape constraints to regularize the partitions analyzed, and simplifying the partition selection process by using utility functions based either on a single feature or probabilities. The performance of the two methods is compared against radiologists' outlines on a screening database of malignant lesions.

II. Lesion Segmentation

Given a subimage or region-of-interest (ROI) of dimension n by m containing the suspect lesion, we define the set of coordinates in this subimage as

$$\mathcal{I} = \{(x, y) : x = 1, 2, \cdots, n \text{ and } y = 1, 2, \cdots, m\}. \quad (1)$$

The function describing the pixel gray levels of this subimage is given by $f(x, y)$ where $(x, y) \in \mathcal{I}$. The values of $f(x, y)$,

Manuscript received February 18, 1998; revised July 15, 1998. This work was supported in part by the U.S. Army Medical Research and Materiel Command under Grants DAMD 17-96-1-6058 and 17-97-1-7202 and USPHS Grant RR11459. The Associate Editor responsible for coordinating the review of this paper and recommending its publication was N. Karssemeijer. *Asterisk indicates corresponding author.*

*M. A. Kupinski is with the Kurt Rossmann Laboratories, Department of Radiology, MC2026, The University of Chicago, 5841 South Maryland Avenue, Chicago, IL 60637 USA (e-mail: m-kupinski@uchicago.edu).

M. L. Giger is with the Kurt Rossmann Laboratories, Department of Radiology, The University of Chicago, Chicago, IL 60637 USA.

Publisher Item Identifier S 0278-0062(98)08046-X.

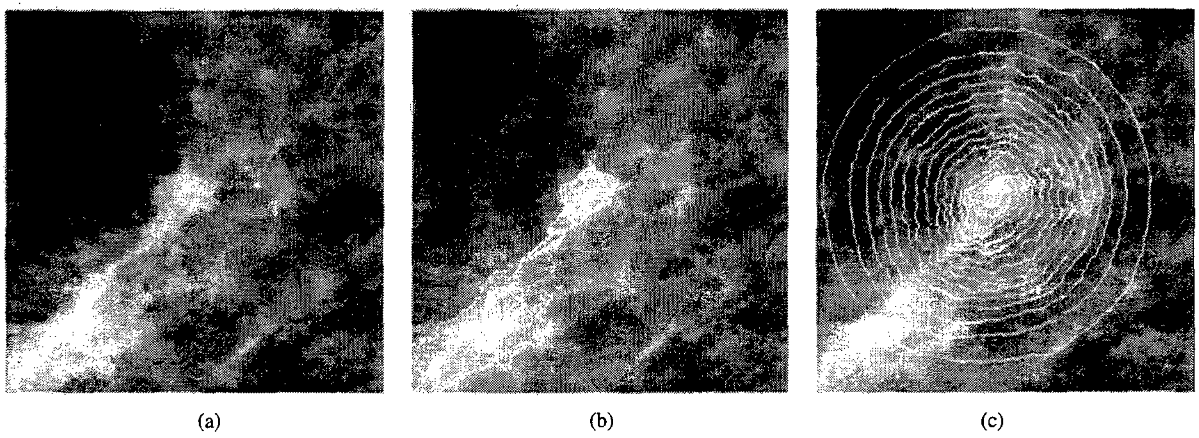


Fig. 1. Partitions that can arise (b) when only gray-level information, $f(x,y)$ is utilized in segmenting lesions and (c) when the gray-level image is multiplied by a constraint function to control the shape of the partitions, $h(x,y)$. The original image is shown in (a).

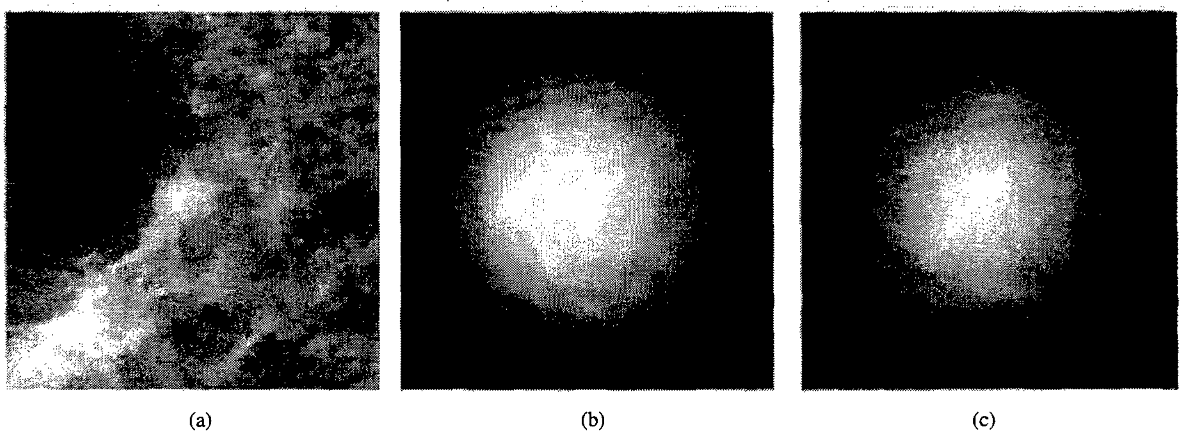


Fig. 2. The image (a) $f(x,y)$ of the lesion is multiplied by the Gaussian function (b) $N(x,y;\mu_x,\mu_y,\sigma_c^2)$ to constrain the partitions to have "lesion-like" shapes, which results in (c) the function $h(x,y)$. The value of σ_c^2 was set to 12.5^2 mm^2 for these images.

for this work, are bound between zero and one, with a zero representing black and a one representing white. The pixel values for all images were normalized to be within this range by dividing by the maximum pixel value possible for the digitizer used. The task of a lesion segmentation algorithm is to partition the set $\mathcal{I}$ into two sets: $\mathcal{L}$ which contains the coordinates of lesion pixels, and $\sim\mathcal{L}$ which contains surrounding background pixels. The lesion segmentation algorithms described in this paper are seeded segmentation algorithms; an initial point is used to start the segmentation. The seed point (μ_x,μ_y) is defined to be within the lesion, i.e., $(\mu_x,\mu_y)\in\mathcal{L}$ for all $\mathcal{L}$. In addition, the perimeter of the set $\mathcal{L}$ must be one continuous closed contour.

In order to segment the potential lesion, the "validity" of various image partitions $\mathcal{L}_i; i=1,\cdots,T$ is evaluated. For conventional region growing segmentation, the partitions are typically defined as

$$\mathcal{L}_i^{(\mathrm{rg})}=\{(x,y):f(x,y)>t_i\} \tag{2}$$

where t_i is a gray-level threshold. This makes use of the fact that lesions tend to be brighter than the surrounding tissue but it does not directly take shape into account, i.e., irregular shapes can be evaluated. Shape is, however, typically indirectly analyzed in these methods when searching for the partition to represent the segmented lesion [10], [11]. Fig. 1(b) shows an example of some of the irregular partitions that can arise in conventional region growing. The partitions are lesion-shaped at high thresholds, but tend to effuse into the background at lower thresholds, and are not representative of the lesion.

Conventional region growing defined the lesion partitions $\mathcal{L}_i^{(\mathrm{rg})}$ based solely on gray-level information in the image. The new algorithms proposed in this paper add additional *a priori* information into the creation of the lesion partitions. Lesions tend to be compact, meaning that their shapes are typically convex. To incorporate this knowledge into the creation of the partitions, the original image is multiplied by a function, called the constraint function, that suppresses distant pixel values. For this study we chose to use an isotropic Gaussian function centered on the seed point location (μ_x,μ_y) with a fixed variance σ_c^2 as the constraint function. The function $h(x,y)$ resulting from the multiplication of the original ROI with the constraint function is given by

$$h(x,y)=f(x,y)N\big(x,y;\mu_x,\mu_y,\sigma_c^2\big) \tag{3}$$

where $N(x,y;\mu_x,\mu_y,\sigma_c^2)$ is a circular normal distribution [see Fig. 2(b)] centered at (μ_x,μ_y) with a variance σ_c^2. Other constraint functions may be more appropriate for different

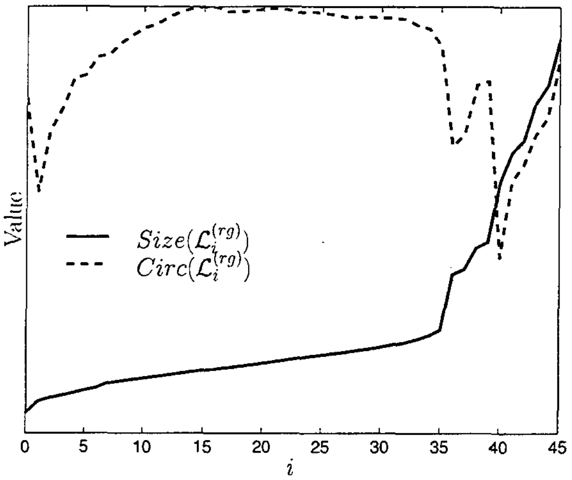


Fig. 3. Features employed in determining the final partition for conventional region growing. Here, i corresponds to the different gray-level intervals.

segmentation tasks. We have found, however, that a Gaussian works well for mammographic lesions. Fig. 2(c) shows an example of the function $h(x,y)$. At a given threshold, the partitions returned by thresholding are more compact than before because distant pixels are suppressed, i.e., a geometric constraint has been applied. The new partitions are defined as

$$\mathcal{L}_i = \{(x,y) : h(x,y) > t_i\}. \tag{4}$$

An example is shown in Fig. 1(c). Note that all of the partitions are now "lesion-like;" they are influenced by both the gray-level information and the geometric constraint. The value of the parameter σ_c^2 will be discussed later.

A. Region-Growing Segmentation

In conventional region growing, a feature or multiple features may be calculated for the partitions described in (2). For example, circularity *Circ*( ) and size *Size*( ) can be calculated for every $\mathcal{L}_i^{(rg)}$ as demonstrated in Fig. 3. The final partition is chosen by analyzing these functions and determining transition points or jumps in the features [4], [10], [11]. As Fig. 3 shows, the data can exhibit multiple transition points, and determining a jump by analyzing the first derivative of noisy data is difficult. If a transition point cannot be found, the segmentation algorithm fails to return a final partition.

B. Radial Gradient Segmentation

Given a series of partitions $\mathcal{L}_i$ from (4), one must determine which of these partitions best delineates the lesion. One method is to apply a utility function. Bick *et al.* [12] employed a *RGI* utility function in his lesion segmentation algorithm that utilized Fourier descriptors to describe the shapes of lesions. We have employed the *RGI* measure on the image $f(x,y)$ around the margin of each partition $\mathcal{L}_i$ as a utility function. For every partition $\mathcal{L}_i$ the *RGI* is calculated (see Fig. 5), and the partition with the maximum *RGI* is returned as the final lesion partition. It is important to note that the partitions $\mathcal{L}_i$ are generated using the processed image $h(x,y)$ while the *RGI* measure is computed on the original image $f(x,y)$.

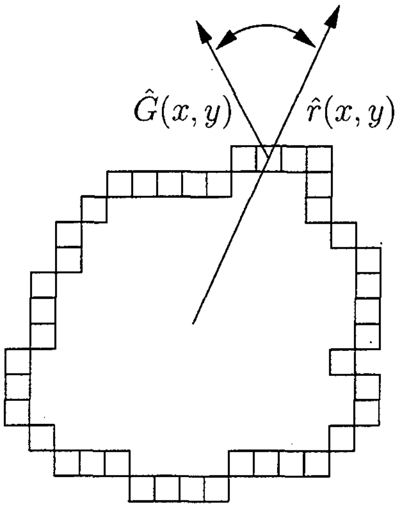


Fig. 4. The geometry used in calculating the *RGI*. The squares represent margin pixels $\mathcal{M}_i$ of the partition $\mathcal{L}_i$ being evaluated.

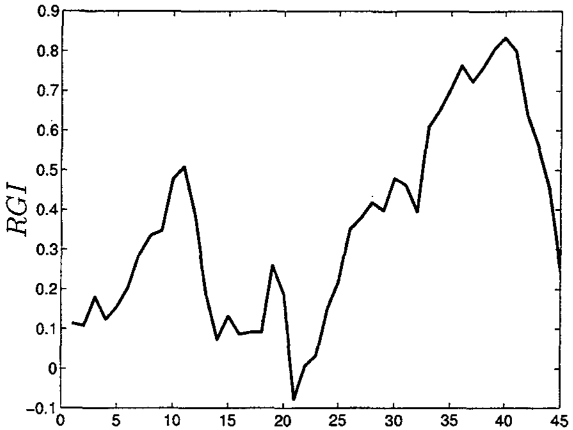


Fig. 5. The *RGI* as a function of the different partition $\mathcal{L}_i$ for the image shown in Fig. 1(a). The partition with the largest *RGI* value is returned as the final lesion partition. Here, i corresponds to the different gray-level intervals.

The *RGI* is computed as follows. Given a partition $\mathcal{L}_i$ (4) we can define the margin as

$$\begin{aligned}\mathcal{M}_i =&\{(x,y) : (x,y) \in \mathcal{L}_i \text{ and either } (x-1,y), \\ &(x+1,y), (x,y+1), \text{ or } (x,y-1) \notin \mathcal{L}_i\}. \end{aligned}\tag{5}$$

This states that a point is on the margin if it has at least one neighbor that is not in the lesion. The *RGI* is given by

$$RGI = \left(\sum_{(x,y)\in\mathcal{M}_i} \|\hat{G}(x,y)\|\right)^{-1} \sum_{(x,y)\in\mathcal{M}_i} \hat{G}(x,y)\,\frac{\hat{r}(x,y)}{\|\hat{r}(x,y)\|} \tag{6}$$

where $\hat{G}(x,y)$ is the gradient vector of $f(x,y)$ at position (x,y) and $(\hat{r}(x,y))/(\|\hat{r}(x,y)\|)$ is the normalized radial vector at the position (x,y) (Fig. 4). The *RGI* is a measure of the average proportion of the gradients directed radially outward. An *RGI* of one signifies that all the gradients around the margin are pointing directly outward along the radius vector and a -1

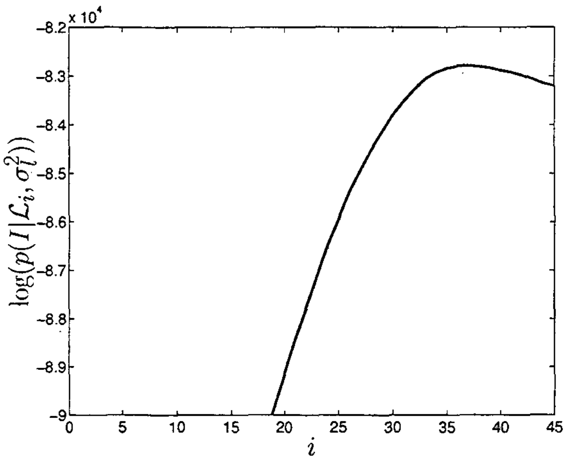


Fig. 6. A plot of the probability that the image occurred at different $\mathcal{L}_i$ for the image shown in Fig. 1(a). The maximum-likelihood estimate of the partition is given by the partition which maximizes this function. Here, i corresponds to the different gray-level intervals.

indicates that all the gradients around the margin are pointing directly inward toward the center of the partition. The *RGI* value around the margin of a circular lesion, for example, is one. If, however, $f(x,y)$ is a uniform image, then the *RGI* value will be zero even if the margin $\mathcal{M}_i$ is a circle.

C. Probabilistic Segmentation

The segmentation method based on probabilistic models is somewhat similar to the *RGI* method, except that the utility function is now a probability. The probability of pixel gray levels given a partition $\mathcal{L}_i$ (4) is modeled as

$$p(f(x,y) \mid \mathcal{L}_i, \sigma_l^2) = \begin{cases} N(f(x,y); f(\mu_x,\mu_y), \sigma_l^2) : (x,y) \in \mathcal{L}_i \\ z(f(x,y)) \qquad : (x,y) \notin \mathcal{L}_i \end{cases} \tag{7}$$

where $N(f(x,y); f(\mu_x,\mu_y), \sigma_l^2)$ is a normal distribution centered at the seed point gray level $f(\mu_x,\mu_y)$, with a variance σ_l^2, and $z(f(x,y))$ is a function to be described later. Lesions will not exhibit a large variation in pixel values, while the tissues surrounding the lesion may show large variation because they may consist of both fatty and dense regions. The uniformity of lesions is accounted for by a small-variance Gaussian function centered around the seed pixel value. The term $z(f(x,y))$ is a function that is estimated for each ROI using the gray levels from all of the pixels within the ROI although it is only employed in calculating $p(f(x,y) \mid \mathcal{L}_i, \sigma_l^2)$ for $(x,y) \notin \mathcal{L}_i$ [see (7)]. Finally, the probability of the image (or ROI) I given a partition $\mathcal{L}_i$ is

$$p(I \mid \mathcal{L}_i, \sigma_l^2) = \prod_{(x,y)\in\mathcal{I}} p(f(x,y) \mid \mathcal{L}_i, \sigma_l^2). \tag{8}$$

The partition $\mathcal{L}_i$ that is chosen is the one that maximizes the probability $p(I \mid \mathcal{L}_i, \sigma_l^2)$, i.e.,

$$p(I \mid \mathcal{L}_{\text{final}}, \sigma_l^2) = \underset{i}{\operatorname{argmax}} \{p(I \mid \mathcal{L}_i, \sigma_l^2)\}. \tag{9}$$

An example plot of $p(I \mid \mathcal{L}_i, \sigma_l^2)$ is shown in Fig. 6. Because there are a finite number of $\mathcal{L}_i$, we avoid the complexity of an optimization problem choosing instead to evaluate all $\mathcal{L}_i$ and determine the maximum.

The probability distribution for the gray levels when the pixels are outside the set $\mathcal{L}_i$ is given by the function $z(f(x,y))$ [see (7)], which is estimated from all gray levels within the ROI. Kernel density estimation using an Epanechnikov kernel was employed to estimate this distribution [13]. The width of the kernel was optimally determined through cross validation [13]. Kernel density estimation is a method similar to histogram analysis except that a nonrectangular kernel is used to bin data and this kernel is swept across the function axis continuously. Histogram analysis, on the other hand, uses a box-function that is moved in increments of the box width. Figs. 9 and 10 show the calculated probability distributions for gray levels inside and outside $\mathcal{L}_i$ for the ROI's shown in Figs. 7 and 8, respectively.

III. Results

A. Parameter Estimation

The width σ_c^2 of the constraint function in (3) was determined based on knowledge of lesions and was not statistically determined. A value of 12.5^2 mm^2 was empirically determined to work well for our purposes. Larger lesions were also segmented with this value but spiculations and small deviations around the edge of the lesion were usually not delineated.

The parameter σ_l^2 in (7) is an unknown quantity and must be determined. The average variation of the gray levels within the radiologist's outlined truth for a screening, malignant database of 118 visible lesions was estimated. Fig. 11 shows the density distribution for these variations as measured by the standard deviation of the gray levels within the radiologist's outlines. A value of 0.038 was determined to be the most common standard deviation of pixel values within the radiologist's outlines. It is important to note that problems may arise when the radiographic presentation of lesions in other databases are substantially different from those in the database employed in this study. We, however, employed a database of 60 malignant, nonpalpable lesions obtained from roughly 700 needle biopsies performed during the years 1987 to 1993 and, thus, should be representative of the actual distribution.

The value of σ_l^2 can also be determined for each lesion individually. Instead of just using the most probable *a priori* value of σ_l^2 (as discussed above) one can apply Bayes' theorem to find that

$$p(\sigma_l^2 \mid I, \mathcal{L}_i) = \frac{p(I \mid \mathcal{L}_i, \sigma_l^2) p(\sigma_l^2 \mid \mathcal{L}_i)}{p(I \mid \mathcal{L}_i)} \tag{10}$$

where $p(I \mid \mathcal{L}_i, \sigma_l^2)$ is given by (8). If we assume that σ_l^2 and $\mathcal{L}_i$ are independent, then $p(\sigma_l^2 \mid \mathcal{L}_i) = p(\sigma_l^2)$. The distribution of $p(\sigma_l^2)$ can be obtained from Fig. 11. Finally, we know that $p(I \mid \mathcal{L}_i) = \int d\sigma_l \, p(I \mid \mathcal{L}_i, \sigma_l^2) p(\sigma_l^2)$ which results in

$$p(\sigma_l^2 \mid I, \mathcal{L}_i) = \frac{p(I \mid \mathcal{L}_i, \sigma_l^2) p(\sigma_l^2)}{\int d\sigma_l \, p(I \mid \mathcal{L}_i, \sigma_l^2) p(\sigma_l^2)}. \tag{11}$$

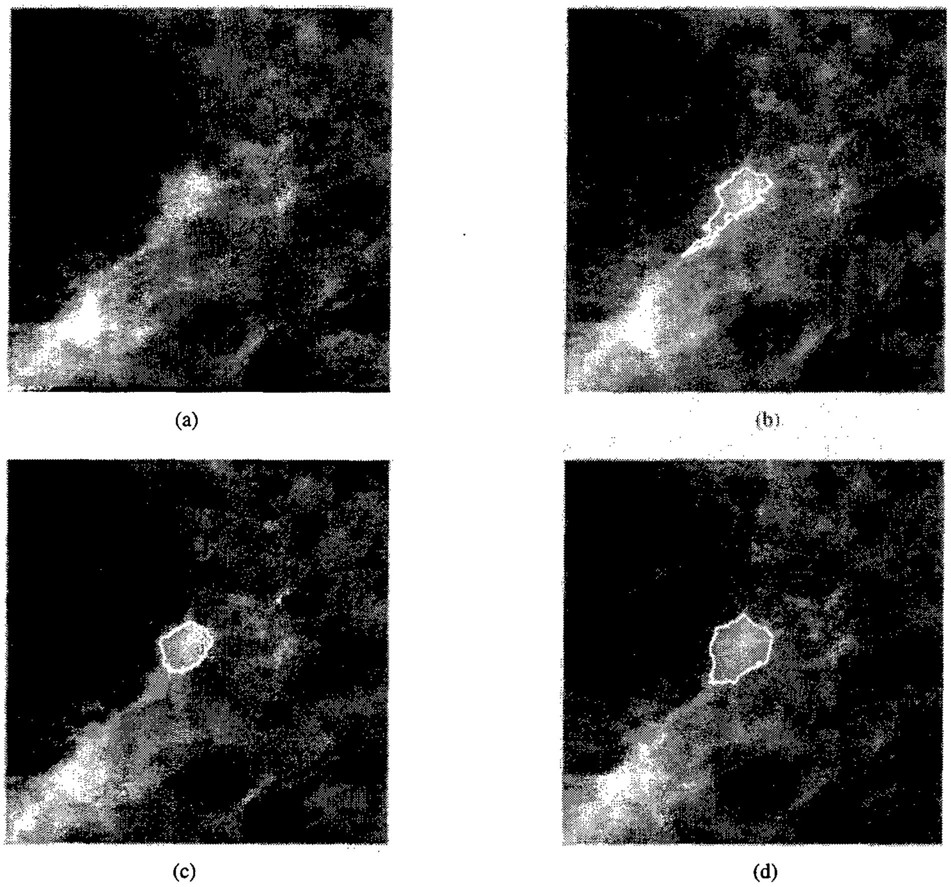


Fig. 7. Segmentation results for (a) a high-contrast lesion using (b) region growing, (c) *RGI*-based segmentation, and (d) probabilistic segmentation.

The probability of various values of σ_l^2 could be compared against each other and the optimal σ_l^2 estimated. Unfortunately, to estimate $p(\sigma_l^2 \mid I, \mathcal{L}_i)$ one must compute

$$\int d\sigma_l \, p(I \mid \mathcal{L}_i, \sigma_l^2) p(\sigma_l^2) \tag{12}$$

which involves integrating over all possible values of σ_l and is very time consuming. Not only do we have the problem of integrating over all σ_l values but the value computed is the probability given a partition $\mathcal{L}_i$. This leaves us with a dual optimization task. For a given σ_l^2 the optimal partition $\mathcal{L}_{\mathrm{final}}$ is determined. This partition is then employed to determine a new optimal σ_l^2. This process continues until there is convergence. For this research, we instead employed a constant value, i.e., the most probable *a priori* value of σ_l^2.

B. Segmentation Performance

Segmentation results for a relatively simple (high contrast) lesion are shown in Fig. 7. All three methods, region growing, *RGI*-based segmentation, and probabilistic segmentation, perform well on this lesion. Region growing has somewhat undergrown the lesion and has a long tail. The *RGI*-based method and the probabilistic method segment the lesion better than region growing. Similar images are shown for a more difficult lesion on a border between a fatty region and the pectoralis muscle (Fig. 8). Because of the brightness of the pectoralis muscle, region growing effuses into the background too soon and thus, the transition point found results in a grossly undergrown lesion. There are also vessels that can be radiographically seen passing through the center of this lesion. The *RGI*-based segmentation algorithm chooses the boundary of a vessel as the best partition because the *RGI* value around the vessel is larger than that around the actual lesion. The probabilistic segmentation algorithm, however, does not get confused by the vessel inside the lesion and correctly segments this difficult lesion.

In order to quantify the performance differences between the three different segmentation methods, the segmentation results were compared against radiologists' outlines of the lesions. The screening database of nonpalpable, biopsy-proven, malignant cancers with a total of 118 visible lesion ROI's was employed. For each lesion the seed point was calculated from the center of mass of the radiologist's outline. Once the lesion was segmented, an overlap measure O was calculated using the set returned from the segmentation algorithm $\mathcal{L}$ and the radiologist's hand-drawn segmentation set $\mathcal{T}$. The overlap O is defined as the intersection over the union, i.e,

$$O = \frac{\mathrm{Area}\,(\mathcal{L} \cap \mathcal{T})}{\mathrm{Area}\,(\mathcal{L} \cup \mathcal{T})}. \tag{13}$$

The value of O is bound between zero (no overlap) and one (exact overlap). A threshold needs to be set in order to classify a result as an "adequate" segmentation, i.e., if O is greater than

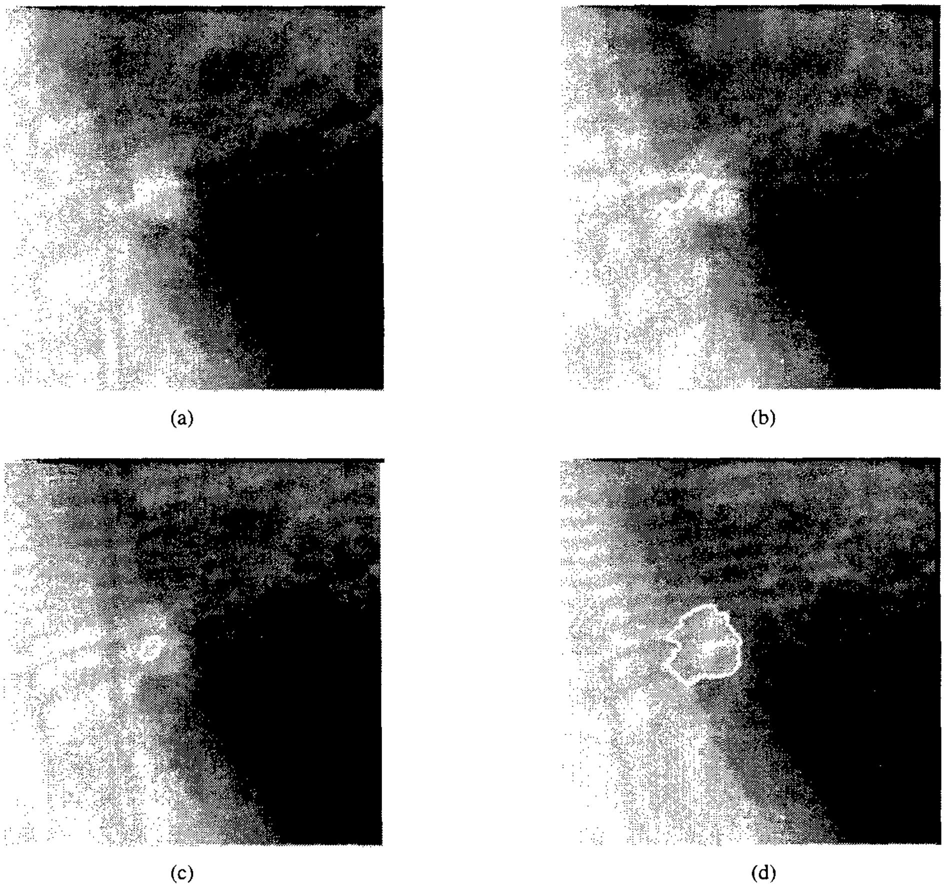


Fig. 8. (a) Segmentation results for a lesion on the boundary between a fatty area and the pectoralis muscle using (b) region growing segmentation, (c) *RGI*-based segmentation, and (d) probabilistic segmentation.

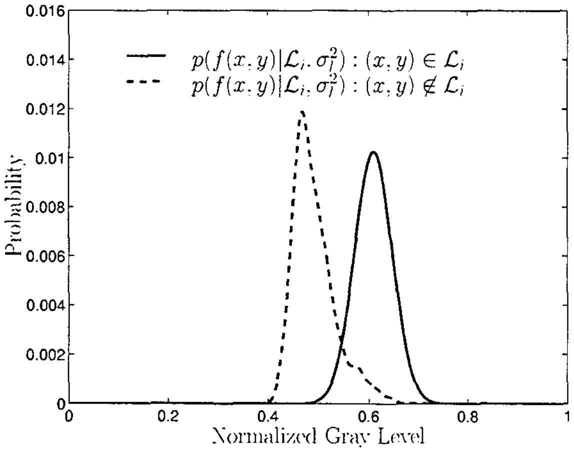


Fig. 9. Probability distributions employed when a pixel is inside or outside of the set in question for the image shown in Fig. 7. The distribution employed when $(x,y) \in \mathcal{L}_i$ Gaussian centered at the seed point gray level with a variance of σ_f^2. The distribution $z(f(x,y))$ is employed when $(x,y) \notin \mathcal{L}_i$ and is estimated from all gray values within the ROI.

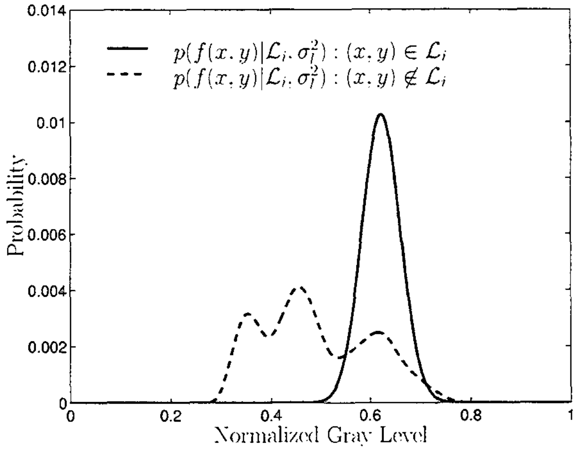


Fig. 10. Probability distributions employed when a pixel is inside or outside of the set in question for the ROI shown in Fig. 8. The distribution employed when $(x,y) \in \mathcal{L}_i$ is a Gaussian centered at the seed point gray level with a variance of σ_f^2. The distribution $z(f(x,y))$ is employed when $(x,y) \notin \mathcal{L}_i$ and is estimated from all gray values within the ROI.

a certain value then the lesion is considered to be correctly segmented.

Fig. 12 shows a plot of the fraction of lesions correctly segmented at various overlap threshold levels. The probabilistic segmentation algorithm outperformed the other methods. Also shown in Fig. 12 is the performance of a different radiologist in extracting the lesions as compared with the first radiologist. It is interesting to note that the performances of the *RGI*-based and probabilistic methods are not too dissimilar from the human performance. Region growing never yielded all lesions correctly segmented even when the overlap threshold was zero because the method failed to find a transition point in many of the images.

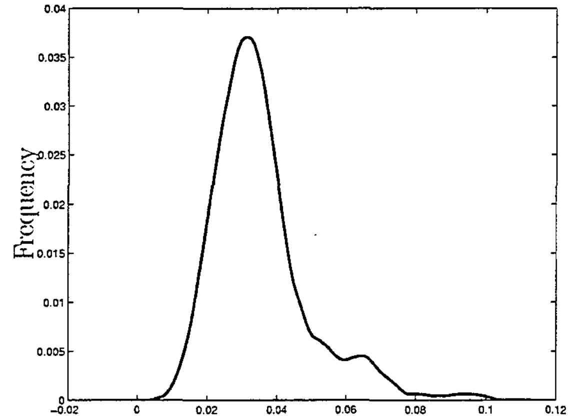


Fig. 11. The distribution of standard deviations of the gray levels within the radiologist's outlined lesions for a database of 60 malignant lesions (118 ROI's). The pixel values of the images were normalized to be between zero and one.

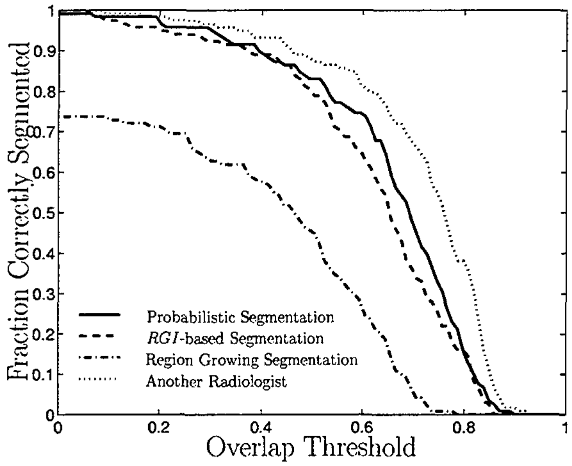


Fig. 12. The performance of the different segmentation methods on a database of malignant lesions as compared with a radiologist's outlines. Also shown is the agreement of another radiologist's outlines of the lesions in the databases with the outlines of the first radiologist.

IV. Discussion

Bayesian analysis could be applied to the probabilistic segmentation algorithm resulting in

$$p(\mathcal{L}_i \mid I, \sigma_l^2) = \frac{p(I \mid \mathcal{L}_i, \sigma_l^2) p(\mathcal{L}_i)}{p(I)}. \tag{14}$$

By analyzing (14) one finds that the $p(\mathcal{L}_i)$ is a term that penalizes partitions which are not "lesion" shaped. The partitions in our study, however, are obtained after the shape constraint function (4) has been applied so every partition analyzed is "lesion" shaped and thus, a Bayesian analysis is not necessary. If deformable contours are employed instead of a series of lesion-shaped partitions, then Bayes' rule (14) should be applied.

The assumption throughout this paper has been that appropriate partitions can be generated by gray-level thresholding the function $h(x, y)$ (3). This assumption, as is shown by the results of this paper, is generally appropriate for most lesions. There are, however, cases where thresholding $h(x, y)$ does not generate adequate partitions for a given lesion. In some cases, oddly shaped lesions may be surrounded by glandular structures which may confuse the algorithm into calling those normal structures part of the lesion. Spiculations, which are common in malignant lesions, are, in general, not included in the final lesion partition because of the application of the constraint function. The purpose of the segmentation algorithms described in this paper, however, is to determine the general shape of the lesions and not necessarily the detailed shape in which all spiculations are demarcated.

There is an implicit model that arises from the density functions employed in the probabilistic segmentation algorithm. Equation (7) assumes that all pixels within the lesion come from a Gaussian distribution centered at the seed point pixel value. The lesion model from which this distribution arises is a very simple one: a lesion has uniform gray levels with fluctuations arising from both noise and structure. In the future, more complex models, such as modeling a lesion as a projection of a sphere, can be implemented. The distributions, however, become more difficult with which to work and the assumption of independence in (8) and (10) is no longer valid.

Different initial seed points will result in different segmentation results. For both the *RGI*-based and probabilistic segmentation algorithms, the results are very similar given small changes in the seed point location. If, however, the seed point is selected to be at the very edge of the lesion, then the final partitions returned by both the *RGI*-based and probabilistic algorithms will be poor.

We comparatively evaluated the three segmentation methods at various overlap criteria (Fig. 12) because different investigators may use different evaluation criteria as well as different databases. Previously, we have shown that the reported performance of a computer detection method can greatly vary depending on the criteria used in tabulating sensitivity and specificity [14].

The performance differences between the probabilistic algorithm and the *RGI*-based method are small. Both, however, substantially outperform conventional region growing. It is expected that this better segmentation performance will, in the future, result in more meaningful features being extracted from potential lesion regions, and, ultimately, in better classification of malignant lesions from normal tissue regions.

V. Conclusion

We have developed two new methods of seeded lesion segmentation for use in digital mammography. These new methods substantially outperform conventional region growing segmentation. At an overlap threshold of 0.3, region growing correctly identified 62% of the lesions in our database, while the *RGI*-based and probabilistic segmentation methods correctly segmented 92% and 96% of the lesions, respectively. With these new segmentation results we hope to find and extract new features that will help differentiate between actual lesions and false detections, thus improving the overall performance of computerized mass detection.

Acknowledgment

The authors would like to thank M. Anastasio, R. Engelmann, and D. Edwards for their many helpful discussions. The authors would also like to thank U. Bick and C. Vyborny for collecting and outlining the lesions in the database.

M. L. Giger is a shareholder in R2 Technology, Inc. (Los Altos, CA). It is the University of Chicago Conflict of Interest Policy that investigators disclose publicly actual or potential significant financial interest which would reasonably appear to be directly and significantly affected by the research activities.

References

[1] M. L. Giger, R. M. Nishikawa, K. Doi, F.-F. Yin, C. J. Vyborny, R. A. Schmidt, C. E. Metz, Y. Wu, H. MacMahon, and H. Yoshimura, "Development of a "smart" workstation for use in mammography," in *SPIE*, vol. 1445, pp. 101–103, 1991.

[2] M. L. Giger, K. Doi, H. MacMahon, R. M. Nishikawa, K. R. Hoffmann, C. J. Vyborny, R. A. Schmidt, H. Jia, K. Abe, X. Chen, A. Kano, S. Katsuragawa, F.-F. Yin, N. Alperin, C. E. Metz, F. M. Behlen, and D. Sluis, "An "intelligent" workstation for computer-aided diagnosis," *Radiographics*, vol. 13, pp. 647–656, 1993.

[3] F.-F. Yin, M. L. Giger, K. Doi, C. E. Metz, C. J. Vyborny, and R. A. Schmidt, "Computerized detection of masses in digital mammograms: Analysis of bilateral subtraction images," *Med. Phys.*, vol. 18, pp. 955–963, 1991.

[4] M. Kupinski, M. L. Giger, P. Lu, and Z. Huo, "Computerized detection of mammographic lesions: Performance of artificial neural network with enhanced feature extraction," in *SPIE*, vol. 2434, pp. 598–605, 1995.

[5] N. Petrick, H. P. Chan, D. Wei, B. Sahiner, M. A. Helvie, and D. D. Adler, "Automated detection of breast masses on mammograms using adaptive contrast enhancement and texture classification," *Med. Phys.*, vol. 23, no. 10, pp. 1685–1696, 1996.

[6] M. L. Comer, S. Liu, and E. J. Delp, "Statistical segmentation of mammograms," in *Digital Mammography*, K. Doi, Ed., International Congress Series. Amsterdam, the Netherlands: Elsevier, 1996, pp. 471–474.

[7] H. D. Li, M. Kallergi, L. P. Clarke, and V. K. Jain, "Markov random field for tumor detection in digital mammography," *IEEE Trans. Med. Imag.*, vol. 14, pp. 565–576, June 1995.

[8] M. Sameti and R. K. Ward, "A fuzzy segmentation algorithm for mammogram partitioning," in *Digital Mammography*, K. Doi, Ed., International Congress Series. Amsterdam, the Netherlands: Elsevier, 1996, pp. 471–474.

[9] J. C. Russ, *The Image Processing Handbook.* Boca Raton, FL: CRC, 1992.

[10] T. Matsumoto, H. Yoshimura, K. Doi, M. L. Giger, A. Kano, H. MacMahon, K. Abe, and S. M. Montner, "Image feature analysis of false-positive diagnoses produced by automated detection of lung nodules," *Investigat. Radiol.*, vol. 27, pp. 587–597, 1992.

[11] Z. Huo, M. L. Giger, C. J. Vyborny, U. Bick, and P. Lu, "Analysis of spiculation in the computerized classification of mammographic masses," *Med. Phys.*, vol. 22, pp. 1569–1579, 1995.

[12] U. Bick, M. L. Giger, R. A. Schmidt, and K. Doi, "A new single-image method for computer-aided detection of small mammographic masses," in *Proc. CAR'95: Int. Symp. Computer and Communication Systems for Image Guided Diagnosis and Therapy*, 1995, pp. 357–363.

[13] T. J. Hastie and R. J. Tibshirani, *Generalized Additive Models*, AT&T Bell Labs, Great Britain, 1990.

[14] M. L. Giger, "Current issues in CAD for mammography," in *Digital Mammography*, K. Doi, Ed., International Congress Series. Amsterdam, the Netherlands: Elsevier, 1996, pp. 53–59.

Prospective Computer Analysis of Cancers Missed on Screening Mammography

Robert M. Nishikawa, Maryellen L. Giger, Robert A. Schmidt,*

Carl J. Vyborny, Ulrich Bick, and Kunio Doi

Department of Radiology, The University of Chicago

*Department of Radiology, New York University

1. Introduction

For a computer-aided detection (CAD) scheme to be an effective aid to radiologists, two conditions must be met. First, the computerized detection scheme must be able to detect cancers that a radiologist would overlook. Second, the radiologist when using the aid must act appropriately (i.e., correctly dismiss computer false positives and call back women with cancer). While at least three studies have indicated that automated detection schemes can find cancers missed on mammograms (Schmidt *et al.*, 1996; te Brake *et al.*, 1998; Warren-Burhenne *et al.*, 2000), these were all done using cases selected retrospectively. In this paper, we expand our study of the first requirement – that the computer can detect cancers overlooked on a screening mammogram – in a prospective study.

We previously reported on our prospective study of computerized detection of cancers on screening mammograms. We found that approximately 50% of cancers missed on a screening mammogram that are apparent in retrospect can be detected by one of our automated detection schemes (Nishikawa *et al.*, 1999). Visually, some of the overlooked cancers were very subtle and did not appear very different from normal breast tissue. In this study, we determined what fraction of these cancers are detectable in a screening-type environment.

2. Materials and Method

Cases and computer outputs used in this study were collected from a prospective study of CAD for screening mammography. At the University of Chicago Hospitals, we have been digitizing all screening mammograms since November 10, 1994. To identify which women in our study cohort have developed breast cancer, we compared the list of all patients included in our study against all breast pathology reports from our Hospital. For all women who had breast cancer, we examined all of their screening mammograms that were included in our study, along with diagnostic exams and, in some cases, needle localization exams. In this way, we were able to identify all cases where a cancer was visible on a screening mammogram. In some cases, these screening mammograms were read as abnormal and the women were called back, and in others, the cancer was overlooked and the mammogram was called normal. Here, we refer to the latter as a missed cancer.

To determine what fraction of these missed cancers can be detected in a screening environment, we conducted an observer study. We asked three radiologists to read 75 screening cases in which the cases containing missed cancers (n=21) were mixed with exams that contained a screen-detected cancer (n=3) and cases without cancer (n=51). The cases were presented in random order on a mammography motorized viewer. Magnifying glasses were available. No time limit was imposed.

The three radiologists were all specialists in breast imaging. Two had over 15 years experience and are MQSA qualified. The third, a European radiologist, with over 10 years of experience, had extensive experience in breast imaging, including digital mammography and breast MRI.

For each case, we included previous exams, when they were available. For each case, we asked the radiologist to give their BI-RADS assessment. Based on this assessment, we determined what fraction of radiologists would call back the cases containing a missed cancer. We also asked the radiologists to give their level of confidence that the patient should be called back for further imaging or for a biopsy. This was done using a visual analog scale with the left end marked as "definitely do not call back" and the right end marked as "definitely call back". The observers were instructed that if they were equivocal about calling the patient back, then they should mark the center of the scale. Short-term follow-up did not count as call back.

Two different detection schemes were used in this study: one for the detection of masses and the other for the detection of clustered microcalcifications. Details of these schemes have been described previously (Bick *et al.*, 1995; Nishikawa *et al.*, 1995; Yin *et al.*, 1993; Zhang *et al.*, 1996). Our prospective study began in November, 1994. The algorithms used throughout the study were kept constant, so those 1994 versions were used. Since then, the false-positive rate has been reduced, but these newer techniques have not been incorporated into the system yet (Anastasio *et al.*, 1998; Kupinski and Giger, 1998; Yoshida *et al.*, 1996).

3. Results

In the first three years of our study, 12,670 exams, which were obtained from 9195 women, were analyzed on our CAD workstation. Of these women, 79 developed breast cancer (minimum two years of follow-up). Sixty-one of the cancers were detected on a screening mammogram. The rest were detected on a diagnostic mammogram, or were palpable or both. Sixty-five cancers were visible mammographically. In the 79 cancer cases, 42 cases had a negative screening mammogram that was included in our study. Of the 42, 19 were mammographically occult in retrospect and 23 had a lesion that was visible at the site where the cancer developed. Examining the prospective computer results for those 23 cases showed that 12 of these cancers were detected by the computer.

All 12 of the computer-detected, radiologist-missed cancers and 9 of the 11 computer-missed, radiologist-missed cancers were used in the observer study. Two computer-missed, radiologist-missed cancer cases were not available for the study. Added to these 21 cases were 3 randomly selected screen-detected cases and 51 normal cases (based on at least two-year follow up) for a total of 75 cases. The normal cases were selected randomly from patients who had a screening mammogram in 1995 and at least one additional exam at least a two years later. Computer sensitivity on the cancer cases used in this study was 62.5% (15/24) and the false positive-rate on all 75 cases was 0.9 per image for calcifications and 2.1 per image for masses.

From the rating data, ROC curves were plotted (see Figure 1). In addition, using the BI-RADS assessment, we determined the sensitivity and specificity for each reader. These are shown as letters in Figure 1 and are reported in Table 1. Also listed in Table 1 are the sensitivity and specificity for the computer schemes and for the clinical interpretation of the screening case. The computer had at least one detection in each case and thus had a specificity of zero. The clinical readings had 100% specificity since the normal cases were found based on a normal screening mammogram. Similarly, the sensitivity of the clinical readings was low since we intentionally included exams where a cancer was overlooked. Note, however, that one of the cases detected clinically was missed by one of the three radiologists.

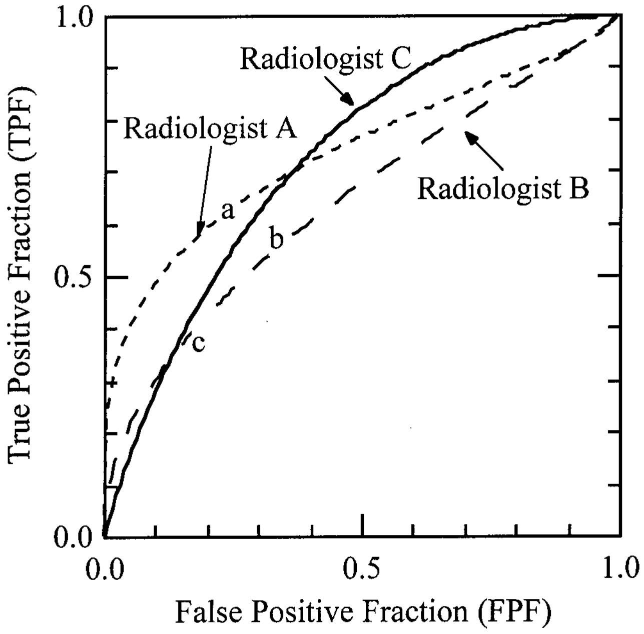


Figure 1. ROC curves for the 3 readers. The lower case letters indicate the operating points (sensitivity and specificity) as determined by their BI-RADS assessment. The areas under the ROC curves, A_z, ± one standard deviation were 0.73±0.06, 0.64±0.07, and 0.73±0.06 for radiologists A, B, and C respectively.

Based on the BI-RADS assessment, we determined the number of times a case was given a 0, 4, or 5 score (call back or biopsy). We then compared the computer performance of sub categories of the data based on the number of times the cases were called abnormal. We included the clinical reading in this analysis, so that there were four assessments made per case (see Table 2). Of the 12 cancer cases that were missed clinically and detected by the computer, 7 were detected by 2 of the 3 readers in this study and 11 were detected by at least one of the readers. One the other hand, some of the computer-detected cancers are below the detection threshold of experienced radiologists – 4 of the 12 cancers were not detected by any of the radiologists.

Table 1. Sensitivity and specificity for the three readers, the clinical reading and the computer.

Reader	Sensitivity	Specificity
A	63%	76%
B	58%	67%
C	38%	82%
Clinical	13%	100%
Computer	63%	0%

4. Discussion and Conclusions

The data in Table 2 show that the computer can detect cancers that are missed by a radiologist and the majority of those computer-detected missed cancers are detectable by a radiologist. When either 2 or 3 of 4 radiologists detected the cancer, the computer had high sensitivity, 89% (8/9). This is in spite of the fact that the overall sensitivity of our two computer schemes is approximately 70% for all cancer cases in our prospective study [Nishikawa, 1999 #5].

A possible drawback of CAD is that computer could increase the call-back rate. Approximately 80% of lesions identified by a radiologist in a normal mammogram were also identified by the computer as a potential lesion. In the same way that we infer that radiologists detecting missed

Table 2. Number of radiologists recommending call back or biopsy for the normal and cancer cases. Also include is the computer performance on those cases.

# of Radiologists Recommending Call Back	Normal Cases	Cancer Cases	# of Cancer Cases Detected by Computer	Computer Sensitivity
0/4	24	4	1	25%
1/4	17	9	4	44%
2/4	9	3	3	100%
3/4	1	6	5	83%
4/4	0	2	2	100%
Total	51	24	15	62%

cancers can lead to improved sensitivity, the high correlation of false-positive lesions between radiologists and the computer would indicate that the call-back rate *may* increase with implementation of CAD. This needs to be confirmed in clinical evaluations. One initial study found no increase in call-back rate when CAD was introduced (Warren-Burhenne *et al.*, 2000). However, the study did not report on whether sensitivity increased with CAD.

Increased call-back rate with CAD must be kept in context. Currently between 5-15% of all screening exams are considered abnormal and the patient is called back for further imaging studies. Since the cancer prevalence rate in a screening population is only 0.5%, approximately 10 to 30 women are called back for every cancer detected. If CAD can detect what would have been otherwise a missed cancer for every 10-30 extra women called back because of CAD, then the "cost/benefit ratio" remains unchanged, but a cancer would have been detected at an earlier stage. Because it is difficult to differentiate benign from malignant lesions mammographically, it is not reasonable to expect CAD to increase sensitivity, without increasing the number of call backs.

The data presented in this paper provide some evidence that computer-detected cancers can help radiologists avoid overlooking cancers. We plan to conduct an observer study to determine the number of cancers initially missed by a reader that are detected when the computer results are available. To determine the actual benefits and costs of using CAD, clinical trials need to be performed. As more systems become commercially available and more widely disseminated, these questions can readily be answered.

5. Acknowledgments

The authors thank Charles Metz for use of his ROC software. This work was done as part of the International Digital Mammography Development Group (IDMDG), and was supported in part by grants from the NCI (CA 60187 and T32 CA09649, the US Army (DAMD17-96-1-6058 and DAMD17-96-1-6228). The contents of this paper are solely the responsibility of the authors and do not necessarily represent the official views of any of the supporting organizations. All authors are shareholders in R2 Technology, Inc. (Los Altos, CA). It is the

University of Chicago Conflict of Interest Policy that investigators disclose publicly actual or potential significant financial interests that may appear to be affected by the research activities.

6. References

Anastasio, M. A., H. Yoshida, R. Nagel, R. M. Nishikawa, and K. Doi. 1998. A genetic algorithm-based method for optimizing the performance of a computer-aided diagnosis scheme for detection of clustered microcalcifications in mammograms. *Medical Physics* 25:1613-1620.

Bick, U., M. L. Giger, R. A. Schmidt, R. M. Nishikawa, D. E. Wolverton, P. Lu, C. J. Vyborny, and K. Doi. 1995. Automated segmentation of digitized mammograms. *Acad. Rad.* 2:1-9.

Kupinski, M., and M. L. Giger. 1998. Automated seeded lesion segmentation on digital mammograms. *IEEE Transactions on Medical Imaging* 17:510-517.

Nishikawa, R. M., M. L. Giger, K. Doi, C. J. Vyborny, and R. A. Schmidt. 1995. Computer-aided detection of clustered microcalcifications using digital mammograms. *Medicine and Biology in Engineering and Computing* 33:174-178.

Nishikawa, R. M., M. L. Giger, R. A. Schmidt, D. E. Wolverton, and K. Doi. 1999. Prospective testing of a clinical CAD workstation for the detection of breast lesions on mammograms. In: *Computer Aided Diagnosis in Medical Imaging*. K. Doi, H. MacMahon, M. L. Giger and K. R. Hoffmann, eds. Amsterdam: Elsevier.

Schmidt, R. A., R. M. Nishikawa, R. Osnis, K. L. Schreibman, M. L. Giger, and K. Doi. 1996. Computerized detection of lesions missed by mammography. In: *Digital Mammography '96*. K. Doi, M. L. Giger, R. M. Nishikawa and R. A. Schmidt, eds. Amsterdam: Elsevier Science.

te Brake, G. M., N. Karssemeijer, and J. H. C. L. Hendriks. 1998. Automated detection of breast carcinomas not detected in a screening program. *Radiology* 207:465-471.

Warren-Burhenne, L. J., S. A. Wood, C. J. D'Orsi, S. A. Feig, D. B. Kopans, K. O'Shaughnessy, E. A. Sickles, L. Tábar, C. J. Vyborny, and R. A. Castellino. 2000. Potential contribution of computer-aided detection to the sensitivity of screening mammography. *Radiology* 215:554-562.

Yin, F. F., M. L. Giger, C. J. Vyborny, K. Doi, and R. A. Schmidt. 1993. Comparison of bilateral-subtraction and single-image processing techniques in the computerized detection of mammographic masses. *Invest Radiol* 28:473-81.

Yoshida, H., K. Doi, R. M. Nishikawa, M. L. Giger, and R. A. Schmidt. 1996. An improved computer-assisted diagnostic scheme using wavelet transform for detecting clustered microcalicifications in digital mammograms. *Academic Radiology* 3:621-627.

Zhang, W., K. Doi, M. L. Giger, R. M. Nishikawa, and R. A. Schmidt. 1996. An improved shift-invariant artificial neural network for computerized detection of clustered microcalcifications in digital mammograms. *Medical Physics* 23:595-601.