AD_____

Award Number: DAMD17-96-1-6254

TITLE: Computer-Aided Diagnosis and Feature-Guided Data Reduction Systems in Mammography

PRINCIPAL INVESTIGATOR: Heang-Ping Chan, Ph.D.

CONTRACTING ORGANIZATION: University of Michigan
Ann Arbor, Michigan 48103-1274

REPORT DATE: October 2001

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.

20020828 059

REPORT DOCUMENTATION PAGE

Form Approved
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	3. REPORT TYPE AND DATES COVERED
	October 2001	Final (23 Sep 96 - 22 Sep 01)

4. TITLE AND SUBTITLE
Computer-Aided Diagnosis and Feature-Guided Data Reduction Systems in Mammography

5. FUNDING NUMBERS
DAMD17-96-1-6254

6. AUTHOR(S)
Heang-Ping Chan, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Michigan
Ann Arbor, Michigan 48103-1274

E-Mail: chanhp@umich.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
Report contains color.

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 Words)*

We have completed the pilot clinical study to evaluate the effects of CAD on radiologists' reading of screening mammograms. We have analyzed the results of about 2,400 cases of screening mammograms. The CADView system detected 90% of the lesions that were recommended for biopsy, and 86% of the fine needle biopsy cases in the two sites. Both the CAD system and the radiologists detected 11 of the 12 malignant cases. However, the missed cancer was not the same one for the computer and the radiologists. Therefore, the cancer detection by radiologists increased from 11 to 12 when they worked with CAD. CAD caused 34 additional callbacks and 2 additional benign biopsies. Although the number of cancers in this pilot study is small and the statistical uncertainty is large, our results indicate that the CAD system can increase the sensitivity of breast cancer detection for screening mammography in academic centers.

Two observer performance studies have been conducted for the CAD-guided image compression project. It was found that the proposed method with adequate bit rate will fully preserve the quality of microcalcifications and suspected microcalcifications without sacrificing the edge sharpness and overall image quality at an area-equalized bit-rate of about 0.4 bit/pixel. The CAD-guided compression can therefore reduce the image transmission and storage requirements for digital mammograms by a factor of about 30 without causing observable degradation of image quality. It can be an effective image compression method for picture archiving and communication and facilitate the implementation of telemammography and digital mammography.

14. SUBJECT TERMS
Mammography, Computer-aided diagnosis, breast cancer detection, Data compression

15. NUMBER OF PAGES
362

16. PRICE CODE

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

(3) Table of Contents

(1) FRONT COVER 1
(2) STANDARD FORM (SF) 298 2
(3) Table of Contents 3
(4) Introduction 4
(5) Body 5
  (a) CADView workstation 5
  (b) Collection of screening mammograms 6
  (c) Image compression of mammograms using a CAD-guided wavelet compression method 18
  (d) Improvement of microcalcification detection by optimizationof the neural network for pattern recognition 21
  (e) Evaluation of CAD mass detection algorithm with independent cases 22
  (f) Improvement of computerized mass detection on mammograms: fusion of two-view information 22
  (g) Optimization of wavelet decomposition for image compression and feature preservation 23
  (h) Predictive decomposition as a framework in dyadic transforms- A unified theory for wavelet and subband decompositons 23
(6) Key Research Accomplishments 24
(7) Reportable Outcomes 25
(8) Conclusions 34
(9) References 35
(10) Appendix 36

(4) Introduction

We have been developing CAD algorithms in detection of microcalcifications and masses using advanced image processing and computer vision techniques. Our CAD algorithms have provided very promising results in laboratory tests. The goals in this project are to implement our CAD algorithms in a fast workstation, develop user interfaces for efficient operation of the CAD programs, and conduct a pilot clinical trial of the CAD schemes at two mammographic screening sites. Based on the results of the pilot clinical trial, we can evaluate the sensitivity and specificity of the CAD algorithms, analyze the effects of the CAD schemes on mammographic screening, identify potential problems in a clinical environment, and develop methods to further improve the CAD schemes in the future. We believe that this is a crucial step to develop a clinically practical CAD workstation.

It has been recognized that digital mammography is one of the key research areas for improvement in the diagnosis of breast cancer. Two of the major issues in digital mammography are the technological requirements in developing high resolution digital detectors and the transmission and archiving the large amount of data. Data compression can reduce the amount of data for transmission and storage. However, there is often a tradeoff between compression ratio and image fidelity. Data compression in mammography is especially difficult because of the very subtle image details such as microcalcifications and mass margins that need to be preserved. In this project, we have developed a CAD guided data compression technique that preserves the original image information by lossless compression in potentially important regions on the mammograms indicated by the CAD programs. For breast areas outside these regions, the most efficient lossy compression technique that does not cause noticeable degradation of image details is applied. This image compression method will maximize the compression efficiency with a minimum loss of information.

With the support of this grant from the USAMRMC Breast Cancer Research Program, we have developed a CAD workstation with a proper graphical user interface for a pilot clinical trial. CAD workstations have been implemented at the University of Michigan and at the Georgetown University. We have recruited about 2,500 patients whose mammograms were read with and without CAD by radiologists. The effects of CAD reading have been evaluated. We have also implemented the CAD guided data compression technique for a data set of mammograms and conducted subjective image quality ranking studies to compare observer performance on the uncompressed images with that on images compressed with the selected lossy technique. Details of these studies have been described in previous annual progress reports and are summarized in this final report.

(5) Body

During the non-cost-time-extension period of 9/23/00 to 9/22/01, our goal is to continue to collect patient cases for the pilot study of the effects of CAD in mammographic screening. The images are read by radiologists without and with CAD using the CAD workstations at the University of Michigan and the Georgetown University:

In this final report, we summarize the major studies performed in the entire funding period (9/23/96-9/22/01) and the significant results obtained under the support of this grant. Some of the details have been reported in the previous years.

(a) CADView workstation

In the previous reports, we have discussed the design and operation of our PC-based CAD workstation, "CADView", and its graphical user interface (GUI) in detail. We will review briefly the operation of the CADView system used in the pilot clinical study, as shown in Figure 1. The radiologist will read the original film mammograms on the alternator as in their daily clinical practice. They will then retrieve the patient 4-view mammogram to be displayed on the CADView monitor by scanning the barcode of the patient folder. The mammograms displayed on the screen are arranged in exactly the same way as the films mounted on the alternator to facilitate the radiologist to compare the corresponding locations marked on the images. The display is placed next to the offline alternator and the radiologist can easily access the keyboard and mouse. The reading process is shown in Figure 2. The radiologist will mark any potential masses on the displayed images and record their impression of the most suspicious mass using the BI-RADS lexicon. They also select the BI-RADS action category for the mass that is recorded by the CAD system. Any potential microcalcification locations will then be marked and the BI-RADS impression and action category for the microcalcifications are recorded. The computer then displays the detected suspicious masses on the images. The radiologist will read the original films again based on the computer prompts. The radiologist can change their initial markings of masses on the displayed images if they are influenced by the computer output. They can also change the BI-RADS impression and the action category for the mass. The same procedure will also be performed for microcalcifications. The markings and action categories of the radiologist before and after CAD display are both recorded in a database file.

Figure 3 illustrates an example of the radiologist's markings on the displayed images. The double circles marked the location of the most suspicious mass in Figure 3(a) and the location of the most suspicious microcalcification clusters in Figure 3(b). The sliders on the right indicated the BI-RADS impression of the marked lesions. The right and left breasts were recorded separately. The BI-RADS action categories for the lesions were also selected on the sliders.

Figure 4 illustrates the same example after the CADView displayed the computer detection output. The computer detected masses were marked by arrowheads and the computer detected clusters were marked by dots. The radiologist's original marks were superimposed on the computer output. If there were disagreements, the radiologist could double-check the film mammograms on the alternator to resolve the discrepancy. If the radiologist found additional suspicious locations, he/she would add new marks on the displayed images. If the new locations were deemed more suspicious than the ones that he/she marked before the computer output was displayed, they could move the double circles to the new locations. The radiologist could also change their BI-RADS impression and action categories on the lesions by moving the pointers on the sliders.

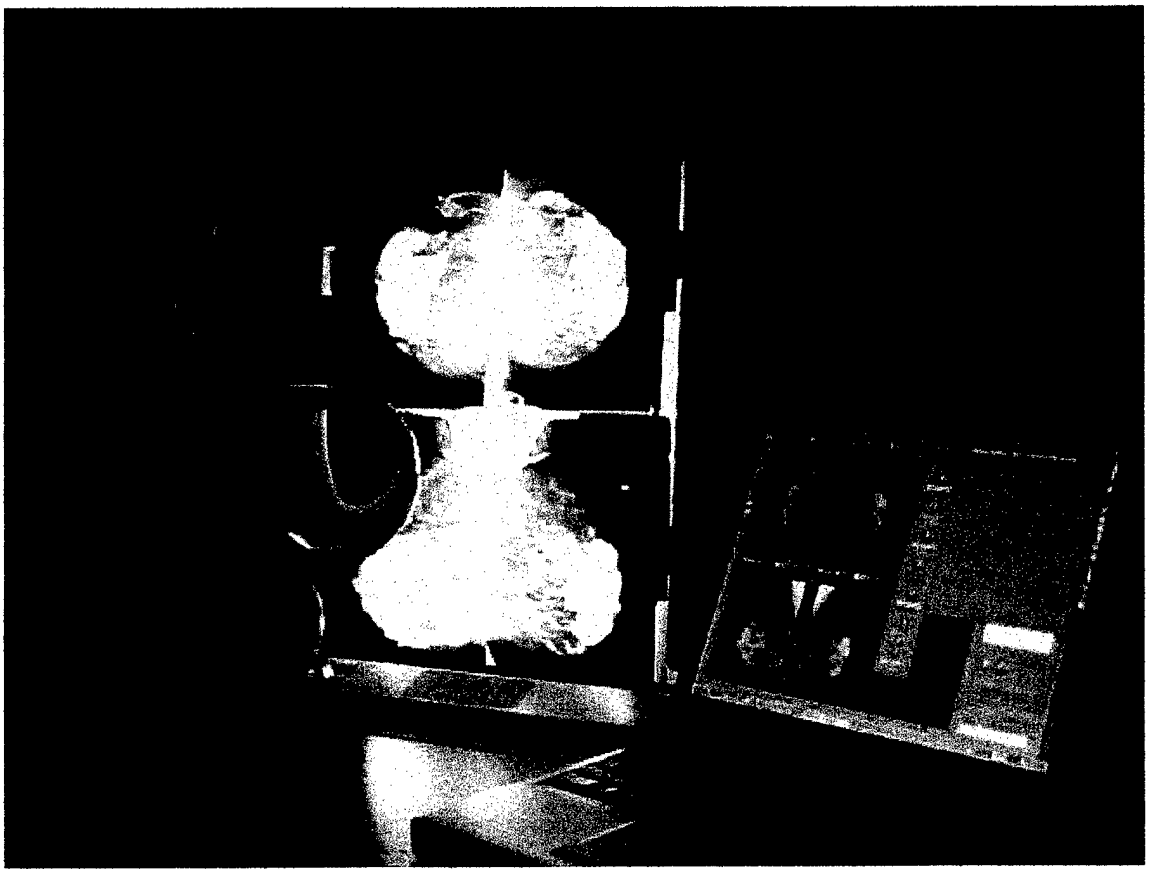


Figure 1. The setup for the CAD reading of off-line screening mammograms. The radiologist reads the original film mammmograms on the alternator while using the computer output as a second opinion.

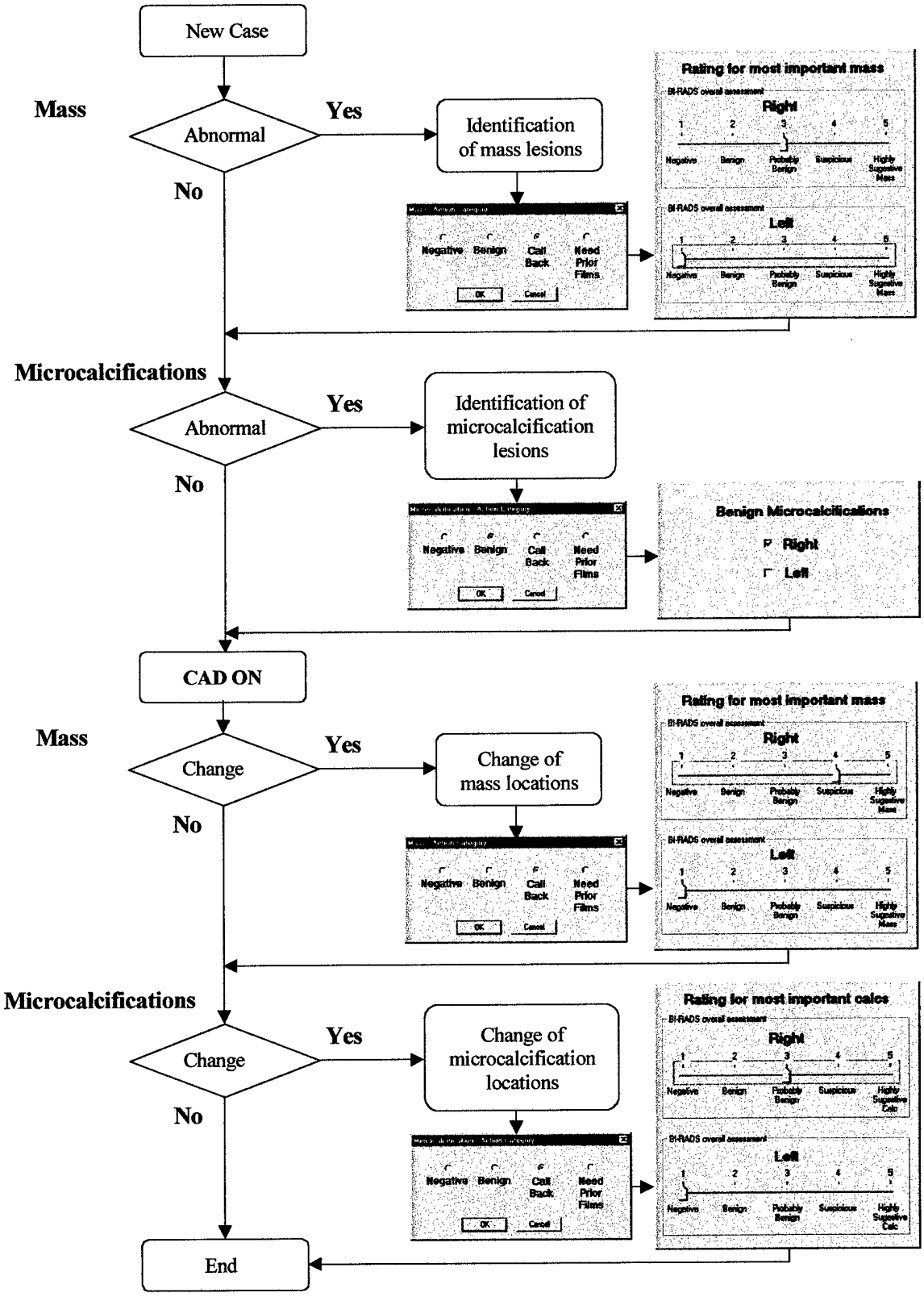


Figure 2. Sequence of reading and collection of the radiologist's BI-RADS assessment of a mammographic case.

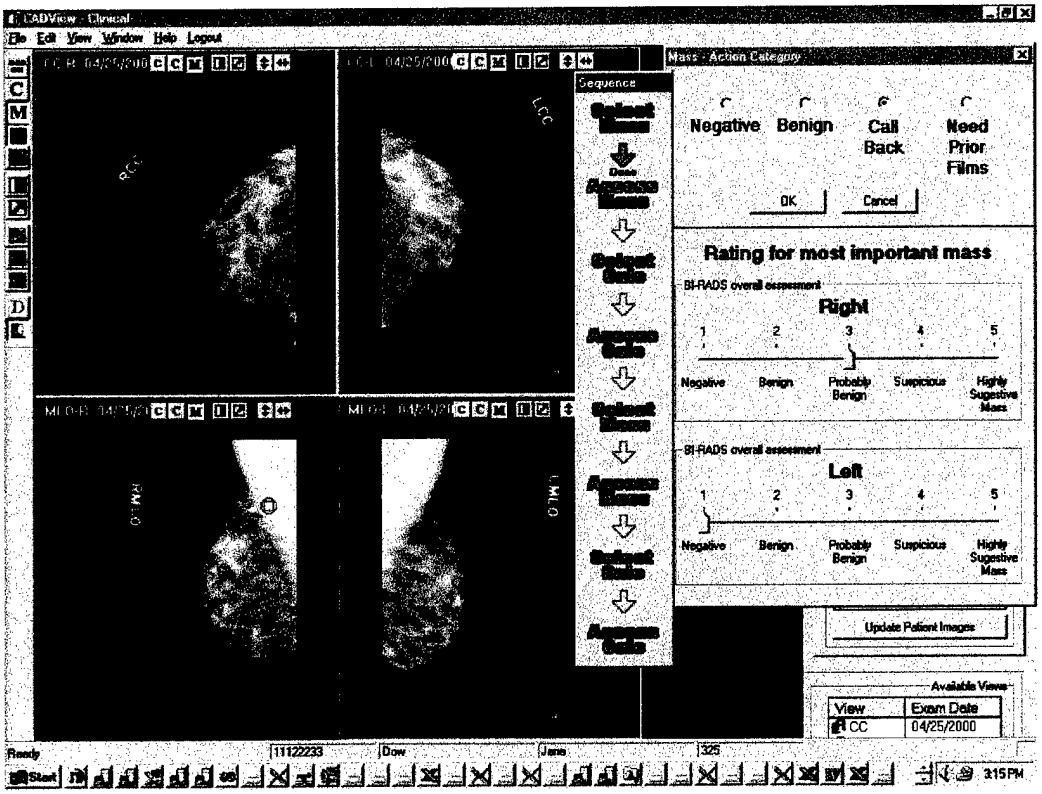


Figure 3(a). Radiologist's assessment of mass before CAD display.

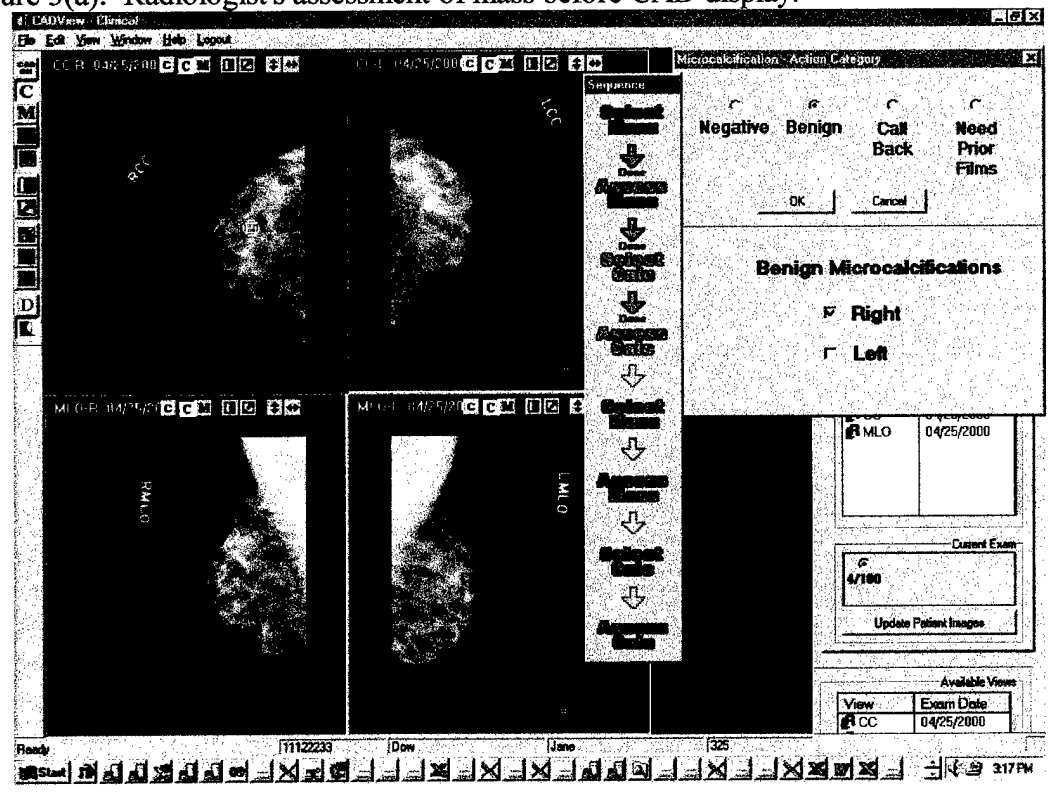


Figure 3(b). Radiologist's assessment of microcalcifications before CAD display.

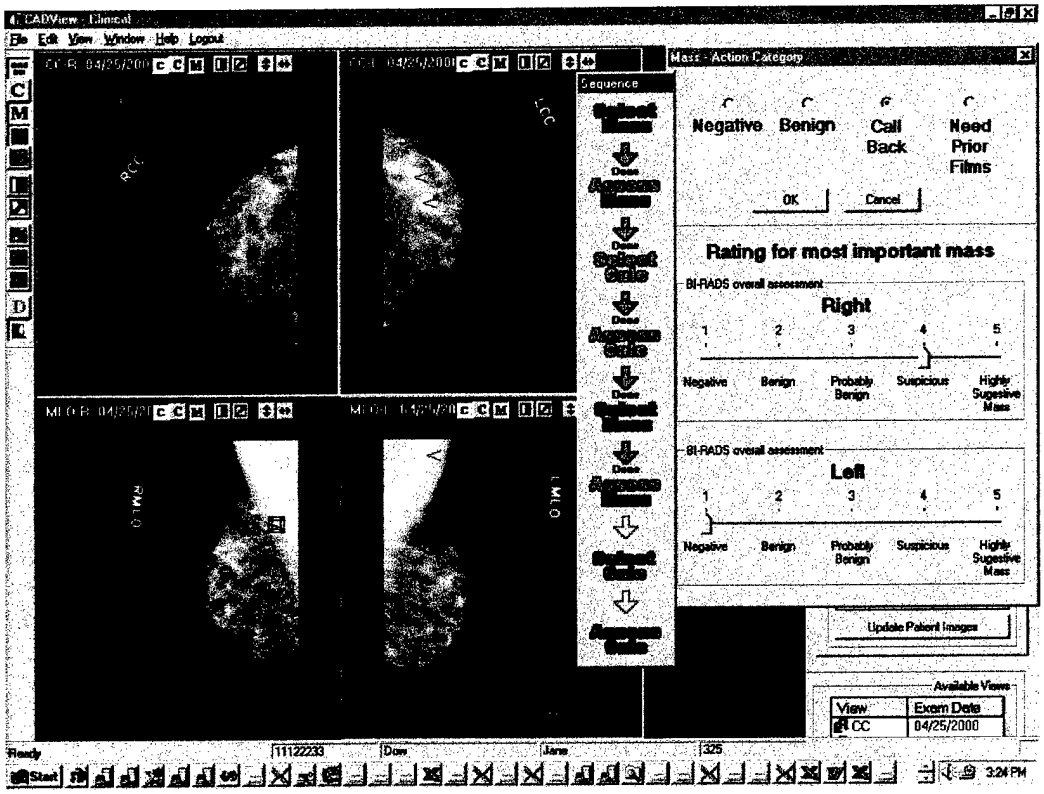


Figure 4(a). Radiologist's assessment of mass with CAD display.

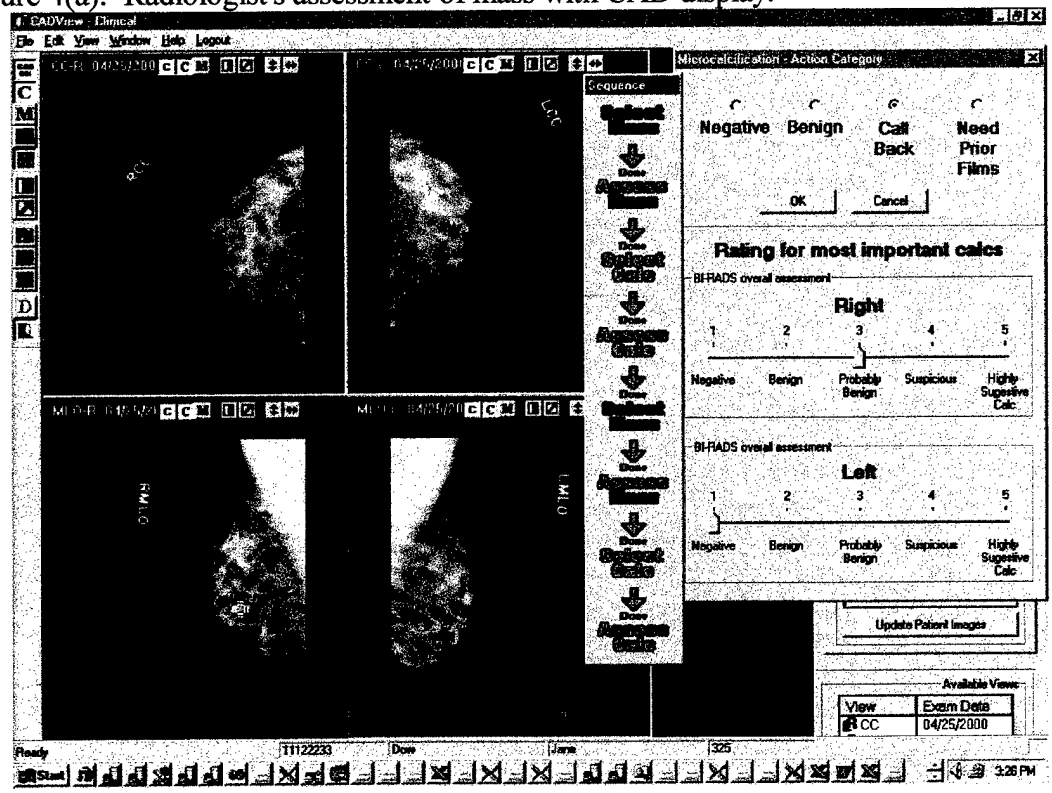


Figure 4(b). Radiologist's assessment of microcalcifications with CAD display.

(b) Collection of screening mammograms

To date, we have collected over 1700 cases from the University of Michigan (UM) breast imaging off-line screening sites, and over 1500 cases from the Georgetown University (GU) Breast Imaging clinic. In each site, there were many radiologists involved in the reading with the screening mammography cases with CAD.

(b.1) University of Michigan cases

We have analyzed the first 1665 cases. We do not have the callback results and follow up information on the other more recent cases yet because of the time delay between a decision to call back and the scheduled call back exam. The number of callbacks, biopsies, and follow-up cases within the first 1665 participating patients at the UM are summarized in Table 1. The results are compared with the GU data and the overall data later in Table 4. We can make the following observations from the UM cases:

1. For the cases that the radiologists recommended biopsy, the computer program detected 88% (23/26) of the lesions.
2. For the cases that radiologists recommended fine needle biopsy, the computer program detected 71% (5/7) of the lesions.
3. The computer detected 83% of the malignant cases (5/6) found in this patient group. One was mass cases, one was microcalcification case, and the other three were cases manifesting both mass and microcalcifications. The computer missed one malignant mass.
4. The computer caused 30 additional call backs, of which 5 were recommended 6 month follow-up, indicating that the computer found some areas of concern that the radiologists would not have called without the computer output. The development of the 6-month follow-up cases will be followed.
5. The computer caused 2 additional biopsies and 2 fine needle aspiration, all were found to be benign.
6. The computer has a detection sensitivity of 75% for masses, 80% for microcalcifications, and 97% for mixed mass and microcalcification cases, slightly lower than our predicted performance in laboratory tests but similar to the detection sensitivity found at the Georgetown University site. These results confirm that the performance of the CAD system is consistent in the patient population, although the two sites use different digitizers and different mammography systems.
7. The computer missed 10 cases that were recommended for 6 month short-term follow up. The development of these follow-up cases will be followed.
8. The computer missed 2 cases of fine needle aspiration and 3 biopsy cases, one of which was malignant.
9. The majority (41/56) of the false negative cases were found to be normal or benign after call back.

Table 1. The performance of the CADView detection system and its effects on radiologists' reading on the callback cases from the first 1665 off-line screening cases at the University of Michigan. The 12 month follow up indicates a regular annual screening schedule and these cases are thus generally considered to be normal. FN=false negative; FU=follow up.

	Biopsy	Fine needle aspiration	6 month FU	12 month FU	Overall	Cancer
Call Backs for Mass						
Radiologist detection	13	6	39	132	190	2
Computer detection	10	4	31	97	142	1
Sensitivity of computer (%)	77%	67%	79%	73%	75%	50%
Call Backs for Calcs						
Radiologist detection	7	0	15	13	35	1
Computer detection	7	0	13	8	28	1
Sensitivity of computer (%)	100%		87%	62%	80%	100%
Call Backs for Mass and Calcs						
Radiologist detection	6	1	10	12	29	3
Computer detection	6	1	10	11	28	3
Sensitivity of computer (%)	100%	100%	100%	93%	97%	100%
Overall Call Backs						
Radiologist detection	26	7	64	157	254	6
Computer detection	23	5	54	116	198	5
Sensitivity of computer (%)	88%	71%	84%	74%	78%	83%
Call Backs caused by CAD						
Mass	1	1	3	19	24	0
Microcalcifications	1	1	2	2	6	0
Computer False Negatives						
FN for Calcs	0	0	2	5	7	0
FN for Mass	3	2	8	35	48	1
FN for Mass and Calcs	0	0	0	1	1	0

(b.2) Georgetown University cases

At the GU site, a total of 1574 cases were digitized and processed by the CAD system. However, only 731 cases (46%) were reviewed in conjunction with the clinical reading by the radiologists due to some operation issues and mismatch of scheduling. The number of callbacks, biopsies, and follow-up cases within the 731 patients read with CAD at the GU are summarized in Table 2. The data are compared with the UM data and the overall data in Table 4.

1. For the cases that the radiologists recommended biopsy, the computer program detected 93% (13/14) of the lesions.
2. For the cases that radiologists recommended fine needle biopsy, the computer program detected 100% (7/7) of the lesions.
3. The computer detected all three malignant cases (6/6) found in this patient group. Two were mass cases, one was microcalcification case, and the other three were cases manifesting both mass and microcalcifications.
4. The computer caused 4 additional call backs, of which 1 was recommended biopsy and found to be malignant.
5. The computer has a detection sensitivity of 71% for masses, 81% for microcalcifications, and 91% for mixed mass and microcalcification case, slightly lower than our predicted performance in laboratory tests but similar to the detection sensitivity found at the UM site. These results confirm that the performance of the CAD system is consistent in the patient population, although the two sites use different digitizers and different mammography systems.
6. The computer missed 5 cases that were recommended for 3 to 6 month short-term follow up. The development of these follow-up cases will be followed.
7. The computer missed 1 biopsy mass cases, which was found to be benign.
8. The majority (25/31) of the false negative cases were found to be normal or benign after call back.

Table 2. The performance of the CADView detection system and its effects on radiologists' reading on the callback cases from the first 731 off-line screening cases at the Georgetown University. The 12 month follow up indicates a regular annual screening schedule and these cases are thus generally considered to be normal. FN=false negative; FU=follow up.

	Biopsy	Fine needle aspiration	3 month FU	6 month FU	12 month FU	Overall	Cancer
Call Backs for Mass							
Radiologist detection	7	6	1	7	65	86	2
Computer detection	6	6	0	5	44	61	2
Sensitivity of computer (%)	86%	100%	0%	71%	68%	71%	100%
Call Backs for Calcs							
Radiologist detection	4	0	2	6	15	27	0
Computer detection	4	0	2	4	12	22	1
Sensitivity of computer (%)	100%		100%	67%	80%	81%	
Call Backs for Mass and Calcs							
Radiologist detection	3	1	1	2	4	11	3
Computer detection	3	1	1	2	3	10	3
Sensitivity of computer (%)	100%	100%	100%	100%	75%	91%	100%
Overall Call Backs							
Radiologist detection	14	7	4	15	84	124	5
Computer detection	13	7	3	11	59	93	6
Sensitivity of computer (%)	93%	100%	75%	73%	70%	75%	120%
Call Backs caused by CAD							
Mass	0	0	0	0	3	3	0
Microcalcifications	1	0	0	0	0	1	1
Computer False Negatives							
FN for Calcs	0	0	0	2	3	5	0
FN for Mass	1	0	1	2	21	25	0
FN for Mass and Calcs	0	0	0	0	1	1	0

Table 3 and Table 4 summarize the overall results from the two institutions. Table 5 shows the ethnic composition of the mammography patient populations at UM and GU. The ethnic composition may affect the cancer prevalence rate in the patient population. The overall performance of the CADView system can be seen in the fourth column of Table 4. The sensitivities for the detection of masses (74%) and microcalcifications (81%) are slightly lower, as expected, than the sensitivities in laboratory tests. However, the detection sensitivity for malignant cases at 92% is higher than that in our laboratory data sets. Since the number of cancer cases is small, the statistical uncertainty will be large. The most promising result is that one additional cancer (detected cancer cases increased from 11 to 12) was detected when the radiologist used CAD. Although the computer and the radiologist both missed one cancer, the missed cancers were not the same one. When they worked together, the cancer detection sensitivity was increased. This result is consistent with that of a prospective clinical trial conducted by Freer et al. [1] in a community hospital. They found that their commercial CAD system increased the cancer detection rate of mammographic screening by 19.5% (from 41 to 49) in 12,860 patients. Our pilot results indicate that CAD may also be useful in academic institutions, although the gain in cancer detection may not be as high.

We estimated the change in the call back rate, the biopsy rate of the call back cases, and the biopsy rate relative to the number of screening cases, when radiologists read the mammograms with and without the influence of CAD. The call back rates without CAD were estimated from the statistics of the general off-line screening mammography patients. The results are tabulated in Table 6 for UM and Table 7 for GU. One interesting observation is that, at the UM, the call back rate without CAD was 10.4%. For the study group with CAD, the call back rate increased to 15.3%. If the cases caused by CAD were excluded, the call back rate was still high at 13.5%. The substantial increase in the call back rate seems to indicate that the radiologists at UM lowered their threshold for call back, either intentionally or unintentionally, when they worked with the CAD system even if the computer did not point to additional suspicious locations. This may cause an increase in their sensitivity for cancer detection with a tradeoff in increasing the call back rate. This may be one of the reasons that the CADView system did not increase the cancer detection rate at the UM because the radiologists were already highly alert in reading the study cases without CAD. Fortunately, the biopsy rate did not seem to increase substantially because many of the call back cases were found to be negative or benign. This can be seen from the decrease in the biopsy-to-callback ratio from 19.8% to 10.2%, as shown in Table 6.

It may also be noted that the call back rate in the community hospital where the prospective CAD study was conducted was only 6.5% without CAD. This is substantially lower than the call back rate without CAD at UM and GU, which is 10.4% and over 20%, respectively. The lower call back rate at the community hospital may reduce the sensitivity of the radiologists when CAD was not used. The gain in the radiologists' sensitivity by using the CAD system can be expected to be greater when the sensitivity of the radiologists without CAD is lower.

Table 3. The performance of the CADView detection system and its effects on radiologists' reading on the callback cases from the total of 2396 off-line screening cases at UM and GU. The 12 month follow up indicates a regular annual screening schedule and these cases are thus generally considered to be normal. FN=false negative; FU=follow up.

	Biopsy	Fine needle aspiration	6 month FU	12 month FU	Overall	Cancer
Call Backs for Mass						
Radiologist detection	20	12	47	197	276	4
Computer detection	16	10	36	141	203	3
Sensitivity of computer (%)	80%	83%	77%	72%	74%	75%
Call Backs for Calcs						
Radiologist detection	11	0	23	28	62	1
Computer detection	11	0	19	20	50	2
Sensitivity of computer (%)	100%		83%	71%	81%	200%
Call Backs for Mass and Calcs						
Radiologist detection	9	2	13	16	40	6
Computer detection	9	2	13	14	38	6
Sensitivity of computer (%)	100%	100%	100%	88%	95%	100%
Overall Call Backs						
Radiologist detection	40	14	83	241	378	11
Computer detection	36	12	68	175	291	11
Sensitivity of computer (%)	90%	86%	82%	73%	77%	92%*
Call Backs caused by CAD						
Mass	1	1	3	22	27	0
Microcalcifications	2	1	2	2	7	1
Computer False Negatives						
FN for Calcs	0	0	4	8	12	-1**
FN for Mass	4	2	11	56	73	1
FN for Mass and Calcs	0	0	0	2	2	0

Note:

*The total number of cancers in this patient cohort is 12. Both the computer and the radiologist missed one case but they were not the same case. The sensitivities of the computer and the radiologists were therefore both 92%

**The "-1" means that the lesion was a false-negative of the radiologist without CAD.

Table 4. Summary of the performance of the CADView system and the effects of CAD on radiologists' cancer detection. Note: 12 month FU = negative or benign finding.

Cases	UM	GU	Total
Recommended biopsy	88% (23/26)	93% (13/14)	90% (36/40)
Fine needle aspiration	71% (5/7)	100% (7/7)	86% (12/14)
Cancer	83% (5/6)	100% (6/6)	92% (11/12)
Mass	75% (142/190)	71% (61/86)	74% (203/276)
Microcalcification	80% (28/35)	81% (22/27)	81% (50/62)
Mass+Microcalcifications	97% (28/29)	91% (10/11)	95% (38/40)
Additional call backs caused by CAD	Total = 30 6 month FU = 5 12 month FU = 21	Total = 4 6 month FU = 0 12 month FU = 3	Total = 34 6 month FU = 5 12 month FU = 24
Additional biopsies caused by CAD	Total = 2 Benign = 2 Malgnant = 0	Total = 1 Benign = 0 Malgnant = 1	Total = 3 Benign = 2 Malgnant = 1
Additional fine needle aspiration caused by CAD	2 (Benign)	0	2 (Benign)
"False negative" by computer	Biopsy = 3 Fine needle asp = 2 6 month FU = 10 12 month FU = 41	Biopsy = 1 Fine needle asp = 0 3-6 month FU = 5 12 month FU = 25	Biopsy = 4 Fine needle asp = 2 6 month FU = 15 12 month FU = 66
Additional cancer found by CAD	**0**	**1**	**1**
Missed cancer by computer	**1**	**0**	**1**
Total cancer found by Radiologist alone	**6**	**5**	**11**
Total cancer found by Radiologist + CAD	**6**	**6**	**12**

Table 5. Summary of the ethnic compositions of the patient populations at the University of Michigan (UM) and the Georgetown University (GU). The statistics are based on the general mammographic patient populations, not from the particular patient cohort in this study

Ethnicity*	UM	GU
American Indian or Alaskan native (American Indian)	0.2%	
Asian or Pacific islander (Asian)	2.8%	1.9%
Black, not of Hispanic origin (African American)	7.0%	19.6%
Hispanic (Spanish Surname)	0.5%	
White, not of Hispanic origin (Caucasian)	83.5%	78.4%
Other or Unknown	6.0%	0.1%

*The ethnicity text is the label used by UM. The text in parentheses is the corresponding label used by GU.

Table 6. The average call back rate and biopsy rate. The "without CAD" rates are estimated off-line screening mammograms at UM from 1997-1999. The "with CAD" rates are estimated from the patient cohort in the study.

	Call Back Rate	Biopsy Rate for Call Back patients	Biopsy Rate for screened patients
Without CAD	10.4%	19.8%	2.1%
With CAD	15.3%	10.2%	1.6%

Table 7. The average call back rate and biopsy rate. The "without CAD" rates are estimated off-line screening mammograms at GU. The "with CAD" rates are estimated from the patient cohort in the study. The biopsy rate for screened patients is not available yet.

	Call Back Rate	Biopsy Rate for Call Back patients	Biopsy Rate for screened patients
Without CAD	22-27%	10%	-
With CAD	17.0%	11.3%	2.1%

(c) Image compression of mammograms using a CAD-guided wavelet compression method

For the second subproject, we collected a database of mammograms containing subtle mammographic lesions and digitized them with a high resolution LUMISYS laser scanner. We investigated the application of different image compression techniques to mammograms and selected the most promising method for the mammograms. The images were processed with the selected CAD-guided wavelet compression method, as shown in Figure 5, and two observer studies were conducted to evaluate the image quality of the compressed mammograms in comparison with the uncompressed mammograms. Detailed discussion of the image compression methods and the observer studies can be found in our progress report last year. We briefly summarize the results in the following.

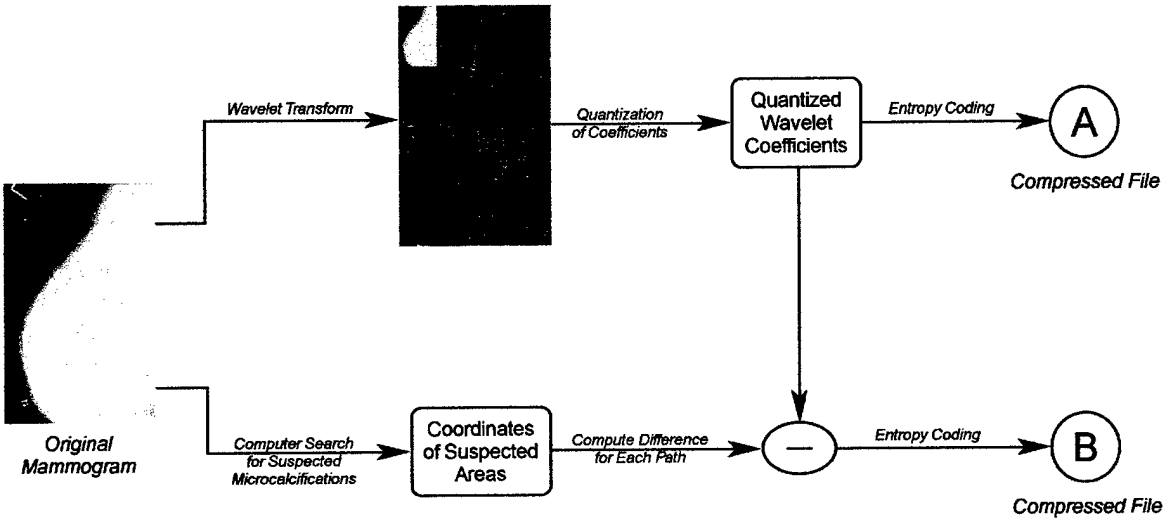


Figure 5: A CAD guided compression scheme based on integer wavelet decomposition.

(c.1) Observer experiments and results

(1). The First Observer Study

An experienced breast radiologist viewed a hundred sets of images with four different compression modes and to rate their subjective impressions on the relative quality of the images. Each set of images is a pair of original and one of three compression modes. The three compression modes are: (i) 0.3 bit/pixel date wavelet encoded in compressed file A (i.e., entire breast) with the residual data for lossless compression of suspected calcifications in file B (i.e., suspicious locations), (ii) 0.1 bit/pixel wavelet encoded in compressed file A with the residual data for lossless compression of suspected calcifications in file B, and (iii) 0.1 bit/pixel data wavelet encoded in file A only. Each set of decompressed and original images were randomly displayed on two monitors (right or left) as a pair. The reader was asked to rate image quality in terms of calcification observability, edge sharpness, overall image quality, and noise appearance for all images. If reader is in favor of one image for its specific feature, one of the two boxes (left and right) can be checked to indicate his/her preference.

The average compression ratios and computed mean-square-errors (MSE) between the original and decompression are shown in Table 8. We found that the CAD guided compression method received very small MSE improvement although it used a significant number of computer space (i.e., Bit rate = total number of bits used to encode the data / total number of pixels in the image) to preserve the full data accuracy of the suspected calcifications. This mainly is because that the suspected microcalcifications occupy very small area as compared to the whole breast region.

Table 8. Compression Ratios and Mean-Square-Errors of the Three Compression Modes in the First Observer Study. The conventional bit rate and area-equalized (AEQ) bit rate are defined as:

Mode	A	B	C
Procedure	0.3 bit/pixel + lossless for spots	0.1 bit/pixel + lossless for spots	0.1 bit/pixel
Average Bit Rate	0.43 bit/pixel	0.23 bit/pixel	0.1 bit/pixel
Compression Ratio	27:1	52:1	120:1
Mean Square Error	50.73 (36.81)	102.72 (62.48)	105.63 (63.97)

From the radiologist's qualitative measures in comparing the original and compressed image pair, we found that no difference could be observed between the original and decompressed images at a bit rate of 0.43 bit/pixel. In fact, it is interesting that the radiologist seemed slightly in favor of the appearances of microcalcifications and edges in the compressed mammograms. The radiologist identified 20% of the compressed images at 0.1 bit rate suffering from minor blurring artifacts and 6% of the compressed images possessing greater edge sharpness. Without using lossless compression for microcalcifications, the radiologist could identify 20% of the less sharp microcalcifications on the compressed mammograms at 0.1 bit rate. The radiologist also identified that 18% and 6% of the compressed images at 0.1 bit rate possess degraded overall image quality and higher image noise, respectively. Degradation of image quality in compressed images at 0.1 bit rate is highly associated with unsharpness of microcalcifications and edges. The image quality degradation at 0.1 bit rate is also correlated with the size of breast area. It is estimated that if the size of the breast takes more than one half of the entire mammogram, degradation in image quality and edge unsharpness would be observed by the radiologist.

We also studied the relationship between compression rate and breast area. For compression rates higher than or equal to 0.1bit/pixel and breast area less than or equal to 40%, no degradation can be identified relative to their original counterpart in overall image quality, overall noise pattern, and edge sharpness. For compression rates higher than or equal to 0.1bit/pixel and breast area less than or equal to 25%, no degradation can be identified as inferior microcalcifications. Therefore, we estimated that the threshold of area-equalized compression rate (AEQ bit rate = total number of bits used to encode the data / total number of pixels within the breast) for the background including edges is 0.25 bit/pixel (0.1bit/pixel divided by 40%) and the threshold of AEQ compression rate for the microcalcification is approximately 0.4 bit/pixel (0.1 bit/pixel divided by 25%).

(2). The Second Observer Study

In this experiment, we compared two different compression methods: (1) using an area-equalized compression rate at 0.25 bit/pixel with preservation of microcalcifications to compress and decompress the mammograms and (2) using an area-equalized compression rate at 0.4 bit/pixel to compress and decompress the mammograms.

Table 9. Qualitative measures by comparing the paired images in the Second Observer Study. (Compression Methods 1 and 2).

Category	Micro-calcifications	Edge Sharpness	Overall Image Quality	Overall Noise Pattern
Total	100	100	100	100
of which:				
In favor of the first method	55	8	0	0
In favor of the second method	0	0	0	0
No Difference	45	92	100	100

Table 10. Compression Ratios and Mean-Square-Errors of the Two Compression Methods in the Second Observer Study.

Category	The First Method (0.25 AEQ bit/pixel + Lossless spots)		The Second Method (0.40 AEQ bit/pixel)	
	Bit Rate (Bit/pixel) Mean (SD)	MSE	Bit Rate (Bit/pixel) Mean (SD)	MSE
All	0.149(0.05)	94	0.141(0.05)	55
Micro-calcifications:				
In favor of the first method	0.146(0.06)	94	0.135(0.05)	65
No Difference	0.152(0.04)	93	0.148(0.05)	53
Edge Sharpness:				
In favor of the first method	0.195(0.09)	92	0.159(0.08)	49
No Difference	0.145(0.05)	95	0.140(0.05)	55

The image display and rating method were similar to the first experiment. The results indicated that no image was rated better than its counterpart by the radiologist. However, the radiologist favored microcalcifications of 55 cases that were compressed and decompressed through the first method (i.e., 0.25 AEQ bit/pixel with preservation of microcalcifications). The radiologist also favored edge characteristics of 8 cases that were compressed and decompressed through the first method. No image was identified as a higher quality image over its counterpart by the radiologist in terms of overall image quality and overall noise pattern. No image compressed by the second compression method (i.e., 0.4 AEQ bit/pixel) was in favor by the radiologists. Table 9 shows the summary results of the observer study. Table 10 shows the bit rate used and the average MSE of the decompression images for each category. Note that the bit rate of the first method includes the wavelet compressed data and the lossless

compressed data of the suspected calcification areas. Although the first compression method spent less computer space to code the overall breast area than the second method did, the first compression method used more computer space to preserve the 10x10 pixels area of all suspected microcalfications, the effective compression bit rates were approximately the same for both methods. We found that the first method produced higher quality for clinically significant features. Although the overall MSEs produced by the first compression method were markedly worse than those produced by the second method, the degradation was not observable by the breast radiologist, indicating that the first compression method generates error-free suspected calcifications that were appreciable and in favor by the radiologist.

(c.2) Conclusions and discussion of the compression studies

In this study, we used conventional compression testing methods with and without the CAD guidance to evaluate the decompressed images. We were able to identify the threshold of area-equalized bit rate for overall breast area and the threshold for encoding quality microcalcifications. We used these two thresholds to compress the mammograms. All four image-quality categories of all compression images were deemed more than adequate. However, the radiologist favored fully preserved microcalcifications on 55 out of 100 images (55% of the test database). This study also showed that neither edge nor overall image quality degradation could be observed by the radiologist using area-equalized bit-rate of 0.25 AEQ bit/pixel and 0.4 AEQ bit/pixel. Therefore, CAD can be used to guide image processing method to preserve or enhance clinically significant features. Our results clearly indicate that the CAD guided compression method with adequate bit rate will fully preserve the quality of microcalcifications and suspected microcalcifications without sacrificing the edge sharpness and overall image quality. The radiologist could not recognize any blocky artifact between lossless and lossy boundaries even on magnified view with contrast adjustable display.

(d) Improvement of microcalcification detection by optimization of the neural network for pattern recognition

The computer program that we developed to automatically detect microcalcification clusters on digitized mammograms has four stages: signal-to-noise ratio enhancement of the mammogram, prescreening for suspicious locations of microcalcifications, rule-based false positive (FP) reduction, pattern recognition with an artificial convolution neural network (CNN), and regional clustering for identifying suspicious clustered microcalcifications. With the support in part from this grant, we evaluated the effectiveness of optimal neural network architecture selection on the performance this microcalcification detection CAD system.

In this study, we evaluated the effectiveness of using an automated optimization technique in selecting the optimal CNN architecture in comparison with the previously manual optimization. Three automated optimization methods were compared: steepest gradient descent (SD), genetic algorithm (GA) and simulated annealing (SA), for their efficiency in reaching the optimum in the multidimensional parameter space of the CNN architectures. It was found that both the GA and SA could reach the global optimum whereas the SD was often trapped in local optima. The SA with the Boltzmann annealing schedule was the most efficient for this optimization problem. We conducted a study to evaluate the improvement in the accuracy of the microcalcification detection system by the optimized CNN in comparison to that with the manually optimized CNN [2] (enclosed in Appendix). For this evaluation, we used a three-stage approach: training, validation, and testing. Three independent data sets were used in the three stages. The test data set for the testing stage included 472 mammograms selected from the University of South Florida public digital mammography database and contained a total of 253 biopsy-

proven malignant clusters. Free-response receiver operating characteristic (FROC) analysis was used to evaluate the tradeoff between detection sensitivity and the number of FPs per image. At an FP rate of 0.7 per image, the microcalcification detection program achieved a film-based sensitivity of 84.6% with the optimized CNN, in comparison with 77.2% with the manually selected CNN. If clusters having images in both craniocaudal (CC) and mediolateral oblique (MLO) views were analyzed and a cluster was considered to be detected when it was detected in one or both views, at 0.7 FPs/image, the sensitivity was 93.3% with the optimized CNN and 87.0% with the manually selected CNN. This study indicated that an optimized CNN can effectively reduce FPs and improve the detection accuracy of the computer-aided detection system.

(e) Evalution of CAD mass detection algorithm with independent cases

In this study, we analyzed the performance of our CAD algorithm for detection of breast masses on independent clinical mammograms [3] (enclosed in Appendix). A digitized mammogram is processed with an adaptive enhancement filter followed by a local border refinement stage. Features are then extracted from each detected structure and used to identify potential masses. We evaluated the performance of the algorithm on independent cases obtained from 263 patients from two institutions. The CAD marker rate was estimated by applying the algorithm to 503 normal films. The computer detected a malignant mass in 83% (130/156) of the malignant cases at a marker rate of 1.0 marks per film. The detection accuracy for benign lesions was lower than that for malignant masses. FROC performance curves were obtained and the tradeoff between detection sensitivity and the number of CAD marks was analyzed. A performance comparison between cases collected at the two different institutions was also included.

In an additional study, we evaluated the performance of the mass detection program on prior mammograms in which the mass was not sent for biopsy in that year. These patients were found to have a biopsy-proven malignant mass on their mammograms in a future year. A data set of 38 patients with mammograms from 1 to 4 years prior to biopsy was collected. The computer detected the malignant mass in 48% (13/27) of the prior cases at a marker rate of 1.0 marks per film. This preliminary result indicates that the mass detection program can detect a substantial fraction of the malignant masses in a prior year, demonstrating the potential that the CAD system may be able to alert radiologists to suspicious masses and lead to earlier breast cancer detection.

(f) Improvement of computerized mass detection on mammograms: fusion of two-view information

Recent clinical studies have proved that CAD systems are helpful for improving lesion detection by radiologists in mammography. However, these systems would be more useful if the FP rate is further reduced. Current CAD systems generally detect and characterize suspicious abnormal structures in individual mammographic images. Clinical experiences by radiologists indicate that screening with two mammographic views improves the detection accuracy of abnormalities in the breast. It is expected that fusion of information from different mammographic views will improve the performance of CAD systems. With the support in part from this grant, we are developing a two-view matching method that utilizes the geometric locations, and morphological and textural features to correlate objects detected in two different views using a prescreening program [4] (enclosed in Appendix). First, a geometrical model is used to predict the search region for an object in a second view from its location in the first view. The distance between the object and the nipple is used to define the search area. After pairing the objects in two views,

textural and morphological characteristics of the paired objects are merged and similarity measures are defined. Linear discriminant analysis is then employed to classify each object pair as a true or false mass pair. The resulting object correspondence score is combined with its one-view detection score using a fusion scheme. The fusion information was found to improve the lesion detectability and reduce the number of FPs. In a preliminary study, we used a data set of 169 pairs of cranio-caudal (CC) and mediolateral oblique (MLO) view mammograms. For the detection of malignant masses on current mammograms, the film-based detection sensitivity was found to improve from 62% with a one-view detection scheme to 73% with the new two-view scheme, at a false-positive rate of 1 FP/image. The corresponding cased-based detection sensitivity improved from 77% to 91%.

(g) Optimization of wavelet decomposition for image compression and feature preservation

As a step to the subproject of CAD-guided data compression for mammography, we investigated different data compression techniques for mammograms and other medical images. We have developed a neural network system that can search for an optimal wavelet kernel for a specific image processing task. In this study, a linear convolution neural network was employed to obtain a wavelet that minimizes errors and maximizes compression efficiency for an image or a defined image pattern such as microcalcifications on mammograms. We have used this method to evaluate the performance of tap-4 wavelets on mammograms, computed tomograms (CTs), magnetic resonance images (MRIs), and the Lena images. We found that Daubechies wavelet or those wavelets possessing similar filtering characteristics produces a high compression efficiency with the smallest mean-square-error. However, Haar wavelet produces the best results on sharp edges and low-noise smooth areas. We also found that a special wavelet, whose low-pass filter coefficients are (0.32252136, 0.85258927, 0.38458542, -0.14548269), can greatly preserve the microcalcification features in peak signal-to-noise ratio, contrast, and figure of merit during a course of compression. The technical details of this study can be referred to the paper by Lo et al.[5] (enclosed in Appendix).

(h) Predictive decomposition as a framework in dyadic transforms - A unified theory for wavelet and subband decompositions

In this research, we found that a generalized decomposition method, Haar + Prediction + Composite (H+PC), based on Haar transform has been derived. This general form can exactly describe dyadic transforms. Another general form Biorthogonal + Prediction + Composite (B+PC), which is a subset of the doublet system, based on the binomial filter can describe triplet-type decompositions including whole point symmetric biorthogonal transformations. Both systems can be unified by the delta function basis decomposition system, Delta + Prediction + Composite (D+PC). We also found that these three bases and their expansions using predictive approximation form the dyadic decomposition family. Wavelet and integer wavelet based decomposition methods can also be included in this unified framework. This framework clearly bridges the relationship among various types of dyadic transforms. To confirm this theory, we perform a computational exercise and found that almost all dyadic decompositions can be directly computed from their basis. A paper based on this research has been submitted to IEEE Signal Processing for review. The technical details of this study can be referred to the paper by Lo et al.[6] (enclosed in the Appendix).

(6) Key Research Accomplishments

- Developed the graphical user interface for the CADView system, and implement the mass detection and microcalcification detection programs in the system for automatic image processing of the digitized mammogram.

- Installed the CADView workstations at the Breast Imaging clinics of University of Michigan Health System and at the Georgetown University Medical Center, and conduct the pilot clinical study.

- Completed the pilot clinical study at the UM and GU mammography screening sites with 2,400 patient mammograms read with CAD.

- Analyzed the data and found that CAD increased the number of cancer detection from 11 to 12.

- Analyzed the effects of CAD on radiologists' reading based on the data collected from the pilot clinical study.

- Analyzed the performance of the CADView system in the patient population, compare the performances at the two sites and those in laboratory tests.

- Continued improvement of the mass and microcalcification detection programs, independent of the versions implemented in the CADView system, which were fixed throughout the pilot study.

- Investigated various image compression approaches for mammography and selected a wavelet compression method for the CAD-guided compression of mammograms

- Conducted two observer performance studies to compare microcalcification detection on mammograms without compression, with conventional compression, and with CAD-guided compression

- Analyzed the results of the observer performance studies and estimated the best compression rate for the CAD-guided compression method

- Published a number of peer-reviewed papers in the various topics related to this project.

(7) Reportable Outcomes

Publications related to the development of the CAD system and the evaluation of the effects of the CAD system:

Peer-Reviewed Journal Articles

1. Sahiner B, Chan HP, Petrick N, Helvie MA, Goodsitt MM. Development of a high-sensitivity classifier for computer-aided diagnosis: Application to classification of malignant and benign masses. Physics in Medicine and Biology 1998; 43: 2853-2871.

2. Chan HP, Sahiner B, Lam KL, Petrick N, Helvie MA, Goodsitt MM, Adler DD. Computerized analysis of mammographic microcalcifications in morphological and texture feature spaces. Medical Physics 1998; 25: 2007-2019.

3. Petrick N, Chan HP, Sahiner B, Helvie MA, Goodsitt MM. Combined adaptive enhancement and object-based region growing for automated detection of masses on mammograms. Medical Physics 1999; 26: 1642-1654.

4. Chan HP, Sahiner B, Helvie MA, Petrick N, Roubidoux MA, Wilson TE, Adler DD, Paramagul C, Newman JS, Sanjay-Gopal S. Improvement of radiologists' characterization of mammographic masses by computer-aided diagnosis: an ROC Study. Radiology 1999; 212: 817-827.

5. Chan HP, Sahiner B, Wagner RF, Petrick N. Classifier design for computer-aided diagnosis: Effects of finite sample size on the mean performance of classical and neural network classifiers. Medical Physics 1999; 26: 2654-2668.

6. Sanjay-Gopal S, Chan HP, Wilson TE, Helvie MA, Petrick N, Sahiner B. A regional registration technique for automated interval change analysis of breast lesions on mammograms. Medical Physics 1999; 26: 2669-2679.

7. Hadjiiski LM, Sahiner B, Chan HP, Petrick N, Helvie MA. Classification of malignant and benign masses based on hybrid ART2LDA approach. IEEE Transactions on Medical Imaging 1999; 18: 1178-1187.

8. Chan HP, Helvie MA, Petrick N, Sahiner B, Adler DD, Paramagul C, Roubidoux MA, Blane CE, Joynt LK, Wilson TE, Hadjiishi LM, Goodsitt MM. Digital mammogramphy: observer performance study of effects of pixel size on radiologists' characterization of malignant and benign microcalcifications. Academic Radiology 2001; 8: 454-466.

9. Hadjiiski LM, Chan HP, Sahiner B, Petrick N, Helvie MA. Automated registration of breast lesions in temporal pairs of mammograms for interval change analysis – local affine transformation for improved localization. Medical Physics 2001; 28: 1070-1079.

Articles Accepted for Publication:

1. Hadjiiski L, Sahiner B, Chan HP, Petrick N, Helvie MA, Gurcan MN. Analysis of Temporal Change of Mammographic Features: Computer-Aided Classification of Malignant and Benign Breast Masses. Medical Physics. May 2001.

2. Petrick N, Sahiner B, Chan HP, Helvie MA, Paquerault S, Hadjiiski LM. Breast cancer detection: Evaluation of a CAD mass detection algorithm with independent cases. Radiology. September 2001.

3. Paquerault S, Petrick N, Chan HP, Sahiner B, Helvie MA. Improvement of Computerized Mass Detection on Mammograms: Fusion of Two-View Information. Medical Physics. September 2001.

Articles Submitted for Publication:

1. Gurcan MN, Chan HP, Sahiner B, Hadjiiski LM, Petrick N, Helvie MA. Optimal neural network architecture selection: Improvement in computerized detection of microcalcifications. Academic Radiology. October, 2001.

Conference Proceedings

1. Chan HP, Sahiner B, Wagner RF, Petrick N, Mossoba JT. Effects of sample size on classifier design: Quadratic and neural network classifiers. Proc. SPIE 1997; 3034: 1102-1103.

2. Sahiner B, Chan HP, Petrick N, Goodsitt MM, Helvie MA. Characterization of masses on mammograms: Significance of the use of the rubber-band straightening transform. Proc. SPIE 1997; 3034: 491-500.

3. Petrick N, Chan HP, Sahiner B, Helvie MA, Goodsitt MM. Unitary ranking for automated detection mammographic masses. Proc. SPIE 1997; 3034: 522-525.

4. Sahiner B, Chan HP, Petrick N, Sanjay-Gopal S, Goodsitt MM. Neural network design for optimization of the partial area under the receiver operating characteristic curve. Proc. of the 1997 International Conference on Neural Networks (ICNN'97) 1997; 4: 2468-2471.

5. Sanjay-Gopal S, Sahiner B, Chan HP, Petrick N. Neural network based segmentation using *a priori* image models. Proc. of the 1997 International Conference on Neural Networks (ICNN'97) 1997; 4: 2455-2459.

6. Sahiner B, LeCarpentier GL, Chan HP, Petrick N, Goodsitt MM, Sanjay-Gopal S, Carson PL Computerized characterization of breast masses using three-dimensional ultrasound images. Proc. SPIE 1998; 3338: 301-312.

7. Sanjay-Gopal S, Chan HP, Petrick N, Wilson T, Sahiner B, Helvie MA, Goodsitt MM. A regional mammogram registration technique for automated analysis of interval changes of breast lesions. Proc. SPIE 1998; 3338: 118-129.

8. Chan HP, Sahner B, Wagner RF, Petrick N. Effects of sample size on classifier design for computer-aided diagnosis. Proc. SPIE 1998; 3338: 845-858.

9. Chan HP, Helvie MA, Petrick N, Sahiner B, Roubidoux MA, Wilson TE, Joynt LK, Hadjiiski LM, Paramagul C, Adler DD, Goodsitt MM. Digital Mammography: observer performance study of the effects of pixel size on radiologists' characterization of malignant and benign microcalcifications. Proc. SPIE 1999; 3659: 394-397.

10. Sahiner B, Chan HP, Petrick N, Wagner RF, Hadjiiski LM. Stepwise linear discriminant analysis in computer-aided diagnosis: the effect of finite sample size. Proc. SPIE 1999; 3661: 499-510.

11. Hadjiiski LM, Sahiner B, Chan HP, Petrick N, Helvie MA. Hybrid unsupervised-supervised approach for computerized classification of malignant and benign masses on mammograms. Proc. SPIE 1999; 3661: 464-473.

12. Petrick N, Chan HP, Sahiner B, Helvie MA, Paquerault S. Evaluation of an automated computer-aided diagnosis system for the detection of masses on prior mammograms. Proc. SPIE 3979. 2000: 967-973.

13. Hadjiiski LM, Chan HP, Sahiner B, Petrick N, Helvie MA, Paquerault S, Zhou C. Interval change analysis in temporal pairs of mammograms using a local affine transformation. Proc. SPIE 3979. 2000: 847-853.

14. Petrick N, Sahiner B, Chan HP, Helvie, MA, Paquerault S. Preclinical evaluation of a CAD algorithm for early detection of breast cancer. Presented at the IWDM-2000. Toronto, Canada. June 11-14, 2000. In: Digital Mammography IWDM 2000: 5th International Workshop on Digital Mammography. Ed. Yaffe MJ. (Medical Physics Publishing, Madison, WI) 2001: 328-333.

15. Gurcan MN, Sahiner B, Chan H-P, Hadjiiski L, Petrick N, "Optimal Selection of Neural Network Architecture for CAD using Simulated Annealing," Proceedings of the 22nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2000, vol. 4, pp. 3052–3055, July 23-28, 2000, Chicago, Illinois.

16. Sahiner B, Petrick N, Chan HP, Paquerault S, Helvie MA, Hadjiiski LM. Recognition of lesion correspondence on two mammographic views - a new method of false-positive reduction for computerized mass detection. Proc SPIE 4322, 2001: 649-655.

17. Gurcan M, Petrick N, Sahiner B, Chan HP, Cascade P, Kazerooni E, Hadjiiski LM. Computerized lung nodule detection on thoracic CT images: combined rule-based and statistical classifier for false positive reduction. Proc SPIE 4322, 2001: 686-692.

18. Paquerault S, Petrick N, Chan HP, Sahiner B, Dolney AY. Improvement of mammographic lesion detection by fusion of information from different views. Proc SPIE 4322, 2001: 1883-1889.

19. Hadjiiski LM, Sahiner B, Chan HP, Petrick N, Helvie MA, Gurcan M. Analysis of temporal change of mammographic features for computer-aided characterization of malignant and benign masses. Proc SPIE 4322, 2001: 661-666.

Scientific Exhibits

1. Chan HP, Petrick N, Sanjay-Gopal S, Wilson TE, Roubidoux MA, Adler DD, Sahiner B, Helvie MA, Paramagul C, Newman JS. Observer performance studies of the effects of computer-aided diagnosis (CAD) on radiologists' classification of malignant and benign masses on mammograms. Scientific Exhibit at the 83rd Scientific Assembly and Annual Meeting of the Radiological Society of North America, Nov 30-Dec 5, 1997, Chicago, Illinois. Radiology 1997; 205(P): 655.

2. Chan HP, Petrick N, Hadjiiski LM, Wilson TE, Paramagul C, Adler DD, Helvie MA, Sahiner B, Roubidoux MA, Sanjay-Gopal S, Joynt LK. ROC study of the effects of pixel size on radiologists' classification of microcalcifications on digitized mammograms. Scientific Exhibit at the 84th Scientific Assembly and Annual Meeting of the Radiological Society of North America, Nov 29-Dec 4, 1998, Chicago, Illinois. Radiology 1998; 209(P): 518.

3. Chan HP, Petrick N, Sahiner B, Hadjiiski LM, Helvie MA, Zhou C, Paquerault S. Computer-aided breast cancer diagnosis. Exhibited at the Biomedical Imaging Symposium: Visualizing the Future of Biology and Medicine. National Institutes of Health Bioengineering Consortium (BECON), Bethesda, Maryland. June 25-26, 1999.

Abstracts and Presentation

1. Sahiner B, Chan HP, Chenevert T, Petrick N, Helvie MA, Sanjay-Gopal S. Computer-aided characterization of malignant and benign lesions on breast MR images using texture features. Presented at the 83rd Scientific Assembly and Annual Meeting of the Radiological Society of North America, Nov 30-Dec 5, 1997, Chicago, Illinois. Radiology 1997; 205(P): 520.

2. Petrick N, Chan HP, Sahiner B, Helvie MA, Sanjay-Gopal S, Goodsitt MM. Computer-Aided Detection of Breast Masses: Evaluation of a Fuzzy Morphological Classifier. Presented at the 83rd Scientific Assembly and Annual Meeting of the Radiological Society of North America, Nov 30-Dec 5, 1997, Chicago, Illinois. Radiology 1997; 205(P): 216.

3. Sanjay-Gopal S, Chan HP, Sahiner B, Petrick N, Wilson TE, Helvie MA. Evaluation of interval change in mammographic features for computerized classification of malignant and benign masses. Presented at the 83rd Scientific Assembly and Annual Meeting of the Radiological Society of North America, Nov 30-Dec 5, 1997, Chicago, Illinois. Radiology 1997; 205(P): 216.

4. Sanjay-Gopal S, Sahiner B, Petrick N, Chan HP, Helvie MA, Wilson TE. Evaluation of automated methods for the segmentation of mass boundaries on mammograms for computer aided diagnosis (CAD) applications. Presented at the 83rd Scientific Assembly and Annual Meeting of the Radiological Society of North America, Nov 30-Dec 5, 1997, Chicago, Illinois. Radiology 1997; 205(P): 215.

5. Chan HP, Sahiner B, Helvie MA, Paramagul C, Newman JS, Sanjay-Gopal S, Petrick N, Adler DD, Roubidoux MA, Wilson TE. Effects of computer-aided diagnosis (CAD) on radiologists' classification of malignant and benign masses on mammograms: an ROC study. Presented at the

83rd Scientific Assembly and Annual Meeting of the Radiological Society of North America, Nov 30-Dec 5, 1997, Chicago, Illinois. Radiology 1997; 205(P): 275.

6. Sanjay-Gopal S, Chan HP, Petrick N, Wilson T, Sahiner B, Helvie MA, Goodsitt MM. A regional mammogram registration technique for automated analysis of interval changes of breast lesions. Presented at the SPIE International Symposium on Medical Imaging, San Diego, CA, Feb. 21-27, 1998.

7. Hadjiiski LM, Sahiner B, Chan HP, Petrick N, Helvie MA, Goodsitt MM. Characterization of malignant and benign masses on mammograms based on a hierarchical classifier. Presented at the 84th Scientific Assembly and Annual Meeting of the Radiological Society of North America, December 1998, Chicago, Illinois. Radiology 1998; 209(P): 354.

8. Sahiner B, Chan HP, Helvie MA, Wilson TE, Sanjay-Gopal S, Petrick N. Computerized classification of mammographic masses using morphological features. Presented at the 84th Scientific Assembly and Annual Meeting of the Radiological Society of North America, December 1998, Chicago, Illinois. Radiology 1998; 209(P): 353.

9. Petrick N, Chan HP, Sahiner B, Helvie MA, Hadjiiski LM, Goodsitt MM. Comparison of local clustering and gradient-based region growing segmentation for the automated detection of mass on digitized mammograms. Presented at the 84th Scientific Assembly and Annual Meeting of the Radiological Society of North America, December 1998, Chicago, Illinois. Radiology 1998; 209(P): 353-354.

10. Chan HP, Helvie MA, Petrick N, Sahiner B, Roubidoux MA, Wilson TE, Joynt LK, Hadjiiski LM, Paramagul C, Adler DD, Goodsitt MM. Digital Mammography: observer performance study of the effects of pixel size on radiologists' characterization of malignant and benign microcalcifications. Presented at the SPIE International Symposium on Medical Imaging, San Diego, CA, February 20-26, 1999.

11. Hadjiiski LM, Sahiner B, Chan HP, Petrick N, Helvie MA. Hybrid unsupervised-supervised approach for computerized classification of malignant and benign masses on mammograms. Presented at the SPIE International Symposium on Medical Imaging, San Diego, CA, February 20-26, 1999.

12. Chan HP, Hadjiiski LM, Petrick N, Helvie MA, Roubidoux MA, Sahiner B. Performance evaluation of an automated microcalcification detection system. Presented at the 41st Annual Meeting of the American Association of Physicists in Medicine. Nashville, Tennessee, July 25-29, 1999. Medical Physics 1999; 26: 1080.

13. Chan HP, Sahiner B, Helvie MA, Petrick N, Hadjiiski LM, Roubidoux MA. Computer-aided breast cancer diagnosis: Comparison of computerized classification with radiologists' performance. Presented at the 85th Scientific Assembly and Annual Meeting of the Radiological Society of North America, Nov. 28-Dec. 3, 1999, Chicago, Illinois. Radiology 1999; 213(P): 322-323.

14. Hadjiiski LM, Chan HP, Sahiner B, Petrick N, Helvie MA, Sanjay-Gopal S. Automated identification of breast lesions in temporal pairs of mammograms for interval change analysis. Presented at the 85th Scientific Assembly and Annual Meeting of the Radiological Society of North America, Nov. 28-Dec. 3, 1999, Chicago, Illinois. Radiology 1999; 213(P): 229-230.

15. Petrick N, Chan HP, Sahiner B, Helvie MA, Paquerault S. Evaluation of an automated computer-aided diagnosis system for the detection of masses on prior mammograms. Presented at the SPIE International Symposium on Medical Imaging, San Diego, CA, February 12-18, 2000. (Honorable Mention Citation)

16. Hadjiiski LM, Chan HP, Sahiner B, Petrick N, Helvie MA, Paquerault S, Zhou C. Interval change analysis in temporal pairs of mammograms using a local affine transformation. Presented at the SPIE International Symposium on Medical Imaging, San Diego, CA, February 12-18, 2000.

17. Paquerault S, Chan HP, Sahiner B, Petrick N, Hadjiiski L, Gurcan M-N, Zhou C, Helvie M. Prediction of object location in different mammographic views using geometrical models. Presented at The 5th International Workshop on Digital Mammography. IWDM-2000. Toronto, Canada. June 11-14, 2000.

18. Petrick N, Sahiner B, Chan HP, Helvie, MA, Paquerault S. Preclinical evaluation of a CAD algorithm for early detection of breast cancer. Presented at The 5th International Workshop on Digital Mammography. IWDM-2000. Toronto, Canada. June 11-14, 2000.

19. Chan HP, Hadjiiski L, Petrick N, Helvie MA, Sahiner B, Paramagul C, Gurcan MN, Lo SCB., Freedman MT, Dorfman DD, Berbaum KS. Pilot clinical study of a computer-aided diagnosis workstation for mammography. Presented at the Era of Hope Meeting, U. S. Army Medical Research and Materiel Command, Department of Defense, Breast Cancer Research Program, Atlanta, Georgia, June 8-12, 2000.

20. Gurcan MN, Sahiner B, Chan HP, Hadjiiski L, Petrick N. Optimal selection of neural network architecture for CAD using simulated annealing. Presented at the Chicago 2000-World Congress on Medical Physics and Biomedical Engineering. Chicago, Illinois, July 23-28, 2000.

21. Hadjiiski L, Petrick N, Chan HP, Sahiner B, Helvie MA, Zhou C, Gurcan MN, Paquerault S. Regional registration of masses on current and prior mammograms using DWCE segmentation. Presented at the Chicago 2000-World Congress on Medical Physics and Biomedical Engineering. Chicago, Illinois, July 23-28, 2000.

22. Chan HP, Sahiner B, Hadjiiski LM, Petrick N, Helvie MA, Goodsitt MM. Computer-aided breast cancer diagnosis: Effects of pixel size on computerized classification of microcalcifications in comparison with radiologists' performance. Presented at the 86th Scientific Assembly and Annual Meeting of the Radiological Society of North America, Nov. 26-Dec. 1, 2000, Chicago, Illinois. Radiology 2000; 217(P): 401.

23. Hadjiiski LM, Sahiner B, Chan HP, Petrick N, Helvie MA, Gurcan MN. Computer-aided classification of malignant and benign breast masses by analysis of interval change of features in temporal pairs of mammograms. Presented at the 86th Scientific Assembly and Annual Meeting of the Radiological Society of North America, Nov. 26-Dec. 1, 2000, Chicago, Illinois. Radiology 2000; 217(P): 435.

24. Gurcan MN, Sahiner B, Chan HP, Hadjiiski LM, Petrick N. Selection of an optimal neural network architecture for computer-aided diagnosis - comparison of automated optimization techniques.

Presented at the 86th Scientific Assembly and Annual Meeting of the Radiological Society of North America, Nov. 26-Dec. 1, 2000, Chicago, Illinois. Radiology 2000; 217(P): 436.

25. Paquerault S, Petrick N, Chan HP, Sahiner B, Helvie MA. Computer-Aided Breast Cancer Diagnosis: Fusion of Information from Two Mammographic Views. Presented at the Univ. of Michigan Cancer Research Symposium, December 8, 2000. Ann Arbor, Michigan.

26. Sahiner B, Petrick N, Chan HP, Paquerault S, Helvie MA, Hadjiiski LM. Recognition of lesion correspondence on two mammographic views - a new method of false-positive reduction for computerized mass detection. Presented at the SPIE International Symposium on Medical Imaging, San Diego, CA, February, 2001.

27. Paquerault S, Petrick N, Chan HP, Sahiner B, Dolney AY. Improvement of mammographic lesion detection by fusion of information from different views. Presented at the SPIE International Symposium on Medical Imaging, San Diego, CA, February, 2001.

28. Hadjiiski LM, Sahiner B, Chan HP, Petrick N, Helvie MA, Gurcan M. Analysis of temporal change of mammographic features for computer-aided characterization of malignant and benign masses. Presented at the SPIE International Symposium on Medical Imaging, San Diego, CA, February, 2001.

29. Gurcan MN, Chan HP, Sahiner B, Hadjiiski LM, Petrick N, Helvie MA. Improvement of computerized detection of microcalcifications using a convolution neural network architecture selected by an automated optimization algorithm. Accepted for presentation at the Medical Image Perception Conference IX. Airlie Conference Center, Warrenton, VA, September 20-23, 2001.

30. Sahiner B, Petrick N, Chan HP, Paquerault S, Helvie MA, Hadjiiski LM. A new two-view correspondence approach to computerized mass detection on mammograms - Performance on an independent data set. Accepted for presentation at the 87th Scientific Assembly and Annual Meeting of the Radiological Society of North America, Chicago, IL, November 25-30, 2001.

31. Hadjiiski LM, Chan HP, Petrick N, Sahiner B, Gurcan M, Helvie MA, Paramagul C, Roubidoux MA. Computerized regional registration of corresponding microcalcification clusters on temporal pairs of mammograms for interval change analysis. . Accepted for presentation at the 87th Scientific Assembly and Annual Meeting of the Radiological Society of North America, Chicago, IL, November 25-30, 2001.

32. Petrick N, Chan HP, Sahiner B, Helvie MA, Hadjiiski LM. Evaluation of CAD mass detection with mammograms containing preoperative and prior cancers. Accepted for presentation at the 87th Scientific Assembly and Annual Meeting of the Radiological Society of North America, Chicago, IL, November 25-30, 2001.

33. Petrick N, Chan HP, Sahiner B, Helvie MA, Hadjiiski LM. Automated mass detection system for CAD in mammography: Algorithm design and performance on independent clinical cases from two institutions. Accepted for poster presentation at the 87th Scientific Assembly and Annual Meeting of the Radiological Society of North America, Chicago, IL, November 25-30, 2001.

34. Gurcan MN, Chan HP, Sahiner B, Hadjiiski LM, Petrick N, Helvie MA. Optimal neural network architecture selection: Effects on computer-aided detection of mammographic microcalcifications. Accepted for presentation at the SPIE International Symposium on Medical Imaging, San Diego, CA, February, 2002.

35. Sahiner B, Gurcan MN, Chan HP, Hadjiiski LM, Petrick N, Helvie MA. The use of joint two-view information for computerized lesion detection on mammograms: Improvement of microcalcification detection accuracy. Accepted for presentation at the SPIE International Symposium on Medical Imaging, San Diego, CA, February, 2002.

36. Hadjiiski LM, Chan HP, Gurcan MN, Sahiner B, Petrick N, Helvie MA, Roubidoux MA. Computer-aided characterization of malignant and benign microcalcification clusters based on the analysis of temporal change of mammographic features. Accepted for presentation at the SPIE International Symposium on Medical Imaging, San Diego, CA, February, 2002.

Publications related to the development of the CAD-guided data compression method and the evaluation of the image quality by observer studies:

Peer-Reviewed Journal Articles

1. Lo SCB, Lin JS, Freedman MT, and Mun SK, Application Of Artificial Neural Networks To Medical Image Pattern Recognition: Detection of Clustered Microcalcifications on Mammograms and Lung Cancer on Chest Radiographs," J. of VLSI in Sign. Proc., Vol. 18, 1998, pp. 263-274.

2. Li H, Wang Y, Liu KJR, Lo SCB, and Freedman MT, ``Computerized Radiographic Mass Detection -- Part I: Lesion Site Selection by Morphological Enhancement and Contextual Segmentation," IEEE Trans. on Medical Imaging, Vol. 20, No. 4, April. 2001, pp. 289-301.

3. Li H, Wang Y, Liu KJR, Lo SCB, and Freedman MT, ``Computerized Radiographic Mass Detection -- Part II: Decision Support by Featured Database Visualization and Modular Neural Networks," IEEE Trans. on Medical Imaging, Vol. 20, No. 4, April. 2001, pp. 302-313.

Articles Submitted for Publication

1. Lo SCB, Li H., Hasegawa A, Wang Y, and Freedman MT, "A Multiple Circular Paths Convolution Neural Network System for Detection of Mammographic Masses," (Submitted to) IEEE Trans. on Medical Imaging, 2001.

2. Lo SCB, Li H, Freedman MT, and Mun SK, "Optimization of Wavelet Decomposition for Image Compression and Feature Preservation, (Submitted to) IEEE Trans. Med. Imaging.

3. Lo SCB, Xuan J, Li H, "Predictive Decomposition as a Framework in Dyadic Transforms," (Submitted to) IEEE Trans. Signal Processing.

4. Lo SCB and MT Freedman "Wavelet-Based Artificial Convolution Neural Network For Image Pattern Recognition: Applications to Microcalcification Detection on Mammograms," (Submitted to) IEEE Trans. Med. Imaging.

Conference Proceedings

1. Li H, Liu KJ, Wang Y, and Lo SCB, "Nonlinear filtering enhancement and histogram modeling segmentation of masses for digital mammograms", Proc. 18th Annual Int. Conf. of IEEE EMBS, 145, 1996.

2. Li H, Lo SCB, Wang Y, Freedman MT, and Mun SK, "Mammographic Mass Detection by Stochastic Modeling and a Mutli-Modular Neural Network, SPIE Proceedings on Medical Imaging 1997, vol. 3034, pp. 480-490.

3. Lo SC, Xuan J, Li H, Wang Y, Freedman MT, and Mun SK, "Dyadic Decomposition: A Unified Perspective on Predictive, Subband, and Wavelet Transforms", SPIE Proceedings on Medical Imaging, 1997, vol. 3031, pp. 286-301.

4. Freedman MT, Lo SCB, Artz DS, and Mun SK, "Classification of False-Positive Findings on Computer-Aided Detection of Breast Microcalcifications," SPIE Proceedings on Medical Imaging 1997, vol. 3034, pp.853-859.

5. Lo SCB, Li H, Hasegawa A, Wang YJ, Freedman MT, Mun SK, "Detection of Mammographic Masses Using Sector Features with a Multiple Circular Path Neural Network," SPIE Proceedings on Medical Imaging 1998, vol. 3338, pp. 1205-1214.

6. Lo SCB, Delegacz A, Freedman MT, Chan HP, Dorfman DD, and Berbarn K, "Evaluation of Digitized Mammograms Compressed by an Optimized Wavelet Technique and Computer-Aided System," Presented at the Era of Hope Meeting, U. S. Army Medical Research and Materiel Command, Department of Defense, Breast Cancer Research Program, Atlanta, Georgia, June 8-12, 2000.

7. Lo SCB, Makariou E, Delegacz A, Chan HP, Dorfman DD, Freedman MT, and Berbaum K. "Integer Wavelet Compression Guided by a Computer-Aided Detection System in Mammography", SPIE Med. Imaging, Vol. 4322, 2001, pp. 643-648.

(8) Conclusions

We have completed the pilot clinical study of the effects of CAD on radiologists' reading of screening mammograms. We have collected over 2,500 cases and analyzed the results of about 2,400 cases. The overall sensitivity of the CADView system was found to be reasonably close to our prediction based on laboratory tests and also is consistent between the two sites. The computer detected 90% of the lesions that were recommended for biopsy in both sites, and 86% of the fine needle biopsy cases in the two sites. 82% of the short-term follow up cases were detected by the CAD system. Whether any of the missed short-term follow up cases will turn out to be malignant remains to be followed. The CAD system caused 30 additional callbacks at the UM site, of which 5 were recommended short-term follow up. We will track these follow-up cases to determine if any of them will turn out to be malignant. The CAD system only caused 4 additional callbacks at the GU site and one of these was found to be malignant. The CAD system detected 5 of the 6 malignant cases at the UM site, whereas causing 2 additional benign biopsies. It detected all 6 malignant cases at the GU site, one of which was not originally called by the radiologist. The total number of cancers detected was therefore increased from 11 to 12 in this patient cohort. Although the number of cancers in this pilot study is small and the statistical uncertainty is large, our results indicate that the CAD system can increase the sensitivity of breast cancer detection for screening mammography in academic centers. This information is complementary to the findings of a larger study of the effects of a commercial CAD system on screening in a community hospital, in which CAD was found to increase cancer detection substantially from 41 to 49 in 12,860 patients.

Since the cancer rate in the screening population is low, the number of patients recruited for this pilot clinical study is not sufficient to draw statistically significant conclusion on the effects of CAD on the sensitivity of mammographic screening. However, this pilot study provides an evaluation of the performance of the CAD system in the clinical screening environment and, more importantly, an assessment of the effects of CAD on the callback rate of the radiologists for reading screening mammograms. At the UM site, the call back rate seemed to increase substantially when the radiologists read mammograms with the CAD system in this study. However, the majority of the call backs was not caused by the markers by the computer. The radiologists seemed to have reduced their threshold for call back, probably they do not want to miss lesions that may be pointed out by the computer. Whether this competitive phenomenon may persist if the radiologists have to read every screening mammogram with CAD routinely remains to be seen. However, this heightened alert level will reduce the probability of false-negative diagnosis by radiologists anyway, serving partly the purpose of CAD. Nevertheless, the increased call back rate did not seem to increase the biopsy rate substantially because most of call back cases were found to be benign or negative upon workup. The results obtained from this pilot study will be important for the design of a large-scale pivotal clinical study in the future to further investigate these issues.

Two observer performance studies have been conducted for the CAD-guided image compression project. It was found that the CAD guided compression method with adequate bit rate will fully preserve the quality of microcalcifications and suspected microcalcifications without sacrificing the edge sharpness and overall image quality. Neither edge nor overall image quality degradation could be observed by the radiologist using area-equalized bit-rate of 0.25 bit/pixel and 0.4 bit/pixel. The CAD-guided compression can therefore reduce the image transmission and storage requirements for digital mammograms by a factor of 30 to 50 without causing perceivable degradation of image quality. An effective image compression method for picture archiving and communication will facilitate the implementation of telemammography and digital mammography. Both approaches are expected to improve patient care, especially in remote and rural areas.

(9) References

[1]T. W. Freer and M. J. Ulissey, "Screening mammography with computer-aided detection: Prospective study of 12,860 patients in a community breast center," Radiology 220, 781-786 (2001).

[2]M. N. Gurcan, H. P. Chan, B. Sahiner, L. Hadjiiski, N. Petrick, and M. A. Helvie, "Optimal neural network architecture selection: Improvement in computerized detection of microcalcifications," Academic Radiology (Submitted) (2001).

[3]N. Petrick, H. P. Chan, B. Sahiner, M. A. Helvie, S. Paquerault, and L. M. Hadjiiski, "Breast cancer detection: Evalution of a CAD mass detection algorithm with independent cases.," Radiology (Accepted) (2001).

[4]S. Paquerault, N. Petrick, H. P. Chan, B. Sahiner, and M. A. Helvie, "Improvement of Computerized Mass Detection on Mammograms: Fusion of Two-View Information," Med Phys (Accepted) (2001).

[5]Lo SCB, Li H, Freedman MT, and Mun SK, "Optimization of Wavelet Decomposition for Image Compression and Feature Preservation, IEEE Trans. Med. Imaging (Submitted) (2001).

[6]Lo SCB, Xuan J, Li H, "Predictive Decomposition as a Framework in Dyadic Transforms," IEEE Trans. Signal Processing (Submitted) (2001).

(10) Appendix

Publications enclosed

Journal Articles

1. Sahiner B, Chan HP, Petrick N, Helvie MA, Goodsitt MM. Development of a high-sensitivity classifier for computer-aided diagnosis: Application to classification of malignant and benign masses. Physics in Medicine and Biology 1998; 43: 2853-2871.

2. Chan HP, Sahiner B, Lam KL, Petrick N, Helvie MA, Goodsitt MM, Adler DD. Computerized analysis of mammographic microcalcifications in morphological and texture feature spaces. Medical Physics 1998; 25: 2007-2019.

3. Petrick N, Chan HP, Sahiner B, Helvie MA, Goodsitt MM. Combined adaptive enhancement and object-based region growing for automated detection of masses on mammograms. Medical Physics 1999; 26: 1642-1654.

4. Chan HP, Sahiner B, Helvie MA, Petrick N, Roubidoux MA, Wilson TE, Adler DD, Paramagul C, Newman JS, Sanjay-Gopal S. Improvement of radiologists' characterization of mammographic masses by computer-aided diagnosis: an ROC Study. Radiology 1999; 212: 817-827.

5. Chan HP, Sahiner B, Wagner RF, Petrick N. Classifier design for computer-aided diagnosis: Effects of finite sample size on the mean performance of classical and neural network classifiers. Medical Physics 1999; 26: 2654-2668.

6. Sanjay-Gopal S, Chan HP, Wilson TE, Helvie MA, Petrick N, Sahiner B. A regional registration technique for automated interval change analysis of breast lesions on mammograms. Medical Physics 1999; 26: 2669-2679.

7. Hadjiiski LM, Sahiner B, Chan HP, Petrick N, Helvie MA. Classification of malignant and benign masses based on hybrid ART2LDA approach. IEEE Transactions on Medical Imaging 1999; 18: 1178-1187.

8. Chan HP, Helvie MA, Petrick N, Sahiner B, Adler DD, Paramagul C, Roubidoux MA, Blane CE, Joynt LK, Wilson TE, Hadjiishi LM, Goodsitt MM. Digital mammogramphy: observer performance study of effects of pixel size on radiologists' characterization of malignant and benign microcalcifications. Academic Radiology 2001; 8: 454-466.

9. Hadjiiski LM, Chan HP, Sahiner B, Petrick N, Helvie MA. Automated registration of breast lesions in temporal pairs of mammograms for interval change analysis – local affine transformation for improved localization. Medical Physics 2001; 28: 1070-1079.

10. Lo SCB, Lin JS, Freedman MT, and Mun SK, Application Of Artificial Neural Networks To Medical Image Pattern Recognition: Detection of Clustered Microcalcifications on Mammograms and Lung Cancer on Chest Radiographs," J. of VLSI in Sign. Proc., Vol. 18, 1998, pp. 263-274.

11. Li H, Wang Y, Liu KJR, Lo SCB, and Freedman MT, ``Computerized Radiographic Mass Detection -- Part I: Lesion Site Selection by Morphological Enhancement and Contextual Segmentation," IEEE

Trans. on Medical Imaging, Vol. 20, No. 4, April. 2001, pp. 289-301.

12. Li H, Wang Y, Liu KJR, Lo SCB, and Freedman MT, ``Computerized Radiographic Mass Detection -- Part II: Decision Support by Featured Database Visualization and Modular Neural Networks," IEEE Trans. on Medical Imaging, Vol. 20, No. 4, April. 2001, pp. 302-313.

Articles Accepted for Publication

1. Hadjiiski L, Sahiner B, Chan HP, Petrick N, Helvie MA, Gurcan MN. Analysis of Temporal Change of Mammographic Features: Computer-Aided Classification of Malignant and Benign Breast Masses. Medical Physics. May 2001.

2. Petrick N, Sahiner B, Chan HP, Helvie MA, Paquerault S, Hadjiiski LM. Breast cancer detection: Evaluation of a CAD mass detection algorithm with independent cases. Radiology. September 2001.

3. Paquerault S, Petrick N, Chan HP, Sahiner B, Helvie MA. Improvement of Computerized Mass Detection on Mammograms: Fusion of Two-View Information. Medical Physics. September 2001.

Articles Submitted for Publication

1. Gurcan MN, Chan HP, Sahiner B, Hadjiiski LM, Petrick N, Helvie MA. Optimal neural network architecture selection: Improvement in computerized detection of microcalcifications. Academic Radiology. October, 2001.

2. Lo SCB, Li H., Hasegawa A, Wang Y, and Freedman MT, "A Multiple Circular Paths Convolution Neural Network System for Detection of Mammographic Masses," (Submitted to) IEEE Trans. on Medical Imaging, 2001.

3. Lo SCB, Li H, Freedman MT, and Mun SK, "Optimization of Wavelet Decomposition for Image Compression and Feature Preservation, (Submitted to) IEEE Trans. Med. Imaging.

4. Lo SCB, Xuan J, Li H, "Predictive Decomposition as a Framework in Dyadic Transforms," (Submitted to) IEEE Trans. Signal Processing.

5. Lo SCB and MT Freedman "Wavelet-Based Artificial Convolution Neural Network For Image Pattern Recognition: Applications to Microcalcification Detection on Mammograms," (Submitted to) IEEE Trans. Med. Imaging.

Conference Proceedings

1. Freedman MT, Lo SCB, Artz DS, and Mun SK, "Classification of False-Positive Findings on Computer-Aided Detection of Breast Microcalcifications," SPIE Proceedings on Medical Imaging 1997, vol. 3034, pp.853-859.

2. Lo SCB, Li H, Hasegawa A, Wang YJ, Freedman MT, Mun SK, "Detection of Mammographic Masses Using Sector Features with a Multiple Circular Path Neural Network," SPIE Proceedings on Medical Imaging 1998, vol. 3338, pp. 1205-1214.

3. Lo SCB, Makariou E, Delegacz A, Chan HP, Dorfman DD, Freedman MT, and Berbaum K. "Integer Wavelet Compression Guided by a Computer-Aided Detection System in Mammography", SPIE Med. Imaging, Vol. 4322, 2001, pp. 643-648.

Submitted to
Acad. Radiol.

Optimal neural network architecture selection: Improvement in computerized detection of microcalcifications

Metin N. Gurcan, Ph.D.
Heang-Ping Chan, Ph.D.
Berkman Sahiner, Ph.D.
Lubomir Hadjiiski, Ph.D.
Nicholas Petrick, Ph.D.
Mark A. Helvie, M.D.

Department of Radiology
University of Michigan, Ann Arbor

Correspondence:

Metin N. Gurcan, Ph.D.
C/o Heang-Ping Chan, Ph.D.
Department of Radiology
University of Michigan
1500 E. Medical Center Drive
UH B1F510B
Ann Arbor, MI 48109-0030
Telephone: (734) 936-4357
Fax: (734) 615-5513
E-mail:chanhp@umich.edu

Short title: Microcalcification detection improvement

RATIONALE AND OBJECTIVES: To evaluate the effectiveness of optimal neural network architecture selection on the performance of a CAD system designed for the detection of microcalcification clusters on digitized mammograms.

MATERIALS AND METHODS: We have developed a computer program to automatically detect microcalcification clusters on digitized mammograms. Previously, we have found that a properly selected and trained convolution neural network (CNN) could reduce false positives (FPs) and therefore improve the accuracy of microcalcification detection. In this work, we evaluated the effectiveness of the CNN optimized with an automated optimization technique in improving the accuracy of the microcalcification detection program in comparison with the previously manually selected CNN. For this evaluation, an independent test data set was used, which included 472 mammograms selected from the University of South Florida public database and contained a total of 253 biopsy-proven malignant clusters.

RESULTS: At an FP rate of 0.7 per image, the film-based sensitivity was 84.6% with the optimized CNN, in comparison with 77.2% with the manually selected CNN. If clusters having images in both craniocaudal (CC) and mediolateral oblique (MLO) views were analyzed and a cluster was considered to be detected when it was detected in one or both views, at 0.7 FPs/image, the sensitivity was 93.3% with the optimized CNN and 87.0% with the manually selected CNN.

CONCLUSION: Classification of true and false signals is an important step in the microcalcification detection program. An optimized CNN can effectively reduce FPs and improve the detection accuracy of the computer-aided detection system.

Keywords: Computer-aided diagnosis, convolution neural network, microcalcification detection

INTRODUCTION

Although the five-year survival rate for breast cancer has improved over the years possibly due to screening programs, breast cancer remains to be one of the most common cancers among women in the Western world (1). When breast cancer is detected in its localized stage, the five-year survival rate is 97% (2). The five-year survival rate drops to 20% if it has metastasized. Screening mammography is currently best tool available for the early detection of breast cancer (3). Although the sensitivity of mammography is relatively high compared with other breast imaging modalities, the false negative detection rate is still as high as 15 to 30%. Double reading has been shown to improve the sensitivity (4), however, it is not cost effective in a clinical setting. Computer-aided diagnosis (CAD) can provide a second opinion and can improve the detection accuracy significantly (5-8).

Several research groups have developed CAD programs for the detection of microcalcifications using different approaches. Each approach employs a number of parameters that are usually determined during the development of the CAD program. For instance, the neighborhood size for the normalization of local contrast in reference (9), and the signal-to-noise ratio (SNR) to determine the locally adaptive threshold in reference (10) are some of the CAD program parameters. Generally, these parameters are chosen by experimenting with their values manually until a satisfactory performance is achieved. However, there is no guarantee that these parameters will reach their optimum values with the trial-and-error approach.

In order to set the parameters of the CAD systems automatically in an optimal manner, several approaches have been proposed. Anastasio *et al.* used a genetic algorithm (GA) based optimization method to select the values of 10 parameters in a rule-based microcalcification detection system (11). GA searches a parameter space using an *ad hoc*

cost function. By performing training and resubstitution on a data set with 89 images, they observed that the optimization increased the sensitivity of the CAD system from 80% to 87% at an FP rate of 1.0 per image. Many CAD systems are composed of several independent yet interrelated parts. Some optimization studies in the CAD area involved optimizing one part of the CAD system. For example, Sahiner *et al.* used GA and a specially designed cost function to select features that could enhance the high sensitivity performance of a classifier for classification of malignant and benign masses (12). Chan *et al.* used GA to optimize features for differentiation of malignant and benign microcalcifications (13). Leichter *et al.* used feature selection to optimize the performance of microcalcification characterization (14). Yoshida *et al.* optimized the wavelet transform for microcalcification detection based on supervised learning (15). Tsai *et al.* used GA to determine the optimal set of fuzzy membership functions to classify myocardial heart disease from ultrasound images (16). Recently, we proposed and compared several automated techniques for the selection of optimal neural network architecture for CAD (17-20). In this work, we evaluated the effect of the CNN architecture selected by the automated optimization technique on microcalcification detection performance in comparison with our previously manually selected architecture. The performances were evaluated using a publicly available, relatively large and completely independent data set of digitized mammograms.

MATERIALS AND METHODS

A. Data Set

The data set of 108 mammograms used for the optimization and training of the CNN architecture was part of our own database collected with Institutional Review Board approval at the University of Michigan. For validation purposes we used another data set consisting of 152 mammograms, which was also part of our own database but different

from the set used for training. The mammograms in our database were randomly selected from the files of screening and diagnostic patients who had undergone biopsy at the University of Michigan so that they included microcalcifications with a wide range of characteristics similar to those encountered in clinical practice. The training and the validation data sets were digitized with a Lumisys 85 laser scanner (Lumisys, Sunnyvale, CA). For test purposes, an independent data set was used. This data set included 472 digitized mammograms, selected from the University of South Florida (USF) digitized mammogram database, which is publicly available over the internet (21). From all the available cases in this database, only malignant cases that were digitized with the Lumisys 200 laser scanner were selected (volumes: cancer_01, cancer_02, cancer_05, cancer_09, and cancer_15). The mammograms were digitized at a pixel resolution of 0.05 mm x 0.05 mm with 4096 gray-levels. We converted these images to 0.1mm x 0.1 mm resolution by averaging adjacent 2x2 pixels and subsampling. The detection was carried out on these 0.1-mm resolution images. The optical density (OD) range of the scanner was 0-3.6. The digitizer was calibrated so that the gray values were linearly and inversely proportional to the OD with a slope of –0.001 OD/pixel value. Details of the case collection method are described in the USF website (21)

Types of the microcalcifications in the selected cases included punctate, amorphous, pleomorphic, round and regular, fine linear branching, round, dystrophic. The distributions of the calcifications were clustered, linear, segmental, and regional. The lesion types, the assessment, the subtlety, and the pathology were provided with the database. The cluster location(s) was marked on each image as an overlay file. There were a total of 272 microcalcification clusters, 253 of which were biopsy-proven malignant. Figure 1 shows the distribution of the assessment code for the malignant clusters used in this study. The assessment was provided in the USF database and it follows the American College of Radiology breast imaging reporting and data system

(BIRADS) categories. This distribution shows that most of the clusters are in the "actionable lesion" category, which is defined as BIRADS assessment score of 0, 4, and 5. These scores require patient callback for additional mammograms or biopsy. Figure 2 shows the breast density information. A majority of the clusters comes from breasts with densities of 2 and 3. Figure 3 shows the distribution of the subtlety rating for the malignant clusters. There is no BIRADS standard for the subtlety rating. In the USF database, a cluster with a subtlety rating of 1 is the most obvious while a cluster with a subtlety rating of 5 is the subtlest. It may be concluded from this distribution that the majority of the clusters used in this study could be classified as subtle.

B. Microcalcification Detection Program

We have developed a computer program to automatically detect microcalcification clusters on digitized mammograms (5, 6). The program has three major steps. The first step is preprocessing in which the breast boundary is automatically determined and the breast region is filtered with a band-pass filter to obtain a signal-to-noise (SNR) enhanced image. The second step is segmentation. In this step, potential microcalcification locations are determined using global and locally adaptive thresholding methods. The local threshold is calculated as the product of the local root-mean-square (RMS) noise and an input SNR threshold. The microcalcification size, maximum contrast and SNR are also calculated. In the third step, the extracted signals are classified as either a true microcalcification (TP) or a false-positive (FP) signal. The first stage is a rule-based classification that uses the size, contrast and SNR information to generate decision rules. The second-stage classification uses a trained convolution neural network (CNN) classifier to recognize the abnormal patterns. Finally, regional clustering is used to identify clusters of signals. If a TP signal is within a neighborhood of other TP signals, they are combined to form a cluster. Previously, we have found that the CNN could

effectively reduce the number of FPs and therefore improve the accuracy of the microcalcification detection program (10).

C. Convolution Neural Network

The CNN is based on the neocognitron structure of Fukushima (22). It was previously used for detection of lung nodules on chest radiographs, detection of microcalcifications on mammograms, and classification of mass and normal breast tissue on mammograms (10, 23, 24). Figure 4 shows a schematic representation of the CNN structure. The input to the CNN is a region of interest (ROI) image, extracted for each of the detected signals. The nodes in the hidden layers are arranged in groups; each group functions like a filter kernel. The CNN classifies the input ROI as a TP or an FP. The output node value is close to one for true microcalcifications and is close to zero for FP signals. In this work, the CNN had one input node, two hidden layers and one output node. All node groups in the two hidden layers were fully connected. The images in each layer were convolved with the filter kernels to obtain the pixel values in the images to be transferred to the following layer. There were N_1 node groups in the first layer, and N_2 node groups in the second hidden layer. The kernel sizes of the first group of filters between the input node and the first hidden layer were K_1x K_1, and those of the second group of filters between the first and second hidden layer were K_2x K_2. Sigmoidal activation functions were used and the CNN was trained using the error back-propagation rule.

D. Neural Network Architecture Selection

The CNN architecture used in our earlier studies was selected using a manual optimization technique (10). We recently evaluated the use of automated optimization methods for selecting an optimal CNN architecture. Details of the automated architecture selection study have been described in the literature (20). Briefly, three automated methods, the steepest descent (SD), the simulated annealing (SA), and the genetic

algorithm (GA) were compared. Four main parameters of the CNN architecture, N_1, N_2, K_1, and K_2, were considered for optimization. The area, A_z, under the receiver operating characteristic (ROC) curve was used to design a cost function. The SA experiments were conducted with four different annealing schedules. Three different parent selection methods were compared for the GA experiments. Our training data set consisted of region-of-interest (ROI) images extracted from 108 mammograms, described above. The locations of individual microcalcifications in these images were manually identified and saved in a truth file. After the prescreening steps of the microcalcification detection program (10), the detected signals were labeled as TP or FP automatically by comparing with the truth file. A 16x16-pixel ROI was then extracted for each of the detected signals and these ROI images were used for training and testing the CNN. Either a true or a false microcalcification was located at the center of the ROI. The microcalcification detection program detected more FP ROIs than TP ROI images at the prescreening stage. In order to have approximately equal numbers of TP and FP ROIs, only a randomly selected subset of FP ROI images was used. The selected ROIs were divided into two separate groups. For the first part of the experiments, the first group, G1, was used for training the CNN and the second group, G2, was used for testing the trained CNN. For the second part of the experiment, the roles of G1 and G2 were switched. The first group, G1, consisted of 533 TP and 553 FP ROIs. The second group G2 had 547 microcalcification ROIs, and 570 FP ROIs. Therefore, G1 contained a total of 1086 ROIs and G2 contained 1117 ROIs. The optimal architecture (N_1-N_2-K_1-K_2) was determined to be 14-4-5-5 when the architecture was trained with G1 and tested with G2, and 14-10-5-7 when the training and the test sets were switched. In our previous study (10), the optimal architecture was determined to be 12-8-5-3 using a manual search technique.

RESULTS

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

In order to test the performance of the selected optimal architecture, the detection program was run at seven SNR threshold values varying between 2.6 and 3.2 at

increments of 0.1. Figure 6 shows the FROC curves of the microcalcification detection program using both the manually optimized and automatically optimized CNN architectures. The FP rate was estimated by performing the detection on the normal mammograms. The automatically optimized architecture again outperformed the manually optimized architecture. At an FP rate of 0.7 cluster per image, the film-based sensitivity is 84.6% with the optimized CNN, in comparison to 77.2% with the manually selected CNN. Figure 7 shows the FROC curves for the microcalcification detection programs if clusters having images in both craniocaudal (CC) and mediolateral oblique (MLO) views are analyzed and a cluster is considered to be detected when it is detected in one or both views. This "case-based" scoring has been adopted for the evaluation of some CAD systems (8). The rationale is that if the CAD system can bring the radiologist's attention to the lesion on one of the views, it will be unlikely that the radiologist will miss the lesion. For case-based scoring the sensitivity at 0.7 FPs/image is 93.3% with the automatically optimized CNN and 87.0% with the manually selected CNN.

DISCUSSION

Classification of true and false signals is an important step in the microcalcification detection program. An optimized CNN can effectively reduce FPs and improve the detection accuracy of the CAD system. Manually searching for the optimal CNN architecture often results in a local optimum because it is difficult to explore adequately a high-dimensional parameter space with manual experimentation. We have demonstrated previously that an automated optimization algorithm such as simulated annealing can find the global optimum efficiently (17-20).

Our optimization is currently limited to one stage, FP reduction with the CNN, of the detection program. Our cost function was based on the A_z of the CNN classifier for its performance in differentiating the TP and FP signals. Ideally, one would prefer to optimize all parameters in the detection program together. In such a case, optimizing the performance in terms of the FROC curve will be necessary. In order to take advantage of some well-established automated optimization methods such as GA or SA, it is necessary to define a scalar cost function. However, there is no widely accepted form of a scalar cost function for the comparison of FROC curves obtained as a result of different detection methods. In an alternative form of FROC analysis, known as AFROC analysis, a scalar A_1 is calculated, which can be considered a form of cost function, but AFROC analysis requires a special experimental setting (25). Anastasio *et al.* (11) proposed an *ad hoc* cost function, *C(f,s)*, in which they incorporated their preferences about their sensitivity-specificity tradeoff into a discrete grid of numbers on the sensitivity-specificity plane; the values in between these grid values were determined by means of bilinear interpolation. The fitness of each solution during their GA evolution process was assigned by evaluation of the cost function for the solution. Since the cost function optimized the FROC curve only at an individual operating point that corresponded to a sensitivity-specificity pair, it did not provide sufficient information to compare two different FROC curves. Moreover, the choice of the preference values is quite subjective. For our optimization study, the ROC methodology, a commonly accepted form of comparing overall classifier performance, was used; therefore, the cost definition was based on the area under the ROC curve, A_z. To extend this definition for FROC curves, we propose the following cost function

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

where l and u are the lower and upper limits of the FP range of interest, respectively; f is the FP per image and $s(f)$ is the sensitivity at an FP rate of f. This cost function will compare two FROC curves in a chosen range of FP rates. A similar function was proposed by te Brake *et al.* to measure the quality of a feature for the discrimination of malignant masses from normal tissue in digital mammograms (26). In their definition, the area under the logarithmically plotted FROC curve between 0.05 and 4.0 FPs per image was used as a quality measure:

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

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

Of all the available images in the USF database, we used only those scanned by the Lumisys scanner because this was similar to the scanner that we used to acquire digitized mammograms for developing our CAD programs and setting its parameters. It is not uncommon to see drastic performance decreases if different types of scanners are used for the development and testing of a CAD system. For instance, Velthuitzen *et al.*

developed a microcalcification detection program using mammogram images digitized with a DBA ImageClear R3000 (DBA Systems Inc, Melbourne, FL) scanner and achieved 94% sensitivity at an FP rate of 1.23 per image on a database of 26 images (27). When they scanned the same images with a Lumiscan 50 (Lumisys, Sunnyvale, CA) scanner, and evaluated the detection performance, the sensitivity dropped to 28% and the FP rate increased to 2.19 per image. In this study, since we were interested in evaluating the performance change due to CNN architecture selection, we limited ourselves to those images in the USF database that were scanned by a similar scanner, thereby keeping the effects of other factors on the performance change to a minimum. The dependence of our detection program on data set acquired with different film scanners will be investigated in the future.

For this optimization study, we followed a three-stage (training-validation-test) CAD development and evaluation methodology. This methodology requires separate data sets for each stage. Table I summarizes the information about the images in these data sets. The images in the first two data sets came from the patient files at the University of Michigan. However, these two data sets were mutually exclusive; they did not share any common images. The data set for training was used to find the parameters of the optimal neural network architecture and neural network weights. The images in the validation set were used to evaluate the performance of the selected architectures and identify the best performing architecture for an independent data set. Once the architecture was selected using the validation set, the parameters of the detection program were fixed and no further changes were made either to the program, or to the CNN architecture and its weights. Using this CAD program, microcalcification detection was carried out on a completely independent and publicly available test data set. The images in this set were used only to assess the performance of the fully specified optimal architecture. If only a small training set and an "independent" test set are used, and the detection performance

on the test set is used as a guide to adjust the parameters of the detection program, there is always a bias due to fine tuning the CAD system to this particular "test" data set that is essentially a validation set. The results achieved with that test set may not be generalizable to other data sets. This is especially an important consideration for CAD system development. Before a CAD system can be considered for clinical implementation, it is advisable to follow this three-stage methodology and to evaluate the system with an independent random test set that contains a large number of mammograms with a wide spectrum of characteristics. Otherwise, the test results may not truly reflect the actual performance of the CAD program in the unknown patient population.

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

The scoring of the microcalcification detection program was performed automatically. Figure 9 demonstrates how our automatic scoring scheme was designed. There are two sets of inputs to the automatic scoring program. The first set consists of the overlay files where the extent of each microcalcification cluster is drawn by an expert radiologist as a polygon. The second set consists of outputs of the automated microcalcification detection program, which are the smallest rectangular bounding boxes enclosing the detected microcalcification clusters. The scoring program automatically calculates the intersection of the areas enclosed by these rectangles and the polygons. If the ratio of the intersection area to either the rectangle or the polygon area is more than 40%, then the cluster enclosed by the polygon is considered to be detected. If a polygon area is detected by more than one rectangular region, only one TP is recorded. The sensitivity for the film-based FROC curve was determined based on the number of malignant clusters detected relative to the total number of malignant clusters present in the data set, considering different views of the same cluster to be independent. For case-based scoring, the corresponding clusters in the two views are used to determine if the same cluster is detected by the CAD system in at least one view. Detection of the same cluster in one or both views will be scored as one TP and the sensitivity is normalized to the total number of different malignant clusters in the data set.

At present, there is no established statistical test for comparing the significance in the differences between two FROC curves. Therefore, we cannot provide a statistical significance evaluation on the improvement in the FROC curves with the optimized CNN. However, since the increase in the sensitivity is substantial, from 77.2% to 84.6% at 0.7 FP per image, and is consistent over the range of FP studied, the effectiveness of the CNN is evident. Furthermore, since the improvement is observed for a relatively large independent test set, and is consistent with the performance observed with the

validation set, this reduces the likelihood that the improvement is biased to the specific data set.

CONCLUSION

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

ACKNOWLEDGMENTS

This work is supported by a USPHS Grant CA 48129 and USAMRMC grant DAMD 17-96-6254. The content of this publication does not necessarily reflect the position of the government and no official endorsement of any equipment and product of any companies mentioned in the publication should be inferred. The authors are grateful to Charles E. Metz, Ph.D., for providing the LABROC program.

FIGURE CAPTIONS:

Figure 1. Distribution of the assessment rating for the clusters used in our test data set. The assessment follows the ACR BIRADS standard and was provided with the USF database. Because only biopsy-proven malignant clusters were included in this test set, the clusters have a BIRADS evaluation of 3 (Probably benign finding - short interval follow-up suggested), 4 (Suspicious abnormality - biopsy should be considered), and 5 (Highly suggestive of malignancy - appropriate action should be taken).

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

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

Figure 4. Schematic diagram of the architecture of a convolution neural network. The input to the CNN is a region of interest image extracted for each of the detected signals. The output is a scalar that is the relative rating by the CNN representing the likelihood that the input ROI contains a true microcalcification or a false-positive signal.

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

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

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

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

Figure 9. In this schematic mammogram, there are four microcalcification clusters, (C_1, C_2, C_3, C_4), the extents of which are drawn by radiologists. The microcalcification detection program detects five clusters (D_1, D_2, D_3, D_4, D_5). D_1 is a TP detection. D_2 and D_3 are FP detections because D_2 does not intersect with any cluster and D_3's intersection with C_3 is less than 40%, which was chosen as the threshold for detection during training and validation of the automatic scoring criteria. D_4 and D_5 are considered to be detecting the same cluster, C_4. Therefore, for this example, the number of TP's is 2 (C_1, C_4), the number of false-negatives is 2 (C_2, C_3), and the number of FP's is 2 (D_2 and D_3).

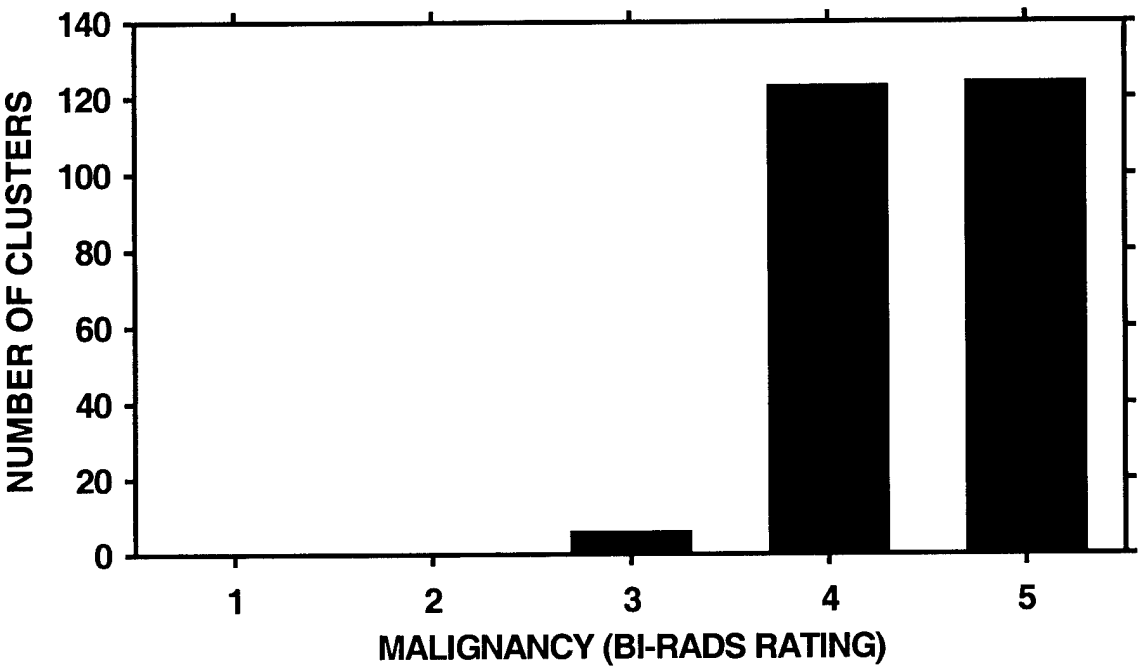


Figure 1. Distribution of the assessment rating for the clusters used in our test data set. The assessment follows the ACR BIRADS standard and was provided with the USF database. Because only biopsy-proven malignant clusters were included in this test set, the clusters have a BIRADS evaluation of 3 (Probably benign finding - short interval follow-up suggested), 4 (Suspicious abnormality - biopsy should be considered), and 5 (Highly suggestive of malignancy - appropriate action should be taken).

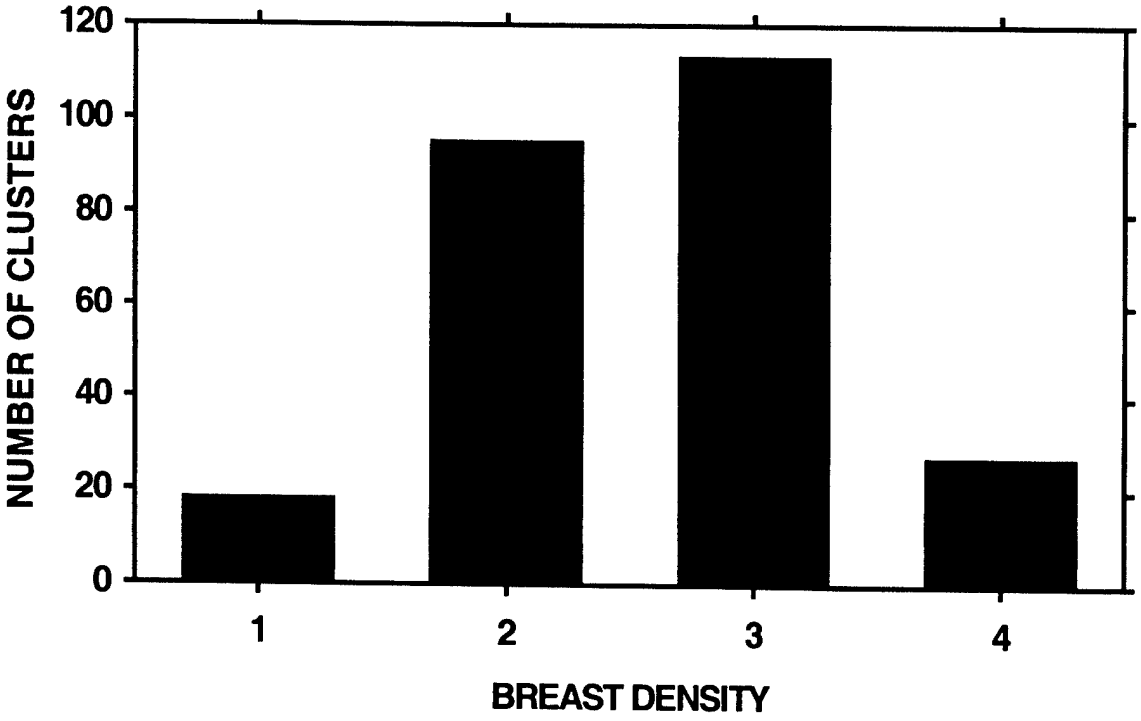


Figure 2. Breast density information for the mammograms included in the test data set. The breast density information follows the BI-RADS standard and was provided with the USF database: 1 = almost entirely fat, 2 = with scattered fibroglandular densities, 3 = heterogeneously dense, 4 = extremely dense.

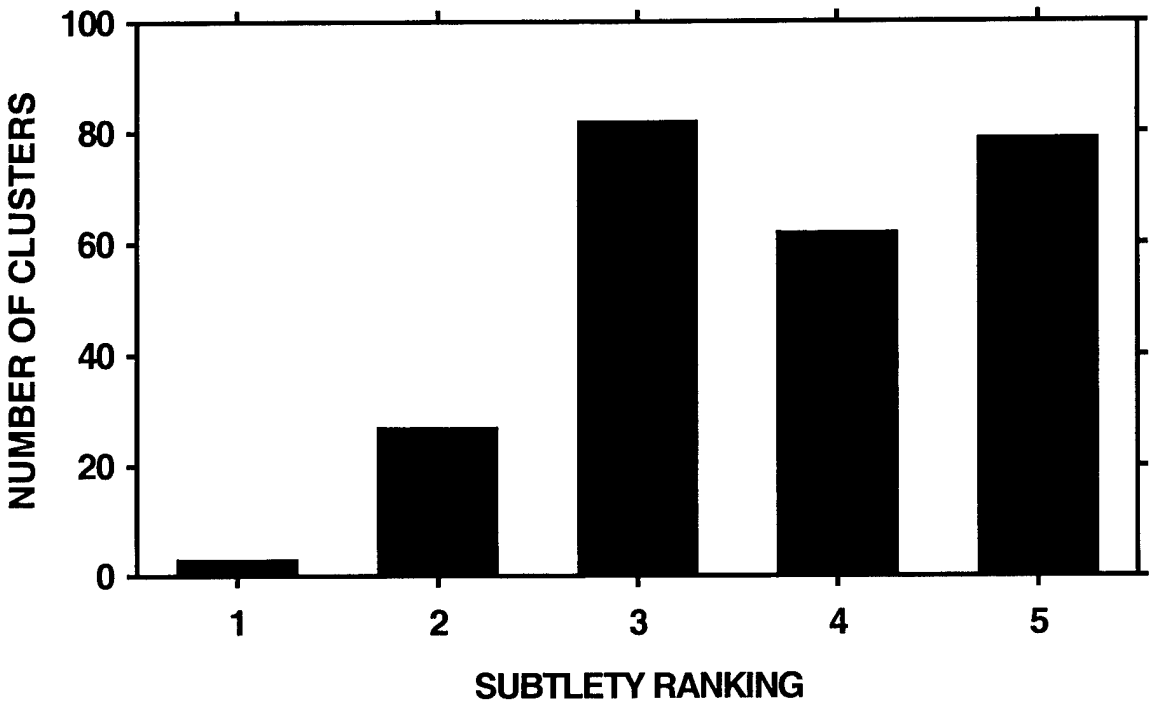


Figure 3. Subtlety ranking (1 = obvious, 5 = subtle) of the 253 clusters provided with the USF data set.

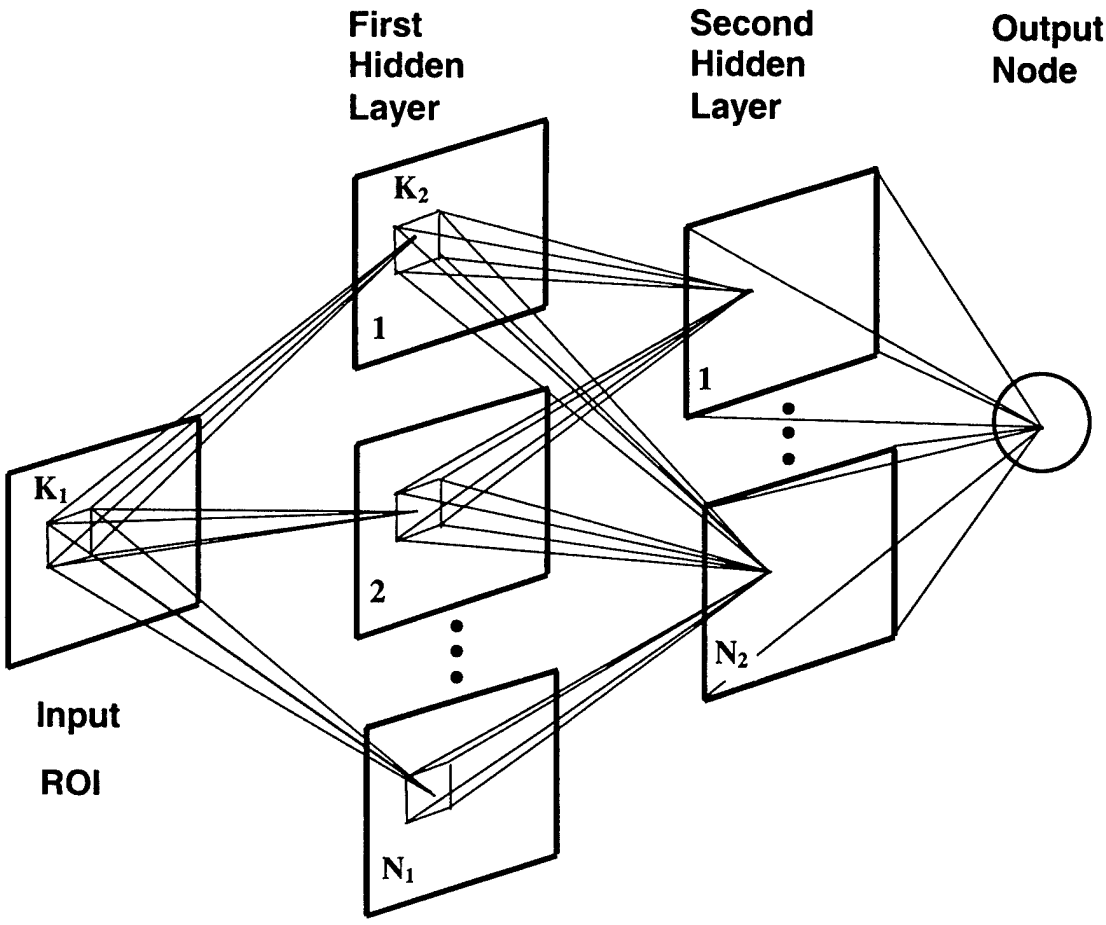


Figure 4. Schematic diagram of the architecture of a convolution neural network. The input to the CNN is a region of interest image extracted for each of the detected signals. The output is a scalar that is the relative rating by the CNN representing the likelihood that the input ROI contains a true microcalcification or a false-positive signal.

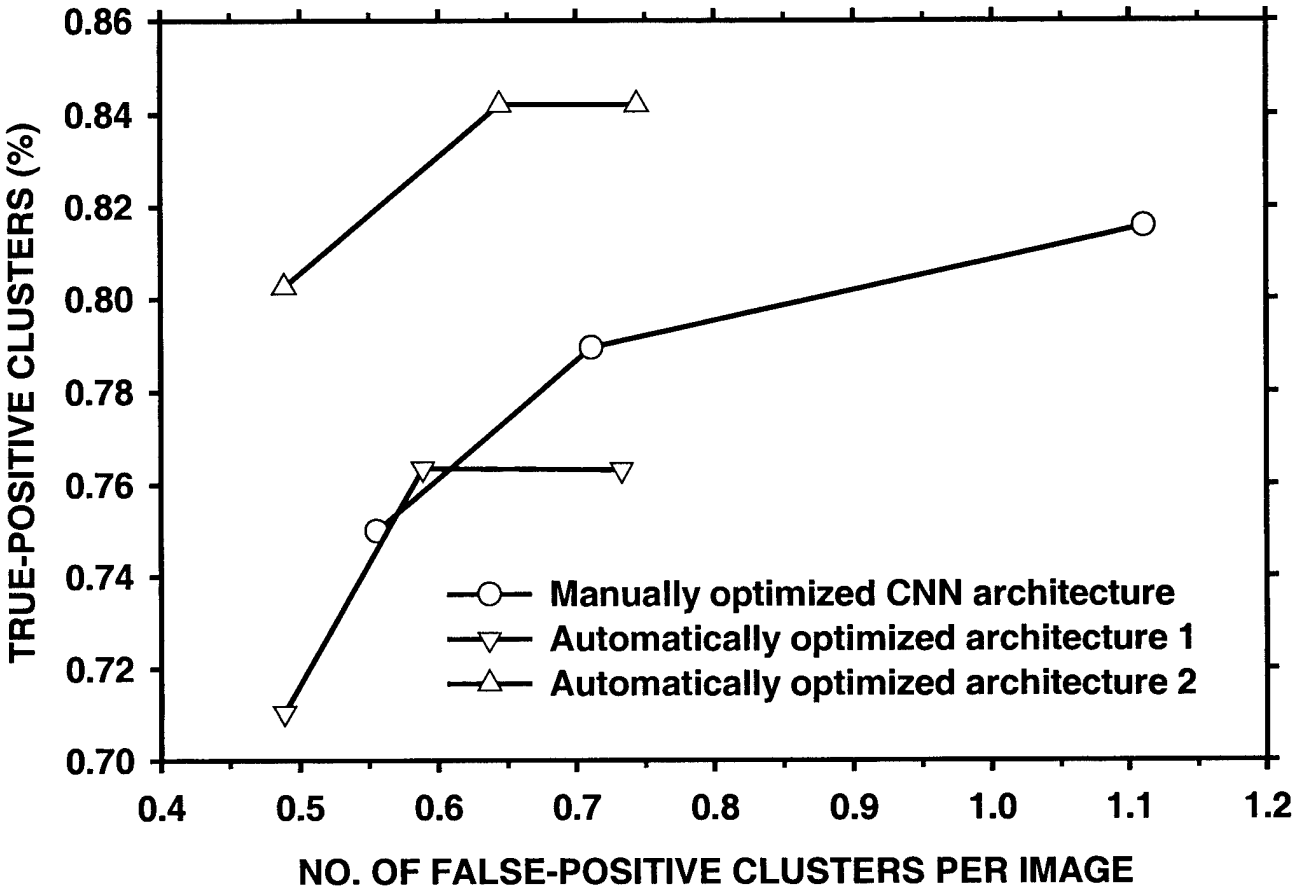


Figure 5. Comparison of validation FROC curves for detection of clustered microcalcifications using different CNN architectures: (a) manually optimized architecture (12-8-5-3), (b) automatically optimized architecture 1 (14-4-5-5), (c) automatically optimized architecture 2 (14-10-5-7). The evaluation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

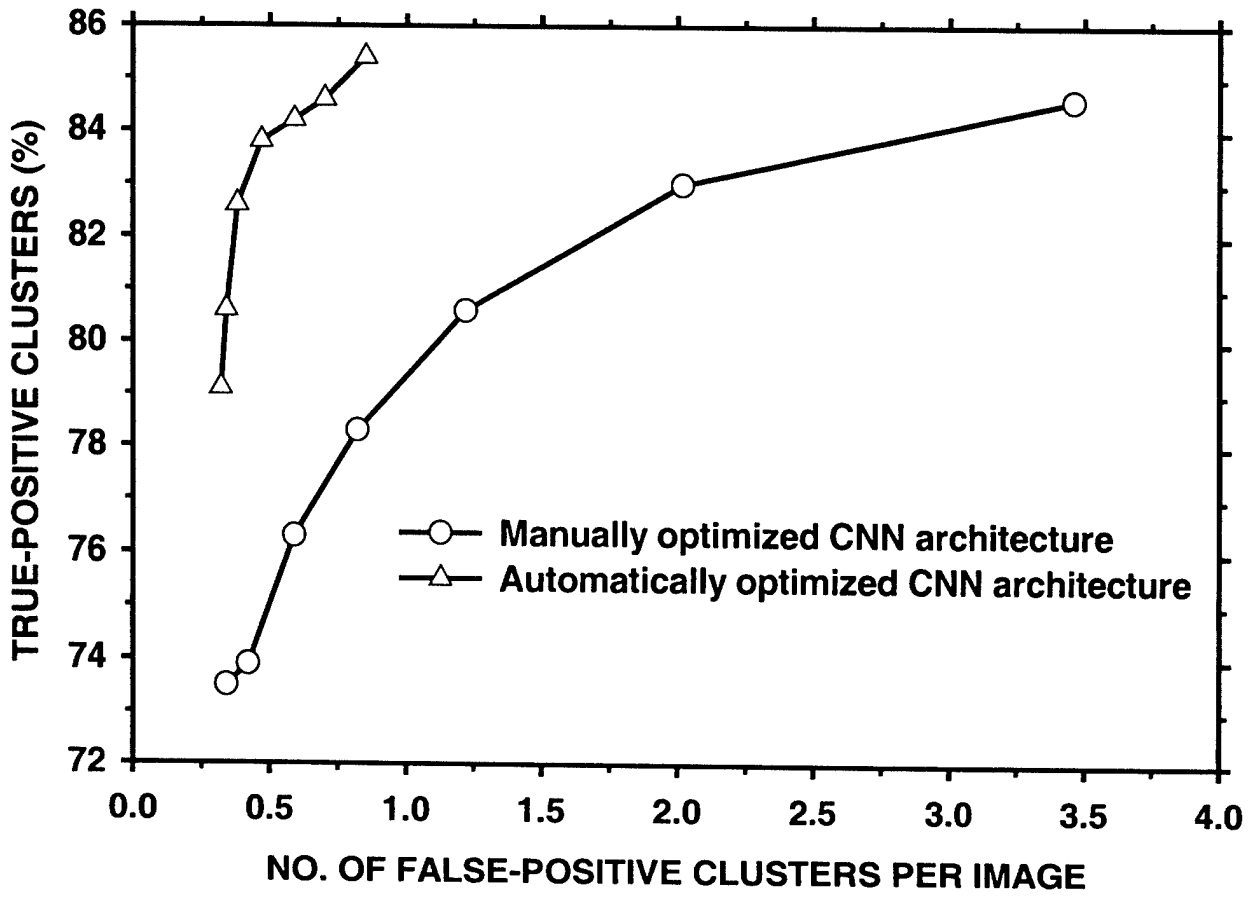


Figure 6. Comparison of test FROC curves for detection of clustered microcalcifications with manually and automatically optimized CNN architectures for film-based (single view) scoring. The automatically optimized architecture is 14-10-5-7. The evaluation was performed using the 472-image test data set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

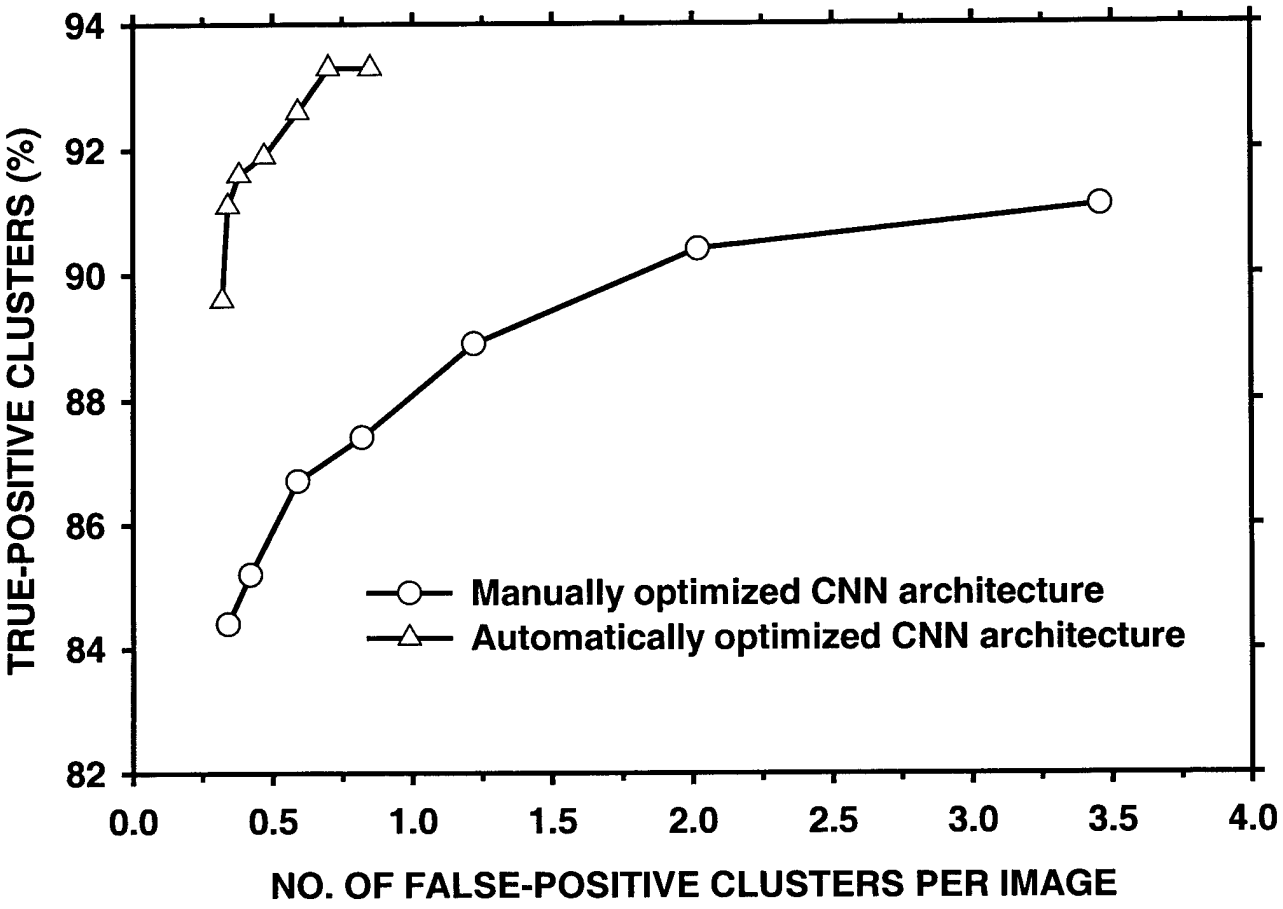


Figure 7. Comparison of test FROC curves for detection of clustered microcalcifications with manually and automatically optimized CNN architectures for case-based scoring. In case-based scoring, if clusters having images in both CC and MLO views are analyzed, a cluster is considered to be detected when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

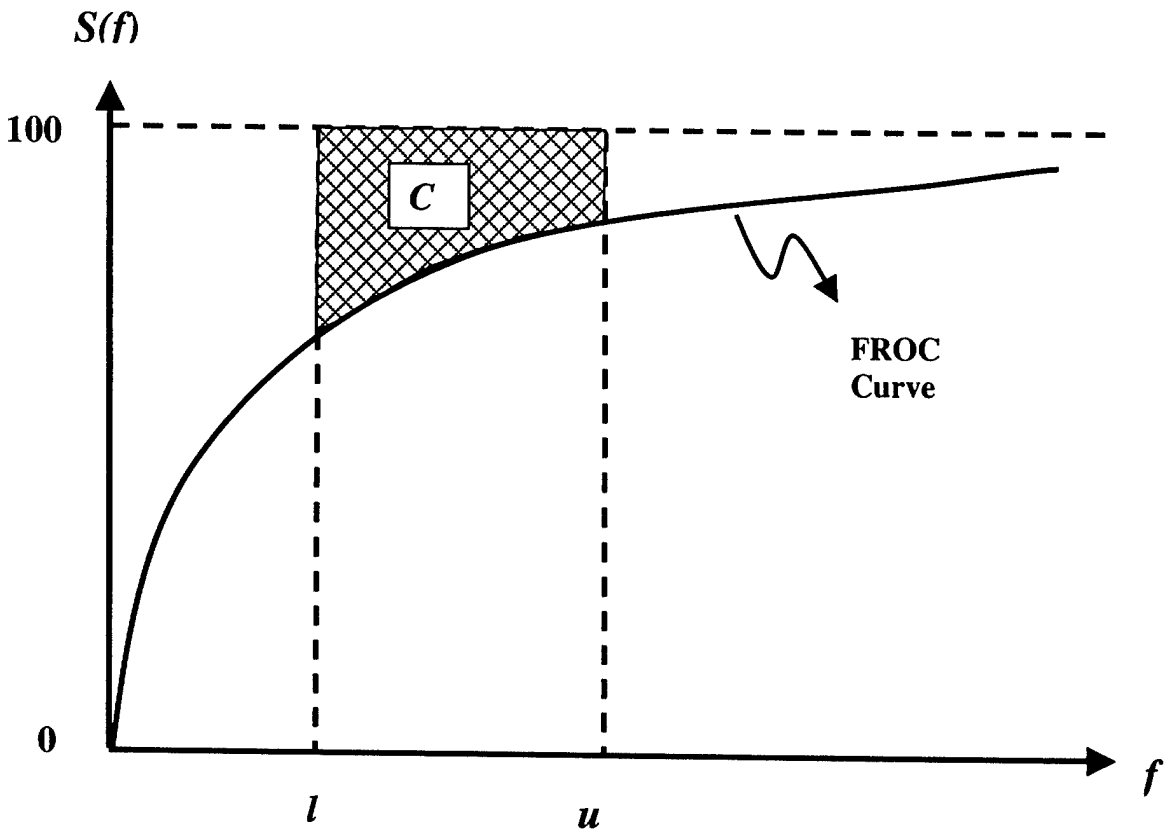


Figure 8: Definition of a scalar cost function for optimization of CAD system. l and u are the lower and upper limits of the range of the number of FPs per image on the FROC curve, respectively, f is the number of FP per image, $s(f)$ is the sensitivity at an FP rate of f. The cost, C, is determined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

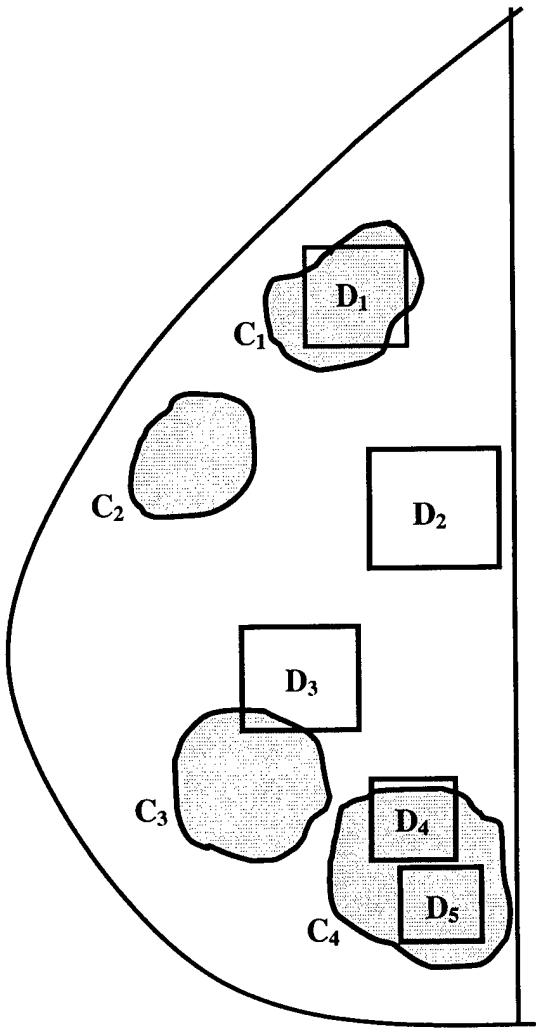


Figure 9. In this schematic mammogram, there are four microcalcification clusters, (C_1, C_2, C_3, C_4), the extents of which are drawn by radiologists. The microcalcification detection program detects five clusters (D_1, D_2, D_3, D_4, D_5). D_1 is a TP detection. D_2 and D_3 are FP detections because D_2 does not intersect with any cluster and D_3's intersection with C_3 is less than 40%, which was chosen as the threshold for detection during training and validation of the automatic scoring criteria. D_4 and D_5 are considered to be detecting the same cluster, C_4. Therefore, for this example, the number of TP's is 2 (C_1, C_4), the number of false-negatives is 2 (C_2, C_3), and the number of FP's is 2 (D_2 and D_3).

DATA SET	SOURCE	NO. OF IMAGES	NO. OF MALIGNANT MICROCALC. CLUSTERS
TRAINING	University of Michigan	108	29
VALIDATION	University of Michigan	152	76
TEST	University of South Florida	472	253

Table I. Summary of the data sets used in the different stages (training, validation, test) of this study. These data sets are mutually exclusive, *i.e.*, there is no overlap of images in the three data sets.

REFERENCES

1. Vogel V. Breast cancer prevention: A review of current evidence. CA Cancer J Clin 2000; 50: 156-170.

2. National Cancer Institute. Surveillance, Epidemiology, and End Results (SEER) Program Public-Use CD-ROM (1973-1997). Journal 2000;

3. American Cancer Society: Mammography guidelines 1983: Background statements and update of cancer-related checkup guidelines for breast cancer detection in asymptomatic women age 40 to 49. CA Cancer J Clin 1983; 33: 225.

4. Thurfjell EL, Lernevall KA, Taube AAS. Benefit of independent double reading in a population-based mammography screening program. Radiology 1994; 191: 241-244.

5. Chan HP, Doi K, Vyborny CJ, Schmidt RA, Metz CE, Lam KL, et al. Improvement in radiologists' detection of clustered microcalcifications on mammograms. The potential of computer-aided diagnosis. Investigative Radiology 1990; 25: 1102-1110.

6. Kegelmeyer WP, Pruneda JM, Bourland PD, Hillis A, Riggs MW, Nipper ML. Computer-aided mammographic screening for spiculated lesions. Radiology 1994; 191: 331-337.

7. Freer TW, Ulissey MJ. Screening mammography with computer-aided detection: Prospective study of 12,860 patients in a community breast center. Radiology 2001; 220: 781-786.

8. Warren Burhenne LJ, Wood SA, D'Orsi CJ, Feig SA, Kopans DB, O'Shaughnessy KF, et al. Potential contribution of computer-aided detection to the sensitivity of screening mammography. Radiology 2000; 215: 554-562.

9. Veldkamp WJH, Karssemeijer N. An improved method for detection of microcalcification clusters in digital mammograms. Proc. SPIE 1999; 3661: 512-522.

10. Chan HP, Lo SCB, Sahiner B, Lam KL, Helvie MA. Computer-aided detection of mammographic microcalcifications: Pattern recognition with an artificial neural network. Medical Physics 1995; 22: 1555-1567.

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

12. Sahiner B, Chan HP, Petrick N, Helvie MA, Goodsitt MM. Design of a high-sensitivity classifier based on a genetic algorithm: Application to computer-aided diagnosis. Physics in Medicine and Biology 1998; 43: 2853-2871.

13. Chan HP, Sahiner B, Lam KL, Petrick N, Helvie MA, Goodsitt MM, et al. Computerized analysis of mammographic microcalcifications in morphological and texture feature space. Medical Physics 1998; 25: 2007-2019.

14. Leichter I, Lederman R, Buchbinder S, Bamberger P, Novak B, Fields S. Optimizing parameters for computer-aided diagnosis of microcalcifications at mammography. Academic Radiology 2000; 7: 406-412.

15. Yoshida H, Zhang W, Cai W, Doi K, Nishikawa RM, Giger ML. Optimizing wavelet transform based on supervised learning for detection of microcalcifications in digital mammograms. Proc. IEEE International Conference on Image Processing, 1995; 3: 152-155.

16. Tsai D-Y, Watanabe S. A method for optimization of fuzzy reasoning by genetic algorithms and its application to discrimination of myocardial heart disease. IEEE Transactions on Nuclear Science 1999; 46: 2239-2246.

17. Gurcan MN, Sahiner B, Chan HP, Hadjiiski LM, Petrick N. Optimal selection of neural network architecture for CAD using simulated annealing. Proc. 22nd Annual International Conference of IEEE Engineering in Medicine and Biology Society, Chicago, IL, 2000; 3052-3055.

18. Gurcan MN, Sahiner B, Chan HP, Hadjiiski LM, Petrick N. Selection of an optimal neural network architecture for computer-aided diagnosis - comparison of automated optimization techniques. Radiology 2000; 217(P): 436.

19. Gurcan MN, Chan HP, Sahiner B, Hadjiiski LM, Petrick N, Helvie MA. Improvement of computerized detection of microcalcifications using a convolution neural network architecture selected by an automated optimization algorithm. Medical Image Perception Conference IX, Airlie Conference Center, Warrenton, VA, 2001;

20. Gurcan MN, Sahiner B, Chan HP, Hadjiiski LM, Petrick N. Selection of an Optimal Neural Network Architecture for Computer-aided Detection of Microcalcifications -

Comparison of Automated Optimization Techniques. Medical Physics 2001; 28: 1937-1948.

21. Heath M, Bowyer K, Kopans D, Kegelmeyer P, Moore R, Chang K, et al. Current status of the digital database for screening mammography. In: Karssemeijer N, Thijssen M, Hendriks J,van Erning L, ed. Digital Mammography. Dordrecht, Kluwer Academic, 1998; 457-460.

22. Fukushima K, Miyake S, Ito T. Neocognitron: A neural network model for a mechanism of visual pattern recognition. IEEE Trans Systems Man Cybernetics 1983; SME-13: 826-834.

23. Lo SCB, Chan HP, Lin JS, Li H, Freedman M, Mun SK. Artificial Convolution neural network for medical image pattern recognition. Neural Networks 1995; 8: 1201-1214.

24. Sahiner B, Chan HP, Petrick N, Wei D, Helvie MA, Adler DD, et al. Classification of mass and normal breast tissue: A convolution neural network classifier with spatial domain and texture images. IEEE Transactions on Medical Imaging 1996; 15: 598-610.

25. Chakraborty DP, Winter LHL. Free-response methodology: Alternate analysis and a new observer-performance experiment. Radiology 1990; 174: 873-881.

26. te Brake GM, Karssemeijer N, Hendriks JHCL. An automatic method to discriminate malignant masses from normal tissue in digital mammograms. Physics in Medicine and Biology 2000; 45: 2843-2857.

27. Velhuitzen RP, Clarke LP. Image Standardization for digital mammography. In: Karrssemeijer N,Thijssen M, ed. Dordrecht, Kluwer Academic Publishers, 1998; 461-464.

Accepted by Radiology

BREAST CANCER DETECTION: EVALUATION OF A COMPUTER-AIDED MASS DETECTION ALGORITHM WITH INDEPENDENT MAMMOGRAPHIC CASES

Author:

Nicholas Petrick, Ph.D.
Berkman Sahiner, Ph.D.
Heang-Ping Chan, Ph.D.
Mark A. Helvie, M.D.
Sophie Paquerault, Ph.D.
Lubomir M. Hadjiiski, Ph.D.

Institutional Affiliations:

All Authors: Department of Radiology
The University of Michigan
CGC B2102, Box 0904
1500 East Medical Center Drive
Ann Arbor, MI 48109-0904

Corresponding Author:

Nicholas Petrick, Ph.D.
Phone: 734-647-7428
Fax: 734-647-8557
Email: petrick@umich.edu

Grant Supporting this Research:

This work is supported by the Whitaker Foundation (N.P.), USPHS Grant CA 48129, USPHS Grant CA 79943 (N.P.), USAMRMC Grant DAMD 17-96-1-6012 (B.S.), USAMRMC Grant DAMD 17-98-1-8211 (L.M.H.) and USAMRMC DAMD 17-96-1-6254.

RSNA Presentations:

This work has been submitted to RSNA 2001.

Type of Manuscript:

Original Research

ABSTRACT

Purpose: We have developed a computer-aided detection (CAD) algorithm to detect breast masses on digitized mammograms. In this study, we analyzed the performance of the algorithm with clinical cases.

Materials and Methods: A digitized mammogram is processed with an adaptive enhancement filter followed by a local border refinement stage. Features are then extracted from each detected structure and used to identify potential masses. We evaluated the algorithm's performance on independent cases obtained from 263 patients from two institutions. The CAD marker rate was estimated by applying the algorithm to 503 normal films.

Results: The computer detected a malignant mass in 83% (130/156) of the malignant cases at a marker rate of 1.0 marks per film. The detection accuracy for benign lesions was lower than that for malignant masses. Free-response receiver operating characteristic (FROC) performance curves are obtained and the tradeoff between detection sensitivity and the number of CAD marks is analyzed. A performance comparison between cases collected at the two different institutions is also included.

Conclusion: Our mass detection algorithm has a high sensitivity for detection of malignant masses . It may be useful as a second opinion in mammographic interpretation.

Keywords: Computer-Aided Diagnosis, Preclinical Evaluation, Mass Detection, Breast Cancer, Mammography

INTRODUCTION

Breast cancer is one of the leading causes of death among American women between 40 to 55 years of age (1). Women who are in a regular mammographic screening have a statistically significant reduction in breast cancer mortality compared to women who do not undergo screening (2). In addition, independent double reading by two radiologists increases the sensitivity of mammographic screening (3). Results of studies indicate that a 4%-15% increase in detected cancers is possible with double reading (3-5). However, the higher cost and increased workload may make double reading by two radiologists impractical in a general screening situation. Computer-aided diagnosis (CAD) is a cost-effective alternative to double reading.

Efforts to evaluate the usefulness of CAD in reducing missed cancers are ongoing. A prospective study of 12,860 patients in a community breast cancer center using a commercial CAD system (ImageChecker V2.0, R2 Technologies, Los Altos, CA) reported a cancer detection rate of 81.6% (40/49) with 8 of the cancers initially detected by computer only. This corresponds to a 20% (41 vs 49) increase in the number of cancers detected (6). These results demonstrate that a CAD system can reduce the missed cancer rate when used as a second opinion even if it does not detect all cancers.

The above results do not distinguish between cancers presenting as masses alone, microcalcification cluster alone, or a combination on the mammograms. In this study, we focus on the detection of preoperative mammographic masses. We define a preoperative mass as a mass that is identified during clinical evaluation and either undergoes biopsy based on this exam, or is followed and determined to be benign. Castellino *et. al.* found that the latest version of the

R2 ImageChecker achieved a mass detection sensitivity of 85.7% with 0.5 marks per image for 677 preoperative mass cases, up from a sensitivity of 74.7% and 1.0 marks per image in a previous (V2.0) release (7). A second study evaluating the Second Look system (CADx Medical Systems, Laval, Quebec, Canada) reported a mass detection sensitivity of 84% with 1.1 marks per image on a database of 149 preoperative mass cases (8).

The purpose of this paper is to define the performance of our CAD mass detection algorithm in marking preoperative masses. This paper differs from previous publications in that the performance is given for both malignant and benign lesions, and on a per-film and a per-case basis. The benign lesion performance is included because the prevalence of CAD markers for benign lesions is different from that for normal structures. In addition, the detection of benign lesions may affect the performance of radiologists using CAD differently from the influence of normal markers. Since an independent data set was used for the evaluation (i.e., the algorithm parameters were fixed without any influence from the test data set), the results presented here should give an indication of the potential benefits of our algorithm when used in clinical practice. We also discuss the major factors that lead to differences in detection performance between malignant and benign masses, and performance differences from cases collected from different institutions.

MATERIALS AND METHODS

Data Sets

This study involved the collection of mammographic films and biopsy information for the evaluation of a CAD mass detection algorithm. These cases were collected with Institutional

Review Board (IRB) approval and the IRB determined that, with our protocol of maintaining patient confidentiality, no patient consent was needed for data collection.

Training Cases

The clinical mammograms used for training the algorithm were selected from the files of patients who had undergone biopsy. The mammograms were acquired with MinR/MRE screen/film systems (Eastman Kodak, Rochester, NY) using dedicated processing. The selection criterion used by the radiologists was that a biopsy-proven mass existed on the mammogram. The data set consisted of 253 mammograms from 102 patients, and it included 128 malignant and 125 benign masses. Sixty-three of the malignant and six of the benign masses were judged to be spiculated by an MQSA approved radiologist.

The mammograms were digitized with a DIS-1000 laser film scanner (Lumisys Inc, Sunnyvale, CA) with a pixel size of 100 μm and 12 bit gray level resolution. The gray levels were linearly proportional to optical density in the 0.1 to 2.8 optical density unit (O.D.) range and gradually fell off in the 2.8 to 3.5 O.D. range.

Independent Test Cases

We analyzed the performance of the trained mass detection algorithm with independent clinical cases. These mammograms were not used in the training process. Cases were collected from two different institutions. The first set of cases, referred to as Group 1, was selected from the files of 127 patients who had undergone biopsy at our institution. The mammograms were acquired with MinR/MRE screen/film systems using dedicated processing in the years prior to 1997 and a Kodak 2000 screen/film system (Eastman Kodak, Rochester, NY) from 1997 on.

Each case consisted of a single craniocaudal (CC) and either a mediolateral oblique (MLO), or lateral view of the breast containing the mass. For simplicity, we will refer to all views other than the CC view as the MLO view in the following discussions with the understanding that this also includes some lateral views. If both breasts of a patient had a mass, each breast was considered to be an independent case. Using these breast-based definitions, a total of 138 cases (with 276 mammograms) were available. Each case contained preoperative breast masses that were identified by a radiologist during clinical evaluation. The independent Group 1 mammograms were digitized with a Lumisys LS 85 laser film scanner (Lumisys Inc, Sunnyvale, CA) that digitized the images at 50 μm and 12 bit gray level resolution. The gray levels were calibrated to be linearly proportional to optical density in the 0.1 to 4.0 O.D. range. The images were reduced to a 100 μm pixel size by averaging 2×2 pixel neighborhoods before performing mass detection.

Clinical cases from the public database available from the University of South Florida (USF) were also analyzed (9). We evaluated an additional 142 CC/MLO pairs from 136 patients collected by USF. For compatibility with the Group 1 database, we only selected USF cases digitized with the Lumisys 200 laser film scanner (Lumisys Inc, Sunnyvale, CA). This scanner again digitized the images at 50 μm and 12 bit gray level resolution but the gray levels were calibrated to be linearly proportional to optical density in the 0.1 to 3.6 O.D. range. The 142 USF cases will be referred to as the Group 2 cases in the following discussions.

Table 1 summarizes the Group 1 and 2 test cases used to evaluate the mass detection algorithm. It includes the number of malignant and benign masses separated by whether they were visible in both views or only in a single view. Fig. 1 shows the distributions of lesion

subtlety (1:subtle to 5:obvious) for the Group 1 and 2 databases as ranked by a radiologist evaluating each individual mass. The rankings of all Group 2 masses were retrieved from the USF database. The mammographic size for the Group 1 masses was measured by the radiologist during initial case evaluation. The malignant Group 1 masses had a mean size, standard deviation and median size of 15.4 mm, 12.0 mm, 12.0 mm, respectively. The benign Group 1 masses had a mean size, standard deviation and median size of 13.4 mm, 11.8 mm and 10.0 mm, respectively. Radiologist-measured mass sizes were not available for the Group 2 cases and we found that the annotations outlining the mass locations in these cases were much larger than the actual mammographic lesion size. Therefore, mass size information is not reported for the Group 2 cases.

The IRB did not require the collection of racial or ethnic information from the subjects at our institution so no statistics on the racial or ethnic composition are available for the Group 1 cases. However, since the cases were randomly sampled from the patients undergoing mammographic exams in our hospital, the composition is expected to be similar to that of our patient population. The ethnicity statistics for our mammography screening patient population in 1998 and 1999 are given in Table 2. Table 2 also includes the patient ethnicity statistics for the Group 2 cases, which were collected in the USF public database.

Mass Detection Algorithm

Algorithm Description

Our mass detection scheme uses adaptive enhancement, object-based border refinement and feature classification to identify potential breast masses. The block diagram for the scheme is

shown in Fig. 2. The first step is to digitize a film mammogram. The digitized mammogram is then processed by an initial segmentation step, in which a density-weighted contrast enhancement (DWCE) filter is utilized for preprocessing. The DWCE filter was developed to accentuate mammographic structures before edge detection by adaptively enhancing the local contrast. After DWCE filtering, edge detection is employed to define the borders of the enhanced structures. This results in a set of detected structures. Each of these structures is then processed by a local refinement stage. First, seed locations are identified by finding all local maxima within each object, using an ultimate erosion technique (10) and then selecting all connected pixels with gray values in the range $M_i \pm 0.01 \cdot M_i$ where M_i is the gray level of the i^{th} local maximum. K-means clustering is then applied to a 25 mm $\times$ 25 mm background-corrected region of interest (ROI) (11) centered on each seed object to refine the initial object border (12). The purpose of the local refinement stage is to improve the accuracy of object borders found by the DWCE segmentation because DWCE segmentation tends to underestimate the size of breast structures. The local refinement was also found to be effective in splitting large connected regions into smaller breast structures. The final stage is to classify each detected object as a breast mass or normal structure based on extracted morphological and texture features. In order to overcome the problems associated with the large number of initial structures, we perform the feature classification in two stages. Eleven morphological features are initially used with a threshold and a linear classifier to remove detected normal structures that are significantly different from breast masses. Texture-based classification then follows this morphological reduction stage. Fifteen global and local multiresolution texture features, based on the spatial gray level dependence (SGLD) matrices are used as inputs to a linear discriminant classifier, which merges the input feature into a single discriminant score for each detected

object. Decision thresholds based on this score and on the maximum number of marks allowed per image are then used to identify potential breast masses. Further details on the mass detection algorithm can be found in the literature (13-16).

Algorithm Training

The computer program was trained using the entire training data set of 253 mammograms. This included adjusting the filters, clustering, selected features and classification thresholds. Once training was completed the parameters and all thresholds were fixed for testing. The training data set was then resubstituted into the algorithm and was found to have a film-based (i.e., each mass on each film was considered as an independent sample) training sensitivity of 81% (85% for malignant masses). The mass detection algorithm produced 2.9 marks per film on average for the training cases. It is important to note that the detection classifiers considered only classification between breast masses and normal tissue, not between malignant and benign masses. Therefore, no distinction was made between malignant and benign masses in the training process.

Definition of TP and FP Markers

For the Group 1 cases, the smallest bounding box containing the entire mass identified by a radiologist was used as the truth. For Group 2, we used a bounding box around the annotated region provided with each image. Our definition of a TP was based on the percentage of overlap between the bounding box of an identified structure and the bounding box of the true mass. Based on the training set, we chose an overlap threshold of 25%. This value corresponds to the minimum overlap between the bounding box of a detected object and the bounding box of a mass in order for the object to be considered as a TP detection. All detected objects that did not meet

this criterion were considered as FPs. The 25% threshold was selected because it was found to match well with TPs identified visually. The detected objects were first labeled automatically by the computer using this criterion. All of the TPs were then visually reviewed to make sure that the program highlighted the true lesion and not a neighboring structure. Marks that were found to match neighboring structures were changed to FPs.

The number of false positive (FP) marks produced by the algorithm was determined by counting the markings produced in normal cases. We used lesion-free films of the breast contralateral to the breast containing an abnormality as normal cases. Since some cases contained lesions in both breasts, we have fewer normal films than abnormal films. We used a total of 251 normal films from Group 1 and 252 normal films from Group 2 to define the marker rate. The TPF, calculated from the abnormal cases, and the average number of marks per image, calculated from the normal cases, were defined for a fixed set of thresholds at the final texture classification stage. The TPF and the average number of marks per image as the threshold varied were then used to plot the FROC performance curves for malignant and benign masses in the different data sets.

RESULTS

Test performance results are presented on a per-film and per-case basis. In the former, the CC and MLO views are considered independently so that a lesion visible in the CC view is considered as a TP and the same lesion in the MLO view is a second TP. In the latter, a mass is considered detected if it is detected on either the CC or the MLO view. The latter evaluation takes into consideration that, in clinical practice, once the computer alerts the radiologist to a

cancer in one view, it is unlikely that the radiologist will miss the cancer. The per-case approach is often used by other researchers in reporting their CAD performance (5, 8, 17). Results are also presented for two different TP scoring methods. The individual scoring method considers each mass in a film or case as a different TP. The grouped scoring method considers all malignant masses in a film or case as a single TP (5). The rationale for group scoring is that a radiologist may not need to be alerted to all malignant lesions in a film or case before taking action. Therefore, multiple detections in a film or case may not significantly enhance the power of CAD.

The FROC curves for the mass detection with the individual data sets are shown in Figs. 3-5. Figs. 3 and 4 contain the FROC performance curves for Group 1 and 2 based on individual mass scoring. The FROC performance of the algorithm for malignant masses based on grouped mass scoring is shown in Fig. 5. We also analyzed the sensitivity achieved by the mass detection algorithm at three fixed normal marker rates. These marker rates were selected because they represent potential clinical implementations for a CAD algorithm based on previously published studies (7, 8). The results at these fixed marker levels are summarized in Table 3.

DISCUSSION

The detection performance curves shown in Figs. 3-5 clearly indicate that our mass detection algorithm is effective in detecting breast masses and that the detection performance for malignant masses is better than that for benign masses. Since these results were based on a large independent test set from two institutions, and the algorithm parameters were not adjusted based on any characteristics of the test set, the performance results estimated in this study should be close to the true performance in the patient population. The malignant mass detection fractions

of 77%, 83% and 87% at marker rates of 0.5, 1.0 and 1.5, respectively, are currently the best estimates for the clinical performance of our mass detection algorithm. These performance values compare quite well with the published performance results from the commercial CAD vendors (85.7% TPF at 0.5 marks/image by R2 and 84% TPF at 1.1 marks/image by CADx (7, 8) as well as with other algorithms currently being development in research laboratories. This indicates that our mass detection could be beneficial to radiologists as a second opinion.

We first compare the performance of the algorithm on malignant and benign lesions in each database. From Table 3 and Fig. 3, we see a somewhat consistent difference in the TPFs between the malignant and benign masses in Group 1. The per-case difference is in the range of 9% to 12% from 3.0 to about 0.25 marks/image. Per-film results for Group 1 follow the same trend. The difference between the malignant and benign masses is larger in the USF database (see Table 3 and Fig. 4). Here the per-case difference starts at about 12% at 3.0 marks/image, increases to 19% at 1.5 marks/image and then to 34% at 1.0 mark/image. The per-film performance again follows a similar trend as shown in Fig 4. It is clear that the performance on the Group 2 benign cases is much lower than that for the Group 1 benign cases. However, the performance difference between the Group 1 and Group 2 malignant masses is small.

One possible cause for the performance differences between benign and malignant masses is a difference in lesion subtlety. From Fig. 1, we observe a small difference in rated subtlety between the benign and malignant masses in the Group 1 database, with the malignant masses being slightly more obvious than the benign masses. This same trend holds for the Group 2 masses. However, it should be noted that the subtlety distributions between Group 1 and Group 2 differ considerably as will be discussed below. The observed difference in subtlety between

benign and malignant masses for both Groups 1 and 2 is not particularly large so a subtlety difference does not seem to fully explain the large disparities observed in Figs. 3 and 4. Another factor that likely contributes to the observed difference is that malignant masses are more likely to be spiculated than benign masses and our algorithm's performance on spiculated masses is superior to its performance on non-spiculated masses. In the Group 1 database, 34% (49/146) of the malignant and 5% (8/159) of the benign masses were spiculated. There were 33% (65/197) and 0% (0/132) spiculated masses in the Group 2 malignant and benign cases, respectively. In our training set, 49% (63/128) of the malignant and 6% (8/125) of the benign lesions were judged as spiculated by radiologists. A comparison between spiculated and non-spiculated mass performances is shown in Fig. 6. The spiculated benign mass curve is not included because of the small number of lesions in this category. It is clear from Fig. 6 that the algorithm is better suited to detect spiculated masses, especially at the lower marker rates, although no special efforts was made to train the algorithm to detect spiculated masses. We surmise that the texture analysis acquired a higher sensitivity to spiculated masses during the training process because of the relatively large fraction of spiculated lesions in the training set. Even though the detection algorithm had a higher sensitivity in detecting spiculated masses, the large number of non-spiculated masses in the training set (182/253) still trained the algorithm to be sensitive to non-spiculated malignant masses. The sizable difference between the malignant and benign non-spiculated mass curves suggesting that some additional, yet undetermined, factors may also be contributing to the observed performance difference between malignant and benign masses.

We also observe performance differences between masses in Groups 1 and 2. The malignant mass detection performances are quite similar as shown in Fig. 5, but the detection of benign

lesions differs considerably between the groups. One potential factor is that 94% (147/157) of the benign masses in the Group 1 database underwent biopsy. This high rate of benign biopsy suggests that the Group 1 masses were judged by the radiologist to be similar enough too malignant masses to warrant biopsy (i.e., the vast majority of the lesions were ACR BI-RADS category 4 and 5). We therefore expect the detection performance of these benign masses to be somewhat similar to that of the malignant masses for Group 1. The number of benign biopsies was not available for the Group 2 database, but it is likely that a smaller fraction of the benign lesions underwent biopsy leading to a larger fraction of ACR BI-RADS category 2 or 3 lesions. If this is true, then the Group 2 benign masses would not match the characteristics of our training set as well and may therefore be more difficult to detect. Another factor that may have contributed to this performance difference is a difference in the optical density ranges of the digitizers used to acquire the cases at each institution. The O.D. ranges were 0-3.5, 0-4.0 and 0-3.6 for the Lumisys digitizers used to digitize the training, Group 1, and Group 2 mammograms, respectively. The smaller O.D. range of the digitizer used to digitize the Group 2 mammograms may have caused a decrease in the detection performance for subtle low-density lesions compared with the Group 1 performance in similar cases. However, the Group 2 digitizer has an advantage in many of the cases because it better matches the O.D. range of the digitizer used to acquire the training set. Because of the presence of other factors such as case variability, it is difficult to differentiate the relative importance of these competing effects on the algorithm's performance.

Comparing the subtlety between the Group 1 and Group 2 databases, we observed a large disparity in the radiologists' rankings. From the subtlety histograms in Fig. 1, one may conclude

that the Group 2 cases are much easier than the Group 1 cases for both malignant and benign masses. However, this does not agree with our detection results. The detection performance for the Group 1 benign cases was much better than that for the Group 2 benign cases even though the Group 1 lesions were ranked as more subtle. The much more "obvious" malignant masses in the Group 2 database resulted in only a small 1%-2% gain in the detection performance when compared with the Group 1 malignant cases (Table 3). Likewise, visual comparison of the cases did not reveal such a large difference between the databases. The Group 2 subtlety distribution does not match well with what is expected in clinical practice because it is highly skewed towards obvious. One would expect that a randomly drawn sample from the patient population would follow a distribution much more similar to the Group 1 histogram. Therefore, the subtlety difference is most likely caused by a difference in the subjective criteria used to define lesion subtlety instead of a true difference in subtlety between the cases. It is likely that the individual radiologists used different scales at the different institutions. The radiologist reading cases from institution 1 appeared to have spread their subtlety ratings across the multiple categories while the radiologists at institution 2 seemed to have used basically a binary decision of visible or not visible. The results suggest that caution must be taken when comparing detection results using different databases. Even if subtlety ratings are available, the rating criteria may be subjected to large inter- and intra-observer variations.. This is especially true if the databases are collected by different institutions. Comparisons between lesions rated at a single institution using a consistent rating criterion (e.g., comparing malignant and benign lesions from the same data set) are much less problematic.

The results show that our automated mass detection algorithm is capable of detecting

malignant masses in mammograms with a low FP marker rate, suggesting that this CAD algorithm may be useful as a second reader in the clinical interpretation of mammograms. Further studies are underway to both improve the detection performance and reduce the marker rate of the algorithm with single-view information. Studies are also underway to determine how well the mass detection performs on prior mammograms in which the lesion was not sent for biopsy. Good performance in prior cases may lead to earlier cancer detection. We are also developing a new technique that will incorporate information from different mammographic views of the same breast (18, 19). Our preliminary results indicate that two-view information fusion will improve sensitivity and reduce FPs in our detection algorithm. Studies will also be conducted to determine if our CAD algorithm aids radiologists in detecting breast cancer earlier and if it affects their recall rate.

ACKNOWLEDGMENTS

This work was supported by the Whitaker Foundation (N.P.), USPHS Grant CA 48129, USPHS Grant CA 79943 (N.P.), a Career Development Award DAMD 17-96-1-6012 (B.S.), USAMRMC Grant DAMD 17-98-1-8211 (L.M.H.) and a research grant DAMD 17-96-1-6254 from the U.S. Army Medical Research and Materiel Command. The content of this publication does not necessarily reflect the position of the government, and no official endorsement of any equipment or product should be inferred. A special thanks to Christopher Washington for downloading the USF database and converting the data format so that the cases could be included in this study.

REFERENCES

1. Landis SH, Murray T, Bolden S, Wingo PA. Cancer statistics, 1998. CA Cancer J Clin 1998; 48:6-29.

2. Tabar L, Fagerberg C, Gad A, et al. Reduction in mortality from breast cancer after mass screening with mammography. Lancet 1985; 1:829-832.

3. Thurfjell EL, Lernevall KA, Taube AAS. Benefit of independent double reading in a population-based mammography screening program. Radiology 1994; 191:241-244.

4. Beam V, Sullivan D, Layde P. Effect of human variability on independent double reading in screening mammography. Academic Radiology 1996; 3:891-897.

5. Warren Burhenne LJ, Wood SA, D'Orsi CJ, et al. Potential contribution of computer-aided detection to the sensitivity of screening mammography. Radiology 2000; 215:554-562.

6. Freer TW, Ulissey MJ. Computer-aided detection (CAD) in screening mammography: A prospective study of 12,860 patients in a community breast center (abstr). Radiology 2001; 217(P):400.

7. Castellino RA, Roehrig J, Zhang W. Improved computer-aided detection (CAD) algorithm for screening mammography (abstr). Radiology 2000; 217(P):400.

8. Brem RF, Schoonjans JM, Hoffmeister J, Raza S, Baum JK. Evaluation of breast cancer with a computer-aided detection system by mammographic appearance, histology and lesion size (abstr). Radiology 2000; 217(P):400.

9. Heath M, Bowyer K, Kopans D, et al. Current status of the digital database for screening mammography. In: Digital Mammography. Karssemeijer N, Thijssen M, Hendriks J, van

Erning L, eds. Dordrecht: Kluwer Academic, 1998; 457-460.

10. Russ JC. The Image Processing Handbook. Boca Rato, FL: CRC Press, 1992

11. Sahiner B, Chan HP, Petrick N, et al. Classification of mass and normal breast tissue: A convolution neural network classifier with spatial domain and texture images. IEEE Transactions on Medical Imaging 1996; 15:598-610.

12. Chan H-P, Petrick N, Sahiner B. Chapter 6. Computer-aided breast cancer diagnosis. In: Artificial Intelligence Techniques in Breast Cancer Diagnosis and Prognosis. Jain A, Jain A, Jain S, Jain L, eds. New Jersey: World Scientific, 2000; 179-264.

13. Petrick N, Chan HP, Sahiner B, Wei D. An adaptive density-weighted contrast enhancement filter for mammographic breast mass detection. IEEE Transactions on Medical Imaging 1996; 15:59-67.

14. Petrick N, Chan HP, Sahiner B, Helvie MA. Combined adaptive enhancement and region-growing segmentation of breast masses on digitized mammograms. Medical Physics 1999; 26:1642-1654.

15. Petrick N, Chan H-P, Sahiner B, Helvie MA, Paquerault S. Evaluation of an automated computer-aided diagnosis system for the detection of masses on prior mammograms. Proc. SPIE 2000; 3979: 967-973.

16. Petrick N, Sahiner B, Chan HP, Helvie MA, Paquerault S. Preclinical evaluation of a CAD algorithm for early detection of breast cancer. 2000; Proc. IWDM-2000: (in press).

17. Paquerault S, Petrick N, Chan HP, Sahiner B, Dolney AY. Improvement of mammographic lesion detection by fusion of information from different views. Proc. SPIE 2001; 4322:(in press).

18. Sahiner B, Petrick N, Chan HP, Paquerault S, Helvie MA, Hadjiiski LM. Recognition of

lesion correspondence on two mammographic views - A new method of false-positive reduction for computerized mass detection. Proc . SPIE 2001; 4322: (in press).

TABLES

Table 1
Summary of the cases, patients and masses in the Group 1 and Group 2 databases.

	Abnormal							
	Total		Malignant		Benign		Normal	
Database	Films	Patients	One View Masses*	Two Views Masses†	One View Masses*	Two View Masses†	Films	Patients
Individual Masses‡								
Group 1	276	127	2	72	3	78	251	93
Group 2	284	136	5	96	6	63	252	128
Grouped Masses§								
Group 1	128	64	-	64	-	-	251	93
Group 2	184	92	-	92	-	-	252	128

*One view masses correspond to the masses that are visible in only one mammographic view in the pair.

†Two view masses correspond to masses that are visible in both mammographic views in the pair.

‡The individual masses category considers each mass in a film or case as a TP during scoring. The number of abnormal films and patients include cases with both malignant and/or benign masses.

§The grouped masses category considers all malignant masses for a film or case together as one TP during scoring. The number of abnormal films and patients include only cases with malignant masses.

Table 2
Summary of the ethnic composition of the Group 1 and 2 patient populations.

Ethnicity*	Percentage of Population	
	Group 1†	Group 2†
American Indian or Alaskan native (American Indian)	0.2	0.1
Asian or Pacific islander (Asian)	2.8	0.2
Black, not of Hispanic origin	7.0	20.4
Hispanic (Spanish Surname)	0.5	1.8
White, not of Hispanic origin (White)	83.5	77.0
Other/Unknown	6.0	0.4

Note. — The sum of all categories may not equal 100% because of rounding errors.
*The ethnicity text is the actual label used by Institution 1 in the Group 1 description. The text in parentheses is the corresponding label used by Institution 2 in the Group 2 description, if it differed.
†Percentages without fractions are given because ethnicity is based on a larger mammographic patient population, not from the particular cases used in this study.

Table 3

Summary of the per-case mass detection performance at marker rates of 0.5, 1.0, and 1.5 marks per image.

Data Set	TPF*		
	0.5 Marks	1.0 Marks	1.5 Marks
Individual Malignant†			
Group 1	55/74 (74)	59/74 (80)	63/74 (85)
Group 2	76/101 (75)	83/101 (82)	84/101 (83)
Combined	131/175 (75)	142/175 (81)	147/175 (84)
Individual Benign†			
Group 1	51/81 (63)	58/81 (72)	60/81 (74)
Group 2	23/68 (34)	33/68 (49)	44/68 (65)
Combined	74/149 (50)	91/149 (61)	104/149 (70)
Grouped Malignant‡			
Group 1	49/64 (77)	53/64 (83)	55/64 (86)
Group 2	71/92 (77)	77/92 (84)	80/92 (87)
Combined	120/156 (77)	130/156 (83)	135/156 (87)

Note. — The numbers in parentheses are percentages

*The TPFs are given for the three different normal film marker rates

†Each individual mass in a film or case is considered as a TP for the individual malignant and benign categories.

‡All malignant masses for a film or case are considered together as one TP for the grouped malignant category.

CAPTIONS FOR ILLUSTRATIONS

Fig 1: Histogram of lesion subtlety for the 138 and 142 cases in the Group 1 and Group 2 databases, respectively, as ranked by the radiologist reviewing the cases. Each mass in each film was rated independently by the radiologist. For comparison purposes, the plot is of the percentage of masses falling within each category. The total number of masses for each group can be found in Table 1.

Fig 2: The block diagram for the mass detection scheme evaluated in this study.

Fig 3: The Group 1 FROC performance curves for malignant and benign masses. The per-case curve is obtained by defining a TP as the detection of the mass in either the CC or MLO view mammogram of a breast. The per-film curve treats the same mass on the CC and MLO films independently. Individual mass scoring is used in the figure so each individual mass in a film or case was considered as a TP.

Fig 4: The Group 2 FROC performance curves for malignant and benign masses. Individual mass scoring is used in the figure so each individual mass in a film or case was considered as a TP.

Fig 5: The Group 1 and 2 FROC performance curves for malignant masses. Grouped mass scoring is used in the figure so all malignant masses in a film or case were considered together as one TP.

Fig 6: The combined Group 1 and 2 FROC performance curves for spiculated and non-spiculated masses. The benign spiculated mass curve is not shown because of the small

number of cases in this category. Individual mass scoring is used in the figure so each individual mass in a film was considered as a TP.

ILLUSTRATIONS

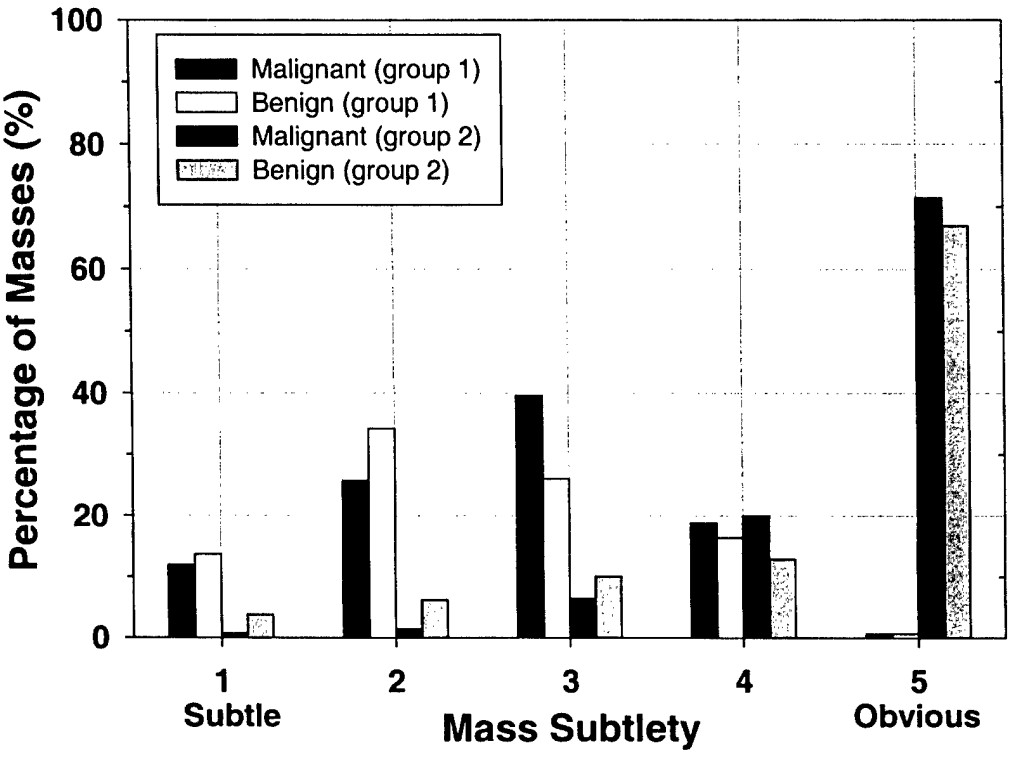


Fig 1: Histogram of lesion subtlety for the 138 and 142 cases in the Group 1 and Group 2 databases, respectively, as ranked by the radiologist reviewing the cases. Each mass in each film was rated independently by the radiologist. For comparison purposes, the plot is of the percentage of masses falling within each category. The total number of masses for each group can be found in Table 1.

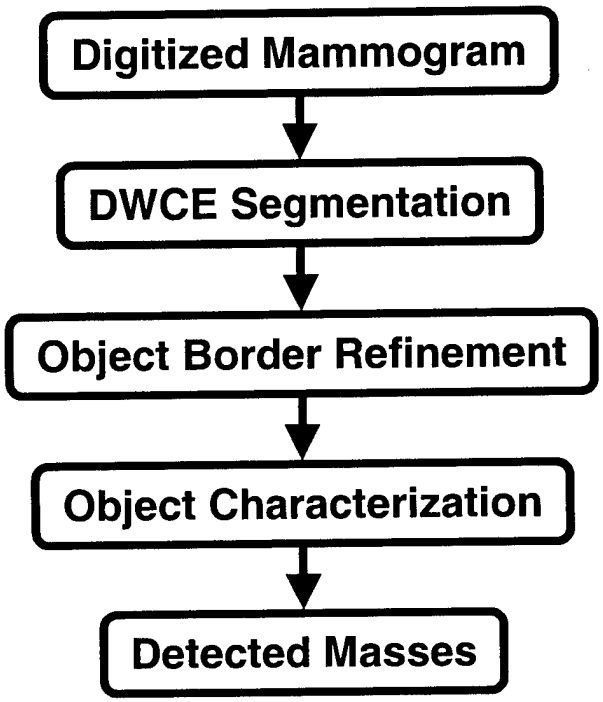


Fig 2: The block diagram for the mass detection scheme evaluated in this study.

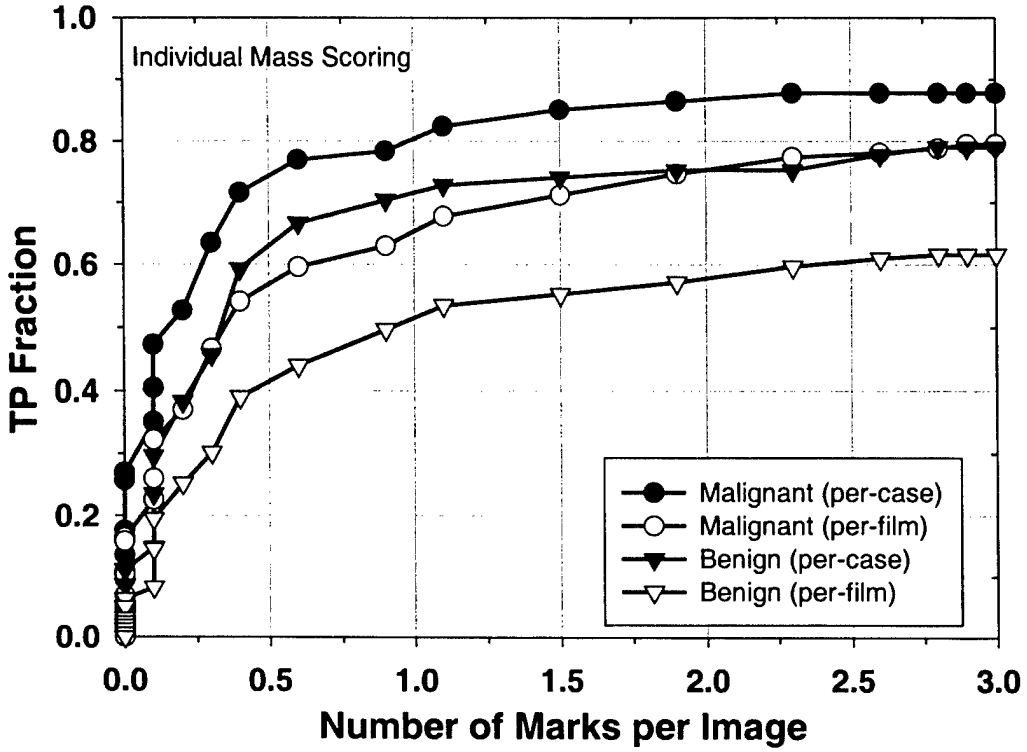


Fig 3: The Group 1 FROC performance curves for malignant and benign masses. The per-case curve is obtained by defining a TP as the detection of the mass in either the CC or MLO view mammogram of a breast. The per-film curve treats the same mass on the CC and MLO films independently. Individual mass scoring is used in the figure so each individual mass in a film or case was considered as a TP.

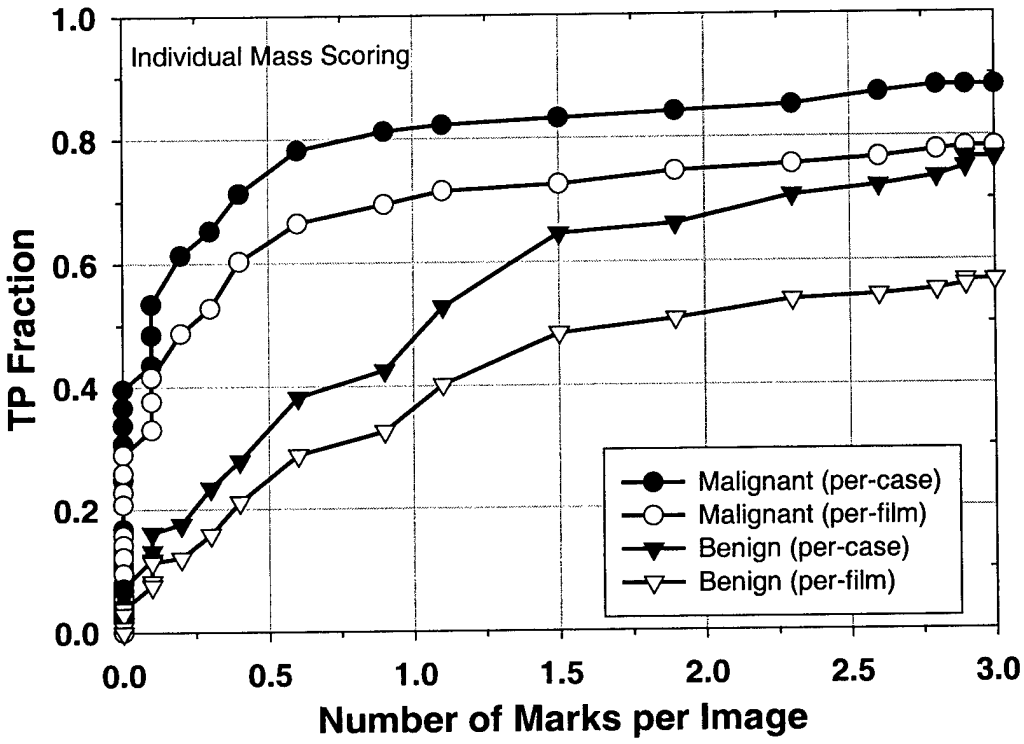


Fig 4: The Group 2 FROC performance curves for malignant and benign masses. Individual mass scoring is used in the figure so each individual mass in a film or case was considered as a TP.

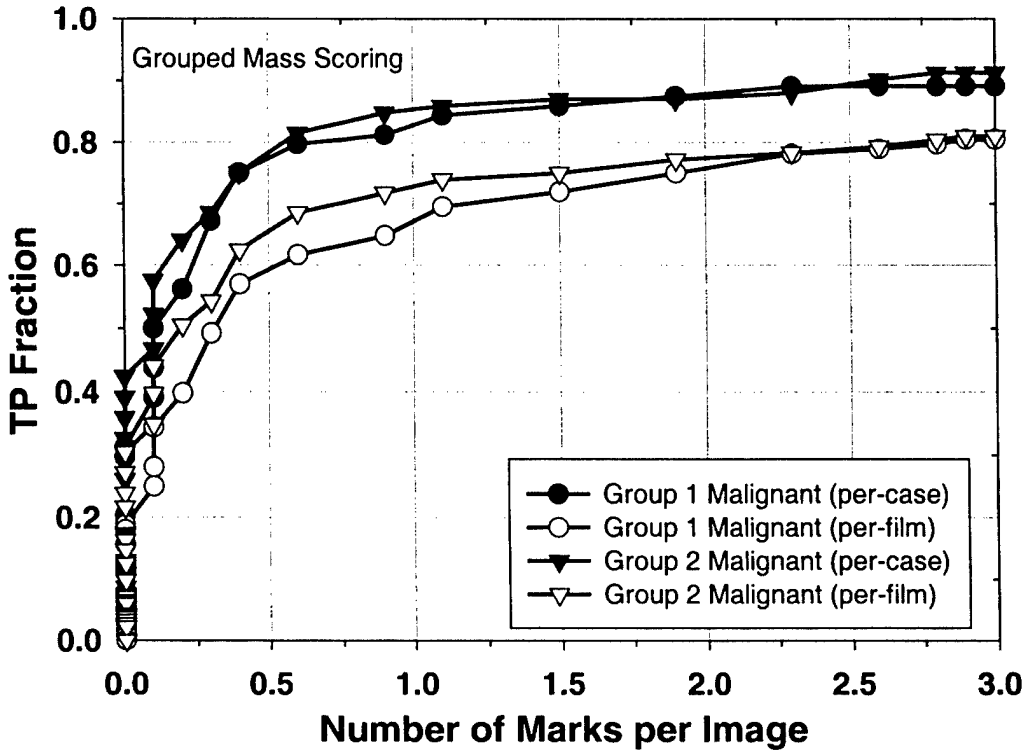


Fig 5: The Group 1 and 2 FROC performance curves for malignant masses. Grouped mass scoring is used in the figure so all malignant masses in a film or case were considered together as one TP.

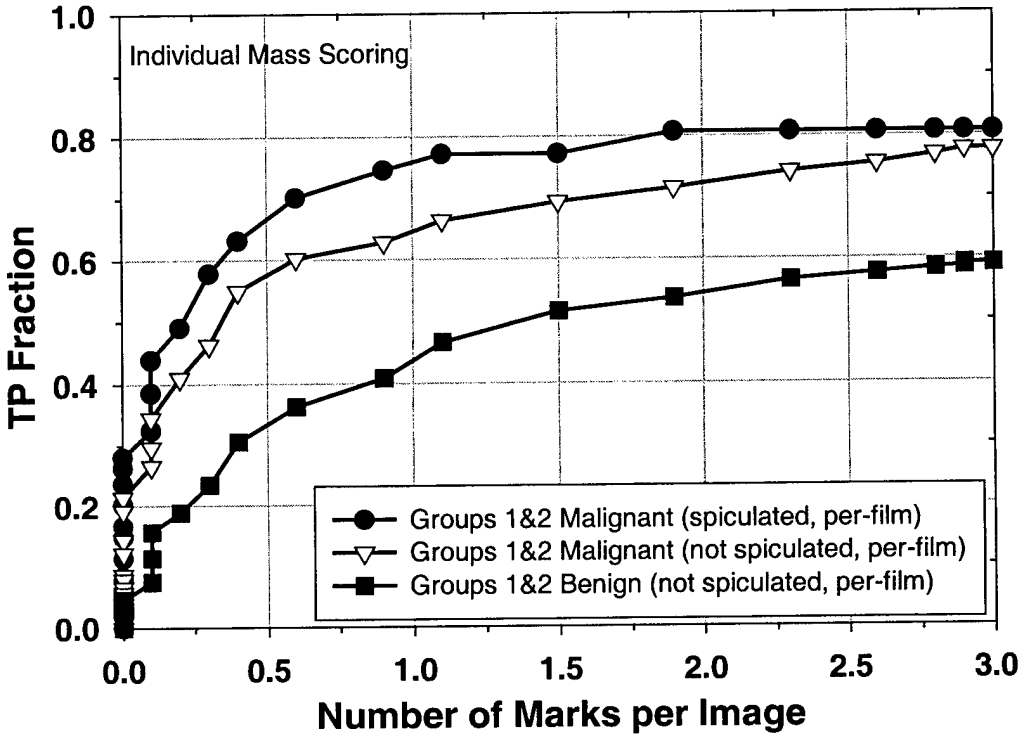


Fig 6: The combined Group 1 and 2 FROC performance curves for spiculated and non-spiculated masses. The benign spiculated mass curve is not shown because of the small number of cases in this category. Individual mass scoring is used in the figure so each individual mass in a film was considered as a TP.

Accepted by Medical Physics

Improvement of Computerized Mass Detection on Mammograms: Fusion of Two-View Information

Sophie Paquerault

Nicholas Petrick

Heang-Ping Chan

Berkman Sahiner

Mark A. Helvie

Department of Radiology

University of Michigan, Ann Arbor, MI.

Correspondence:

Sophie Paquerault, Ph.D.
c/o Heang-Ping Chan, Ph.D.
Department of Radiology
University of Michigan
1500 E. Medical Center Drive
B1F510B
Ann Arbor, MI 48109-0030
Telephone: (734) 936-4357
Fax: (734) 615-5513
Email: chanhp@umich.edu

ABSTRACT

Recent clinical studies have proved that computer-aided diagnosis (CAD) systems are helpful for improving lesion detection by radiologists in mammography. However, these systems would be more useful if the false-positive rate is reduced. Current CAD systems generally detect and characterize suspicious abnormal structures in individual mammographic images. Clinical experiences by radiologists indicate that screening with two mammographic views improves the detection accuracy of abnormalities in the breast. It is expected that fusion of information from different mammographic views will improve the performance of CAD systems. We are developing a two-view matching method that utilizes the geometric locations, and morphological and textural features to correlate objects detected in two different views using a prescreening program. First, a geometrical model is used to predict the search region for an object in a second view from its location in the first view. The distance between the object and the nipple is used to define the search area. After pairing the objects in two views, textural and morphological characteristics of the paired objects are merged and similarity measures are defined. Linear discriminant analysis is then employed to classify each object pair as a true or false mass pair. The resulting object correspondence score is combined with its one-view detection score using a fusion scheme. The fusion information was found to improve the lesion detectability and reduce the number of FPs. In a preliminary study, we used a data set of 169 pairs of cranio-caudal (CC) and mediolateral oblique (MLO) view mammograms. For the detection of malignant masses on current mammograms, the film-based detection sensitivity was found to improve from 62% with a one-view detection scheme to 73% with the new two-view scheme, at a false-positive rate of 1 FP/image. The corresponding cased-based detection sensitivity improved from 77% to 91%.

Keywords: computer-aided diagnosis, mammography, mass detection, classification, fusion of information.

I. INTRODUCTION

X-ray mammography is the only proven diagnostic technique for detecting breast cancer in its early stages.[1, 2] In mammographic screening, a cranio-caudal (CC) and a mediolateral oblique (MLO) view are routinely taken for each breast. The two views not only allow most of the breast tissue to be imaged but also improve the chance that a lesion will be seen in at least one of the views. Radiologists analyze the different mammographic views to detect calcifications and masses that may be a sign of breast cancer and to decide whether to call the patient back for further diagnostic evaluations. They also use the two views to reduce false positives such as overlapping dense tissue in one view that mimics masses. Their interpretation integrates complex criteria of human vision and intelligence, including morphology, texture, and geometric location of any suspicious structures of the imaged breast, combining information from different views, checking differences between the two breasts, and looking for changes between the prior and current mammograms when available. Clinical studies indicate that lesion detectability in two-view mammograms is more accurate than when only one view is available.[3,4,5]

It has also been shown that, independent double reading by two radiologists significantly increases the sensitivity of mammographic screening.[6, 7] However, the increased cost and workload to the radiologists make double reading impractical in most screening situations. To provide a second opinion to the radiologists, computer-aided diagnosis (CAD) systems have been developed using computer vision and pattern recognition techniques to automatically detect and characterize abnormal lesions on mammograms. Although it has been reported that these systems are useful in reducing the error rate in mammographic screening,[8,9,10] the detection sensitivity of these systems needs to be improved and the false-positive (FP) rate reduced to provide maximum benefit to the radiologist and the patient. CAD algorithms reported in the literature so far use one-view information for detection of lesions even though the accuracy may be scored and reported using two views. Yin et al.[11] used bilateral subtraction in a prescreening step of a mass detection program to locate mass candidates, but the subsequent image analysis was performed based only on a single view. Recently, Hadjiiski et al.[12, 13] have developed an interval change analysis of masses on current and prior mammograms and found that the classification accuracy of malignant and benign masses can be improved significantly in comparison to single image classification. These

studies demonstrated the potential of using multiple image information for CAD. However, current CAD algorithms have not utilized one of the most important pieces of information available in a mammographic examination - the correlation of computer-detected lesions between the two standard views. This is a very difficult problem for computer vision because the breast is elastic and deformable. The overlapping tissue and the relative position of the breast structures are generally different even when the breast is compressed in the same view two different times. The change in geometry for an elastic object and lack of invariant "landmarks" make difficult, if not impossible, to correctly register two breast images in the same view by any established image warping technique or by using an analytic model to predict corresponding object locations in the different views of the same breast.

Few studies have been conducted on how to find the relationship between structures in different mammographic views. Highnam et al.[14] proposed a breast deformation model for compressed breasts and Kita et al.[15] used the model for finding corresponding points in two different views. They demonstrated with a data set of 26 cases (a total of 37 lesions) that this method allowed prediction of location in a second view within a band of pixels ±26 mm from an epipolar line. However, assumptions on the parameters and the deformation of a compressed breast had to be made and the robustness of the model has yet to be validated. More practical approaches, which do not depend on a large number of assumptions, may be preferable. Good et al. and Chang et al. recently reported preliminary attempt of matching computer-detected objects in two views.[16, 17] They demonstrated the feasibility of identifying corresponding objects (A_z=0.82) in the two views by exhaustive pairing of the detected objects and feature classification. None of these studies attempted to use the two-view correspondence information to improve lesion detection or classification.

During mammographic interpretation, if a suspicious breast mass is found in one view, the radiologist will attempt to find the same object in the other available views in order to identify the object as a true or a false mass. Radiologists commonly consider the distance from the nipple to the center of the suspicious lesion in one view and then search the corresponding object in the second view in an annular region at about the same radial distance from the nipple. Based on this approach, we previously developed a regional registration technique to identify corresponding lesion locations on current and prior mammograms of the same view.[18,19,13] We have also designed geometric models that can localize corresponding

lesions within a search region when two-view or three-view mammograms are available for lesion localization.[20] With the geometric information, the computer searches for a corresponding lesion in the other view within a limited search region. The object of interest can then be matched with possible corresponding objects in the search region using the similarity of feature measures. We have found that the geometric constraints improved the chance of correctly matching lesions in current and prior mammograms for classification of malignant and benign masses.[21] In this study, we explore the use of the regional registration technique as a basis to correlate lesions in the two views. The correspondence information is used to reduce false detections produced by our one-view CAD algorithm. The detection accuracy of the two-view scheme was evaluated and compared to our current one-view CAD scheme using free response receiver operating characteristic (FROC) analysis.

II. MATERIALS AND METHODS

Our approach to improving the accuracy of the mass detection is to merge information from corresponding segmented structures in the two standard views of the same breast. We first assume that a true mass will have a higher chance of being detected in both views. Likewise, we assume that the objects corresponding to the same mass detected in the two different views (a TP-TP pair) will be more similar in their feature measures than a mass object compared to normal tissue (a TP-FP pair), or two false-positives (an FP-FP pair). Object matching is performed in two stages. First, all possible pairing of the detected objects on the two views are determined, taking into account geometric constraints. Second, features are extracted from each object, similarity measures for the features pairs are derived, and a classifier is trained to classify true pairs (TP-TP pairs) from false pairs (TP-FP, FP-TP or FP-FP pairs) using the similarity measures. The two stages are detailed below. The data sets used in the development and evaluation of this approach are described next.

A. Image acquisition and data set

Two data sets of two-view mammograms were collected and separately used to train and test the geometric models and our proposed two-stage information fusion technique. These mammograms were selected from patient files in the Breast Imaging Division at the University of Michigan.

For the geometric modeling of object location on two views, the database consisted of 116 cases with masses, large benign calcifications, or clustered microcalcifications identifiable on both views of the same breast. The mammograms were digitized with a LUMISYS 85 film scanner with a pixel size of 50 μm and 12-bit gray levels. The gray levels were calibrated to be linearly proportional to optical density in the 0.1 to 4.0 O.D. range. The images were reduced to a pixel resolution of 800 μm × 800 μm by averaging 16 × 16 neighboring pixels and down-sampling. For each case, the two standard mammographic views were available. A total of 177 objects were manually selected and marked by an expert radiologist on each of these two views. The nipple location was also identified for each breast image. The radial distance of the selected objects was calculated and the prediction model of an object location in one view from its location in the other view was estimated, as described above.

For the evaluation of the two-view mass detection scheme, a data set of 169 pairs of mammograms containing masses on both the CC and MLO views was used. The mammograms were obtained from 117 patients, of which 128 pairs were current mammograms (defined as mammograms from the exam before biopsy) and 41 pairs were from exams 1 to 4 years prior to biopsy. 58 of the 128 current and 26 of the 41 prior image pairs contained a malignant mass. The 338 mammograms were also digitized with the LUMISYS 85 film scanner. The true mass locations on both views were identified and rated by a radiologist approved by the Mammography Quality Standards Act (MQSA). The histograms of the size (longest dimension) and the visibility (subtlety) rating of the benign and malignant masses contained in this data set are shown in Fig. 1 and 2, respectively. The subtlety of the masses was estimated subjectively on a 10-point scale by the experienced radiologist relative to the masses encountered in clinical practice.

B. Geometrical Modeling

We will first describe the geometric models that we developed for predicting the location of an object in the MLO view from that in the CC view or vice versa. For the purpose of studying the geometric relationship between the locations of an object imaged on the two mammographic views, any identifiable objects can be used. We therefore chose two-view mammograms that contained masses, microcalcification clusters, and large benign calcifications identifiable on both views. This data set was different from that used for mass detection to be described below. The locations of the corresponding objects on the two views

and the nipple locations were identified on the mammograms by the MQSA-approved radiologist. For a large object such as a mass or a microcalcification cluster, the manually identified "centroid" was taken as its location. A breast boundary tracking program was used to segment the breast area from the mammogram.[22, 23] Using the nipple location as the origin, concentric circles were drawn, each of which intersected the breast boundary at two points and defined an arc. The locus of the mid-points of these arcs was considered to be the breast midline. The breast length was defined as the distance from the nipple to the point where the midline intersected the chest wall. From these parameters, the polar coordinates (R_x, θ_x) with x = C (CC view), or M (MLO view), as shown in Fig. 3, were defined, where R_x was the distance from the nipple to the object center and θ_x, the angle between R_x and the line from the nipple to the mid-point of the arc intersecting the object. We investigated the relationship between the coordinate of the object on one view and that on the other view in this coordinate system.

Scatter plots of the radial distance and the angle of the radiologist-identified objects on the two views in the data set are shown in Fig. 4 and Fig. 5, respectively. It can be seen that there is a high correlation (correlation coefficient=0.94) of the radial distances of the corresponding objects in the two views. However, the angular coordinates in the two views are basically uncorrelated (correlation coefficient=0.42). We therefore chose a linear model for predicting the radial distance of an object in a second view from that in the first view:

$$R_y = a_r \cdot R_x + b_r \tag{1}$$

Because of the variability of the breast tissue caused by compression, the predicted location for an individual case could deviate from its "true" location, as determined by the radiologist, by a wide range. Therefore, we estimated a global model using a set of training cases with radiologist-identified object locations on both views. The model coefficients were obtained by minimizing the mean square error between the true and the predicted coordinates in the second view. The error in this estimation was then used to define an annular search region, which had a center at a radial distance R_y from the nipple as predicted by the model, and a width of $\pm \Delta R$ as estimated from the localization errors observed in the training set. This search region avoids using the entire area of the breast and eliminates many inappropriate pairings between detected objects on the CC view and the MLO view in the second stage, discussed in Section II.D below.

We randomly divided the available data set into a training set and a test set in a 3:1 ratio. The training set was used for the estimation of the model coefficients and the search region width. The test set was used for evaluating the prediction accuracy of the model. Four non-overlapping partitions separating the database into training and test sets were considered. The model performance was then obtained by combining the results of the four test sets.

The geometrical analysis is then used for pairing objects detected on the two views of the same breast in the prescreening stage of our mass detection program as detailed below.

C. One-view analysis

The one-view approach is used to identify potential breast masses among the suspicious objects. The one-view prescreening used in this study is similar to that discussed previously.[24, 25, 26] The only difference is that the false positive (FP) reduction step was modified such that a slightly different object overlap criterion was employed. The block diagram for the one-view mass detection scheme is shown in Fig. 6. A density-weighted contrast-enhancement (DWCE) filter is first applied to each digitized mammogram. The DWCE filter enhances mammographic structures in the breast image. Following this preprocessing filtering, edge detection is employed to refine the borders of the detected regions. K-means clustering is then applied to a 25 mm × 25 mm, background-corrected region of interest centered on each initially detected object to improve the object border. This segmentation process extracts a large number of objects, including masses and normal breast structures. In order to reduce the number of non-mass objects, different FP reduction stages based on morphological features, overlap of the detected regions, and texture features were designed and trained using an independent set of mammograms in a previous study.[26, 27] It was found that 11 morphological features composed of shape descriptors and 15 spatial gray level dependence (SGLD) texture features extracted for each object were useful for FP reduction.[28, 29] In this study, rule-based classification using the 11 morphological features reduced the average number of objects from 37 to about 29 per image and lowered the TP detection sensitivity from 91.1% to 87.9% at this stage. The 15 texture features were then used as the input variables for a linear discriminant analysis (LDA) classifier. A texture score

for each object was obtained from the classifier. Overlap reduction was then applied using these texture scores as discussed below.

During object segmentation, the border of an object is obtained by K-means clustering in a fixed sized region centered on a "seed" object. If the seeds from two objects are close to each other, the two segmented objects can overlap each other. This occurs when the two detected objects are neighboring structures that overlap in the mammographic view or they may be part of a large single structure that was initially detected in multiple pieces. An overlap criterion based on the texture scores is imposed to select one of the two overlapping objects as a mass candidate. In this study, we used the shape of the segmented objects to estimate the overlapping area between the two neighboring objects on the mammogram. An overlap fraction was defined as:

$$Overlap = \frac{O_1 \cap O_2}{O_1 \cup O_2}$$

where O_1 and O_2 are the segmented areas of the overlapping objects. A threshold on the overlap fraction was chosen such that if the overlap fraction of two objects exceeded the threshold, the object with the higher texture score (i.e., more likely to be a mass candidate) was kept and the other was discarded as an FP. The sensitivity and the specificity of differentiating true and false masses depend on the selection of the overlap threshold. We chose an overlap threshold of 15% which led to an average of 15 objects per image at a detection sensitivity of about 85%. As shown later in the Results section, the overall detection accuracy was relatively independent of the FP rate in this intermediate stage so that the selection of the 15% overlap threshold was not a critical factor.

After overlap reduction, our current one-view algorithm employed a final stage of FP reduction based on the texture scores, as illustrated in the block diagram in Fig. 7. A decision threshold was applied to the texture scores such that objects with scores lower than the threshold were excluded as FPs. In addition, another criterion was imposed so that no more than three objects were kept on each image. By comparing the retained objects with the true mass locations on each image for a range of decision thresholds, an FROC curve characterizing the sensitivity as a function of the number FPs per image could be generated.

D. Two-view analysis

The block diagram in Fig. 7 illustrates our two-view mass detection scheme and its relationship to our current one-view approach. The detection algorithm described above was used as a prescreening stage in our two-view fusion approach. The only difference was that the operating threshold that limits the maximum number of objects on an image was relaxed to increase sensitivity while retaining a larger number of FPs. The remaining objects after this threshold will be still referred to as the prescreening objects in the following discussions. To investigate the dependence of the overall detection accuracy of our two-view detection scheme on the initial number of prescreening objects, three different decision thresholds were selected to obtain a maximum of either 5, 10, or 15 objects per image.

To further perform the two-view information fusion analysis, an expanded set of morphological features was extracted from each prescreening object. These morphological features included the 11 shape descriptors discussed previously, and 13 new contrast measures [30] and 7 new shape features. In order to evaluate the new method, we randomly divided the available cases into a training and a test set using a 3:1 training/test ratio. The training set was used to select a subset of useful morphological features using stepwise feature selection and to estimate the coefficients of an LDA classifier. To reduce biases in the classifier, 50 random 3:1 partitions of the cases were employed. A morphological score was obtained for each individual object by averaging the test score of the object obtained from the different partitions. The morphological score was then combined with the one-view texture score by averaging the two scores. A single combined score thus characterized each prescreening object. This one-view score was further fused with the discriminant score obtained by the two-view scheme, as described below.

The prescreening objects wereanalyzed by the two-view method shown in the right branch of the diagram in Fig. 7. All possible pairing between the prescreening objects in the two views of the same breast was determined using the distance from the nipple to the centroid of each object and the geometrical model described above. Since the location of a given object detected in one view cannot be uniquely identified in the other view, as described in Section II.B, an object was initially paired with all objects with centroids located within its defined annular region in the other view. The geometric constraints reduced the number of object pairs that needed to be classified as true or false correspondences in the subsequent steps. A true pair (TP-TP) was defined as the correspondence between the same true masses on the two

mammographic views, and a false pair is defined as any other object pairing (TP-FP, FP-TP and FP-FP). For each object pair, the set of 15 texture and 31 morphological features described above were used to form similarity measures. In this preliminary study, two simple measures, the absolute difference and the mean, were used. A total of 30 texture measures and 62 morphological measures were thus obtained for each object pair. The absolute difference between the nipple-to-object distances in the CC and MLO views was also included in both the texture and morphological feature sets as a feature for differentiating true from false object pairs. Two separate LDA classifiers with stepwise feature selection were trained to classify the true and false pairs using the similarity features in the morphological and texture feature spaces, respectively.

For training the classifiers, the data set was randomly divided into a training set and a test set again using a 3:1 training/test ratio. Fifty random 3:1 partitions of the cases were used to reduce bias. Individual morphological and texture scores were obtained for each object pair by averaging the test scores of each object pair obtained from the different partitionings. The two classification scores were then averaged to obtain one "correspondence" score for each object pair. This score along with the one-view prescreening score were used in the following fusion step.

E. Fusion analysis

The fusion of the one-view prescreening scores with the two-view correspondence scores was the final step in our two-view detection scheme. In this study, we designed a fusion scheme that combines ranking and averaging of the prescreening and correspondence scores. We first ranked all prescreening object scores within a given film from the largest to the smallest. The correspondence scores were ranked in a similar way. These two new rank scores were then merged into a single score for each object in each view. Since an object could have more than one correspondence score, its two-view correspondence score was taken to be the maximum correspondence score among all object pairs in which this object was a member. There can be many variations for the fusion step.[31, 32] In this preliminary study, the final discriminant score for an object was obtained by averaging its two-view correspondence score rank with its one-view prescreening score rank.

The FROC performance curve for the two-view analysis was generated by varying the decision threshold on the final discriminant score for each object and determining the sensitivity and FP per image at each threshold. . We compared the FROC performance curves obtained by the two-view scheme when starting with 5, 10, and 15 prescreening objects per image and that obtained with the one-view detection scheme.

III. RESULTS

A. Geometrical Modeling

In the geometrical analysis experiments, we first estimated a prediction model of the radial distance of an object in a second view from its radial distance in the first view using the training set. The model was then used to predict object location from one view to the other for the independent test cases. Since the model did not provide an exact solution, a search region, $R \pm \Delta R$, where R was the predicted radial distance and ΔR the half width of an annular region, was defined. The percentage of the true object centroids enclosed within the search region was measured as a function of the size of $2\Delta R$. Fig. 8 shows the prediction accuracy as a function of $2\Delta R$ for estimating the object radial distance in the MLO view from that in the CC view. Fig. 9 shows the corresponding results for predicting the object radial distance in the CC view from that in the MLO view. The training and test curves almost overlap in each case. The difference in the accuracy between searching the object centers in the CC or MLO views is small. About 83% of the object centers are within the search region when the radial width of the search region is about 40 pixels (32 mm) for either the CC view or the MLO view. These results indicate that the search region, although large, is much smaller than the entire area of the breast. The limited search region size reduces the number of object pairs to be analyzed in the two-view detection scheme. To avoid missing any pairs of true masses in the two-view scheme, we chose to set the radial width of the annular search region to about 80 pixels. This led to a larger number of false pairs, but it was substantially less than that if the entire breast area was considered.

B. One-View Analysis

The FROC curve obtained from our current one-view mass detection algorithm [26] applied to the data set of 338 images is shown in Fig. 10. The FROC curves for detection of the malignant masses on the current and prior mammograms are also plotted for comparison.

In clinical application, if the mass is detected on one-view by the computer and the radiologist is alerted to the mass, the radiologist will likely find the mass on the other view, if it is visible, even if the CAD algorithm misses it on the other view. Some researchers therefore consider a true-positive as the detection of the mass on one or two views of the breast. We refer to this as case-based analysis. In this situation, the total number of masses or cases in this study was 169. For comparison purposes, we plot the case-based FROC curves for all masses, malignant masses on current mammograms, and malignant masses on prior mammograms in Fig. 11.

C. Fusion Analysis

Three different decision thresholds that retained a maximum of 5, 10, and 15 objects per image after the one-view prescreening stage were used to select mass candidates as inputs to the two-view detection scheme. Table I summarizes the characteristics of these three object sets. The average number of prescreening objects per image was smaller than the maximum number allowed per image because the total number of objects in some images was smaller than the maximum number.

The FROC curves for the detection of malignant and benign masses on each image, using our two-view fusion technique, are shown in Fig. 12. The curves are similar for the three thresholds of 5, 10, 15 prescreening objects per image. This similarity also holds for the FROC curves for detection of malignant masses as illustrated in Fig. 13. The improvement in detection by our current two-view fusion method therefore seems to be independent of the operating threshold when the maximum number of objects retained per image in the prescreening stage is between 5 and 15.

Fig. 14 compares the film-based FROC curves for detection of malignant masses by the one-view and two-view fusion methods obtained from the condition of 10 prescreening objects per image. Fig. 15 compares the corresponding case-based FROC curves. A comparison of the detection sensitivity at 1 FP/image between the one-view and two-view fusion methods is given in Table II for both film-based and case-based detection.

IV. DISCUSSION

In this work, we propose a new technique based on fusion of one-view and two-view information to improve the performance of mammographic mass detection. The results of our preliminary study show that including correspondence information from two mammographic views is an effective technique for reducing FPs. At a case-based detection sensitivity of 75% for all masses, the number of FPs per image was reduced from 1.5 FPs/image using the one-view detection technique to 1.13 FPs/image using the two-view fusion technique. The results also indicate that our proposed method is more effective in reducing FPs in the subset of cases containing malignant masses on current mammograms. At a case-based sensitivity of 85% for malignant masses on current mammograms, the number of FPs per image was reduced from 1.5 FPs/image to 0.5 FPs/image using the two-view fusion technique (Fig. 15). Alternatively, at 1 FPs/image, the two-view algorithm achieved a case-based detection sensitivity of 91% whereas the current one-view scheme had a 77% sensitivity at the same number of FPs/image (Table II).

The two-view correspondence analysis is more useful for mammogram pairs for which the mass is detected on both views in the prescreening stage. The fusion process is designed to both increase the scores for the TPs and reduce the scores for FPs for such cases. For the data set of 169 pairs of mammograms under the condition of 10 prescreening objects per image, the mass was detected on both CC and MLO views in a subset of 120 cases and on only one view in another subset of 32 cases. If we analyzed the subset of cases in which the mass was detected in both views, at 1 FP/image, the case-based detection sensitivity increased from 82.5% for the current one-view algorithm to 93.3% using the two-view fusion technique. However, for the subset of cases in which the mass was detected on only one view at the prescreening stage, the fusion analysis reduced the scores for TPs. At 1 FP/image, the case-based detection sensitivity was reduced from 50% for the current one-view algorithm to 43.7% using the two-view fusion process. Similar trends for the detection results were observed when 5 and 15 objects per image were retained in the prescreening stage.

In this study, we chose the radial width of the annular search region to be 80 pixels for all mammograms. This radial width reduced the search region to only a fraction of the breast

area for large breasts but it covered most of the breast area in smaller breasts. Therefore, the advantage of geometric correlation has not been fully utilized in small breasts. One approach to reducing the search region size for small breasts would be to choose the region size as a percentage of the breast area so that the actual width of the annular region will be different for each pair of mammograms. This will lead to a reduction in the number of false object pairs for small breasts. The second approach would be to use a third mammographic view when it is available. As we discussed previously[20], using the three standard views (CC, MLO, and Lateral) of the breast allow more accurate localization of a lesion to within a small fan-shaped region. This approach would require further adaptation of our two-view scheme to a three-view fusion scheme. Although 3-view mammograms are not generally available for screening, it will be of interest to investigate how 3-view mammograms will improve the detection of malignancy in the breast by the computer.

In this study, we used radiologist-identified nipple locations for the geometric correlation process. In a fully automated mass detection program, this step will have to be automated. We are developing an automated nipple detection program. This detection program could identify the nipple within 1 cm of the true location in 88% of the 311 mammograms in a study set.[23] For the purpose of this study, we did not use automated nipple detection because it will complicate our analysis of the two-view fusion techniques if errors in nipple detection have to be taken into account. We therefore isolated the latter effects by using manually identified nipple locations. We will continue to improve the automated nipple detection algorithm and incorporate this step into the two-view mass detection scheme in the future.

In this preliminary study, we used two simple similarity measures for classification of object correspondence. The fusion of the two-view and one-view scores for the individual objects was performed with a relatively simple ranking and averaging methods. These approaches already provided substantial improvement in the detection accuracy, indicating the promise of the two-view method for mass detection and FP reduction. Further studies are being conducted to optimize the various steps in the two-view classification and fusion schemes.

V. CONCLUSION

We are developing a two-view fusion technique to improve computerized mass detection on mammograms. Starting from objects detected in a prescreening stage, we defined all possible pairing based on geometry and then combined morphological and textural characteristics from these paired objects into a correspondence score for each object. A classifier was trained to differentiate the true mass pairs from the false pairs. A final fusion stage combined the two-view object pair information with the one-view object scores. Our preliminary results demonstrate that the proposed two-view scheme can reduce FPs in comparison with our current one-view method. The mass detection sensitivity is also improved by using information from the two-views. Further studies are underway to optimize the pre-screening process, the design of the similarity measures, as well as the two-view fusion scheme. When fully developed and integrated into the CAD system, it is expected that our proposed two-view technique will improve upon the current one-view scheme and provide a useful second opinion to radiologists in the detection of breast cancer on mammograms.

ACKNOWLEDGEMENTS

This work is supported by USPHS grant CA 48129, USAMRMC grant DAMD 17-96-1-6254, and a Career Development Award (B.S.) from the USAMRMC DAMD 17-96-1-6012. The content of this publication does not necessarily reflect the position of the government and no official endorsement of any equipment and product of any companies mentioned in the publication should be inferred.

REFERENCES

[1]L. Tabar, C. Fagerberg, A. Gad, L. Baldetorp, L. Holmberg, O. Grontoft, U. Ljungquist, B. Lundstrom, J. Manson, G. Eklund, et al., "Reduction in mortality from breast cancer after mass screening with mammography," Lancet 1, 829-832 (1985).

[2]H. C. Zuckerman, *The role of mammography in the diagnosis of breast cancer. In: Breast Cancer, Diagnosis and Treatment*, I. M. Ariel and J. B. Cleary, (McGraw-Hill, New York, 1987).

[3]L. W. Bassett, D. H. Bunnell, R. Jahanshahi, R. H. Gold, R. D. Arndt, and J. Linsman, "Breast cancer detection: one versus two views," Radiology 165, 95-97 (1987).

[4]E. Thurfjell, "Mammography screening: One versus two views and independent double reading," Acta Radiologica 35, 345-50 (1994).

[5]R. G. Blanks, M. G. Wallis, and R. M. Given-Wilson, "Observer variability in cancer detection during routine repeat (incident) mammographic screening in a study of two versus one view mammography," J. Medical Screening 6, 152-158 (1999).

[6]E. D. C. Anderson, B. B. Muir, J. S. Walsh, and A. E. Kirkpatrick, "The efficacy of double reding mammograms in breast screening.," Clin Radiol 49, 248-251 (1994).

[7]E. L. Thurfjell, K. A. Lernevall, and A. A. S. Taube, "Benefit of independent double reading in a population-based mammography screening program," Radiology 191, 241-244 (1994).

[8]H. P. Chan, K. Doi, C. J. Vyborny, R. A. Schmidt, C. E. Metz, K. L. Lam, T. Ogura, Y. Wu, and H. MacMahon, "Improvement in radiologists' detection of clustered microcalcifications on mammograms. The potential of computer-aided diagnosis," Investigative Radiology 25, 1102-1110 (1990).

[9]W. P. Kegelmeyer, J. M. Pruneda, P. D. Bourland, A. Hillis, M. W. Riggs, and M. L. Nipper, "Computer-aided mammographic screening for spiculated lesions," Radiology 191, 331-337 (1994).

[10]L. J. Warren Burhenne, S. A. Wood, C. J. D'Orsi, S. A. Feig, D. B. Kopans, K. F. O'Shaughnessy, E. A. Sickles, L. Tabar, C. J. Vyborny, and R. A. Castellino, "Potential contribution of computer-aided detection to the sensitivity of screening mammography," Radiology 215, 554-562 (2000).

[11]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 18, 955-963 (1991).

[12]L. M. Hadjiiski, B. Sahiner, H. P. Chan, N. Petrick, M. A. Helvie, and M. Gurcan, "Analysis of temporal change of mammographic features for computer-aided characterization of malignant and benign masses," Proc. SPIE 4322, 661-666 (2001).

[13]L. M. Hadjiiski, H. P. Chan, B. Sahiner, N. Petrick, and M. A. Helvie, "Automated registration of breast lesions in temporal pairs of mammograms for interval change analysis - local affine transformation for improved localization," Medical Physics 28, 1070-1079 (2001).

[14]R. P. Highnam, Y. Kita, J. M. Brady, B. J. Shepstone, and R. English, "Determining correspondence between views.," 4th International Workshop on Digital Mammography, Nijmegen, Netherlands, June 1998, Digital Mammography, Kluwer Academic Publisher,

[15]Y. Kita, R. P. Highnam, and J. M. Brady, "Correspondence between two different views of x-ray mammograms using simulation of breast deformation," Proc. IEEE Conference on Computer Vision and Pattern Recognition, Santa Barbara, CA, June 1998, 700-707.

[16]W. F. Good, B. Zheng, Y. H. Chang, Z. H. Wang, G. S. Maitz, and D. Gur, "Multi-image CAD employing features derived from ipsilateral mammographic views," Proceedings of the SPIE - Medical Imaging 3661, 474-485 (1999).

[17]Y. H. Chang, W. F. Good, J. H. Sumkin, B. Zheng, and D. Gur, "Computerized localization of breast lesions from two views - An experimental comparison of two methods," Investigative Radiology 34, 585-588 (1999).

[18]S. S. Gopal, H.-P. Chan, B. Sahiner, N. Petrick, T. E. Wilson, and M. A. Helvie, "Evaluation of interval change in mammographic features for computerized classification of malignant and benign masses," Radiology 205(P), 216 (1997).

[19]L. M. Hadjiiski, H. P. Chan, B. Sahiner, N. Petrick, M. A. Helvie, and S. Sanjay-Gopal, "Automated identification of breast lesions in temporal pairs of mammograms for interval change analysis," Radiology 213(P), 229-230 (1999).

[20]S. Paquerault, B. Sahiner, N. Petrick, L. M. Hadjiiski, M. N. Gurcan, C. Zhou, and M. A. Helvie, "Prediction of object location in different views using geometrical models," The 5th International Workshop on Digital Mammography, Toronto, Canada, June 11-14, 2000, Proc IWDM-2000, Madison, WI: Medical Physics Publishing, 748-755.

[21]L. M. Hadjiiski, B. Sahiner, H. P. Chan, N. Petrick, M. A. Helvie, and M. Gurcan, "Computer-aided classification of malignant and benign breast masses by analysis of interval change of features in temporal pairs of mammograms," Radiology 217(P), 435 (2000).

[22]A. R. Morton, H. P. Chan, and M. M. Goodsitt, "Automated model-guided breast segmentation algorithm," Medical Physics 23, 1107-1108 (1996).

[23]C. Zhou, H. P. Chan, N. Petrick, M. M. Goodsitt, C. Paramagul, and L. M. Hadjiiski, "Computerized image analysis: breast segmentation and nipple identification on mammograms," Chicago 2000-World Congress on Medical Physics and Biomedical Engineering., Chicago, Illinois, July 23-28, 2000,

[24]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," Medical Physics 23, 1685-1696 (1996).

[25]N. Petrick, H. P. Chan, B. Sahiner, and D. Wei, "An adaptive density-weighted contrast enhancement filter for mammographic breast mass detection," IEEE Transactions on Medical Imaging 15, 59-67 (1996).

[26]N. Petrick, H. P. Chan, B. Sahiner, and M. A. Helvie, "Combined adaptive enhancement and region-growing segmentation of breast masses on digitized mammograms," Medical Physics 26, 1642-1654 (1999).

[27]N. Petrick, B. Sahiner, H. P. Chan, M. A. Helvie, and S. Paquerault, "Preclinical evaluation of a CAD algorithm for early detection of breast cancer," The 5th International Workshop on Digital Mammography, Toronto, Canada, June 11-14, 2000, Proc. IWDM-2000, Madison, WI: Medical Physics Publishing, 328-333.

[28]R. M. Haralick, K. Shanmugam, and I. Dinstein, "Texture features for image classification," IEEE Transactions on Systems, Man, and Cybernetics SMC-3, 610-621 (1973).

[29]R. M. Haralick, *Chaper 11. Statistical Image Texture Analysis. In: Handbook of Pattern Recognition and Image Processing*, (Academic Press, New York, 1986).

[30]G. M. te Brake, N. Karssemeijer, and J. H. C. L. Hendriks, "An automatic method to discriminate malignant masses from normal tissue in digital mammograms," Physics in Medicine and Biology 45, 2843-2857 (2000).

[31]A. Kandel, *Fuzzy Techniques in Pattern Recognition*, (John Wiley and Sons, New York, 1982).

[32]B. Sahiner, N. Petrick, H. P. Chan, S. Paquerault, M. A. Helvie, and L. M. Hadjiiski, "Recognition of lesion correspondence on two mammographic views - A new method of false-positive reduction for computerized mass detection," Proc . SPIE 4322, 649-655 (2001).

FIGURE CAPTIONS

Fig. 1: Histograms of the size (the longest dimension) of the benign and malignant masses contained in the data set of 338 one-view mammograms and rated by an MSQA-radiologist. Eight masses in the prior mammograms of the data set did not receive a rating because the radiologist could not delineated the mass even in retrospect, although a focal density could be seen.

Fig. 2: Histograms of the visibility (1= most obvious, 10=subtlest) of the benign and malignant masses contained in the data set of 338 one-view mammograms and rated by an MSQA-radioliogist. Eight masses in the prior mammograms of the data set did not receive a rating because the radiologist could not delineated the mass even in retrospect, although a focal density could be seen.

Fig. 3: Example of the coordinate system used to localize an object in a mammographic view. An automatic boundary tracking process is used to segment the breast. The nipple location was identified by an MQSA-approved radiologist. The distance of the object from the nipple location is defined by $R=\|\overrightarrow{MN}\|$. The angle of the mass from the midline of the breast is defined by the angle between the vectors $\overrightarrow{MN}$ and $\overrightarrow{ON}$.

Fig. 4: CC view versus MLO view of the radial distances of the identified objects from the nipple location.

Fig. 5: CC view versus MLO view of the angular coordinates of the identified objects from the breast midline.

Fig. 6: Schematic diagram for the current one-view prescreening detection algorithm.

Fig. 7: Schematic diagram for the proposed two-view fusion scheme.

Fig. 8: Prediction of the center of an object in the MLO view from its location in the CC view. Training and test performances are given as a function of the radial width of the annular search region.

Fig. 9: Prediction of the center of an object in the CC view from its location in the MLO view. Training and test performances are given as a function of the radial width of the annular search region.

Fig. 10: Film-based performances of the current one-view mass detection algorithm applied to the data set of 338 one-view (169 pairs) mammograms. The FROC curves are plotted

for detection of all malignant and benign masses, and of the malignant masses on the current and the prior mammograms. Higher sensitivity was obtained for the detection of malignant masses on current mammograms.

Fig. 11: Case-based performances of the current one-view mass detection algorithm applied to the data set of 169 pairs of mammograms. The FROC curves are plotted for detection of all malignant and benign masses, and of the malignant masses on the current and the prior mammograms. Higher sensitivity was obtained for the detection of malignant masses on current mammograms.

Fig. 12: Film-based performances of the proposed two-view detection scheme for all masses. Three initial conditions depending on the maximum number of retained objects per image (5, 10, and 15 objects per image) at the prescreening stage were evaluated.

Fig. 13: Film-based performances of the proposed two-view detection scheme applied to the current malignant masses. Three initial conditions depending on the maximum number of retained objects per image (5, 10, and 15 objects per image) at the prescreening stage were evaluated.

Fig. 14: Comparison of the film-based performance of the one-view and two-view detection methods for the detection of malignant masses on current mammograms and prior mammograms.

Fig. 15: Comparison of the case-based performance of the one-view and two-view detection methods for the detection of malignant masses on current mammograms and prior mammograms.

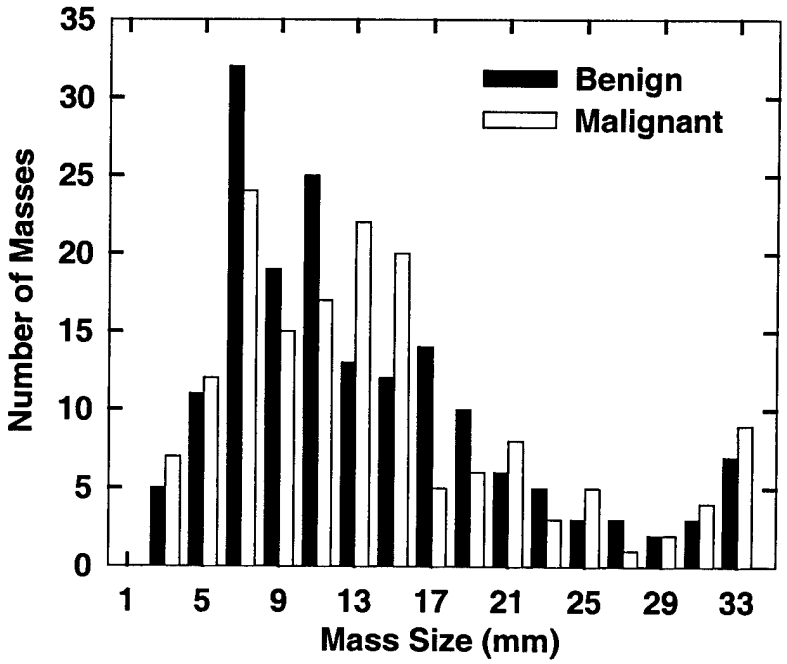


Fig. 1: Histograms of the size (the longest dimension) of the benign and malignant masses contained in the data set of 338 one-view mammograms and rated by an MSQA-radiologist. Eight masses in the prior mammograms of the data set did not receive a rating because the radiologist could not delineated the mass even in retrospect, although a focal density could be seen.

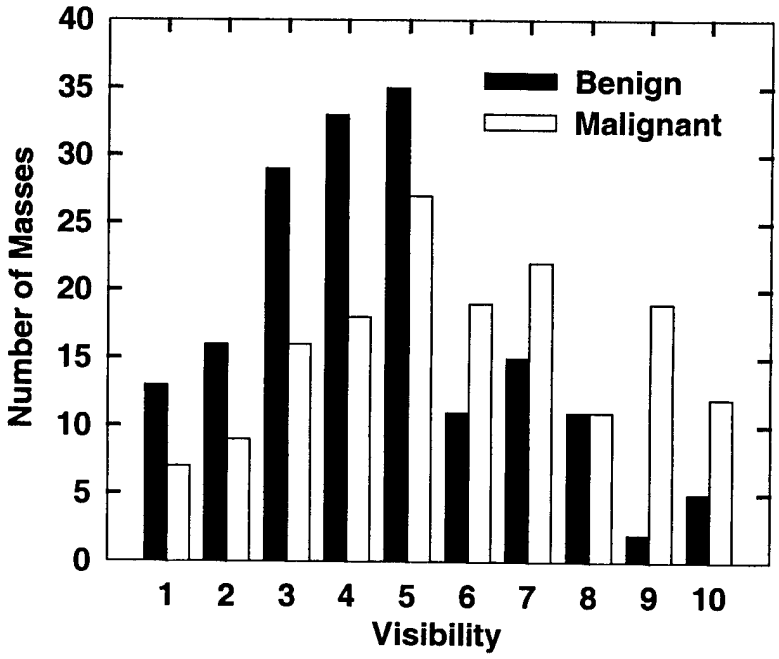


Fig. 2: Histograms of the visibility (1= most obvious, 10=subtlest) of the benign and malignant masses contained in the data set of 338 one-view mammograms and rated by an MSQA-radioliogist. Eight masses in the prior mammograms of the data set did not receive a rating because the radiologist could not delineated the mass even in retrospect, although a focal density could be seen.

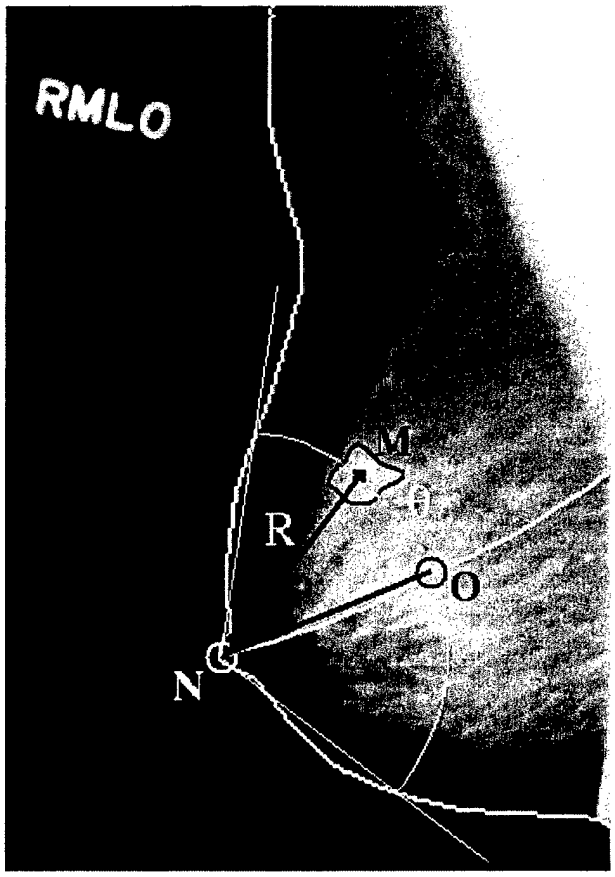


Fig. 3: Example of the coordinate system used to localize an object in a mammographic view. An automatic boundary tracking process is used to segment the breast. The nipple location was identified by an MQSA-approved radiologist. The distance of the object from the nipple location is defined by R=$\|\overrightarrow{MN}\|$. The angle of the mass from the midline of the breast is defined by the angle between the vectors $\overrightarrow{MN}$ and $\overrightarrow{ON}$.

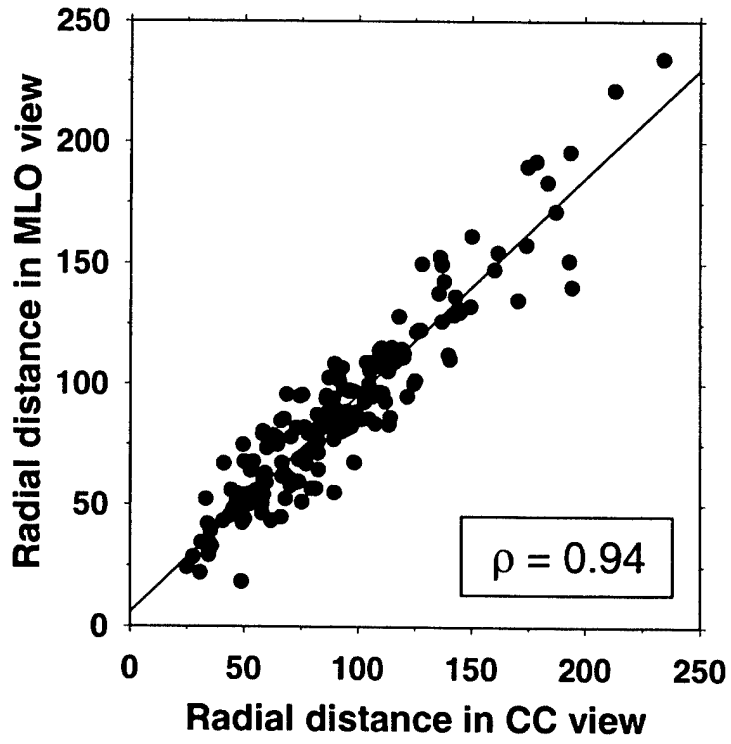


Fig. 4: CC view versus MLO view of the radial distances of the identified objects from the nipple location.

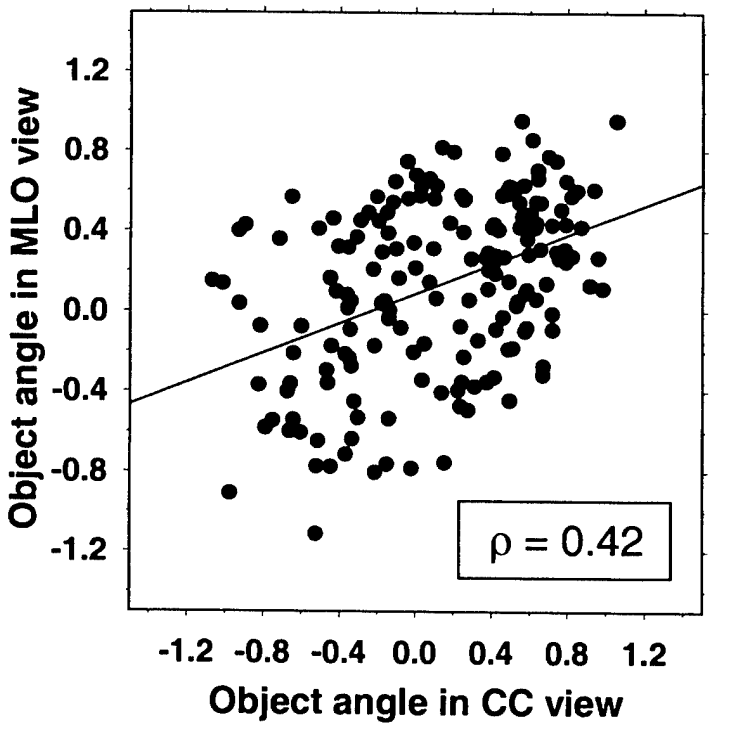


Fig. 5: CC view versus MLO view of the angular coordinates of the identified objects from the breast midline.

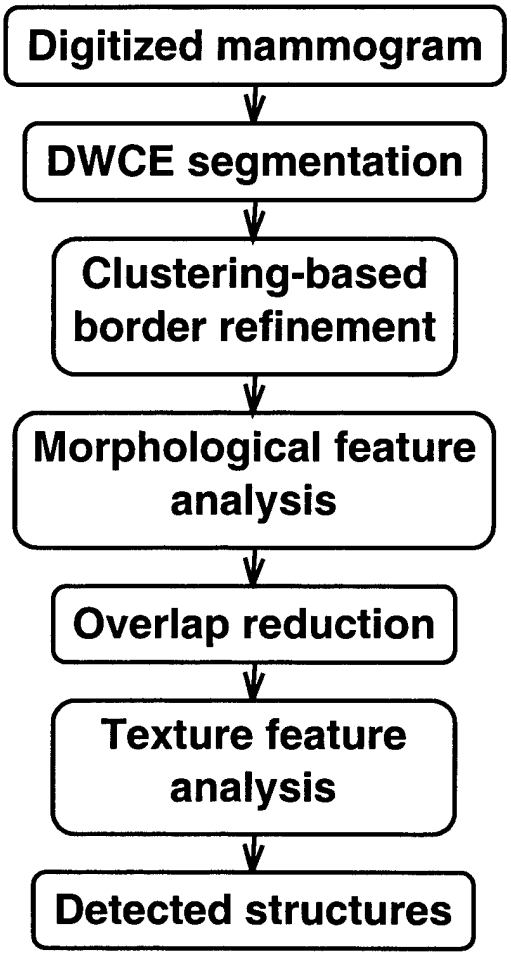


Fig. 6: Schematic diagram for the current one-view prescreening detection algorithm.

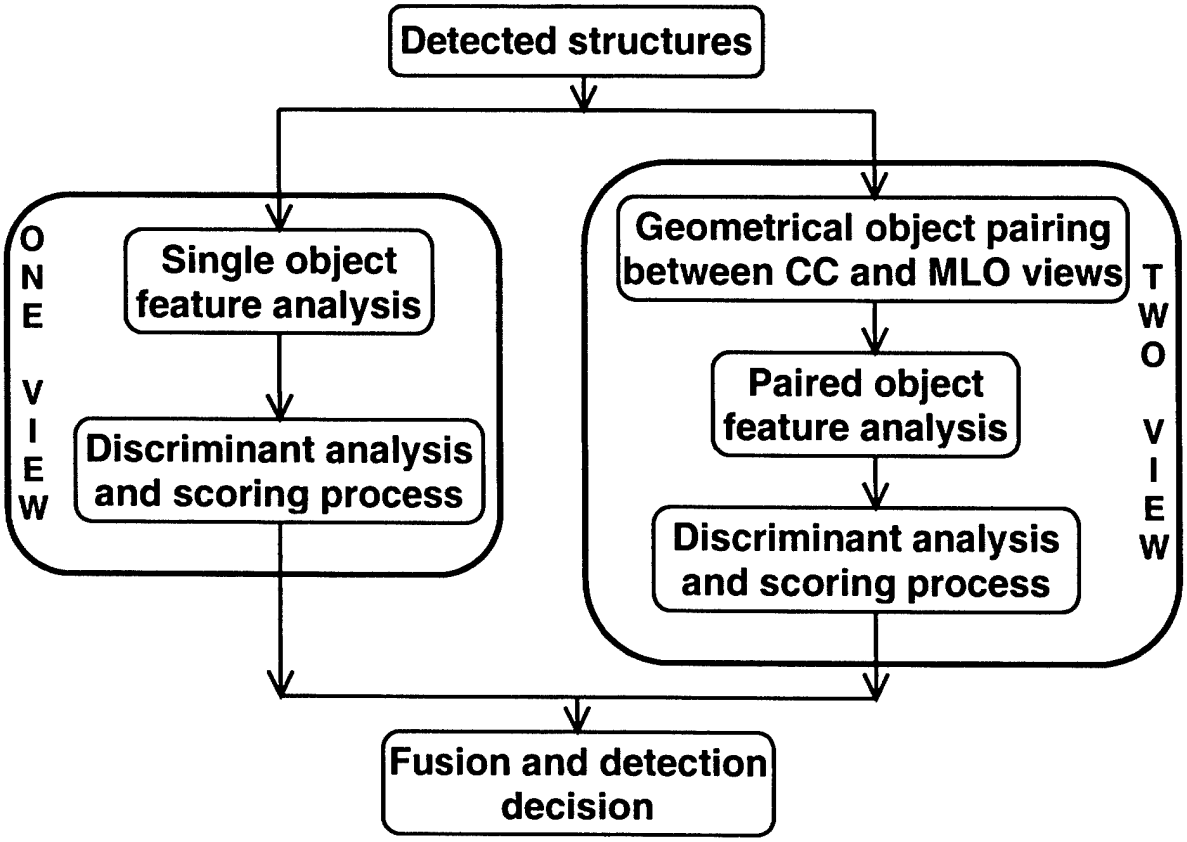


Fig. 7: Schematic diagram for the proposed two-view fusion scheme.

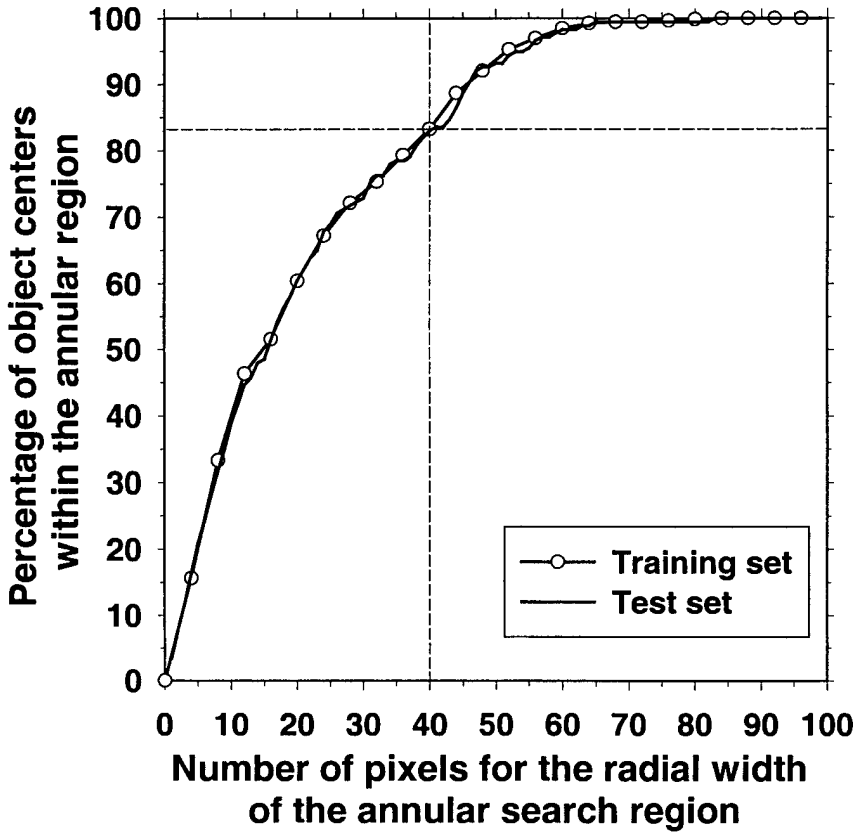


Fig. 8: Prediction of the center of an object in the MLO view from its location in the CC view. Training and test performances are given as a function of the radial width of the annular search region.

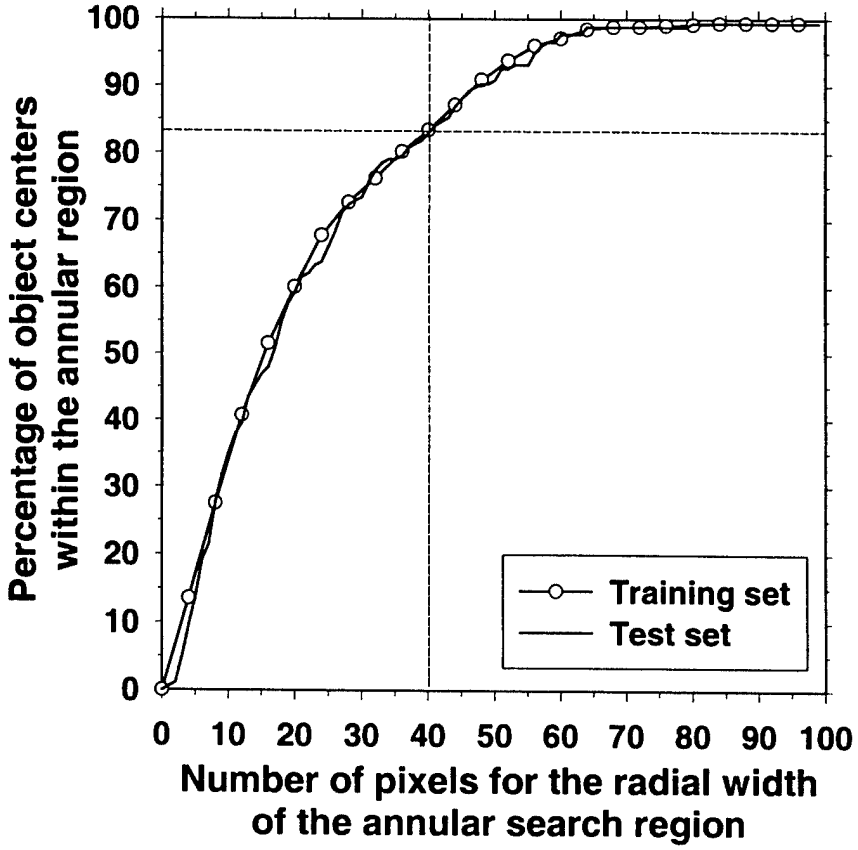


Fig. 9: Prediction of the center of an object in the CC view from its location in the MLO view. Training and test performances are given as a function of the radial width of the annular search region.

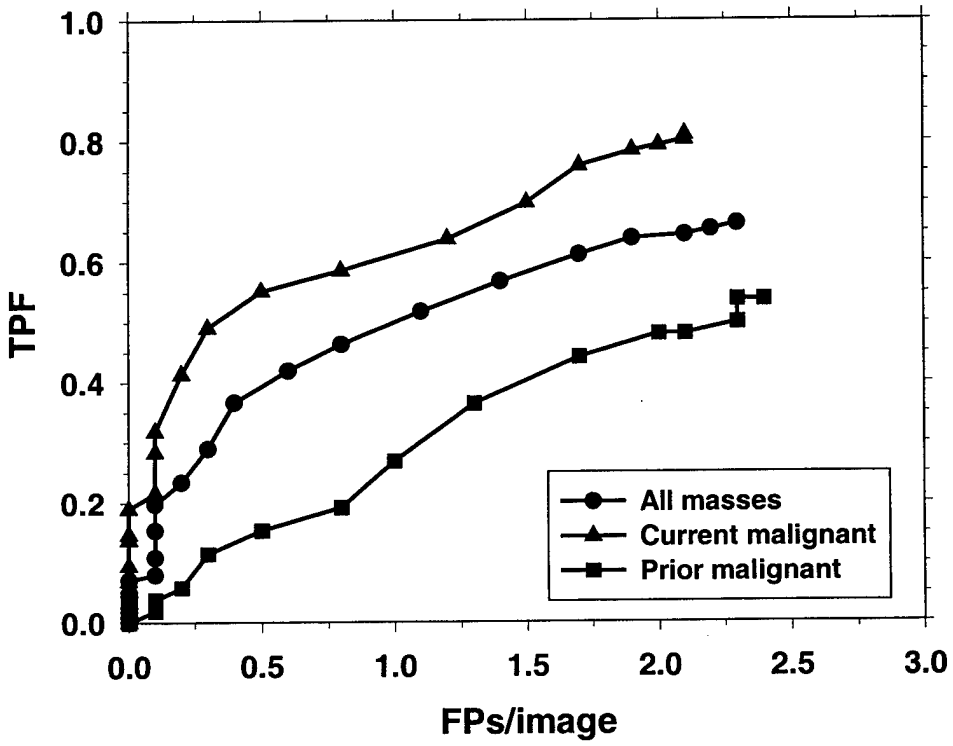


Fig. 10: Film-based performances of the current one-view mass detection algorithm applied to the data set of 338 one-view (169 pairs) mammograms. The FROC curves are plotted for detection of all malignant and benign masses, and of the malignant masses on the current and the prior mammograms. Higher sensitivity was obtained for the detection of malignant masses on current mammograms.

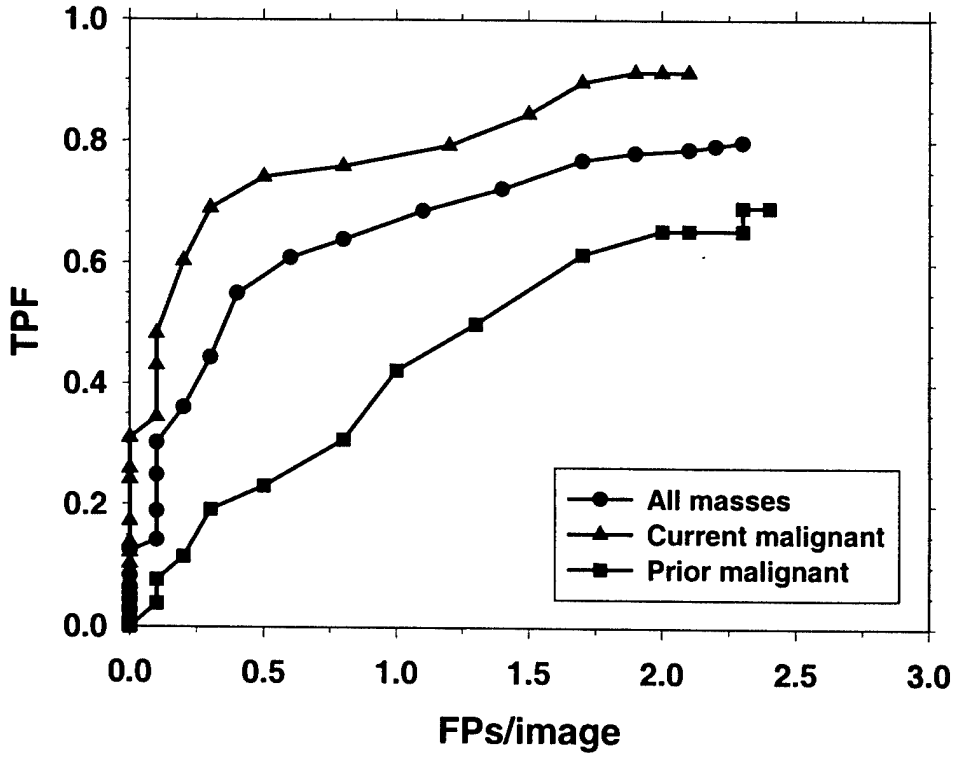


Fig. 11: Case-based performances of the current one-view mass detection algorithm applied to the data set of 169 pairs of mammograms. The FROC curves are plotted for detection of all malignant and benign masses, and of the malignant masses on the current and the prior mammograms. Higher sensitivity was obtained for the detection of malignant masses on current mammograms.

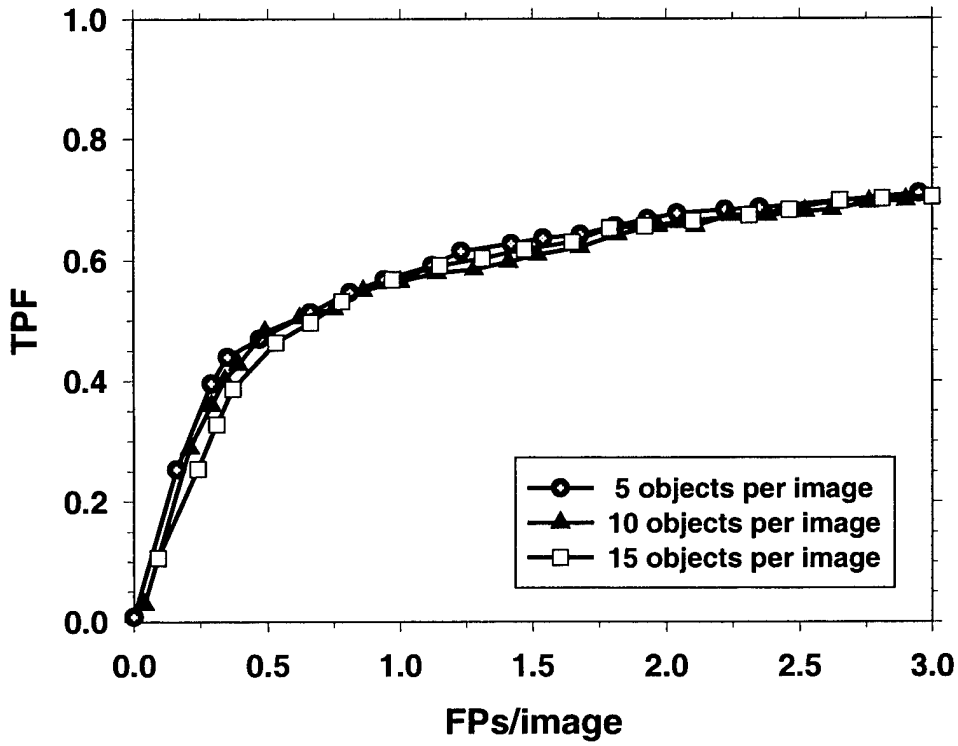


Fig. 12: Film-based performances of the proposed two-view detection scheme for all masses. Three initial conditions depending on the maximum number of retained objects per image (5, 10, and 15 objects per image) at the prescreening stage were evaluated.

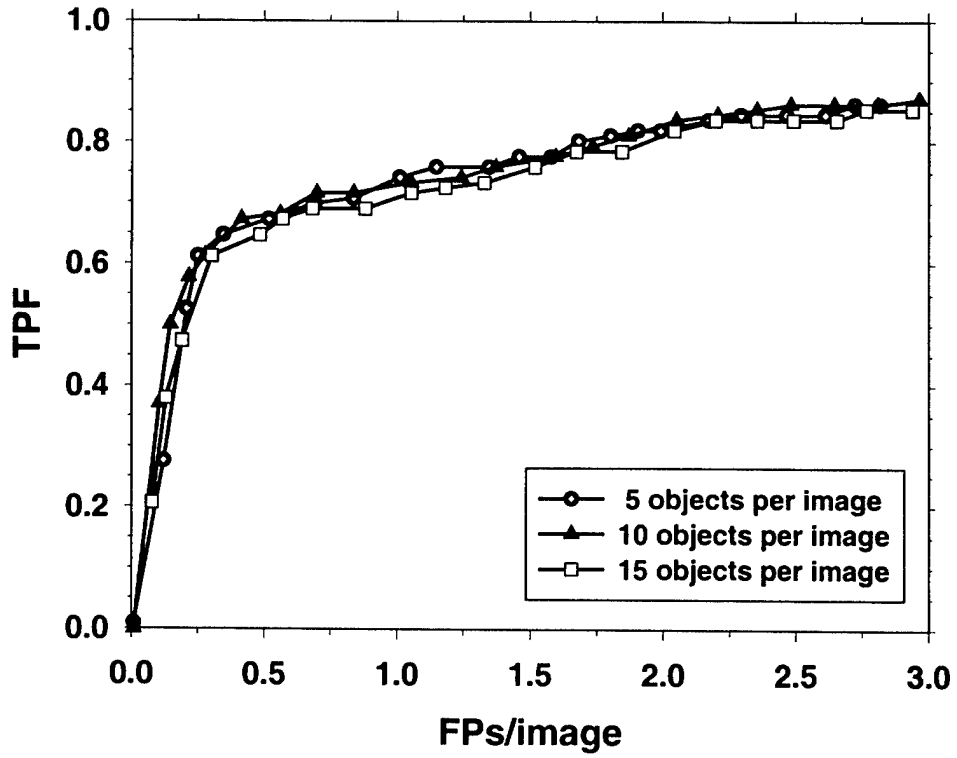


Fig. 13: Film-based performances of the proposed two-view detection scheme applied to the current malignant masses. Three initial conditions depending on the maximum number of retained objects per image (5, 10, and 15 objects per image) at the prescreening stage were evaluated.

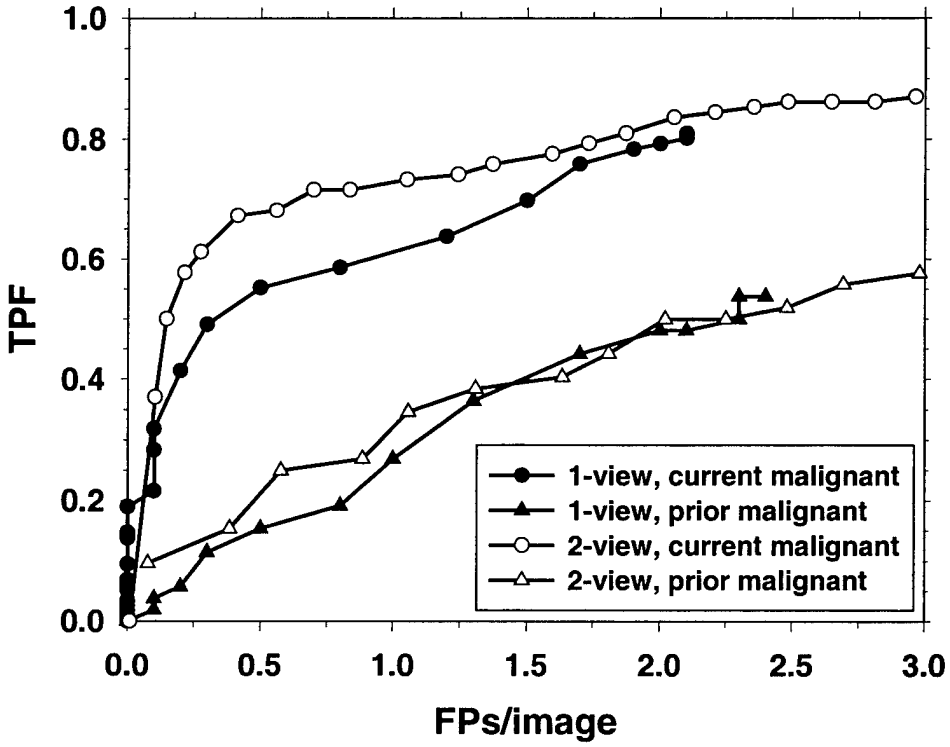


Fig. 14: Comparison of the film-based performance of the one-view and two-view detection methods for the detection of malignant masses on current mammograms and prior mammograms.

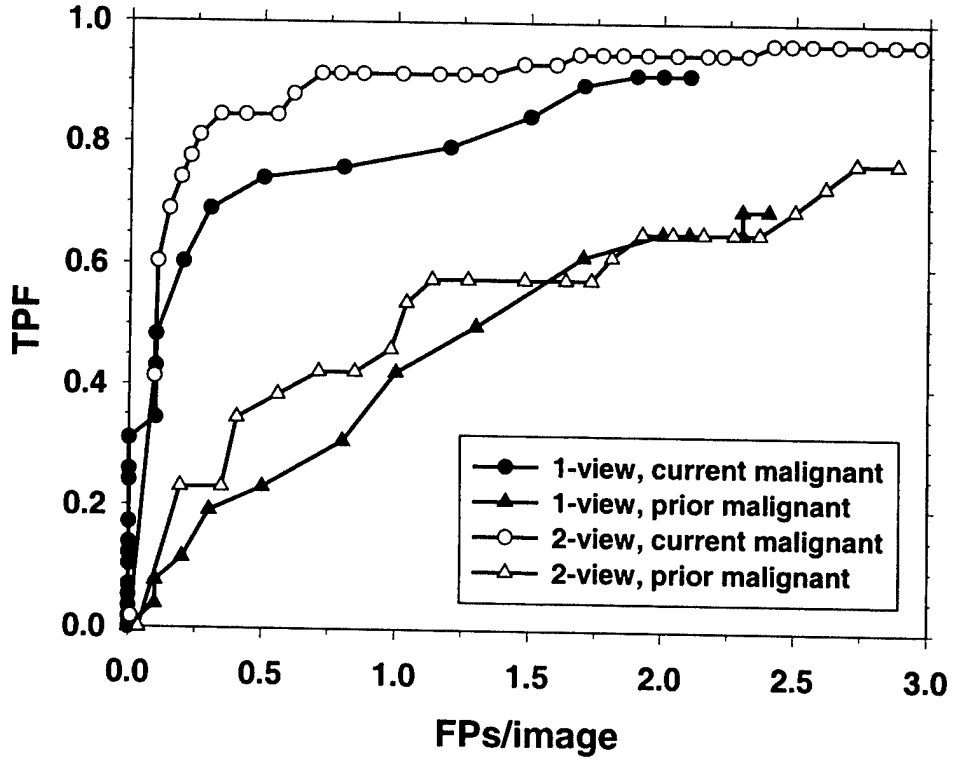


Fig. 15: Comparison of the case-based performance of the one-view and two-view detection methods for the detection of malignant masses on current mammograms and prior mammograms.

Table I: Characteristics of the 3 sets of objects to be input to the two-view scheme. The objects were obtained by applying a detection threshold at the prescreening stage to extract a maximum of 5, 10, and 15 objects per image.

Prescreening threshold objs/image	**Avg. objs/image**	**Sensitivity film-based (%)**	**Sensitivity case-based (%)**	**No. of pairs/case**
5	4.9	72.7	85.2	14.2
10	9.4	79.8	89.3	49.4
15	12.6	83.4	92.3	85.9

Table II: Comparison of detection sensitivities obtained by the one-view and the two-view fusion schemes for film-based and case-based detection.

Mass type	**Sensitivity - film-based (1 FPs/image)**		**Sensitivity - case-based (1 FPs/image)**	
	1-view	2-view	1-view	2-view
All	50%	56%	67%	73%
Current malignant	62%	73%	77%	91%
Prior malignant	27%	33%	42%	54%

Analysis of temporal changes of mammographic features: Computer-aided classification of malignant and benign breast masses

Lubomir Hadjiiski,[a] Berkman Sahiner, Heang-Ping Chan, Nicholas Petrick, Mark A. Helvie, and Metin Gurcan
Department of Radiology, The University of Michigan, Ann Arbor, Michigan 48109-0904

(Received 22 May 2001; accepted for publication 27 August 2001)

A new classification scheme was developed to classify mammographic masses as malignant and benign by using interval change information. The masses on both the current and the prior mammograms were automatically segmented using an active contour method. From each mass, 20 run length statistics (RLS) texture features, 3 speculation features, and 12 morphological features were extracted. Additionally, 20 difference RLS features were obtained by subtracting the prior RLS features from the corresponding current RLS features. The feature space consisted of the current RLS features, the difference RLS features, the current and prior speculation features, and the current and prior mass sizes. Stepwise feature selection and linear discriminant analysis classification were used to select and merge the most useful features. A leave-one-case-out resampling scheme was used to train and test the classifier using 140 temporal image pairs (85 malignant, 55 benign) obtained from 57 biopsy-proven masses (33 malignant, 24 benign) in 56 patients. An average of 10 features were selected from the 56 training subsets: 4 difference RLS features, 4 RLS features, and 1 speculation feature from the current image, and 1 speculation feature from the prior, were most often chosen. The classifier achieved an average training A_z of 0.92 and a test A_z of 0.88. For comparison, a classifier was trained and tested using features extracted from the 120 current single images. This classifier achieved an average training A_z of 0.90 and a test A_z of 0.82. The information on the prior image significantly ($p=0.015$) improved the accuracy for classification of the masses. © *2001 American Association of Physicists in Medicine.* [DOI: 10.1118/1.1412242]

Key words: computer-aided diagnosis, interval change, classification, feature analysis, mammography, malignancy

I. INTRODUCTION

Mammography is currently the most effective method for early breast cancer detection.[1,2] Analysis of interval changes is an important method used by radiologists in mammographic interpretation to detect developing malignancy.[3,4] A variety of computer-aided diagnosis (CAD) techniques have been developed to detect abnormalities and to distinguish malignant and benign lesions on mammograms. We are studying the use of CAD techniques to assist radiologists in interval change analysis.

Commonly used lesion classification methods for CAD employ information from a single image. These methods have been shown to perform well in lesion classification problems.[5–12] However, when mammograms from multiple examinations are available, it can be expected that even higher accuracy may be achieved if the computer can utilize the interval change information for classification. New computer vision methods will have to be designed to extract features characterizing temporal changes and to improve the differentiation between benign and malignant masses.

A number of researchers have developed algorithms to register the mass on current and prior mammograms. Sallam *et al.*[13] have proposed a warping technique for mammogram registration based on manually identified control points. A mapping function was calculated for matching each point on the current mammogram to a point on the prior mammogram. Brzakovic *et al.*[14] have investigated a three-step method for comparison of the most recent and the prior mammograms. They first registered two mammograms using the method of principal axis, and partitioned the current mammogram using a hierarchical region-growing technique. Translation, rotation, and scaling were then used for registration of the partitioned regions. Vujovic *et al.*[15] have proposed a multiple-control-point technique for mammogram registration. They first determined several control points independently on the current and prior mammograms based on the intersection points of prominent anatomical structures in the breast. A correspondence between these control points was established based on a search in a local neighborhood around the control point of interest.

The previous techniques depend on the identification of control points. Furthermore, these studies aimed at registration without using the results for interval change analysis.

Gopal *et al.*[16,17] and Hadjiiski *et al.*[18–20] have developed a multistage technique that defines a transformation to locally map the position of the mass on a current mammogram to a search region on the prior mammogram. A local search for the exact mass location is then performed on the prior mammogram. Good *et al.*[21] have developed a technique that defines a transformation to map all points from the current mammogram onto a prior mammogram. The current mammogram is then subtracted from the prior mammogram.

PROOF COPY 011111MPH

Few studies have been performed so far in the area of automated classification of breast masses based on the interval change information. Gopal *et al.*[22] and Hadjiiski *et al.*[23,24] have carried out a preliminary study of the classification scheme that combines prior and current information automatically extracted from masses on prior and current mammograms, respectively. The classifier using the combined prior and current information performed better than the classifier using current information alone. To our knowledge, no other studies that describe automated classification of malignant and benign breast lesions based on temporal changes of mammographic features have been reported.

The goal of our research is to develop a CAD method for automated analysis of interval changes to be used as an aid to radiologists for detection and classification of malignant and benign lesions on mammograms. In this study, we conducted a preliminary investigation to demonstrate the feasibility of analyzing temporal differences in the texture and morphological features between a mass on the most recent mammogram and a prior mammogram of the same view for the classification task. Additionally, we compared this method with two classification methods, one of which is based on information extracted from the current mammograms alone, the other one is based on information extracted from the prior mammograms alone.

II. MATERIALS AND METHODS

The new classification technique is based on the design of features that characterize the temporal change in the lesion of interest between two mammographic examinations. The mass to be analyzed can either be identified manually by a radiologist or automatically by a computerized detection program. In this study, the mass on each mammogram was identified by an MQSA certified radiologist. The masses on both the current and the prior mammograms were automatically segmented using an active contour method that has been discussed in detail elsewhere.[25,26] Examples of the segmentation are shown in Figs. 2 and 3 for a malignant and a benign mass, respectively. Features that characterized mammographic masses including texture features, morphological features, and spiculation features were extracted from each mass. Three of the morphological features are related to the mass size. Additionally, difference features were obtained by subtracting a feature of the prior mass from the corresponding feature of the current mass. The current, prior, and difference features formed a multidimensional feature space for the classification task. Stepwise feature selection applied to linear discriminant analysis (LDA) was used to select the most useful features. The selected features were then used as the input predictor variables for the LDA classifier (Fig. 1). The classifier was trained and tested by a leave-one-case-out resampling scheme. A case was considered to contain all regions of interest from a given patient. In each resampling step, the temporal pairs from 55 cases were used for feature selection and formulation of the linear discriminant function, while the temporal pairs from the left-out case were used for testing the trained classifier. A total of 56 training and testing steps were obtained from the 56 cases. The classification results from the 56 test cases were accumulated to evaluate the classifier performance. Since the data set in this study was still small, we chose the feature selection parameters such that the dimensionality of the input feature vector for the LDA classifier was small in order to reduce the possibility of over-training. The feature selection procedure is discussed in Sec. II C.

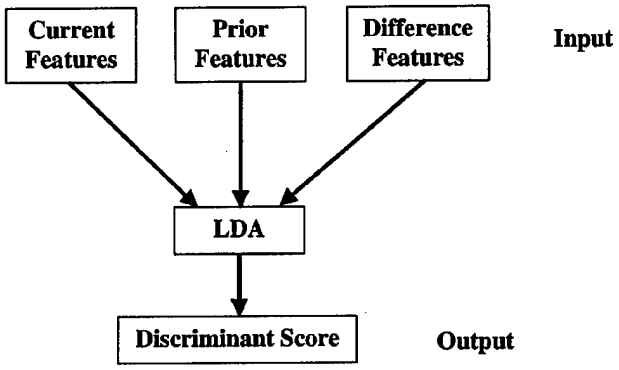


FIG. 1. Block diagram of the classification method.

To evaluate the improvement in the classifier performance designed by using the temporal change information, two additional classifiers were obtained. One of them was trained using the information extracted from the current single images of the temporal pairs. We will refer to these images as current images. The other classifier was trained using the information extracted from the prior single images of the temporal pairs and we will refer to these images as prior images. Comparison of the three classifiers will reveal the effectiveness of interval change analysis for the classification of malignant and benign masses.

A. Data set

A set of 140 temporal pairs of mammograms containing biopsy-proven masses on the current mammograms was used to examine the performance of this approach. The data set consisted of 241 mammograms from 56 patients. The mammograms were digitized with a LUMISCAN 85 laser scanner at a pixel resolution of $50\,\mu m \times 50\,\mu m$ and 4096 gray levels. The digitizer was calibrated so that gray level values were linearly proportional to the optical density (OD) within the range of 0–4 OD units, with a slope of 0.001 OD/pixel value. The digitizer output was linearly converted so that a large pixel value corresponded to a low optical density. The image matrix size was reduced by averaging every 2×2 adjacent pixels and downsampled by a factor of 2, resulting in images with a pixel size of $100\,\mu m \times 100\,\mu m$ for further analysis.

There were 57 biopsy-proven masses (33 malignant and 24 benign) in the 56 cases. The 241 mammograms contained different mammographic views (CC, MLO, and lateral views) and multiple examinations of the masses including the examination when the biopsy decision was made. By matching masses of the same view from two different examinations, a total of 140 temporal pairs were formed, of which

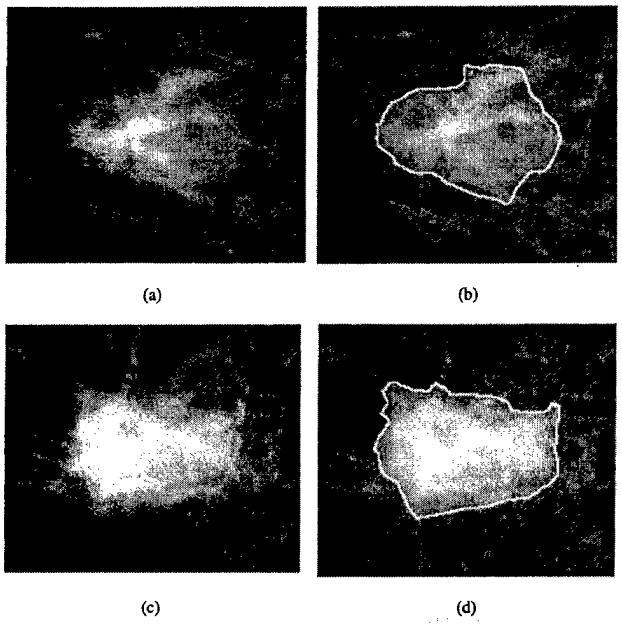


FIG. 2. A malignant mass: (a) the mass in a prior year mammogram (1997), (b) mass outline obtained by active contour segmentation, (c) the mass in a current year mammogram (1998), (d) mass outline obtained by active contour segmentation.

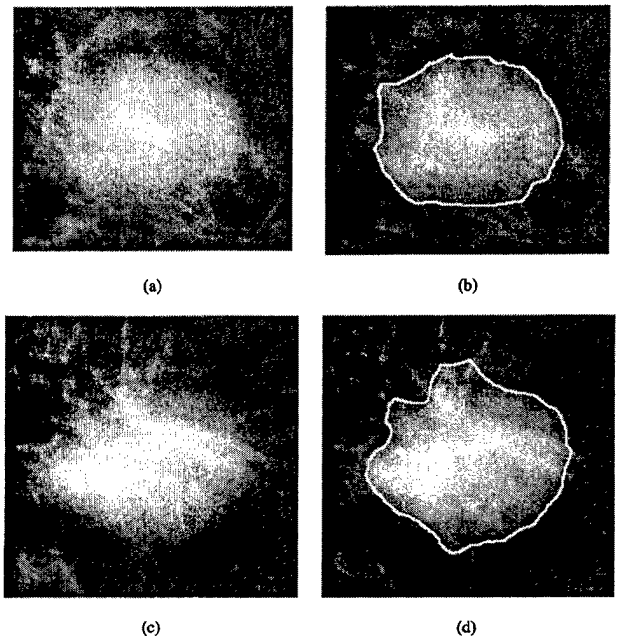


FIG. 3. A benign mass: (a) the mass on a prior year mammogram (1995), (b) mass outline obtained by active contour segmentation, (c) the mass on a current year mammogram (1996), (d) mass outline obtained by active contour segmentation.

85 were malignant and 55 benign. A malignant temporal pair consisted of a biopsy-proven malignant mass or a mass that was initially not recommended for biopsy and later found to be malignant by biopsy in a future year. A similar definition was used for the benign temporal pairs. Within a pair, the current mammogram was defined as the mammogram with the later date, and the prior mammogram was defined as the one with the earlier date. Therefore, in cases with three consecutive exams, more than one temporal pair could be formed and two of the mammograms could be called "current." Among the 140 temporal pairs, we had 120 unique current mammograms. Of the masses in the 120 current mammograms, 70 were malignant and 50 benign.

Since all cases in this data set had undergone biopsy, the benign masses in this set could not be distinguished easily from the malignant ones based on current mammographic criteria. Changes occurred for the benign masses that prompted the radiologists to recommend biopsy. Examples of such cases are shown in Figs. 2 and 3. The malignant mass in Fig. 2 did not increase in size but changed its density. The benign mass (Fig. 3), on the other hand, appeared to have spicules. For the malignant masses in this data set, the average mass size, estimated by the radiologist as the longest dimension of the mass on the mammogram, was 8.2 mm on the prior mammograms and 12.7 mm on the current mammograms. The corresponding sizes were 10.6 and 12.2 mm, respectively, for the benign masses. As discussed in Sec. IV, 25 of the masses on the prior mammograms were too subtle for the radiologist to estimate their sizes. The average sizes given previously were obtained after excluding all temporal pairs that involved these masses.

The radiologist also rated the visibility of the masses on the mammograms relative to those encountered in clinical practice on a 10-point scale, with 1 representing the most obvious and 10 representing the most subtle masses. The visibility of the masses on the current mammogram is plotted against those on the prior mammogram in Fig. 4 for the malignant and benign temporal pairs. Generally the malignant masses were less visible on the prior than on the current mammograms while the visibility of the benign masses was found to be more similar on the current and prior mammograms. The mean difference in the visibility rating between the prior and the current mammograms for the malignant masses is 2.8 compared to 1.2 for the benign masses ($p = 0.0007$ with an unpaired t-test between the malignant and benign masses). The correlation coefficient is 0.02 for malignant masses [Fig. 4(a)] and 0.37 for benign masses [Fig. 4(b)]. In addition, the radiologist also estimated the likelihood of malignancy of the current masses on a 10-point confidence scale (1—definitely benign and 10—definitely malignant) based on the 120 current mammograms alone without comparison with the prior (Fig. 5). The temporal pairs had a time interval of 6–36 months (Fig. 6). More than 70% of the pairs had a time interval of 12 months.

B. Feature extraction

A rectangular region of interest (ROI) was defined to include the radiologist-identified mass with an additional surrounding breast tissue region of at least 40 pixels wide from any point of the mass border. A fully automated method was then used for segmentation of the mass from the breast tissue background within the ROI. The masses on both the current and the prior mammograms were automatically segmented

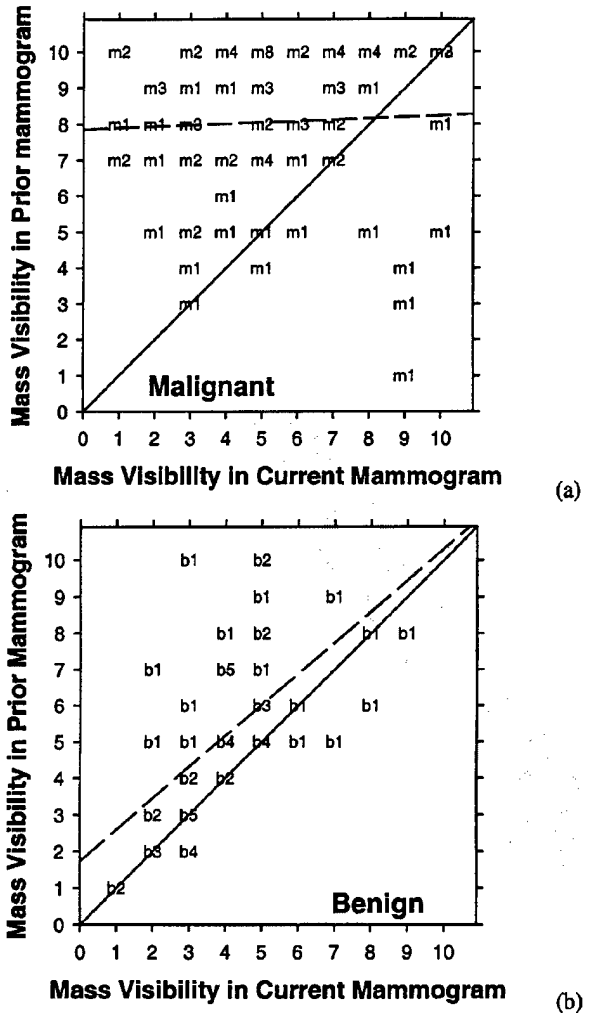


FIG. 4. Visibility of the masses on the current mammogram plotted against those on the prior mammogram for (a) malignant and (b) benign temporal pairs. The visibility was rated on a 10-point discrete scale (1 = most obvious, 10= most subtle). Because many of the data points overlap, we indicate the number of points with the same rating by a number next to the symbol (m or b). The diagonal line on the graph represents the cases when the current and the prior mass sizes are identical. The dashed lines are the linear regression lines for the data defined by $y=0.038x+7.86$ for (a) and by $y=0.857x+1.742$ for (b). The correlation coefficient for malignant masses is 0.02 and for benign masses is 0.37.

within the ROI using a two-dimensional active contour method that was initialized by K-mean clustering.[25,26]

The texture features used in this study were calculated from run-length statistics (RLS) matrices.[27] The RLS matrices were computed from the images obtained by the rubber band straightening transform (RBST).[6] The RBST maps a band of pixels surrounding the mass onto the Cartesian plane (a rectangular region). In the transformed image, the mass border appears approximately as a horizontal edge, and spiculations appear approximately as vertical lines. A complete description of the RBST can be found in the literature.[6]

RLS texture features were extracted from the vertical and horizontal gradient magnitude images, which were obtained by filtering the RBST image with horizontally or vertically

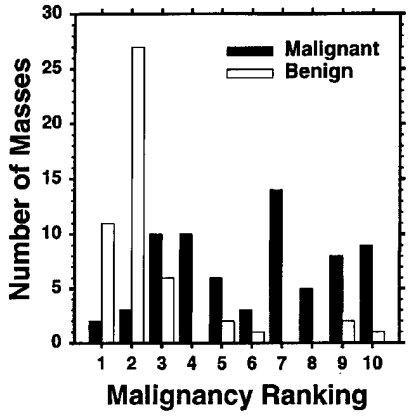


FIG. 5. The distribution of the malignancy ranking of the masses in the 120 current mammograms. The rating was performed by an experienced MQSA radiologist (1: definitely benign, 10: definitely malignant).

oriented Sobel filters and computing the absolute gradient values of the filtered image.[6] Five texture measures, namely, short run emphasis (SRE), long run emphasis (LRE), gray level nonuniformity (GLN), run length nonuniformity (RLN), and run percentage (RP) were extracted from the vertical and horizontal gradient images in two directions, $\theta=0°$, and $\theta=90°$. Therefore, a total of 20 RLS features were calculated for each ROI. The definition of the RLS feature measures can be found in the Appendix and in the literature.[27]

Morphological features were extracted from the automatically segmented mass shape. Five of the morphological features were based on the normalized radial length (NRL), defined as the Euclidean distance from the object's centroid to each of its edge pixels, i.e., the radial length, and normalized relative to the maximum radial length for the object.[11] The following five NRL features were extracted: mean (NRLAVG), standard deviation (NRLSD), entropy (NRLENT), area ratio (NRLAREAR), zero crossing count (NRLZCC). In addition, the perimeter (PERIM), area (AREA), circularity (CIRC), rectangularity (SQR), contrast (CONT), perimeter-to-area ratio (CRR), and Fourier descriptor (FF)

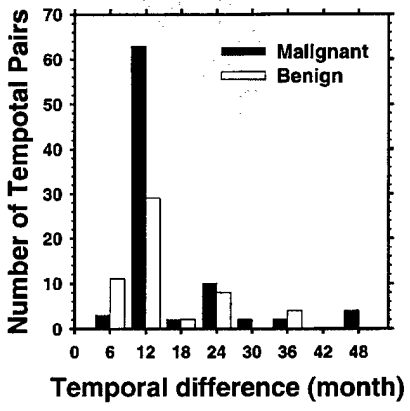


FIG. 6. Temporal interval between the current and the prior mammograms for the 140 temporal pairs in our data set.

features were extracted. The definitions of the morphological features can be found in the literature.[26,28] Three of the morphological features (perimeter, area, and perimeter-to-area ratio) are related to the mass size and thus are feature descriptors of the mass size.

A spiculation measure was defined for each pixel on the mass border by using the statistics of the image gradient direction relative to the normal direction to the mass border. The statistics was determined in a 90° sector centered about the normal at the border pixel and outside of the mass border.[25,26] The spiculation measure for each border pixel was normalized to be between 0 and $\pi/2$, with a value of $\pi/4$ indicating a random orientation of image gradients, and larger values indicating a higher likelihood of spiculation. Three features were extracted from the spiculation measure. The first feature (AVG) was the average of the spiculation measure for all pixels on the mass boundary. The second feature (PERC_ABV) was the percentage of border pixels with a spiculation measure larger than $\pi/4$, and the third feature (AVE_ABV) was the average of the spiculation measure for those pixels with a spiculation measure larger than $\pi/4$.

A total of 35 features (20 RLS, 12 morphological, and 3 spiculation) were therefore extracted from each ROI. Additionally, difference features were obtained by subtracting a prior feature from the corresponding current feature. Therefore, 35 difference features were derived from the 20 RLS, 12 morphological, and 3 spiculation features.

C. Feature selection

In order to reduce the number of the features and to obtain the best feature subset to design an effective classifier, feature selection with stepwise linear discriminant analysis[29] was applied. At each step of the stepwise selection procedure one feature is entered or removed from the feature pool by analyzing its effect on the selection criterion. In this study, the Wilks' lambda (the ratio of within-group sum of squares to the total sum of squares[30]) was used as a selection criterion. The optimization procedure used a threshold F_{in} for feature entry, a threshold F_{out} for feature removal, and a tolerance threshold T for measuring feature correlation with the other features. In a feature entry step, the features not yet selected are entered into the selected feature pool one at a time, the significance of the change in the Wilks' lambda caused by this feature is estimated based on F statistics. The feature with the highest significance is entered into the feature pool if its significance is higher than F_{in} and its correlation value with the rest of the features in the pool is below T. In a feature removal step, the features that have already been entered in the selected feature pool are removed one at a time and the significance of the change in the Wilks' lambda is estimated. The feature with the least significance is removed from the selected feature pool if the significance is less than F_{out}. Since the appropriate values of F_{in}, F_{out} and T are not known *a priori*, we examined a range of F_{in}, F_{out}, and T values using an automated simplex optimization method.[31,32] The appropriate thresholds were chosen in such a way that a minimum number of features were selected to achieve a high accuracy of classification by LDA. More details about the stepwise linear discriminant analysis and its application to CAD can be found elsewhere.[5,6]

TABLE I. Classification results for the classifier based on the temporal change information, the classifier based on current single image information, and the classifier based on prior single image information.

Classification	Avg. No. of selected features	Training A_z	Test A_z	Test partial $A_z^{(0.9)}$
Temporal pairs	10	0.92	0.88±0.03	0.37±0.10
Current images	11	0.90	0.82±0.04	0.32±0.08
Prior images	4	0.78	0.76±0.04	0.24±0.08

The feature selection in this study was performed by applying the stepwise feature selection to the entire feature space (combination of texture, spiculation, and morphological features altogether) as well as subspaces obtained by different combinations of the three feature subspaces: texture, spiculation, and morphological features. The stepwise feature selection uses a sequential forward inclusion and backward elimination approach. The procedure does not exhaustively evaluate all possible combinations of individual features. It is therefore not optimal, especially when the feature space is large and the training sample is small. By limiting the input to the feature subspaces, the dimensionality was reduced compared to the entire feature space. We found that better feature subsets could be selected by the stepwise feature selection in the subspaces than in the entire feature space.

D. Evaluation methods

To evaluate the classifier performance, the training and test discriminant scores were analyzed using receiver operating characteristic (ROC) methodology.[33] The discriminant scores of the malignant and benign masses were used as decision variables in the LABROC1 program,[34] which fits a binormal ROC curve based on maximum likelihood estimation. The classification accuracy was evaluated as the area under the ROC curve, A_z. The performances of the classifiers were also assessed by estimating the partial area index ($A_z^{(0.9)}$). The partial area index ($A_z^{(0.9)}$) is defined as the area that lies under the ROC curve but above a sensitivity threshold of 0.9 ($\text{TPF}_0=0.9$) normalized to the total area above TPF_0 ($1-\text{TPF}_0$). The partial $A_z^{(0.9)}$ indicates the performance of the classifier in the high sensitivity (low false negative) region which is most important for a cancer detection task.

III. RESULTS

The performances of the classifiers based on the temporal pairs, the current images, and the prior images are summarized in Table I. The classifiers that achieved the highest test A_z values with a small average number of features were presented here. Table II is a summary of the features selected for each classifier. For the 56 training subsets of temporal pairs used in this study, an average of 10 features were selected for

TABLE II. Selected features for classifiers based on temporal pairs, current images, and prior images. The letter "H" or "V" at the beginning of the texture feature labels indicates that the features were extracted from the horizontal or vertical gradient magnitude images, respectively. The number (0 or 90) at the end of the texture feature labels shows the direction at which the features were extracted.

			Temporal pairs			Current images	Prior images
Feature type	Group	Features	Curr	Pr	Diff	Curr	Pr
Texture	SRE	H_SRE_0			×		
		H_SRE_90	×		×		
		V_SRE_0	×		×	×	×
		V_SRE_90				×	
	LRE	V_LRE_0			×		×
		H_LRE_0				×	
	RLN	V_RLN_0	×			×	
	RP	H_RP_0	×				×
Spiculation		PERC_ABV	×			×	
		AVG		×			
		AVG_ABV					×
Morphological		CRR				×	
		NRLZCC				×	
		PERIM				×	
		NRLAVG				×	
		SQR				×	
		CONT				×	

the classification task. The most frequently selected features included 4 difference RLS features (3 SRE and 1 LRE), 4 RLS features (2 SRE, 1 RLN and 1 RP), 1 spiculation feature from the current image, and 1 spiculation feature from the prior image (Table II). The LDA classifier achieved an average training A_z of 0.92 and a test A_z of 0.88. The test partial $A_z^{(0.9)}$ was 0.37.

For classification of malignant and benign masses using the current single images (the current images of the temporal pairs), the LDA classifier selected an average of 11 features for the 56 training subsets. The most frequently selected features were 4 RLS features (2 SRE, 1 LRE and 1 RLN), 1 spiculation feature, and 6 morphological features (Table II). The classifier achieved an average training A_z of 0.90, a test A_z of 0.82, and a test partial $A_z^{(0.9)}$ of 0.32.

For the classification of masses based on the prior single images alone, an average of 4 features were selected for the 56 training subsets. The most frequently selected features were 3 RLS features (1 SRE, 1 LRE, and 1 RP) and 1 spiculation feature. The LDA classifier achieved an average training A_z of 0.78, test A_z of 0.76, and test partial $A_z^{(0.9)}$ of 0.24.

The test ROC curves for the three classifiers are compared in Fig. 7. The difference in the test A_z between the classifier based on the temporal pairs and that based on the current images alone is statistically significant ($p=0.015$). The difference in the test A_z between the classifier based on the temporal pairs and that based on the prior images alone is also statistically significant ($p=0.001$). The partial area index for the classifier based on the temporal pairs is also improved compared to the classifiers based on the current or the prior images alone, although the differences did not achieve statistical significance.

IV. DISCUSSION

Texture and spiculation features were important for malignant and benign classification of mammographic masses for all three types of classifiers: the classifier based on temporal pair information, the classifier based on current image information, and the classifier based on prior image information. One or more of the spiculation features were always selected in all training partitions for all three classifiers. The most frequently selected texture features were the short run emphasis (SRE) features. They comprised more than 50% of the texture features selected for the three classifiers (Table II).

The temporal-information-based classifier showed improved performance compared to the classifiers based on cur-

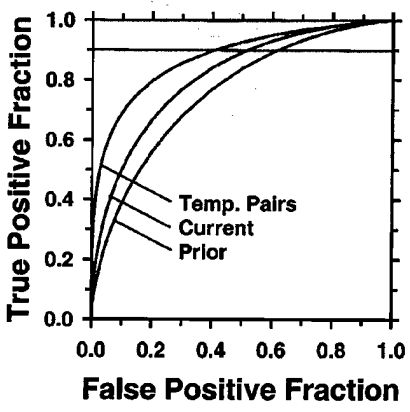


FIG. 7. The test ROC curves for the classifiers based on temporal pair information, current image information, and prior image information.

rent or prior image information alone. The input feature space to the temporal-information-based classifiers included the current, prior, and difference features. This allows the classifier to choose the individual features or the difference features. Using the stepwise feature selection procedure and the linear discriminant classifier, it was found that the texture and the spiculation features contained useful temporal information to perform malignant and benign mass classification. Texture features appeared to provide the best information by the difference features obtained from subtracting the prior from the corresponding current features (SRE and LRE difference features). On the other hand, the best use of the spiculation features appeared to be a direct combination of current and prior features in the input feature vector by the LDA since the individual features were chosen.

We found that better feature subsets could be selected by the stepwise feature selection in the subspaces than in the entire feature space. For example, for the temporal-information-based classifier, a better feature subset with a higher test A_z at 0.88 was found when the input feature space included only the texture and spiculation subspaces. The addition of the morphological feature subspace to the input feature space reduced the highest test A_z to 0.84. Similarly, in the case of the classifier based on prior image information, a better feature subset was obtained when the texture and spiculation feature subspaces were used in the input feature space for stepwise feature selection. Again the addition of the morphological feature subspace to the input feature space reduced the highest test A_z to 0.72. The classifier based on current image information was the only one, among the three, that obtained a better result, as shown in Table I, when the morphological feature subspace was included in the input feature space.

One reason for the poor performance of the morphological features may be due to the fact that the masses were more subtle in the prior images. In fact, the experienced MQSA mammographer was not confident in seeing 25 of the "masses" on the prior images and could not provide a mass size estimation for them. Although the active contour model would stop the iteration based on the preset criteria and found an "outline" of the masses on the prior mammograms, generally these mass outlines were less reliable than those on the current masses in providing morphological characteristics of the masses. Texture features did not depend as strongly on the precise mass boundary as morphological features. Three out of the four features selected for classification of the malignant and benign masses on the prior images were RLS texture features. A spiculation feature was also found to be a good discriminator.

We also performed ROC analysis of the malignancy confidence ratings provided by the experienced MQSA radiologist for the current image data set (120 images). The distribution of the malignancy ratings is shown in Fig. 5, which resulted in an A_z value of 0.80 ± 0.04. This indicates that the masses in the current mammograms cannot be easily distinguished as malignant or benign even by an experienced radiologist, consistent with the fact that all lesions had indeed undergone biopsy. The classifier based on the current image information has an A_z value of 0.82 ± 0.04, similar to the accuracy of the radiologist for this data set.

In this study, the locations of the masses were identified manually on both the current and the prior mammograms by a radiologist. This simulated the situation when a radiologist finds a mass either in a diagnostic or a screening setting and call upon the CAD algorithm to seek a second opinion on the likelihood of malignancy of the mass based on the interval change information. We are developing an automated regional registration technique that can automatically locate the mass on the prior mammogram based on its location on the current mammogram. The location of the mass on the current mammogram can be identified by a radiologist or by an automated mass detection algorithm. In the latter case, the process of mass detection, current and prior mass registration, and classification can be fully automated. The analysis of interval change can be incorporated as one of the functions provided by a CAD system for interpretation of mammograms.

In this study, we employed a simple measure of temporal change by taking the difference between the feature from the current mass and the corresponding feature from the prior mass. We observed improvement in classification with this simple temporal information. It will be important to evaluate other similarity measures that can characterize small difference in image features of the object of interest. It can be expected that a more sensitive similarity measure will provide a better measurement of dissimilarity, or difference, between the current and prior masses and further improve the utilization of the temporal change information on mammograms.

V. CONCLUSION

We performed a preliminary study to evaluate the effectiveness of interval change analysis for classification of malignant and benign masses on mammograms. It was found that the difference RLS texture features and spiculation features were useful for identification of malignancy in temporal pairs of mammograms. The information on the prior image was important for characterization of the masses; 5 out of the 10 selected features contained prior information. We found that the mass size descriptors were not discriminatory features for these difficult cases because many of the benign masses also grew over time. In comparison with the classification based on image information from the current images alone, the temporal change information significantly ($p = 0.015$) improved the accuracy for classification of the masses in terms of the total area under the ROC curve (A_z). The partial area under the ROC curve for the classifier based on the temporal pairs ($A_z^{(0.9)} = 0.37$) is also improved compared to the classifier based only on the current images ($A_z^{(0.9)} = 0.32$), although the difference did not achieve statistical significance. Further studies are under way to improve this temporal change classification technique and to evaluate its performance on a larger data set.

ACKNOWLEDGMENTS

This work is supported by a Career Development Award from the USAMRMC (No. DAMD 17-98-1-8211) (L.H.), USPHS Grant No. CA 48129, and a USAMRMC grant (No. DAMD 17-96-1-6254). The content of this publication does not necessarily reflect the position of the government and no official endorsement of any equipment and product of any companies mentioned in the publication should be inferred. The authors are grateful to Charles E. Metz, Ph.D., for the LABROC program.

APPENDIX: RUN LENGTH STATISTICS TEXTURE FEATURES

A gray level run length is a set of consecutive collinear pixels all having the same gray level value. The length of the run is the number of pixels in the run. For a given image it is possible to compute a gray level run length matrix for runs in any given direction. In this study, two directions are used: $\theta=0°$, and $\theta=90°$. Let $p(i,j)$ be the number of times there is a run of length j that has a gray level i. Let N_g be the number of gray levels and N_r be the number of runs. The short run emphasis is defined as

$$\mathrm{SRE}=\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}\frac{p(i,j)}{j^2}}{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}p(i,j)}.$$

This feature divides the frequency of each run length by the length of the run squared. This tends to emphasize short runs. The denominator is the total number of runs in the image and serves as a normalizing factor. The long run emphasis is defined as

$$\mathrm{LRE}=\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}j^2p(i,j)}{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}p(i,j)}.$$

This feature multiplies the frequency of each run length by the length of the run squared. This tends to emphasize long runs.

The gray level nonuniformity is defined as

$$\mathrm{GLN}=\frac{\sum_{i=1}^{N_g}(\sum_{j=1}^{N_g}p(i,j))^2}{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}p(i,j)}.$$

This feature squares the number of run lengths for each gray level. This measures the gray level nonuniformity of the image. If the runs are equally distributed over all gray levels, the feature takes on its lowest values. A larger run length contributes more to the feature value.

Run length nonuniformity is defined as

$$\mathrm{RLN}=\frac{\sum_{j=1}^{N_r}(\sum_{i=1}^{N_r}p(i,j))^2}{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}p(i,j)}.$$

This feature measures the nonuniformity of the run lengths. If the runs are equally distributed over all lengths, the feature will have a low value. A larger run contour contributes more to the feature value.

Run percentage is defined as

$$\mathrm{RP}=\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}p(i,j)}{P}.$$

This feature is a ratio of the total number of runs to the total number of possible runs (P) if all runs have a length of one.

The above-given definitions are based on Galloway[27] and more details can be found in this reference.

[a]Electronic mail: lhadjisk@umich.edu

[1] H. C. Zuckerman, "The role of mammography in the diagnosis of breast cancer," in *Breast Cancer, Diagnosis and Treatment*, edited by I. M. Ariel and J. B. Cleary (McGraw–Hill, New York, 1987).
[2] L. Tabar and P. B. Dean, "The control of breast cancer through mammography screening," Radiol. Clin. North Am. **25**, 961–■■■ (1987).
[3] L. W. Bassett, B. Shayestehfar, and I. Hirbawi, "Obtaining previous mammograms for comparison: Usefulness and costs," AJR, Am. J. Roentgenol. **163**, 1083–1086 (1994).
[4] E. A. Sickles, "Periodic mammographic follow-up of probably benign lesions: Results in 3183 consecutive cases," Radiology **179**, 463–468 (1991).
[5] H. P. Chan, D. Wei, M. A. Helvie, B. Sahiner, D. D. Adler, M. M. Goodsitt, and N. Petrick, "Computer-aided classification of mammographic masses and normal tissue: Linear discriminant analysis in texture feature space," Phys. Med. Biol. **40**, 857–876 (1995).
[6] B. Sahiner, H. P. Chan, N. Petrick, M. A. Helvie, and M. M. Goodsitt, "Computerized characterization of masses on mammograms: The rubber band straightening transform and texture analysis," Med. Phys. **25**, 516–526 (1998).
[7] H. P. Chan, B. Sahiner, N. Petrick, M. A. Helvie, K. L. Leung, D. D. Adler, and M. M. Goodsitt, "Computerized classification of malignant and benign microcalcifications on mammograms: Texture analysis using an artificial neural network," Phys. Med. Biol. **42**, 549–567 (1997).
[8] L. M. Hadjiiski, B. Sahiner, H. P. Chan, N. Petrick, and M. A. Helvie, "Classification of malignant and benign masses based on hybrid ART2LDA approach," IEEE Trans. Med. Imaging **18**, 1178–1187 (1999).
[9] Y. Wu, M. L. Giger, K. Doi, C. J. Vyborny, R. A. Schmidt, and C. E. Metz, "Artificial neural networks in mammography: Application to decision making in the diagnosis of breast cancer," Radiology **187**, 81–87 (1993).
[10] V. Goldberg, A. Manduca, D. L. Evert, J. J. Gisvold, and J. F. Greenleaf, "Improvement in specificity of ultrasonography for diagnosis of breast tumors by means of artificial intelligence," Med. Phys. **19**, 1475–1481 (1992).
[11] J. Kilday, F. Palmieri, and M. D. Fox, "Classifying mammographic lesions using computer-aided image analysis," IEEE Trans. Med. Imaging **12**, 664–669 (1993).
[12] Z. M. Huo, M. L. Giger, C. J. Vyborny, D. E. Wolverton, R. A. Schmidt, and K. Doi, "Automated computerized classification of malignant and benign masses on digitized mammograms," Acad. Radiol. **5**, 155–168 (1998).
[13] M. Sallam and K. Bowyer, "Detecting abnormal densities in mammograms by comparison with previous screenings," in *Digital Mammography 96*, edited by K. Doi, M. L. Giger, R. M. Nishikawa, and R. A. Schmidt (Elsevier, Amsterdam, 1996).
[14] D. Brzakovic, N. Vujovic, M. Neskovic, P. Brzakovic, and K. Fogarty, "Mammogram analysis by comparison with previous screenings," in *Digital Mammography*, edited by A. G. Gale, S. M. Astley, D. R. Dance, and A. Y. Cairns (Elsevier, Amsterdam, 1994).
[15] N. Vujovic and D. Brzakovic, "Establishing the correspondence between control points in pairs of mammographic images," IEEE Trans. Med. Imaging **6**, 1388–1399 (1997).
[16] S. S. Gopal, H.-P. Chan, N. Petrick, T. E. Wilson, B. Sahiner, M. A. Helvie, and M. Goodsitt, "A regional registration technique for automated analysis of interval changes of breast lesions," Proc. SPIE **3338**, 118–131 (1998).
[17] S. S. Gopal, H.-P. Chan, T. E. Wilson, M. A. Helvie, N. Petrick, and B.

Sahiner, "A regional registration technique for automated interval change analysis of breast lesions on mammograms," Med. Phys. **26**, 2669–2679 (1999).

[18] L. M. Hadjiiski, H. P. Chan, B. Sahiner, N. Petrick, M. A. Helvie, and S. Sanjay-Gopal, "Automated identification of breast lesions in temporal pairs of mammograms for interval change analysis," Radiology **213(P)**, 229–230 (1999).

[19] L. M. Hadjiiski, H. P. Chan, B. Sahiner, N. Petrick, M. A. Helvie, S. Paquerault, and C. Zhou, "Interval change analysis in temporal pairs of mammograms using a local affine transformation," Proc. SPIE **3979**, 847–853 (2000).

[20] L. M. Hadjiiski, H. P. Chan, B. Sahiner, N. Petrick, and M. A. Helvie, "Automated registration of breast lesions in temporal pairs of mammograms for interval change analysis—Local affine transformation for improved localization," Med. Phys. **28**, 1070–1079 (2001).

[21] W. F. Good, B. Zheng, Y. H. Chang, Z. H. Wang, and G. S. Maitz, "Generalized procrustean image deformation for substraction of mammograms," Proc. SPIE **3661**, 1562–1573 (1999).

[22] S. S. Gopal, H.-P. Chan, B. Sahiner, N. Petrick, T. E. Wilson, and M. A. Helvie, "Evaluation of interval change in mammographic features for computerized classification of malignant and benign masses," Radiology **205(P)**, 216–■■■ (1997).

[23] L. M. Hadjiiski, B. Sahiner, H. P. Chan, N. Petrick, M. A. Helvie, and M. Gurcan, "Computer-aided classification of malignant and benign breast masses by analysis of interval change of features in temporal pairs of mammograms," Radiology **217(P)**, 435–■■■ (2000).

[24] L. M. Hadjiiski, B. Sahiner, H. P. Chan, N. Petrick, M. A. Helvie, and M. Gurcan, "Analysis of temporal change of mammographic features for computer-aided characterization of malignant and benign masses," Proc. SPIE **4322**, 661–666 (2001).

[25] B. Sahiner, H. P. Chan, N. Petrick, L. M. Hadjiiski, M. A. Helvie, and S. Paquerault, "Active contour models for segmentation and characterization of mammographic masses," The Fifth International Workshop on Digital Mammography Proceedings, 2000, IWDM-2000, pp. 357–362.

[26] B. Sahiner, H.-P. Chan, N. Petrick, M. A. Helvie, and L. M. Hadjiiski, "Improvement of mammographic mass characterization using spiculation measures and morphological features," Med. Phys. **28**, 1455–1465 (2001).

[27] M. M. Galloway, "Texture classification using gray level run lengths," Comput. Graph. Image Process. **4**, 172–179 (1975).

[28] N. Petrick, H. P. Chan, B. Sahiner, and M. A. Helvie, "Combined adaptive enhancement and region-growing segmentation of breast masses on digitized mammograms," Med. Phys. **26**, 1642–1654 (1999).

[29] M. J. Norusis, *SPSS for Windows Release 6 Professional Statistics* (SPSS, Chicago, IL, 1993).

[30] M. M. Tatsuoka, *Multivariate Analysis, Techniques for Educational and Psychological Research*, 2nd ed. (Macmillan, New York, 1988).

[31] S. S. Rao, *Optimization: Theory and Applications* (Wiley Eastern, 1979).

[32] F. A. Lootsma, *Numerical Methods for Non-linear Optimization* (Academic, New York, 1972).

[33] C. E. Metz, "ROC methodology in radiologic imaging," Invest. Radiol. **21**, 720–733 (1986).

[34] C. E. Metz, B. A. Herman, and J. H. Shen, "Maximum-likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data," Stat. Med. **17**, 1033–1053 (1998).

Automated registration of breast lesions in temporal pairs of mammograms for interval change analysis—local affine transformation for improved localization

Lubomir Hadjiiski,[a] Heang-Ping Chan, Berkman Sahiner, Nicholas Petrick, and Mark A. Helvie
Department of Radiology, University of Michigan, Ann Arbor, Michigan 48109

(Received 15 August 2000; accepted for publication 9 April 2001)

Analysis of interval change is important for mammographic interpretation. The aim of this study is to evaluate the use of an automated registration technique for computer-aided interval change analysis in mammography. Previously we developed a regional registration technique for identifying masses on temporal pairs of mammograms. In the current study, we improved lesion registration by including a local alignment step. Initially, the lesion position on the prior mammogram was estimated based on the breast geometry. An initial fan-shaped search region was then defined on the prior mammogram. In the second stage, the location of the fan-shaped region on the prior mammogram was refined by warping, based on an affine transformation and simplex optimization in a local region. In the third stage, a search for the best match between the lesion template from the current mammogram and a structure on the prior mammogram was carried out within the search region. This technique was evaluated on 124 temporal pairs of mammograms containing biopsy-proven masses. Eighty-seven percent of the estimated lesion locations resulted in an area overlap of at least 50% with the true lesion locations and an average distance of 2.4 ± 2.1 mm between their centroids. The average distance between the estimated and the true centroid of the lesions on the prior mammogram over all 124 temporal pairs was 4.2 ± 5.7 mm. The registration accuracy was improved in comparison with our previous study that used a data set of 74 temporal pairs of mammograms. This improvement in accuracy resulted from the improved geometry estimation and the local affine transformation. © *2001 American Association of Physicists in Medicine.*
[DOI: 10.1118/1.1376134]

Key words: mammography, interval change, computer-aided diagnosis, breast cancer, affine transformation

I. INTRODUCTION

Mammography is currently the most effective method for early breast cancer detection.[1,2] One of the important techniques used by radiologists in mammographic interpretation to detect developing malignancy is analysis of interval changes.[3,4] A variety of computer-aided diagnosis (CAD) techniques have been developed to detect mammographic abnormalities and to distinguish between malignant and benign lesions. We are studying the use of CAD techniques to assist radiologists in interval change analysis.

Sallam *et al.*[5] have proposed a warping technique for mammogram registration based on manually identified control points. A mapping function was calculated for mapping each point on the current mammogram to a point on the prior mammogram. Brzakovic *et al.*[6] have investigated a three-step method for comparison of the most recent and the prior mammograms. They first registered two mammograms using the method of principal axis, and partitioned the current mammogram using a hierarchical region-growing technique. Translation, rotation, and scaling were then used for registration of the partitioned regions. Vujovic *et al.*[7] have proposed a multiple-control-point technique for mammogram registration. They first determined several control points independently on the current and prior mammograms based on the intersection points of prominent anatomical structures in the breast. A correspondence between these control points was established based on a search in a local neighborhood around the control point of interest.

The previous techniques depend on the identification of control points. However, because the breast is mainly composed of soft tissue that can change over time, there are no obvious landmarks on mammograms. The crossing line structures are often fibrous tissue from different depths of the breast which overlap in a projection image. These crossing points are not invariant landmarks on different mammograms. Because of the elasticity of the breast tissue, there is large variability in the positioning and compression used in mammographic examination. As a result, the relative positions of the breast tissues projected onto a mammogram vary from one examination to the other. Techniques that depend on identification of control points may not be generally applicable to registration of breast images.

Gopal *et al.*[8–10] and Hadjiiski *et al.*[11] have developed a multistage technique that defines the transformation to locally map the position of the mass on a current mammogram to that of the prior mammogram. A local search for the mass is then performed on the prior mammogram. Good *et al.*[12] also have developed a technique that defines a transforma-

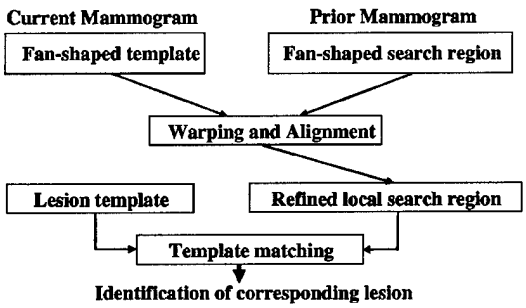


FIG. 1. Block diagram of the regional registration technique.

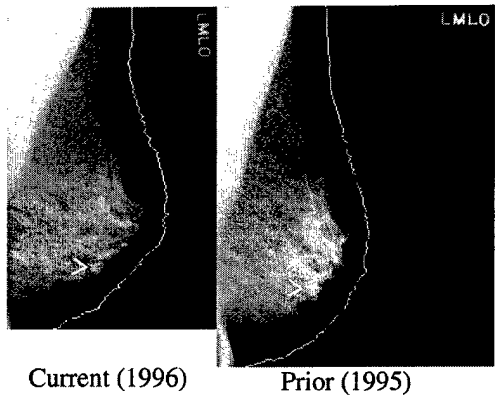


FIG. 2. An example of a pair of current and prior mediolateral oblique mammograms in our data set. The arrows point to the masses on the current and the prior mammograms. The white lines represent the breast boundary determined by the automated boundary detection procedure.

tion to map all points from the current mammogram onto a prior mammogram. The current mammogram is then subtracted from the prior mammogram.

The goal of our research is to develop a technique for computerized analysis of temporal differences between a mass on the most recent mammogram and a prior mammogram of the same view. The computer algorithm will assist radiologists in quantifying interval changes and thus distinguishing between benign and malignant masses for CAD. When fully developed, the technique will be applied to a mass on the current mammogram either identified by the radiologist or by an automated mass detection program, thus the interval change analysis can be an integrated part of an automated CAD system. In this study, we focused on the development of an automated registration technique that localizes the corresponding mass on the prior mammogram when the mass on the current mammogram is known. Therefore, we used radiologist-identified mass location on the current mammogram as a starting point and that on the prior mammogram as the ground truth for evaluation of the registration technique. A local registration technique was developed based on an affine transformation and simplex optimization and its usefulness in improving the localization of the mass on the prior mammogram was investigated.

II. REGISTRATION TECHNIQUE

A multistage regional registration technique was developed for identifying corresponding masses on temporal pairs of mammograms. The block diagram of the regional registration technique is shown in Fig. 1. In the first stage, an initial fan-shaped search region was defined on the prior mammogram based on the mass location on the current mammogram. In the second local alignment stage, the location of the search region on the prior mammograms was first refined by maximizing a correlation measure between a template of the fan-shaped region centered at the mass extracted from the current mammogram and the breast structures on the prior mammogram. The affine transformation in combination with simplex optimization was then employed to warp this local region and further improve the correlation. In the final stage, a search for the best match between the lesion template from the current mammogram and a structure on the prior mammogram was carried out within the refined search region. A more detailed explanation for each of the stages will be presented in the following subsections.

A. Stage 1—Initial estimate of search region

We have modified our previous method to define a fan-shaped search region on the prior mammogram. Initially an automated procedure is used to detect the breast boundary on the mammograms (Fig. 2). The location of the mass on the current mammogram is determined in a polar coordinate system with the nipple as the origin. By using the radial distance R_{curr} between the nipple and mass centroid, $|\mathbf{NM}|$, an arc is drawn which intersects the breast boundary at points **A** and **B** (Fig. 3). Three angles are estimated at the radial distance R_{curr}: The angle β between **NM** and **NA**, the angle φ between **NM** and **NB**, and the angle θ between **NA** and **NB** ($\theta=\beta+\varphi$). The location of the mass is determined by R_{curr} and the angle β or φ. The angle θ is the breast width at the radial distance R_{curr}. Using the radial distance R_{curr} to draw an arc centered at the nipple centroid on the prior mammogram, $\mathbf{N}'$, the two intersect points $\mathbf{A}'$ and $\mathbf{B}'$ with the breast boundary on the prior mammogram are determined. The

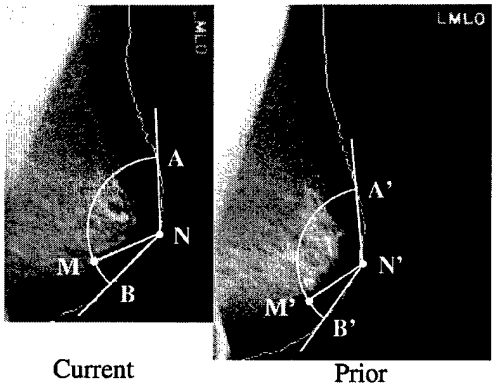


FIG. 3. Initial estimation of the mass location on the prior mammogram, based on the nipple-mass centroid distance and an angular distance from the breast periphery on the current mammogram.

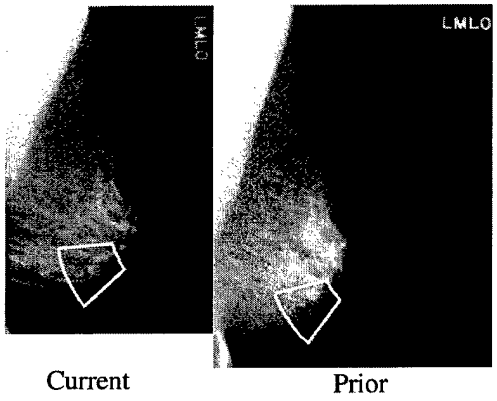


FIG. 4. Definition of an initial fan-shaped search region on the prior mammogram and a fan-shaped template on the current mammogram.

angle θ_p between the axes $|\mathbf{N}'\mathbf{A}'|$ and $|\mathbf{N}'\mathbf{B}'|$ is estimated. An angular scaling factor α can be calculated as the ratio of the prior and the current angles, $\alpha=\theta_p/\theta$.

In order to predict the angular location of the mass on the prior mammogram, the smaller angle between β and φ is selected as the angular coordinate of the mass on the current mammogram. The smaller angle is used because we found by experiment that it produces a smaller angular deviation error than using the larger angle. The angular deviation error is defined as the angle between the axis connecting the nipple and the true mass centroid and the axis connecting the nipple and the predicted mass centroid on the prior mammogram. The selected angle, multiplied by the angular scaling factor α, is used as the predicted angle from the corresponding axis on the prior mammogram. The radial distance R_{curr} is used to predict the radial position of the mass on the prior mammogram.

An initial fan-shaped search region is then defined on the prior mammogram centered at the predicted location of the mass centroid (Fig. 4). The size of the fan-shaped region is estimated previously[10] to have the form $\epsilon=k_1+k_2/R_{\text{curr}}$ and $\delta=k_3$, where 2ϵ determines the angular width and 2δ determines the radial length of the fan-shaped region. The constants k_1,k_2, and k_3 were chosen experimentally such that the estimated fan-shaped regions will essentially include all mass centroids on the prior mammograms. A fan-shaped template centered at the mass is also defined on the current mammogram. More details on defining the fan-shaped region can be found in Appendix A and in Ref. 10.

B. Stage 2—Refinement of search region by warping and alignment

The second stage combined two procedures. First, the location of the search region on the prior mammograms was refined by maximizing a correlation measure between the fan-shaped template extracted from the current mammogram and the breast structures on the prior mammogram. The template was shifted pixel by pixel within the initial fan-shaped search region and a correlation measure was calculated at each pixel location. The pixel location providing the maximum correlation is used as the center of a refined search region. This is basically a template matching operation. Second, the affine transformation in combination with simplex optimization was iteratively used to warp the fan-shaped template and further maximize the correlation measure with the breast structures on the prior mammogram.

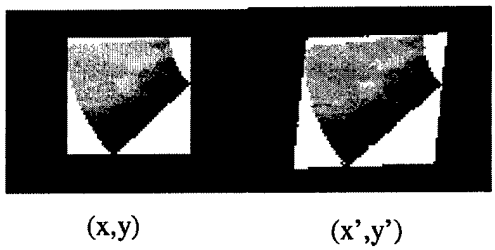


FIG. 5. The fan-shaped template (x,y) and the warped fan-shaped template (x',y') by the affine transformation.

1. Affine transformation

An affine transformation[13] is a linear transformation combining scaling, rotation, and translation. A two-dimensional affine transformation is defined as follows:

$$x'=ax+by+c,$$
$$y'=dx+ey+f, \tag{1}$$

where (x,y) are the original coordinates, (x',y') are the transformed coordinates, and a, b, d, e, c, f are the transformation coefficients. The coefficients a, b, d, e determine a scaling and a rotation, and the coefficients c and f determine a translation. The result of applying the affine transformation of Eq. (1) in combination with the simplex optimization (described below) to refine the fan-shaped search region is shown in Fig. 5. Since the affine transformation is linear, the transformed object is linearly resized and rotated. This can be observed from the edges of the bounding box of the fan-shaped region (white box in Fig. 5). After the transformation the edges are still straight lines, however, the corner angles are different from 90 degrees and the lengths of the lines are linearly scaled.

2. Nonlinear simplex optimization

The nonlinear simplex optimization by Nelder and Mead[14,15] is used to adjust the coefficients a, b, c, d, e, and f and to warp the fan-shaped template, thereby maximizing the correlation between the template and a breast structure on the prior mammogram. This optimization defines a hyper-polygon. For each vertex an error function is calculated. The polygon is then "rolled" towards the minimum. The movement of the polygon (towards the minimum) is obtained by reflection in the direction opposite to the vertex with the maximal error. Figure 5 shows the result of application of the affine transformation whose coefficients were obtained by the nonlinear simplex optimization. A more detailed discussion on this optimization method can be found in Appendix B and Refs. 14 and 15.

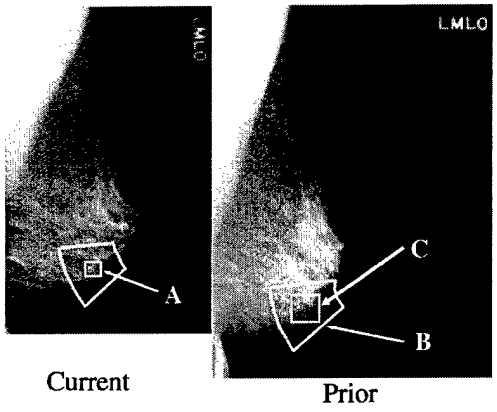


FIG. 6. A refined search region was defined on the prior mammogram. A search for the best match between the mass template from the current mammogram and a structure on the prior mammogram was carried out within the refined search region. (A—mass template on current mammogram, B—warped fan-shaped region from current mammogram, C—refined search region).

3. Stage 3—Mass template matching and localization of corresponding lesion

At this stage a new search region with a reduced size is defined on the prior mammogram (Fig. 6). The reduced size of the search region is determined experimentally by iterative adjustment of the size of the rectangular region targeting the improvement of the final result. A template containing the mass is extracted from the current mammogram. The mass location on the prior mammogram is then determined by maximizing the correlation between the template and a structure within the search region (Fig. 7).

III. DATA SET

A set of 124 temporal pairs of mammograms containing biopsy-proven masses on the current mammograms was used to examine the performance of this approach. Different mammographic views of the same breast were also included. There were a total of 221 mammograms obtained from 54 cases. Temporal pairs were formed using the temporal sequence from the corresponding view. Some cases contained mammograms of multiple years and a combination of the mammograms from different prior years with the current-year mammogram formed multiple temporal pairs. Thirty five of the mammograms were digitized with a LUMISYS DIS-1000 laser scanner at a pixel resolution of 100 μm$\times$100 μm and 4096 gray levels. The digitizer was calibrated so that gray level values were linearly proportional to the optical density (OD) within the range of 0.1–2.8 OD units, with a slope of 0.001 OD/pixel value. Outside this range, the slope of the calibration curve decreased gradually. The OD range of the digitizer was 0–3.5. The remaining 186 mammograms were digitized with a LUMISCAN 85 laser scanner at a pixel size of 50 μm$\times$50 μm and 4096 gray levels. The digitizer was calibrated so that the gray level values were linearly proportional to the OD within the range of 0–4 OD units, also with a slope of 0.001 OD/pixel value. Output from both digitizers was linearly converted so that large pixel value corresponded to a low-optical density. In order to process the mammograms digitized with these two different digitizers, the images were first averaged using a filter that has constant weights over the entire filter kernel and then were down-sampled. This filter will be referred to as a box filter. The images digitized with the LUMISCAN 85 digitizer were averaged with a 16$\times$16 box filter and then were down-sampled by a factor of 16. The images digitized with the LUMISYS DIS-1000 digitizer were averaged with an 8$\times$8 box filter and then were down-sampled by a factor of 8. Therefore, all resulting images had a pixel size of 800 μm$\times$800 μm.

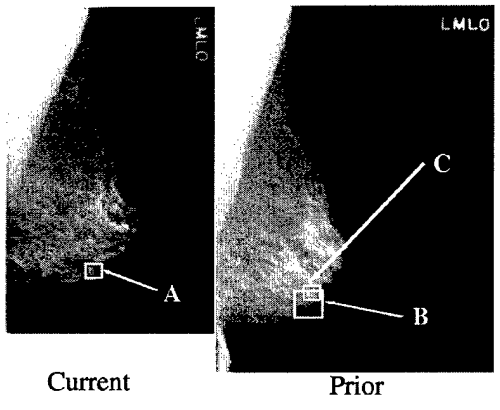


FIG. 7. Final identification of the corresponding mass on the prior mammogram. (A—Mass template on current mammogram, B—Refined search region, C—Identified mass location).

The 54 cases contained 53 biopsy proven and one follow-up masses. The 221 mammograms contained different mammographic views and multiple years of the masses including the year when the biopsy was performed. Of the 124 temporal pairs of mammograms 73 were malignant and 51 benign. A malignant temporal pair consists of a biopsy proven malignant mass or a mass that was followed up and was found to be malignant when a biopsy was performed in a future year. Of the 124 temporal pairs of mammograms, 63 were CC-view pairs, 48 were MLO-view pairs, and 13 were lateral-view pairs. A Mammography Quality Standards Act (MQSA)-approved radiologist read the original mammogram to identify the mass and provide description of its characteristics. The radiologist defined a bounding box around the mass and marked the nipple location on every film.

The radiologist also measured the mass sizes, defined as the longest dimension of the mass, both on the current and prior mammograms. In Figs. 8(a) and 8(b) the mass sizes on the current mammograms were plotted against those on the prior mammograms for the malignant and the benign temporal pairs, respectively. Only 103 temporal pairs were plotted (54 malignant and 49 benign) due to the fact that the masses on the prior mammograms in the remaining 21 temporal pairs were too subtle for the radiologist to estimate their boundaries. On average the malignant masses appear to have a larger increase in size than the benign masses. The mean increase in size from prior to current for the malignant masses is 4.2 mm compared to 1.6 mm for the benign masses (p=0.008). The correlation coefficient is 0.71 for the malig-

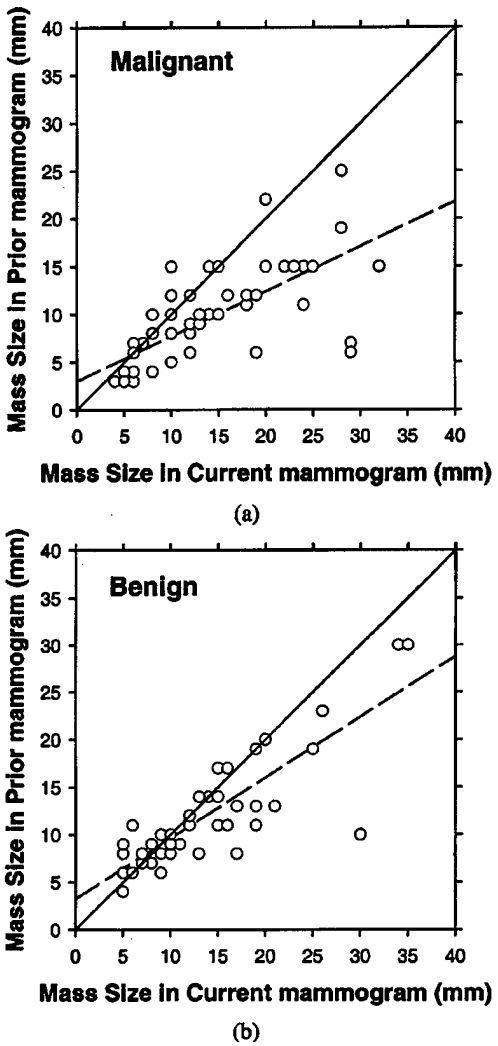


FIG. 8. Mass sizes measured by an MQSA-approved radiologist on the current mammograms plotted against those on the prior mammograms for (a) 54 malignant and (b) 49 benign temporal pairs. The diagonal line on the graph represents the case when the current and the prior mass sizes are identical. The dashed lines are the linear regression lines defined by $y=0.469x+3.012$ for (a) and by $y=0.638x+3.242$ for (b). The correlation coefficient for malignant masses is 0.71 and for benign masses is 0.83.

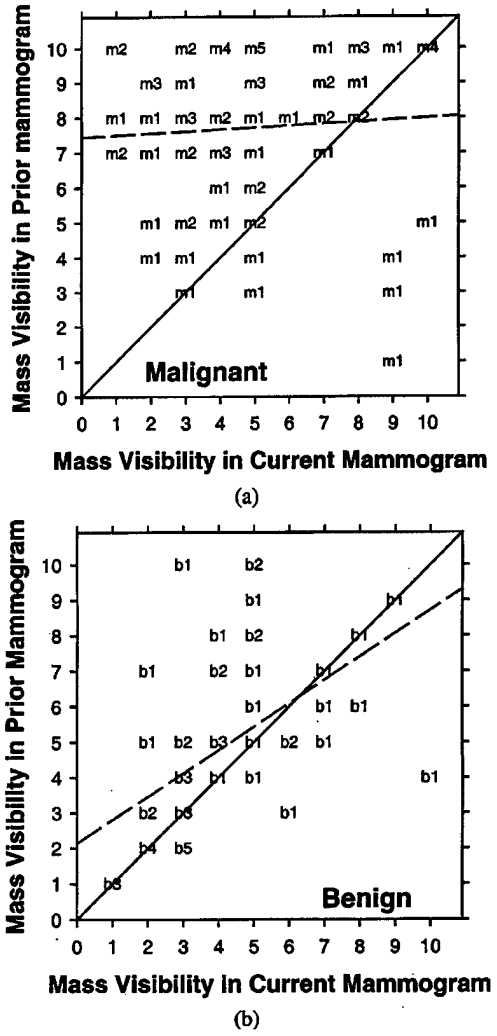


FIG. 9. Visibility of the masses on the current mammogram plotted against those on the prior mammogram for (a) malignant and (b) benign temporal pairs. The visibility was rated on a 10-point discrete scale (1=most obvious, 10=subtlest). Because many of the data points overlap, we indicate the number of points with the same rating by a number next to the symbol (*m* or *b*). The diagonal line on the graph represents the case when the current and the prior mass sizes are identical. The dashed lines are the linear regression lines defined by $y=0.055x+7.44$ for (a) and by $y=0.658x+2.138$ for (b). The correlation coefficient for malignant masses is 0.06 and for benign masses is 0.54.

nant masses and 0.83 for the benign masses [Fig. 8(a) and 8(b)].

The radiologist also rated the visibility of the masses on the mammograms relative to those encountered in clinical practice on a 10-point scale, with one represents the most obvious and 10 the subtlest masses. The visibility of the masses on the current mammogram is plotted against those on the prior mammogram in Fig. 9 for the 73 malignant and 51 benign temporal pairs. Generally, the malignant masses were less visible on the prior mammograms while the visibility of the benign masses was found to be more similar. The mean difference in visibility between the prior and the current mammograms for the malignant masses is 2.8 compared to 0.7 mm for the benign masses ($p=0.0002$). The correlation coefficient is 0.06 for malignant masses and 0.54 for benign masses [Figs. 9(a) and 9(b)]. For most of the temporal pairs the time interval between the current and the prior mammogram was 12 months (Fig. 10).

IV. EVALUATION METHODS

The accuracy of the multistage regional registration was analyzed in terms of two measures. The first measure is the overlap area between the estimated and the true lesions on the prior mammogram. The fractions of registered temporal pairs that could provide an accuracy of over 50% area overlap and over 75% area overlap were examined. The second measure is the average Euclidean distance between the centroids of the estimated and the true lesion locations.

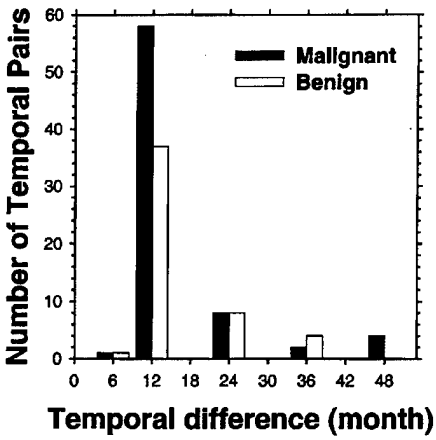


FIG. 10. Temporal interval between the current and the prior mammograms for the 124 temporal pairs in our data set.

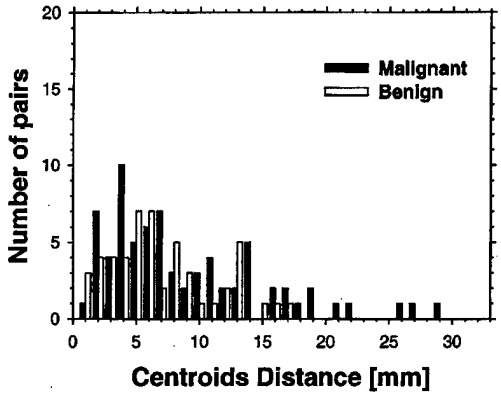


FIG. 11. Distribution of Euclidean distance error between the initial estimate of the mass centroid location on the prior mammogram and the center of the bounding box of the mass provided by the radiologist for the malignant and benign pairs after the first detection stage.

V. REGISTRATION RESULTS

A. Stage 1—Initial estimate of search region

At this stage an initial estimation of the mass location on the prior mammogram was carried out based on the geometrical position of the mass on the current mammogram. Based on observation of the radial deviation errors and the angular deviation errors, the fan-shaped search region was estimated to be $\epsilon = 0.25 + 5/R_{curr}$ radians and $\delta = 20$ mm. This definition of the fan-shaped search region resulted in an average search area of 1462 mm^2 on the prior mammograms. For the 124 temporal image pairs used in this study, the Euclidean distance between the initial estimate of the centroid location of the corresponding structure on the prior mammogram and the center of the bounding box of the mass provided by the radiologist was estimated. For the 124 temporal image pairs, the average Euclidean distance error of the initial estimate was 8.4±5.4 mm. The error distributions for both the malignant and the benign pairs are shown in Fig. 11. At this initial stage, 57% of the estimated lesion locations resulted in an area overlap of at least 50% with the true lesion locations and 27% resulted in an area overlap of at least 75% (Fig. 12).

B. Stage 2—Refinement of search region by warping and alignment

At the second stage, the location of the search region on the prior mammogram was first refined by maximizing a correlation measure between the fan-shaped template extracted from the current mammogram and the breast structures on the prior mammogram. The affine transformation in combination with simplex optimization was then employed to warp this local region. For the 124 temporal image pairs, the average Euclidean distance error after the second stage was 7.5±5.4 mm. At this stage, 59% of the estimated lesion locations resulted in an area overlap of at least 50% with the true lesion locations, and 36% resulted in an area overlap of at least 75%. The average Euclidean distance error at this stage was reduced compared to that of the first stage, however, it did not achieve statistical significance ($p = 0.07$).

After the simplex optimization, the search region was reduced to a constant size of 24 mm×24 mm (=576 mm^2) centered at the refined fan-shaped region for every prior mammogram.

C. Stage 3—Mass template matching and localization of corresponding lesion

At this final stage, a search for the best match between the lesion template from the current mammogram and a structure on the prior mammogram was carried out within the refined search region. This template matching resulted in 87% of the estimated lesion locations having an area overlap of at least 50% with the true lesion locations. The distributions of the Euclidean error for the malignant and the benign temporal pairs are shown in Fig. 13. The average distance between the estimated and the true centroids of the lesions on the prior mammogram for all 124 pairs was 4.2±5.7 mm with a maximum of 31.6 mm. These results are summarized in Table I. For the 87% of the temporal pairs with 50% overlap, the

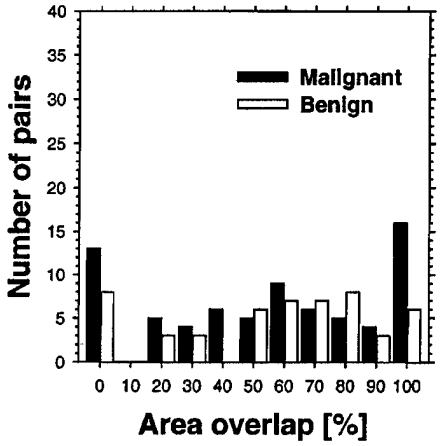


FIG. 12. Distribution of the area overlap between the estimated and the true lesion locations for 124 temporal pairs after the first detection stage.

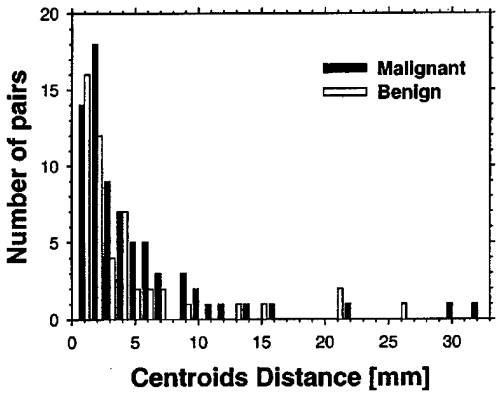


FIG. 13. Distribution of Euclidean distance error between the estimate of the mass centroid location on the prior mammogram and the center of the bounding box of the mass provided by the radiologist for the malignant and benign pairs after the final detection stage.

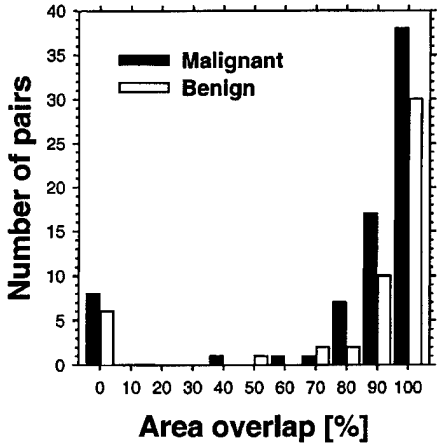


FIG. 14. Distribution of the area overlap between the estimated and the true lesion locations for 124 temporal pairs after the final detection stage.

average distance between the estimated and the true centroids of the lesions on the prior mammogram was 2.4±2.1 mm with a maximum of 10.2 mm. When a more stringent criterion of 75% overlap is imposed, 82% of the masses on the prior mammograms are considered to be localized (Fig. 14). For the 82% of the temporal pairs with 75% overlap, the average distance between the estimated and the true centroids of the lesions on the prior mammogram was 2.2±1.9 mm with a maximum of 10.2 mm. The average Euclidean distance error at this stage was significantly reduced compared to the error of the first stage (p=0.000 001) and the error of the second stage (p=0.000 001).

D. Study of the importance of the stage 2 procedures

The effect of the two procedures at Stage 2 on the registration accuracy was studied. We removed them one at a time and evaluated the registration results. When the first correlation procedure was removed, the average Euclidean distance error increased to 5.6±8.2 mm in the final stage. Only 81% of the estimated lesion locations resulted in an area overlap of at least 50% with the true lesion locations and 75% resulted in an area overlap of at least 75% with the true lesion locations. When the second warping procedure was removed, the average Euclidean distance error increased to 5.0±6.3 mm in the final stage. Only 82% of the estimated lesion locations resulted in an area overlap of at least 50% with the true lesion locations and 76% resulted in an area overlap of at least 75% with the true lesion locations.

TABLE I. The Euclidean distance between the true and the estimated centroids of the mass on the prior mammogram for the three detection stages.

		Overall	50% overlap	75% overlap
Stage 1	Mean distance	8.4 mm	5.6 mm	4.5 mm
	Standard. Deviation.	5.4 mm	2.8 mm	2.6 mm
	Max. distance	29.0 mm	16.2 mm	13.8 mm
Stage 2	Mean distance	7.5 mm	4.9 mm	3.9 mm
	Standard. Deviation.	5.4 mm	3.0 mm	2.6 mm
	Max. distance	32.0 mm	16.9 mm	11.6 mm
Stage 3	Mean distance	4.2 mm	2.4 mm	2.2 mm
	Standard. Deviation	5.7 mm	2.1 mm	1.9 mm
	Max. distance	31.6 mm	10.2 mm	10.2 mm

VI. DISCUSSION

The approach proposed here has simplified the first stage compared to our previous method.[10] In the previous method, the distances between the nipple and the breast centroid on the current and prior mammograms were determined and used to estimate a radial scaling factor. The angular location of the mass was measured from the nipple–breast centroid axis. A global alignment procedure was used for determination of the breast centroids. With our new approach we eliminated the scaling for the radial distance between the nipple and the mass location of the prior mammogram. The breast periphery was used as a reference for the estimation of the angular position of the mass. Therefore, there was no need to determine the breast centroids on the current and the prior mammograms and the global alignment procedure could be eliminated. This is possible because the local alignment step provides better compensation for the displacement of the corresponding masses on the current and the prior mammogram caused by different compression and positioning of the breast.

It was found that the estimation of the angular position from the breast periphery allowed more precise localization of the mass position on the prior mammogram compared to our previous method where the angular position of the mass was estimated based on the nipple–breast centroid axis.[10] There is a large variability in the estimation of the breast centroid location because the extend of the breast imaged on the mammogram at the chest wall and at the axillary tail in the MLO view depends on the breast positioning and compression. This causes an uncertainty in defining the region to calculate the breast centroid. In the previous study using 74 temporal pairs, the estimated Euclidean distance error at the first stage was 9.8±6.0 mm. The fan-shaped search region was defined as $\epsilon=0.35+5/r$, resulting in an average area of 1865 mm^2 for the fan-shaped search region. In the current

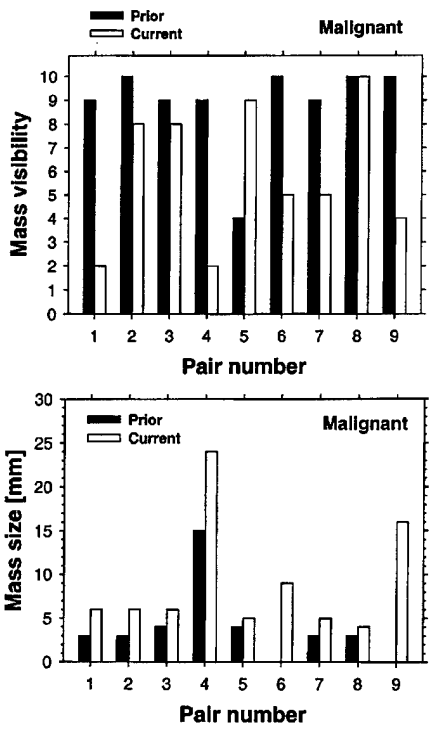


FIG. 15. The visibility and the mass size of nine malignant temporal pairs having area overlap less than 50%. The radiologist was unable to define the prior mass sizes of pairs 6 and 9 due to the subtlety of these masses.

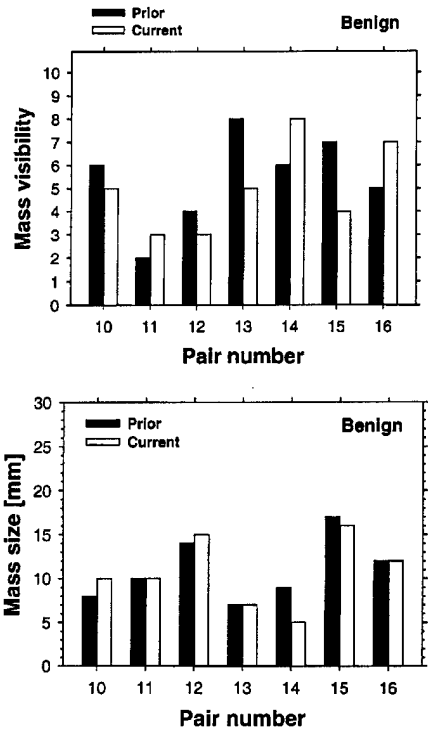


FIG. 16. The visibility and the mass size of seven benign temporal pairs having area overlap less than 50%.

study, the estimated Euclidean distance error at the first stage was reduced to 8.4±5.4 mm even though the data set was increased to 124 temporal pairs of mammograms. This allows the fan-shaped region to be reduced to $\epsilon=0.25+5/r$, resulting in an average fan-shaped search area of 1462 mm^2 on the prior images. The reduction of the search area improves the chance of correctly localizing the mass on the prior mammogram.

The second stage combined two procedures: First the localization of the search region on the prior mammograms was refined by maximizing a correlation measure between the fan-shaped template extracted from the current mammogram and the breast structures on the prior mammogram. The affine transformation in combination with simplex optimization was then employed to warp and locally align the template with the breast structures. Both procedures improved the detection process. When one of these procedures was removed the registration results deteriorated, as discussed in the Results section.

With these improvements, the accuracy of the current regional registration technique is improved over the previous method.[10] The current technique produced an average Euclidean distance error of 4.2±5.7 mm, compared to 5.4±7.5 mm when the previous technique was applied to the current data set. This difference is statistically significant ($p=0.03$). 82% of the estimated lesion locations resulted in an area overlap of at least 75% with the true lesion locations compared with 72% when applying the previous technique to the current data set. It is interesting to note that, of the 21 "masses" on the prior mammograms that the experienced radiologist could not confidently define the mass and measure its size, our registration technique localize 19 of them with an area overlap greater than 50%.

The average distance between the estimated and the true centroid of the lesions on the prior mammogram for the subset of temporal pairs having 50% overlap is about half of that of the entire data set (Table I). The maximum distance for this subset is about 1/3 of that for the entire data set.

With the current regional registration technique, 16 temporal pairs (13% of 124 temporal pairs) have an area overlap less than 50%. Twelve of the 16 computer estimated locations do not overlap at all with the radiologist's identified locations, and the other four pairs have an overlap between 1% and 49%. Seven of them are benign and nine are malignant. A major cause of the misregistration was that the mass was small and subtle and a breast structure within the search region had a higher correlation with the mass template from the current mammogram. Figures 15 and 16 show the visibility ratings and sizes of these misregistered masses. Eight of the nine misregistered malignant masses have visibility ratings of 9 or 10 and sizes below 5 mm. The misregistered benign masses are somewhat more obvious and larger in sizes than the malignant ones. Since many of the masses on the prior mammograms were not interpreted as a mass without reference to the current mammograms, the automatic registration with template matching would be difficult with these masses if the search region contains normal, but dense breast structures. We are currently investigating the application of local mass detection in the search region to focus

template matching to a few suspicious areas. Morphological and texture features will be extracted from the potential mass areas to provide additional matching information in the feature space.

The interval change analysis, when fully developed, will be one of the functions provided in an integrated CAD system. The mass on the current mammogram can be detected by an automated mass detection algorithm or identified by a radiologist. The CAD system will then analyze whether the mass is an existing or a newly developed lesion and will estimate its likelihood of malignancy. We are developing methods for characterization of malignant and benign masses based on analysis of interval changes in the mass features.[16] Investigation of criteria to determine whether a mass exists on the prior mammogram is underway. If the mass is a newly developed lesion on the current mammogram, it will then undergo a single-exam analysis by the CAD system.

VII. CONCLUSION

We are developing an automated registration technique for analysis of interval change of a mass from a previous mammographic exam to the current one. In this study we found that a local affine transformation in combination with nonlinear simplex optimization can improve the localization and reduce the size of the search region. With the improved method, 87% of the estimated lesion locations in 124 randomly selected temporal pairs resulted in an area overlap of at least 50% with the true lesion locations. When the threshold for correct localization was set to 75% area overlap, 82% of the temporal pairs still exceeded this threshold. The average distance between the estimated and the true centroids of the lesions on the prior mammogram over all pairs was 4.2 $\pm$5.7 mm. The registration accuracy of the current method has been improved in comparison with that of our previous method[10] even though the data set was increased from 74 pairs to 124 pairs. This improvement is obtained mainly from the second stage affine transformation and simplex optimization. Additional studies are currently underway to develop a feature matching method to further improve lesion localization.

ACKNOWLEDGMENTS

This work is supported by a Career Development Award from the U.S. Army Medical Research and Material Command (DAMD 17-98-1-8211) (L.H.), a USPHS Grant CA 48129, a USAMRMC grant (DAMD 17-96-1-6254), and a USAMRMC Career Development Award DAMD 17-96-1-6012 (B.S.). The content of this publication does not necessarily reflect the position of the government and no official endorsement of any equipment or product of any companies mentioned in the publication should be inferred.

APPENDIX A: DEFINITION OF THE FAN-SHAPED REGION ON THE PRIOR MAMMOGRAM

Refer to Figs. 3 and 4, the fan-shaped region on the prior mammogram is drawn based on the nipple centroid on the prior mammogram, $\mathbf{N}'$, as the center of the coordinate system. The two bounding arcs are drawn using the radial distances $R_{curr}+\delta$ and $R_{curr}-\delta$, both centered at $\mathbf{N}'$. The two sides of the fan-shaped region are bounded by two radial lines that form angles ϵ and $-\epsilon$ with the line $|\mathbf{N}'\mathbf{M}'|$. Thus the initial fan-shaped search region is centered as the predicted location of the mass centroid $\mathbf{M}'$ on the prior mammogram (Fig.4).

The constants k_1, k_2, and k_3 were chosen experimentally based on analysis of the angular deviation errors and the corresponding radial deviation errors for the 124 temporal pairs. The radial deviation error is defined as the difference between the predicted and the true distance of the mass from the nipple on the prior mammogram. The constants k_1, k_2 are obtained in such a way that ϵ is the smallest upper bound that can enclose all angular deviation errors for all radial distances (R_{cur}) and all temporal pairs. The selection of the parametric form of ϵ was discussed in detail in Ref. 10. It reduced ϵ at larger R_{cur}. The constant k_3 was chosen to be equal to the maximum radial deviation error.

APPENDIX B: SIMPLEX OPTIMIZATION

An optimization problem can be defined as an error function that has to be minimized by iterative selection of the values of the function parameters n. We can define $n+1$ dimensional space, where n dimensions (degree of freedom) correspond to the error function parameters, and one dimension is the error function itself. When the optimization function is calculated for all possible values of the n parameters, and error surface in $(n+1)$-dimensional space will be obtained. Usually the error functions for the real world applications are complex and nonlinear and the corresponding error surfaces contain local minima.

The nonlinear simplex optimization by Nelder and Mead[14,15] defines a hyper-polygon with $n+1$ vertexes in a $(n+1)$ dimensional space. For each vertex the error function is calculated. The polygon is then "rolled" towards the minimum. The movement of the polygon (towards the minimum) is obtained by reflection in the direction opposite to the vertex ($\mathbf{K}$) with the maximal error. To achieve this the center of masses ($\mathbf{L}$) of the hyper-polygon vertexes is calculated. A line $\mathbf{KL}$ connects the center of the masses with the vertex with the maximal error. The new vertex ($\mathbf{K}'$) is obtained by central projection of the vertex $\mathbf{K}$ on the line $\mathbf{KL}$ with center $\mathbf{L}$ and $|\mathbf{K}'\mathbf{L}|=t|\mathbf{KL}|$. The coefficient t determines how far the new vertex will be projected and what the corresponding size of the hyper-polygon will be. The larger the hyper-polygon is, the easier it will avoid ("roll over") the local minima on the error surface. However, it will be difficult to get close to the global minimum if its size is too large. On the other hand, although a small hyper-polygon will allow it to get to a close proximity to the global minimum, it is more likely to be trapped in a local minimum. The magnitude of the coefficient t is controlled adaptively by the Nelder and Mead algorithm. In case a large reduction in the error is detected for the new vertex, the magnitude of t is increased. In case the error is found to be increased for the new vertex, the magnitude of t is decreased.

The this paper, the nonlinear simplex optimization by Nelder and Mead was used to adjust the coefficients *a, b, c, d, e,* and *f* and to warp the fan-shaped template, thereby maximizing the correlation (C) between the template and a breast structure on the prior mammogram. Therefore, the dimensionality of the space was 7: Six parameters to be adjusted and the error function to be minimized was defined as $1-C$.

[a]Author to whom correspondence should be addressed. Telephone: (734) 647-8552. Fax: (734) 647-8557. Electronic mail: lhadjisk@umich.edu

[1] H. C. Zuckerman, "The role of mammography in the diagnosis of breast cancer," in *Breast Cancer, Diagnosis and Treatment*, edited by I. M. Ariel and J. B. Cleary (McGraw-Hill, New York, 1987), pp. 152–172.

[2] L. Tabar and P. B. Dean, "The control of breast cancer through mammographic screening: What is the evidence," Radiol. Clin. N. Amer. **25**, 993–1005 (1987).

[3] L. W. Bassett, B. Shayestehfar, and I. Hirbawi, "Obtaining previous mammograms for comparison: usefulness and costs," Amer. J. Roentgenology **163**, 1083–1086 (1994).

[4] E. A. Sickles, "Periodic mammographic follow-up of probably benign lesions: results in 3183 consecutive cases," Radiology **179**, 463–468 (1991).

[5] M. Sallam and K. Bowyer, "Detecting abnormal densities in mammograms by comparison with previous screenings," in *Digital Mammography '96*, edited by K. Doi, M. L. Giger, R. M. Nishikawa, and R. A. Schmidt (Elsevier, Amsterdam, 1996).

[6] D. Brzakovic, N. Vujovic, M. Neskovic, P. Brzakovic, and K. Fogerty, "Mammogram analysis by comparison with previous screenings," in *Digital Mammography '96*, edited by K. Doi, M. L. Giger, R. M. Nishikawa, and R. A. Schmidt (Elsevier, Amsterdam, 1996).

[7] N. Vujovic and D. Brzakovic, "Establishing the correspondence between control points in pairs of mammographic images," IEEE Trans. Image Process. **6**, 1388–1399 (1997).

[8] S. Sanjay-Gopal, H. P. Chan, B. Sahiner, N. Petrick, T. Wilson, and M. Helvie, "Evaluation of interval change in mammographic features for computerized classification of malignant and benign masses," Radiology **205(P)**, 216 (1997).

[9] S. Sanjay-Gopal, H. P. Chan, N. Petrick, T. Wilson, B. Sahiner, M. Helvie, and M. Goodsitt, "A regional registration technique for automated analysis of interval changes of breast lesions," Proc. SPIE **3338**, 118–131 (1998).

[10] S. Sanjay-Gopal, H. P. Chan, T. E. Wilson, M. A. Helvie, N. Petrick, and B. Sahiner, "A regional registration technique for automated interval change analysis of breast lesions on mammograms," Med. Phys. **26**, 2669–2679 (1999).

[11] L. Hadjiiski, H. P. Chan, B. Sahiner, N. Petrick, M. A. Helvie, and S. S. Gopal, "Automated identification of breast lesions in temporal pairs of mammograms for interval change analysis," Radiology **213(P)**, 229–230 (1999).

[12] W. Good, B. Zheng, Y. H. Chang, X. Wang, and G. Maitz, "Generalized procrustean image deformation for subtraction of mammograms," Proc. SPIE **3661**, 1562–1573 (1999).

[13] L. Quan and T. Kanade, "Affine structure from line correspondence with uncalibrated affine cameras," IEEE Trans. Pat. Anal. Machine Intel. **19**(8), 834–845 (1997).

[14] S. S. Rao, *Optimization: Theory and Applications* (Wiley, New York, 1979).

[15] *Numerical Methods for Non-Linear Optimization*, edited by F. A. Lootsma (Academic, New York, 1972).

[16] L. Hadjiiski, B. Sahiner, H. P. Chan, N. Petrick, M. A. Helvie, and M. Gurcan, "Computer-aided classification of malignant and benign breast masses by analysis of interval change of features in temporal pairs of mammograms," Radiology **217(P)**, 435 (2000).

Digital Mammography:

Observer Performance Study of the Effects of Pixel Size on the Characterization of Malignant and Benign Microcalcifications[1]

Heang-Ping Chan, PhD, Mark A. Helvie, MD, Nicholas Petrick, PhD, Berkman Sahiner, PhD, Dorit D. Adler, MD
Chintana Paramagul, MD, Marilyn A. Roubidoux, MD, Caroline E. Blane, MD, Lynn K. Joynt, MD
Todd E. Wilson, MD, Lubomir M. Hadjiiski, PhD, Mitchell M. Goodsitt, PhD

Rationale and Objectives. The authors performed this study to evaluate the effects of pixel size on the characterization of mammographic microcalcifications by radiologists.

Materials and Methods. Two-view mammograms of 112 microcalcification clusters were digitized with a laser scanner at a pixel size of 35 μm. Images with pixel sizes of 70, 105, and 140 μm were derived from the 35-μm–pixel size images by averaging neighboring pixels. The malignancy or benignity of the microcalcifications had been determined with findings at biopsy or 2-year follow-up. Region-of-interest images containing the microcalcifications were printed with a laser imager. Seven radiologists participated in a receiver operating characteristic (ROC) study to estimate the likelihood of malignancy. The classification accuracy was quantified with the area under the ROC curve (A_z). The statistical significance of the differences in the A_z values for different pixel sizes was estimated with the Dorfman-Berbaum-Metz method and the Student paired t test. The variance components were analyzed with a bootstrap method.

Results. The higher-resolution images did not result in better classification; the average A_z with a pixel size of 35 μm was lower than that with pixel sizes of 70 and 105 μm. The differences in A_z between different pixel sizes did not achieve statistical significance.

Conclusion. Pixel sizes in the range studied do not have a strong effect on radiologists' accuracy in the characterization of microcalcifications. The low specificity of the image features of microcalcifications and the large interobserver and intraobserver variabilities may have prevented small advantages in image resolution from being observed.

Key Words. Breast neoplasms, calcification; breast radiography, comparative studies; breast radiography, technology; receiver operating characteristic curve (ROC).

Breast cancer is one of the leading causes of death in women between the ages of 40 and 55 years. In the United States, the mortality rate for breast cancer in women is the second highest of all cancers, and breast cancer was estimated to account for 16% of all cancer deaths in 1998 (1). Studies have indicated that early detection and treatment improve the chances of survival for breast cancer patients. At present, mammography is the only proven method that consistently demonstrates minimal breast cancers (2,3). The image quality with conventional mammography, however, is limited by the dynamic range of screen-film systems. The contrast sensitivity of screen-film mammograms is very poor in the overpen-

Acad Radiol 2001; 8:454–466

[1] From the Department of Radiology, University of Michigan Hospital, UH B1F510, Ann Arbor, MI 48109-0030. Received September 6, 2000; revision requested October 3; revision received January 10, 2001; accepted January 11. Supported by U.S. Public Health Service grant CA 48129 and by a grant from the U.S. Army Medical Research and Materiel Command DAMD 17-96-1-6254. B.S. and L.M.H. supported by Career Development Awards DAMD 17-96-1-6012 and 17-98-1-8211, respectively. **Address correspondence to** H.P.C.

The content of this publication does not necessarily reflect the position of the government and no official endorsement of any equipment and product of any companies mentioned in the publication should be inferred.

© AUR, 2001

etrated periphery and the underpenetrated dense fibroglandular tissue regions on the breast image. Recently, a digital mammography system has received U.S. Food and Drug Administration clearance for clinical use. Digital mammography detectors are expected to provide a wider dynamic range than screen-film systems and, thus, increase the contrast sensitivity in the periphery and dense regions of the breast. The improved image quality is expected to lead to an improvement in the accuracy of breast cancer diagnosis.

The spatial resolution of current digital detectors is generally lower than that of screen-film systems. Digital detectors used in the full-field digital mammography systems that are commercially available or under development have pixel sizes in the range of 40 × 40 μm to 100 × 100 μm, which correspond to nominal spatial resolution of about 12 line pairs per millimeter to 5 line pairs per millimeter. In contrast, the spatial resolution of mammographic screen-film systems generally exceeds 20 line pairs per millimeter. Higher-resolution digital detectors require smaller pixel sizes. The development of digital detectors with small pixel sizes, however, is not only technologically demanding, but the requirements for image transmission, archiving, and display increase rapidly as the matrix size increases. The trade-offs between spatial resolution and cost and efficiency are important considerations in the development of digital mammography systems. The maximum pixel size acceptable for performing mammography without reducing the detectability of subtle breast cancers is unknown.

One of the important signs of breast cancer is clustered microcalcifications (4), which can be seen on mammograms in 30%–50% of breast cancers (5–8). Microcalcifications associated with early breast cancers are usually smaller than about 500 μm. Among the image features that may indicate the presence of breast cancer, microcalcifications are the smallest. Therefore, the spatial resolution required for the detection and characterization of subtle microcalcifications on mammograms may be regarded as the lower bound for the resolution of a mammographic detector. In a previous receiver operating characteristic (ROC) study (9), we compared the detectability of subtle microcalcifications on original screen-film mammograms with that on mammograms digitized at a pixel size of 100 μm with an optical drum scanner. We found that the detection accuracy of subtle microcalcifications decreased when radiologists read the digitized images. Although the detection accuracy improved after the digitized images were enhanced with unsharp mask filtering, it remained lower than that with the original screen-film mammograms. In another study (10), we investigated the detectability of individual microcalcifications on digitized mammograms by using a computer program. Those results also indicated a reduction in detectability when the digitization pixel size increased from 35 to 140 μm.

Malignant microcalcifications may exhibit linear and branching shapes, as well as variations in shape and size within a cluster. Benign microcalcifications tend to be round and smooth, with relatively uniform shapes and sizes within a cluster. The visibility of the detailed shapes is dependent on the spatial resolution of the image recording system. Therefore, it is generally believed that a higher spatial resolution is required to differentiate malignant from benign microcalcifications than to detect microcalcifications. Results of some recent studies, however, indicate that this may not be the case. Karssemeijer et al (11) performed an ROC study to compare the accuracy of classifying microcalcifications on original screen-film mammograms with that on images digitized at a pixel size of 100 μm and viewed on a display monitor. They found that there was no statistically significant difference in the classification accuracy between the two reading conditions. Kallergi et al (12) also performed an ROC study to compare the detection and classification of clustered microcalcifications at three reading conditions: screen-film mammograms, images digitized at a pixel size of 105 μm and displayed on a monitor, and wavelet-enhanced digitized images displayed on a monitor. They found that the detection with the original mammograms was much better than that with the digitized mammograms displayed on a monitor; the use of wavelet enhancement, however, reduced the difference. The characterization of microcalcifications was not substantially different among the three reading conditions.

We performed this ROC study to evaluate the effects of pixel size on the characterization of malignant and benign microcalcifications on digitized mammograms. Two-view mammograms were digitized and displayed as laser-printed film images at four pixel sizes ranging from 35 to 140 μm. Seven radiologists experienced in mammography estimated the likelihood of malignancy. The dependence of classification accuracy on pixel size was analyzed with ROC methodology.

MATERIALS AND METHODS

Data Set

Digital mammograms were obtained by digitizing screen-film mammograms with a laser film scanner. One

hundred twelve microcalcification clusters were selected from 100 patient cases in the Breast Imaging Division at the University of Michigan with approval from the Institutional Review Board. Two-view mammograms of each cluster were digitized. The two views included a craniocaudal view and a mediolateral oblique or lateral view.

Forty of the microcalcification clusters were proved at biopsy to be malignant, and 65 were proved at biopsy to be benign. The other seven clusters were considered to be benign based on findings of at least 2 years of follow-up. Of the 40 malignant clusters, 25 were ductal carcinoma in situ. The distribution of the sizes (the longest dimension) of the microcalcification clusters is shown in Figure 1. The longest dimension of the clusters ranged from 2.0 to 18.0 mm (mean, 6.4 mm). Seven of the benign microcalcifications and five of the malignant microcalcifications were spread over an area larger than 20 mm in diameter and, thus, were considered to be diffuse. The data set included microcalcifications with a range of subtleties. The subtlety of the microcalcifications was rated by a radiologist experienced in mammography (M.A.H.) on a scale of 1 (obvious) to 10 (subtle) relative to the visibility range of microcalcifications encountered in clinical practice. The subtlety ratings are shown in Figure 2. The malignant and benign microcalcifications were similarly distributed, with the benign microcalcifications slightly more subtle than the malignant clusters.

All mammograms were digitized at a pixel size of 35 × 35 μm with 12-bit gray levels by using a laser scanner (DIS-1000, Lumisys, Los Altos, Calif). The digitizer had an optical density range of about 0 to 3.5. It was calibrated such that the optical density on film was linearly proportional to the pixel value at 0.001 optical density units per pixel value in the optical density range of about 0–2.8. The pixel values of the images were linearly inverted so that large pixel values represented a low optical density. The resolution of the scanner was evaluated by digitizing test film images with line pair patterns. It was found that line pair patterns up to 14.3 line pairs per millimeter could be resolved on the digitized image (10).

A 1,024 × 1,024-pixel region of interest (ROI) containing the microcalcifications was extracted from the digitized image. Except for clusters that were close to the chest wall or in the breast periphery, the extracted cluster was usually centered within the ROI. Diffuse microcalcifications that were larger than the ROI were truncated to the size of the ROI. Microcalcification images digitized with pixel sizes of 70, 105, and 140 μm were simulated from the image with the 35-μm pixel size by averaging

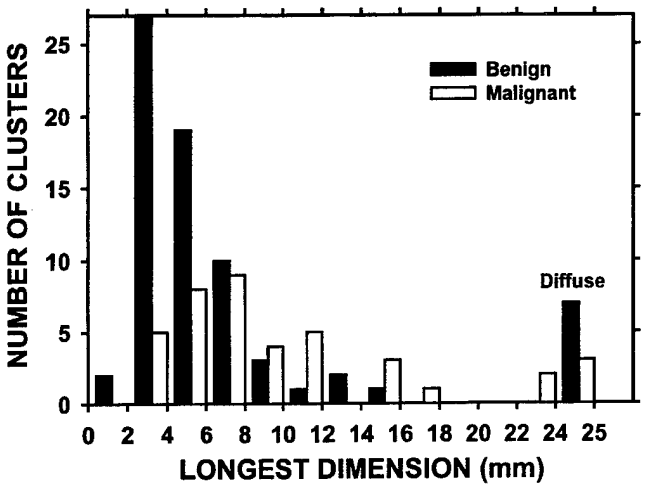


Figure 1. The size distribution of the microcalcification clusters.

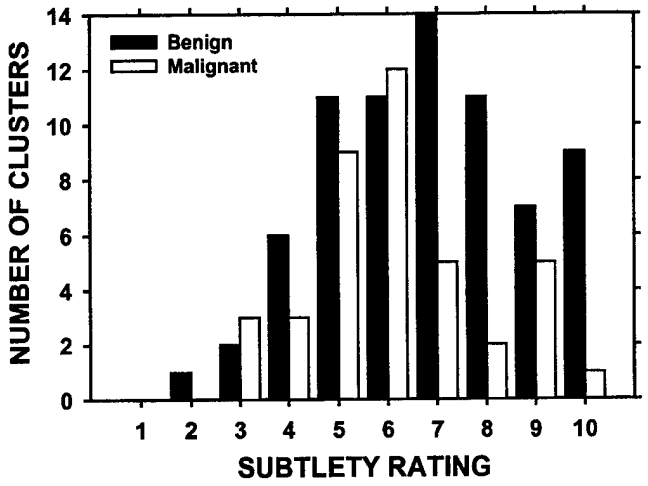


Figure 2. Distribution of the subtlety ratings for the microcalcification clusters. *1* = most obvious, *10* = most subtle.

2 × 2, 3 × 3, and 4 × 4 neighboring pixels, respectively. Because ROIs of different pixel sizes were derived from the same digitized image, there would not be differences in image quality caused by the reproducibility of digitization. The actual size of all ROIs corresponded to an area of 35.8 × 35.8 mm on the original mammograms, regardless of the pixel sizes.

Because the use of display monitors to view images can introduce variables that may be difficult to control, we printed the ROI images on film with a laser imager (model 969HQ; Imation, Oakdale, Minn) for the observer performance study. To reduce the effects of image size on characterization, the ROIs with the three larger pixel sizes (ie, smaller matrix sizes for the same ROI image) were enlarged to the same printed size as that of the 35-μm

Table 1
Confidence Rating Scale

Rating	Likelihood of Malignancy (%)	Suspicion Level	BI-RADS Category
1	0–2	Benign, probably benign	2, 3
2	3–20	Suspicious, with low probability of malignancy	4
3	21–30	Suspicious, with low probability of malignancy	4
4	31–40	Suspicious, with moderate probability of malignancy	4
5	41–50	Suspicious, with moderate probability of malignancy	4
6	51–60	Suspicious, with moderate probability of malignancy	4
7	61–70	Suspicious, with moderate probability of malignancy	4
8	71–80	Highly suggestive (high probability) of malignancy	5
9	81–90	Highly suggestive (high probability) of malignancy	5
10	91–100	Highly suggestive (high probability) of malignancy	5

pixel size images by means of interpolation. Sixteen interpolation schemes were available from the laser imager interface software. To choose the best interpolation scheme for this study, we printed an image of a cluster containing microcalcifications of different sizes and shapes at pixel sizes of 70, 105, and 140 μm by using the 16 interpolation schemes. The images of 35-μm pixel size were also printed. A radiologist who was qualified under the requirements of the Mammography Quality Standards Act visually compared the printed images and numbered his top three choices for each set of images. The radiologist was not aware of the specific schemes. After the decision was made, he informed us that his criteria were a balance between blockiness and blurriness on the enlarged image and its similarity to the 35-μm image. The experiment was repeated two times, with the sessions separated by more than a month. The top two choices obtained from the two readings were consistent. The top two choices were essentially indistinguishable so that one of them was used to print the images. The chosen scheme was a convolution interpolation that filled the interpolated pixels with smooth weighted gray levels of the adjacent pixels.

The printed ROIs measured 84 × 84 mm, which corresponded to a pixel pitch of about 82 μm for the laser imager. The printed ROIs were therefore magnified by a factor of about 2.3 compared with their size on the original screen-film mammograms. Because radiologists routinely view microcalcifications with a magnifying lens or on a magnified spot mammogram, however, the magnification should not affect the classification of the microcalcifications. To maintain the same displayed contrast for images of different pixel sizes, the four ROIs of different pixel sizes were printed on the same piece of film and, thus, developed at the same time. This minimized the effects of any potential fluctuations in the printer calibration and in the development conditions of the laser film on the relative density and contrast of the printed images.

Observer Performance Study

Seven radiologists, all of whom were qualified under the requirements of the Mammography Quality Standards Act to read and routinely interpret mammograms, participated as observers. The radiologists had 3–20 years experience in mammographic interpretation. Because there were 112 ROIs and four pixel sizes for each ROI, a total of 448 images were read by each observer. The two views of each cluster at the same pixel size were read side by side. The observers were not informed of the prevalence of malignant cases or the proportion of biopsy cases. Each observer read the ROI images in four reading sessions. Every reading session was separated from the previous one by at least 2 weeks. In each session, one-quarter of the images of each pixel size were read. Each case appeared once and only once in each session. The reading orders of the images in each pixel size were counterbalanced such that, on average, no images of a given pixel size were read in a given order (eg, read first by the observers) more often than images of any other pixel sizes. The reading order of the images was randomized differently for each observer. This systematic randomization reading scheme minimized any potential learning effects on the reading results (13). The observers were allowed unlimited reading time.

The likelihood that the microcalcifications were malignant was rated with a 10-point confidence rating scale. The confidence rating scale was designed and related to the Breast Imaging Reporting and Data System (BI-RADS) ratings by an experienced radiologist, as shown in Table 1. A likelihood of malignancy of less than 2% for benign or probably benign mammographic abnormalities was chosen on the basis of the studies by Sickles (14,15). The observers also rated the subtlety of each case according to a 10-point scale (1 = most obvious, 10 = most subtle)

on the basis of their perception of the cluster relative to their experience with clinical cases.

A table showing the rating scale and the corresponding BI-RADS category was available to the observers for reference during the reading sessions. A training session was conducted before each reading session to familiarize the observers with the rating scales. Three malignant and three benign clusters not included in the test set were used in the training session. After the rating scales were explained to the observer, he or she rated each cluster as described earlier. They were told the biopsy outcome of the cluster after rating each training case. There was no "truth" for the subtlety rating. The subtlety rating was recorded as additional information about each radiologist's subjective impression of a cluster.

Analysis of Classification Accuracy

The confidence ratings of the likelihood of malignancy were analyzed with ROC analysis (13). The two class distributions were assumed to be binormal, and an ROC curve was fitted to the confidence ratings on the basis of maximum likelihood estimation. The ROC curve represents the relationship between the true-positive fraction (sensitivity) and the false-positive fraction (1 − specificity) as the confidence threshold varies. An ROC curve was generated for each observer and for images of each pixel size. The classification accuracy was quantified by using the area under the ROC curve (A_z). The average ROC curve for each reading condition was derived by averaging the slope and intercept parameters of the individual observers' fitted ROC curves. The statistical significance of the differences in the ROC curves for two pixel sizes was estimated by using the Dorfman-Berbaum-Metz (DBM) method for multireader, multicase ROC data (16) and the Student paired t test for the observer-specific paired A_z values. The paired t test takes into account the statistical variation of the readers, whereas the DBM method includes both the reader variation and case sample variation with an analysis-of-variance approach. Therefore, the results with the DBM method can be generalized to the population of readers as well as the case samples. In addition, the bootstrap method developed by Beiden et al (17) was used to analyze the components of variances in this classification task.

RESULTS

Images of a small malignant microcalcification cluster and a benign cluster from our data set obtained with a pixel size of 35 μm are shown in Figure 3a and 3b, respectively. The craniocaudal and mediolateral oblique views of the same cluster are shown side by side. Figure 4 shows one view of a malignant cluster with all four pixel sizes. Slight blurring of the image details and the noise can be observed as the pixel size increases from 35 to 140 μm.

The ROC curves for the seven radiologists reading the images with 35-μm pixel size are shown in Figure 5. The ROC curves are spread over a relatively wide range. The A_z values for the radiologists are listed in Table 2 and plotted in Figure 6. The standard deviation of the A_z ranges from 0.05 to 0.07, as estimated with the LABMRMC program. Only one of the seven radiologists demonstrated a higher classification accuracy with the 35-μm images than with the 70- or 105-μm images. The A_z-versus–pixel size curve for this radiologist (reader 6) had a different trend from that of other radiologists. The A_z of another radiologist (reader 7) was basically constant over the entire range of pixel sizes studied. The average ROC curves for each pixel size were derived from the average slope and intercept parameters of the seven individual ROC curves and are plotted in Figure 7. The dependence of average A_z on pixel size is shown in Table 2. The average A_z showed a higher classification accuracy with pixel sizes of 70 and 105 μm. The differences in A_z between the different pixel sizes did not achieve statistical significance with either the DBM method (16) or the Student paired t test. Table 3 shows the P values obtained with the DBM and the paired t test when images with a pixel size of 35 μm were compared with those with pixel sizes of 70, 105, and 140 μm. The P values obtained with the two methods are very similar, which indicates that the reader variation is dominant over case variation in this classification task.

Because of the outlying trend of reader 6, we performed the analysis of the classification accuracy without this reader in an attempt to evaluate the dependence of A_z on pixel size for the majority of radiologists in our study. For these six readers, the average A_z for the four pixel sizes was 0.71, 0.74, 0.75, and 0.71, respectively. Although the trend that the radiologists had a higher classification accuracy with pixel sizes of 70 and 105 μm became more apparent, the difference in the A_z between the pixel sizes still fell short of statistical significance. The P value determined with the DBM method was .11 for the difference in A_z between 35- and 70-μm images and .12 for that between 35- and 105-μm images. The corre-

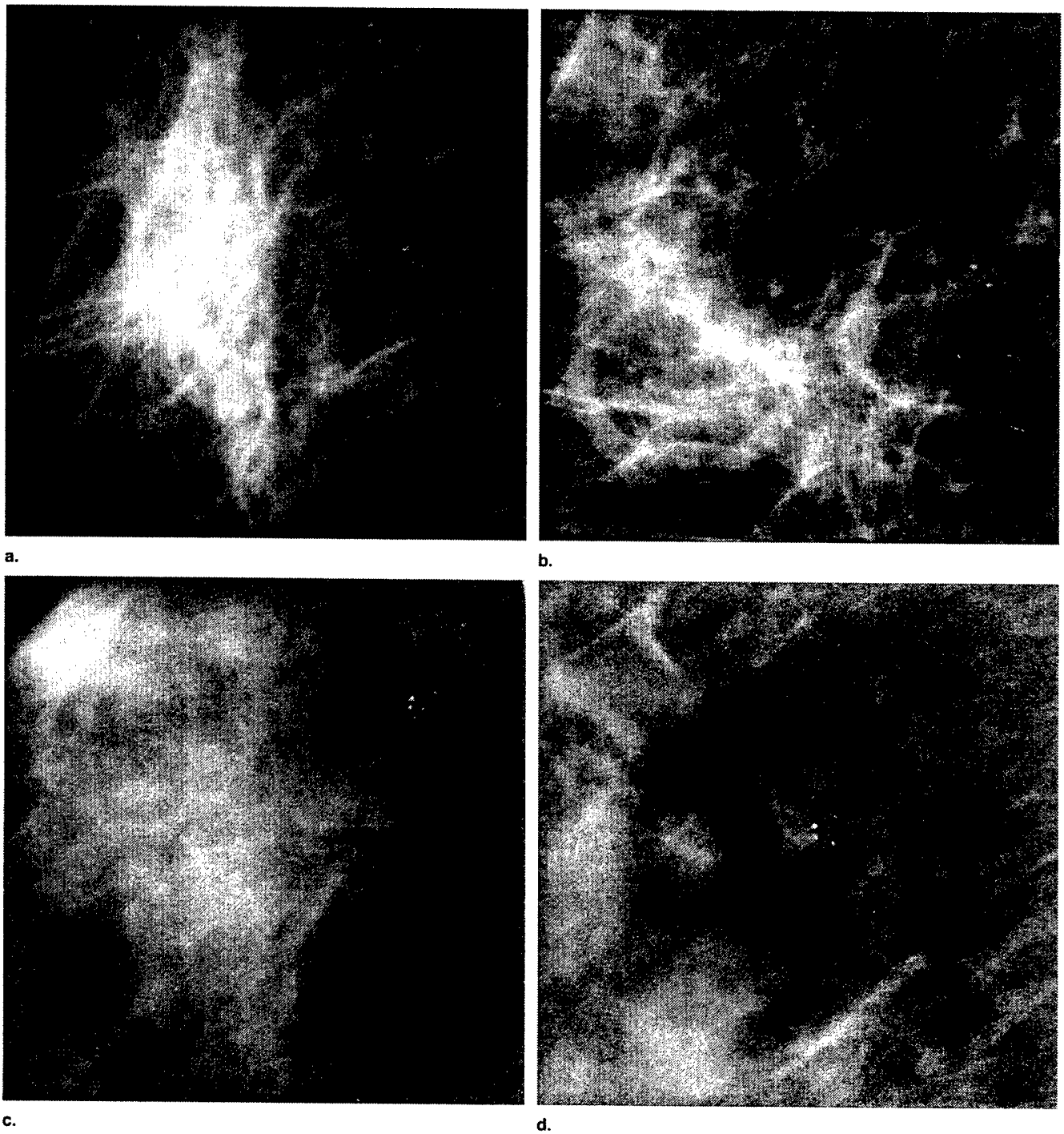


Figure 3. **(a, c)** Craniocaudal and **(b, d)** mediolateral oblique images of **(a, b)** a malignant microcalcification cluster (intraductal carcinoma) and **(c, d)** a benign cluster (sclerosing adenosis) digitized with a pixel size of 35 μm.

sponding two-tailed P values with the Student paired t test were .10 and .12, respectively.

We also analyzed the percentages of positive and negative cases for which the observers gave a confidence rating of 1 in each pixel size. A confidence rating of 1 corresponded to a 0%–2% likelihood of malignancy and BI-RADS categories of benign or probably benign (Table 1). These cases would be returned to a regular screening schedule or undergo short-interval follow-up without biopsy. The results are shown in Table 4. Each observer

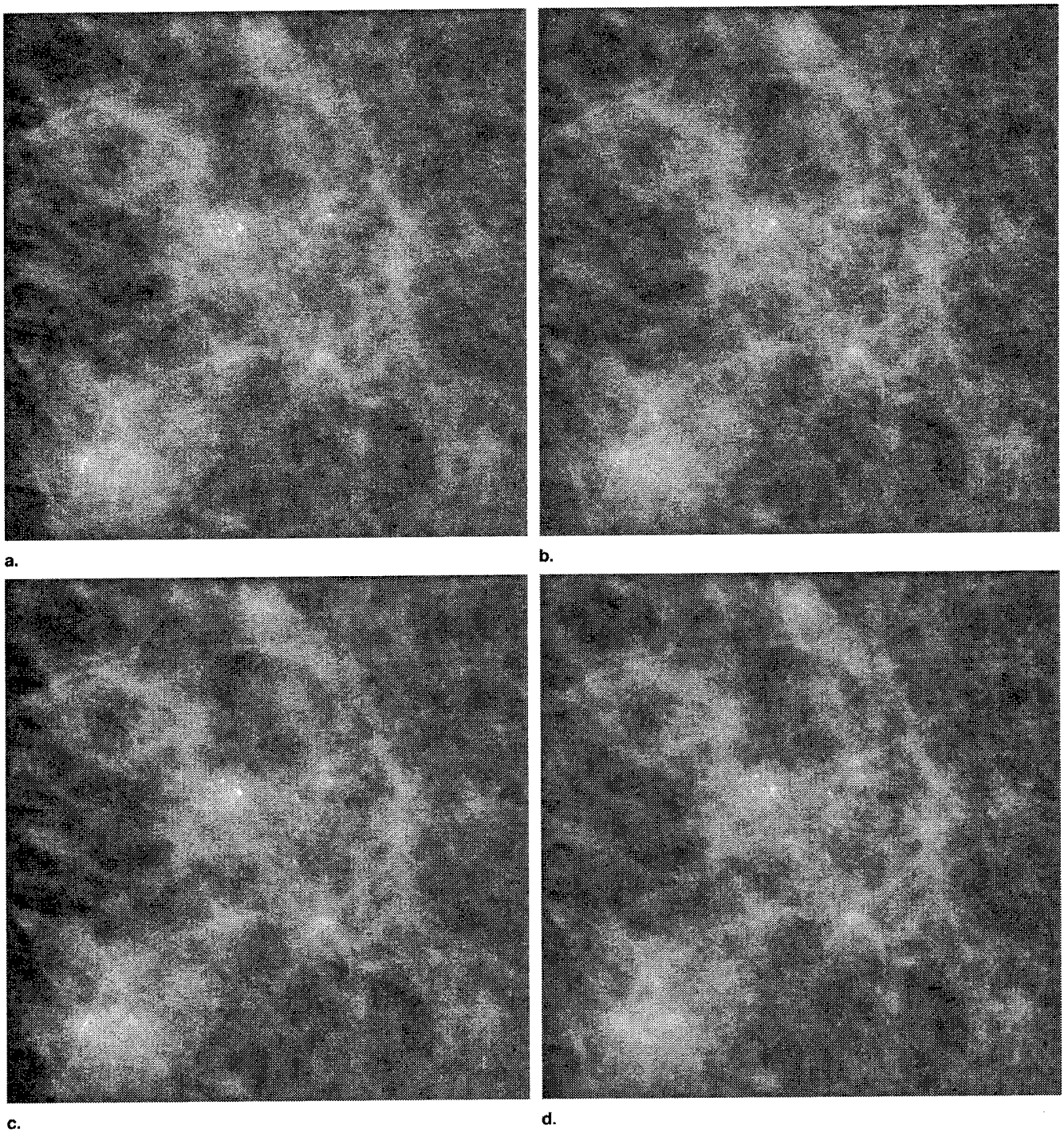


Figure 4. Lateral images of a malignant cluster (intraductal carcinoma, comedo type) at pixel sizes of **(a)** 35, **(b)** 70, **(c)** 105, and **(d)** 140 μm.

appeared to have a different threshold for suspicion. For a given observer, however, the threshold was relatively consistent among the different pixel sizes. There was no obvious trend that this threshold depended on pixel size.

DISCUSSION

The results of the ROC study indicate that the differences in the classification accuracy of microcalcifications,

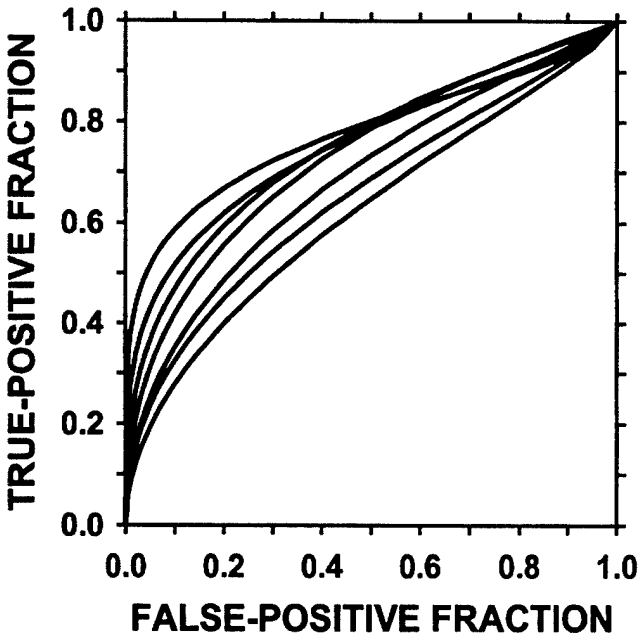


Figure 5. The ROC curves for seven radiologists in the evaluation of the images with 35-μm pixel size. The standard deviation of the A_z ranges from 0.05 to 0.07.

if any, with pixel sizes of 35–140 μm did not achieve statistical significance. Of the A_z-versus–pixel size curves from seven radiologists, only one showed that a pixel size of 35 μm provided a larger A_z than did pixel sizes of 70 and 105 μm. Although the variances in the A_z were large, this consistent trend indicates a strong likelihood that images with a 35-μm pixel size may not provide a higher accuracy in the differentiation of malignant from benign microcalcifications than those with a pixel size of 70 or 105 μm. This finding differs from the expectation that a smaller pixel size would better preserve the shape information of microcalcifications and, consequently, provide higher accuracy in the differentiation of microcalcifications on mammograms. Our findings are consistent with those of Karssemeijer et al (11) and Kallergi et al (12) who, in their ROC studies, compared the classification accuracy of microcalcifications on original screen-film mammograms with that on images digitized at a pixel size of 100 μm and viewed on a display monitor.

Beiden et al (17) recently developed a bootstrap method for analyzing the variance components in an ROC experiment. They analyzed our ROC data set and estimated the variance components and the total variance of the difference in A_z, $\sigma^2(\Delta A_z)$, for any pairing of modalities (pixel sizes), as shown in Table 5. We used these variances to estimate whether the finite sample size in our ROC study is the main factor that caused the insignificant differences between pixel sizes.

Equation (21) in the article by Beiden et al (17) shows that the total variance of ΔA_z is given as $\sigma^2(\Delta A_z) = 2[\sigma^2_{mc}(N_0/N) + \sigma^2_{mr}/R + \sigma^2_{\epsilon}(N_0/N)/R]$, where R is the number of readers; N_0 is the sample size of the current experiment; N is the sample size of a future experiment; and σ^2_{mc}, σ^2_{mr}, σ^2_{ϵ} are the modality-by-case, modality-by-reader, and effective error components of the variance, respectively. The total variance at an infinite sample size, $N \to \infty$, is thus caused only by the reader variance, as follows: $\sigma^2(N \to \infty) = 2\sigma^2_{mr}/R$. Therefore, if we can repeat the ROC experiment with an infinite sample size, the minimum observed difference in A_z between two modalities, [min $\Delta A_z(N \to \infty)$], that will allow rejection of the null hypothesis, A_z(small pixel) = A_z(large pixel), with $P < .05$ can be estimated as [min $\Delta A_z(N \to \infty)$] = $1.645 \cdot \sigma(N \to \infty)$. The values of $\sigma(N \to \infty)$ and [min $\Delta A_z(N \to \infty)$] are shown in Table 5. The z value of 1.645, which corresponds to the one-tailed P value of .05 for a normal distribution, was used in these estimations because it is expected that a smaller pixel size would provide better performance than a larger pixel size.

From the standard deviation, $\sigma(\Delta A_z)$, and the observed difference in A_z, we can estimate the maximum mean ΔA_z between two modalities. In our ROC experiment, we observed a difference of ΔA_z(observed) = A_z(small pixel) − A_z(large pixel). Because of the variance, we do not know the true population mean ΔA_z(mean) of the normal distribution from which the ΔA_z(observed) was sampled. It can be estimated, however, that we have a less than 5% chance of observing this ΔA_z value if the population mean ΔA_z(mean) of the distribution is greater than [ΔA_z(observed) − $(-1.645) \cdot \sigma(\Delta A_z)$]. This estimated bound of mean ΔA_z is denoted as [max ΔA_z(mean)] and tabulated in Table 5.

Because an increasing sample size reduces only the variance while the population mean of the distribution of ΔA_z remains the same, the [max ΔA_z(mean)] estimated earlier for a finite sample size may also be considered to be the maximum mean ΔA_z for $N \to \infty$. Comparison of the values of [max ΔA_z(mean)], $\sigma(N \to \infty)$, and [min $\Delta A_z(N \to \infty)$] in Table 5 shows that the [max ΔA_z(mean)] is approximately equal to $\sigma(N \to \infty)$ and is thus smaller than [min $\Delta A_z(N \to \infty)$] for the 35- versus 70-μm and 35- versus 105-μm image pairs when the sample size approaches infinity. Therefore, the finite sample size in our

Table 2
Summary of A_z Values

Pixel Size (μm)	Reader 1	Reader 2	Reader 3	Reader 4	Reader 5	Reader 6	Reader 7	Average*
35	0.68	0.62	0.75	0.75	0.65	0.74	0.77	0.71
70	0.73	0.71	0.77	0.80	0.64	0.65	0.77	0.73
105	0.80	0.63	0.73	0.81	0.73	0.60	0.77	0.73
140	0.69	0.64	0.68	0.80	0.68	0.74	0.76	0.71

Note.—The standard deviations of the A_z values ranged from 0.05 to 0.07.
*A_z of average ROC curve, which was obtained by averaging the slope and intercept parameters of the individual ROC curves.

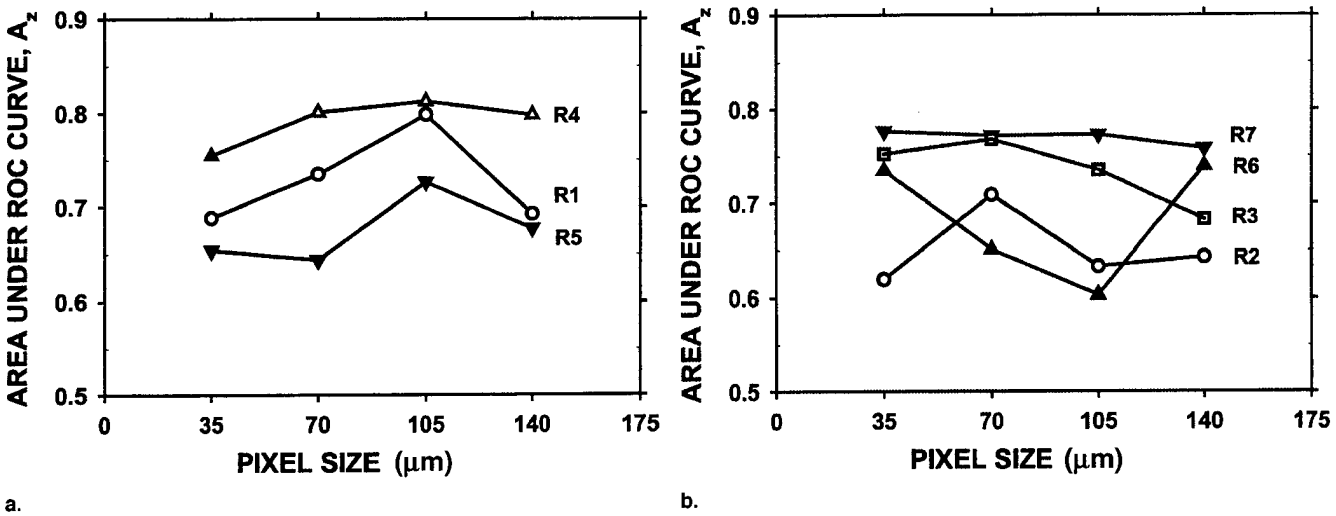


Figure 6. Dependence of the A_z on pixel size for readers **(a)** 1, 4, and 5 and **(b)** 2, 3, 6, and 7.

Table 3
Comparison of 35-μm Images with 70, 105, and 140-μm Images

	All Readers		All Readers Except Reader 6	
Pixel Size (μm)	DBM Method	Paired *t* Test	DBM Method	Paired *t* Test
35 vs 70	.51	.51	.11	.10
35 vs 105	.65	.65	.12	.12
35 vs 140	.93	.91	.96	.96

Note.—Data are two-tailed *P* values.

current ROC study is not the main contributor to the lack of statistical significance in the difference for the 35- versus 70-μm and 35- versus 105-μm image pairs. The small difference in A_z relative to the large reader variance may be the main reason we did not observe a statistically significant advantage of the 35-μm pixel size over 70- or 105-μm pixel sizes in the characterization of malignant and benign microcalcifications.

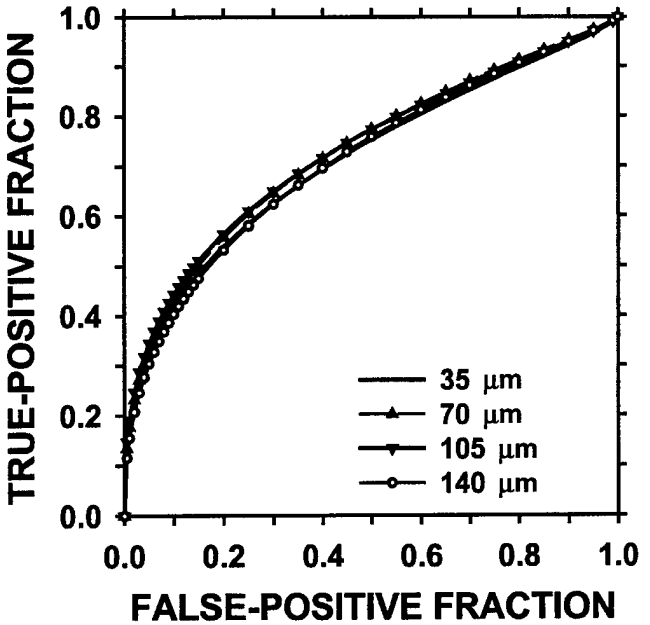


Figure 7. The average ROC curves for the four pixel sizes. Each curve was derived from the average slope and intercept parameters of the individual ROC curves from the seven radiologists.

Table 4
Percentage of Positive and Negative Cases that Received a Confidence Rating of 1

	35-μm Pixel Size		70-μm Pixel Size		105-μm Pixel Size		140-μm Pixel Size	
Reader	Negative	Positive	Negative	Positive	Negative	Positive	Negative	Positive
1	9.7	5.0	4.2	2.5	5.6	0.0	4.2	0.0
2	41.7	30.0	44.4	20.0	37.5	25.0	38.9	22.5
3	16.7	7.5	13.9	2.5	12.5	5.0	8.3	7.5
4	25.0	10.0	23.6	7.5	30.6	5.0	26.4	5.0
5	40.3	20.0	43.1	25.0	50.0	20.0	45.8	25.0
6	62.5	27.5	52.8	35.0	54.2	40.0	63.9	30.0
7	23.6	10.0	18.1	5.0	22.2	7.5	22.2	10.0

Table 5
Variance Components of the ROC Experiment

Modalities (μm)	σ^2_{mc}	σ^2_{mr}	σ^2_ϵ	$\sigma(\Delta A_z)$	ΔA_z(obs)	Max $\Delta A_z(m)$ at one-tailed, $P = .05$	$\sigma(N \to \infty)$	Min $\Delta A_z(N \to \infty)$ at one-tailed, $P = .05$
35 vs 70	−0.000009*	0.000867	0.000833	0.0216	−0.02	0.016	0.016	0.026
35 vs 105	−0.000014*	0.001803	0.000778	0.0266	−0.02	0.024	0.023	0.038
35 vs 140	0.000024	0.000488	0.000928	0.0213	−0.00	0.035	0.012	0.020
70 vs 105	0.000002	0.001213	0.000728	0.0236	−0.00	0.039	0.019	0.031
70 vs 140	0.000077	0.001195	0.000825	0.0270	0.02	0.064	0.018	0.030
105 vs 140	0.000031	0.001888	0.000763	0.0286	0.02	0.067	0.023	0.038

Note.—Data were estimated with the bootstrap method of Beiden et al (17). The total variance $\sigma^2(\Delta A_z)$ is computed from the variance components and Eq (21) of Beiden et al as $\sigma^2(\Delta A_z) = 2(\sigma^2_{mc} + \sigma^2_{mr}/R + \sigma^2_\epsilon/R)$, where R is the number of readers. Max $\Delta A_z(m)$ is the maximum mean difference in A_z between two modalities. $\sigma(N \to \infty) = (2\sigma^2_{mr}/R)^{1/2}$ is the standard deviation and Min $\Delta A_z(N \to \infty)$ is the minimum difference in A_z between two modalities that will allow rejection of the null hypothesis, A_z (small pixel) = A_z (large pixel) with $P < .05$ when the sample size N approaches infinity. The variance component σ^2_{mc} is negative in some cases due to the variance of the bootstrap estimation; the error bars tightly bracket the neighborhood of zero.

*Data are negative owing to the variance of the bootstrap estimation; their error bars tightly bracket the neighborhood of zero.

Another interesting observation can be made from the analysis of the variance components. In this classification task, the modality-by-case variance component σ^2_{mc} is consistently near zero for any of the paired comparisons. This means that even with an infinite number of readers, the variations in the two modalities will completely follow each other. It is still possible that the two modalities will have different mean performances, but cases that are more (or less) difficult with one modality will completely follow in the direction of cases that are more (or less) difficult with the other modality. This again seems to imply that the nature of the classification task is more dominant than the appearance of the image with each modality.

One aspect of the interobserver variabilities is demonstrated in Table 4, where the radiologists' decision thresholds for biopsy varied over a wide range. The large variation among the ROC curves in Figure 6 indicates that the variation among the radiologists' biopsy recommendation is not entirely caused by the use of a different decision threshold by each radiologist along similar ROC curves. This suggests that the estimation of the likelihood of malignancy of microcalcifications based on their mammographic features such as morphologic characteristics and spatial distribution pattern is very different among the radiologists. It may be noted, however, that the majority of the cases used in this ROC study had undergone biopsy so that easily distinguished benign cases had already been excluded from the case samples.

To investigate the intraobserver variabilities in the classification of microcalcifications, we repeated one reading session with three observers (readers 1–3). The distributions of the differences in the confidence ratings between the two readings of the same film of a cluster by

each radiologist are shown in Figure 8. The differences in the ratings range from −5 to +3 for reader 1, −5 to +6 for reader 2, and −3 to +4 for reader 3. This is consistent with the results of the variance analysis with the method of Beiden et al, where the reader variance was found to be an important component of the total variance for the classification of microcalcifications.

We also attempted to analyze the correlation of the estimated likelihood of malignancy when the same images were read by different radiologists. The scatter plots of the malignancy ratings by every two radiologists (not shown) were, in general, spread over wide ranges without obvious correlation. The histograms of the difference in the malignancy ratings for the same cluster between two radiologists were similar to those of the intraobserver variability shown in Figure 8, with ranges as wide as −6 to +6. There were some trends that some radiologists (eg, reader 1) tended to have higher likelihood of malignancy estimates for most clusters than did other radiologists, and some radiologists (eg, reader 6) tended to have lower suspicion for malignancy than did the others. These trends are consistent with the lower biopsy threshold of reader 1 and the higher biopsy threshold of reader 6 (Table 4).

We investigated whether the intraobserver variability in the malignancy ratings depended on the perceived subtlety of the microcalcification cluster. Because reader 3 demonstrated the smallest range of variability in the malignancy ratings among the three radiologists with whom we repeated the experiment, we plotted the relationship of the difference in the malignancy ratings between the two readings of the same cluster against the subtlety rating of the cluster for reader 3, as shown in Figure 9. There was no obvious correlation between the variability in the malignancy ratings and the perceived subtlety of the clusters.

The large inter- and intraobserver variabilities in the malignancy ratings may be a result of the fact that the radiologists usually do not have to estimate specifically the likelihood of malignancy of the clusters when they read mammograms in clinical practice. However, because their decision threshold for biopsy recommendation also varied over a wide range, as discussed above, the variabilities were not simply caused by their unfamiliarity in the estimation of the likelihood of malignancy. The variabilities may again reflect the low specificity of the image features of the microcalcifications. As can be seen from the examples in Figure 3, the appearance of a cluster of benign microcalcifications from sclerosing adenosis can be very similar to that of a malignant cluster from intraductal carcinoma.

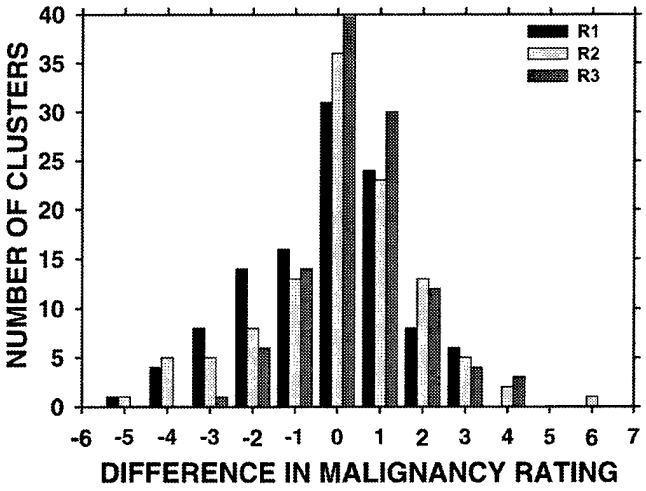


Figure 8. The distributions of the differences in the confidence ratings between the two readings of the same film of a cluster by the same radiologist for readers 1–3.

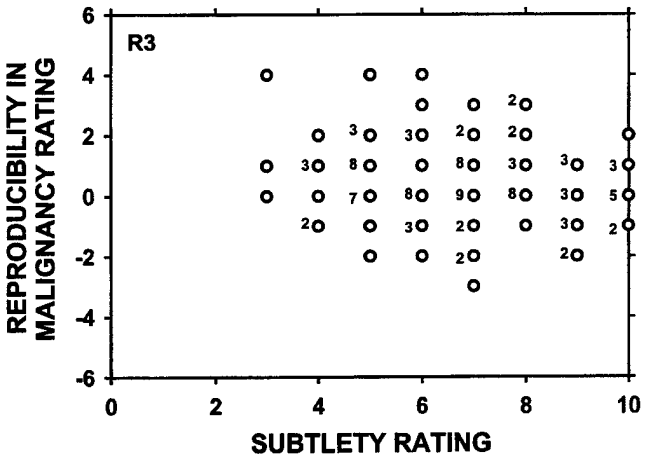


Figure 9. Scatter plot shows the relationship between the differences in confidence ratings between the two readings of the same cluster and the subtlety ratings of the cluster, as rated by reader 3. The number next to a data point indicates the number of cases that overlap at the same point. Data points without a number indicate that there is only one case at that point.

The dependence of classification accuracy on pixel size may be further weakened when other patient information is available for making diagnostic decisions. In clinical practice, the decision for biopsy is not dependent on the mammographic appearance alone. When the morphologic information is nonspecific, other patient information (eg, age, family history, and personal history) becomes important for estimating the likelihood of breast cancer. Because our goal was to evaluate whether the classification accuracy of microcalcifications depended on the pixel size of the digitized images, we did not provide such patient

information to the observers. Our results indicate that the mammographic information that a radiologist assesses from the displayed images, such as the morphologic characteristics and spatial distribution pattern of the microcalcifications, does not have a strong dependence on pixel size in the range studied.

It may be noted that in our current ROC study we concentrated on the effect of pixel size on the classification of malignant and benign microcalcifications according to their mammographic features. We previously conducted an ROC study (9) to compare the detectability of subtle microcalcifications on original screen-film mammograms with that on mammograms digitized at a pixel size of 100 μm by using an optical drum scanner. We found that the detection accuracy for the subtle microcalcifications decreased when radiologists read the 100-μm pixel size digitized images. Results of another previous study (10), in which we investigated the detection of microcalcifications by a computer program, also indicated a reduction in detectability when the digitization pixel size increased from 35 to 140 μm. The results from these experiments indicate that spatial resolution may be more important for the detection than for the classification of microcalcifications in mammographic imaging.

In clinical practice, an important technique used by radiologists to estimate the likelihood of malignancy of a microcalcification cluster is to evaluate its interval change between examinations. The number of microcalcifications in a cluster is an important feature for characterizing changes. High-quality mammograms that can provide sensitive detection of new, subtle microcalcifications are crucial for such a task. The results of our previous studies (9,10) indicate that the spatial resolution of mammographic images will affect the detectability of subtle microcalcifications. The pixel size of digital mammograms may, therefore, affect the evaluation of interval changes, although the effect will be reduced with the use of magnification views. Because the radiologists in our current study were not provided with images from previous examinations for comparison, the effects of pixel size on the detection of interval change will warrant further investigation.

Another possible reason that the images with a 35-μm pixel size did not provide better classification accuracy for malignant and benign microcalcifications than did images with 70- or 105-μm pixel sizes, as observed in this study, is the higher noise level in the digitized images at this small pixel size. A higher noise level will reduce the signal-to-noise ratio of the image and may interfere with the perception of image features. It is possible that if the radiation dose to the patient is unlimited, a digital mammography system with a smaller pixel size can provide better classification. In the current study, we investigated the dependence of classification accuracy on pixel size under the constraint of equal radiation dose. The trade-off between image quality and radiation dose and the acceptability of higher-dose techniques are beyond the scope of this study. Furthermore, because digitized mammograms and mammograms acquired with digital detectors have different noise, contrast sensitivity, and resolution properties, further investigations are needed to determine whether a similar trend holds for mammograms acquired with different types of digital detectors.

In conclusion, we performed an ROC study to investigate the effects of pixel size on the classification of malignant and benign microcalcifications on digitized mammograms. Our results indicate that the differences in the A_z between pairs of pixel sizes ranging from 35 to 140 μm do not achieve statistical significance. The pixel sizes in this range therefore do not have a strong effect on radiologists' accuracy in the characterization of microcalcifications. The low specificity of the image features of microcalcifications and the large interobserver and intraobserver variabilities may have prevented small advantages in image resolution from being observed.

ACKNOWLEDGMENTS

The authors are grateful to Sergey V. Beiden, PhD, for analysis of the variance components, Charles E. Metz, PhD, for the LABMRMC program, and Robert F. Wagner, PhD, and Charles E. Metz, PhD, for helpful discussion on statistical analysis.

REFERENCES

1. Landis SH, Murray T, Bolden S, Wingo PA. Cancer statistics, 1998. CA Cancer J Clin 1998; 48:6–29.
2. Byrne C, Smart CR, Cherk C, Hartmann WH. Survival advantage differences by age: evaluation of the extended follow-up of the Breast Cancer Detection Demonstration Project. Cancer 1994; 74:301–310.
3. Feig SA, Hendrick RE. Risk, benefit, and controversies in mammographic screening. In: Haus AG, Yaffe MJ, eds. Syllabus: a categorical course in physics—technical aspects of breast imaging. Oak Brook, Ill: Radiological Society of North America, 1993; 119–135.
4. Tabar L, Dean PB. Teaching atlas of mammography. New York, NY: Thieme, 1985.
5. Wolfe JN. Analysis of 462 breast carcinomas. AJR Am J Roentgenol 1974; 121:846–853.
6. Murphy WA, DeSchryver-Kecskemeti K. Isolated clustered microcalcification in the breast: radiologic-pathologic correlation. Radiology 1978; 127:335–341.

7. Millis RR, Davis R, Stacey AJ. The detection and significance of calcifications in the breast: a radiological and pathological study. Br J Radiol 1976; 49:12–26.
8. Sickles EA. Mammographic features of 300 consecutive nonpalpable breast cancers. AJR Am J Roentgenol 1986; 146:661–663.
9. Chan HP, Vyborny CJ, MacMahon H, Metz CE, Doi K, Sickles EA. Digital mammography: ROC studies of the effects of pixel size and unsharp-mask filtering on the detection of subtle microcalcifications. Invest Radiol 1987; 22:581–589.
10. Chan HP, Niklason LT, Ikeda DM, Lam KL, Adler DD. Digitization requirements in mammography: effects on computer-aided detection of microcalcifications. Med Phys 1994; 21:1203–1211.
11. Karssemeijer N, Frieling JTM, Hendriks JHCL. Spatial resolution in digital mammography. Invest Radiol 1993; 28:413–419.
12. Kallergi M, Clarke LP, Qian W, et al. Interpretation of calcifications in screen-film, digitized, and wavelet-enhanced monitor-displayed mammograms: a receiver operating characteristic study. Acad Radiol 1996; 3:285–293.
13. Metz CE. Some practical issues of experimental design and data analysis in radiological ROC studies. Invest Radiol 1989; 24:234–245.
14. Sickles EA. Periodic mammographic follow-up of probably benign lesions: results in 3,184 consecutive cases. Radiology 1991; 179:463–468.
15. Sickles EA. Nonpalpable, circumscribed, noncalcified solid breast masses: likelihood of malignancy based on lesion size and age of patient. Radiology 1994; 192:439–442.
16. Dorfman DD, Berbaum KS, Metz CE. ROC rating analysis: generalization to the population of readers and cases with the jackknife method. Invest Radiol 1992; 27:723–731.
17. Beiden SV, Wagner RF, Campbell G. Components-of-variance models and multiple-bootstrap experiments: an alternative methodology for random-effects ROC analysis. Acad Radiol 2000; 7:341–349.

Classifier design for computer-aided diagnosis: Effects of finite sample size on the mean performance of classical and neural network classifiers

Heang-Ping Chan[a] and Berkman Sahiner
Department of Radiology, University of Michigan, Ann Arbor, Michigan 48109-0030

Robert F. Wagner
Center for Devices and Radiology Health, FDA, Rockville, Maryland 20852

Nicholas Petrick
Department of Radiology, University of Michigan, Ann Arbor, Michigan 48109-0030

(Received 14 June 1999; accepted for publication 16 September 1999)

Classifier design is one of the key steps in the development of computer-aided diagnosis (CAD) algorithms. A classifier is designed with case samples drawn from the patient population. Generally, the sample size available for classifier design is limited, which introduces variance and bias into the performance of the trained classifier, relative to that obtained with an infinite sample size. For CAD applications, a commonly used performance index for a classifier is the area, A_z, under the receiver operating characteristic (ROC) curve. We have conducted a computer simulation study to investigate the dependence of the mean performance, in terms of A_z, on design sample size for a linear discriminant and two nonlinear classifiers, the quadratic discriminant and the backpropagation neural network (ANN). The performances of the classifiers were compared for four types of class distributions that have specific properties: multivariate normal distributions with equal covariance matrices and unequal means, unequal covariance matrices and unequal means, and unequal covariance matrices and equal means, and a feature space where the two classes were uniformly distributed in disjoint checkerboard regions. We evaluated the performances of the classifiers in feature spaces of dimensionality ranging from 3 to 15, and design sample sizes from 20 to 800 per class. The dependence of the resubstitution and hold-out performance on design (training) sample size (N_t) was investigated. For multivariate normal class distributions with equal covariance matrices, the linear discriminant is the optimal classifier. It was found that its A_z-versus-$1/N_t$ curves can be closely approximated by linear dependences over the range of sample sizes studied. In the feature spaces with unequal covariance matrices where the quadratic discriminant is optimal, the linear discriminant is inferior to the quadratic discriminant or the ANN when the design sample size is large. However, when the design sample is small, a relatively simple classifier, such as the linear discriminant or an ANN with very few hidden nodes, may be preferred because performance bias increases with the complexity of the classifier. In the regime where the classifier performance is dominated by the $1/N_t$ term, the performance in the limit of infinite sample size can be estimated as the intercept ($1/N_t = 0$) of a linear regression of A_z versus $1/N_t$. The understanding of the performance of the classifiers under the constraint of a finite design sample size is expected to facilitate the selection of a proper classifier for a given classification task and the design of an efficient resampling scheme. © *1999 American Association of Physicists in Medicine.*
[S0094-2405(99)00212-6]

Key words: computer-aided diagnosis, classifier design, linear classifier, quadratic classifier, neural network, sample size, feature space dimensionality, ROC analysis

I. INTRODUCTION

With the advent of digital imaging modalities, computer-aided diagnosis (CAD) is becoming an important area of research in medical imaging. A CAD algorithm can detect abnormalities and classify disease or normal cases based on image and/or patient information, and thus provide a second opinion to the radiologist in the detection or diagnostic decision making process.

Design of classifiers that can accurately distinguish normal and abnormal features is a critical step in the development of CAD algorithms. It has been shown that the performance of a classifier for unknown cases depends on the sample size used for training.[1] When a finite design (training) sample size is used, the performance is pessimistically biased in comparison to that obtained from an infinitely large design sample. In order to design a classifier with a performance generalizable to the population at large, one has to use a sufficient number of case samples that are representative of the population. However, the availability of case samples is often limited in medical imaging research. It is therefore important to study the sample-size dependence of different classifiers and determine the most efficient way of training a classifier, under the constraint of a finite sample size.

 0094-2405/99/26(12)/2654/15/$15.00 © 1999 Am. Assoc. Phys. Med.

We note that the concept of generalizability may be used in several technical senses when assessing the performance of a classifier: one with respect to mean classifier performance, the other with respect to the variance of classifier performance. In many classifier design problems, one is most interested in investigating if the mean performance of a classifier estimated from a given set of finite design samples can be generalized to classification performance with unknown test samples drawn from the same population of cases. The generalizability in this regard can be observed from the biases of the mean performances in the finite design set and in the test set in comparison to the optimal performance estimated from an infinite design set. The bias in the mean performance of different classifiers under various input conditions is the subject of investigation in this study. We will discuss further other interpretation of generalizability in the Discussion section of this paper.

A number of investigators have studied the finite-sample-size problem[1-9] Fukunaga[1,3] derived a general formulation for the bias and variance of a function, f, which is to be estimated from the available samples. When f is a nonlinear function of the mean vectors and covariance matrices of two feature distributions, it has been shown that a bias results from the nonlinear propagation of the finite-sample variances in the estimates of the mean vectors and covariance matrices of the distributions through this function. For multivariate-normal data, these variances are proportional to $1/N_i$, where N_i is the design sample size, and this dependence propagates into the lowest-order terms in the bias. The bias is independent of the test sample size, N_{test}. All measures of classifier performance that count the fraction of times the decision value for an abnormal case exceeds that for a normal case (independent of underlying distribution), and various measures of error for normally distributed decision functions, are nonlinear functions of the parameters of the underlying distributions. They are thus subject to this effect. Fukunaga and Hayes[3] analyzed the finite sample effects on the probability of misclassification (PMC) of a classifier and suggested a technique that makes use of the linear dependence of PMC on $1/N_i$ to estimate the performance at $N_i \to \infty$ with a finite sample set.

For the evaluation of medical diagnostic systems, the most commonly used performance index is the area under the receiver operating characteristic (ROC) curve, A_z. We have derived analytically that, for linear discriminant classifiers, the classifier performance in terms of A_z can be approximated by a linear function in $1/N_i$, under conditions when higher order terms in N_i can be neglected. We have been investigating the dependence of A_z on sample size by simulation studies[7-9] Wagner *et al.*[10,11] have also analyzed the effects of design and test sample sizes on the variance components of the classifier performance. Although these behaviors depend strongly on the class distributions and the properties of the classifier, the studies will provide some insight into the sample size requirements for the design of different classifiers. This work may eventually lead to the selection of an efficient resampling scheme for classifier design, as well as the development of a statistical test of the sample size requirements and the generalizability of the trained classifier.

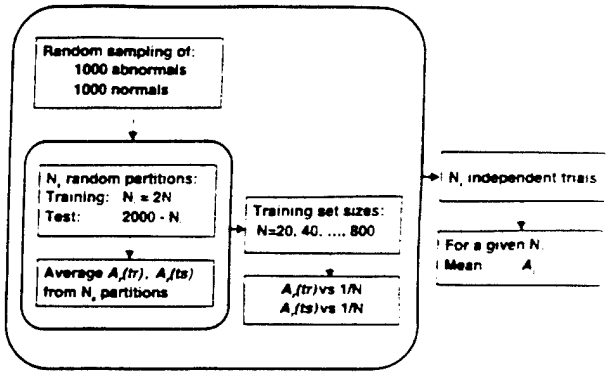


FIG. 1. The sampling and evaluation scheme of the simulation study.

In this paper, we will describe the simulation studies and analyze the effects of sample size on classifier performance. Several commonly used classifiers, including the linear discriminant, the quadratic discriminant, and the back-propagation neural network will be studied and compared under different input conditions. Feature distributions with markedly different characteristics will be used to represent a variety of situations that may be encountered in classification problems for many detection or diagnostic tasks.

II. MATERIALS AND METHODS

We performed simulation studies to evaluate the effects of sample size on classifier design. Normal and abnormal case samples were randomly drawn from known probability distributions of the two classes. These samples were then used to design classifiers for differentiation of normal and abnormal cases. The simulation approach assures that any number of case samples can be obtained from populations with known statistical properties. It thus allows evaluation of the dependence of classifier performance on design sample size and comparison of the performance with theoretically predicted optimal classification based on the chosen probability distributions.

A. Simulation study

The sampling and evaluation scheme of the simulation study is shown in Fig. 1. In this study, we considered only the situation in which equal numbers ($=N_{total}/2$) of normal and abnormal cases randomly drawn from the class distributions were available in our data set. A resampling strategy similar to the technique suggested by Fukunaga and Hayes was devised to generate the A_z-vs-$1/N_i$ curve. Subsets of $N_{i_1}, N_{i_2}, \ldots, N_{i_j}$ design samples were randomly drawn from the available sample set, again under the constraint that the numbers of normal and abnormal samples were equal in each subset, i.e., $N_{i_k,\text{normal}} = N_{i_k,\text{abnormal}} = N_{i_k}/2\,(i=1,\ldots,j)$. A classifier was designed by using each subset of samples. The random sampling of a given subset from the available set of N_{total} samples was performed without replacement, whereas the random sampling of different subsets always started from

the same set of N_{total} samples. Therefore, after drawing a given design subset N_{t_i}, the remaining samples, $N_{total}-N_{t_i}$ were independent of the design samples and used as the test samples. For simplicity, the number of design samples per class is denoted as N in the following discussion.

In general, there are two methods, resubstitution and hold-out, for testing classifier performance. In the resubstitution method, the design sample set is resubstituted into the trained classifier to test its performance, whereas in the hold-out method, an independent test set is used. It has been shown[1] that, for a Bayes classifier, if the classifier is trained with a finite number of design samples, the resubstitution estimate of the classifier performance is optimistically biased whereas the hold-out estimate is pessimisticaly biased in comparison to that achievable with an infinite design sample set. The mean performance obtained from the former estimation provides an upper bound and that from the latter provides a lower bound on the true classifier performance. When the design sample size is limited, it is important to evaluate the hold-out performance to avoid an overly optimistic prediction of the classifier performance. In the limit of very large sample size, the upper and lower bounds converge towards the unbiased estimate.

In this study, we evaluated the performance of the classifier using both the resubstitution and the hold-out methods as a function of finite design sample size N_t. In order to reduce the variances in the estimates of A_z, we randomly resampled without replacement each N_{t_i} from the same N_{total} samples N_p times, trained and tested the classifier, and estimated the average A_z from the N_p individual A_z's as shown in Fig. 1. The resubstitution or hold-out A_z-vs-$1/N_t$ curve was plotted from the j points and the unbiased estimate of A_z in the limit of $N_t \to \infty$ could be extrapolated from either curve.

This method of estimating classifier performance at large N_t by generating a few data points at finite sample sizes is similar to the Fukunaga and Hayes technique. However, we did not assume that the j points were in the linear region of the A_z-vs-$1/N_t$ curve and we used resampling to reduce the variances. In fact, one of the goals of this study was to investigate the range of design sample size in which the performance curve was approximately linear for various classifiers and probability distributions of the class populations. Therefore, we used a much larger total number of samples ($N_{total}=2000$) in our simulation study than was generally available for classifier design. We could then choose N_t over a wide range and study the behavior of the entire A_z-vs-$1/N_t$ curve.

To estimate the population mean of A_z at each N_{t_i}, we repeated the above experiment N_e times, each with 2000 independently drawn samples from the population. The population mean of A_z was estimated by averaging the A_z values obtained from the N_e experiments. We did not analyze the variances in this study because of the complication in the correlation among the N_p values of A_z introduced by resampling. A detailed analysis of the variances and its modeling was performed in a separate study by Wagner *et al.*[10,11] in which a different study design was used.

By varying the number of design samples per class, N, over a large range from 20 to 800, the regime where the $1/N$ dependence dominated could be observed from the A_z (population mean)-vs-$1/N_t$ (or $1/N$) curves. It is important to note that, although the number of test samples, $N_{test_i}=2000-N_{t_i}$, varied from point to point on both the resubstitution and the hold-out curves, the bias in A_z is independent of N_{test_i}.[1] The shape of the A_z-vs-$1/N$ curve is independent of N_{test_i} after N_{t_i} is fixed. However, the variance of a given A_z does depend on the test sample size.

For simplicity, we will refer to these estimates of A_z (population mean) as $A_z(\mathrm{tr})$ for the resubstitution and as $A_z(\mathrm{ts})$ for the hold-out performance in the following discussions.

B. Class distributions

1. Multivariate normal distributions

For three of the four types of class distributions, we assumed that the normal and abnormal classes followed multivariate normal distributions in the feature space. The dimensionality of the feature space, k, was varied from 3 to 15. The characteristics of the multivariate normal distributions can be completely specified by the multivariate mean vector of the rth class, denoted as μ_r ($r=1,2$) and its covariance matrix, denoted as Σ_r. The separation of the normal and abnormal classes is measured by the Bhattacharyya distance, B, defined as[1,12]

$$B=\frac{1}{8}\Delta+\frac{1}{2}\ln\frac{\det[(\Sigma_1+\Sigma_2)/2]}{\sqrt{\det\Sigma_1}\sqrt{\det\Sigma_2}}, \tag{1}$$

where $\det\Sigma_r$ denotes the determinant of Σ_r, and Δ is the squared Mahalanobis distance,[12] defined as

$$\Delta=(\mu_2-\mu_1)^T\left(\frac{\Sigma_1+\Sigma_2}{2}\right)^{-1}(\mu_2-\mu_1). \tag{2}$$

The Mahalanobis distance is the Euclidean distance between the means of the two distributions, normalized by the square root of the average of their covariance matrices. It can therefore be considered to be a measure of the signal-to-noise ratio (SNR) between the abnormal and the normal distributions. The second term of B is the contribution from the difference in the covariance matrices of the two class distributions. If the covariance matrices are equal, the second term will be zero and the Bhattacharyya distance will be equal to 1/8 of the squared Mahalanobis distance.

In the current study, three types of multivariate normal class distributions were considered. In the following discussion, we shall refer to the use of simultaneous diagonalization for the two covariance matrices of the class distributions. This operation leaves the normal-based decision functions unchanged because the distance measures that arise in these decision functions are invariant to any non-singular linear transformation.[1]

(1) Equal covariance matrices and unequal means: In this case, the covariance matrices of the normal and abnormal class distributions can be simultaneously diagonalized

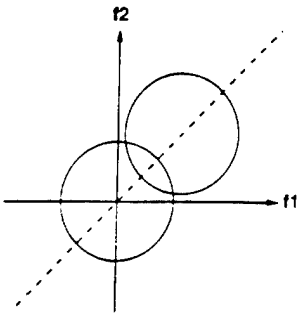


FIG. 2. A schematic illustration of the two class distributions with equal covariance matrices and unequal means in a 2D feature space. The circles represent contours of equal probability in each distribution.

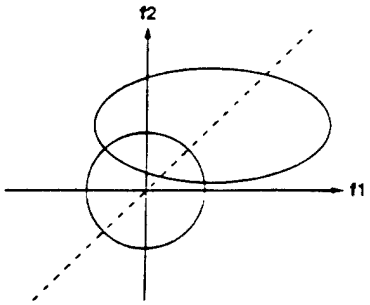


FIG. 3. A schematic illustration of the two class distributions with unequal covariance matrices and unequal means in a 2D feature space. The closed curves represent contours of equal probability in each distribution.

and the variances of the individual feature components can be scaled to unity. Therefore, without loss of generality, the covariance matrices of the two classes could be assumed to be equal to identity matrices, $\Sigma_1 = \Sigma_2 = I$. The mean feature vector for the first class was assumed to be zero, $\mu_1 = \mathbf{0}$, and the mean feature vector for the second class, $\mu_2 = \mathbf{M}$ with all components of $\mathbf{M}$ equal to a constant m. The magnitude of m could be adjusted to obtain a desired separation of the two classes. For the purpose of this simulation study, we chose m such that the squared Mahalanobis distance was 3, i.e., the Bhattacharyya distance was 3/8, for feature spaces of any dimensionality. As discussed below, this separation corresponds to a theoretical A_z of 0.89, which is in the performance range of many classification problems in CAD applications. An example of the two class distributions in a 2D feature space is shown schematically in Fig. 2.

(2) Unequal covariance matrices and unequal means: The covariance matrix of the first class was again diagonalized and scaled to be an identity matrix, $\Sigma_1 = I$, and the mean feature vector for the first class was assumed to be zero, $\mu_1 = \mathbf{0}$. The covariance matrix of the second class, Σ_2, was simultaneously diagonalized to have eigenvalues λ_i, $i = 1,\ldots,k$. For this study, we generated the values of λ_i with the simple relationship

$$\lambda_i = \lambda_{\min} + \frac{(i-1)(\lambda_{\max} - \lambda_{\min})}{(k-1)}, \quad i = 1,\ldots,k \tag{3}$$

and evaluated one condition where $\lambda_{\min} = 1$, and $\lambda_{\max} = 2$ for all dimensionalities of the feature spaces. We also assumed that the components of the mean feature vector μ_2 were equal, the values of which were adjusted to achieve a Bhattacharyya distance of 3/8. For the purpose of demonstrating the general trends of the A_z-vs-$1/N$ curves and comparing the relative performance of the different classifiers under the various conditions, the specific choices of these values are not critical. Figure 3 illustrates an example of the two class distributions in a 2D feature space.

(3) Unequal covariance matrices and equal means: The covariance matrix of the first class was the same as that in the first two cases described above. The covariance matrix of the second class was proportional to the identity matrix, $\Sigma_2 = \alpha I$, where the proportionality constant α was adjusted to provide a Bhattacharyya distance of 3/8. The mean feature vectors of the two classes were equal, $\mu_1 = \mu_2 = \mathbf{0}$. In this case, the discriminatory power of the two classes comes entirely from the difference in the covariance matrices. A schematic of the two class distributions in a 2D feature space is shown in Fig. 4.

2. Checkerboard distributions

The fourth type of class distributions was a checkerboard where the normal and abnormal classes were located in alternate square box regions of the feature space. Within each box of the checkerboard, the feature vectors were uniformly distributed. The two classes did not overlap with each other so that they could be perfectly separated by an "ideal" classifier with $A_z = 1$. We considered a 2×3 checkerboard in a 2D feature space and a $2\times2\times2$ checkerboard in a 3D feature space. The example of a 2×3 checkerboard in a 2D feature space is shown in Fig. 5. Such class distributions may not be common in actual classification problems encountered in CAD. However, it was included in this study to demonstrate the capability and limitations of the different classifiers when the class distributions were not multivariate normal.

C. Classifiers

We studied three types of classifiers: the linear discriminants, the quadratic discriminants, and the back-propagation neural networks. They represent a range of classifiers commonly used in the field of pattern recognition at present.

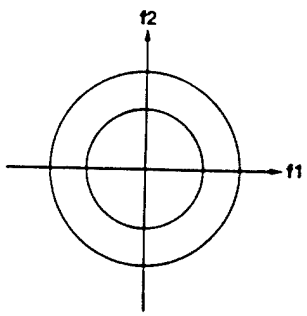


FIG. 4. A schematic illustration of the two class distributions with unequal covariance matrices and equal means in a 2D feature space. The circles represent contours of equal probability in each distribution.

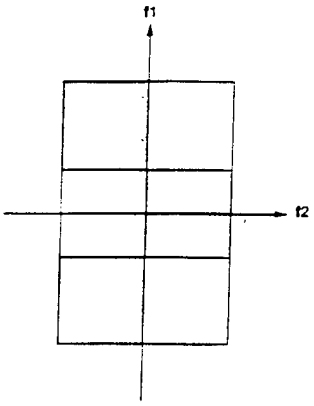


FIG. 5. An example of a 2×3 checkerboard in a 2D feature space

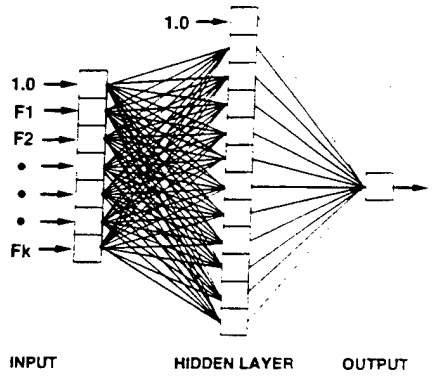


FIG. 6. A schematic diagram of a backpropagation neural network with one hidden layer.

(1) Linear discriminant classifier: The linear discriminant classifier can be derived from the means and the covariance matrices of the class distributions as follows:[1,13]

$$h_l(X) = (\mu_2 - \mu_1)^T \bar{\Sigma}^{-1} X - \tfrac{1}{2}(\mu_1^T \bar{\Sigma}^{-1} \mu_1 - \mu_2^T \bar{\Sigma}^{-1} \mu_2), \quad (4)$$

where $\bar{\Sigma} = (\Sigma_1 + \Sigma_2)/2$, and X is the feature vector to be classified. The means and covariance matrices have to be estimated as the sample means and sample covariance matrices from the available design samples. The sample means and covariance matrices undergo a nonlinear transformation to become the discriminant scores, which in turn are transformed nonlinearly into a measure of the performance. The variances in the estimated parameters propagate into the mean classifier performance and result in a bias through the second derivative of the transformation function.

It is known that, for multivariate normal distributions with equal covariance matrices, the linear discriminant classifier is optimal and the classifier performance in the limit of large design samples is determined by the Mahalanobis distance, given by

$$A_z = \frac{1}{\sqrt{2\pi}} \int_{-\Delta/2}^{\infty} e^{-u^2/2}\, du. \quad (5)$$

For the class distributions with $\Delta = 3$ to be used in this study, it can be derived from Eq. (5) that the maximum A_z that the optimal linear discriminant can achieve in the limit of large design samples is 0.89.

(2) Quadratic discriminant classifier: The quadratic discriminant classifier can be expressed as[1]

$$h_q(X) = \frac{1}{2}(X - \mu_1)^T \Sigma_1^{-1} (X - \mu_1) - \frac{1}{2}(X - \mu_2)^T \Sigma_2^{-1} (X - \mu_2) + \frac{1}{2} \ln \frac{\det \Sigma_1}{\det \Sigma_2}. \quad (6)$$

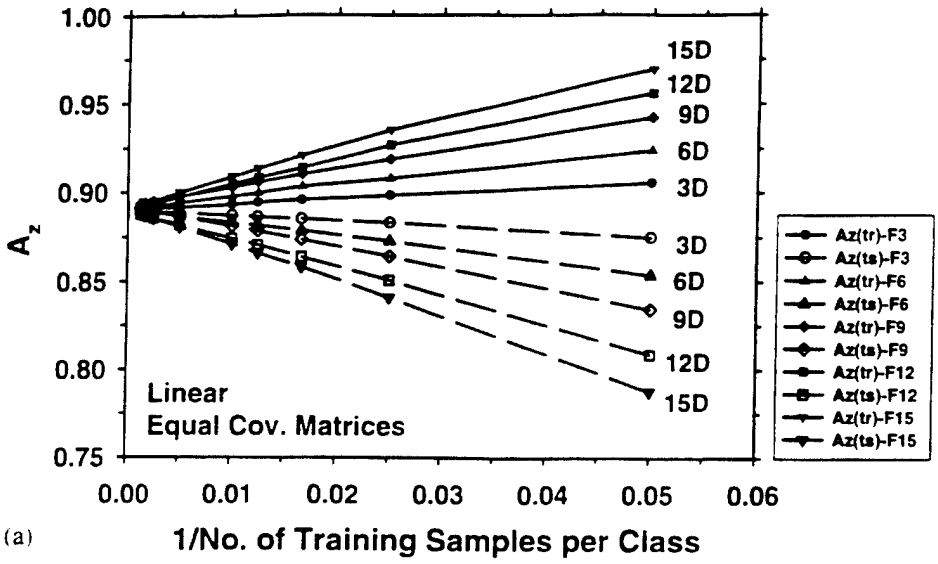
When the class distributions are multivariate normal with unequal covariance matrices, the quadratic discriminant classifier is optimal in the limit of large training samples. The Bhattacharyya distance gives an upper bound on the Bayes error.[1] The general properties of the linear and quadratic classifiers have been described in the literature (for example, Fukunaga[1]).

(3) Back-propagation neural network: Many different architectures and training methods have been developed for artificial neural networks (ANN)[14] in various applications. In this study, we considered only a three-layered neural network trained with a feed-forward back-propagation method. The neural network has k input nodes, n hidden nodes, one output node, and a bias node in both the input and the hidden layers. The ANN architecture is denoted as $k-n-1$. The nodes in the ANN are fully connected and are trained with a minimum sum-of-squares-error criterion. The number of weights to be estimated is equal to $n(k+1)+(n+1)$. A schematic diagram of an ANN is shown in Fig. 6.

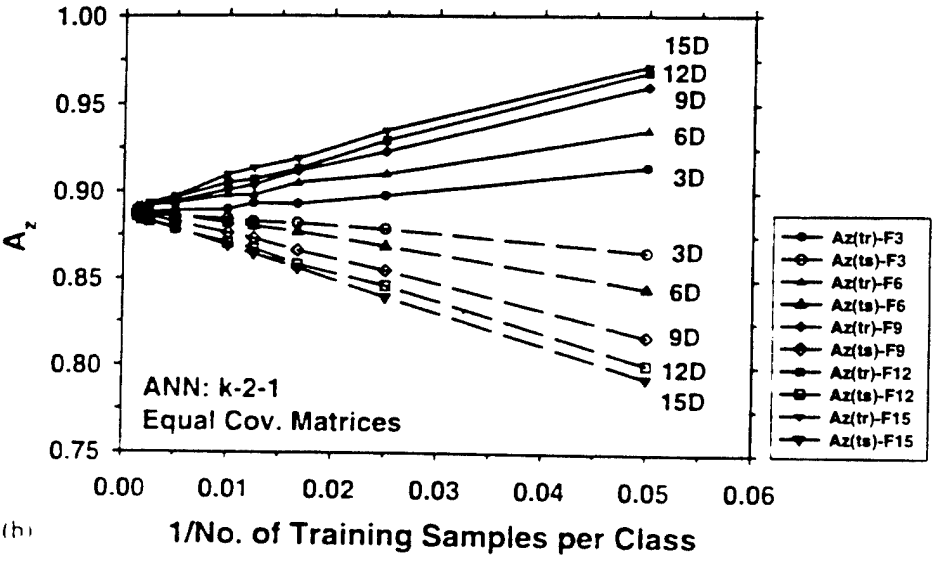
III. RESULTS

In our simulation study, we compared the performance of the linear, quadratic, and backpropagation neural network classifiers for the different class distributions in the feature spaces of dimensionality ranging from 3 to 15. The number of repeated experiments N_e was chosen to be 20 for all cases in the multivariate normal feature spaces and 100 in the checkerboard feature space. The number of data set partitionings N_p in each experiment ranged from 1 to 20. These choices are a compromise between computation time and estimation accuracy, especially for ANN classifiers with a large number of hidden nodes in high dimensional feature spaces. As shown in the graphs discussed below, some of the performance curves may exhibit fluctuations that could be reduced by a larger number of experiments. However, the general trend of the performance curves should not be changed by the statistical uncertainties.

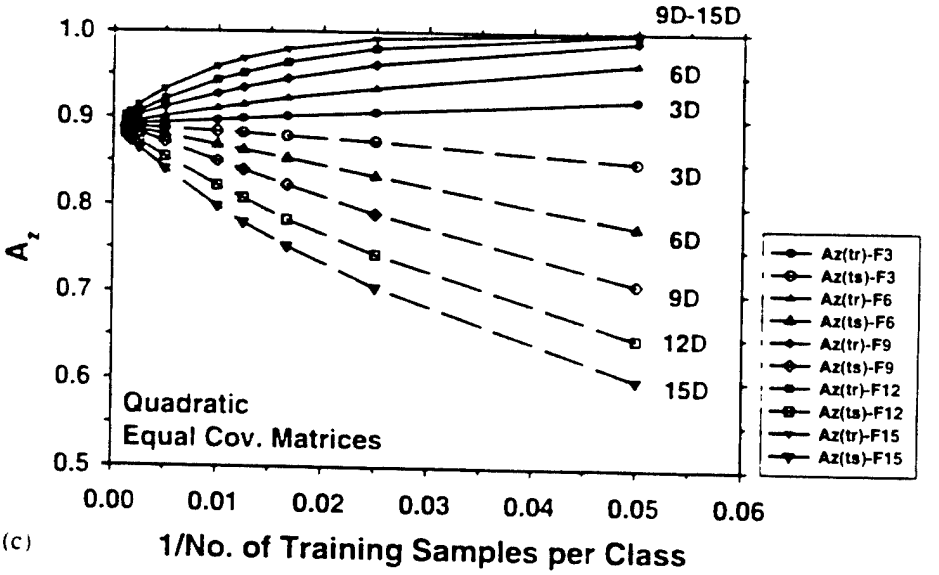
(1) Multivariate normal distributions—Equal covariance matrices and unequal means: For class distributions with equal covariance matrices, the linear discriminant is theoretically the optimal classifier when the design sample size is large. However, when the design sample size is small, the performances of all classifiers are biased. Figures 7(a)–7(c) show the dependence of the A_z obtained from resubstitution (training), $A_z(\mathrm{tr})$, and the A_z obtained from the hold-out method (testing), $A_z(\mathrm{ts})$, on $1/N$ for the linear, ANN, and



(a)



(b)



(c)

FIG. 7. The dependence of the A_z obtained from resubstitution (training-solid lines), A_z(tr), and the A_z obtained from the hold-out method (testing—dashed lines), A_z(ts), on $1/N$ for the class distributions with equal covariance matrices and unequal means. (a) Linear, (b) ANN, and (c) quadratic classifier. Legend: F3 = 3D feature space, etc.

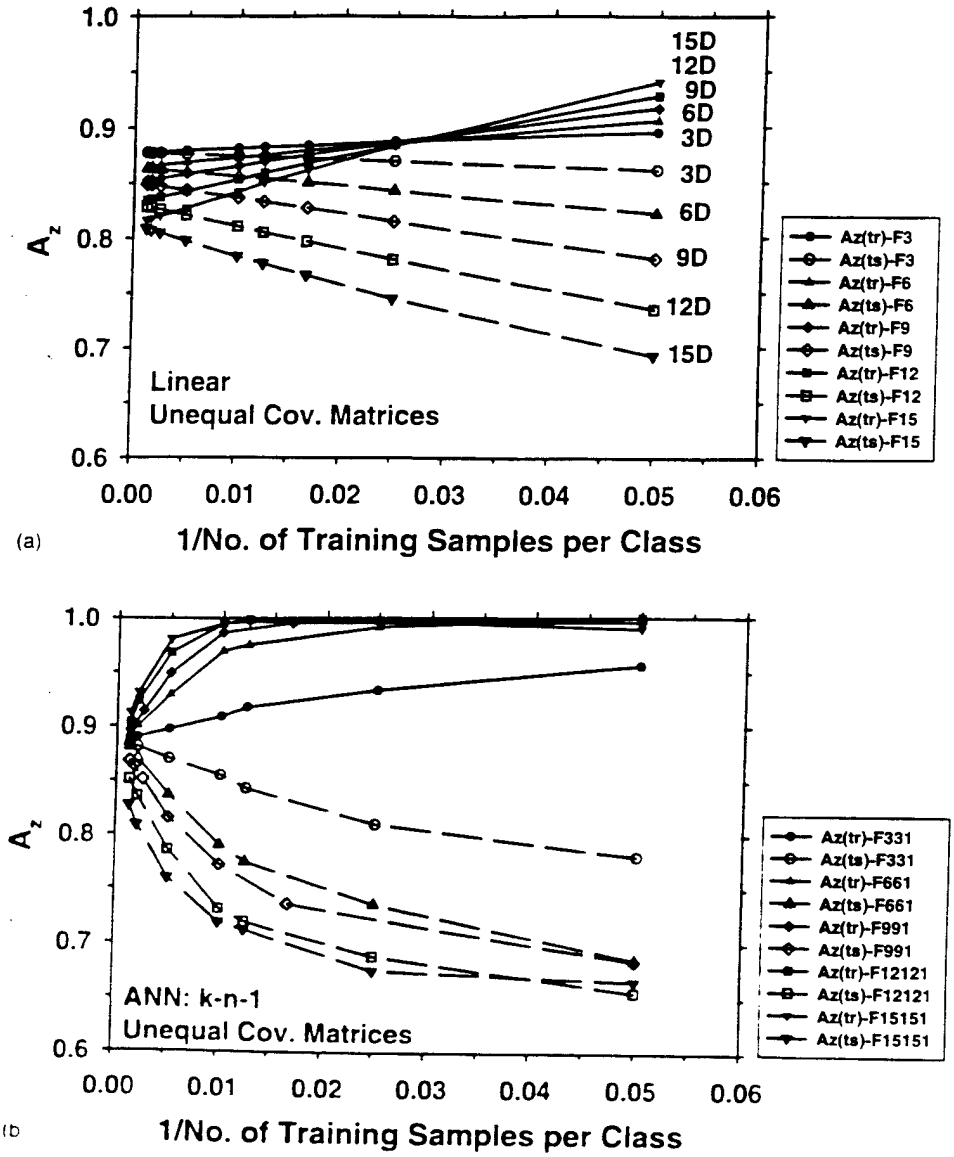


FIG. 8. The performances of the classifiers for class distributions with unequal covariance matrices and unequal means. (a) Linear, (b) ANN classifier. Legend: F3=3D feature space, etc., solid lines $=A_z$(tr), dashed lines $=A_z$(ts).

quadratic classifier, respectively. Two hidden nodes were used for the ANN $(k-2-1)$ because it is the smallest number of hidden nodes in a nonlinear ANN. An ANN with only one hidden node will be a linear classifier and behave in a similar manner as the linear discriminant. On the other hand, ANNs with a large number of hidden nodes (not shown) will overfit the design samples and have poor generalizability to the unknown cases, similar to the ANN curves to be discussed below. All three classifiers can reach the optimal classification accuracy of $A_z=0.89$ in the limit of large N. The curves for the linear classifier and the ANN $(k-2-1)$ at 400 training epochs (iterations) are approximately linear over the entire range. The quadratic classifier does not reach the approximately linear region until N is greater than about 100 $(1/N<0.01)$ in the higher-dimensional feature space. The biases on both the resubstitution and hold-out curves for the quadratic classifier are greater than those for the linear classifier and the ANN $(k-2-1)$. The large biases again indicate overfitting and poor generalization by the quadratic classifier in the equal-covariance-matrices situation.

(2) Multivariate normal distributions—Unequal covariance matrices and unequal means: The performances of the classifiers for class distributions with unequal covariance matrices are shown in Figs. 8(a)–8(b). The linear discriminant and the ANN $(k-2-1)$ classifier (not shown) are again approximately linear over the entire range of N studied. However, the A_z at $1/N=0$ decreases as the dimensionality of the feature space increases. This is because both the linear discriminant and the near-linear ANN $(k-2-1)$ cannot make use of the class separability due to the differences in the covariance matrices which is the second term in the Bhattacharyya distance. The second term increases relative to the first term, the squared Mahalanobis distance, when the Bhattacharyya distance is fixed and the dimensionality of the feature space increases.

The performance curves of the ANN at large N improve when a greater number of hidden nodes and a sufficient number of training epochs are used. The number of hidden nodes required to reach the optimal classification of $A_z=0.89$ at $1/N=0$ increases with the dimensionality of the feature

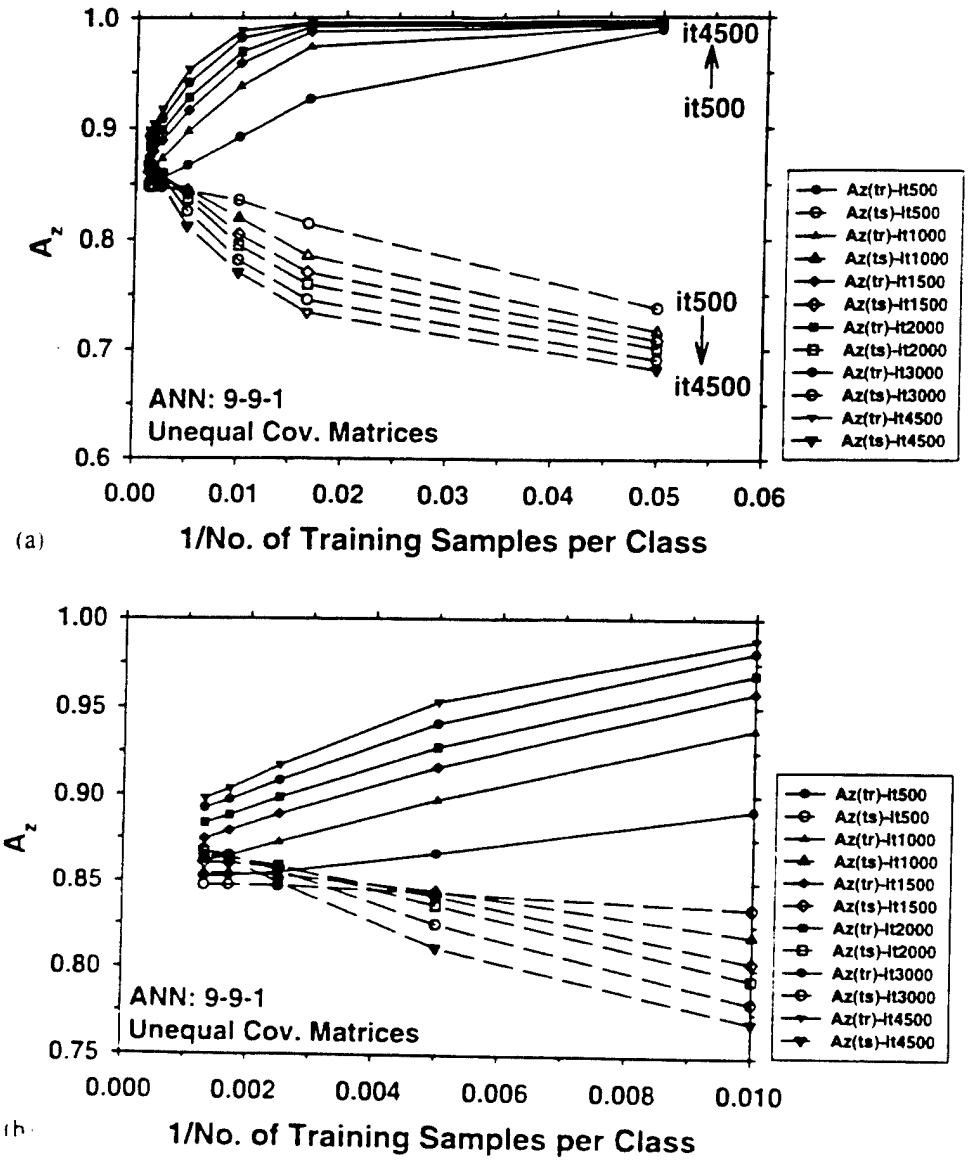


FIG. 9. The dependence of the performance curves on the number of training epochs for an ANN with nine hidden nodes in a 9D feature space: ANN(9−9−1). Legend: it500 = 500 training epochs, etc., solid lines = A_z(tr), dashed lines = A_z(ts). The expanded view in (b) shows the trend of the curves at large sample sizes.

space. Figure 8(b) shows the performance of the ANNs when the number of hidden nodes is equal to the dimensionality in each feature space. Since the number of weights to be trained increases rapidly with increasing number of nodes in an ANN, the number of epochs required for training the ANN to achieve a reasonable classification accuracy increases accordingly. The resubstitution and hold-out performance curves of each ANN shown in Fig. 8(b) were chosen at the smallest number of training epoch that resulted in approximately the highest A_z value when the hold-out curve was extrapolated to $1/N = 0$. The number of training epochs required to reach the highest A_z increased as the dimensionality and the number of hidden nodes in the ANN increased. It ranged from about 4000 to 10 000 for the conditions shown in Fig. 8(b). We did not attempt to perform an exhaustive search for the "optimal" number of hidden nodes in each feature space because of the extensive computation time required for the search. Instead, we evaluated ANNs with a few different numbers of hidden nodes in each feature space and chose the "best" ANN within those studied. With this approximation we observed that, in a k-dimensional feature space and with these class distributions, an ANN with approximately k hidden nodes can approach the optimal performance when the design sample size and the number of training epochs are sufficiently large, as shown in Fig. 8(b).

To illustrate the training of an ANN with a large number of hidden nodes, we show the dependence of the resubstitution and the hold-out curves on the number of training epochs for ANN (9−9−1) in Fig. 9. A number of commonly discussed problems of an ANN can be observed. In the small N region below about 60 samples per class, overparametrization and over-training are obvious, i.e., near perfect classification during training [A_z(tr) greater than 0.95] and poor generalization [A_z(ts) below about 0.8]. The problem becomes more pronounced with an increasing number of training epochs. In the middle range of 200 to 400 samples per class where A_z(ts) increases to a maximum then decreases with further training, an "optimal" number of training epoch exists. Only in the region with a sufficiently large N (greater than about 500 per class), A_z(ts) increases with

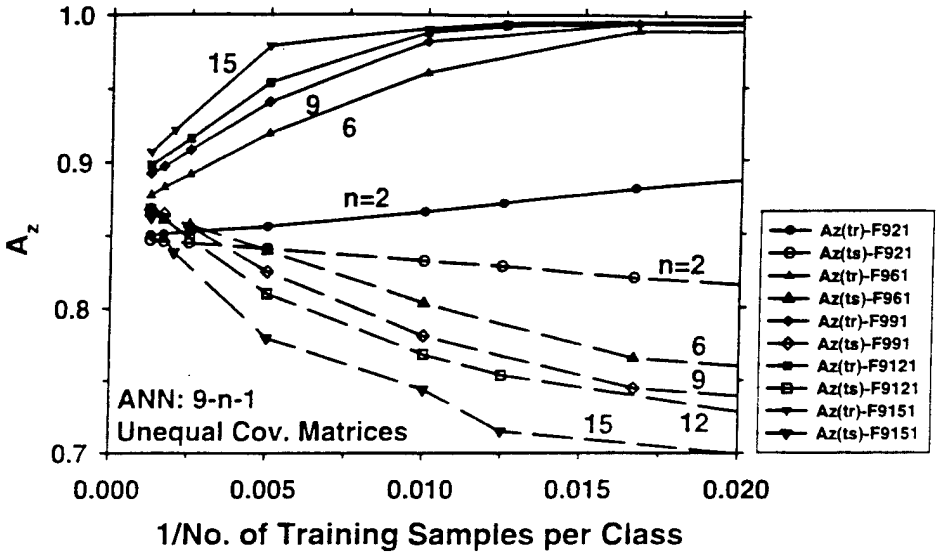


FIG. 10. The dependence of the performance curves of an ANN on the number of hidden nodes in the 9D feature space for class distributions with unequal covariance matrices and unequal means. Legend: F921 = ANN with two hidden nodes, etc., solid lines = A_z(tr), dashed lines = A_z(ts).

increasing number of training epochs within the range studied. The A_z(ts)-vs-$1/N$ curve becomes linear for N greater than about 200. This dependence of ANN on training epoch is generally observed for ANNs with a large number of hidden nodes and in high-dimensional feature spaces, although the design sample size required in order to avoid overtraining and over-parametrization varies. It reinforces our general experience that the ANNs with a large number of weights can overfit the design samples easily and provide poor generalization when the sample size is small.

The performance curves of ANNs with different numbers of hidden nodes in the 9D feature space are shown in Fig. 10. The curves for a given ANN were again chosen at a training epoch in which the hold-out curve approached approximately the highest performance at $1/N = 0$. The chosen training epoch ranged from 600 to 12 000 for the 2- to 15-hidden-node ANNs shown. When the number of hidden nodes is small, the highest A_z obtained by extrapolation to $1/N = 0$ appears to be below the theoretical optimum of 0.89. For example, the A_z extrapolated to $1/N = 0$ is about 0.85 for ANN (9−2−1), and is about 0.87 for ANN (9−6−1). The ANN with nine hidden nodes appears to approach the optimal A_z of 0.89 in the limit of $1/N = 0$. However, the ANN (9−9−1) does not reach the approximately linear region until N is greater than about 200 (easier to see in Fig. 9). As can be seen from the hold-out curves, increasing the number of hidden nodes further will increase overfitting, reduce generalizability, and increase train time without gaining true improvement in performance for classification of unknown case samples.

The quadratic classifier is the theoretically optimal classifier for the class distributions with unequal covariance matrices. It can optimally utilize the class separability contributed by both the differences in the means and the covariance matrices. The performance curves for the quadratic classifier (not shown) in feature spaces of different dimensionalities are very similar to those obtained for the equal covariance matrices situation [Fig. 7(c)]. The A_z of the quadratic classi-

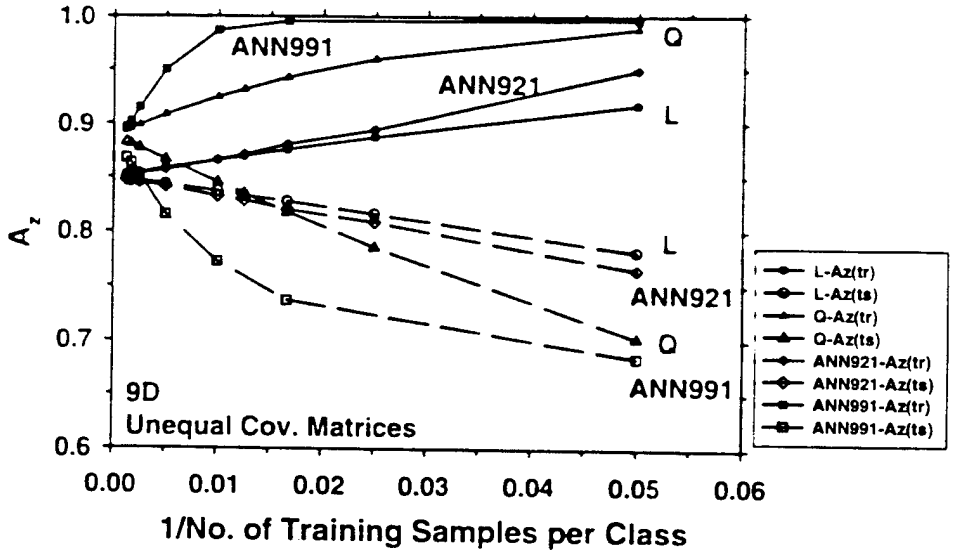


FIG. 11. Comparison of the performance curves of the linear, quadratic, ANN(9−2−1), and ANN(9−9−1) classifiers in the 9D feature space for class distributions with unequal covariance matrices and unequal means. Legends: L = linear; Q = quadratic, ANN = neural network, solid lines = A_z(tr), dashed lines = A_z(ts).

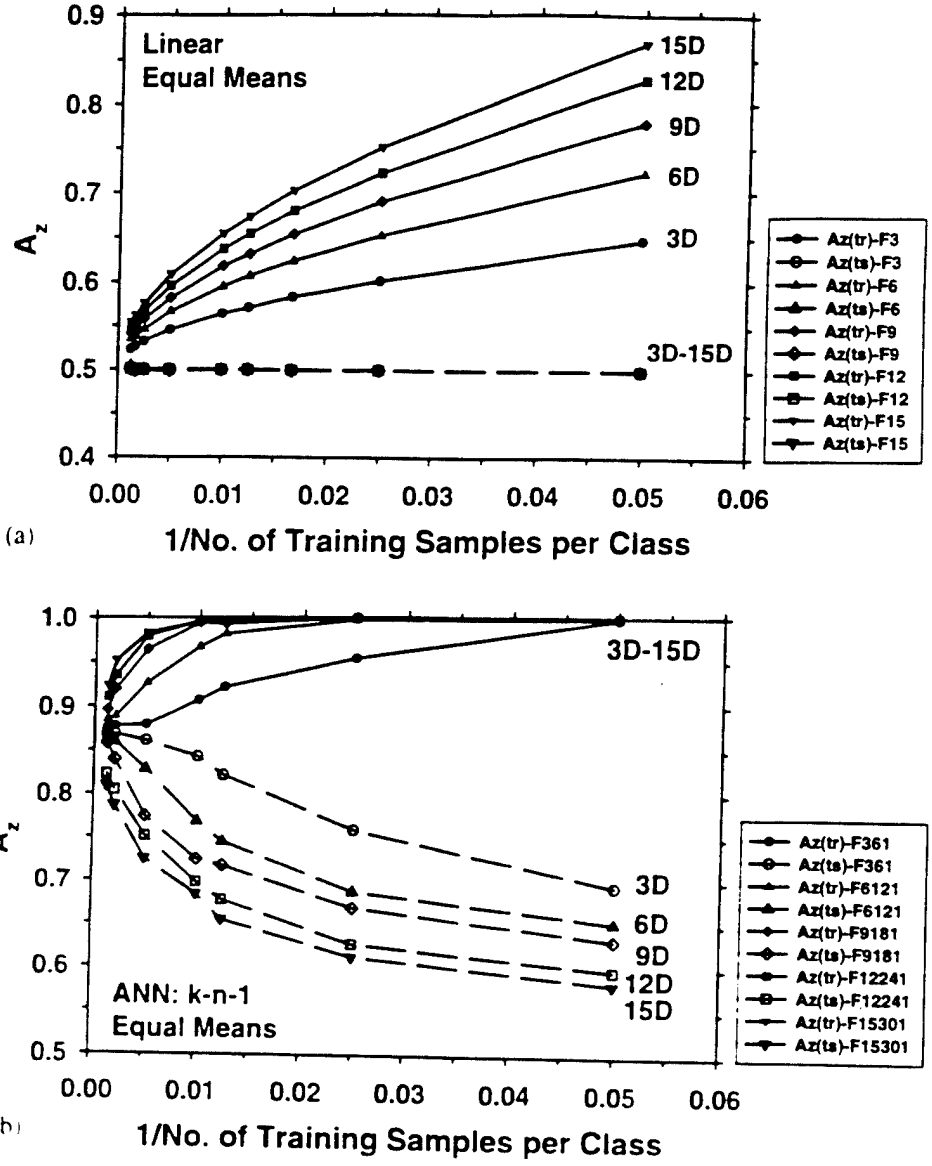


FIG. 12. The dependence of the performance curves on dimensionality of feature space for the class distributions with unequal covariance matrices and equal means. (a) Linear, (b) ANN classifier. Legend: F3=3D feature space, etc. F921 =ANN with two hidden nodes, etc. solid lines=A_z(tr), dashed lines=A_z(ts).

fier reaches the optimal value of 0.89 in the limit of large N for all dimensionalities studied.

Figure 11 shows a comparison of the performance of the linear, quadratic, and the ANN classifiers with two and nine hidden nodes. The biases on the resubstitution and the hold out curves of the quadratic classifier are not as large as those of the ANN (9−9−1) classifier. However, in the regime of small design sample sizes, the hold-out curve of the optimal quadratic classifier can be much lower than the corresponding curves of the linear classifier or ANN with one or two hidden nodes. This result indicates that the theoretically optimal classifier may not be the optimal choice when the available design sample size is small and overparametrization becomes an important consideration.

(3) Multivariate normal distributions—Unequal covariance matrices and equal means: Figure 12(a) shows the dependence of A_z on $1/N$ for the linear classifiers for the class distributions with equal means. Since the Mahalanobis distance is zero when the means of the two class distributions are equal, the linear classifier performs no better than random guessing in the hold-out situation (A_z(ts)=0.5). However, it is somewhat surprising that the resubstitution curve can be biased to very high A_z values, when the design sample is small. The bias increases with increasing dimensionality of the feature space because the severity of overfitting to the design samples worsens with increased parameterization in the linear discriminant function. This indicates that the predicted performance of a classifier can be unrealistically optimistic if the test samples are not independent of the design samples.

For the class distributions with equal means, it is much more difficult to train the ANN classifier. The number of hidden nodes and the number of training epochs required for the ANN to approximate the decision surfaces, which are spherical hypersurfaces in the k-dimensional feature space, increase as k increases. Figure 12(b) shows the A_z-vs-$1/N$ curves for the ANNs in which the number of hidden nodes is 2 times the dimensionality of the feature space. The number of training epochs required to approach the highest perfor-

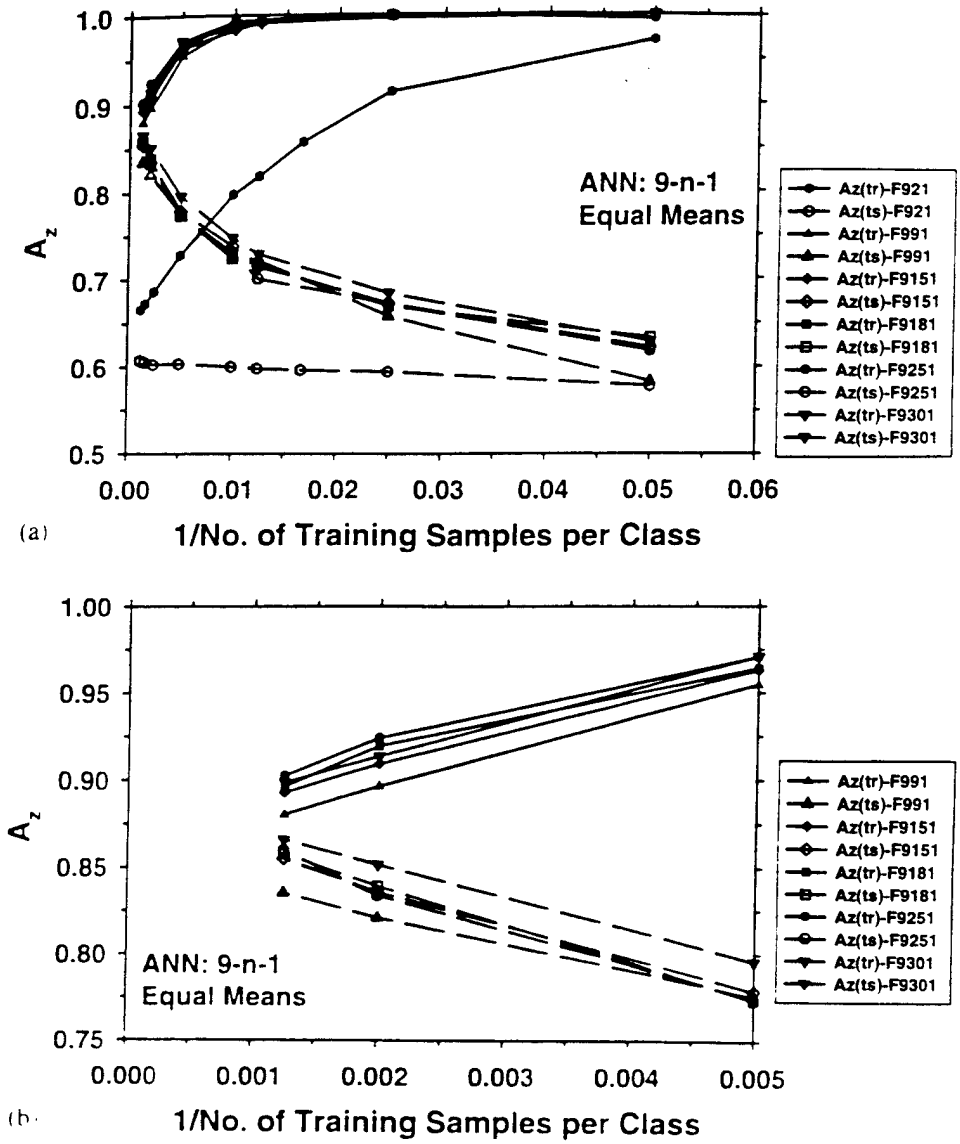


FIG. 13. (a) The dependence of the performance curves of an ANN on the number of hidden nodes in the 9D feature space for class distributions with unequal covariance matrices and equal means. In the expanded scale (b), the approximately linear regions of the curves can be observed. Solid lines = A_z(tr), dashed lines = A_z(ts).

mance for a given ANN architecture ranges from about 1800 to 20000 in these cases. Again we did not attempt an exhaustive search for the "optimal" number of hidden nodes in each case. These ANNs were chosen because they appear to approach the maximum performance of $A_z = 0.89$ in the limit of large N and their number of hidden nodes is a simple multiple of the dimensionality. Compared to the class distributions with unequal means, for a given dimensionality, the number of hidden nodes and the number of training epochs required for achieving the near maximum performance at large N are greater in this equal-mean situation. Figure 13(a) shows an example of the dependence of the performance curves on the number of hidden nodes in the 9D feature space. Figure 13(b) is an enlarged view of the curves in Fig. 13(a) in the range where the sample size is greater than 200 per class. The hold-out performance of ANN(9−9−1) at $1/N = 0$ reaches about 0.85. When the number of hidden nodes is greater than nine, the performances of the ANNs at $1/N = 0$ are similar and approach the optimal A_z.

The quadratic discriminant is again the theoretically optimal classifier for the class distributions with unequal covariance matrices. Its performance curves (not shown) are very similar to those plotted in Fig. 7(c), except that the extrapolated A_z values at $1/N = 0$ do not reach as high as those in the equal covariance matrices situation. By using the approximately linear region of the A_z-vs-$1/N$ curve at N greater than 100, the extrapolated A_z ranges from about 0.873 to 0.885 for the 3D to 15D feature spaces. In this case, it is much more efficient to train a quadratic discriminant than the ANN. Since the linear discriminant and ANNs with few hidden nodes cannot provide effective classification regardless of the design sample size, the quadratic discriminant is obviously the optimal classifier both in terms of performance and training efficiency.

(4) Checkerboard distributions: In a feature space with checkerboard class distributions, classification is difficult for many classifiers because of the disjoint clusters of samples belonging to the same class. We compared the three classifiers in such a situation by two examples. Figure 14 shows the performance curves of the three classifiers in a 2D feature

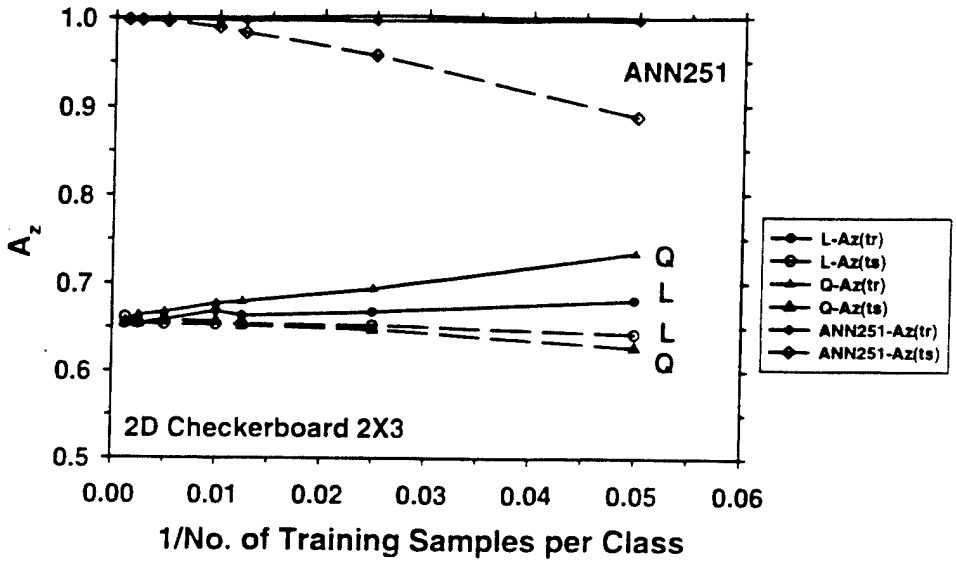


FIG. 14. Performance curves of the three classifiers for a 2×3 unit checkerboard in a 2D feature space. L=linear, Q=quadratic, ANN251=backpropagation neural network with five hidden nodes. Solid lines=A_z(tr), dashed lines=A_z(ts).

space with a 2×3 unit checkerboard distribution. Both the linear and the quadratic discriminants perform poorly even for the resubstitution method where A_z values are in the range of 0.6 to 0.7. However, the ANN(2−3−1) can achieve an A_z of 0.96 (not shown) and the ANN(2−5−1) a near-perfect classification at a training epoch of about 1200.

In a 3D feature space with a 2×2×2 unit checkerboard distribution, the difficulty in classification experienced by the linear and quadratic discriminants is even more apparent. Figure 15 shows that the hold-out curve of the linear classifier is basically the same as random guessing. The hold-out curve of the quadratic classifier is slightly higher than 0.5 at small design sample sizes but approaches 0.5 as the design sample increases. On the other hand, the ANN(3−3−1) can attain a test A_z of 0.9 (not shown) and the ANN(3−5−1) can reach near-perfect classification at large design sample sizes after about 1500 training epochs. These two examples demonstrate that an ANN classifier can be superior to the linear or quadratic classifiers for class distributions that are very different from the idealized multivariate normal distributions.

IV. DISCUSSION

Classifier design is an important field of research in computer-aided diagnosis. Yet many of the issues related to classifier design have not been explored systematically. This simulation study is a part of our on-going investigation of the sample size effects on classifier design.[7–11,15] In this study, we evaluated classifier performance for three multivariate normal class distributions with specific properties: equal covariance matrices, unequal covariance matrices, and equal means. These distributions are idealized but they do approximate a range of situations that may occur in real classification problems. Since the optimal classifier and the upper bound of classification accuracy in the limit of $1/N=0$ are

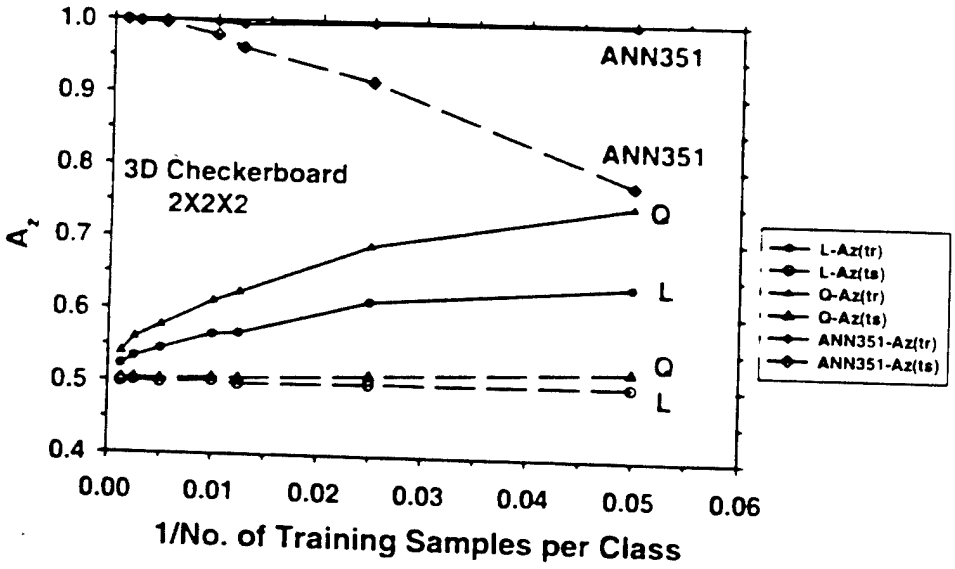


FIG. 15. Performance curves of the three classifiers for a 2×2×2 unit checkerboard distribution in a 3D feature space. Legend: L=linear, Q=quadratic, ANN351=back-propagation neural network with five hidden nodes.

known for each of these cases, we can compare the performances of the classifiers under each condition with the optimum. In addition, a checkerboard class distribution was included in the study. A comparison of the performances of the different classifiers for this class distribution can illustrate their effectiveness when the distributions are very different from multivariate normal.

For all three classifiers, the A_z(tr) obtained by resubstitution is biased optimistically while the A_z(ts) obtained by testing with an independent test set is biased pessimistically, relative to the A_z in the limit of $N \to \infty$, except for the situations when A_z(tr) is bounded from above by perfect classification ($A_z = 1$) or when A_z(ts) is bounded from below by random guessing ($A_z = 0.5$). The magnitude of the biases increases as the design sample size decreases and as the dimensionality of the feature space increases. In the cases where a given classifier has no discriminatory power for a given class distribution, for example, the linear discriminant for the equal-mean or checker-board class distributions, or the quadratic discriminant for the 3D checker-board class distribution, the test A_z(ts) remains almost constant at 0.5, independent of the design sample size. In many cases, the A_z-vs-$1/N$ curve cannot be approximated by a straight line that extrapolates to the A_z at $1/N = 0$ until the design sample sizes are very large, beyond the range of sample sizes that are generally available for CAD classifier design. To estimate the performance of a classifier at large N under the constraint of a small design sample, one may use the Fukunaga and Hayes resampling scheme[3] to derive several points along the A_z-vs-$1/N$ curves in the small sample size region. If the extrapolated resubstitution and hold-out curves do not converge to approximately the same A_z at $1/N = 0$, an average of the points on the two curves which correspond to the same design sample size may be a closer estimate of A_z than either A_z(tr) or A_z(ts). It may be noted that the resubstitution and the hold-out curves are not biased symmetrically from the A_z at infinite N, the average thus obtained will only be a rough estimate. It is also not valid in cases when the classifier has no discriminatory power with A_z(ts) constant at about 0.5 or when the resubstitution curve is overly optimistic with A_z(tr) constant at about 1.

In any case, caution should be taken in estimating classifier performance by extrapolation to $1/N = 0$ or by averaging the resubstitution and hold-out performance as discussed above. The estimated performance contains variances that have to be estimated using further tools. One such attempt in estimating the components of variance by a bootstrapping resampling scheme has been studied recently by Wagner *et al.*[11] These estimates reveal the amount of bias and variance in the classifier performance obtained with the finite design samples, thus allowing estimation of the sample size required to achieve a desired degree of generalizability, rather than replacing the need for a larger sample set and further studies.

With the equal-covariance-matrix class distributions, the linear discriminant is the optimal classifier as expected. The biases are low and the computation is efficient. Moreover, since the A_z-vs-$1/N$ relationship is linear over almost the entire range of design sample sizes, the classifier performance at very large N can be estimated from the small sample size performance by linear interpolation, as suggested by Fukunaga and Hayes[3] and demonstrated previously by Wagner *et al.*[9]

With the unequal-covariance-matrices and equal-mean class distributions, the linear discriminant and the back-propagation neural network with one hidden layer are inferior to the quadratic classifier when the design sample size is large. The linear discriminant cannot utilize the difference in the covariance matrices and underestimates the class separability even when an infinite number of design samples is available. The ANN needs a relatively large number of hidden nodes and a large number of training epochs in order to reach the optimal performance. Its hold-out performance and the computation efficiency are both inferior to those of the quadratic classifier. However, for the unequal-covariance-matrices and unequal-mean case and a small design sample size, the linear classifier or an ANN with very few hidden nodes, e.g., $n = 2$, provides better hold-out performance than the more complex ANNs or the optimal quadratic classifiers. These results indicate that the bias on classifier performance increases with increasing complexity (loosely related to the number of parameters to be estimated) of the classifier. The linear classifier contains $(k+1)$ independent parameters and the quadratic classifier contains $(k+1)(k+2)/2$ independent parameters in their formulations. The number of weights to be estimated for the ANN depends on the number of hidden nodes as $n(k+1)+(n+1)$. The number of weights in an ANN can therefore easily exceed that of a quadratic classifier, although the estimation of the mean and covariance matrices for the linear and quadratic discriminants may contribute additional "complexity" to the classifier design. Two observations can be made. First, when the available sample size is small, a simple classifier will have better generalization than a more complex classifier. Second, a complex ANN or a quadratic classifier trained with an insufficient number of design samples generalizes poorly, even if it is the optimal classifier for the class distributions. It is therefore important to select an appropriate classifier by taking into consideration the design sample size.

A further problem in classifier design is that the true population distributions of the classes in the feature space are generally unknown. It was suggested that the quantile–quantile (Q–Q) plot and the chi-square plot may be used for investigating the normality of univariate and multivariate sample distributions, respectively.[16] However, it is still unknown under what criteria the chi-square plot will indicate that it is optimal to use a classifier designed under the normality assumption. For any measure of goodness-of-fit, when the sample size is small, only the most aberrant deviations from the normal distribution can be identified as a lack of fit from these plots.[16] Therefore, there is often no *a priori* knowledge to select an "optimal" classifier or to predict whether the observed performance is caused by the sample size, the choice of an overly complex classifier, or by an actual poor separation of the classes in the feature space. If one observes poor generalization of a trained classifier in a

truly independent test set, it will be important to take into consideration all these factors and redesign the classifier.

In this study, we assumed that the best features have already been determined for the classification task. In a general classifier design problem, the best set of features usually has to be selected based on the available design samples. The feature selection step will introduce additional biases to the classifier performance. The number of features selected also has a strong influence on the classifier design, as can be seen from the dependence of the bias on the dimensionality of the feature space. The investigation of this more complex situation including both the feature selection and classifier training steps is underway.[17]

The term generalizability is nonspecific and needs to be qualified here. The present paper is concerned with the generalizability of the mean performance of classifiers to unknown test samples drawn from the same population of cases. We have shown in this paper that the mean performance of a classifier depends on the number of samples used to train the classifier, the architecture of the classifier, and—for multivariate-normal data—the means and covariances of the population distributions. Suppose in this context that a classifier is trained on a given finite number of design samples (patients). The mean performance of the classifier over independent replications with the same number of design samples is generalizable to studies characterized by the same number of design samples. In other words, the mean resubstitution or hold-out performance is an unbiased estimate for repeated sampling of independent design and test sample sets, respectively, when the same number of design samples is used. The classifier performance may not, however, be generalizable to studies characterized by a different number of design samples. In particular, when a very large and representative design sample size is used, the mean performance may be very different from the mean performance that characterizes the finite-training-sample condition. When the mean performance under the conditions of a finite design sample size is close to that expected with a very large design sample size, the finite-training sample performance is said to be generalizable to the population performance.

The term generalizability is not only used with respect to mean performance, it is also used with respect to uncertainty in performance, as reflected in estimates of error bars (standard deviations, or the corresponding variances). For example, if we think of repeating a given training and testing experiment on a classifier and if only the test samples are drawn independently on the repeated trials, then the estimated uncertainties are said to be generalizable only to a population of test samples. If, however, we think of repeating the experiment and independently drawing new training samples as well as new test samples, then the estimated uncertainties are said to be generalizable to a population of trainers and a population of testers.[17] Models for the components of variance in both paradigms are the subjects of current work in progress.[10,11] A key point of this latter work is the fact that for computer-aided diagnosis, most available software for ROC analysis only provides estimates of uncertainty that are generalizable to a population of test samples.

In this investigation, we have limited our study to only three types of classifiers: the linear discriminant, the quadratic discriminant, and the backpropagation ANNs with one hidden layer. There are, of course, many other variations of the ANN architecture and other parametric or non-parametric classifiers available for feature classification tasks. The purpose of our work is not to exhaustively evaluate all possible combinations of class distributions and classifiers. Rather, by limiting our investigation to some well-known situations, we can perform systematic analyses and gain some insights into the classifier design problems. Furthermore, we have limited our discussion here to the estimates of the mean classifier performance. Wagner *et al.*[10,11] have investigated the variances of classifier performance estimated from a finite sample set and developed models to study the relative importance of the sizes of the training and test samples. It has been demonstrated that a components-of-variance model can be estimated with a finite sample set by using a bootstrap method. More importantly, the analysis of variances can reveal the generalizability of the performance estimates to other training and test sample sets in the population. Our long term goals are to find some guidelines for designing efficient resampling schemes that can minimize the bias and variance of a trained classifier using the available samples, and to provide a quantitative design tool that can estimate the design sample size requirement for a larger "pivotal" study from the results of a smaller "pilot" study in order to achieve a desired precision in A_z and the desired generalizability.

ACKNOWLEDGMENTS

This work is supported in part by USPHS Grant No. CA 48129 and by a grant from the U.S. Army Medical Research and Materiel Command DAMD 17-96-1-6254, a Career Development Award (B.S.) DAMD 17-96-1-6012 from the U.S. Army Medical Research and Materiel Command and a Whitaker Foundation Grant (N. P.). The content of this paper does not necessarily reflect the position of the government and no official endorsement of any equipment and product of any companies mentioned in this paper should be inferred. The authors are grateful to Charles E. Metz, Ph. D., for providing the LABROC1 programs.

* Author to whom correspondence should be addressed. Department of Radiology, University of Michigan, 1500 E. Medical Center Drive, UHB1 F510B, Ann Arbor, MI 48109-0030; Phone: 734-936-4357; Fax: 734-936-7948. Electronic mail: chanhp@umich.edu

[3] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. (Academic, New York, 1990).

[5] S. Raudys and V. Pikelis, "On dimensionality, sample size, classification error, and complexity of classification algorithm in pattern recognition," IEEE Trans. Pattern. Anal. Mach. Intell. **PAMI-2**, 242–252 (1980).

[6] K. Fukunaga and R. R. Hayes, "Effects of sample size on classifier design," IEEE Trans. Pattern. Anal. Mach. Intell. **11**, 873–885 (1989).

[4]R. F. Wagner, D. G. Brown, J.-P. Guedon, K. J. Myers, and K. A. Wear, in *Information Processing in Medical Imaging*, edited by H. H. Barrett and A. F. Gmitro (Springer-Verlag, Berlin, 1993).

[5]R. F. Wagner, D. G. Brown, J.-P. Guedon, K. J. Myers, and K. A. Wear, "On combining a few diagnostic tests or features," Proc. SPIE **2167**, 503–512 (1994).

[6]D. G. Brown, A. C. Schneider, M. P. Anderson, and R. F. Wagner, "Effect of finite sample size and correlated/noisy input features on neural network pattern classification," Proc. SPIE **2167**, 180–190 (1994).

[7]H. P. Chan, B. Sahiner, R. F. Wagner, N. Petrick, and J. Mossoba, "Effects of sample size on classifier design: quadratic and neural network classifiers," Proc. SPIE **3034**, 1102–1113 (1997).

[8]H. P. Chan, B. Sahner, R. F. Wagner, and N. Petrick, "Effects of sample size on classifier design for computer-aided diagnosis," Proc. SPIE **3338**, 845–858 (1998).

[9]R. F. Wagner, H. P. Chan, B. Sahiner, N. Petrick, and J. T. Mossoba, "Finite-sample effects and resampling plans: applications to linear classifiers in computer-aided diagnosis," Proc. SPIE **3034**, 467–477 (1997).

[10]R. F. Wagner, H. P. Chan, J. T. Mossoba, B. Sahiner, and N. Petrick, "Components of variance in ROC analysis of CADx Classifier performance," Proc. SPIE **3338**, 859–875 (1998).

[11]R. F. Wagner, H. P. Chan, B. Sahiner, N. Petrick, and J. T. Mossoba, "Components of variance in ROC analysis of CADx classifier performance. II: Applications of the bootstrap," Proc. SPIE **3661**, 523–532 (1999).

[12]D. J. Hand, *Discrimination and Classification* (Wiley, New York, 1981).

[13]P. A. Lachenbruch, *Discriminant Analysis* (Hafner, New York, 1975).

[14]J. A. Freeman and D. M. Skapura, *Neural Networks-Algorithms, Applications, and Programming Techniques* (Addison-Wesley, Reading, 1991).

[15]H. P. Chan, B. Sahiner, R. F. Wagner, and N. Petrick, "Classifier design for computer-aided diagnosis in mammography: effects of finite sample size," Med. Phys. **24**, 1034–1035 (1997).

[16]R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis* (Prentice-Hall, Englewood Cliffs, NJ, 1982).

[17]C. A. Roe and C. E. Metz, "Variance-component modeling in the analysis of receiver operating characteristic index estimates," Acad. Radiol. **4**, 587–600 (1997).

Improvement of Radiologists' Characterization of Mammographic Masses by Using Computer-aided Diagnosis: An ROC Study[1]

Heang-Ping Chan, PhD
Berkman Sahiner, PhD
Mark A. Helvie, MD
Nicholas Petrick, PhD
Marilyn A. Roubidoux, MD
Todd E. Wilson, MD
Dorit D. Adler, MD
Chintana Paramagul, MD
Joel S. Newman, MD
Sethumadavan Sanjay-Gopal, PhD

Index terms:
Breast neoplasms, 00.31, 00.32
Breast neoplasms, radiography, 00.111, 00.119
Breast radiography, 00.111, 00.119
Computers, diagnostic aid
Receiver operating characteristic curve (ROC)

Radiology 1999; 212:817–827

Abbreviations:
CAD = computer-aided diagnosis
PPV = positive predictive value
ROC = receiver operating characteristic

[1] From the Department of Radiology, University of Michigan Hospital, UH B1F510, 1500 E Medical Center Dr, Ann Arbor, MI 48109-0030. From the 1997 RSNA scientific assembly. Received August 10, 1998; revision requested September 8; revision received November 30; accepted January 21, 1999. Supported in part by United States Public Health Service grant CA 48129 and by U.S. Army Medical Research and Materiel Command grant DAMD 17-96-1-6254. B.S. supported by Career Development award DAMD 17-96-1-6012 from the U.S. Army Medical Research and Materiel Command. N.P. supported by a grant from the Whitaker Foundation. **Address reprint requests to** H.P.C. (e-mail: *chanhp@umich.edu*).

The content of this article does not necessarily reflect the position of the funding agencies, and no official endorsement of any equipment or product of any companies mentioned in this article should be inferred.

© RSNA, 1999

Author contributions:
Guarantor of integrity of entire study, H.P.C.; study concepts and design, H.P.C., M.A.H., B.S., N.P.; literature research, H.P.C., M.A.H.; experimental studies, M.A.H., M.A.R., T.E.W., D.D.A., C.P., J.S.N.; data acquisition, all authors; data analysis, H.P.C., B.S., N.P.; statistical analysis, H.P.C.; manuscript preparation, editing, and review, H.P.C., B.S., M.A.H., N.P., M.A.R., T.E.W., D.D.A., C.P., J.S.N.

PURPOSE: To evaluate the effects of computer-aided diagnosis (CAD) on radiologists' classification of malignant and benign masses seen on mammograms.

MATERIALS AND METHODS: The authors previously developed an automated computer program for estimation of the relative malignancy rating of masses. In the present study, the authors conducted observer performance experiments with receiver operating characteristic (ROC) methodology to evaluate the effects of computer estimates on radiologists' confidence ratings. Six radiologists assessed biopsy-proved masses with and without CAD. Two experiments, one with a single view and the other with two views, were conducted. The classification accuracy was quantified by using the area under the ROC curve, A_z.

RESULTS: For the reading of 238 images, the A_z value for the computer classifier was 0.92. The radiologists' A_z values ranged from 0.79 to 0.92 without CAD and improved to 0.87–0.96 with CAD. For the reading of a subset of 76 paired views, the radiologists' A_z values ranged from 0.88 to 0.95 without CAD and improved to 0.93–0.97 with CAD. Improvements in the reading of the two sets of images were statistically significant (P = .022 and .007, respectively). An improved positive predictive value as a function of the false-negative fraction was predicted from the improved ROC curves.

CONCLUSION: CAD may be useful for assisting radiologists in classification of masses and thereby potentially help reduce unnecessary biopsies.

Breast cancer is the most prevalent non–skin cancer in women; 178,700 new cases are estimated to have occurred in 1998 (1). The mortality of breast cancer is the second highest among all cancer deaths in women (1). At present, there is no effective method to prevent breast cancer. The best approach to reducing the breast cancer mortality rate is early detection and treatment. Because the mammographic features of early-stage breast cancers are not very specific, the need for high detection sensitivity leads to biopsy of many low-suspicion lesions. The positive predictive values (PPVs) of mammographic signs are, therefore, often below 30% (2,3).

Computer-aided diagnosis (CAD) is considered to be one of the approaches that may improve the efficacy of mammography (4). With CAD, a computerized detection algorithm alerts a radiologist to the location of the suspicious lesions, and/or a trained computer classifier provides the radiologist with an estimate of the likelihood of malignancy of a lesion. The radiologist takes into consideration the information provided by the computer before making a decision. This "second opinion" may improve the diagnostic accuracy because it serves as a form of double reading (5). Furthermore, a computer evaluation is often more consistent and reproducible than a human decision maker (6).

Considerable research has been devoted to the development of computerized schemes for the detection and classification of mammographic abnormalities. These efforts have advanced the CAD technology such that clinical application appears to be possible in the

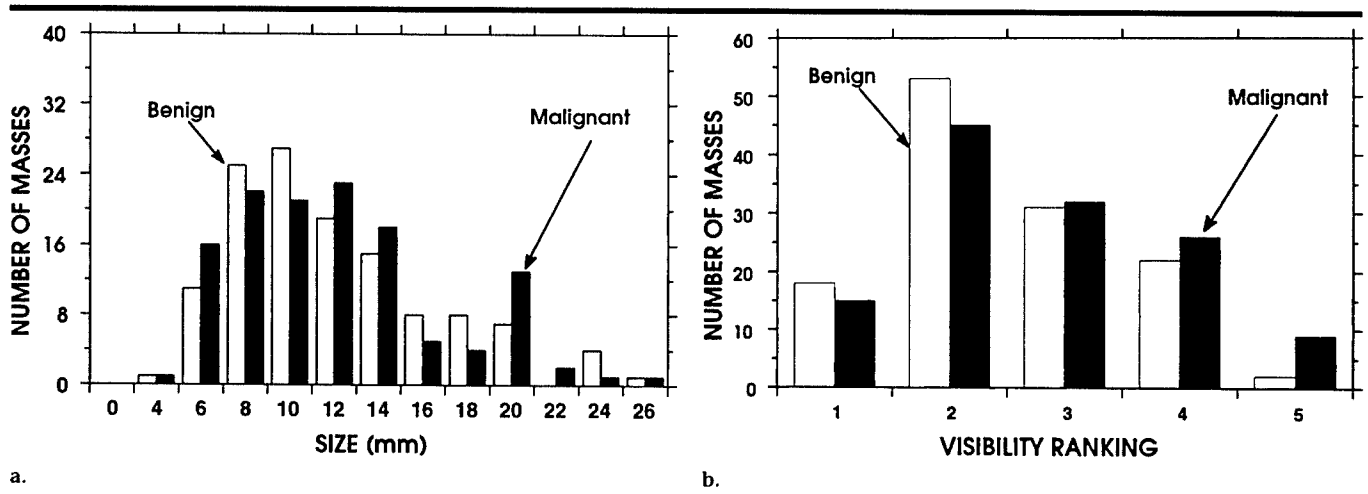


Figure 1. Histograms illustrate the distributions of (a) size (ie, length of the long axis) and (b) visibility ranking (1 = obvious, 5 = subtle) of the 253 masses included in the data set. Because classification accuracy depends on the case mix, these distributions provided some information on the masses in the data set.

near future. It is, therefore, necessary to evaluate the effects of CAD on radiologists' detection and diagnosis of mammographic lesions. In a previous receiver operating characteristic (ROC) study, we demonstrated that CAD could improve radiologists' accuracy in the detection of subtle microcalcifications on mammograms (7). Kegelmeyer et al (8) also reported an improvement in radiologists' sensitivity for the detection of spiculated masses with use of a computer aid. For the classification of mammographic lesions, it has been shown that a computer classifier that estimated the likelihood of malignancy on the basis of mammographic features extracted by radiologists could improve radiologists' accuracy in distinguishing malignant from benign lesions (9–11).

We previously conducted ROC studies to compare the performance of radiologists with that of the computer (12) and to compare radiologists' ability to classify masses with and without CAD (13). Jiang et al (14) also performed an ROC study of the effect of CAD on radiologists' performance in classifying microcalcifications. The results of all of these observer performance studies indicate the potential to improve mammographic interpretation with a computer aid.

We have developed an automated method to analyze masses seen on mammograms (15–17). A mass is segmented from its surrounding breast tissue, and an image transformation technique is used to transform the mass margin from the polar coordinate system to the Cartesian coordinate system. A linear discriminant classifier then extracts the useful texture features from the transformed image and merges them into a relative malignancy rating. Our approach is different from others that use a trained classifier to merge radiologist-extracted image features or feature codes by using the American College of Radiology Breast Imaging Reporting and Database System lexicon (9–11). Our fully automated method has the advantage that, unlike a human reader, it does not have variability in feature recognition and coding. In addition, the computer may be able to extract some information, such as texture features, that may not be readily perceived by human eyes. We conducted an ROC study to evaluate whether this computer aid can improve radiologists' performance in the classification of mammographic masses (13). The results of our observer performance study are described in this article.

Other investigators also have reported on automated algorithms for the classification of mammographic masses (18–21). The methods used in these algorithms varied, and their accuracy in classification cannot be compared directly because of the differences in the data sets. However, the effects of CAD on radiologists' performance are not expected to depend strongly on the specific algorithm if different computer aids of comparable accuracy are used. Therefore, the applications of the findings of this study should not be limited to our computerized classification aid.

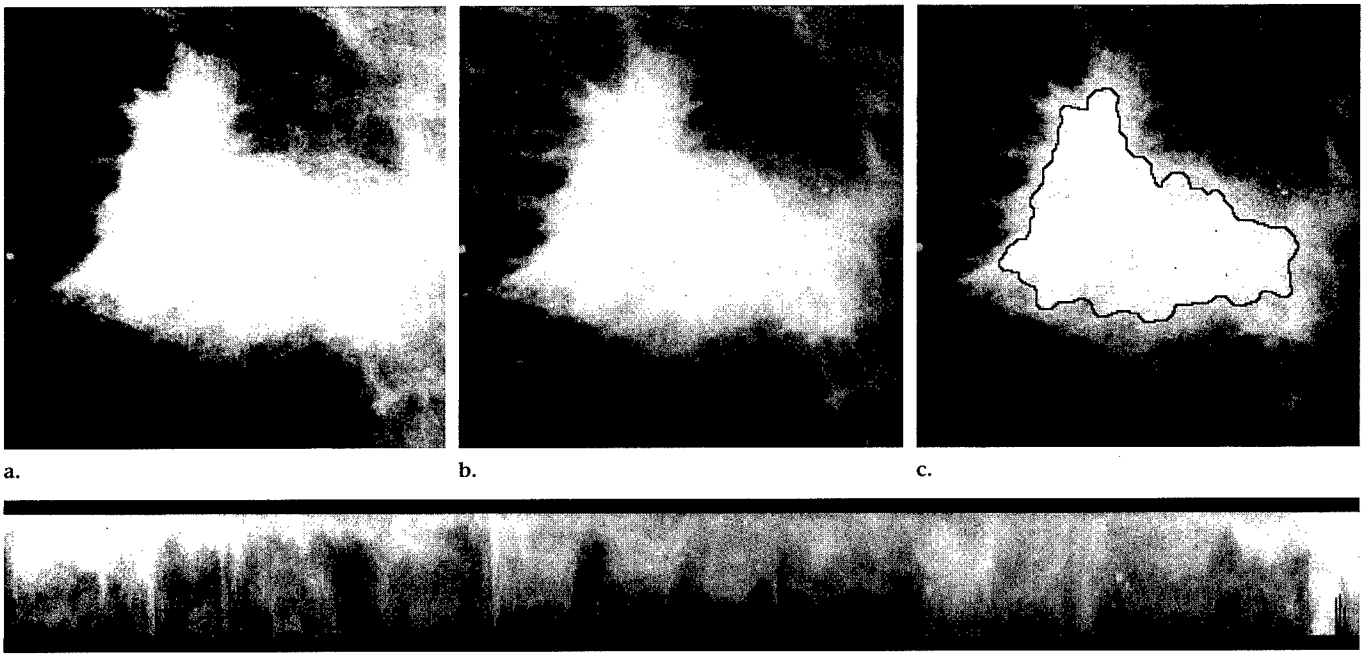
MATERIALS AND METHODS

Data Set

The data set for this study consisted of 253 mammograms obtained in 103 patients. Each image contained a biopsy-proved mass that was evaluated in this study. Some cases involved multiple views or images from multiple examinations. The cases were randomly selected from patient files from the breast imaging division of a National Cancer Institute–designated national cancer center with the approval of the Institutional Review Board. The PPV of masses recommended for biopsy at this center is about 25%–30%, but an approximately equal number of malignant and benign masses (127 and 126, respectively) were chosen to enhance the statistical power in this observer performance study. Any images that were judged to be technically poor were excluded.

The mammograms were acquired with a contact technique. The dedicated mammographic systems had a molybdenum anode and molybdenum filter, a 0.3-mm nominal focal spot, and a reciprocating grid. MinR/MinR-E screen-film systems (Eastman-Kodak, Rochester, NY) were used with these units. Sixty-two of the malignant masses and six of the benign masses were judged to be spiculated by a radiologist (M.A.H.) experienced in mammography. The radiologist also measured the size (ie, longest dimension) and ranked the visibility of the masses on a scale of 1 (obvious) to 5 (subtle) relative to the range of visibility of masses encountered in clinical practice. For a description of the masses included in the data set, histograms of the size and visibility of the masses are shown in Figures 1a and 1b, respectively.

For the computer analysis, the selected mammograms were digitized with a laser



a. b. c.

d.

Figure 2. Example of rubber-band-straightening transform for extraction of texture features in the margin region surrounding a mass. **(a)** Original and **(b)** background-corrected images showing the region of interest with the mass, **(c)** mammogram showing an outline of the segmented mass, and **(d)** rubber-band-straightening–transformed image of a 40-pixel-wide region surrounding the segmented mass.

imager (Lumisys DIS-1000, Los Altos, Calif) at a pixel size of 0.1 × 0.1 mm and 12-bit gray levels. This imager has an optical density range of about 0.0–3.5. The optical density on the film was digitized linearly to pixel value at a calibration of 0.001 optical density unit/pixel value in the optical density range of about 0.0–2.8. The digitizer deviated from a linear response at an optical density higher than 2.8.

For the observer experiments, we used laser-printed images of the digitized mammograms for all readings. The images were printed with a 969HQ laser imager (Imation, Oakdale, Minn) that was connected to a Macintosh computer (Apple Computer, Cupertino, Calif) through a special digital interface. The interface provided a 12-bit in, 10-bit out look-up table and allowed images to be scaled to different factors with 15 interpolation methods. Because this laser imager has a pixel size of about 0.085 mm, we enlarged the images by about 18% during printing to maintain them at the same size as the original mammograms. One of the interpolation methods was chosen by an experienced radiologist (M.A.H.), who inspected the printed images with a magnifier and evaluated the sharpness of the spicules and mass boundaries. Because of the small pixel size used for both digitization and printing, basically no noticeable blurring of the masses could be seen with the chosen interpolation method. The images were also inspected for the potential contouring effect of 10-bit output images, but no noticeable artifacts could be found. A linear pixel value–to–output optical density calibration curve of the laser imager was used for the printing. All images were printed with the same settings.

Computerized Classification of Masses

Our computerized method of classifying mammographic masses has been described in detail previously (15–17). The method is summarized as follows: A region of interest that contained the biopsy-proved mass was identified on the mammogram by the radiologist. Background correction based on a distance-weighted estimation method was applied to the region of interest to reduce the low-frequency density variation in the region. A median-filtered smoothed image and two high-frequency enhanced images were generated from the background-corrected region of interest. The smoothed and enhanced gray-level values at each pixel were used as features in a k-means clustering algorithm to classify the pixels into two clusters; one was the mass, and the other was the surrounding breast tissue background. By choosing an appropriate criterion, a mass region slightly smaller than the actual mass that was visible on the image was segmented.

The boundary of the segmented region was smoothed by morphologic filtering. A new image transformation technique, referred to as the rubber-band-straightening transform, was used to transform a 40-pixel-wide region that surrounded the segmented mass boundary into a rectangular region. After transformation, the mass margin became approximately parallel, and any spicules that were radiating from the mass became approximately perpendicular, to the long dimension of the rectangular region. The rubber-band-straightening transform enabled the spicules to be aligned approximately in a uniform direction and thus facilitated the extraction of texture features from the margin of the mass. An example of a rubber-band-straightening–transformed image is shown in Figure 2.

Two types of texture features were found to be useful for classification. The first set of features included eight texture measures derived from the spatial gray-level dependence matrices of the rubber-band-straightening–transformed image. A spatial gray-level dependence matrix ele-

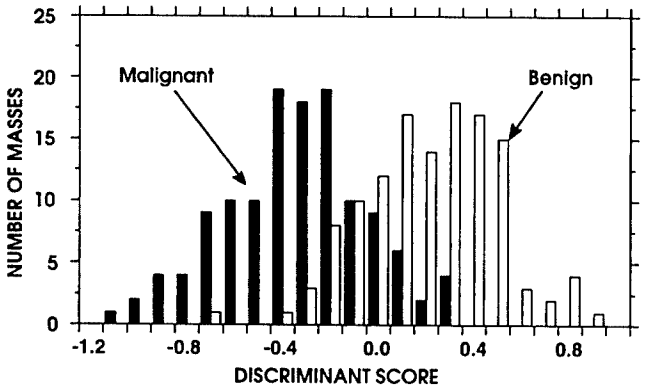


Figure 3. Histogram of the test discriminant scores of the 253 masses obtained from the linear discriminant classifier by using a "leave one case out" training and test resampling scheme. For this classifier, a smaller discriminant score corresponded to a higher likelihood of malignancy. The discriminant scores were used as the decision variable in the ROC analysis of classification performance.

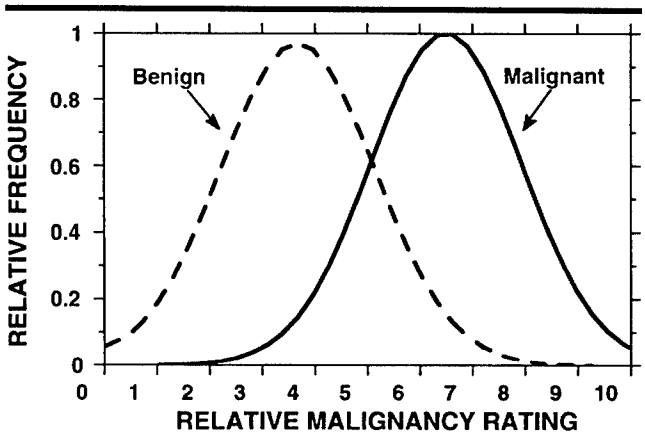


Figure 4. Binormal distribution fitted to the histogram of the discriminant scores of the malignant and benign masses. The discriminant scores were linearly transformed into a relative malignancy rating ranging from 1 to 10, where 1 corresponded to the most benign rating and 10 corresponded to the most malignant rating. This binormal distribution was shown to the observers during the training session to explain the rating scale of the computer classifier.

ment $p_{\theta,d}(i,j)$ is the joint probability of the occurrence of gray levels i and j for pixel pairs that are separated by a distance d and at a direction θ (22). For analysis of the masses, the spatial gray-level dependence matrices were constructed for 10 pixel distances (d = 1, 2, 3, 4, 6, 8, 10, 12, 16, 20 pixels) and in four directions (0°, 45°, 90°, 135°) relative to the mass boundary. Therefore, a total of 320 spatial gray-level dependence texture features were extracted.

The second set of texture features was derived from the run length statistics matrices of the horizontal and vertical gradient images of the rubber-band-straightening–transformed margin region. Five texture measures were extracted from the run length statistics matrix in each of the two directions (0° or 90°) on each gradient image. A total of 20 run length statistics texture features were thus obtained. Therefore, we had a total of 340 features from the two types of texture measures.

A stepwise linear discriminant feature selection procedure (23) was used to select the most effective features from the available feature set. A total of 41 features were selected. The selected features were input into the Fischer linear discriminant classifier (24) as predictor variables. A "leave one case out" resampling scheme was used to train and test the classifier. A histogram illustrating the test discriminant scores of the 253 masses is shown in Figure 3. For this classifier, a smaller discriminant score corresponded to a higher likelihood of malignancy. By using the test discriminant score as the decision variable, the performance of the computer classifier could be evaluated by using ROC analysis (17,25,26) and compared with that of the radiologists, as described later.

Relative Malignancy Rating of the Masses

For the observer performance study, we provided a relative malignancy rating of each mass to the observer during the reading session with CAD. The relative malignancy rating was obtained by taking a linear transformation of the computer classifier's decision variable to a range of 1–10 and rounding the value to the nearest integer. The transformation also reversed the relative magnitude of the decision variables so that 1 corresponded to the highest benignity rating, and 10 corresponded to the highest malignancy rating.

The purpose of the transformation was to provide a simple and intuitive relative scale for the observer. Because the transformation was linear and monotonic, the distributions of the normal and abnormal samples, as well as their ROC curves, were not affected, with the exception of a small error caused by making the decision variables discrete. Furthermore, the slope a and intercept b parameters that were fitted to the transformed discriminant scores for the normal and abnormal samples by using the LABROC program (26) were used to generate a binormal distribution. The fitted binormal distribution with the relative malignancy rating on a 1–10 scale (Fig 4), together with the computer's ROC curve, were shown and explained to the observers during a training session.

Observer Performance Study

Two ROC experiments (27) were conducted: The masses were evaluated from a single view in the first experiment and from two views in the second experiment. The location of the biopsy-proved mass was marked on each image so that the correct mass was evaluated by all observers. The observers were instructed to ignore any other possible masses on the images. Six radiologists (M.A.H., M.A.R., T.E.W., D.D.A., C.P., J.S.N.) who are approved by the Mammography Quality Standards Act and have 7–20 years of experience in interpreting mammograms participated in the observer performance experiments.

There were two reading sessions in each experiment—one with CAD and the other without CAD. The observers were asked to rate the likelihood of malignancy of the masses on a 10-point confidence rating scale under all reading conditions. In the first session, half the observers interpreted the images without CAD, and the other half interpreted them with CAD. The two reading sessions in the same experiment were separated by at least 3 weeks, and the two experiments were separated by 6 months. For all four reading sessions, the observer had unlimited time to read each case. To estimate the average reading time per case for each observer, the reading time for each case was recorded by using a stopwatch.

In the first experiment, the data set of 253 single-view mammograms was divided into a training set of 15 mammograms and a study set of 238 mammo-

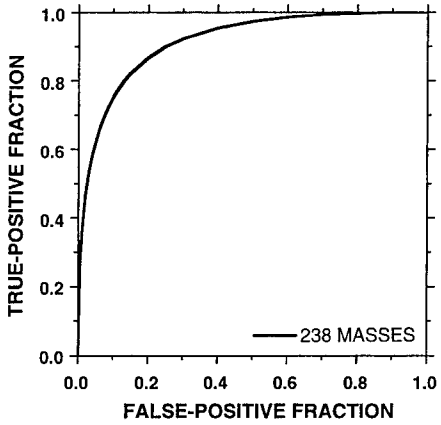


Figure 5. ROC curve for computerized classification of the 238 masses used in the observer performance study with single-view reading. The computer's ROC curve can be compared with the radiologists' ROC curves obtained from the single-view reading experiment illustrated in Figures 6 and 8.

grams (117 benign, 121 malignant). In each reading session, training was conducted before the reading of the study images. For the reading session with CAD, the fitted binormal distributions of the computer rating scores (Fig 4) for the entire data set were explained to the observer during training to familiarize the observer with the computer's rating scale. The computer rating of the mass was displayed on each image. After reading each training image, the observer was told the results of biopsy of the mass.

Each observer read the entire data set in one reading session. The order of the study images was randomized by a random number generator. The random sequence was different for each observer and for each reading session by the same observer. For the reading session with CAD, the observer was free to look at the computer rating, which was displayed on the image, either before or after estimating the likelihood of malignancy of the mass. However, each observer was asked to always read the computer rating before making a final decision. The observer was not informed of the pathologic results of any mass on the study images.

The second experiment was very similar to the first experiment. From the 238 single-view mammograms, 76 matched pairs (37 benign, 39 malignant) of craniocaudal and mediolateral oblique or lateral views were found. Another six pairs of two-view mammograms were identified from the rest of the images and used as training cases. The remaining mammograms were either single-view images or additional views of the pairs already chosen, so they were not used in this experiment. In this experiment, the observers were not informed of the pathologic results of any study case in any reading session. The 76 pairs of mammograms were read in one reading session by each observer.

For the reading session with CAD, the rating of the mass in each view was displayed on the respective image. The computer ratings of the mass on the two views were generally different. It was up to the observer to decide how to merge the two-view information. Observers were asked to give a single rating of the mass after reading both views.

ROC Analysis

The confidence ratings of each observer obtained from each reading condition were analyzed by using ROC methodology, and the classification accuracy was quantified by using the area under the ROC curve, A_z. A maximum likelihood estimation of the binormal distribution was fitted to the confidence ratings by using the LABROC program. This program provides an estimate of the A_z and of the a and b parameters of the ROC curve. The statistical significance of the difference in A_z between the reading with CAD and that without CAD was estimated with two methods: One was the Student paired t test for observer-specific paired data; the other was the Dorfman-Berbaum-Metz method for analysis of multireader, multicase ROC data (28). The statistical significance of the difference in $A_z$for reading single-view and two-view mammograms was estimated by using the Student paired t test for the six observers. The Student paired t test takes into account the statistical variation of readers, whereas the Dorfman-Berbaum-Metz method considers both reader variation and case sample variation by means of an analysis of variance approach. Therefore, the results of Dorfman-Berbaum-Metz analysis can be generalized to the population of readers as well as to the population of case samples.

Positive Predictive Value

An ROC curve represents the entire range of operating conditions of a diagnostic process and is independent of disease prevalence. When the disease prevalence is known, any operating point on an ROC curve can be used to derive the PPV and the corresponding false-negative fraction (false-negative fraction = 1 − true-positive fraction) on the basis of the following relationship: PPV = TPF × P(M)/[TPF × P(M) + FPF × P(B)], where TPF is the true-positive fraction, FPF is the false-positive fraction at the chosen decision threshold, and P(M) and P(B) are the prevalences of malignant and benign cases, respectively. By varying the decision threshold, the dependence of the PPV on the false-negative fraction can be derived.

Because our data set did not include masses on which biopsy had not been performed, the ROC curves obtained in this study cannot be generalized to predict the performance of the computer classifier and radiologists in clinical practice. However, to demonstrate the possible effect of CAD on the PPV in the population of masses in which biopsy is likely to be performed under the current clinical criteria, we can estimate the PPV by using the prevalence of the malignant and benign masses in this patient group. Because the PPV of masses sent for biopsy ranges from about 25% to 44% in general and from about 25% to 30% at our institution, for the purposes of our estimation, we assumed that the P(M) was 25% and the P(B) was 75% in this population. A higher prevalence of malignant cases would cause an increase in the PPV, but the trend between the PPV curves with and without CAD would be similar.

RESULTS

The ROC curve illustrating the performance of the computer classifier for the 238 study mammograms is shown in Figure 5. The ROC curve for the entire set of 253 mammograms (not shown) was almost identical to that of the 238 study cases; this indicates that the 15 training cases were typical of the 238 cases used in the study. The A_z values (± SD) for both ROC curves were 0.92 ± 0.02.

For the first experiment of reading the 238 single-view mammograms, the ROC curves for the readings by the six radiologists both without and with CAD are shown in Figures 6a and 6b, respectively. The A_z values of the six radiologists for the readings with and without CAD are listed in Table 1.

For the second experiment of reading the 76 pairs of two-view mammograms, the ROC curves for the readings by the six radiologists both without and with CAD are shown in Figures 7a and Figure 7b, respectively. The A_z values of the six radiologists in this experiment are also listed in Table 1.

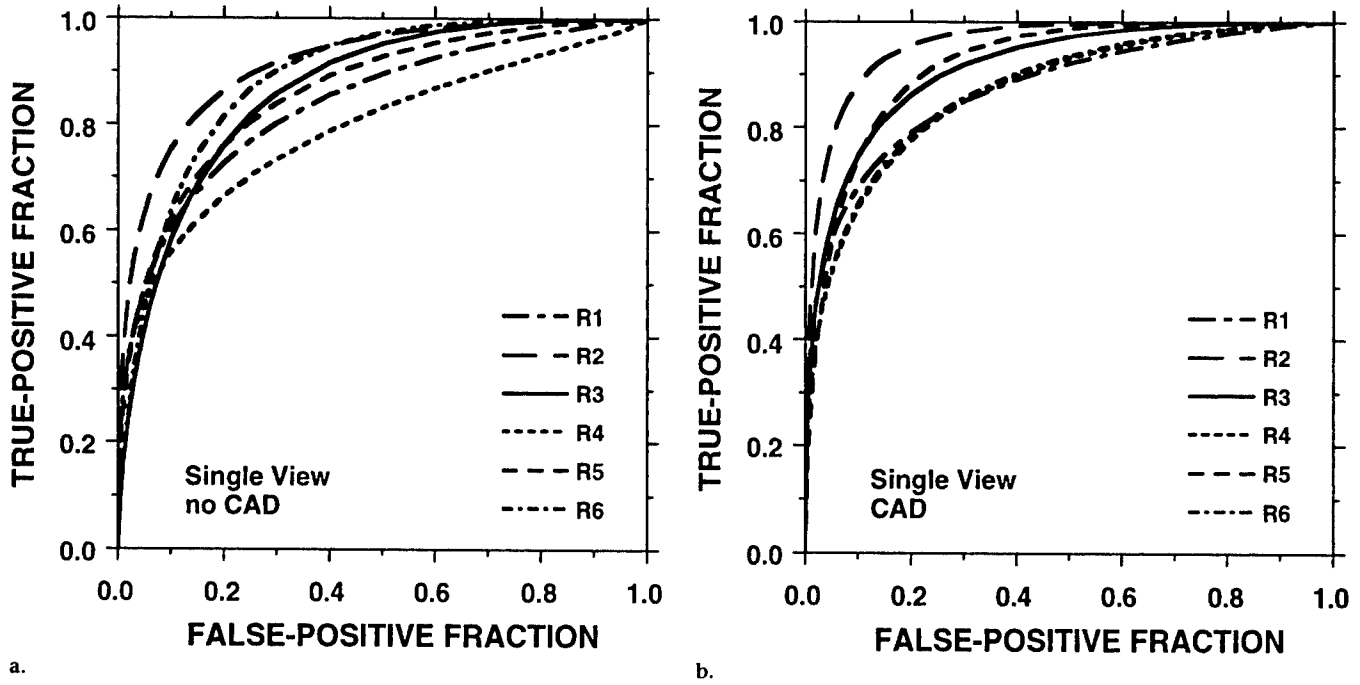


Figure 6. ROC curves for the six observers for single-view reading of the masses (a) without CAD and (b) with CAD. (a, b) *R1* = reader 1, *R2* = reader 2, *R3* = reader 3, *R4* = reader 4, *R5* = reader 5, *R6* = reader 6. Five of the six observers achieved an increase in the area under the ROC curve, A_z, with CAD.

The average ROC curve was derived from the average *a* and *b* parameters of the six individual ROC curves for a given reading condition (27). The average ROC curves for the four reading conditions are shown in Figure 8. The A_z values of the average ROC curves are listed in Table 1.

For the reading of the single-view mammograms, the performance of the computer classifier was comparable to that of the radiologist (reader 2) who had the highest classification accuracy (compare Figs 5 and 6) and higher than the average performance of the six radiologists (compare Figs 5 and 8). When the radiologists read the images with the computer aid, the classification accuracy of five radiologists improved (Table 1); the improvement in their A_z values ranged from 0.04 to 0.08. The average performance of the six radiologists became comparable to that of the computer classifier. The improvement in the radiologists' classification accuracy by using CAD was statistically significant ($P = .022$, Student paired *t* test; $P = .020$, Dorfman-Berbaum-Metz method). Reader 2 with CAD obtained an A_z value of 0.96, which was higher than that obtained by the radiologist alone or by the computer alone.

A trend similar to that with the single-view readings was observed with the two-view readings. The A_z value of the computer classifier for the corresponding 152 single-view masses was 0.91 ± 0.02. The classification accuracy of all six radiologists improved when they read the mammograms with the computer aid. The increase in the A_z values ranged from 0.01 to 0.07. The improvement was statistically significant ($P = .007$, Student paired *t* test; $P = .026$, Dorfman-Berbaum-Metz method). With CAD, two radiologists achieved an A_z value of 0.97, which was higher than that obtained by the radiologists alone or by the computer alone. These results indicate that the second opinion provided by the computer classifier might have strengthened the radiologists' confidence in the interpretation of some difficult cases but had less influence on the radiologists' decision when the computer made mistakes or when the radiologists were confident about their decision.

As can be seen from the data in Table 1,

TABLE 1
Areas under the ROC Curves for the Classification of Masses with and without CAD by the Six Radiologists

Radiologist No.	A_z (Single View)*		A_z (Two View)†	
	Without CAD	With CAD	Without CAD	With CAD
1	0.84 ± 0.03	0.87 ± 0.02	0.90 ± 0.03	0.93 ± 0.03
2	0.92 ± 0.02	0.96 ± 0.01	0.95 ± 0.02	0.97 ± 0.02
3	0.86 ± 0.02	0.91 ± 0.02	0.92 ± 0.03	0.93 ± 0.03
4	0.79 ± 0.03	0.87 ± 0.02	0.88 ± 0.04	0.95 ± 0.03
5	0.86 ± 0.02	0.92 ± 0.02	0.93 ± 0.03	0.97 ± 0.02
6	0.89 ± 0.02	0.87 ± 0.02	0.89 ± 0.04	0.93 ± 0.03
A_z from average a, b parameters	0.87	0.91	0.92	0.96

Note.—Data are the mean ± SD.

* $P = .022$ for the difference between the A_z values measured with CAD and those measured without CAD, as determined by using the Student two-tailed *t* test. $P = .020$ for this difference, as determined by using the Dorfman-Berbaum-Metz method.

† $P = .007$ for the difference between A_z values measured with CAD and those measured without CAD, as determined by using the Student two-tailed *t* test. $P = .026$ for this difference, as determined by using the Dorfman-Berbaum-Metz method.

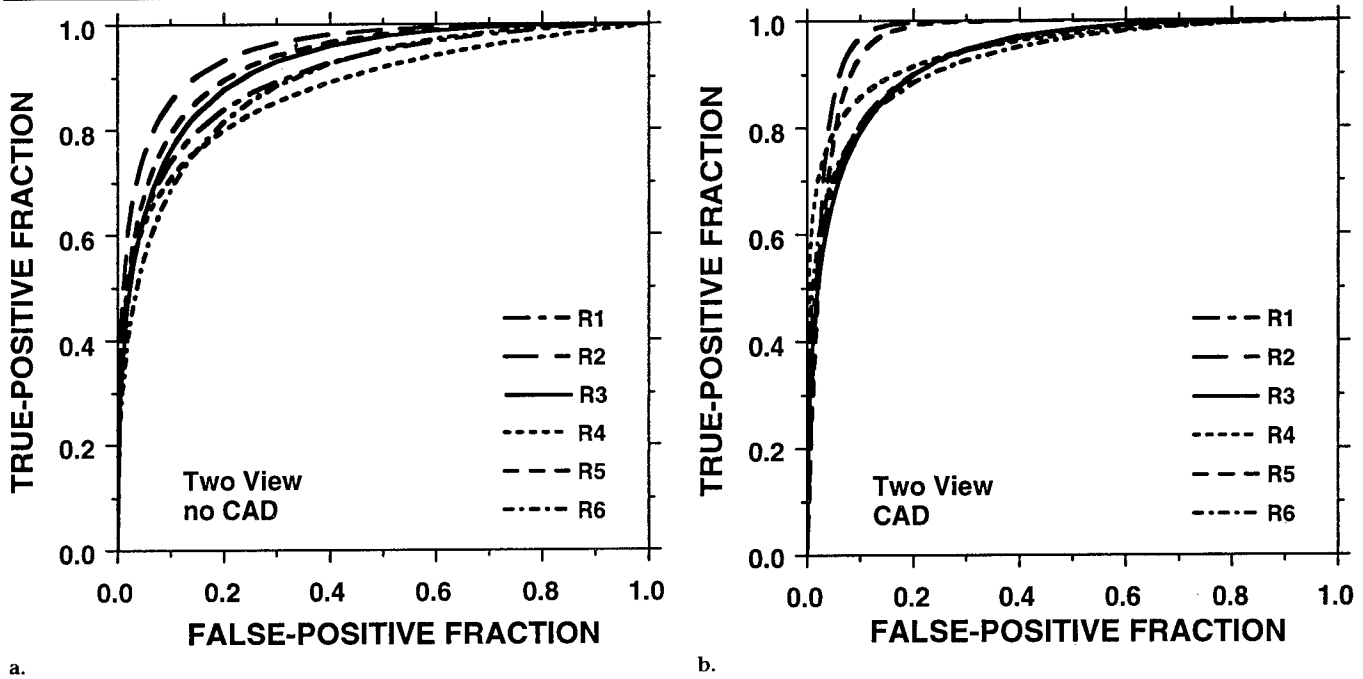


Figure 7. ROC curves for the six observers for two-view reading of the masses **(a)** without CAD and **(b)** with CAD. **(a, b)** *R1* = reader 1, *R2* = reader 2, *R3* = reader 3, *R4* = reader 4, *R5* = reader 5, *R6* = reader 6. All six observers achieved an increase in the area under the ROC curve, A_z, with CAD.

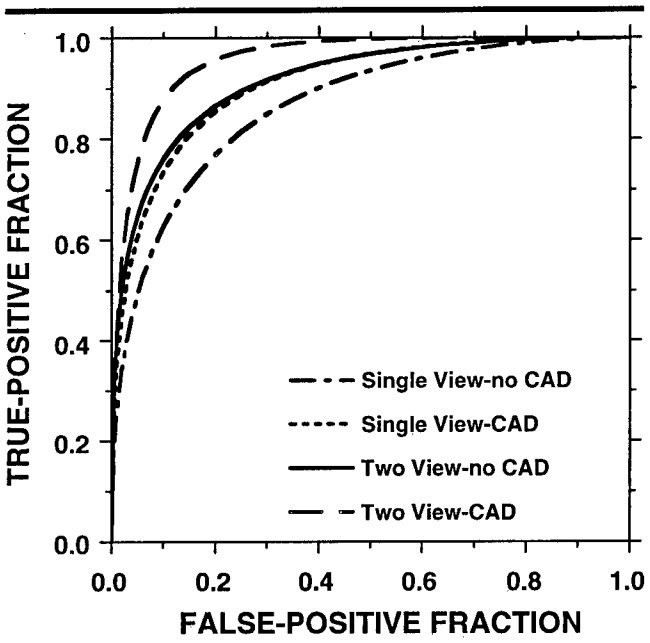


Figure 8. Average ROC curve obtained from the average *a* and *b* parameters of the six individual ROC curves for each of the four reading conditions. An improved ROC curve was achieved with CAD in both the single-view and two-view reading experiments.

the radiologists' accuracy in classifying masses by reading two-view mammograms was consistently higher than that by reading single-view mammograms (P = .008). This trend remained when they read the mammograms with CAD (P = .007). These findings are consistent with the clinical experience of the radiologists that at least two views of mammograms are needed to effectively evaluate a suspicious lesion.

The PPV as a function of the false-negative fraction was derived from the fitted ROC curves under the assumption that the prevalence of malignant masses was 25% in the population of masses sent for biopsy. The PPVs estimated for the six observers who read the two-view mammograms with and without CAD are plotted in Figure 9. CAD would provide an improvement in the PPV in the high false-negative fraction range for all observers except readers 2 and 5. The increase in the PPV at a decision threshold of "no missed malignant mass" (ie, false-negative fraction = 0) varied over a wide range; the largest gain, 39%, would be achieved by reader 2, and the smallest gain, 0%, would be achieved by reader 4.

DISCUSSION

In the observer experiment of reading two-view mammograms with CAD, we presented the computer's rating of each view separately. The decision of how to merge the computer ratings of the two views was left to the radiologist. It is likely that the radiologists took the conservative approach of using the highest malignancy rating of the two as the computer's overall rating. However, it also might have depended on whether the relative ranking between the two computer ratings agreed with the observer's opinion. In some cases, we observed that the radiologist's rating was very different from the computer's rating of either view.

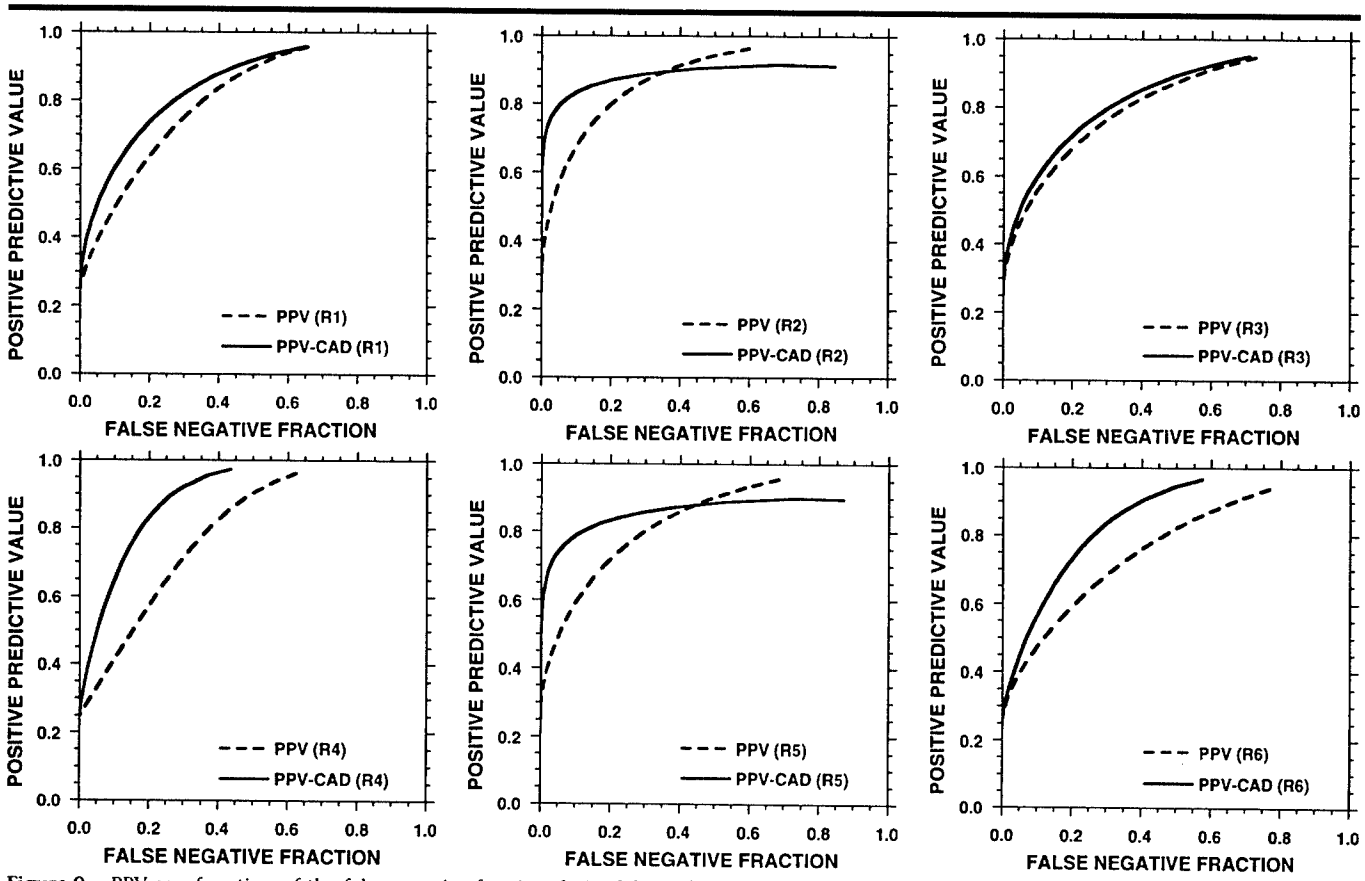


Figure 9. PPV as a function of the false-negative fraction derived from the ROC curves for the six observers (Fig 7). The PPV was predicted for a population of masses in which biopsy was likely to be performed under current clinical criteria and by assuming the prevalence of malignant masses to be 25%. *R1* = reader 1, *R2* = reader 2, *R3* = reader 3, *R4* = reader 4, *R5* = reader 5, *R6* = reader 6.

Because decision making is a complex process, the simple approach of using the highest malignant rating or the average rating from multiple views may not be the method preferred by radiologists. The separate ratings that we used in this study would provide less biased information. Further investigation is needed to determine the best approach of presenting the computer's ratings to radiologists in clinical practice.

To obtain insight into how the radiologists might use the two-view information, we compared the classification results from their true two-view reading with those from a simulated two-view reading without the computer aid. The latter results were derived from ratings of single-view readings of the same 76 pairs of mammograms interpreted in experiment 2 by assuming two strategies—one in which the highest malignancy rating between the two ratings was used, and the other in which the average of the two ratings was used (Table 2). The A_z values for these classification ratings derived from the single-view reading are listed in Table 2. The corresponding A_z values for the computer classifier are also given in Table 2 for comparison.

The A_z values for the maximal rating and the average rating were similar. Four of the radiologists obtained higher A_z values at the true two-view reading; the A_z values obtained by the remaining two radiologists were lower than those obtained at the simulated two-view reading. Although the difference did not achieve statistical significance ($P = .37$) and both readings included intraobserver variations, there seemed to be a slight trend toward the true two-view reading being more accurate than the simulated two-view reading. This may indicate that the radiologists used a more complex decision-making process to interpret the two views of the masses than that of simply maximizing or averaging the ratings from each view.

In this study, the discriminant scores of the masses given by the computer classifier were transformed into a relative malignancy rating. The relative malignancy rating scale and the distribution of the malignant and benign masses along the relative rating scale were explained to the observers in the training sessions. A relative malignancy rating scale was used because the true likelihood of malignancy of the masses could not be estimated from a small data set, as will be explained. However, the relative rating scale provided by the computer was ad-

TABLE 2
Estimation of the Malignancy Classification of 76 Masses by Two-View Reading, as Simulated from Single-View Reading of Mammograms by Radiologists without CAD

Radiologist No.	A_z	
	Maximal Rating	Average Rating
1	0.94 ± 0.03	0.93 ± 0.03
2	0.94 ± 0.03	0.94 ± 0.03
3	0.84 ± 0.05	0.86 ± 0.04
4	0.85 ± 0.04	0.83 ± 0.05
5	0.88 ± 0.04	0.89 ± 0.04
6	0.91 ± 0.03	0.92 ± 0.03
Computer	0.96 ± 0.02	0.96 ± 0.02

Note.—Data are the mean ± SD. Two strategies were used: In one, the highest of the malignancy ratings on each view was used; in the other, the average between the two ratings was used.

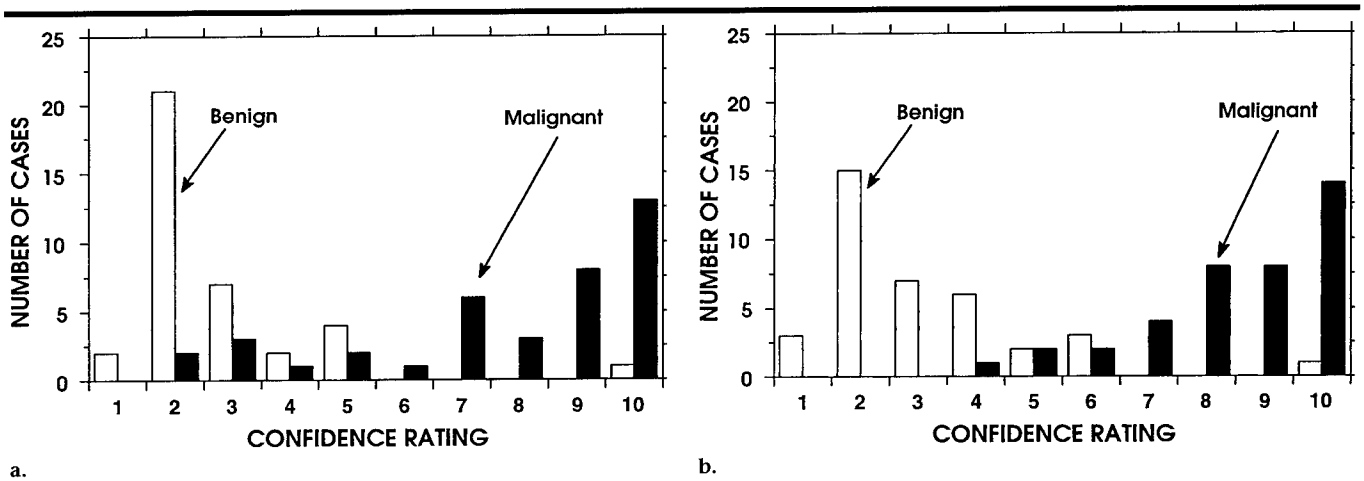


Figure 10. Histograms illustrate the confidence ratings of reader 5 obtained by reading 76 two-view mammograms **(a)** without CAD and **(b)** with CAD. The specificity of reader 5 at 100% sensitivity would increase from 5% (two of 37 masses) without CAD to 68% (25 of 37 masses) with CAD if an appropriate decision threshold were chosen.

equate for measuring the relative performance of classification with and without CAD in an ROC study.

If a computer classifier is trained and tested with very large data sets, and if both the malignant and benign cases represent random samples of the population, then the likelihood of malignancy of a classified mass can be estimated on the basis of the probability distributions of the classifier's test output scores and the prevalence of the two classes of masses in the patient population. However, with a relatively small data set, such as that used in this and other observer studies (14), there are limitations. First, the performance of a classifier trained with a small sample set may have large bias and variance (29–31). Second, the data set in this study did not include masses on which biopsy was not performed, so it did not represent a random sample of the masses in the patient population. If our classifier were applied to all cases of solid masses in clinical practice, the probability distribution of the test scores for the two classes of masses would be different from that of the current data set.

If we ignore the patient population at large, it is possible to estimate the likelihood of malignancy of a mass on the basis of the probability distribution of the classifier output scores by using the prevalence of the two classes of masses in this specific data set. However, the likelihood of malignancy derived in this way will be completely different from the true likelihood of malignancy of a mass in the patient population. This can be easily seen if one considers that the same mass with the same discriminant score will have a smaller likelihood of malignancy if it is analyzed within a data set that has a lower prevalence of malignant cases than that in the current data set.

Training the participating radiologists with a "likelihood of malignancy" derived from a small data set for the observer experiment may mislead them if they encounter a similar mass in their clinical practice. We, therefore, preferred to use a "relative malignancy rating," which is independent of the prevalences of malignant and benign masses in the data set. As long as the same classifier and the same linear transformation are used for classifying masses, the relative malignancy rating for a given mass will remain the same, regardless of the types of other masses in the data set. When a computer classifier is implemented in a clinical setting and its performance can be established in the patient population, the true likelihood of malignancy of a given mass can be estimated and provided to the radiologist. The true likelihood of malignancy may be a more informative measure for radiologists in the clinical application of CAD.

For the reading of the 76 two-view mammograms, the results of the ROC study indicated an improvement in the A_z value for all six radiologists when the computer aid was used. This indicates an overall increase in the separation of confidence rating distributions between the malignant and benign cases. The histograms in Figure 10 illustrate the distributions of confidence ratings with and without CAD for reader 5, who achieved the second greatest improvement in both the A_z value (Table 1) and the separation of malignant from benign distributions. Without CAD, this reader's ratings of the malignant cases ranged from 2 to 10. This is consistent with the fact that biopsy was performed in all masses in the data set to avoid missing the malignant cases. With CAD, there was marked improvement in the separation of the two distributions. It is possible to set a decision threshold at a confidence rating of 4, below which biopsy would not need to be performed and no malignant masses would be missed. The number of benign masses that could be identified without missing a malignant mass by setting an appropriate threshold would increase by 23 (out of 76 cases) for reader 5. Five of the six radiologists in our ROC study achieved an improvement in distinguishing benign from malignant masses, and one radiologist had no difference. Although the improvement of the five radiologists varied over a wide range, from one to 25 cases, this result indicates a strong possibility that CAD can be used to reduce the number of unnecessary biopsies.

The large variation in improvement among the radiologists may have been due to the different degrees of confidence that they had in the computer aid. As with any new diagnostic tool, this confidence is influenced by the experience the radiologist has with the tool. Although the radiologists received training before the reading sessions, the high variability in confidence was not unexpected, because this ROC study was the first instance in which they had worked with the computer aid. Their confidence levels may have also been reflected in the relatively low accuracy of classification by some radiologists with CAD compared with that of the computer classifier alone.

If a radiologist can increase his or her

confidence in the performance of a computer aid by gaining more extensive clinical experience, then he or she will likely be able to find the most effective way of merging his or her judgment with the computer's rating and thus reduce both interobserver and intraobserver variability. Because a radiologist who uses CAD can establish a meaningful decision threshold for biopsy only after becoming familiar with the sensitivity and specificity of working with CAD, the radiologists in this study were not asked to decide whether biopsy should have been performed on a mass. Rather, we focused on the evaluation of changes in the sensitivity and specificity of the radiologists' classification of masses when CAD was used.

In this ROC study, all six observers were attending radiologists with extensive experience in the interpretation of mammograms. It is possible that the computer aid may be even more useful to radiology residents or radiologists with less experience in mammography. The effect of CAD on mammographic interpretation by less-experienced readers will be a subject of investigation in future studies.

The observers were allowed unlimited time to read each case in this ROC study. To obtain an estimate of the change in reading time with CAD, we recorded the reading time of each observer in each reading session by using a stopwatch. For the single-view reading experiment, the average reading time per image without CAD varied from 4.3 seconds to 17.1 seconds (mean time for the six observers, 7.8 seconds). The average reading time per image with CAD varied from 4.2 seconds to 17.3 seconds (mean time, 7.3 seconds). For the two-view reading experiment, the average reading time per pair of images without CAD varied from 6.6 seconds to 16.0 seconds (mean time, 10.4 seconds). The average reading time per pair of images with CAD varied from 7.6 seconds to 27.1 seconds (mean time, 13.5 seconds).

The reading time essentially did not change with use of the computer aid for the single-view readings. For the two-view readings, the radiologists took longer with CAD, probably because they had to merge the two computer ratings and merge the computer ratings with their own evaluations. Further investigation is needed to determine whether there is a trade-off between the radiologist's efficiency and the method of presenting the computer rating and whether the reading time with CAD will depend on the experience that the radiologist has with the computer information.

In the observer study, we used laser-printed mammograms instead of the original mammograms for the reading experiments. A major reason is that it is difficult to keep all the original mammograms together for the entire period of the study because they are part of active patient files and thus often recalled for comparison with new studies or for other clinical reasons. Because the maximum optical density of laser-printed images was 3.1 for the laser imager used, the contrast on the printed mammograms was about 20% lower than that on the original mammograms. Although the image quality was slightly lower than that of the original, the laser-printed digitized images were judged to be adequate for reading the details of the masses by the participating radiologists. The laser-printed image set might also be considered as one that had slightly more subtle masses than the original set of images. Because the relative performance of two modalities is measured in ROC experiments, and because the readings both with and without CAD in this study were conducted with the same set of printed images, the relative performance of the two readings should be valid. It should also be noted that in order for a computer aid that uses automated image analysis to be widely accepted, direct digital mammography would have to be the imaging modality in clinical use. Laser-printed images or soft-copy monitors will be the display medium for the digital mammograms. The use of laser-printed images for this ROC study was therefore practical.

In our observer performance experiment, we found that CAD improved the radiologists' ability to distinguish malignant and benign masses. This is consistent with the results of other studies (11,14) in which a statistically significant improvement ($P < .001$ in both studies) in the radiologists' classification accuracy by using CAD was found. The results of the former study (11) further showed that the PPV of a recommendation for biopsy by the radiologists was significantly increased ($P < .001$). In our approach, the computer classifier automatically extracted image features, whereas in the other studies, the computer classifier used the radiologist's evaluation and other patient information as input. Therefore, it appears that CAD can provide a useful second opinion to radiologists, either by consistently extracting and analyzing the image features or by optimally weighting various diagnostic factors and thereby improving the consistency in the decision-making process. This suggests that a computer classifier that combines both approaches—that is, automatically extracts image features and optimally merges them with the radiologist's evaluation and patient information—may be even more effective for breast cancer diagnosis. The latter step will also improve the radiologist's utilization of the computer rating on the basis of the computer-extracted features; this utilization was found to have large interobserver variation in our ROC experiment.

In conclusion, an ROC study of the effects of CAD on radiologists' classification of malignant and benign masses on mammograms was conducted. The results showed that CAD can provide a statistically significant improvement in the classification accuracy—that is, in the A_z value—for both single-view reading ($P = .022$) and two-view reading ($P = .007$). The improved separation between the confidence ratings of the malignant masses and those of the benign masses indicates the potential that CAD may reduce the rate of biopsy of benign masses when decision thresholds are properly chosen by the radiologists. The decision threshold may vary among radiologists, as in the case of mammographic interpretation without CAD, and can be set after the radiologist working with CAD has established his or her sensitivity and specificity with this approach through clinical experience.

Further studies are needed to evaluate the effects of CAD on the accuracy of radiologist classification of masses in clinical settings in which the prevalence of malignant masses is different from that in a laboratory data set and the likelihood of malignancy of a mass can be estimated by the computer classifier. In the two-view reading ROC experiment, the reading time per case increased by about 30% with the use of CAD. The dependence of the radiologist's efficiency in reading with CAD on the presentation method and on the reader's experience in using the computer information also warrants further investigation.

Acknowledgments: The authors are grateful to Charles E. Metz, PhD for useful discussions and for the use of the LABROC and LABMRMC programs.

References

1. Landis SH, Murray T, Bolden S, Wingo PA. Cancer statistics 1998. CA Cancer J Clin 1998; 48:6–29.
2. Adler DD, Helvie MA. Mammographic biopsy recommendations. Curr Opin Radiol 1992; 4:123–129.

3. Kopans DB. The positive predictive value of mammography. AJR 1991; 158:521–526.
4. Shtern F. Digital mammography and related technologies: a perspective from the National Cancer Institute. Radiology 1992; 183: 629–630.
5. Thurfjell EL, Lernevall KA, Taube AAS. Benefit of independent double reading in a population-based mammography screening program. Radiology 1994; 191:241–244.
6. Vyborny CJ. Can computers help radiologists read mammograms? Radiology 1994; 191:315–317.
7. Chan HP, Doi K, Vyborny CJ, et al. Improvement in radiologists' detection of clustered microcalcifications on mammograms: the potential of computer-aided diagnosis. Invest Radiol 1990; 25:1102–1110.
8. Kegelmeyer WP, Pruneda JM, Bourland PD, Hillis A, Riggs MW, Nipper ML. Computer-aided mammographic screening for spiculated lesions. Radiology 1994; 191: 331–337.
9. Getty DJ, Pickett RM, D'Orsi CJ, Swets JA. Enhanced interpretation of diagnostic images. Invest Radiol 1988; 23:240–252.
10. D'Orsi CJ, Getty DJ, 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.
11. Baker JA, Kornguth PJ, Lo JY, Floyd CE. Artificial neural network: improving the quality of breast biopsy recommendations. Radiology 1996; 198:131–135.
12. Chan HP, Sahiner B, Petrick N, et al. Observer performance study of radiologists' reading of mammographic masses and comparison with computerized classification (abstr). Radiology 1996; 201(P):370.
13. Chan HP, Sahiner B, Helvie MA, et al. Effects of computer-aided diagnosis (CAD) on radiologists' classification of malignant and benign masses on mammograms: an ROC study (abstr). Radiology 1997; 205(P):275.
14. Jiang Y, Nishikawa R, Schmidt RA, Metz CE, Doi K. Improving breast cancer diagnosis with computer-aided diagnosis (CAD): an observer study (abstr). Radiology 1997; 205(P):274.
15. Sahiner B, Chan HP, Petrick N, Helvie MA, Adler DD, Goodsitt MM. Classification of masses on mammograms using rubberband straightening transform and feature analysis. Proc SPIE 1996; 2710:44–50.
16. Sahiner B, Chan HP, Petrick N, Helvie MA, Goodsitt MM. Computerized characterization of masses on mammograms: the rubber-band straightening transform and texture analysis. Med Phys 1998; 25:516–526.
17. Sahiner B, Chan HP, Petrick N, Helvie MA, Goodsitt MM. Design of a high-sensitivity classifier based on a genetic algorithm: application to computer-aided diagnosis. Phys Med Biol 1998; 43:2853–2871.
18. Ackerman LV, Gose EE. Breast lesion classification by computer and xeroradiograph. Cancer 1972; 30:1025–1035.
19. Kilday J, Palmieri F, Fox MD. Classifying mammographic lesions using computerized image analysis. IEEE Trans Med Imaging 1993; 12:664–669.
20. Pohlman S, Powell KA, Obuchowshi NA, Chilote WA, Grundfest-Broniatowski S. Quantitative classification of breast tumors in digitized mammograms. Med Phys 1996; 23:1337–1345.
21. Huo Z, Giger ML, Vyborny CJ, Wolverton DE, Schmidt RA, Doi K. Automated computerized classification of malignant and benign masses on digitized mammograms. Acad Radiol 1998; 5:155–168.
22. Haralick RM, Shanmugam K, Dinstein I. Texture features for image classification. IEEE Trans Syst Man Cybernetics 1973; 3:610–621.
23. Norusis MJ. SPSS for Windows release 6: professional statistics. Chicago, Ill: Statistical Product for Service Solutions, 1993.
24. Lachenbruch PA. Discriminant analysis. New York, NY: Hafner, 1975; 8–19.
25. Metz CE. ROC methodology in radiologic imaging. Invest Radiol 1986; 21:720–733.
26. Metz CE, Herman BA, Shen JH. Maximum-likelihood estimation of receiver operating characteristic (ROC) curves from continuously distributed data. Stat Med 1998; 17:1033–1053.
27. Metz CE. Some practical issues of experimental design and data analysis in radiological ROC studies. Invest Radiol 1989; 24:234–245.
28. Dorfman DD, Berbaum KS, Metz CE. ROC rating analysis: generalization to the population of readers and cases with the jackknife method. Invest Radiol 1992; 27:723–731.
29. Fukunaga K, Hayes RR. Effects of sample size on classifier design. IEEE Trans Pattern Analysis and Machine Intelligence 1989; 11:873–885.
30. Chan HP, Sahiner B, Wagner RF, Petrick N, Mossoba J. Effects of sample size on classifier design: quadratic and neural network classifiers. Proc SPIE 1997; 3034:1102–1113.
31. Chan HP, Sahiner B, Wagner RF, Petrick N. Classifier design for computer-aided diagnosis in mammography: effects of finite sample size. Med Phys 1997; 24:1034–1035.

Classification of Malignant and Benign Masses Based on Hybrid ART2LDA Approach

Lubomir Hadjiiski,* *Member, IEEE*, Berkman Sahiner, *Member, IEEE*, Heang-Ping Chan, Nicholas Petrick, *Member, IEEE*, and Mark Helvie

***Abstract*—A new type of classifier combining an unsupervised and a supervised model was designed and applied to classification of malignant and benign masses on mammograms. The unsupervised model was based on an adaptive resonance theory (ART2) network which clustered the masses into a number of separate classes. The classes were divided into two types: one containing only malignant masses and the other containing a mix of malignant and benign masses. The masses from the malignant classes were classified by ART2. The masses from the mixed classes were input to a supervised linear discriminant classifier (LDA). In this way, some malignant masses were separated and classified by ART2 and the less distinguishable benign and malignant masses were classified by LDA. For the evaluation of classifier performance, 348 regions of interest (ROI's) containing biopsy proven masses (169 benign and 179 malignant) were used. Ten different partitions of training and test groups were randomly generated using an average of 73% of ROI's for training and 27% for testing. Classifier design, including feature selection and weight optimization, was performed with the training group. The test group was kept independent of the training group. The performance of the hybrid classifier was compared to that of an LDA classifier alone and a backpropagation neural network (BPN). Receiver operating characteristics (ROC) analysis was used to evaluate the accuracy of the classifiers. The average area under the ROC curve (A_z) for the hybrid classifier was 0.81 as compared to 0.78 for the LDA and 0.80 for the BPN. The partial areas above a true positive fraction of 0.9 were 0.34, 0.27 and 0.31 for the hybrid, the LDA and the BPN classifier, respectively. These results indicate that the hybrid classifier is a promising approach for improving the accuracy of classification in CAD applications.**

***Index Terms*— Computer-aided diagnosis, hybrid classifier, mammography, neural networks.**

I. Introduction

MAMMOGRAPHY is the most effective method for detection of early breast cancer [1]. However, the specificity for classification of malignant and benign lesions from mammographic images is relatively low. Clinical studies have shown that the positive predictive value (i.e., ratio of the number of breast cancers found to the total number of biopsies) is only 15% to 30% [2]–[4]. It is important to increase the positive predictive value without reducing the sensitivity of breast cancer detection. Computer-aided diagnosis (CAD) has the potential to increase the diagnostic accuracy by reducing the false-negative rate while increasing the positive predictive values of mammographic abnormalities.

Classifier design is an important step in the development of a CAD system. A classifier has to be able to merge the available input feature information and make a correct evaluation. Commonly used classifiers for CAD include linear discriminants (LDA) [5], [6] and backpropagation neural networks (BPN) [7]–[9] which have been shown to perform well in lesion classification problems [10]–[22]. These classifiers are generally designed by supervised training. However, these types of classifiers have limitations dealing with the nonlinearities in the data (in case of LDA) and in generalizability when a limited number of training samples are available (especially BPN). Another classification approach is based on unsupervised classifiers, which cluster the data into different classes based on the similarities in the properties of the input feature vectors. Therefore, unsupervised classifiers can be used to analyze the similarities within the data. However, it is difficult to use them as a discriminatory classifier [29], [30]. They also have limited generalizability when the training sample set is small.

We propose here a hybrid unsupervised/supervised structure to improve classification performance. The design of this structure was inspired by neural information processing principles such as self organization, decentralization and generalization. It combines the adaptive resonance theory network (ART2) [26], [27] and the LDA classifier as a cascade system (ART2LDA). The self-organizing unsupervised ART2 network automatically decomposes the input samples into classes with different properties. The ART2 network has been found to perform better compared to conventional clustering techniques in terms of learning speed and discriminatory resolution for the detection of rare events in many classification tasks [28]–[30]. The supervised LDA then classifies the samples belonging to a subset of classes that have greater similarities. By improving the homogeneity of the samples, the classifier designed for the subset of classes may be more robust.

The ART2LDA design implements both structural and data decomposition. Decomposition is a powerful approach that can reduce the complexity of a problem. Both structural decom-

Manuscript received January 27, 1999; revised October 26, 1999. This work was supported by in part by the USPHS under Grant No. CA 48129 and in part by the U.S. Army Medical Research and Materiel Command (USAMRMC) under Grant DAMD 17-96-1-6254. The work of L. Hadjiiski was supported in part by the USAMRMC under Career Development Award DAMD 17-98-1-8211. The work of B. Sahiner was supported in part by the USAMRMC under Career Development Award DAMD 17-96-1-6012. The work of Nicholas Petrick was supported in part by a grant from The Whitaker Foundation. The Associate Editor responsible for coordinating the review of this paper and recommending its publication was N. Karssemeijer. *Asterisk indicates corresponding author.*

*L. Hadjiiski, B. Sahiner, H.-P. Chan, N. Petrick, and M. Helvie are with the Department of Radiology, The University of Michigan, Ann Arbor, MI 48109-0904 USA.

Publisher Item Identifier S 0278-0062(99)10410-5.

0278–0062/99$10.00 © 1999 IEEE

position and data decomposition can improve classification accuracy [23] as well as model accuracy [24]. However, decomposition can also reduce the prediction accuracy due to overfitting the training data. We will demonstrate in this paper that the proposed hybrid structure can reduce the overfitting problem and improve the prediction capabilities of the system. The performance of the hybrid ART2LDA classifier will be compared with those of an LDA alone or a BPN classifier.

The rest of the paper is organized as follows. In Section II the ART2 unsupervised network is described. A hybrid ART2LDA classifier is introduced in Section III. Section IV describes the data set used in this study. The results are presented in Section V. Section VI contains discussion of these results. Finally, Section VII concludes this investigation.

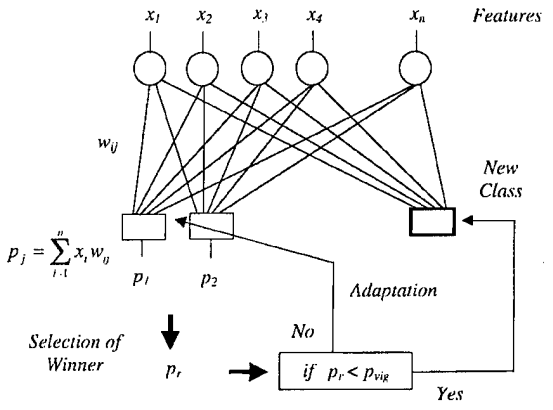


Fig. 1. Structure of the ART2 network.

II. ART2 Unsupervised Neural Network

The ART2 is a self-organizing system that can simulate human pattern recognition. ART2 was first described by Grossberg [25] and a series of further improvements were carried out by Carpenter, Grossberg, and coworkers [26]–[28]. The ART2 network clusters the data into different classes based on the properties of the input feature vectors. The members within a class have similar properties. The process of ART2 network learning is a balance between the plasticity and stability dilemma. Plasticity is the ability of the system to discover and remember important new feature patterns. Stability is the ability of the system to remain unchanged when already known feature patterns with noise are input to the system. The balance between plasticity and stability for the ART2 training algorithm allows fast learning [28], i.e., rare events can be memorized with a small number of training iterations without forgetting previous events. The more conventional training algorithms, such as back propagation [7]–[9], perform slow learning, i.e., they tend to average over occurrences of similar events and require many training iterations.

The structure of the ART2 system is shown in Fig. 1. It consists of two parts: the ART2 network and the learning stage. Suppose that there are n input features x_i $(i = 1, \cdots, n)$ and k classes in the ART2 network. When a new vector is presented to the input of the ART2 network, an activation value p_j for class j is calculated as

$$p_j = \sum_{i=1}^{n} x_i w_{ij}, \qquad j = 1, \cdots, k \tag{1}$$

where w_{ij} is the connection weight between input i and class j. The activation value is a measure of the membership of the particular input feature vector to class j. The higher the value p_j is, the better the input vector matches class j. The maximum value p_r is selected from all p_j $(j = 1, \cdots, k)$ to find the best class match. Furthermore, in order to balance the contribution to the activation value from all feature components, the input feature values applied to the ART2 system are scaled between zero and one [30]. This normalization will allow detection of similar feature patterns even when the magnitudes of the input feature components are very different.

The learning stage of the ART2 system can influence the weights of the selected class or the complete ART2 network structure by adding a new class. An additional parameter, the vigilance, is used to determine the type of learning [26]. The vigilance parameter p_{vig} is a threshold value that is compared to the maximum activation value p_r. If p_r is larger than p_{vig} then the input vector is considered to belong to class r. The adaptation of the weights connected with class r is performed as follows:

$$w_{ir}^{new} = w_{ir}^{\text{old}} + \eta(x_i - w_{ir}^{\text{old}}), \qquad \text{for } i = 1, \cdots, n \tag{2}$$

where η is a learning rate. The adaptation of the class r weights (2), aims at maximization of the p_r value for the particular input vector. In an iterative manner the weights are adjusted so that the activation values produced for similar input vectors will be maximum only for the class to which they belong and these maximum activation values will be higher than p_{vig}.

If the maximum activation value p_r is smaller than p_{vig}, it is an indication that a novelty has appeared and a new class will be added to the ART2 structure. The new weights connecting the input with the new class $(k + 1)$ are initialized with the scaled input feature values of this novelty. In such a way, the activation value p_{k+1} will be maximum $(p_r = p_{k+1})$ higher than p_{vig} when computed for this novelty in further training iterations. The value of the vigilance parameter p_{vig} determines the resolution of ART2. It can be chosen in the range between zero and one. In the case that p_{vig} is relatively small, only very different input feature vectors will be distinguished and separated in different classes. If p_{vig} is relatively large, the input feature vectors that are more similar will be separated into different classes. The value of p_{vig} is selected differently depending on the particular application.

III. ART2LDA Classifier

Despite the good performance of ART2 for efficient clustering and detection of novelties, the fast learning approach can cause problems associated with the generalization capability of the system and the correct classification of unknown cases. Supervised classifiers such as linear discriminants or backpropagation neural network classifiers can have better generalization capability than ART2, because they are trained by averaging over similar event occurrences. However, the learning process in these traditional learning algorithms tends

to erase the memory of previous expert knowledge when a new type of expertise is being learned. Therefore, these classifiers do not have as good an ability to correctly classify rare events as ART2 [28], [29].

In order to improve the accuracy and generalization of a classifier, we propose to design a hybrid classifier that combines the unsupervised ART2 network and a supervised LDA classifier. This hybrid classifier (ART2LDA) utilizes the good resolution capability of ART2 and the good generalization capability of LDA. The ART2 first analyzes the similarity of the sample population and identifies a subpopulation that may be separated from the main population. This will improve the performance of the second-stage LDA if the subpopulation causes the sample population to deviate from multivariate normal distributions for which LDA is an optimal classifier. Therefore, the ART2 serves as a screening tool to improve the homogeneity of the sample distributions by classifying outlying samples into separate classes.

The ART2LDA hybrid classifier can be described as

$$y_{AL} = g(f_2(x))f_1(x) + 1 - g(f_2(x)) \tag{3}$$

where x is the input vector, $f_1(\cdot)$ is the LDA classifier, $f_2(\cdot)$ is the ART2 classifier, and $g(\cdot)$ is a binary membership function, which labels the classes identified by ART2 to be one of the two types: malignant class or mixed class. A particular class is defined as malignant if it contains only malignant members. It is defined as mixed if it contains both malignant and benign members. The membership function is defined as follows:

$$g(c) = \begin{cases} 0, & \text{if } c \text{ is a malignant class} \\ 1, & \text{if } c \text{ is a mixed class.} \end{cases} \tag{4}$$

The type of a given class is determined based on ART2 classification of the training data set.

The structure of the ART2LDA classifier is shown in Fig. 2. The ART2 classifies the input sample x into either a malignant or a mixed class. Depending on the class type the function $g(\cdot)$ determines whether the LDA classifier will be used. If x is classified into a mixed class, the final classification will be obtained based on the LDA classifier. However, if x is classified by ART2 into a malignant class, then the mass will be considered malignant, without using the LDA classifier. Therefore, in the ART2LDA structure, the ART2 is used both as a classifier and a supervisor. This can be seen in (3). The first term in (3), $g(f_2(x))f_1(x)$, is the LDA classifier multiplied by the ART2 control part $g(f_2(x))$. The second term in (3), $(1 - g(f_2(x)))$, gives the classification result of the ART2 stage. If $f_2(x)$ is a malignant class, then $g(f_2(x)) = 0$, the LDA stage is eliminated, and the classifier output y_{AL} is equal to 1. On the other hand, if $f_2(x)$ is a mixed class, then $g(f_2(x)) = 1$, the ART2 term is eliminated, and the final classification is determined by the LDA classifier $(y_{AL} = f_1(x))$.

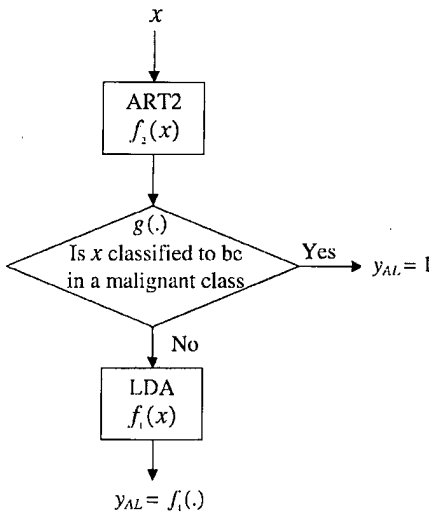


Fig. 2. Structure of the ART2LDA classifier.

IV. Methods

A. Data Set

The mammograms used in this study were randomly selected from the files of patients who had undergone biopsies at the University of Michigan. The criterion for inclusion of a mammogram in the data set was that the mammogram contained a biopsy-proven mass. The data set contained 348 mammograms with a mixture of benign ($n = 169$) and malignant ($n = 179$) masses. On each mammogram, a region of interest (ROI) containing the mass was identified by a radiologist experienced in breast imaging. The visibility of the masses was rated by the radiologist on a scale of 1 to 10, where the rating of 1 corresponds to the most visible category. The distributions of the visibility rating for both the malignant and benign masses are shown in Fig. 3. The visibility ranged from subtle to obvious for both types of masses. It can be observed that the benign masses tend to be more obvious than the malignant ones. Additionally the likelihood of malignancy for each mass was estimated based on its mammographic appearance. The radiologist rated the likelihood of malignancy on a scale of 1 to 10, where 1 indicated a mass with the most benign appearance. The distribution of the malignancy rating of the masses is shown in Fig. 4.

The data set can be considered as representative of the patient population that is sent for biopsy under current clinical criteria. Some characteristics of many malignant and benign masses can be visually distinguished by radiologists. However, there is also a nonnegligible fraction of malignant masses that are very similar to benign masses (the low malignancy rating region in Fig. 4). The estimated likelihood of malignancy of malignant and benign masses that are sent for biopsy basically overlaps over the entire range. This is consistent with the fact that in order not to miss malignant masses radiologists must recommend biopsy for even very low suspicion lesions.

Three hundred and five of the mammograms were digitized with a LUMISYS DIS-1000 laser scanner at a pixel resolution of 100 μm $\times$ 100 μm and 4096 gray levels. The digitizer was calibrated so that gray level values were linearly and inversely proportional to the optical density (OD) within the range of 0.1 to 2.8 OD units, with a slope of -0.001 OD/pixel value. Outside this range, the slope of the calibration curve decreased gradually. The OD range of the digitizer was 0

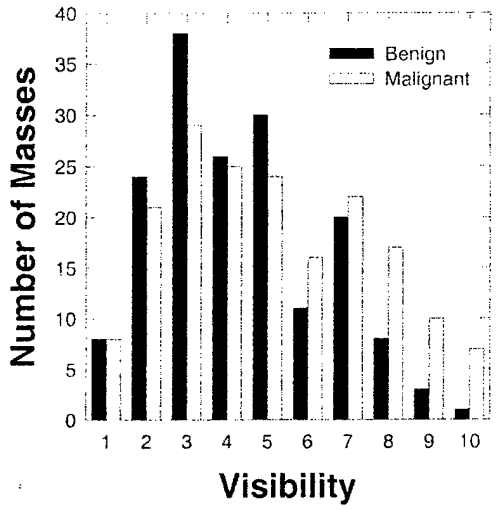


Fig. 3. The distribution of the visibility ranking of the masses in the dataset. The ranking was performed by an experienced breast radiologist (1: very obvious, 10: very subtle).

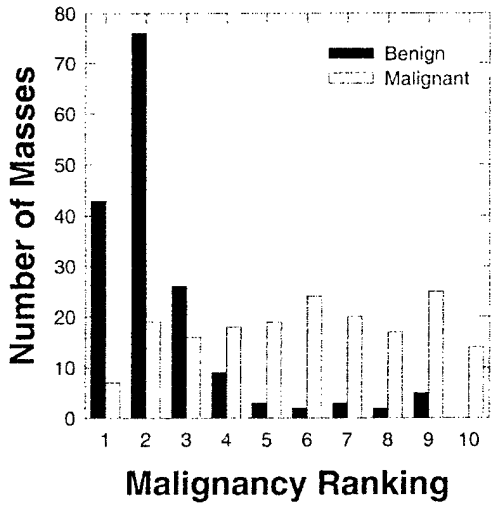


Fig. 4. The distribution of the malignancy ranking of the masses in the dataset. The ranking was performed by an experienced breast radiologist (1: very likely benign, 10: very likely malignant).

to 3.5. The remaining 43 mammograms were digitized with a LUMISCAN 85 laser scanner at a pixel resolution of 50 μm $\times$ 50 μm and 4096 gray levels. The digitizer was calibrated so that gray level values were linearly and inversely proportional to the OD within the range of 0 to 4 OD units, with a slope of -0.001 OD/pixel value. In order to process the mammograms digitized with these two different digitizers, the images digitized with LUMISCAN 85 digitizer were averaged with a 2 $\times$ 2 box filter and subsampled by a factor of two, resulting in 100 μm images.

In order to validate the prediction abilities of the classifier, the data set was partitioned randomly into training and test subsets on a 3:1 ratio, under the constraints that both the malignant and the benign samples were split with the 3:1 ratio and that the images from the same patient were grouped into the same (training or test) subset. These constraints caused the subsets to deviate from an exact 3:1 ratio. The data set was repartitioned randomly ten times. On average, 73% of the samples were grouped into the training set and 27% into the test set. The training and test results from the ten partitions were averaged to reduce their variability.

B. Feature Extraction

A rectangular ROI was defined to include the radiologist-identified mass with an additional surrounding breast tissue region of at least 40 pixels wide from any point of the mass border. A fully automated method was then used for segmentation of the mass from the breast tissue background within the ROI. The rubber band straightening transform (RBST) was previously developed [12] to map a band of pixels surrounding the mass onto the Cartesian plane (a rectangular region). In the transformed image, the border of mass appears approximately as a horizontal edge and spiculations appear approximately as vertical lines. The transformation of the radially oriented textures surrounding the mass margin to a more uniform orientation facilitates the extraction of texture features.

The texture features used in this study were calculated from spatial gray-level dependence (SGLD) matrices [10]–[12], [31], and run-length statistics (RLS) matrices [32] computed from the RBST images. The (i,j)th element of the SGLD matrix is the joint probability that gray levels i and j occur in a direction at a distance of θ pixels apart in an image. Based on our previous studies [10], a bit depth of eight was used in the SGLD matrix construction, i.e., the four least significant bits of the 12-bit pixel values were discarded. Thirteen texture measures, including correlation, energy, difference entropy, inverse difference moment, entropy, sum average, sum entropy, inertia, sum variance, difference average, difference variance, and two types of information measure of correlation were used. These measures were extracted from each SGLD matrix at ten different pixel pair distances ($d = 1, 2, 3, 4, 6, 8, 10, 12, 16$ and 20) and in four directions (0°, 45°, 90°, and 135°). Therefore, a total of 520 SGLD features were calculated for each image. The definitions of the texture measures are given in the literature [10]–[12], [31]. These features contain information about image characteristics such as homogeneity, contrast, and the complexity of the image.

RLS texture features were extracted from the vertical and horizontal gradient magnitude images, which were obtained by filtering the RBST image with horizontally or vertically oriented Sobel filters and computing the absolute gradient value of the filtered image. A gray level run is a set of consecutive, collinear pixels in a given direction which have the same gray level value. The run length is the number of pixels in a run [32]. The RLS matrix describes the run length statistics for each gray level in the image. The (i,j)th element of the RLS matrix is the number of times that the gray level i in the image possesses a run length of j in a given direction. In our previous study, it was found experimentally that a bit depth of five in the RLS matrix computation could provide good texture characteristics [12].

Five texture measures, namely, short run emphasis, long run emphasis, gray level nonuniformity, run length nonuniformity,

and run percentage were extracted from the vertical and horizontal gradient images in two directions, $\theta = 0°$ and $\theta = 90°$. Therefore, a total of 20 RLS features were calculated for each ROI. The formal definition of the RLS feature measures can be found in [32].

A total of 540 features (520 SGLD and 20 RLS) were therefore extracted from each ROI.

C. Feature Selection

In order to reduce the number of the features and to obtain the best feature set to design a good classifier, feature selection with stepwise linear discriminant analysis [33] was applied. At each step of the stepwise selection procedure one feature is entered or removed from the feature pool by analyzing its effect on the selection criterion. In this study, the Wilks' lambda (the ratio of within-group sum of squares to the total sum of squares [34]) was used as a selection criterion. The optimization procedure used a threshold F_{in} for feature entry and a threshold F_{out} for feature removal. On a feature entry step, the features not yet selected are entered into the selected feature pool one at a time, the significance of the change in the Wilks' lambda caused by this feature is estimated based on F statistics. The feature with the highest significance is entered into the feature pool if its significance is higher than F_{in}. On a feature removal step, the features which have already been selected are analyzed one at a time from the selected feature pool and the significance of the change in the Wilks' lambda is estimated. The feature with the least significance is removed from the selected feature pool if the significance is less than F_{out}. Since the appropriate values of F_{in} and F_{out} are not known *a priori*, we examined a range of F_{in} and F_{out} values and chose the appropriate thresholds in such a way that a minimum number of features were selected to achieve a high accuracy of classification by LDA for the training sets. More details about the stepwise linear discriminant analysis and its application to CAD can be found in [10]–[12].

D. Performance Analysis

To evaluate the classifier performance, the training and test discriminant scores were analyzed using receiver operating characteristic (ROC) methodology [35]. The discriminant scores of the malignant and benign masses were used as decision variables in the LABROC1 program [36], which fit a binormal ROC curve based on maximum likelihood estimation. The classification accuracy was evaluated as the area under the ROC curve, A_z. For the ART2LDA classifier, the discriminant scores of all case samples classified in the two stages are combined. All masses classified into the malignant group by the ART2 stage were assigned a constant positive discriminant score higher than or equal to the most malignant discriminant score obtained from the LDA stage .

The performance of ART2LDA was also assessed by estimation of the partial area index ($A_z^{(0.9)}$) and compared with the corresponding performance index of the LDA and BPN classifiers. The partial area index ($A_z^{(0.9)}$) is defined as the area that lies under the ROC curve but above a sensitivity threshold of 0.9 ($\text{TPF}_0 = 0.9$) normalized to the total area above TPF_0, (1-TPF_0). The partial $A_z^{(0.9)}$ indicates the performance of the classifier in the high-sensitivity (low false negative) region which is most important for clinical cancer detection task. In addition, the performance of the LDA stage of the ART2LDA classifier was evaluated by the estimation of the area under the ROC curve, denoted as A_z (LDA), for the case samples passed onto the LDA classifier.

TABLE I
NUMBER OF SELECTED FEATURES FOR THE TEN DATA GROUPS WITH THE CORRESPONDING F_{IN} AND F_{OUT} PARAMETERS

Data Group No.	Number of selected features	F_{in}	F_{out}
1	12	1.8	1.6
2	15	2.4	2.2
3	13	2.4	2.2
4	18	2.4	2.2
5	14	2.4	2.2
6	14	2.1	1.8
7	13	2.4	2.2
8	18	1.8	1.6
9	14	2.4	2.2
10	14	2.4	2.2

V. RESULTS

In this section the ART2LDA classification results for malignant and benign masses will be presented and compared with those of the LDA or BPN classifiers. The important point in this study is the fact that the test subset is truly independent of the training subset. Only the training subset is used for feature selection and classifier training, and only the test subset is used for classifier validation. In order to validate the prediction abilities of the classifier, ten different partitions of the training and test sets were used. A different ART2LDA classifier was trained using each training set and the corresponding set of selected features. The classification result was estimated as the average performance for the ten partitions.

For a given partition of training and test sets, feature selection was performed based on the training set alone. The feature selection results for the ten different training groups are shown in Table I. The average number of selected features was 14. An average of two RLS features and twelve SGLD features were selected for each of the training sets which represented 10% of all RLS features and 2.3% of all SGLD features, respectively. Both types of features (RLS and SGLD) are necessary in order to obtain good classification. The most often selected RLS features for the ten training sets were: horizontal short run emphasis (four times), horizontal long run emphasis (six times), vertical run length nonuniformity (three times), horizontal run length nonuniformity (three times). The most often selected SGLD texture measures for the ten training sets were: inverse difference moment (eight times), information measure of correlations one and two (19 times), difference average (nine times), and correlation (ten times). For a given texture measure, features at different angles or distances may be selected, but these features are usually highly correlated so

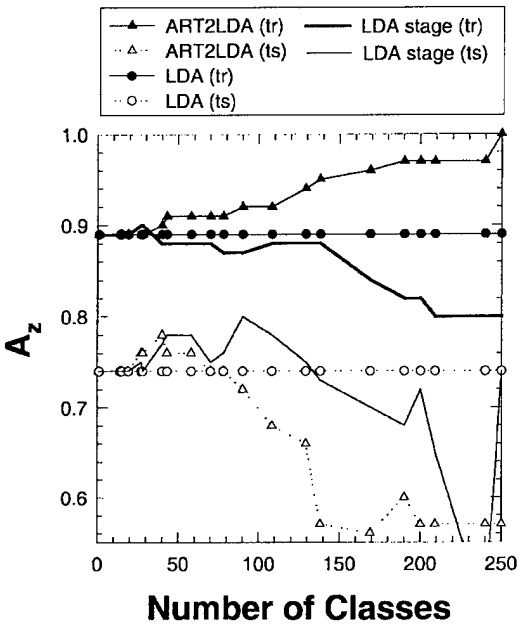


Fig. 5. ART2LDA and LDA classification results for training and test sets from data group three as a function of the generated number of classes. Additionally the results for the LDA stage from the ART2LDA classifier are plotted.

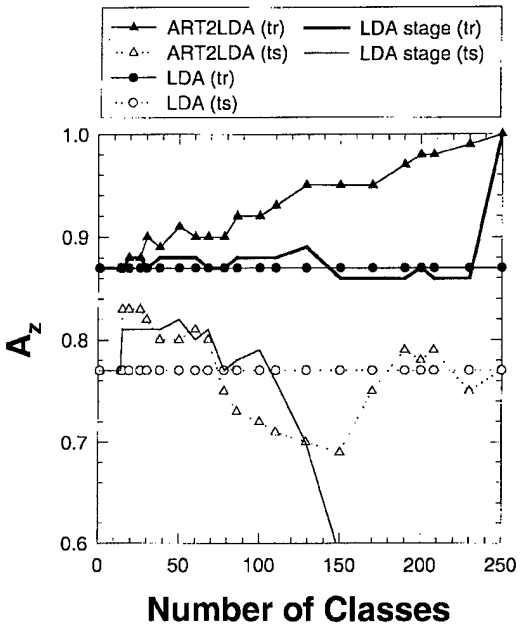


Fig. 6. ART2LDA and LDA classification results for training and test sets from data group one as a function of the generated number of classes. Additionally the results for the LDA stage from the ART2LDA classifier are plotted.

that they can be considered to be similar and counted together as described above.

A. ART2LDA Classification Results

For the ART2LDA classifier, the number of selected features determines the dimensionality of the input vector of the ART2 classifier and the dimensionality of the LDA classifier. By applying different values for the vigilance parameter, ART2 classifiers with different number of classes were obtained. In this study, the vigilance parameter p_{vig} was varied from 0.9 to 0.99, resulting in a range of 10 to 240 classes. The overall performance of the ART2LDA classifier was evaluated for different numbers of ART2 classes because different subset of the samples were separated and classified by ART2 when p_{vig} was varied. In Fig. 5, the classification results for the ART2LDA are compared to the results from LDA alone for the training and test set partition three. The classification accuracy, A_z, was plotted as a function of the number of ART2 classes. For this training and test set partition, when the number of classes was between 20 and 60, the ART2LDA classifier improved the classification accuracy for the test set in comparison to LDA. As the number of classes increased to greater than 60, the A_z value increased for the training data set, but decreased for the test data set and was lower than that of the LDA alone. The two solid lines in Fig. 5 show the A_z values for the LDA stage in the ART2LDA classifier for both the training and test sets. It can be observed that the test A_z for the LDA stage is higher than the A_z for the LDA classifier alone, but not as high as A_z obtained by ART2LDA when the number of classes is small.

In Fig. 6 the classification results of LDA and ART2LDA for the partition one training and test sets are shown. In this case it appeared that in the test set there were two large malignant outliers which degraded the LDA performance. Only 15 classes at the ART2 stage in the ART2LDA was enough to cluster the outliers into a separate malignant class and to improve the performance of the LDA stage and the overall result. The rest of the outliers required more ART2 classes before they were clustered into separate classes and correctly classified as malignant. This is the reason for the similar behavior of the classifiers for partitions three and one in the range of 40 to 70 classes as seen in Figs. 5 and 6. When the number of classes was less than 70, the test A_z for the LDA stage (A_z(LDA)) was higher than the LDA alone, but not as high as the A_z for ART2LDA with less than 30 classes (Fig. 6). The best A_z values for the test data sets of the ten training and test partitions are presented in Table II and Fig. 7. The ART2LDA classifier achieved higher A_z values than the LDA alone in nine of the ten partitions. The average A_z is 0.81 for ART2LDA and 0.78 for LDA alone. The standard deviations of the A_z values for the ten groups range from 0.03 to 0.05 for the ART2LDA classifier and from 0.04 to 0.05 for the LDA classifier.

The performance of ART2LDA was also assessed by estimation of the partial area under the ROC curve $A_z^{(0.9)}$ at a TPF higher than 0.9. The results are presented in Table III and Fig. 7. In the lower part of Fig. 7, the $A_z^{(0.9)}$ values of the test set for the corresponding ten partitions of training and test sets are presented. The average test $A_z^{(0.9)}$ value is 0.34 for the ART2LDA and 0.27 for LDA. For nine of the ten partitions, the $A_z^{(0.9)}$ value was improved at the high-sensitivity operating region (TPF $>$ 0.9) of the ROC curve.

The classifier performance was also evaluated when the ART2LDA classifiers were designed using a fixed number

TABLE II
CLASSIFIERS PERFORMANCE FOR THE TEN TEST SETS. THE A_z VALUES REPRESENT THE TOTAL AREA UNDER ROC CURVE

Data Group No.	LDA	ART2LDA	BPN	ART2LDA(1)
1	0.77	0.83	0.85	0.80
2	0.78	0.80	0.82	0.77
3	0.74	0.78	0.77	0.78
4	0.77	0.77	0.75	0.77
5	0.77	0.78	0.76	0.77
6	0.80	0.83	0.82	0.81
7	0.80	0.81	0.82	0.77
8	0.77	0.80	0.74	0.75
9	0.77	0.80	0.81	0.80
10	0.86	0.89	0.84	0.89
Mean	0.78	0.81	0.80	0.79

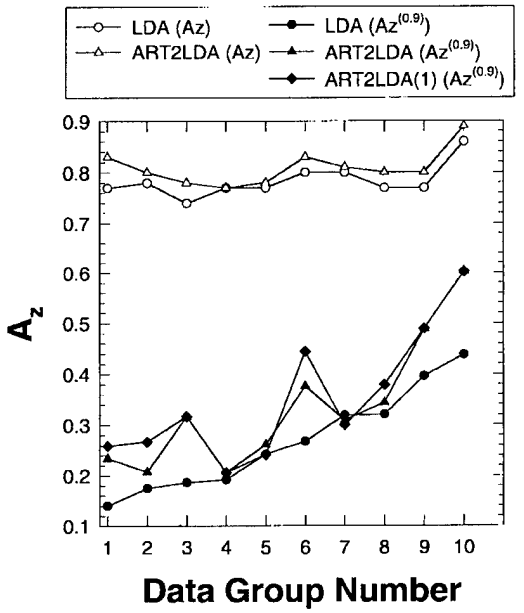


Fig. 7. Average A_z classification results for the 10 test sets. The top graphs represent the ART2LDA and LDA A_z values for the total area under the ROC curve. The bottom graphs represent the ART2LDA, ART2LDA(1) and LDA A_z values for the partial area of the ROC curve above the true positive fraction of 0.9.

TABLE III
CLASSIFIERS RESULTS FOR THE TEN TEST SETS. THE A_z VALUES REPRESENT THE PARTIAL AREA OF THE ROC CURVE ABOVE THE TRUE POSITIVE FRACTION OF 0.9 ($A_z^{(0.9)}$)

Data Group No.	LDA	ART2LDA	BPN	ART2LDA(1)
1	0.14	0.23	0.31	0.26
2	0.17	0.21	0.28	0.27
3	0.19	0.32	0.27	0.32
4	0.19	0.21	0.19	0.21
5	0.24	0.26	0.32	0.24
6	0.27	0.38	0.27	0.44
7	0.32	0.31	0.38	0.30
8	0.32	0.34	0.25	0.38
9	0.40	0.49	0.40	0.49
10	0.44	0.60	0.38	0.60
Mean	0.27	0.34	0.31	0.35

of ART2 classes. The A_z, and $A_z^{(0.9)}$ results, averaged over the ten test partitions, are presented in Table IV. The average A_z with the ART2LDA classifier, compared to that of LDA alone, was again improved between 15 and 40 classes. The maximum average A_z of 0.80 was achieved between 20 and 40 classes. The average $A_z^{(0.9)}$ results are improved for all

TABLE IV
AVERAGE A_z AND AVERAGE $A_z^{(0.9)}$ CLASSIFICATION RESULTS FOR THE TEN TEST SETS. CLASSIFIERS WERE DESIGNED USING A FIXED NUMBER OF ART2 CLASSES

	LDA	ART2LDA					
No. of classes		15	20	30	40	50	60
A_z	0.78	0.80	0.80	0.80	0.80	0.78	0.77
$A_z^{(0.9)}$	0.27	0.30	0.31	0.33	0.33	0.31	0.31

ART2LDA classifiers presented in Table IV. The maximum average $A_z^{(0.9)}$ value is 0.33 and it remains constant between 30 and 40 classes.

An alternative way to evaluate the performance of a classifier is its classification accuracy when a decision threshold for malignancy is selected based on the training set. For instance, a decision threshold may be selected such that all positive samples from the training set are classified correctly i.e., at a sensitivity of 100%. The ART2LDA with this decision threshold is referred to as ART2LDA(1). For a given training and test partitioning, ART2LDA classifiers with different number of classes in the ART2 stage were obtained (Figs. 5 and 6). For each of these models the decision threshold for a sensitivity of 100% was selected from the training set and the corresponding ART2LDA(1) classifier was obtained. Then the ART2LDA(1) classifier (with a specific number of classes in the ART2 stage) that correctly classified the maximum number of malignant masses in the test set is selected. By using all samples of the test set, the A_z value is calculated for the corresponding ART2LDA model. The A_z values for the ART2LDA(1) classifiers for the test sets of the ten data partitionings are shown in Tables II and III. For five of the partitions the overall A_z value for ART2LDA(1) is higher than that of LDA alone (Table II). The average A_z value was 0.79. The partial areas above the TP fraction of 0.9, $A_z^{(0.9)}$, for the ten test data sets obtained by the ART2LDA(1) classifier are also shown in Fig. 7. The ART2LDA(1) achieved the highest average $A_z^{(0.9)}$ value of 0.35 compared to ART2LDA and LDA (Table III).

B. BPN Classification Results

A multilayer perceptron back-propagation neural network with a single hidden layer and a single output node was used for comparison with the ART2LDA classifier. The number of selected features determined the number of input nodes to the BPN. The same ten training/test set partitions (as in the case of ART2LDA) were used for the training and validation of the BPN classifiers. BPN's with their number of hidden nodes ranging from two to ten were evaluated to obtain the best architecture. Back-propagation training was used. Each of the BPN's was trained for up to 18 000 training epochs. At every 1000 epochs the neural network weights were saved and the classification result for the corresponding test set was evaluated. This design procedure was repeated for each of the ten training/test groups. For each group, the best test result among all the BPN architectures (different number of hidden nodes) and all the training epochs examined was selected. The average test A_z over the ten groups for the BPN was 0.80, compared to 0.81 for ART2LDA (Table II). The standard deviations of the A_z values for the ten groups range from 0.04 to 0.05 for the BPN. The average partial $A_z^{(0.9)}$ for the BPN

was 0.31, compared to 0.34 for ART2LDA (Table III). The A_z and $A_z^{(0.9)}$ of the ART2LDA classifier were higher than those of the BPN in six of the ten training/test groups.

VI. Discussion

In the present study, a new classifier (ART2LDA) was designed and applied to the classification of malignant and benign masses. The results indicated that the ART2LDA classifier had better generalizability than an LDA classifier alone. The ART2 classifier grouped the case samples that were different from the main population into separate classes. The minimum number of classes needed to start the clustering of outliers into separate classes depended on how different the outliers were from the rest of the sample population. For the ten different partitions of training and test sets used in this study, the minimum number varied between 13 and 15 classes. When the number of ART2 classes was less than this minimum number of classes, the ART2 classifier generated only mixed malignant-benign classes and all samples were transferred to the LDA stage. In that case, the ART2LDA was equivalent to the LDA classifier alone. When a higher number of classes were generated, an increased number of cases that might be considered outliers of the general data population was removed (clustered in separate classes). For the ten training sets used in this study, the malignant outliers were gradually removed when the number of classes increased. The training accuracy increased when the number of classes increased and A_z could reach the value of 1.0. However, a large number of ART2 classes led to overfitting the training sample set and poor generalization in the test set. The classification accuracy of ART2 for the test set tended to decrease when the number of classes was greater than about 70. The large number of classes also led to a reduction in the generalizability of the second-stage LDA; the training of LDA with a small number of samples would again result in overfitting the training set, and poor generalizability in the test set. This effect was observed when more than 60 or 70 classes were generated by ART2 (see Figs. 5 and 6).

The classification accuracy of ART2LDA increased initially with an increased number of classes and then decreased after reaching a maximum. The correct classification of the outliers by the ART2 in combination with an improvement in the classification by the LDA resulted in the increased accuracy. When the number of ART2 classes was further increased, the effects of overfitting by the ART2 and the LDA became dominant and the prediction ability of the ART2LDA decreased. In some cases the second-stage LDA prediction was much worse than the ART2. In other cases the ART2 could not generalize well. The generation of a high number of classes is therefore impractical and unnecessary both from a computational and a methodological point of view.

For the optimal number of classes (usually less than 50 for the data sets used) the A_z value for the second-stage LDA in the ART2LDA was better than an LDA classifier alone, but it was not as good as the overall A_z from the ART2LDA. It is evident that the ART2 was a useful classifier for improvement of the second-stage classification.

When the partial area of the ROC curve above the true positive fraction (TPF) of 0.9 ($A_z^{(0.9)}$) was considered as a measure of classification accuracy, the advantage of ART2LDA over LDA alone became even more evident. By removing and correctly classifying the outliers, the accuracy of the classification was increased at the high sensitivity end of the curve.

The classifier performance was evaluated when the ART2LDA classifiers were designed using a fixed number of ART2 classes. The results showed improved performance of the ART2LDA in a range between 20 and 40 ART2 classes. Both the average A_z and the average $A_z^{(0.9)}$ reached a maximum within this region, and the maximum average A_z and the average $A_z^{(0.9)}$ values remained unchanged between 30 and 40 classes. These results indicated that the performance of a hybrid ART2LDA classifer was robust and stable and could be potentially useful in real clinical applications.

We have performed statistical tests with the CLABROC program to estimate the significance in the differences between the A_z values from the ART2LDA, the LDA alone, and the BPN, as well as in the differences in the partial $A_z^{(0.9)}$ from the three classifiers. The statistical tests were performed for each individual data set partition because the correlation among the data sets from the different partitions precludes the use of student's paired t test with the ten partitions. We found that the differences in both cases did not reach statistical significance because of the small number of test samples and thus the large standard deviation in the A_z values. However, the consistent improvements in A_z and $A_z^{(0.9)}$ by the ART2LDA (9 out of 10 data set partitions in both cases for LDA and six out of ten data set partitions in both cases for BPN) suggest that the improvement was not by chance alone, and that the accuracy of a classification task could be improved by the use of an ART2 network. In addition, one advantage of the ART2LDA is that the training process is more efficient than that of the BPN, especially when there is a subset of outlying samples. In such a case, the BPN will require a large number of training epochs to minimize the error function.

ART2LDA can be trained to classify the sample cases into more than two classes, such as a class of normal tissue regions in addition to malignant and benign masses. There will be an increase in the complexity of training and a larger training sample size will be desired, but these requirements will be comparable for the different classifiers. In a clinical situation, if the classification task is performed on all computer-detected lesions, the classifier has to distinguish the falsely detected normal tissue from malignant or benign lesions. However, it may be noted that a classifier that can distinguish only malignant and benign masses is applicable to the scenario that the radiologist identifies a suspicious lesion on the mammogram and would like to have a second opinion about its likelihood of malignancy before making a diagnostic decision. Therefore, the development of a classifier that can differentiate malignant and benign masses is the research of interest for many investigators.

Similarly, ART2 can be trained to discover and remove a pure benign mass class. The approach will be similar to the task of classifying and removing the pure malignant classes,

as described in this study. However, our approach of removing the malignant classes will reduce the chance of misclassification of malignant masses. In breast cancer detection, the cost of false-negative (missed cancer) is very high. Therefore, our goal in classifier design is to be conservative. By removing the malignant classes in the first stage, any misclassification to these classes will be regarded as malignant. The remaining classes will be classified again with the second-stage classifier so malignant masses will be less likely to be missed.

The problem of classification of malignant and benign masses has been studied by many investigators. Rangayyan *et al.* [15] used Mahalanobis distance classifer (a modification of an LDA classifier) and the leave-one-out method to evaluate the classification of 54 masses. Fogel *et al.* [16] compared LDA and BPN classifiers using the leave-one-out method and 139 masses (malignant and benign classification). Highnam *et al.* [17] used a morphological feature called a halo to classify 40 masses as malignant and benign. Huo *et al.* [22] employed BPN and a rule-based classifier to classify 95 masses using the leave-one-out evaluation method. Sahiner *et al.* [12] used an LDA classifier and the leave-one-out method to classify 168 masses. An important difference between the classifier designed in this study and the previous studies in the CAD field is the method of feature selection. In the above mentioned studies [12], [15]–[17], [22] and several other published studies [18]–[21] the features were selected from the entire data set first, and then the data set was partitioned into training and test sets. This meant that at the feature selection stage of the classifier design, the entire data set was used as a training set. Depending on the distribution of the features and the total number of samples used, the test results in these studies might be optimistically biased [37]. In our current study, the entire data set was initially partitioned into training and test sets and then feature selection was performed only on the training set. This method will result in a pessimistic estimate of the classifier performance when the training set is small [37]. However, it will provide a more conservative but realistic estimation of the classifier performance in the general patient population. We can expect that the performance would be improved if the classifier in this study were designed using a large data set. Since our main purpose in this study was to compare the ART2LDA classifier with the commonly used LDA and BPN, we did not attempt to quantify how pessimistic our results were in this study.

The most important contribution of this paper is to introduce a new approach that utilizes a two-stage unsupervised–supervised hybrid classifier. We believe that the hybrid approach will improve classification when the sample distribution contains subpopulations that may be difficult for a single classifier to classify. It will be useful for similar classification tasks although different classifiers may be used in each stage of the hybrid structure.

VII. Conclusion

A new classifier combining an unsupervised ART2 and a supervised LDA has been designed and applied to the classification of malignant and benign masses. A data set consisting of 348 films (179 malignant and 169 benign) was randomly partitioned into training and test subsets. Ten different random partitions were generated. For each training set, texture features were extracted and feature selection was performed. An average of features were selected for each group. A hybrid ART2LDA classifier, an LDA, and a BPN were trained by using each of the ten training sets. The A_z value under the ROC curve for the test sets, averaged over the ten partitions, was higher for ART2LDA ($A_z = 0.81$) compared to those of the LDA alone ($A_z = 0.78$) and of the BPN ($A_z = 0.80$). A greater improvement was obtained when the partial ROC area above a true-positive fraction of 0.9 was considered. The average partial A_z for ART2LDA was 0.34, as compared to 0.27 for LDA and 0.31 for BPN. Additionally, for the ART2LDA classifiers that correctly classified the maximum number of malignant masses in the test sets with decision threshold defined with the training set, the average partial A_z was 0.35. These results indicate that the hybrid classifier is a promising approach for improving the accuracy of classifiers for CAD applications.

Acknowledgment

The authors would like to thank Prof. S. Grosberg and Dr. G. Carpenter for providing them with valuable information as well as for the useful discussions. Additionally the authors would like to thank C. E. Metz, Ph.D., for providing the LABROC1 and CLABROC programs.

References

[1] H. C. Zuckerman, "The role of mammography in the diagnosis of breast canser," in *Breast Canser, Diagnosis and Treatment*, I. M. Ariel and J. B. Cleary, Eds. New York: McGraw-Hill, 1987, pp. 152–172.
[2] D. B. Kopans, "The positive predictive value of mammography," *Amer. J. Roentgenol.*, vol. 158, pp. 521–526, 1992.
[3] D. D. Adler and M. A. Helvie, "Mammographic biopsy recommendations," *Curr. Opin. Radiol.*, vol. 4, pp. 123–129, 1992.
[4] M. Moskowitz, "Impact of a priory medical detection on screening for breast cancer," *Radiology*, vol. 184, pp. 619–622, 1989.
[5] P. A. Lachenbruch, *Discriminant Analysis*. New York: Hafner, 1975.
[6] R. O. Duda, and P.E. Hart, *Pattern Classification and Scene Analysis*. New York: Wiley, 1973.
[7] P. J. Werbos, "Beyond regression: New tools for prediction and analysis in the behavioral sciences," Ph.D. dissertation, Harvard Univ., Cambridge, MA, 1974.
[8] D. Rumelhart, G. E. Hinton, and R. J. Williams, in D. E. Rumelhart, Ed., *Parallel and Distributed Processing*. Cambridge, MA: MIT Press, 1986, vol. 1, p. 318.
[9] J. Herz, A. Krogh, and R. Palmer, *Introduction to the Theory of Neural Computation*. Reading, MA: Addison-Wesley, 1991.
[10] H. P. Chan, D. Wei, M. A. Helvie, B. Sahiner, D. D. Adler, M. M. Goodsitt, and N. Petrick, "Computer-aided classification of mammographic masses and normal tissue: Linear discriminat analysis in texture feature space," *Phys. Med. Biol.*, vol. 40, pp. 857–876, 1995.
[11] D. Wei, H. P. Chan, M. A. Helvie, B. Sahiner, N. Petrick, D. D. Adler, and M. M. Goodsitt, "Classification of mass and normal breast tissue on digital mammograms: Multiresolution texture analysis," *Med. Phys.*, vol. 22, pp. 1501–1513, 1995.
[12] B. Sahiner, H. P. Chan, N. Petick, M. A. Helvie, and M. M. Goodsitt, "Computerized characterization of masses on mamograms: The rubber band sraightening transform and texture analysis," *Med. Phys.*, vol. 25, no. 4, pp. 516–526, Apr. 1998.
[13] B. Sahiner, H. P. Chan, D. Wei, N. Petick, M. A. Helvie, D. D. Adler, and M. M. Goodsitt, "Image feature selection by a genetic algorithm: Application to classification of mass and normal breast tissue," *Med. Phys.*, vol. 23, no. 10, pp. 1671–1683, Oct. 1996.
[14] H. P. Chan, B. Sahiner, N. Petrick, M. A. Helvie, K. L. Lam, D. D. Adler, and M. M. Goodsitt, "Computerized classification of malignant

and benign microcalsifications on mammograms: Texture analysis using an artificial neural network," *Phys. Med. Biol.*, vol. 42, pp. 549–567, 1997.

[15] R. M. Rangayyan, N. M. El-Farmawy, J. E. Desautels, and O. A. Alim, "Measures of acutance and shape for classification of breast tumors," *IEEE Trans. Med. Imag.*, vol. 16, pp. 799–810, Dec. 1997.

[16] D. B. Fogel, E. C. Wasson, E. M. Boughton, V. W. Porto, and P. J. "Angeline, linear and neural model for classifying breast masses," *IEEE Trans. Med. Imag.*, vol. 17, pp. 485–488, June 1998.

[17] R. P. Highnam, J. M. Brady, and B. J.Shepstone, "A quantitative feature to aid diagnosis in mammography," in *Proc. Digital Mammography'96*, pp. 201–206.

[18] Y. Wu, M. L. Giger, K. Doi, C. J. Vyborny, R. A. Schmidt, and C. E. Metz, "Artificial neural networks in mammography: Application to decision making in the diagnosis of breast cancer," *Radiology*, vol. 187, pp. 81–87, 1993.

[19] V. Goldberg, A. Manduca, D. L. Evert, J. J. Gisvold, and J. F. Greenleaf, "Improvements in specificity of ultrasonography for diagnosis of breast tumors by means of artificial intelligence," *Med. Phys.*, vol. 19, pp. 1475–1481, 1992.

[20] J. Kilday, F. Palmieri, and M. D. Fox, "Classifying mammographic lesions using computerized image analysis," *IEEE Trans. Med. Imag.*, vol. 12, pp. 664–669, Dec. 1993.

[21] M. F. McNitt-Gray, H. K. Huang, and J. W. Sayre, "Feature selection in the pattern classification problem of digital chest radiograph segmentation," *IEEE Trans. Med. Imag.*, vol. 14, pp. 537–547, Sept. 1995.

[22] Z. Huo, M. L. Giger, C. J. Vyborny, D. E. Wolverton, R. A. Schmidt, and K. Doi, "Automated computerized classification of malignant and benign masses on digitized mammograms," *Acad. Radiol.*, vol. 5, pp. 155–168, 1998.

[23] M. Jordan, and R. A. Jacobs, "Hierarchical mixture of experts and EM algorithm," *Neural Comput.*, vol. 6, pp. 181–214, 1994.

[24] L. Hadjiiski and P. Hopke, "Design of large scale models based on multiple neural network approach," *Intelligent Engineering Systems Through Artificial Neural Networks*. ASME, 1997, vol. 7, pp. 61–66.

[25] S. Grossberg, "Adaptive pattern classification and universal recoding. I. Parallel development and coding of neural feature detectors," *Biolog. Cybern.*, vol. 23, no. 3, pp. 121–134, 1976.

[26] G. A. Carpenter and S. Grossberg, "ART 2: Self-organization of stable category recognition codes for analog input patterns," *Appl. Opt.*, vol. 26, no. 23, 1, pp. 4919–4930, Dec. 1987.

[27] G. A. Carpenter, S. Grossberg, and D. B. Rosen, "ART 2-A: An adaptive resonance algorithm for rapid category learning and recognition," *Neural Networks*, vol. 4, no. 4, pp. 493–504, 1991.

[28] G. A. Carpenter and S. Grossberg, "Integrating symbolic and neural processing in a self-organizing architeture for pattern recognition and prediction," in *Artificial Intelligence and Neural Networks: Steps toward Principled Integration*. New York: Academic, 1994.

[29] G. A. Carpenter and N. Markuzon, "ARTMAP-IC and medical diagnosis: Instance counting and inconsistent cases," *Neural Networks*, vol. 11, no. 2, pp. 323–336, Mar. 1998.

[30] Y. Xie, P. K. Hopke, and D. Wienke, "Airborne particle classification with a combination of chemical composition and shape index utilizing an adaptive resonance artificial neural network," *Environ. Sci. Technol.*, vol. 28, no. 11, pp. 1921–1928, 1994.

[31] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Texture features for image classification," *IEEE Trans. Syst., Man, Cybern.*, vol. 3, pp. 610–621, Nov. 1973.

[32] M. M. Galloway, "Texture analysis using gray level run length," *Comput. Graph. Image Processing*, vol. 4, pp. 172–179, 1975.

[33] M. J. Norusis, *SPSS Professional Statistics 6.1*. Chicago, IL: SPSS, 1993.

[34] M. M. Tatsuoka, "Multivariate Analysis," *Techniques for Educational and Psychological Research*. New York: Macmillan, 1988.

[35] C. E. Metz, "ROC methodology in radiographic imaging," *Invest. Radiol.*, vol. 21, pp. 720–733, 1986.

[36] C. E. Metz, J. H. Shen, and B. A. Herman, "New methods for estimating a binomial ROC curve from continuously distributed test results," presented at the *1990 Annu. Meeting American Statistical Association*, Anahaim, CA, 1990.

[37] B. Sahiner, H. P. Chan, N. Petrick, R. Wagner, and L. Hadjiiski, "The effect of sample size on feature selection in computer-aided diagnosis," *Proc. SPIE*, vol. 3661, pp. 499–510, 1999.

A regional registration technique for automated interval change analysis of breast lesions on mammograms

S. Sanjay-Gopal, Heang-Ping Chan,[a] Todd Wilson, Mark Helvie, Nicholas Petrick, and Berkman Sahiner
Department of Radiology, University of Michigan, Ann Arbor, Michigan 48109-0030

(Received 11 November 1998; accepted for publication 13 September 1999)

Analysis of interval change is a useful technique for detection of abnormalities in mammographic interpretation. Interval change analysis is routinely used by radiologists and its importance is well-established in clinical practice. As a first step to develop a computerized method for interval change analysis on mammograms, we are developing an automated regional registration technique to identify corresponding lesions on temporal pairs of mammograms. In this technique, the breast is first segmented from the background on the current and previous mammograms. The breast edges are then aligned using a global alignment procedure based on the mutual information between the breast regions in the two images. Using the nipple location and the breast centroid estimated independently on both mammograms, a polar coordinate system is defined for each image. The polar coordinate of the centroid of a lesion detected on the most recent mammogram is used to obtain an initial estimate of its location on the previous mammogram and to define a fan-shaped search region. A search for a matching structure to the lesion is then performed in the fan-shaped region on the previous mammogram to obtain a final estimate of its location. In this study, a quantitative evaluation of registration accuracy has been performed with a data set of 74 temporal pairs of mammograms and ground-truth correspondence information provided by an experienced radiologist. The most recent mammogram of each temporal pair exhibited a biopsy-proven mass. We have investigated the usefulness of correlation and mutual information as search criteria for determining corresponding regions on mammograms for the biopsy-proven masses. In 85% of the cases (63/74 temporal pairs) the region on the previous mammogram that corresponded to the mass on the current mammogram was correctly identified. The region centroid identified by the registration technique had an average distance of 2.8 ± 1.9 mm from the centroid of the radiologist-identified region. These results indicate that our new registration technique may be useful for establishing correspondence between structures on current and previous mammograms. Once such a correspondence is established an interval change analysis could be performed to aid in both detection as well as classification of abnormal breast densities. © *1999 American Association of Physicists in Medicine.* [S0094-2405(99)00612-4]

Key words: image registration, computer-aided diagnosis, computer vision, interval change, breast cancer

I. INTRODUCTION

Mammography is currently the most effective method for early breast cancer detection.[1,2] A variety of computer-aided diagnosis (CAD) techniques have recently been developed to detect mammographic abnormalities and to distinguish between malignant and benign lesions.[3–8] Knowledge from diverse areas such as signal and image processing, pattern recognition, computer vision, artificial intelligence, and neural networks has been used to develop algorithms to be implemented within a CAD scheme. Varying degrees of success for these approaches have been reported in the literature. One common feature of most of these CAD techniques is that they use a single mammogram for analysis. However, some malignancies may only manifest as a new density on mammograms without associated calcifications or masses, others distinguish themselves from benign lesions only by their relatively rapid changes in sizes. Therefore, radiologists routinely use several mammographic views along with mammograms obtained in previous years for detecting and evaluating breast lesions and for identifying interval changes. The importance of interval change analysis in mammographic interpretation has been established in clinical practice.[9,10] It can be expected that analysis of changes in mammographic features between current and previous mammograms of the patient will also be an important component of a CAD system for both the detection and the classification tasks. The ability for automated analysis of interval changes would further the ability of CAD to offer an objective second opinion. This improvement, in turn, could increase the positive predictive value of mammography, reduce the number of benign biopsies, and hence reduce both cost and patient morbidity.

While a number of CAD schemes use only a single mammogram, the simultaneous use of more than one mammogram has been under investigation for some time. Several researchers have used views of the contra-lateral breast for detecting masses and developing densities. For instance, Yin

et al.[11,12] have utilized architectural asymmetry between the right and left breasts to detect masses. While it is widely accepted that interval changes in mammographic features are very useful for both detection and classification of breast abnormalities, the development of CAD techniques to use this information has achieved limited success.[13–18] Sallam and Bowyer[13] have proposed a warping technique for mammogram registration. They manually obtained control points and calculated a mapping function for mapping each point on the current mammogram to a point on the previous mammogram. The mapping function was obtained based on local affine transformations, as well as interpolation and surface fitting techniques. A drawback of this technique is the need for manual demarcation of control points. Brzakovic *et al.*[14] have investigated a three-step method for comparison of most recent and previous mammograms. They first registered two mammograms using the method of principal axis, and partitioned the current mammogram using a hierarchical region-growing technique. The breast regions in the two mammograms were aligned with respect to each other by means of translation, rotation, and scaling. Although the technique was evaluated on a total of 64 images obtained from eight cases, this work mainly aimed toward detecting cancerous changes in breast tissue and, therefore, no quantitative analysis of registration accuracy was presented. Vujovic and co-workers[15,16] have proposed a multiple-control-point technique for mammogram registration. They first determined several control points independently on the current and previous mammograms based on the intersection points of prominent anatomical structures in the breast. A correspondence between these control points was established based on a search in a local neighborhood around the control point of interest. In a more recent publication,[17] they have evaluated their approach for establishing the correspondence between control points extracted from two mammograms using 29 temporal image pairs, and presented a qualitative evaluation based on an observer study. They have demonstrated that 91% of 103 computer-matched control points were in agreement with those matched by a radiologist. An important assumption of their work was that the distances between the control points did not change significantly between the two mammograms. However, this assumption is not necessarily a valid one. Variations in compression could potentially cause a large variation in the relative distances between the control points. Furthermore, the control points representing the intersections of elongated structures do not always have correspondences on the two mammograms. Most of these points are two-dimensional projection image of structures at different depths of an elastic and compressible three-dimensional breast. The projected intersection points can thus vary from image to image and are not invariant lankmarks. As noted by the authors, the potential control points are not points that are naturally selected by a radiologist when examining mammograms. Hence, the significance of these points is debatable.

An important factor that may limit the success of the above-mentioned techniques is that the extraction of any meaningful information from previous mammograms first requires a common frame of reference between the current and previous mammograms. Several complicating factors confound obtaining such a frame of reference. These factors include differences in breast compression and positioning between the current and previous mammograms, differences in the imaging technique between the two examinations, and changes in breast structure, size, and tissue density between the two images with patient age. As a result, the mammographic appearance of breast tissue on the current and previous mammograms of the same patient may vary considerably. Although these variabilities have not been quantified experimentally, they can be observed easily from most mammograms. Conventional registration techniques work well for applications involving rigid objects. Because of the elasticity of the breast tissue, the absence of obvious landmarks, and the large variability in the relative positions of the breast tissues projected onto the mammogram from one examination to the other, these techniques may not be optimal for registration of breast images.

In mammographic interpretation, a radiologist routinely compares the current mammogram with previous mammograms (if available) of the same view in order to detect changes in mammographic features. For example, if a mass is detected in the current mammogram, the radiologist searches for that mass in the previous mammogram to determine if this is a new or developing density. If the corresponding mass is found on the previous mammogram, then the radiologist compares the current and previous mass size and estimates if the mass has increased in size. To facilitate these comparisons, we plan to develop automated methods to detect the interval changes as a part of a computer-aided diagnostic system. As a first step, we have developed a novel method for automatic registration of lesions on temporal pairs of mammograms. In our approach, the computer emulates the search method used by many radiologists for finding corresponding structures on mammograms. The method aims at registering a small region containing a suspected mass on the most recent mammogram of the patient with one on a mammogram obtained from a previous year. Our regional registration technique involves three steps: (1) identification of a suspicious structure on the most recent mammogram, (2) initial estimation of the location on a previous mammogram of the region corresponding to the suspicious structure and the definition of a search region which encloses the object of interest on the previous mammogram, and (3) accurate identification of the location of the matched object within the search region. After the two matched lesions are identified, their characteristic features can be automatically extracted and interval changes estimated. In the present study, we focused on the development and the evaluation of the regional registration technique, rather than to solve the entire interval change analysis problem. The subsequent steps in the interval change analysis are beyond the scope of this study.

In the following sections we will provide a detailed description of our regional registration technique for temporal registration of mammograms and the results of a quantitative evaluation using a data set of 74 temporal image pairs. Although we evaluated a semiautomated version of the tech-

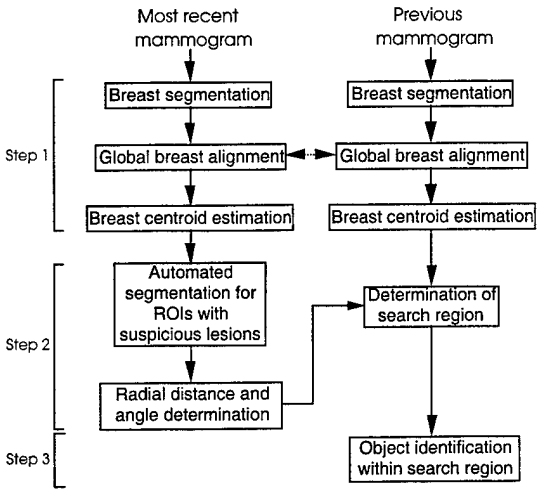


FIG. 1. Regional registration technique for determining an object on the previous mammogram which corresponds to a suspicious object on the most recent or current mammogram.

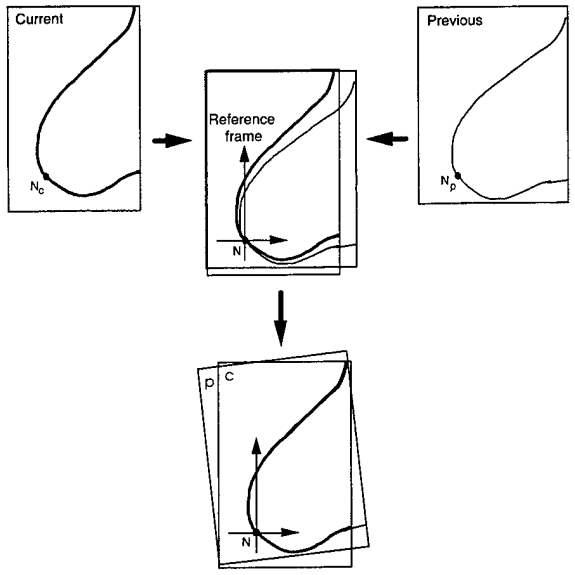


FIG. 2. Global breast alignment based on the mutual information between the two breast regions. N_c—nipple location in current mammogram, N_p—nipple location in previous mammogram, N—nipple location for both current and previous mammograms after translating them to the common frame of reference. The previous mammogram is rotated until the mutual information between the two mammograms is maximized.

nique in this preliminary study, it can be fully automated by incorporating a nipple detection step so that no user interaction will be required.

II. MATERIALS AND METHODS

A. Regional registration and mammogram correspondence

As the term indicates, regional registration is a local rather than a global registration technique. It is a multistep procedure and utilizes computer-detected objects in the most recent (hereafter termed current) mammogram. In the context of this paper, a current mammogram is either the latest mammogram of the patient, or the latest mammogram before biopsy. The detected objects could be either true masses (benign or malignant) or false positives (normal breast structures). Regional registration then finds a matching object on a previous mammogram. The three major steps in regional registration are illustrated in Fig. 1 and details of the technique are described below.

In the first step of regional registration, the breast region is segmented from the background on both the current and the previous mammograms. For this purpose we have used a breast boundary detection algorithm previously developed in our laboratory.[19,20] This algorithm could successfully track the breast boundaries in over 90% of the 1000 mammograms in a previous study. It performed reliably on all the images in our database. After extracting the breast border from the mammogram, the location of the nipple is estimated on both the current and the previous mammograms. Any automated method[21,22] can be used for finding the nipple location. However, in this study, the nipple location was manually identified by a radiologist for all images in our data set. The breast border and the nipple location now form the basis of a global breast alignment (GBA) procedure illustrated in Fig. 2. Since the sizes and the orientations of the two images could vary between the current and previous mammograms, a common frame of reference is needed. The GBA procedure has been devised specifically to provide such a frame of reference. We first define a new frame of reference with the nipple location on the current mammogram (N_c) as the origin. The previous mammogram is translated so that its nipple location (N_p) aligns with the origin in the common frame of reference as shown in Fig. 2. Using the origin as the pivot point, we rotate the previous mammogram to align the breast regions in the two images.

We have evaluated two different methods for estimation of the optimum rotation angle. The first method is based on maximization of the overlap area, and the second method is based on maximization of the mutual information (MI)[23,24] between the two segmented breast regions. To determine the MI, we first rescale the breast portion of both mammograms to a 0–255 gray scale. For a given rotation angle θ, the two-dimensional (2D) histogram $h_\theta(i,j)$ of the gray levels for the corresponding pixels on the current mammogram and the previous mammogram is constructed. Here i refers to the gray level on the current mammogram and j refers to the gray level on the previous mammogram rotated by an angle θ. The probability density of the gray scale co-occurrences is estimated from the 2D histogram as

$$f_\theta(i,j)=\frac{h_\theta(i,j)}{\Sigma_{m,n}h_\theta(m,n)}, \tag{1}$$

where $0\leq i,j\leq 255$, $0\leq m,n\leq 255$. The mutual information (MI_θ) between the two images for a specific rotation angle θ is computed as

$$MI_\theta=\sum_{i,j} f_\theta(i,j)^* \log_2\left\{\frac{f_\theta(i,j)}{\Sigma_m f_\theta(i,m)\Sigma_n f_\theta(n,j)}\right\}. \tag{2}$$

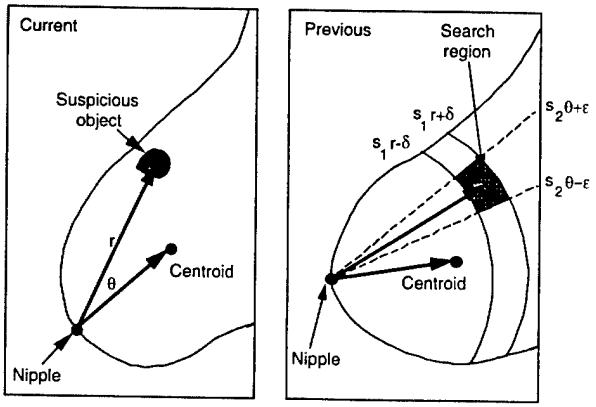


FIG. 3. Polar coordinate system defined using the nipple location and the nipple–centroid axis. The search region for finding a matching object on the previous mammogram is shown as the shaded region.

The above-mentioned procedure is repeated for several rotation angles and the angle θ_{max} which provides the maximum mutual information is chosen for global breast alignment of the previous mammogram and the current mammogram. Note that while the area overlap method for GBA uses the binary image after segmentation, the MI-based method uses the original gray scale image. The effects of the two methods on the accuracy of regional registration will be discussed later in Sec. IV. Once the two images are aligned in the common frame of reference, the centroid of the breast region is estimated, and the nipple–centroid axis is defined for both mammograms. For comparison we also show in Sec. III regional registration results based on computing the centroids of the two breast regions without global breast alignment. The nipple–centroid axis forms the basis for the second step of regional registration.

In the second step, suspicious regions are automatically segmented from the breast region on the current mammogram. This can be accomplished by using a density-weighted contrast enhancement (DWCE) technique[25] previously developed in our laboratory. While the use of the DWCE technique is not critical for regional registration, it does help automate the entire procedure. Alternatively, a radiologist can manually identify a suspicious object or a region of interest on the current mammogram and the regional registration technique can be used to identify a corresponding region on the previous mammogram. Once suspicious objects have been identified on the current mammogram, the centroid of each object is estimated. A polar coordinate system is then defined using the nipple as the origin and the nipple–centroid axis as the 0° axis on both images. This is illustrated in Fig. 3. The location of the centroid of a suspicious object on the current mammogram is determined as (r,θ). We then compute two scale factors—the radial scale factor s_1 and the angular scale factor s_2. These scale factors have been devised to provide a first-order correction for factors such as breast compression differences between the current and previous mammograms, differences in image magnification and size, and changes in overall breast shape between the two images. The radial scale factor s_1 is estimated as the ratio of the nipple–centroid distances on the previous and current images. The angular scale factor s_2 is estimated as the ratio of the angular width of the breast on the previous image at radius $s_1 r$ to that on the current image at radius r. The initial estimate of the corresponding location of the suspicious object on the previous mammogram is then obtained as $(s_1 r, s_2 \theta)$.

Using the initial estimate of the centroid of the object on the previous mammogram, we can define a fan-shaped search region bounded by $s_1 r \pm \delta$ and $s_2 \theta \pm \epsilon$ as illustrated in Fig. 3. The object found on the current mammogram is then used as a template to search for a matching object in the search region on the previous mammogram. The size of the search region (defined by δ and ϵ) depends on the variability between mammograms obtained from one examination to the other. Since it is difficult to predict the variability of an elastic and deformable object such as the breast by any analytical method, we have determined this variability experimentally from the mammograms in our data set. The variation in compression can cause a change in the relative locations of various breast structures on these images as well as a rotation of the breast boundary with respect to the fixed image coordinates. By relating the position of a breast structure to the corresponding nipple–centroid axis, and by performing a search in the corresponding search region, we can reduce the effect of this variability. In this study we have estimated the size of the search region required to enclose all corresponding objects on the previous mammogram using ground truth objects identified on the previous mammograms by a radiologist. The distance of the initial estimate of the center of the search region from the centroid of the ground truth object was also estimated.

The third and final step in the regional registration procedure involves a systematic search to identify a corresponding structure within the fan-shaped search region on the previous mammogram. In this study we have evaluated two different search criteria. The first criterion is based on gray scale template matching. A rectangular gray scale template centered on the mass centroid is extracted from the current mammogram. The choice of the size of the template region can affect the accuracy of the registration technique. The minimum required size of a rectangular template is, of course, a rectangular region which encloses the mass exactly. However, one can also include a small portion of the background region in the template. We have analyzed the performance of our algorithm using two different sizes for this template. The first includes a 1-pixel-wide background region all around the boundary of the suspicious object while the second includes a 5-pixel-wide background region. For each pixel (i,j) in the fan-shaped region on the previous mammogram, a region of interest (ROI) centered on the pixel and of the same size as the mass template is extracted. We denote the (m,n)th pixel in the gray scale template extracted from the current mammogram as $p(m,n)$ and that from the ROI obtained from the fan-shaped region as $q_{i,j}(m,n)$. A correlation measure defined as

$$C_{i,j}=\frac{\Sigma_{m,n}(p(m,n)-\bar{p})(q_{i,j}(m,n)-\bar{q})}{\sqrt{(\Sigma_{m,n}(p(m,n)-\bar{p})^2)(\Sigma_{m,n}(q_{i,j}(m,n)-\bar{q})^2)}} \quad (3)$$

is then obtained for each pixel (i,j) within the search region on the previous mammogram. Here the summation is performed over the mass template, and $\bar{p}$ and $\bar{q}$ denote the average pixel values in the template and ROI, respectively. The correlation values in the search region are then smoothed by a 3×3 averaging kernel to reduce fluctuations. The final estimate of the location of the mass centroid on the previous mammogram is obtained as the location corresponding to maximum correlation. The second search criterion is based on maximizing the mutual information between the mass template and the ROI extracted from within the search region. The MI approach is similar to that described earlier for alignment of the breast regions, except that the regions to be matched are limited to the size of the mass template.

Once a corresponding structure is found on the previous mammogram for a suspicious object on the current mammogram, it can be used for an interval change analysis within a CAD scheme, as we have shown in an independent study.[26] If the search procedure in the fan-shaped region does not yield a corresponding region, then the suspicious object on the current mammogram can be considered as a newly developed density. Objects for which no corresponding object can be found on the previous mammogram can be analyzed with methods designed for single images in an overall CAD scheme. Note that in this study the search techniques are structured in a way to always determine a matching object. Search criteria to identify new densities will be developed in future studies.

B. Image acquisition and data set

The data set for this study consisted of 127 images obtained from the files of 34 patients who had undergone biopsy at the University of Michigan. From these 127 mammograms, 74 temporal pairs of images were obtained. The current mammogram of each temporal pair exhibited a biopsy-proven mass. All previous mammograms in the 74 temporal pairs contained a mass, a structure, or a density which the radiologist could match to the mass detected in the corresponding current image. Since some patient files contained a sequence of mammograms over three years, the number of temporal pairs was larger than half the number of images. The 74 temporal image pairs were comprised of 43 cranio-caudal views and 31 mediolateral-oblique views.

The mammograms of 20 temporal pairs were digitized with a LUMISYS DIS-1000 laser scanner at a pixel resolution of 0.1 mm×0.1 mm and with 12 bit resolution. The digitizer was calibrated so that the gray values were linearly and inversely proportional to the optical density (OD) within the range of 0.1–2.8 OD units, with a slope of −0.001 OD/pixel value. Outside this range, the slope of the calibration curve decreased gradually. The OD range of this digitizer was 0–3.5. The mammograms of the remaining 54 temporal pairs were digitized with a LUMISCAN 85 laser scanner at a pixel resolution of 0.05 mm×0.05 mm and with 12 bit resolution. This digitizer was calibrated so that the gray values were linearly and inversely proportional to the OD within the range 0–4 OD units, with a slope of −0.001 OD/pixel value. All images were subsequently reduced to 0.8 mm resolution by averaging adjacent 8×8 pixels (20 pairs) or 16×16 pixels (54 pairs). Since the same digitizer was used for digitizing all films of the same case, the differences in the digitizers would have no effect on the analysis of each image pair. Given the small differences between the two laser digitizers and the large differences in the imaging technique and in the breast appearance from one case to another, it could be expected that the use of cases collected with the two different digitizers would not affect the evaluation of the registration technique.

While the regional registration technique can be used for determining a corresponding structure or region for any structure (both false positives and masses) in the breast, in this study we have analyzed its accuracy on biopsy-proven masses alone. The location of the mass on the current mammogram was identified by an MQSA-certified radiologist experienced in breast imaging. The radiologist manually identified the corresponding region on the previous mammogram and the nipple location on both the current and the previous mammograms using an interactive image analysis tool on a UNIX workstation. For each current mammogram, the boundary of the mass was manually delineated by the radiologist using an image display program developed in our laboratory. A bounding box enclosing the corresponding object on the previous mammogram was provided by the radiologist for each of the masses. Each mass as well as the corresponding structure on the previous mammogram was rated for its visibility on a scale of 1–10, where the rating of

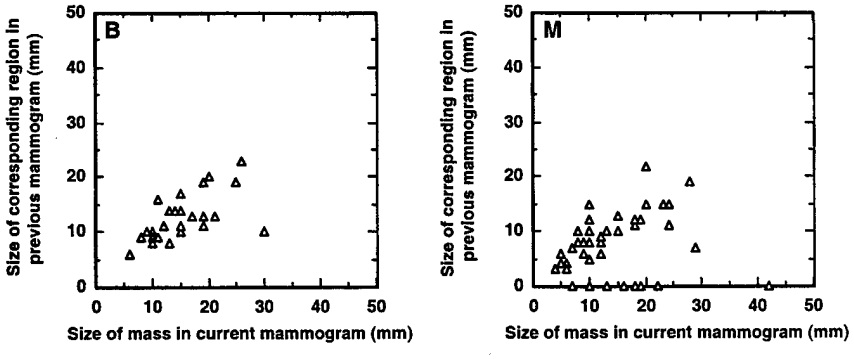


FIG. 4. Distribution of the size of the mass on the current mammogram with respect to the size of the corresponding structure on the previous mammogram as estimated by an experienced breast radiologist for benign (B) and malignant (M) cases in the data set.

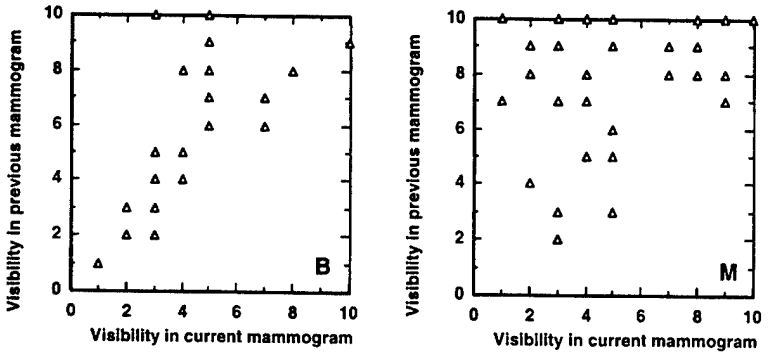


FIG. 5. Distribution of the visibility of the mass on the current mammogram with respect to the visibility of a corresponding structure on the previous mammogram as rated by an experienced breast radiologist for benign (B) and malignant (M) cases. In this rating scale the visibility of the masses decreases from 1 to 10 with 10 being the least visible. The total number of points in these two graphs is less than the total number of mammogram pairs in our database, because mammogram pairs with the same rating appear as a single point.

1 corresponded to the most visible category. The size of the mass on the current mammogram as well as the size of the corresponding structure on the previous mammogram was also provided by the radiologist. For previous mammograms on which the radiologist could not identify a distinct mass, the "mass" size was given a size of 0 mm. The parenchymal density was rated based on the BIRADS lexicon. The distributions of the size and visibility ratings for benign and malignant cases in this data set are shown in Figs. 4 and 5.

C. Evaluation of registration accuracy

The bounding box enclosing the corresponding object on the previous mammogram provided by the radiologist was used as the "ground truth" to evaluate the accuracy of the regional registration technique. We have used two different measures for assessing registration accuracy. The first measure quantifies whether the corresponding region is correctly identified by the registration algorithm. This measure is computed simply as the number of cases in which the estimated centroid location of the mass on the previous mammogram is inside the bounding box provided by the radiologist. The second measure quantifies the error in the estimate of the corresponding region on the previous mammogram and is defined as the Euclidean distance between the estimated centroid of the corresponding region and the center of the bounding box provided by the radiologist. Together these two measures answer the questions: (a) does regional registration work? (b) how well does the technique perform in matching structures between the current and previous mammograms? In Sec. III we provide the results of regional registration with and without global breast alignment and using both correlation and mutual information as the search criterion in step 3.

III. RESULTS

To provide the reader with a qualitative idea of algorithm performance we first illustrate the intermediate results at various stages of the algorithm. Then the results of each of the three steps of the algorithm are presented with an analysis of the dependence of the performance on various algorithm parameters. Also presented is an analysis of the accuracy of regional registration using the error measures defined in Sec. II C. In the following sections, the term "initial estimate" refers to the estimate of the center of the search region in step 2 of regional registration. The term "final estimate" refers to the outcome of the search procedure adopted in step 3 and represents the overall result of regional registration.

A. Intermediate results of regional registration

Figures 6–8 show an example of the intermediate and final results of applying the regional registration technique to a temporal pair of mammograms. The original digitized mammograms—current and previous—with the automati-

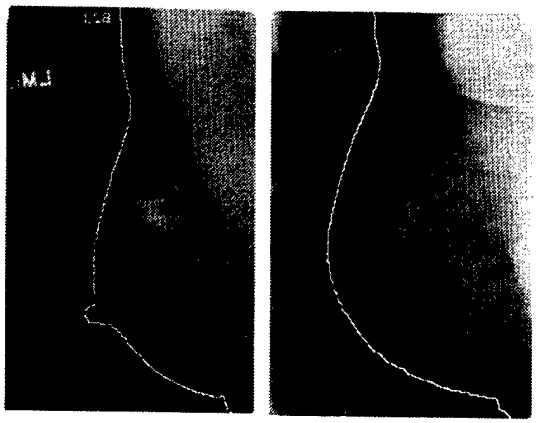


FIG. 6. Left—most recent or current mammogram. Right—previous mammogram. The breast images are superimposed with the breast borders detected by a breast boundary tracking algorithm.

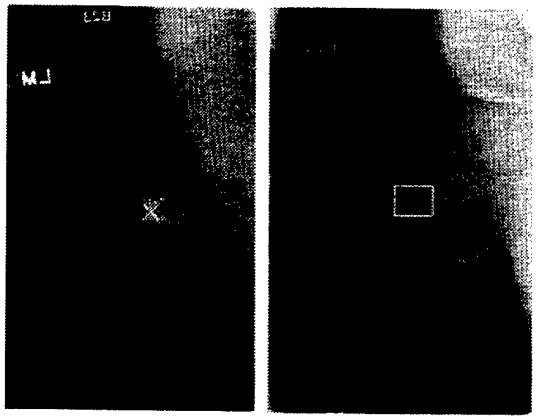


FIG. 7. Left—location of the mass on the current mammogram. Right—radiologist-identified region on previous mammogram corresponding to the mass on the current mammogram.

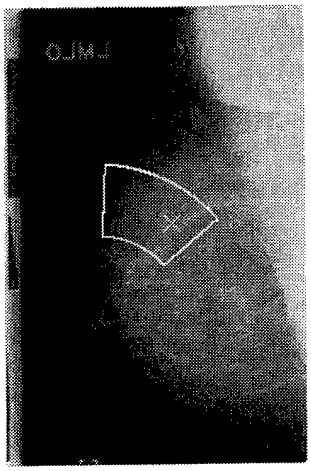


FIG. 8. The fan-shaped search region on the previous mammogram. The initial computer estimate of the centroid location of the region corresponding to the mass is at the center of the search region. The final estimate of the centroid of the corresponding region (indicated by *X*) is obtained by using the correlation criterion within the fan-shaped search region.

cally tracked breast boundaries superimposed, are shown in Fig. 6. The location of the mass on the current mammogram is shown in Fig. 7 along with the corresponding radiologist-identified region on the previous mammogram. Figure 8 shows the fan-shaped search region on the previous mammogram estimated in step 2 of regional registration. The initial estimate is at the center of this search region which is to be used in step 3 for localization of the corresponding mass. The centroid location of the corresponding object estimated by the algorithm using the correlation measure as the search criterion is also shown in Fig. 8.

B. Initial estimates and search regions

Figure 9 shows histograms of the Euclidean distance between the initial estimate of the centroid location of the corresponding structure on the previous mammogram and the center of the bounding box provided by the radiologist. For the 74 temporal image pairs used in this data set, the average Euclidean distance error of the initial estimate was 10.5 mm (std. dev. 6.4 mm) without the GBA procedure and 9.8 mm (std. dev. 6.0 mm) with the GBA procedure. The overall accuracy was 46% in both cases, i.e., in 34 of the 74 temporal image pairs the initial estimate was inside the ground-truth bounding box. Based on observation of the radial deviation errors and the angular deviation errors (defined in Sec. IV) in Figs. 10 and 11, a search region defined by $\epsilon = 0.35 + 5/r$ rad and $\delta = 20$ mm with GBA ($\delta = 25$ mm for no GBA), where r is the radial distance from the nipple, was used for the evaluation of the local search criteria used in step 3 of regional registration.

C. Local search criteria and final estimates

Figure 12 shows the histograms of the Euclidean distance errors of the final estimate of the corresponding structure using the correlation measure as the search criterion. Table I summarizes the results along with the average Euclidean distance errors and standard deviations using both the correlation and the mutual information search criteria and with and without the GBA procedure. The average Euclidean distance errors and deviations for the cases where the final estimate is inside the ground-truth region identified by the radiologist and the cases where it is outside are also listed separately. Regional registration incorporating the GBA procedure and using correlation as a search criterion has an accuracy of 85%. In 63 of the 74 temporal image pairs, the final estimate of the location of the corresponding region was inside the radiologist-identified ground-truth region. The use of mutual information as a search criterion yielded an accuracy of 74% (55 out of 74 temporal pairs). The average Euclidean distance error for regional registration incorporating GBA and correlation was 4.7 mm (std. dev. 5.8 mm) for all 74 temporal pairs and 2.8 mm (std. dev. 1.9 mm) in 85% (63/74) of the temporal pairs. Use of mutual information as a search criterion in step 3 results in values of 7.2 mm (std dev. 8.6 mm) and 3.0 mm (std. dev. 2.0 mm), respectively, for the same quantities.

IV. DISCUSSION

A. Initial estimates and search regions

From the histograms of Fig. 9, we observe that the use of the GBA procedure results only in a marginal improvement in the initial estimate, if the Euclidean distance error is the only measure considered. However, the GBA procedure has a significant effect in reducing the size of the search region required for regional registration. In order to compute the required sizes (δ and ϵ in Fig. 3) of the search region, we computed two quantities—the radial distance deviation and the angular deviation—using the initial estimate obtained from step 2 for the 74 temporal image pairs. The radial distance deviation is defined as the absolute difference between $s_1 r$ and r_c, where r_c is the radial distance of the center of the ground-truth region from the nipple location on the pre-

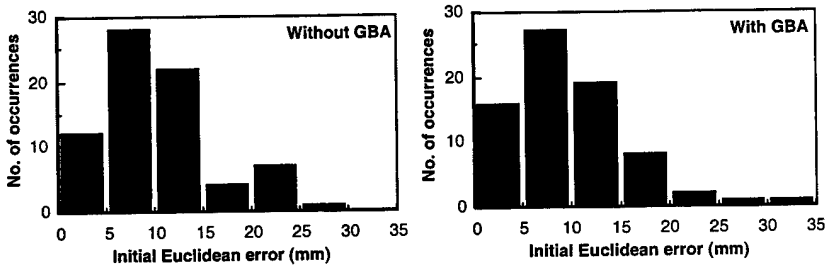


FIG. 9. Histograms of Euclidean distance between the initial estimate of the centroid location of the corresponding object and the center of the radiologist-identified object on the previous mammogram with and without GBA.

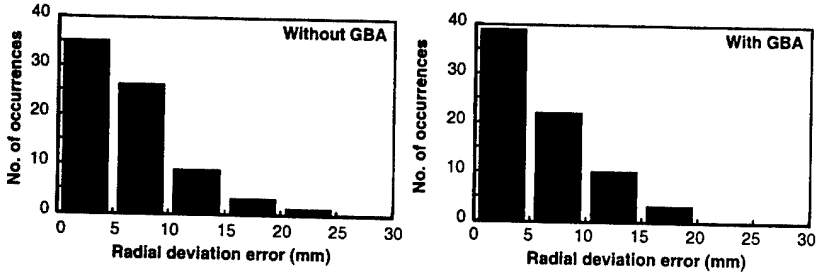


FIG. 10. Histograms of radial distance deviation between the initial estimate of the centroid location of the corresponding object and the center of the radiologist-identified object on the previous mammogram with and without GBA.

vious mammogram. The histograms of radial distance deviations for the 74 temporal image pairs with and without the GBA procedure are shown in Fig. 10. An important observation is that a δ value of 25 mm is needed to include the centers of the ground-truth structures if the GBA procedure is not used in step 1. The use of the GBA procedure results in a decrease in the value of δ to 20 mm. This decrease helps significantly increase the overall accuracy of the regional registration as discussed below.

In Fig. 11 the angular deviation of the initial estimate is plotted against the radial distance of the centers of the ground-truth regions on the previous mammogram. The angular deviation ϵ is defined as $s_2\theta - \theta_c$ where θ_c is the angle between the nipple-ground-truth center vector and the nipple-centroid axis. In an earlier study[27] using both false positives and masses, we have observed that the value of ϵ needed to include the center of the ground-truth region decreases with distance from the nipple, i.e., increases with distance from the chest wall. This may be attributed to the increased deformability of the breast tissue closer to the nipple compared to the tissue closer to the chest wall. This indicates that a possible approach to take into account this variability is to incorporate a variable ϵ, one which is inversely proportional to the radial distance r from the nipple. For the data set in this study, we have investigated several forms for this dependence all of which fit under the general model

$$\epsilon = \epsilon_{th} + K/r.$$

Here ϵ_{th} and K are two constants which affect the form of the dependency. Based on our observation of the angular deviations for the entire data set of 74 temporal pairs we have chosen $\epsilon_{th} = 0.35$ rad and $K = 5$ rad-mm. As can be seen from Fig. 11, with these values of ϵ_{th} and K, all of the centers of the ground-truth regions are within the search region. Therefore, a search region defined by $\epsilon = 0.35 + 5/r$ rad, and $\delta = 20$ mm (if GBA was applied) or $\delta = 25$ mm (if GBA was not applied) was used for evaluation of the local search criteria used in step 3 of regional registration.

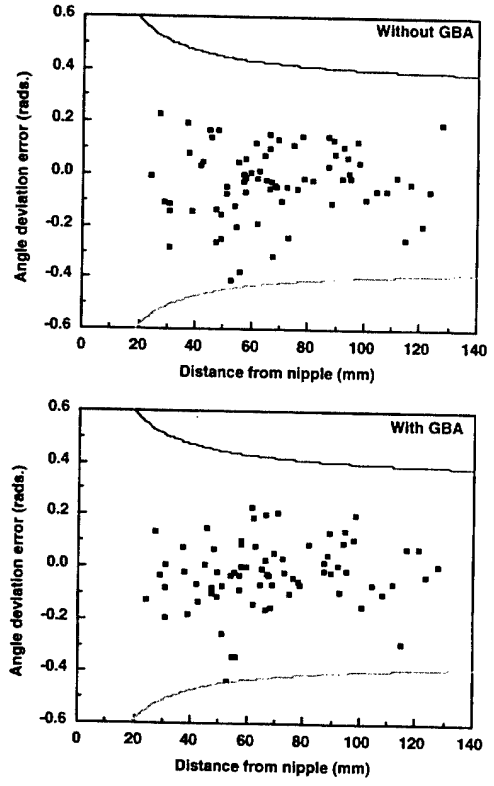


FIG. 11. Angular deviation between the initial estimate of the centroid location of the corresponding object and the center of the radiologist-identified object on the previous mammogram with and without GBA. Also shown are the bounding lines defined using $\epsilon = 0.35 + 5/r$ rad.

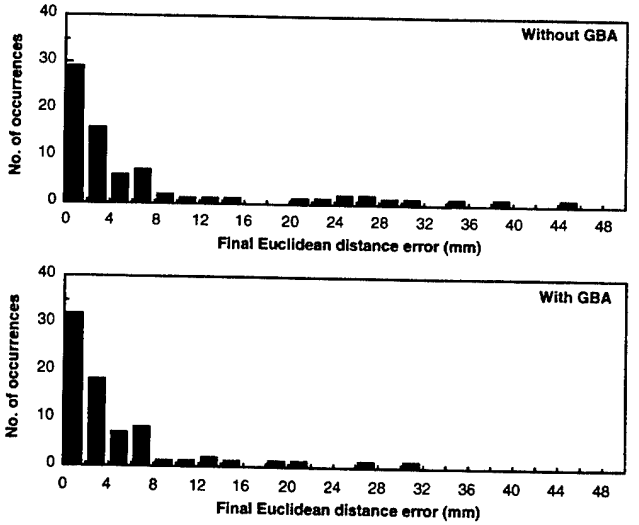


FIG. 12. Histograms of Euclidean distance error for corresponding regions estimated by regional registration using the correlation measure in step 3 with and without GBA. This error is defined as the Euclidean distance between the centroid location of the estimated corresponding region and the center of the radiologist-identified ground-truth corresponding region on the previous mammogram.

TABLE I. Accuracy of regional registration using correlation measure and mutual information measure in step 3 with and without global breast alignment (GBA) and using a 1-pixel-wide background region for the template from the current mammogram. Correct estimates are the cases where the estimated centroid location was within the bounding box of the radiologist-identified object location.

Method	Accuracy	Overall average error (mm)	Average error (mm) for correct estimates	Average error (mm) for incorrect estimates
Correlation without GBA	77% (57/74)	7.4±10.2	2.8±2.0	22.9±11.5
Mutual information without GBA	68% (50/74)	8.8±10.5	3.0±2.0	20.7±11.1
Correlation with GBA	85% (63/74)	4.7±5.8	2.8±1.9	15.7±8.3
Mutual information with GBA	74% (55/74)	7.2±8.6	3.0±2.0	19.4±8.9

B. Local search criteria and final estimates

We have evaluated the use of correlation and mutual information as the local search criteria. From Table I we observe that the GBA procedure results in a higher accuracy irrespective of the search criterion. While the use of mutual information as a search criterion performs reasonably well by itself (74% accuracy with an average error of 7.2 mm) the use of correlation measure was observed to result in more accurate registration. For the images in this data set, the correlation measure outperformed the mutual information measure irrespective of whether the breast centroids were computed with or without the GBA procedure.

A few observations on the 11 cases where the final estimate was outside the radiologist-identified ground-truth corresponding region are in order. In 7 of the 11 cases although the radiologist did provide a region corresponding to the mass on the current mammogram, the corresponding structure on the previous mammogram was very subtle (visibility rating 8 or higher) with indistinct boundaries. The radiologist could only estimate the region where the mass would develop rather than the mass itself, so the truth was uncertain. In one of the remaining 4 cases, the mass was an architectural distortion in the current mammogram. In a second (benign) case the mass shape had changed considerably. Upon consultation of the pathology report, the radiologist concluded that the mass was a benign cyst which had been aspirated in the previous year resulting in a substantial change in its shape. In the third case, the proximity of the mass to the chest wall resulted in it being incompletely imaged in the previous year compared to the current year. In such cases the correlation measure of a neighboring breast structure would tend to be higher than that of the corresponding structure. In the fourth case, an overlap of two vessels was identified as corresponding to the mass on the current mammogram while the region corresponding to the mass was observed to be extremely subtle. In almost all of the 11 cases the proximity of the corresponding region to a dense structure combined with the subtle nature of the structure on the previous mammogram render the correlation measure ineffective in establishing correspondence. However, in clinical practice, these masses will likely be categorized as a newly developed density. Criteria to distinguish a newly developed density will be investigated in further studies.

C. GBA: Area overlap vs mutual information

For the images used in this study, the result of the GBA procedure based on maximizing the area overlap between the breast regions in the two images of a temporal pair is comparable to that based on maximizing the mutual information. However, our observation is that the mutual information criterion is preferable to the area overlap criterion. The area overlap measure suffers from the drawback that if the breast region in one of the mammograms is uniformly smaller than that in the other, i.e., the breast edge in one is completely within the breast edge in the other, then there is no unique rotation angle at which the area overlap is maximized. Although the range of rotation angles over which local maxima of the area overlap occur is small, the resulting estimate of the rotation angle for GBA may be suboptimal. The use of mutual information, however, results in a single unique rotation angle at which MI is maximized. In any case, as discussed earlier, the use of the GBA procedure before computing the breast centroid results in a reduction in the size of the search region. A smaller search region reduces the likelihood that the mass template is matched to an incorrect structure and, therefore, increases the accuracy and reduces the Euclidean distance error.

D. Template size, scale factors, and computation times

The size of the background region in the gray scale template extracted from the current mammogram affects registration accuracy. For the 74 temporal pairs in this data set, the best performance was observed when a 1-pixel-wide background region was included all around the boundary of the mass template. A 5-pixel-wide background region resulted in a decrease in accuracy and an increase in the average Euclidean distance error. The accuracy progressively decreased and the Euclidean distance error increased with an increase in the size of the background region in the template. Figure 13 shows the distributions of the radial and angular scale factors for the images used in this study. The radial scale factor s_1 ranged from 0.94 to 1.05 for this data set. Use of s_1 reduced the size of the search area by decreasing the required value for δ. The angular scale factor s_2 was very close to 1 in all cases and did not seem to make any major difference for the images in this data set. On a final note the computation time required for regional registration incorporating correlation was on the average 2 s without GBA and 4 s with GBA on a UNIX workstation (DEC AlphaStation 600 series).

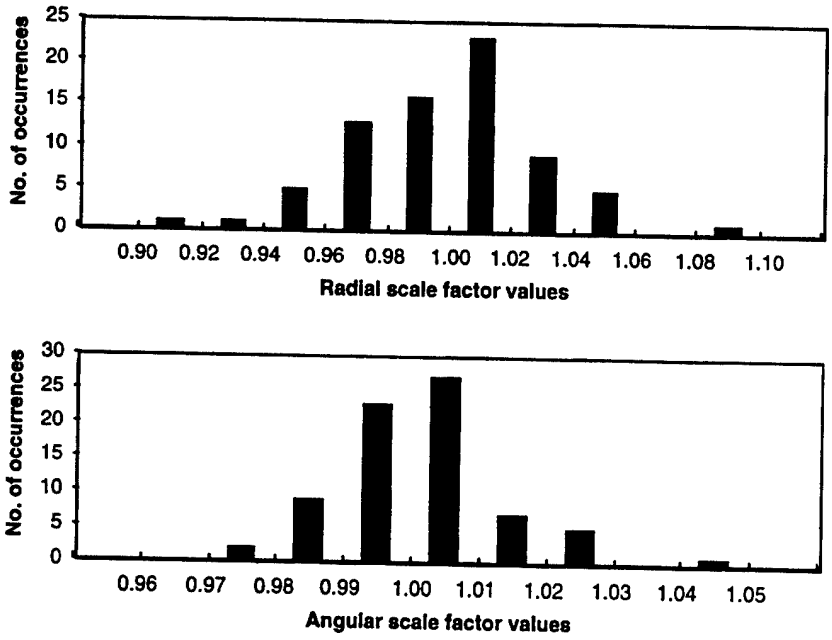


FIG. 13. Histograms of the radial scale factor and the angular scale factor for 74 temporal image pairs. The radial scale factor s_1 is estimated as the ratio of the nipple–centroid distances on the previous and current images. The angular scale factor s_2 is estimated as the ratio of the angular width of the breast on the previous image at radius $s_1 r$ to that on the current image at radius r.

V. CONCLUSIONS

Radiologists are interested in determining any local changes in breast tissue over time which may indicate a developing cancer. We have developed a novel regional registration technique for temporal registration of mammograms. This technique could become an important component of a CAD scheme for mammographic analysis. Unlike other techniques found in the literature, our regional registration technique does not depend on the identification of landmark structures or control points on the mammograms. It is based on a search technique that many radiologists use and has proven to be successful in mammographic interpretation. After corresponding objects are found, they can be analyzed for interval changes in a CAD scheme. Our preliminary results indicate that the regional registration technique is promising in identifying corresponding regions from temporal mammographic pairs. In 85% (63/74) of the cases the regional registration technique correctly identified the corresponding region in the previous mammogram. For these 63 cases, it is highly encouraging to note that the estimated location of the region corresponding to the mass in the current mammogram was less than 3 mm on the average from radiologist-identified corresponding locations.

Areas for future work include the development of an automated technique for identifying the nipple location on the mammograms, investigation of other local search criteria such as Fourier descriptors and shape-invariant moments to be used in the fan-shaped search region, adaptive methods for determining the size of the search region, criteria for identifying newly developed densities, application of regional registration to false positives as well as masses, and studies with a large data set to investigate the robustness of the regional registration technique. It may be noted that the regional registration technique may also be applicable to other related registration problems, such as the registration of left and right mammograms.

ACKNOWLEDGMENTS

This work is supported by a Career Development Award from the USAMRMC Grant No. DAMD 17-98-1-8211, USPHS Grant No. CA 48129, and USAMRMC Grant No. DAMD 17-96-1-6254. The content of this publication does not necessarily reflect the position of the government, and no official endorsement of any equipment or product of any companies mentioned in the publication should be inferred.

[a]Electronic mail: chanhp@umich.edu

[1] S. A. Feig and R. E. Hendrick, "Risk, Benefit, and Controversies in Mammographic Screening," in *Syllabus: A Categorical Course in Physics Technical Aspects of Breast Imaging*, edited by A. G. Haus and M. J. Yaffe (Radiological Society of North America, Oak Brook, IL, 1993).

[2] C. Byrne, C. R. Smart, C. Cherk, and W. H. Hartmann, "Survival advantage differences by age: Evaluation of the extended follow-up of the Breast Cancer Detection Demonstration Project," Cancer (N.Y.) **74**, 301–310 (1994).

[3] Y. Wu, K. Doi, M. L. Geiger, and R. M. Nishikawa, "Computerized detection of clustered microcalcifications in digital mammograms: Applications of artificial neural networks," Med. Phys. **19**, 555–560 (1992).

[4] H. P. Chan *et al.*, "Improvement in radiologists' detection of clustered microcalcifications: The potential of computer-aided diagnosis," Invest. Radiol. **25**, 1102–1110 (1990).

[5] S. M. Lai, X. Li, and W. F. Bischof, "On techniques for detecting circumscribed masses in mammograms," IEEE Trans. Med. Imaging **8**, 377–386 (1989).

[6] J. Kilday, F. Palmieri, and M. D. Fox, "Classifying mammographic lesions using computerized image analysis," IEEE Trans. Med. Imaging **12**, 664–669 (1993).

[7] W. P. Kegelmeyer, J. M. Pruneda, P. D. Bourland, A. Hillis, M. W. Riggs, and M. L. Nipper, "Computer-aided mammographic screening for spiculated lesions," Radiology **191**, 331–337 (1994).

[8] H. P. Chan, D. Wei, M. A. Helvie, B. Sahiner, N. Petrick, D. D. Adler, M. M. Goodsitt, and N. Petrick, "Computer-aided classification of mammographic masses and normal tissue: Linear discriminant analysis in texture feature space," Phys. Med. Biol. **40**, 857–876 (1995).

[9] L. W. Bassett, B. Shayestehfar, and I. Hirbawi, "Obtaining previous mammograms for comparison: Usefulness and costs," Am. J. Roentgenol. **163**, 1083–1086 (1994).

[10] E. A. Sickles, "Periodic mammographic follow-up of probably benign lesions: Results in 3183 consecutive cases," Radiology **179**, 463–468 (1991).

[11]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. **18**, 955–963 (1991).

[12]F. F. Yin, M. L. Giger, K. Doi, C. J. Vyborny, and R. A. Schmidt, "Computerized detection of masses in digital mammograms: Automated alignment of breast images and its effect on bilateral-subtraction technique," Med. Phys. **21**, 445–452 (1994).

[13]M. Sallam and K. Bowyer, "Detecting abnormal densities in mammograms by comparison with previous screenings," in *Digital Mammography '96*, edited by K. Doi, M. L. Giger, R. M. Nishikawa, and R. A. Schmidt (Elsevier, Amsterdam, 1996).

[14]D. Brzakovic, N. Vujovic, M. Neskovic, P. Brzakovic, and K. Fogarty, "Mammogram analysis by comparison with previous screenings" in Ref. 13.

[15]N. Vujovic, D. Brzakovic, and K. Fogarty, "Detection of cancerous changes in mammograms using intensity and texture measures," Proc. SPIE **2434**, 37–47 (1995).

[16]N. Vujovic, P. Bakic, and D. Brzakovic, "Detection of potentially cancerous signs by mammogram followup," in Ref. 13.

[17]N. Vujovic and D. Brzakovic, "Establishing the correspondence between control points in pairs of mammographic images," IEEE Trans. Image Process. **6**, 1388–1399 (1997).

[18]W. K. Zouras, M. L. Giger, P. Lu, D. E. Wolverton, C. J. Vyborny, and K. Doi, "Investigation of a temporal subtraction scheme for computerized detection of breast masses in mammograms," in Ref. 13.

[19]A. R. Morton, "Design of an x-ray beam equalization filter for mammographic imaging," M.S. thesis, Department of Environmental and Industrial Health, University of Michigan, 1996.

[20]A. R. Morton, H. P. Chan, and M. M. Goodsitt, "Automated model-guided breast segmentation algorithm," Med. Phys. **23**, 1107–1108 (1996).

[21]A. J. Mendez, P. G. Tahoces, M. J. Lado, M. Souto, J. L. Correa, and J. J. Vidal, "Automatic detection of breast border and nipple in digital mammograms," Comput. Methods Programs Biomed. **49**, 253–262 (1996).

[22]R. Chandrasekhar and Y. Attikiouzel, "A simple method for automatically locating the nipple on mammograms," IEEE Trans. Med. Imaging **16**, 483–494 (1997).

[23]F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, "Multimodality image registration by maximization of mutual information," IEEE Trans. Med. Imaging **16**, 187–198 (1997).

[24]A. Maintz, E. Meijering, and M. Viergever, "General multimodal elastic registration based on mutual information," Proc. SPIE **3338**, 144–154 (1998).

[25]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 tissue classification," Med. Phys. **23**, 1685–1696 (1996).

[26]S. Sanjay-Gopal, H. P. Chan, B. Sahiner, N. Petrick, T. Wilson, and M. Helvie, "Evaluation of interval change in mammographic features for computerized classification of malignant and benign masses," Radiology **205(P)**, 216 (1997).

[27]S. Sanjay-Gopal, H. P. Chan, N. Petrick, T. Wilson, B. Sahiner, M. Helvie, and M. Goodsitt, "A regional registration technique for automated analysis of interval changes of breast lesions," Proc. SPIE **3338**, 118–131 (1998).

Combined adaptive enhancement and region-growing segmentation of breast masses on digitized mammograms

Nicholas Petrick, Heang-Ping Chan, Berkman Sahiner, and Mark A. Helvie
The University of Michigan, Department of Radiology, CGC B2102, 1500 East Medical Center Drive, Ann Arbor, Michigan 48109-0904

(Received 15 July 1998; accepted for publication 27 April 1999)

As an ongoing effort to develop a computer aid for detection of masses on mammograms, we recently designed an object-based region-growing technique to improve mass segmentation. This segmentation method utilizes the density-weighted contrast enhancement (DWCE) filter as a preprocessing step. The DWCE filter adaptively enhances the contrast between the breast structures and the background. Object-based region growing was then applied to each of the identified structures. The region-growing technique uses gray-scale and gradient information to adjust the initial object borders and to reduce merging between adjacent or overlapping structures. Each object is then classified as a breast mass or normal tissue based on extracted morphological and texture features. In this study we evaluated the sensitivity of this combined segmentation scheme and its ability to reduce false positive (FP) detections on a data set of 253 digitized mammograms, each of which contained a biopsy-proven breast mass. It was found that the segmentation scheme detected 98% of the 253 biopsy-proven breast masses in our data set. After final FP reduction, the detection resulted in 4.2 FP per image at a 90% true positive (TP) fraction and 2.0 FPs per image at an 80% TP fraction. The combined DWCE and object-based region growing technique increased the initial detection sensitivity, reduced merging between neighboring structures, and reduced the number of FP detections in our automated breast mass detection scheme. © *1999 American Association of Physicists in Medicine.* [S0094-2405(99)00808-1]

Key words: computer-aided diagnosis, digital mammography, breast mass detection, density-weight contrast enhancement, region growing

I. INTRODUCTION

Mammographic screening has proven to be an effective method for early detection of breast cancer. Women in a regular mammographic screening program have a statistically significant reduction in breast cancer mortality when compared to women not in such a program.[1] In addition, independent double reading by two radiologists has proven to significantly increase the sensitivity of mammographic screening.[2] Therefore, regular screening and double reading would appear to be a sensible approach for breast cancer detection. While regular screening is emphasized in health care programs, the higher cost and increased workload on the radiologists may make double reading by two radiologists impractical in a general screening situation. Computer-aided diagnosis (CAD) is one alternative that could allow a large number of mammograms to be double read by a single radiologist aided by the computer. This technique may improve the accuracy of both detection and characterization of breast lesions.

Many researchers have been interested in computerized analysis of mammograms[3] and a number of groups have developed algorithms for automated detection of breast masses. The detection of spiculated masses has been of particular importance because of its high likelihood of malignancy. Karssemeijer *et al.*,[4] Kobatake *et al.*,[5] and Kegelmeyer *et al.*[6] have all proposed methods for detecting spiculated masses on digitized mammograms. However, since a number of malignant masses are not spiculated, other groups have tackled the general problem of identifying all types of breast masses on digitized mammograms.[3,7–11]

Our research group has reported on a method for automatically detecting masses on digitized mammograms.[10,12] The method employed multiple stages of density-weighted contrast enhancement (DWCE) segmentation. The DWCE segmentation was first applied to the full mammogram, and then reapplied to local regions within the mammogram to improve object border definition. A final object splitting stage was employed to eliminate merging between neighboring or overlapping breast structures. False positive (FP) reduction based on extracted morphological features was applied after each segmentation step with texture analysis used as a final arbitrator between masses and normal structures. The segmentation was evaluated on 168 digitized mammograms and it achieved a performance of 4.4 FPs per image at a 90% true positive (TP) detection fraction and 2.3 FPs per image at an 80% TP detection fraction.[10]

Our approach to mass detection has been to first identify all significant structures within the breast region using a global segmentation technique and then refine the initial object borders using local processing. Finally, we differentiate between true masses and normal structures using morphological and texture information. Our method is therefore different from other detection algorithms that utilize the object shape information for initial detection. The disadvantage

 0094-2405/99/26(8)/1642/13/$15.00 © 1999 Am. Assoc. Phys. Med.

our combined global and local detection approach is that a large number of normal structures are identified in the initial stage. This can lead to additional FPs if the classification is suboptimal. However, the advantage of this approach is that it can identify difficult masses since the initial detection is not based on shape information. The shape information is still used in the classification stage to reduce FPs.

In this paper, we present an improved version of our two-stage DWCE segmentation approach. This new scheme was designed to both increase specificity and reduce the overall complexity of the segmentation. A primary motivation is to develop a method for eliminating the merging between neighboring structures in the local DWCE processing step and thus improve local segmentation. We introduce an object-based region-growing technique to perform this task. Improved local segmentation serves a number of purposes. First, it improves the morphological and texture information used for FP reduction as well as eliminates the need for the shape-based splitting step. It also enables us to eliminate two morphological FP reduction steps. This significantly reduces the overall complexity of the detection program and should lead to a more practical implementation in a general clinical setting. In this paper, we summarize the intermediate and overall detection performance of the improved mass segmentation algorithm and describe some of its limitations.

II. METHODS

A. Database

The clinical mammograms used in this study were selected from the files of patients who had undergone biopsy at the University of Michigan Hospital. The mammograms were acquired with American College of Radiology (ACR) accredited mammography systems. Kodak MinR/MRE screen/film systems with extended cycle processing were used as the image recorder. The mammography systems have a 0.3-mm focal spot, a molybdenum anode, 0.03-mm thick molybdenum filter, and a 5:1 reciprocating grid. The selection criterion used by the radiologists was simply that a biopsy-proven mass existed on the mammogram. The data set consisted of 253 mammograms from 102 patients, and it included 128 malignant and 125 benign masses. Sixty-three of the malignant and six of the benign masses were judged to be spiculated by a MQSA approved radiologist. The size of the masses ranged from 5 to 29 mm (mean size=12.5 mm), and their visibility ranged from 1 (obvious) to 5 (subtle) (mean=2.1). Figures 1 and 2 show the histograms of mass size and mass visibility for the data set.[13] These distributions characterize the difficulty and diversity of the cases contained in the data set.

The mammograms were digitized with a LUMISYS DIS-1000 laser film scanner with a pixel size of 100 μm and 12 bit gray level resolution. The gray levels were linearly proportional to optical density in the 0.1 to 2.8 optical density unit (O.D.) range. The slope was 0.001 O.D./pixel value. The slope gradually fell off in the 2.8 to 3.5 O.D. range.[10,13] A large pixel value corresponds to a low optical density with this digitizer.

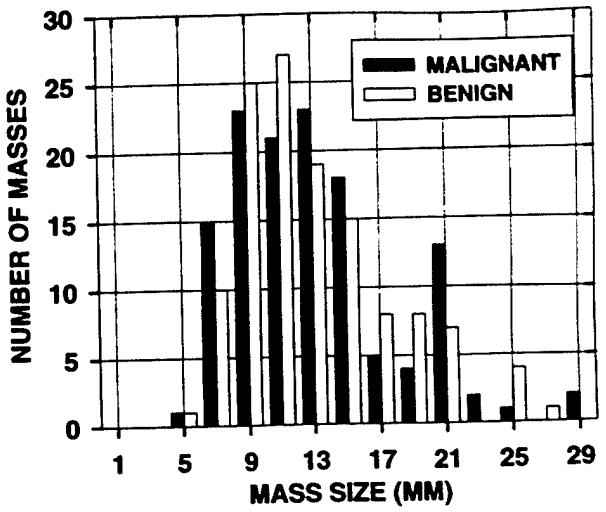


FIG. 1. Histograms of mass size for the 253 masses contained in our data set. Mass sizes were measured as the largest axis of the mass by an experienced breast radiologist.

The location and extent of all the biopsy-proven masses were marked on the original films. The radiologist then identified both the centroid of the lesion and the smallest bounding box containing the entire lesion using an interactive image manipulation tool on a workstation. Both procedures were performed using the original marked film as a guide. The lesion centroid was used to identify TP detections after the morphological FP reduction step. If a segmented object was within 4 mm of the mass centroid, it was considered a TP. All other segmented objects were considered as FPs. The final free-response receiver operating characteristic (FROC) curves following texture-based classification used the more precise mass bounding box for TP identification. A region was considered a TP only when it contained more than 50% of the mass bounding box.

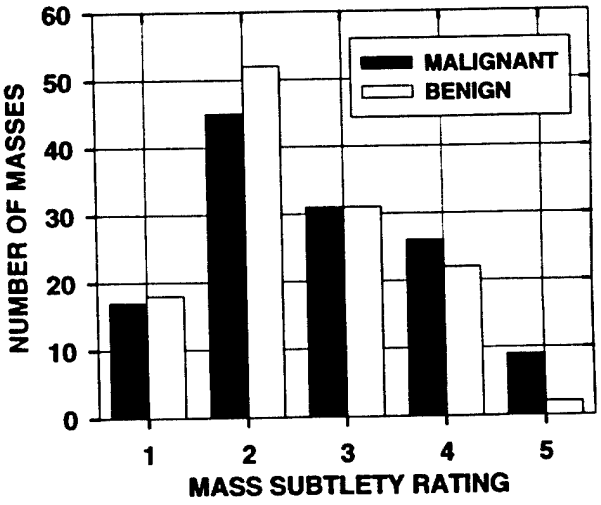


FIG. 2. Histograms of mass subtlety for the 253 masses contained in our data set. Mass subtleties were rated by an experienced breast radiologist from 1 (obvious) to 5 (subtle).

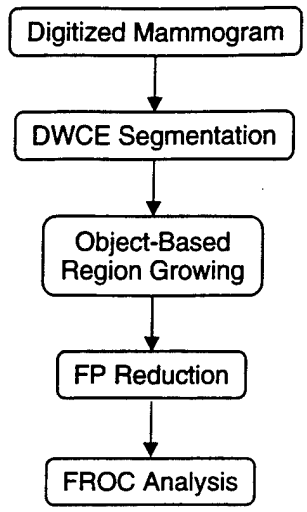


FIG. 3. Block diagram of the breast mass segmentation scheme. A digitized mammogram undergoes DWCE segmentation followed by object-based region growing and then morphological and texture classification. The performance of the segmentation scheme was evaluated by FROC analysis.

B. Density-weighted contrast enhancement segmentation

The block diagram for the proposed detection scheme is shown in Fig. 3. Global DWCE segmentation was used to identify an initial set of breast structures on the digitized mammograms. These objects were then used as seed locations to perform gradient-based region growing. A thorough description of the DWCE technique can be found in the literature.[10,12,14] Briefly, the DWCE technique employs an adaptive filter to enhance the local contrast and thus accentuate mammographic structures in an image. As the term implies, the parameters of the enhancement filter are based on the local density within the image and the filter is applied to the image on a pixel-by-pixel basis. The filter is designed to suppress very low contrast values, to emphasize the low to medium contrast values and to just slightly deemphasize the high contrast values. The effect of suppressing the extremely low contrast values is to reduce bridging between adjacent breast structures. Pixels with low to medium contrast values are enhanced so that more subtle structures can be detected. Finally, the slight deemphasis of the high contrast structures is included to provide a more uniform intensity distribution for detected structures. After contrast enhancement, Laplacian–Gaussian edge detection is applied and all enclosed objects are filled to produce a set of detected structures for the image. The DWCE segmentation is applied to mammograms that have been smoothed and subsampled from their original 100 μm pixel size to an 800 μm pixel resolution.[10] The DWCE stage has been found to be effective in detecting most breast structures including a significant portion of breast masses. However, the DWCE borders usually fall well inside the true borders of an object and a significant number of adjacent structures are merged into single objects. This occurs most frequently when the adjacent breast structures have some tissue overlap.

C. Object-based region-growing segmentation

1. Initial gray-scale region growing

Before gradient-based region growing was applied, an initial set of seed objects was identified. This was accomplished by first identifying all local maxima in the original gray-scale image which occurred within the extent of the DWCE objects. Local maxima were defined using the ultimate erosion technique described by Russ.[15] In simple terms, a pixel was a local maximum if and only if its value was at least as large as all nearest neighbor pixel values. All maxima were identified and grown into larger objects by a simple gray-scale region growing technique as follows. Gaussian smoothing (σ=2.0) was applied to the gray-scale image, and a maximum and a minimum pixel value threshold were specified to select a range of acceptable pixel values. The thresholds were defined as

$$G_i^{\max_1} = 1.01 G_i^{\mathrm{UEP}} \tag{1}$$

and

$$G_i^{\min_1} = 0.99 G_i^{\mathrm{UEP}}, \tag{2}$$

where G_i^{UEP} was the pixel value of the ith maximum and $G_i^{\max_1}$ and $G_i^{\min_1}$ were the maximum and minimum pixel value thresholds, respectively. All pixels within a radius of 20 pixels from a maximum location and with a pixel value inside the defined range were considered to be part of the object. This was repeated for all maxima within an image. Figures 4(a)–4(d) show an original gray-scale image and corresponding images with the DWCE objects, the local maxima, and the gray-scale region-grown objects highlighted. The expanded objects were used as seeds for the gradient-based region growing, described below.

2. Gradient images

A mammogram at 200 μm resolution was used in the gradient-based region-growing stage. The 200 μm resolution image was obtained by averaging 2×2 pixels from the original image. The reduced resolution image had to be smoothed again before gradient filtering because the mammographic tissue produced gradients not only within individual breast structures but also throughout the background portions of the image. Figure 5(b) shows the gradient magnitude image resulting from vertical and horizontal Sobel filtering applied to the 200 μm gray-scale image shown in Fig. 5(a). It clearly demonstrates the large number of gradients throughout the image and the difficulty in applying object-based region growing without additional smoothing. For our application, the smoothing needed to reduce the spurious gradients was accomplished by frequency-weighted Gaussian (FWG) filtering. Frequency-weighted filtering is a technique in which all pixels within the image are split into a base and a residual term. The residual is either positive or negative. This technique produces three subimages from an original image, F, where

$$F = F_F + F_{\mathrm{sub}^+} + F_{\mathrm{sub}^-}. \tag{3}$$

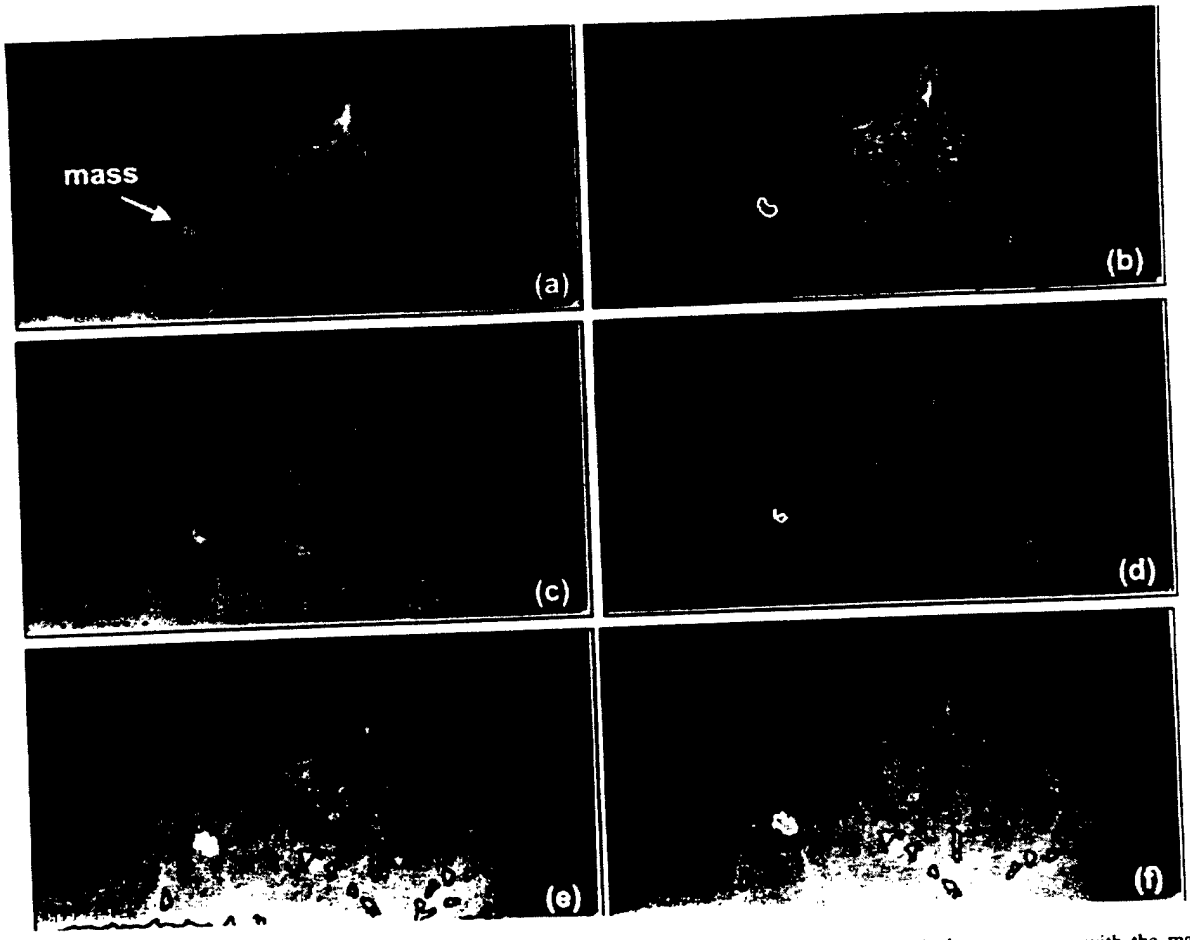


FIG. 4. Objects produced by each segmentation step for a typical mammogram from our data set: (a) the original mammogram with the mass location identified, (b) the DWCE objects, (c) the local maxima, (d) the objects obtained with gray-scale region growing, (e) the objects obtained with gradient-based region growing, and (f) the objects remaining after morphological FP reduction

The first filter component, F_F, is a filtered version of the original image. In our case, a Gaussian filter, $G(\mu=0,\sigma=10)$, was used. The second and third images are the positive and negative residual images of $F-F_F$, respectively. The F_{sub^+} residual is nonzero where the image intensity is larger than the local background and F_{sub^-} is nonzero where the image intensity is smaller than the local background. For a particular image pixel, (x,y), the residual images are defined as

$$F_{sub^+}(x,y)\equiv\begin{cases}F(x,y)-F_F(x,y), & F(x,y)>F_F(x,y),\\ 0, & \text{otherwise.}\end{cases} \tag{4}$$

and

$$F_{sub^-}(x,y)\equiv\begin{cases}F(x,y)-F_F(x,y), & F(x,y)<F_F(x,y),\\ 0, & \text{otherwise.}\end{cases} \tag{5}$$

Two FWG filters were designed for sequentially processing the mammograms. The first FWG filtering step reduced the gradients within the breast structures and produced an intermediate image, F_1, which had the form

$$F_1(F)=\tfrac{3}{4}F_F(F)+\tfrac{1}{4}F_{sub^+}(F), \tag{6}$$

where the F_F and F_{sub^+} images were derived from F, the original 200 μm resolution gray-scale image. A second FWG filtering step was used to eliminate gradients in the breast background. It produced image F_2, which had the form

$$F_2(F_1)=F_{sub^+}(F_1), \tag{7}$$

where the F_{sub^+} image was derived from image F_1. The result of applying the two FWG filters to the original mammogram in Fig. 5(a) is shown in Fig 5(c). In this image, a significant amount of background has been eliminated and the gradients in the remaining structures have been reduced. Horizontal and vertical Sobel filters[15] were then applied to image F_2 and the magnitude calculated to produce a gradient image as shown in Fig. 5(d). Finally, 5×5 median filtering was used to produce the final gradient image shown in Fig. 5(e). This image was used in the gradient-based region-growing step.

3. Final gradient-based region growing

Each initially grown object (described in Sec. II C 1) was again grown by applying an adaptive technique to the gradient image, F_2, described in Sec. II C 2. The region-growing technique was based on the work of Chang and Li[16] and their adaptive homogeneity test for determining the similarity be-

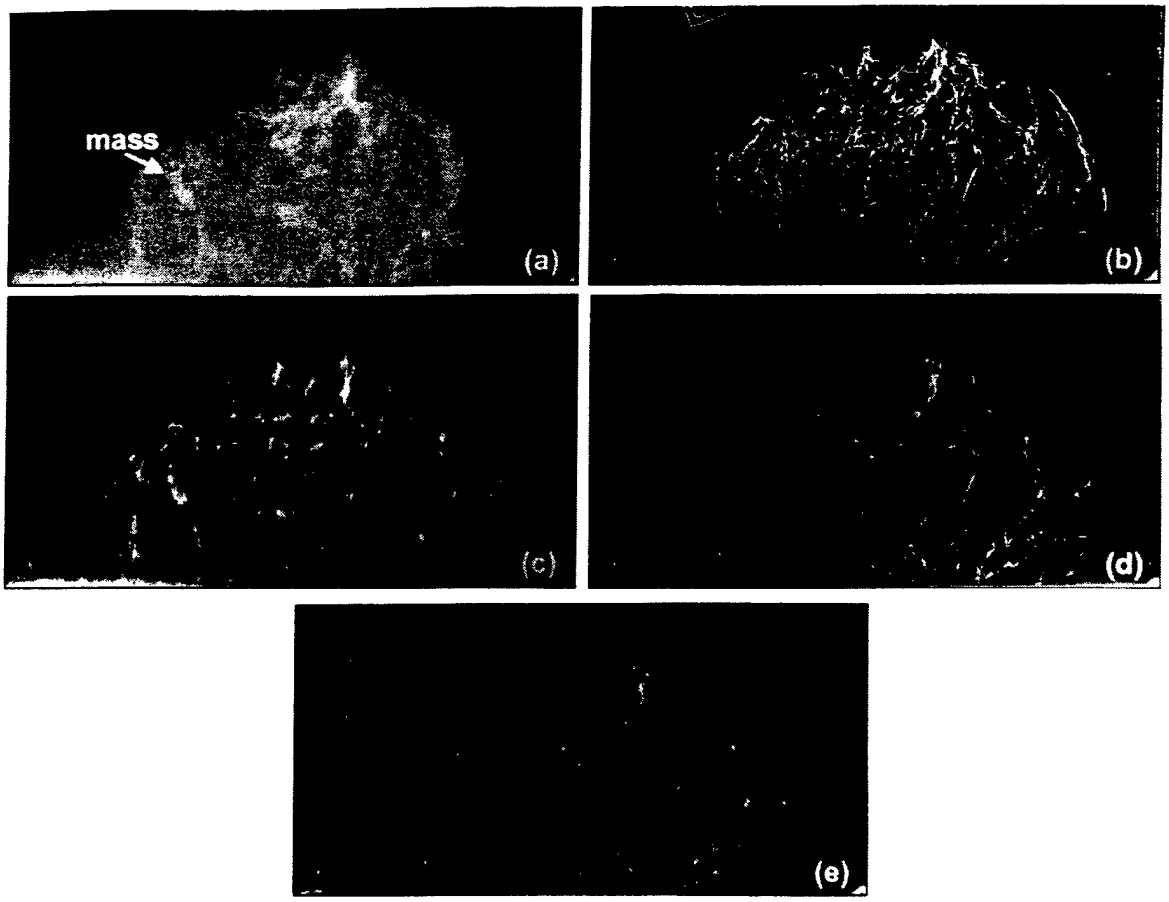


FIG. 5. Processing steps used to define the gradient images: (a) the original mammogram with the mass location identified; (b) the gradient magnitude image obtained from horizontal and vertical Sobel filtering of the original mammogram; (c) the image resulting from FWG filtering of the original mammogram; (d) the gradient magnitude image resulting from horizontal and vertical Sobel filtering of the FWG image; and (e) the image resulting from median filtering of the gradient magnitude image.

tween regions. We have modified this technique to perform object-based region growing. For a mammogram, the corresponding gradient image was smoothed using a Gaussian filter ($\sigma=2.0$). A cumulative distribution function (CDF) of pixel values was then calculated from the smoothed gradient image for each object. For each object, the pixel value thresholds were defined as

$$G_{i,0}^{\max_F} = \{g : \mathrm{CDF}_{i,0}(g) = 1.0\} \tag{8}$$

and

$$G_{i,0}^{\min_F} = \{g : \mathrm{CDF}_{i,0}(g) = 0.0\}, \tag{9}$$

where g was a pixel value and $\mathrm{CDF}_{i,0}(g)$ was the cumulative pixel value distribution within the border of object i and for initial growing iteration 0. The initial growing thresholds simply correspond to the maximum and minimum pixel values within an object. Single-pixel growing was performed on all objects using the thresholds for each individual object to define a range of acceptable pixel values. In this context, single-pixel growing meant growing was limited to only those pixels directly connected to the initial border. Once single-pixel growing was applied to all objects within the image, the thresholds were adjusted and a second iteration of growing was performed. Iterative single-pixel growing was employed to limit the influence of the order that objects were grown within an image. The thresholds used for the ith object during the jth growing iteration were defined as

$$G_{i,j}^{\max_F} = \{g : \mathrm{CDF}_{i,j}(g) = 1.0\} \tag{10}$$

and

$$G_{i,j}^{\max_F} = \left\{ g : \mathrm{CDF}_{i,j}(g) = \frac{j}{30} \right\}, \tag{11}$$

where $\mathrm{CDF}_{i,j}(g)$ was the cumulative pixel value distribution from the smoothed gradient image within the current borders of object i. Single pixel growing was applied to all objects within the image. This iterative procedure was repeated until no more connected pixels had a value within the appropriately defined range. Note that neighboring objects were not allowed to merge together during this region-growing stage so that growing between adjacent objects stopped with at least a one pixel gap between them. Figures 4(d) and 4(e) show the initial seed objects and the final gradient grown objects for the example shown in Fig. 4(a).

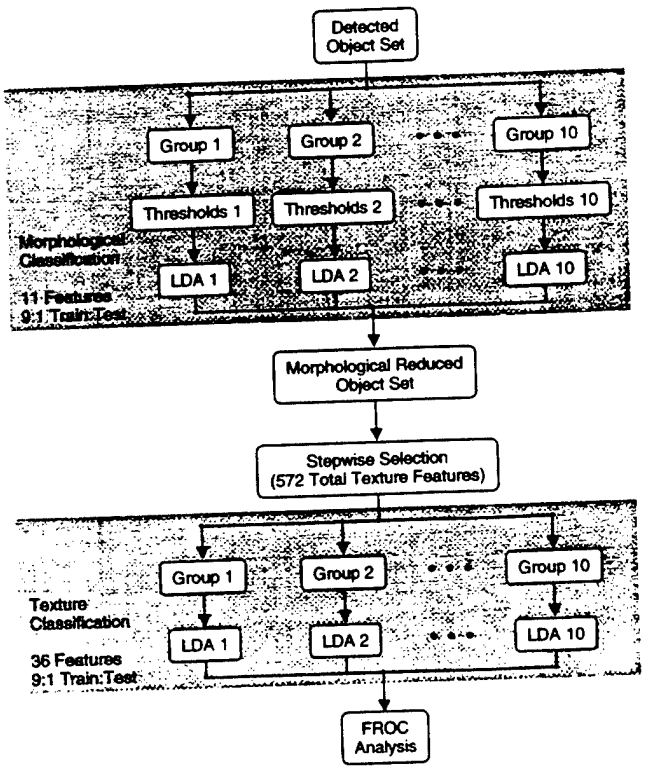


FIG. 6. Flowchart of the FP reduction scheme. The images were separated into ten independent groups. Each group underwent morphological FP reduction with the nine other groups used for classifier training. The reduced objects were recombined and stepwise feature selection was performed. The images were again separated into the ten groups and each group underwent LDA texture classification again using the nine other groups for classifier training. All test scores were then recombined and final FROC analysis was performed

D. False positive reduction

The DWCE segmentation and region growing do not differentiate masses from normal tissues, therefore, a large number of breast structures were usually detected in each mammogram. Since the shape and texture of mass objects, in general, should be different from those of normal breast structures, a set of features was extracted from each detected object and used to differentiate between the detected structures. The feature set included both morphological and texture features. These features were then used in a sequential classification scheme to reduce the number of FP detections in the mammograms. The sequential application of different classifiers has been found to increase classification accuracy,[17] and it also allows more computationally intensive classifiers to be applied to as few objects as possible. A flow chart depicting the general approach employed for FP reduction is shown in Fig. 6. In this study, morphological classification was initially used to eliminate objects that had shapes significantly different from breast masses. Texture features were then computed for all remaining objects and used with a linear classifier as a final arbiter between masses and normal structures. The following sections describe the major components of the FP reduction scheme.

1. Morphological feature-based FP reduction

The mammograms were partitioned into a number of different groups so that the morphological classifiers could be trained and tested to differentiate masses from normal structures. In this study, the 253 mammograms were randomly partitioned into ten independent groups. Each mammogram was allowed to appear in only one group, and all images from the same patient were grouped together. The goal of the partitioning was to have approximately the same number of images in each group under the given constraints. Classification of the objects within each individual group was performed with a classifier trained using the objects from the nine other image groups. This allowed an approximate 9:1 training-to-test ratio for morphological classification. By rotating the test group through all ten image sets, each mammogram served as a test case once.

Eleven morphological features were used in the initial differentiation of the detected structures. These features included the following object-based measures: number of perimeter pixels, area, perimeter-to-area ratio, circularity, rectangularity, and contrast. In addition, five normalized radial length (NRL) features introduced by Kilday *et al.* were also utilized.[18] They included the NRL mean value, standard deviation, entropy, area ratio, and zero-crossing count. The definition for each morphological feature can be found in the literature.[10] They are also included in Appendix A of this paper.

The morphological features were used as input variables for two different classifiers. A simple threshold classifier was followed by a linear discriminant analysis (LDA) classifier in the morphological FP reduction step. The simple threshold classifier set a maximum and minimum value for each morphological feature based on the maximum and minimum feature values found from the breast masses in the data set. The LDA classification was applied to all objects remaining after threshold classification. The LDA classifier is a linear classifier based on Fisher's discriminant, which is optimal for the two-class, multivariate normal, equal covariance problem.[19,20] The LDA classifier was trained for each training set and applied to the appropriate test set. The LDA classifier produced a single discriminant score for each object in the test set. A threshold was defined as the maximum discriminant score of the masses. This threshold was applied to the test set to further differentiate breast masses for normal structures. The threshold was again based on all masses in the data set to ensure that no mass would be lost during this initial stage. Figure 4(f) shows the results of morphological FP reduction for the example depicted in the figure.

2. Texture feature-based FP reduction

Texture-based classification followed the morphological FP reduction. A large set of multiresolution texture features was extracted for each detected object in the mammogram. Stepwise feature selection was then used to choose the most appropriate set of features for linear classification. The selected features were subsequently used with a LDA classifier to produce a single discriminant score for each detected ob-

TABLE I. The number of detected masses and FPs, the single stage reduction, the mean object area (μ_{Area}), and standard deviation of the object areas (σ_{Area}) for the initial stages in the mass detection scheme. Note texture FP reduction followed the morphological FP reduction stage.

Stage	TPs fraction	FPs/image (initial stages)	Reduction	μ_{Area} (mm^2)	σ_{Area} (mm^2)
DWCE	97%	49.1	...	33.6	66.8
Region growing	97%	45.3	0%	52.4	85.1
Morph. FP reduction	97%	35.5	22%	51.9	52.1

ject. The overall performance of the detection scheme was then evaluated with FROC analysis. The texture-based reduction scheme has been documented in the literature; therefore, this paper will only summarize the important components of the texture analysis and point out any differences from the previously described techniques.[10,21,22]

Regions of interest (ROIs) containing each object remaining after morphological FP reduction were extracted from the 100 μm resolution mammograms. The ROIs had a fixed size of 256×256 pixels and the center of each ROI corresponded to the centroid location of a detected object. The only exception was when the object was located near the border of the breast and a complete 256×256 pixel ROI could not be defined. In this case the ROI was shifted until the appropriate edge coincided with the border of the original mammogram.

Global and local multiresolution texture features, based on the spatial gray level dependence (SGLD) matrix,[23,24] were used in texture analysis.[22] An element of the SGLD matrix, $p_{d,\theta}(i,j)$, is defined as the joint probability that gray levels i and j occur at a given interpixel separation d and direction θ. In this study, 13 texture measures were defined for each SGLD matrix. These measures were correlation, energy, entropy, inertia, inverse difference moment, sum average, sum variance, sum entropy, difference average, difference variance, difference entropy, information measure of correlation 1, and information measure of correlation 2. The definition for all texture measures can be found in the literature[22] and are included in Appendix B of this paper.

The wavelet transform with a four-coefficient Daubechies kernel was used to decompose individual ROIs into different scales. For global texture features, four different wavelet scales, 14 different interpixel distances and 2 different angles were used to produce 28 SGLD matrices. This resulted in 364 global multiresolution texture feature for each ROI. To further describe the information specific to the mass and its surrounding normal tissue, a set of local texture features were calculated for each ROI.[10,22,25] Five rectangular subregions were segmented from each ROI; an object subregion defined by the detected object in the center and four peripheral regions at the corners. Eight SGLD (four interpixel distances and two angles) and a total of 208 local features were calculated from the object subregion and the periphery. They included 104 features in the object region and an additional 104 features defined as the difference between the feature values in the object and the periphery.

In order to improve the generalization of the texture classification, stepwise feature selection was used to select a subset of feature from the pool of 572 global and local features. Feature selection was performed using texture features derived from the ROIs obtained from all 253 images. A total of 40 texture features were selected by stepwise feature selection. Details on the application of stepwise feature selection can be found in our previous publications.[21,26]

At this point in texture classification, the mammograms were again divided into the same ten partitions as described in the morphological FP reduction step. Texture classification was performed on each test group with a trained LDA classifier employing the selected features. The training was based on the texture features derived from the ROIs in the nine other image groups. The test scores within each group were combined with the scores from the other groups to form a complete test set of discriminant scores.

The FROC analysis based on the single set of test scores was used to evaluate the overall performance of the segmentation method.[27,28]

III. RESULTS

The number of TP and FP detections found following the DWCE, region-growing, and morphological FP reduction stages of the segmentation algorithm are summarized in Table I. The DWCE segmentation identified 97% of the breast masses. Table I also includes the reduction percentage, the mean object areas (μ_{Area}) and the standard deviations in the object areas (σ_{Area}) for these initial stages. Table II summarizes the mass type, mass size, mass subtlety, and the

TABLE II. The mass type, mass size, mass subtlety, and mammographic tissue density for the mammograms where the mass was not identified by the initial segmentation. In the table, B dentifies a benign lesion, M identifies a malignant lesion, the subtlety is on a scale of 1 (obvious) to 5 (subtle), and breast density uses the BIRADS density scale of 1 (fatty) to 4 (dense). Both the subtlety and density rankings were performed by an experienced breast radiologist.

Mass no.	Type	Size (mm)	Subtlety	Breast density
1	M	6	4	1
2	B	10	2	1
3	B	14	2	2
4	B	10	2	3
5	B	10	2	3
6	B	14	2	3
7	B	12	4	4
Average		10.9	2.6	2.4

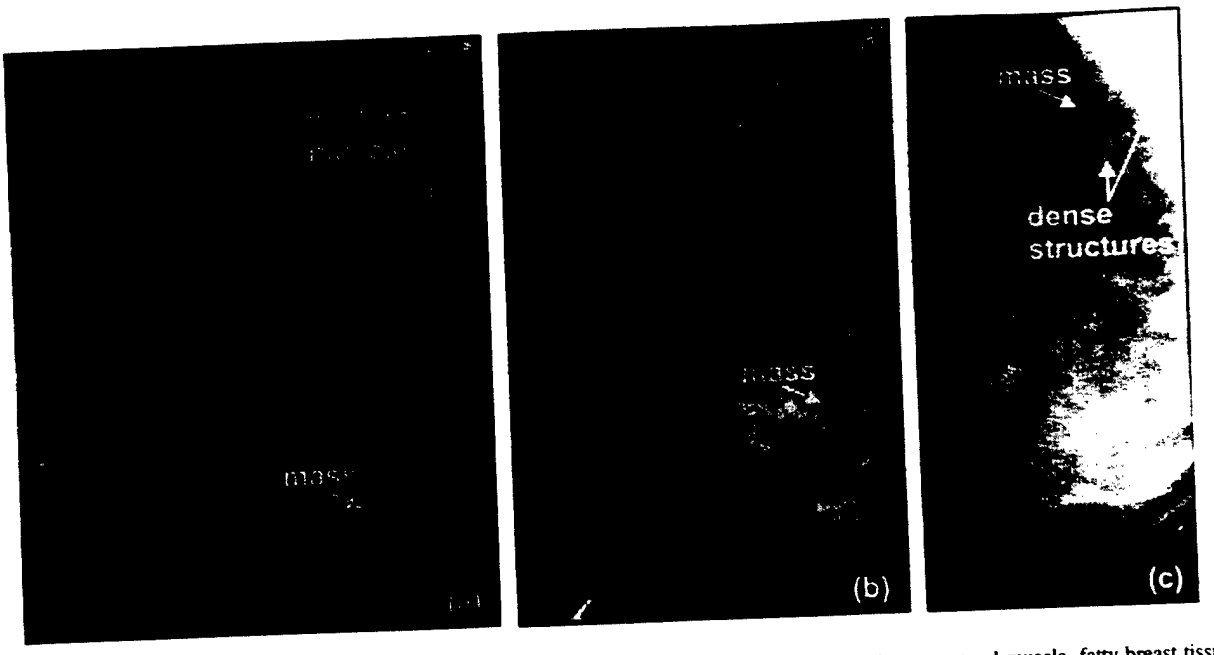


FIG. 7. Examples of masses missed during the initial DWCE segmentation stage: (a) a mammogram with a dense pectoral muscle, fatty breast tissue, and a subtle malignant mass (mass 1 in Table II); (b) a mammogram containing a low contrast benign mass (mass 3 in Table II); and (c) a mammogram with dense structures next to a lower contrast benign mass (mass 4 in Table II).

overall mammographic tissue density for the seven masses missed during the initial DWCE segmentation stage. Figure 7 shows examples of the cases where the mass was missed during the DWCE stage. Figure 8 shows example images with corresponding gradient and object images for cases that had problems during the region-growing stage. This figure contains an example where the mass stopped growing before it reach the correct edge, and an example where the mass was

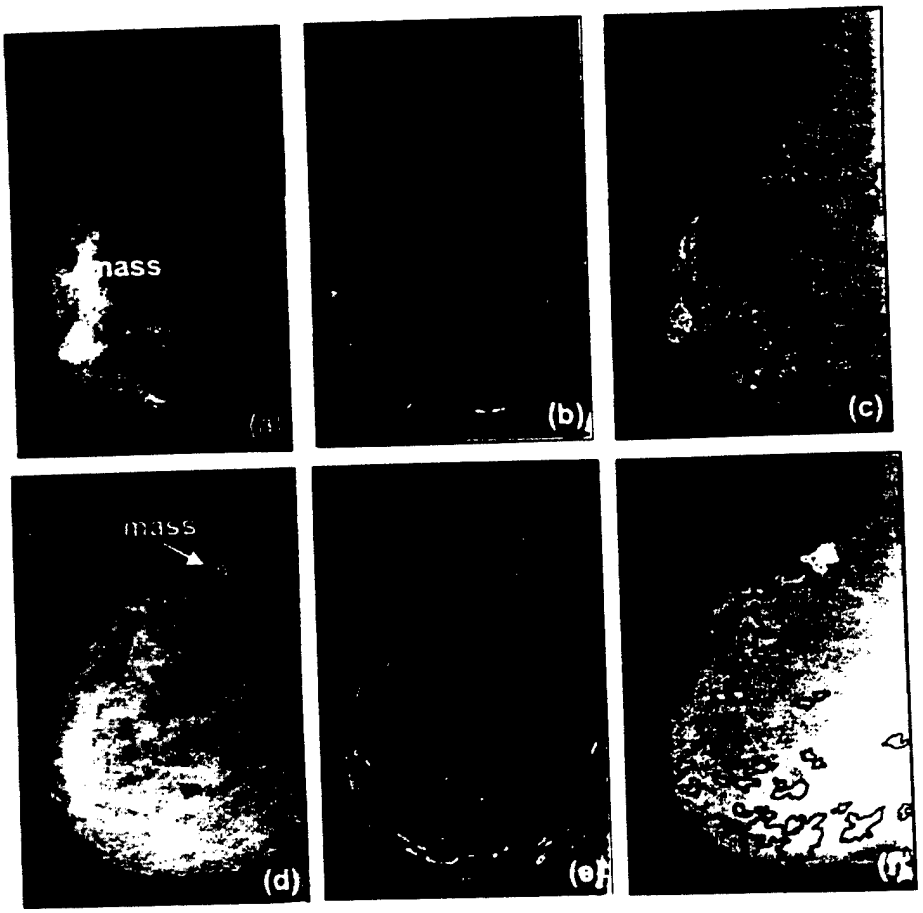


FIG. 8. A mammographic case containing a mass that stopped growing before it reached the correct edge (a)–(c) and a case containing a mass that was split into two pieces during growing (d)–(f). This figure includes (a) and (d) the original mammograms with the mass locations identified, (b) and (e) the corresponding gradient images, and (c) and (f) the final grown objects.

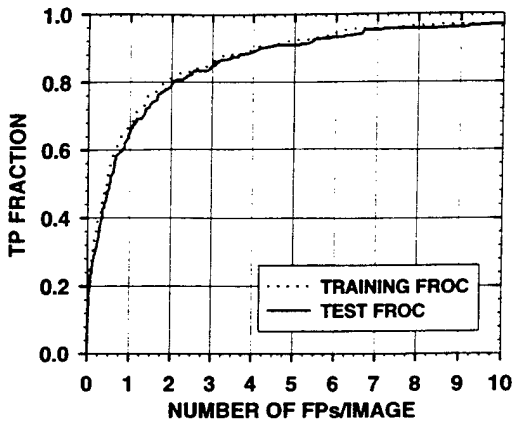


FIG. 9. The training and test FROC curve obtained following LDA classification using 40 selected texture features. The training scores were obtained by averaging the nine training scores from each detected object. The FROC data points were obtained by varying the discriminant decision threshold from the maximum to the minimum value.

split into two pieces during region growing. Finally, Fig. 9 show the FROC training and test performance for the complete segmentation scheme. A summary of the overall performance is given in Tables III and IV for a number of different TP detection fractions. The test performance for the combined DWCE and region-growing segmentation technique at a 90% TP detection level was 4.2 FPs per image and 2.0 FPs per image at an 80% TP level.

IV. DISCUSSION

The purpose of the initial DWCE segmentation stage was to have a method sensitive enough to identify breast masses but which also limited the number of normal structures detected. We have found the DWCE segmentation to be effective in this task. In this study, DWCE segmentation identified 246 of the 253 (97%) masses in the images. Table II summarizes the properties of the masses missed in DWCE segmentation. Masses 1 and 2 were missed because of a dense pectoral muscle visible on the mammogram which overwhelmed all lower-density structures (i.e., both mammograms had BIRADS category 1 breast density). The dense pectoral muscle caused the lower level of the DWCE intensity range to be set so high that lower intensity structures were missed. Figure 7(a) shows the mammogram of the missed malignant mass (mass 1 from Table II). The pectoral muscle is much denser than the mass. This led to the miss. One possible method for eliminating this type of miss may be to identify the pectoral muscle in the mammogram and to apply DWCE segmentation to only the remaining breast region. Mass 3 in Table II was missed because of the small contrast difference between the mass and the background tissue even though the mass was not particularly small or subtle. The mammogram containing this mass is depicted in Fig. 7(b). The remaining masses were missed in mammograms containing denser breast tissue. It was observed that DWCE segmentation had problems detecting masses that were located near much denser normal structures. The dense structures were detected but the masses were missed. Figure 7(c) shows an example of this type of miss. It shows the mammogram containing mass 4 from Table II. Again the dense pectoral muscle may have also hindered detection of the mass in this case. Other than these problems, the DWCE segmentation performed reasonable well as a first stage in mass segmentation. It could identify the majority of the masses while eliminating many of the lower contrast background structures. However, the DWCE segmentation usually underestimated the actual borders of most structures. It also had a tendency to merge the mass with neighboring structures that may have had some tissue overlap with the breast mass. A total of 48 masses had significant merging between the mass and adjacent tissues after DWCE segmentation. This limited the effectiveness of the morphological FP reduction step and limited the localization of the mass during texture-based classification.

The region-growing stage reduced the effects of object merging and significantly increased the size of the initial DWCE objects. This is clearly shown in Table I where the average size of a structure increases from 33.6 mm^2 with DWCE alone to 52.4 mm^2 following region growing. Likewise, a comparison of objects from Figs. 4(b) and 4(e) shows the improvement in border definition following region growing. A combination of gray-scale and gradient-based region growing was used because of the difficulty in stopping gray-scale region growing at the correct edge and the need for large seed objects in gradient-based region growing. The combination approach performed adequately in our detection task and led to an improvement in both morphological and texture-based FP reduction. However, some problems were observed. One problem was that small and low-contrast structures had a tendency to grow into the background and become large regions even though the actual structures were quite small. This did not occur with masses, but it did occur with other breast structures. Another problem was that structures containing internal gradients did not always grow to the correct border, but ended up containing only a section of the true object. This occurred to some mass objects and led to either inaccurate structural information or a mass being split into multiple pieces. Figure 8 shows an example of both incomplete growing and a mass split into pieces during region growing. While these problems reduced the effectiveness of the morphological FP reduction, we have found that the overall benefit of region growing outweights its drawbacks and leads to an improvement in detection accuracy with our segmentation scheme.

The final step in the segmentation was FP reduction. Morphological feature classification was performed first in our reduction scheme. The morphological classification reduced the number of FPs per image from 45.3 to 35.5 as shown in Table I. Following morphological reduction, the average size of the objects was similar to the average size before reduction, but the standard deviation in object size fell from 85.1 mm^2 before reduction to 52.1 mm^2 after reduction. This indicates that morphological reduction eliminated objects that were either much larger or much smaller than the average object size, but had trouble differentiating between TPs and

TABLE III. Summary of the training FROC result depicted in Fig. 9. The table contains the number of FPs per image for different TP fractions along with the percentage of FPs reduced at each TP level relative to the initial value of 19.4 FPs per image. The first entry in the table is the reduction achieved without missing any additional breast masses.

TP fraction	FPs/image	FP reduction
98%	19.4	0%
95%	6.1	69%
90%	4.0	79%
80%	1.9	90%

FPs of similar sizes. Therefore, a classifier that can better differentiate between these similar shaped objects was still necessary. This was achieved, to a large extent, with texture-based feature classification.

A LDA, classifier based on SGLD texture features extracted from ROIs defined by each detected object has proven to be effective in differentiating between similar shaped objects. The training and test FROC performance curves following final texture classification are shown in Fig. 9. In addition, the number of FPs per image for different TP fractions are given in Tables III and IV for the two curves. As discussed in the Methods section, the mammograms were divided into ten independent groups and a 9:1 training-to-test ratio was employed in the classification. Therefore, the test value for an object was its single testing score, and its training value was the average of the scores obtained for the object during training with the nine different training group combinations. The first point to note in Tables III and IV is that the initial TP detection fraction has increased from 97% in Table I to 98% (i.e., 247 total masses were detected). This is due to the change in the definition of a TP with the texture ROIs. The additional mass was detected because in one of the seven mammograms where no object contained the mass centroid, an object ROI overlapped with at least 50% of the mass. The texture classification was able to reduce the number of FPs per image from an initial value of 35.5 to approximately 19 without the loss of any TPs, achieving a 45% reduction. While the number of FPs is still large, it indicates that the more computationally intensive texture classification performs better than morphological reduction. Additional reduction in FPs can be achieved with lower TP detection thresholds. For example, at a 90% TP fraction the FPs decreased to 4.2 per image and at an 80% TP level the FPs decreased to 2.0 per image. Comparing with our previously reported two-stage DWCE edge detection segmentation technique[10] (discussed in Sec. I), we obtained improved performance at all TP levels despite the fact that the data set was increased from 168 to 253 mammograms and two fewer FP reduction stages were used with the new segmentation technique.

The results presented in this paper do not reflect results from a completely independent test set because the feature selection and the selection of morphological classification thresholds were based on the entire image set. This was necessary to obtain the best possible mass statistics from our limited data set at the intermediate stages of the algorithm. A database is currently being collected so that completely independent testing can be performed using the proposed method.

V. CONCLUSION

We have reported on an improved version of a breast mass detection scheme. The scheme employs DWCE segmentation and object-based region growing. Its overall performance has achieved a 90% TP detection level with 4.2 FPs per image and an 80% TP detection level with 2.0 FPs per image with a diverse database of 253 mammograms. The addition of region growing improved the borders of the detected objects and reduced merging between adjacent or overlapping structures. This improved the morphological information extracted from the detected breast masses and thus the differentiation between masses and normal tissues. The FP reduction was also simplified to a single stage of morphological feature classification and a single stage of SGLD texture feature classification. It is expected that a simplified FP reduction scheme has the potential to generalize better than a more complicated scheme when CAD is implemented in a clinical setting. This breast mass segmentation scheme provided improved FROC performance compared to our previously reported two-stage DWCE technique. Further investigations are under way to improve the region-growing segmentation by analyzing different growing methods that may improve the border definition of the detected structures, as well as to develop new object features that may further differentiate masses from normal structures. Preclinical testing of this algorithm on a large set of independent mammograms will also be conducted.

TABLE IV. Summary of the test FROC result depicted in Fig. 9. The table contains the number of FPs per image for different TP fractions along with the percentage of FPs reduced at each TP level relative to the initial value of 19.2 FPs per image. The first entry in the table is the reduction achieved without missing any additional breast masses.

TP fraction	FPs/image	FP reduction
98%	19.2	0%
95%	6.7	65%
90%	4.2	78%
80%	2.0	90%

ACKNOWLEDGMENTS

This work is supported by the Whitaker Foundation (NP), USPHS Grant No. CA 48129, a Career Development Award DAMD 17-96-1-6012 (BS), and research grant DAMD 17-96-1-6254 from the U.S. Army Medical Research and Materiel Command. The content of this publication does not necessarily reflect the position of the government, and no official endorsement of any equipment or product should be inferred.

APPENDIX A: MORPHOLOGICAL FEATURE DEFINITIONS

A set of 11 features is used in morphological FP reduction. Ten of these features are based solely on the binary object defined by the segmentation. The other feature utilizes the original gray scale values inside and surrounding the segmented object. An individual object segmented from image $F(x,y)$ is defined as:

$$F_{\text{obj}_i}(x,y)=\begin{cases}1, & (x,y) \text{ is a pixel in object } i,\\ 0, & \text{otherwise.}\end{cases} \tag{A1}$$

In addition, $F_{BB_i}(x,y)$ defines the pixels contained in the smallest bounding box completely containing object i and $F_{\text{Eqv}_i}(x,y)$ defines the pixels of the circle with the same area as F_{obj_i} and centered at its centroid location. The radius of $F_{\text{Eqv}_i}(x,y)$ is given by

$$r_{\text{Eqv}}=\sqrt{\frac{\text{area }(F_{\text{obj}_i})}{\pi}}. \tag{A2}$$

Five features are based on the normalized radial length (NRL), defined as the Euclidean distance from an object's centroid to each of its edge pixels and normalized relative to the maximum radial length for the object.[18] This results in a NRL vector for each object i given as

$$\mathbf{R}_i=\{r_{i,j}:0\leq j\leq N_e-1\}, \tag{A3}$$

where N_e is the number of edge pixels in the object and $r_{i,j}\leq 1$. The histogram of the normalized radial length is also calculated and is given by

$$\mathbf{P}_i=\{\text{prob}_{i,j}:0\leq j\leq N_h-1\}, \tag{A4}$$

where N_h is the number of bins used in the histogram. Using these basic definitions, the morphological features are defined as follows. Perimeter:

$$\text{Perim}_i=\sum_{\forall x,\forall y} p_i(x,y), \tag{A5}$$

where

$$p_i(x,y)=\begin{cases}1, & F_{\text{obj}_i}(x,y) \text{ is an edge pixel of object } i,\\ 0, & \text{otherwise.}\end{cases}$$

Area:

$$\text{Area}_i=\sum_{\forall x,\forall y} F_{\text{obj}_i}(x,y). \tag{A6}$$

Perimeter-to-area ratio:

$$\text{PAR}_i=\frac{\text{Perim}_i}{\text{Area}_i}. \tag{A7}$$

Circularity:

$$\text{Circ}_i=\frac{\Sigma_{\forall x,\forall y} F_{\text{obj}_i}\cap F_{\text{Eqv}_i}}{\text{Area}_i}. \tag{A8}$$

Rectangularity:

$$\text{Rect}_i=\frac{\text{Area}_i}{\Sigma_{\forall x,\forall y} F_{BB_i}}. \tag{A9}$$

NRL mean:

$$\mu_{\text{NRL}_i}=\frac{1}{N_e}\sum_{j=0}^{N_e-1} r_{i,j}. \tag{A10}$$

NRL standard deviation:

$$\sigma_{\text{NRL}_i}=\sqrt{\frac{1}{N_e}\sum_{j=0}^{N_e-1}(r_{i,j}-\mu_{\text{NRL}_i})^2}. \tag{A11}$$

NRL entropy:

$$E_{\text{NRL}_i}=-\sum_{j=0}^{N_h-1}\text{prob}_{i,j}\cdot\log_2(\text{prob}_{i,j}). \tag{A12}$$

NRL area ratio:

$$\text{AreaR}_i=\left\{\frac{1}{N_e\mu_{\text{NRL}_i}}\sum_{j=0}^{N_e-1}(r_{i,j}-\mu_{\text{NRL}_i}):r_{i,j}>\mu_{\text{NRL}_i}\right\}. \tag{A13}$$

NRL zero-crossing count:

$$\text{ZCC}_i=\sum_{j=0}^{N_e-1} z_{i,j}, \tag{A14}$$

where

$$z_{i,j}=\begin{cases}1, & (r_{i,j-1}>\mu_{\text{NRL}_i})\cap(r_{i,j+1}<\mu_{\text{NRL}_i}),\\ 1, & (r_{i,j-1}<\mu_{\text{NRL}_i})\cap(r_{i,j+1}>\mu_{\text{NRL}_i}),\\ 0, & \text{otherwise.}\end{cases}$$

Contrast:

$$\text{Cont}_i=\frac{g_{\text{in}_i}}{g_{\text{out}_i}}, \tag{A15}$$

where g_{in_i} is the average gray value inside object i and g_{out_i} is the average gray value of the one-pixel wide background surrounding the object.

APPENDIX B: SGLD TEXTURE FEATURE DEFINITIONS

Global and local multiresolution texture features are based on the spatial gray level dependence (SGLD) matrix.[22–24] An element of the SGLD matrix, $p_{d,\theta}(i,j)$, is defined as the joint probability that gray levels i and j occur at a given interpixel separation d and direction θ. In this study, n is defined as the number of gray levels in an image. A total of 13 different texture measures were defined for each SGLD matrix. They were defined as follows.[22]

Energy:

$$E=\sum_{i=0}^{n-1}\sum_{j=0}^{n-1} p_{d,\theta}^2(i,j). \tag{B1}$$

Correlation:

$$R=\frac{\sum_{i=0}^{n-1}\sum_{j=0}^{n-1}(i-\mu_x)(j-\mu_y)p_{d,\theta}(i,j)}{\sigma_x\sigma_y}, \tag{B2}$$

where

$$\mu_x=\sum_{i=0}^{n-1}\sum_{j=0}^{n-1} ip_{d,\theta}(i,j), \tag{B3}$$

$$\mu_y=\sum_{i=0}^{n-1}\sum_{j=0}^{n-1} jp_{d,\theta}(i,j), \tag{B4}$$

$$\sigma_x=\sqrt{\sum_{i=0}^{n-1}\sum_{j=0}^{n-1}(i-\mu_x)^2p_{d,\theta}(i,j)}, \tag{B5}$$

and

$$\sigma_y=\sqrt{\sum_{i=0}^{n-1}\sum_{j=0}^{n-1}(j-\mu_y)^2p_{d,\theta}(i,j)}. \tag{B6}$$

Entropy:

$$H=-\sum_{i=0}^{n-1}\sum_{j=0}^{n-1} p_{d,\theta}(i,j)\log_2(p_{d,\theta}(i,j)). \tag{B7}$$

Inertia:

$$\text{In}=\sum_{i=0}^{n-1}\sum_{j=0}^{n-1}(i-j)^2p_{d,\theta}(i,j). \tag{B8}$$

Inverse difference moment:

$$\text{IDM}=\sum_{i=0}^{n-1}\sum_{j=0}^{n-1}\frac{1}{1+(i-j)^2}p_{d,\theta}(i,j). \tag{B9}$$

Sum average:

$$\mu_{x+y}=\sum_{k=0}^{2n-2} kp_{x+y}(k), \tag{B10}$$

where

$$p_{x+y}(k)=\sum_{i=0}^{n-1}\sum_{j=0}^{n-1} p_{d,\theta}(i,j), \quad i+j=k \quad \text{and} \quad k=0.....2n-2. \tag{B11}$$

Sum variance:

$$\sigma_{x+y}^2=\sum_{k=0}^{2n-2}(k-\mu_{x+y})^2p_{x+y}(k). \tag{B12}$$

Sum entropy:

$$H_{x+y}=-\sum_{k=0}^{2n-2} p_{x+y}(k)\log_2(p_{x+y}(k)). \tag{B13}$$

Difference average:

$$\mu_{x-y}=\sum_{l=0}^{n-1} lp_{x-y}(l), \tag{B14}$$

where

$$p_{x-y}(l)=\sum_{i=0}^{n-1}\sum_{j=0}^{n-1} p_{d,\theta}(i,j), \quad |i-j|=l \quad \text{and} \quad l=0.....n-1. \tag{B15}$$

Difference variance:

$$\sigma_{x-y}^2=\sum_{l=0}^{n-1}(l-\mu_{x-y})^2p_{x-y}(l). \tag{B16}$$

Difference entropy:

$$H_{x-y}=-\sum_{l=0}^{n-1} p_{x-y}(l)\log_2(p_{x-y}(l)). \tag{B17}$$

Information measure of correlation 1:

$$\text{IMC}_1=\frac{H-H_1}{\max\{H_x,H_y\}}. \tag{B18}$$

Information measure of correlation 2:

$$\text{IMC}_2=\sqrt{1-\exp^{-2(H_2-H)}}, \tag{B19}$$

where

$$H_x=-\sum_{i=0}^{n-1} p_x(i)\log_2(p_x(i)), \tag{B20}$$

$$H_y=-\sum_{j=0}^{n-1} p_y(j)\log_2(p_y(j)), \tag{B21}$$

$$H_1=-\sum_{i=0}^{n-1}\sum_{j=0}^{n-1} p_{d,\theta}(i,j)\log_2(p_x(i)p_y(j)) \tag{B22}$$

and

$$H_2=-\sum_{i=0}^{n-1}\sum_{j=0}^{n-1} p_x(i)p_y(j)\log_2(p_x(i)p_y(j)). \tag{B23}$$

[1] L. Tabar *et al.*, "Reduction in mortality from breast cancer after mass screening with mammography," Lancet **1**, 829–832 (1985).

[2] E. L. Thurfjell, K. A. Lernevall, and A. A. S. Taube, "Benefit of independent double reading in a population-based mammography screening program," Radiology **191**, 241–244 (1994).

[3] C. J. Vyborny and M. L. Giger, "Computer vision and artificial intelligence in mammography," AJR, Am. J. Roentgenol. **162**, 699–708 (1994).

[4] N. Karssemeijer and G. te Brake, "Detection of stellate distortions in mammograms," IEEE Trans. Med. Imaging **15**, 611–619 (1996).

[5] H. Kobatake and Y. Yoshinaga, "Detection of spicules on mammogram based on skeleton analysis," IEEE Trans. Med. Imaging **15**, 235–245 (1996).

[6] W. P. Kegelmeyer, J. M. Pruneda, P. D. Bourland, A. Hillis, M. W. Riggs, and M. L. Nipper, "Computer-aided mammographic screening for spiculated lesions," Radiology **191**, 331–337 (1994).

[7] 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. **18**, 955–963 (1991).

[8] D. Brzakovic, X. M. Luo, and P. Brzakovic, "An approach to automated detection of tumors in mammograms," IEEE Trans. Med. Imaging **9**, 233–241 (1990).

[9] H. D. Li, M. Kallergi, L. P. Clarke, V. K. Jain, and R. A. Clark, "Markov random field for tumor detection in digital ammography," IEEE Trans. Med. Imaging **14**, 565–576 (1995).

[10] 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. **23**, 1685–1696 (1996).

[11]H. Kobatake, H. Ron Jin, Y. Yoshinaga, and S. Nawano, "Computer diagnosis of breast cancer by mammogram processing," Radiologia Diagnostica **35**, 29–33 (1994).

[12]N. Petrick, H. P. Chan, B. Sahiner, and D. Wei, "An adaptive density-weighted contrast enhancement filter for mammographic breast mass detection," IEEE Trans. Med. Imaging **15**, 59–67 (1996).

[13]B. Sahiner, H. P. Chan, N. Petrick, M. A. Helvie, and M. M. Goodsitt, "Computerized characterization of masses on mammograms: The rubber band straightening transform and texture analysis," Med. Phys. **25**, 516–526 (1997).

[14]N. Petrick, H. P. Chan, B. Sahiner, M. A. Helvie, M. M. Goodsitt, and D. D. Adler, "Computer-aided breast mass detection: False positive reduction using breast tissue composition," in *Digital Mammography*, edited by K. Doi, M. Giger, R. Nishikawa, and R. Schmidt (Elsevier, New York, 1996).

[15]J. C. Russ, *The Image Processing Handbook* (CRC, Boca Rato, FL, 1992).

[16]Y. L. Chang and X. Li, "Adaptive image region-growing," IEEE Trans. Image Process. **3**, 868–872 (1994).

[17]L. Xu, A. Krzyzak, and C. Y. Suen, "Methods of combining multiple classifiers and their applications to handwriting recognition," IEEE Trans. Syst. Man Cybern. **22**, 418–435 (1992).

[18]J. Kilday, F. Palmieri, and M. D. Fox, "Classifying mammographic lesions using computer-aided image analysis," IEEE Trans. Med. Imaging **12**, 664–669 (1993).

[19]P. A. Lachenbruch, *Discriminant Analysis* (Hafner, New York, 1975).

[20]R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis* (Wiley, New York, 1973).

[21]D. Wei, H. P. Chan, M. A. Helvie, B. Sahiner, N. Petrick, D. D. Adler, and M. M. Goodsitt, "Classification of mass and normal breast tissue on digital mammograms: Multiresolution texture analysis," Med. Phys. **22**, 1501–1513 (1995).

[22]D. Wei, H. P. Chan, N. Petrick, B. Sahiner, M. A. Helvie, D. D. Adler, and M. M. Goodsitt, "False-positive reduction for detection of masses on digital mammograms: Global and local multiresolution texture analysis," Med. Phys. **24**, 903–914 (1997).

[23]R. M. Haralick, K. Shanmugam, and I. Dinstein, "Texture features for image classification," IEEE Trans. Syst. Man Cybern. **SMC-3**, 610–621 (1973).

[24]R. W. Conners, "Towards a set of statistical features which measure visually perceivable qualities of textures," in *Proceedings of the IEEE Conference on Pattern Recognition and Image Processing*, pp. 382–390 (1979).

[25]D. Wei, H. P. Chan, M. A. Helvie, B. Sahiner, N. Petrick, D. D. Adler, and M. M. Goodsitt, "Multiresolution texture analysis for classification of mass and normal breast tissue on digital mammograms," Proc. SPIE **2434**, 606–611 (1995).

[26]H.-P. Chan, D. Wei, M. A. Helvie, B. Sahiner, D. D. Adler, M. M. Goodsitt, and N. Petrick, "Computer-aided classification of mammographic masses and normal tissue: Linear discriminant analysis in texture feature space," Phys. Med. Biol. **40**, 857–876 (1995).

[27]D. P. Chakraborty, "Maximum likelihood analysis of free-response receiver operating characteristic (FROC) data," Med. Phys. **16**, 561–568 (1989).

[28]D. P. Chakraborty and L. H. L. Winter, "Free-response methodology, Alternate analysis and a new observer-performance experiment," Radiology **174**, 873–881 (1990).

Phys. Med. Biol. 43 (1998) 2853–2871. Printed in the UK PII: S0031-9155(98)88223-7

Design of a high-sensitivity classifier based on a genetic algorithm: application to computer-aided diagnosis

Berkman Sahiner†, Heang-Ping Chan, Nicholas Petrick, Mark A Helvie and Mitchell M Goodsitt

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

Received 6 October 1997

Abstract. A genetic algorithm (GA) based feature selection method was developed for the design of high-sensitivity classifiers, which were tailored to yield high sensitivity with high specificity. The fitness function of the GA was based on the receiver operating characteristic (ROC) partial area index, which is defined as the average specificity above a given sensitivity threshold. The designed GA evolved towards the selection of feature combinations which yielded high specificity in the high-sensitivity region of the ROC curve, regardless of the performance at low sensitivity. This is a desirable quality of a classifier used for breast lesion characterization, since the focus in breast lesion characterization is to diagnose correctly as many benign lesions as possible without missing malignancies. The high-sensitivity classifier, formulated as the Fisher's linear discriminant using GA-selected feature variables, was employed to classify 255 biopsy-proven mammographic masses as malignant or benign. The mammograms were digitized at a pixel size of 0.1 mm × 0.1 mm, and regions of interest (ROIs) containing the biopsied masses were extracted by an experienced radiologist. A recently developed image transformation technique, referred to as the rubber-band straightening transform, was applied to the ROIs. Texture features extracted from the spatial grey-level dependence and run-length statistics matrices of the transformed ROIs were used to distinguish malignant and benign masses. The classification accuracy of the high-sensitivity classifier was compared with that of linear discriminant analysis with stepwise feature selection (LDA_{sfs}). With proper GA training, the ROC partial area of the high-sensitivity classifier above a true-positive fraction of 0.95 was significantly larger than that of LDA_{sfs}, although the latter provided a higher total area (A_z) under the ROC curve. By setting an appropriate decision threshold, the high-sensitivity classifier and LDA_{sfs} correctly identified 61% and 34% of the benign masses respectively without missing any malignant masses. Our results show that the choice of the feature selection technique is important in computer-aided diagnosis, and that the GA may be a useful tool for designing classifiers for lesion characterization.

1. Introduction

Due to its high sensitivity, mammography is usually the first radiological examination used for the early detection of malignant breast lesions. However, the positive predictive value (PPV) of mammographic diagnosis (ratio of the number of malignancies to the total number of biopsy recommendations) is not high. Biopsies performed for mammographically suspicious non-palpable breast masses had PPVs of 20 to 30% in three studies (Hermann *et al* 1987, Hall *et al* 1988, Jacobson and Edeiken 1990). To reduce health-care costs and patient morbidity, it is desirable to increase the PPV of mammographic diagnosis

† Address for correspondence: Department of Radiology, University of Michigan, 1500 E. Medical Center Drive, CGC B2102, Ann Arbor, MI 48109-0904, USA. E-mail address: berki@umich.edu

0031-9155/98/102853+19$19.50 © 1998 IOP Publishing Ltd

while maintaining its sensitivity of cancer detection. Computerized mammographic analysis methods can potentially aid radiologists in achieving this goal.

In recent years, several researchers have developed new techniques for the classification of mammographic masses based on computer-extracted features (Brzakovic *et al* 1990, Kilday *et al* 1993, Huo *et al* 1995, Pohlman *et al* 1996, Rangayyan *et al* 1996, Sahiner *et al* 1996a, 1997, 1998). Kilday *et al* (1993) classified masses using morphological features and patient age. Brzakovic *et al* (1990) classified suspected lesions using their shape and intensity variations. Huo *et al* (1995) developed a technique to quantify the degree of spiculation of a lesion, and classified masses as malignant and benign using these spiculation measures. Pohlman *et al* (1996) developed a region growing algorithm for tumour segmentation, and used features describing the tumour shape for classification. Rangayyan *et al* (1996) used an edge acutance measure extracted from the grey-scale intensity along the normal direction to the mass shape, as well as moments to classify masses. We have developed the rubber-band straightening transform (RBST) for facilitating the extraction of effective texture features, and used the texture features extracted from the transformed image for classification (Sahiner *et al* 1996a, 1997, 1998).

A common characteristic of the above approaches is that the lesion is first segmented from the surrounding tissue, and then features are extracted from the shape and grey-level characteristics of the lesion and the surrounding tissue. The extracted features usually represent a mathematical description of characteristics that are helpful for distinguishing malignant and benign lesions. When several features are extracted for classification, it may be difficult to predict which features or feature combinations will result in more accurate classification. For example, it is known that the borders of malignant masses tend to be more irregular than those of benign masses; therefore, it is expected that the normalized radial lengths (Kilday *et al* 1993) carry useful information about the probability of malignancy of a mass. However, since the normalized radial lengths, and especially the features extracted from them (for example variance and entropy), do not exactly measure irregularity but instead merge information from a combination of border characteristics, it is difficult to predict which feature combination will yield the highest classification accuracy when used in a statistical classifier. It is known that the inclusion of inappropriate features may adversely affect classifier performance, especially when the training set is not sufficiently large (Raudys and Jain 1991, Sahiner *et al* 1996c). Therefore, in many situations, one must face the task of selecting a subset of effective features for classification.

One systematic method for feature selection is linear discriminant analysis with stepwise feature selection (LDA_{sfs}), which has been applied to feature selection problems in computer-aided diagnosis (Chan *et al* 1995, Wei *et al* 1995). LDA_{sfs} is an iterative procedure, where one feature is entered into or removed from the selected feature pool at each step by analysing its effect on a selection criterion. The nature of the stepwise selection procedure makes it imperative that the selection criterion be a statistical distance measure between the two groups to be classified. The Wilks lambda and the Mahalanobis distance are commonly used measures. Genetic algorithm (GA) based feature selection, which is capable of using any numerically computed criterion for its fitness function, is a slower but more versatile method than stepwise feature selection. We have demonstrated that when the GA fitness criterion is related to the area A_z under the receiver operating characteristic (ROC) curve, GA-based feature selection yields slightly more effective features than LDA_{sfs} (Sahiner *et al* 1996c).

In the task of lesion characterization, the cost of missing a malignancy is very high. Therefore, the performance of a classifier in the high-sensitivity (high true-positive fraction) region of the ROC curve is more important than the overall area A_z under the ROC curve. In

other words, if a classifier is to be designed for breast lesion characterization, the specificity at high levels of sensitivity is much more important than the specificity at low levels of sensitivity. Recently, Jiang *et al* (1996) developed a method for describing an ROC partial area index that may be useful as a performance measure in lesion characterization problems. Since a feature (or feature combination) that can provide a large overall A_z (or a large Wilks lambda and Mahalanobis distance) may not provide a large partial ROC area, it is important to develop a feature selection method for the design of high-sensitivity classifiers. The partial ROC area is potentially a good feature selection criterion for this application. The flexibility of a GA in the selection of its fitness function allows this index to be incorporated for feature selection.

In this study, we developed a methodology to design high-sensitivity classifiers. The design process was illustrated by the task of classifying masses on digitized mammograms as malignant or benign. A GA-based algorithm with the ROC partial area index as the feature selection criterion, in combination with Fisher's linear discriminant, was used for the design of this classifier. Texture features extracted from RBST images (Sahiner *et al* 1998) were used for classification. The performance of the high-sensitivity classifier was compared with the performance achieved by LDA_{sfs} using the Wilks lambda as the feature selection criterion.

2. Materials and methods

2.1. Data set

The mammograms used in this study were selected from the files of patients at the Radiology Department of the University of Michigan who had undergone biopsy. The mammograms were acquired with dedicated mammographic systems with 0.3 mm focal spots, molybdenum anodes, 0.03 mm thick molybdenum filters and 5:1 reciprocating grids. For recording the images, a Kodak MinR/MRE screen/film system with extended cycle processing was used. The criterion for inclusion of a mammogram in the data set was that the mammogram contained a biopsy-proven mass, and that approximately equal numbers of malignant and benign masses were present in the data set.

Our data set consisted of 255 mammograms from 104 patients. For most of the patients we had two mammograms in the data set, which were the craniocaudal and the mediolateral oblique views. However, for some of the patients, extra views such as lateral and oblique views were included in the data set. There were 128 mammograms with benign masses, of which 8 were spiculated based upon radiologist interpretation, and 127 mammograms with malignant masses, of which 62 were spiculated. Of the 104 patients evaluated in this study, 48 had malignant masses. The probability of malignancy of the biopsied mass on each mammogram was ranked by a Mammography Quality Standards Act (MQSA) approved radiologist experienced in mammographic interpretation on a scale of 1 to 10. A ranking of 1 corresponded to the masses with the most benign mammographic appearance, and a ranking of 10 corresponded to the masses with the most malignant mammographic appearance. The distribution of the malignancy ranking of the masses is shown in figure 1. The true pathology of the masses was determined by biopsy and histological analysis.

The mammograms in the data set were digitized with a Lumisys DIS-1000 laser scanner at a pixel resolution of 0.1 mm $\times$ 0.1 mm and 4096 grey levels. The digitizer was calibrated so that grey-level values were linearly proportional to the optical density (OD) within the range of 0.1 to 2.8 OD units, with a slope of 0.001 OD/pixel value. Outside this range, the slope of the calibration curve decreased gradually, with the OD range extending to 3.5.

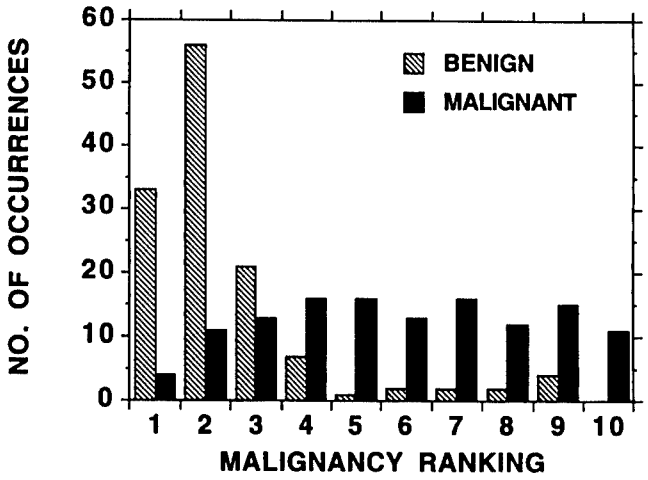


Figure 1. The distribution of the malignancy ranking of the masses in our data set, as determined by a radiologist experienced in mammographic interpretation: 1, very likely benign; 10, very likely malignant.

The pixel values were linearly converted before they were stored on the computer so that a high pixel value represented a low optical density.

The location of the biopsied mass was identified by the radiologist, and a region of interest (ROI) containing the biopsied mass was extracted for computerized analysis. The size of the ROI was allowed to vary according to the lesion size. The extracted ROIs contained a non-uniform background, which depended on the overlapping breast structures and the location of the lesion on the mammogram. The non-uniform background is not related to mass malignancy, but may affect the segmentation and feature extraction results used in our computerized analysis. To reduce the background non-uniformity, an automated background correction technique was applied to each ROI as the very first step in our analysis. Details and examples of our background correction technique can be found in the literature (Sahiner *et al* 1996b).

2.2. *The rubber-band straightening transform (RBST)*

In this study, the classification of malignant and benign masses was based on the textural differences of their mammographic appearance. We have previously designed a rubber-band straightening transform (RBST) which was found to facilitate the extraction of texture features from the region surrounding a mammographic mass. The image transformation performed by the RBST is depicted in figure 2, and a block diagram of different stages of the RBST is given in figure 3. A detailed discussion of the transform can be found in the literature (Sahiner *et al* 1996a, 1997, 1998). For completeness, a brief description is given below.

The RBST transforms a band of pixels surrounding a mass onto the Cartesian plane. The four basic steps in the RBST are mass segmentation, edge enumeration, computation of normals and interpolation. A modified K-means clustering algorithm (Sahiner *et al* 1995) was used for segmentation. The parameters of the segmentation algorithm were chosen so that the segmented region was slightly smaller than the actual size of the mass. After clustering, one to several objects would be segmented in the ROI. If more than one object was segmented, the largest connected object was selected. The selected object

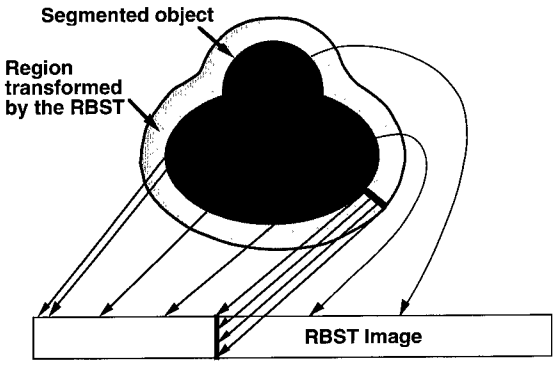


Figure 2. The formation of the RBST image.

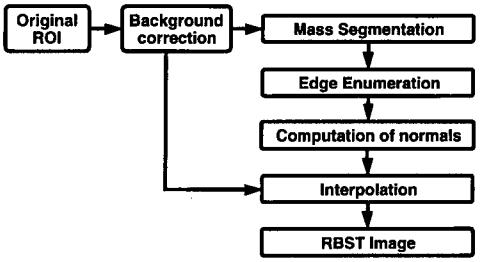


Figure 3. Block diagram of the stages of RBST image computation.

was then filled, grown in a local neighbourhood, and eroded and dilated with morphological operators. The implementation details of these steps have been described elsewhere (Sahiner *et al* 1998). After the outline of the mass was obtained, an edge enumeration algorithm assigned a pixel number to each border pixel of the mass, such that neighbouring pixels were assigned consecutive numbers. The computation of normals depended on the output of the edge enumeration algorithm. The normal $L(i)$ at border pixel i was determined as the normal to the line joining border pixels $i - K$ and $i + K$. The choice of the constant K represents a trade-off between a noisy estimate of the normal direction (small K) and an estimate that misses fine variations in the normal direction (large K). In order to determine the constant K to be used in this study, we selected a small subset of images from our database, and plotted the normal direction obtained by using different values of K superimposed on the segmented image. By performing a visual comparison of the computed normal direction to what was perceived to be the true normal direction, it was empirically found that $K = 12$ resulted in a satisfactory normal estimation. In the interpolation step, the value of the pixel in row j, column i of the RBST image was found as follows. Let $p(i, j)$ denote the location in the original image at a distance j along $L(i)$ from border pixel i. The two closest pixels in the original ROI to location $p(i, j)$ were identified, and the (i, j)th pixel value of the RBST image was defined as the distance-weighted average of these two pixel values.

The width of the band transformed by the RBST was chosen as 40 pixels in this study, which corresponded to 4 mm on the mammogram. An example of the background-corrected ROI, the segmented and morphologically filtered mass shape, and the RBST image are shown in figure 4.

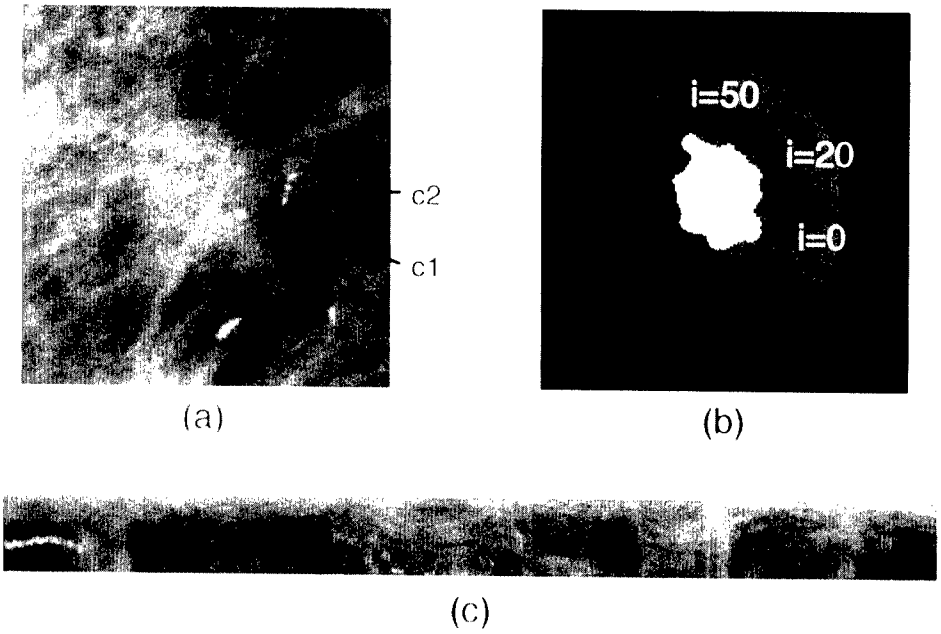


Figure 4. (a) The original mammographic ROI. (b) The segmented and morphologically filtered mass shape (white), and the 40-pixel-wide band around it (grey). For the purpose of illustration, the normals computed at $i = 0$, 20 and 50 are also shown. (c) The RBST image. Notice that due to the position of the first normal location ($i = 0$), the calcifications c1 and c2 on the original ROI appear at the right and the left of the RBST image respectively. The pathological analysis indicated that this was an invasive ductal and intraductal carcinoma.

2.3. Texture features

The texture features used for the classification of the malignant and benign masses were spatial grey-level dependence (SGLD) and run length statistics (RLS) features. These features were extracted from SGLD and RLS matrices, which were constructed from the RBST images as described below.

2.3.1. SGLD features. The (i, j)th element of the SGLD matrix $p_{\theta,d}(i, j)$ represents the probability that grey levels i and j occur at an angle θ and a distance d with respect to each other. The use of SGLD matrices for feature extraction was motivated by the assumption that texture information is contained in the average spatial relationships between the grey-level tones in the image (Haralick *et al* 1973). The features extracted from SGLD matrices of mammographic ROIs have been shown to be useful in classification of mass and normal tissue, and malignant and benign masses or microcalcifications in computer-aided diagnosis (CAD) (Chan *et al* 1995, 1997a, Wei *et al* 1995, Sahiner *et al* 1996b, 1998).

In this study, four different directions ($\theta = 0°$, 45°, 90° and 135°) and ten different pixel pair distances ($d = 1$, 2, 3, 4, 6, 8, 10, 12, 16 and 20) were used for the construction of SGLD matrices from RBST images. The total number of SGLD matrices was therefore 40. Based on our previous studies (Chan *et al* 1995), a bit depth of eight bits was used in the SGLD matrix construction.

A number of SGLD features, which describe the shape of the SGLD matrices, can be extracted from each SGLD matrix. In this study, we extracted eight such features, which

were also used in our previous studies (Chan *et al* 1995, Wei *et al* 1995, Sahiner *et al* 1998). These texture features were correlation, difference entropy, energy, entropy, inertia, inverse difference moment, sum average and sum entropy. This resulted in the computation of 320 SGLD features per RBST image. These features characterize information such as homogeneity, contrast and structural linearity in the images. However, it is difficult to establish a one-to-one correspondence between these qualitative image characteristics and the extracted texture features (Haralick *et al* 1973). The definitions of the SGLD features used in this study can be found in the literature (Haralick *et al* 1973, Chan *et al* 1995, Wei *et al* 1995).

2.3.2. RLS features. The pixels along a given line in an image occasionally contain runs of consecutive pixels that all have the same grey level. A grey-level run is defined as a set of consecutive, collinear pixels in a given direction which have the same grey-level value. A run length is the number of pixels in a grey-level run. The RLS matrix for a given image describes the run length statistics in a given direction for each grey-level value in the image. The (i, j)th element of the RLS matrix $r\theta(i, j)$ represents the number of times that runs of length j in the direction θ consisting of pixels with a grey level i exist in the image (Weszka *et al* 1976).

The RLS matrices in this study were extracted from the vertical and horizontal gradient magnitudes of the RBST images. The vertical and horizontal gradients were obtained by filtering the RBST images with horizontally and vertically oriented Sobel filters (Jain 1989) respectively. Examples of the gradient magnitude images are shown in figure 5. The RLS matrices were obtained from each gradient magnitude image in two directions, $\theta = 0°$ and $\theta = 90°$. Therefore, a total of four RLS matrices were obtained for each RBST image.

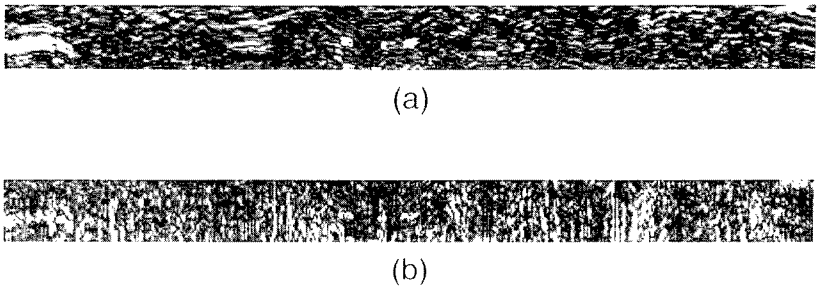


Figure 5. Gradient magnitude images for the RBST image in figure 4: (a) horizontal gradient magnitude image and (b) vertical gradient magnitude image.

Based on our previous study, a bit depth of 5 was used for the computation of RLS matrices (Sahiner *et al* 1998). Five RLS features, namely short runs emphasis, long runs emphasis, grey-level non-uniformity, run length non-uniformity and run percentage were extracted from each RLS matrix. This resulted in the computation of 20 RLS features per RBST image. The definitions of these features can be found in the literature (Galloway 1975). It is possible to describe the general aspects of the relationship between the image characteristics and the RLS feature values. For example, run percentage is low for images with long linear structures, and grey-level non-uniformity is low for images where runs are equally distributed throughout the grey levels (Galloway 1975). However, it is again

difficult to establish a one-to-one correspondence between these texture features and visual image features.

2.4. Fisher's linear discriminant and LDA_{sfs}

For a two-class problem, Fisher's linear discriminant projects the multidimensional feature space onto the real line in such a way that the ratio of between-class sum of squares to within-class sum of squares is maximized after the projection (Duda and Hart 1973). This is the optimal classifier if the features for the two classes have a multivariate Gaussian distribution with equal covariance matrices (Lachenbruch 1975). It has been shown to be a reasonably good classifier even when the feature distributions for the two classes are non-Gaussian (Duda and Hart 1973). Linear discriminant analysis (LDA) is a class of statistical techniques based on Fisher's linear discriminant.

When the training data size is limited, the inclusion of inappropriate features in a classifier may reduce the test accuracy due to overtraining. Therefore, when a large number of features are available for a classification task, it is necessary to select a subset of the most effective features from the feature pool. LDA_{sfs} is a commonly used feature selection method (Lachenbruch 1975). In this study, the performance of a GA-based high-sensitivity feature selection method was compared with that of stepwise feature selection.

Wilks' lambda, which is defined as the ratio of within-group sum of squares to the total sum of squares (Lachenbruch 1975), was used as the selection criterion for the stepwise feature selection method. The stepwise feature selection algorithm starts with no selected features at step 0. At step s of the algorithm, the available features are entered into the selected feature pool one at a time during feature entry, and those already selected are removed one at a time during feature removal. The significance of the change in the Wilks' lambda, as determined by F-statistics, when a new feature is entered into the selected feature pool is compared with a threshold F_{in}. The feature with the highest significance is entered to the selected feature pool only if the significance is higher than F_{in}. Likewise, the significance of the change in the Wilks' lambda when a selected feature is removed from the feature pool is compared with a threshold F_{out}. The feature with the least significance is removed from the selected feature pool only if the significance is lower than F_{out}. This completes step s of the algorithm. The algorithm terminates when no more features can satisfy the criteria for either being added to or removed from the selected feature pool.

2.5. Genetic algorithms for feature selection

Genetic algorithms solve optimization problems by mimicking the natural selection process. A GA follows the evolution of a population of chromosomes which are encoded so that each chromosome corresponds to a possible solution of the optimization problem. The chromosomes consist of genes, which are components of the solution. The goal of a GA is to search for better combinations of the genes, i.e. new chromosomes which are better solutions to the optimization problem. This goal is achieved by evolution. A new generation of chromosomes is produced from the current population by means of parent selection, crossover and mutation. The probability that a chromosome is selected as a parent is related to its ability to solve the optimization problem, i.e. its fitness. Chromosomes which are better solutions to the optimization problem are given a higher chance to reproduce than those which are worse solutions to the problem, similar to the principle of natural selection. The fitness of a chromosome is computed using a fitness function, which is designed on the basis of the optimization criterion for the problem. The probability that a chromosome

is selected as a parent is equal to its normalized fitness, which is defined as the fitness of the chromosome divided by the sum of fitnesses for all chromosomes. The chromosomes of the selected parents are allowed to randomly cross over and mutate, introducing new genes and new chromosomes into the population. This process generates a new population of chromosomes, which tends to evolve towards a better solution.

GAs had been applied to the problem of feature selection (Brill *et al* 1992, Sahiner *et al* 1996c). The most natural way of encoding a chromosome for this problem is as follows (Sahiner *et al* 1996c). Each gene in a chromosome is a bit, which takes a value of either 1 or 0. Each gene location in a chromosome corresponds to a particular feature. If the bit value at a gene location is 1, the corresponding feature is selected for the solution of the classification problem. Otherwise, the corresponding feature is not selected. Each chromosome thus defines a set of selected features. A statistical classifier, such as Fisher's linear classifier or a neural network classifier, is then employed for classification based on the selected feature set. The fitness function reflects the success of the selected feature set for solving the classification problem. The design of the fitness function for a high-sensitivity classifier is described in the next section. The GA training method and the choice of GA parameters are summarized next.

2.5.1. GA training. The GA in this study was trained using a leave-one-case-out paradigm. In this paradigm, all ROIs except those from a particular patient were defined as the training set, and the ROIs from that particular patient were defined as the test set. For each chromosome of the GA, the coefficients of Fisher's linear discriminant function were determined using the features of the training set. The trained discriminant function was then used to classify the test cases using the features of the test cases as the input. In a given generation of the GA, all patients were visited in a round-robin manner, so that test scores were obtained for each ROI in the entire data set. The fitness of the chromosome was computed based on the classification accuracy for the test cases, as described in the next section.

2.5.2. GA parameters. The fundamental parameters of a GA are the number of chromosomes, the chromosome length, the crossover rate, the mutation rate and the stopping criterion. In a GA, the population must contain a large number of chromosomes to provide the variability that offers the opportunity to evolve towards the optimal solution. This requirement and computing speed considerations are trade-offs for selecting the number of chromosomes in a given application. The length of a chromosome is determined by the encoding mechanism which translates the optimization problem into a GA. With the encoding mechanism described earlier in this subsection, the length of each chromosome is equal to the total number of features. The fitness function is the most important component of the GA, and its design is described in the next section. Pairs of chromosomes are probabilistically selected as parents based on their fitness. A selected pair may exchange genes to generate two offspring. The crossover rate determines the probability that parents will exchange genes. After crossover, the binary value of each bit may probabilistically be altered (from 1 to 0, or vice versa), i.e. mutated. The mutation rate determines the probability that genes will undergo mutation. The increase in the fitness of the chromosomes starts to stagnate after a number of generations. The stopping criterion determines when the evolution is terminated. In this study, the GA evolution was terminated after a fixed number of iterations. The appropriateness of this stopping criterion is discussed in section 4. After the termination, the chromosome with the highest fitness value provided the set of selected features.

Table 1 shows the values of each of these parameters, selected based on our previous work. More detailed discussion of these operators and parameters can be found in the literature (Sahiner *et al* 1996c).

Table 1. GA parameters used in this study.

Crossover rate	0.9
Mutation rate	0.0025
Chromosome length	340
Number of chromosomes	200
Stopping criterion	200 iterations

2.6. Design of a high-sensitivity classifier

A widely accepted method for comparing the performance of two classifiers is to consider their ROC curves. The area A_z under the ROC curve is a commonly used index for this comparison. However, for applications where the performance at high sensitivity (or high true-positive fraction) is important, for example breast lesion characterization in CAD, this index may be inadequate. Jiang *et al* (1996) explored this issue, and defined an ROC partial area index that will be denoted as A_{TPF_0} in this paper.

The partial area index A_{TPF_0} summarizes the average specificity above a sensitivity of TPF_0 (figure 6), and can be expressed as (Jiang *et al* 1996)

$$A_{\mathrm{TPF}_0} = 1 - \frac{1}{1-\mathrm{TPF}_0}\int_{\mathrm{TPF}_0}^{1} \mathrm{FPF}(\mathrm{TPF})\,\mathrm{d}(\mathrm{TPF}) \tag{1}$$

which is the ratio of the partial area under the actual ROC curve to the partial area of the perfect ROC curve. The maximum value for A_{TPF_0} is thus 1. The A_{TPF_0} value for a classifier that operates purely on random guessing is $(1-\mathrm{TPF}_0)/2$, which is the area under the chance diagonal normalized to $1-\mathrm{TPF}_0$.

When the conventional binormal model is employed for the computation of the ROC curve, the curve is completely defined by two parameters, a and b, which are determined from the rating data using maximum likelihood estimation. The constant b represents the estimated standard deviation of the actually negative cases, normalized by the estimated standard deviation of the actually positive cases, and the constant a represents the estimated difference between the means of actually positive and negative cases, normalized again by the estimated standard deviation of the actually positive cases. Using the binormality assumption, the partial area index A_{TPF_0} can be expressed as (McClish 1989, Jiang *et al* 1996)

$$A_{\mathrm{TPF}_0} = 1 - \frac{1}{1-\mathrm{TPF}_0}\int_{c_0}^{\infty} \Phi\left(\frac{u-a}{b}\right)\phi(u)\,\mathrm{d}u \tag{2}$$

where

$$\phi(u) = \frac{1}{\sqrt{2\pi}}\exp(-u^2/2)$$

and

$$\Phi(u) = \int_{-\infty}^{u} \phi(x)\,\mathrm{d}x.$$

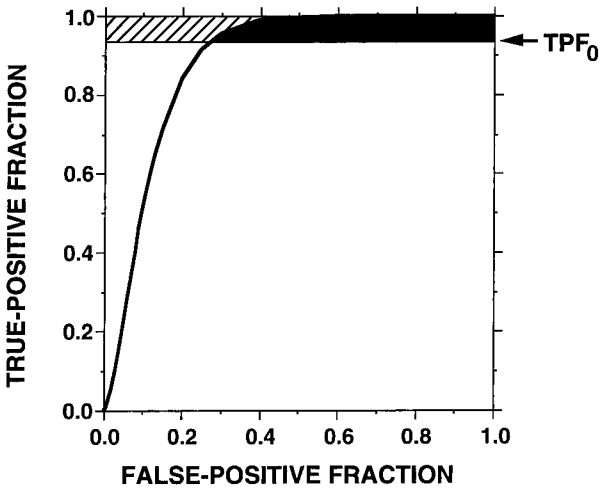


Figure 6. The partial area index A_{TPF_0} is defined as the ratio of the partial area under the ROC curve above a given sensitivity (grey area) to the partial area of the perfect ROC curve (hatched region) above the same sensitivity.

Our goal in this study was to train a GA to select features which would yield high specificity in the high-sensitivity region of the ROC curve. Therefore, the fitness of a chromosome was defined as a monotonic function of A_{TPF_0}, such that the maximization of A_{TPF_0} would maximize the fitness function

$$\text{fitness} = \left(\frac{A_{TPF_0} - A_{min}}{A_{max} - A_{min}}\right)^n \tag{3}$$

where A_{max} and A_{min} were the maximum and minimum values of A_{TPF_0} among all chromosomes in a generation, and n was a power parameter whose effect on GA feature selection was investigated, as discussed in section 3. From equation (3), it is seen that as the power parameter becomes larger the difference in the fitness, and thus the probability of being chosen as parents, between the chromosomes are more amplified. The choice of n is a tradeoff between the goal of promoting chromosomes with high fitness values and the need to retain segments of good genes in other chromosomes.

For a given chromosome, the parameters a and b that are required for the computation of A_{TPF_0} were determined from the distribution of test scores using the LABROC program of Metz *et al* (1998). The partial area index A_{TPF_0} was then computed by numerically integrating equation (2). The classifiers thus designed will be referred to as GA-based high-sensitivity classifiers in the following discussions.

In this study, the significance of the difference in A_{TPF_0} of different classifiers was determined using a recently developed statistical test (Jiang *et al* 1996). The test is analogous to statistical tests involving the area A_z under the entire ROC curve, and is implemented using the covariance estimates of a and b values for the two curves.

3. Results

To demonstrate the training of high-sensitivity classifiers using GA, we chose two levels of sensitivity thresholds, $TPF_0 = 0.50$ and $TPF_0 = 0.95$ in equation (1). The classification results of these classifiers were compared with those of LDA_{sfs}. GA-based feature selection

Table 2. The number of features, the area A_z under the ROC curve, the partial area above the true positive fraction of 0.5 ($A_{0.50}$), and that above 0.95 ($A_{0.95}$) for various values of F_{in} and F_{out} in the stepwise feature selection method.

F_{in}	F_{out}	Number of selected features	A_z	$A_{0.50}$	$A_{0.95}$
3.8	2.7	9	0.84	0.71	0.22
2.6	2.4	13	0.85	0.72	0.27
2.2	2.0	14	0.86	0.73	0.25
1.8	1.6	26	0.89	0.80	0.38
1.4	1.2	41	0.92	0.83	0.47
1.0	1.0	49	0.92	0.83	0.46

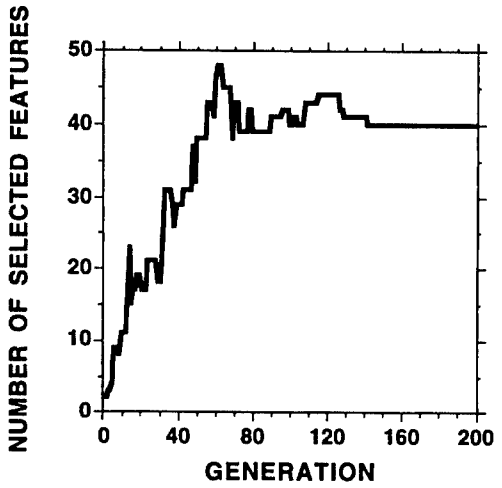


Figure 7. The evolution of the number of selected features for a GA training session ($n = 4$, $TPF_0 = 0.95$).

was also performed with no emphasis on high sensitivity ($TPF_0 = 0$). The classifier designed with the features thus selected will be referred to as an ordinary GA-based classifier. Its performance was compared with those of the GA-based high-sensitivity classifiers and LDA_{sfs}.

In LDA_{sfs}, the optimal values of the F_{in} and F_{out} thresholds are not known *a priori*. We therefore varied these thresholds to obtain the feature subset with the best test performance. Table 2 shows the number of selected features, the area A_z under the ROC curve, the partial area above the true positive fraction of 0.5 ($A_{0.50}$), and that above 0.95 ($A_{0.95}$) as these F thresholds are varied. By comparing the A_z values and the performance at the high-sensitivity portion of the ROC curve, the combination $F_{in} = 1.4$, $F_{out} = 1.2$ was found to provide the best feature subset.

High-sensitivity classifiers with $TPF_0 = 0.50$ and $TPF_0 = 0.95$ were trained with three different values of the power parameter, n ($n = 1$, 2 and 4). Figure 7 shows the evolution of the number of selected features, and figure 8 shows the total area under the ROC curve (A_z) and the partial area above the true positive fraction of 0.95 ($A_{0.95}$) for a typical GA training ($n = 4$, $TPF_0 = 0.95$).

The ROC curve of the best LDA_{sfs} classifier and those of GA-based classifiers ($TPF_0 = 0.50$ and $TPF_0 = 0.95$) with $n = 1$, 2 and 4 are compared in figures 9–11

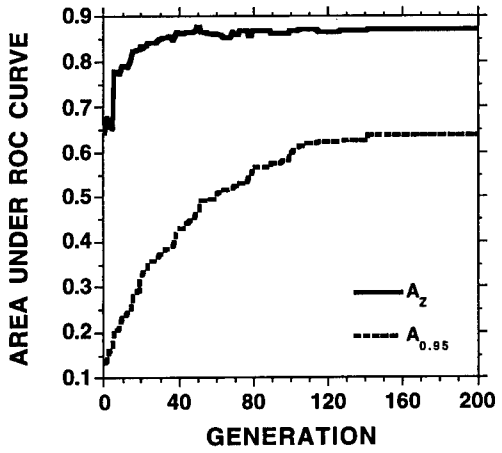


Figure 8. The evolution of the area A_z and the partial area $A_{0.95}$ under the ROC curve for the GA training session of figure 7 ($n = 4$, $TPF_0 = 0.95$).

respectively. It is observed from figures 10 and 11 that for $n = 2$ or 4, the designed high-sensitivity classifiers seem to be superior to the best LDA_{sfs} classifier for large values of true positives. When $n = 1$, the ROC curves of the GA-based high-sensitivity classifiers are still higher than that of the LDA_{sfs} classifier when TPF is very close to 1; however, the difference between the curves is small. To quantify the improvement obtained by the GA-based high-sensitivity classifier, we performed statistical significance tests (Jiang *et al* 1996) on the partial area above a true-positive threshold of 0.95 ($A_{0.95}$) as described in the previous section. With $n = 4$, the difference between the partial areas of the GA-based high-sensitivity classifiers and LDA_{sfs} above a true-positive threshold of 0.95 was statistically significant with two-tailed p-levels of 0.006 and 0.02 for the classifiers trained with $TPF_0 = 0.95$ and $TPF_0 = 0.5$ respectively. For $n = 2$, the corresponding p-levels were 0.01 and 0.07 respectively. For $n = 1$, the difference did not achieve statistical significance ($p = 0.14$ for $TPF_0 = 0.95$ and $p = 0.49$ for $TPF_0 = 0.5$). The difference of the partial area index over a true-positive threshold of 0.5 ($A_{0.50}$) did not achieve statistical significance when the high-sensitivity classifiers trained with $TPF_0 = 0.5$ were compared with LDA_{sfs} for any of the power parameters studied ($n = 1$, 2 and 4).

The performance of the high-sensitivity classifiers and the ordinary GA-based classifiers ($TPF_0 = 0$) are also compared in figures 9–11. It is observed that the difference between the high-sensitivity and the ordinary GA-based classifiers is less than the difference between the high-sensitivity classifiers and the LDA_{sfs}. With a two-tailed significance test, it was found that the difference between the partial areas of the high-sensitivity and the ordinary GA-based classifiers above a true-positive threshold of 0.95 ($A_{0.95}$) did not achieve statistical significance for any of the power parameter values studied ($n = 1$, 2 and 4) with p-levels ranging between 0.06 and 0.5. Similarly, the difference between the ordinary GA-based classifiers and LDA_{sfs} did not achieve statistical significance for any of the power parameter values studied. Table 3 summarizes the A_z, $A_{0.50}$ and $A_{0.95}$ values, as well as the number of features selected by each classifier.

Figures 12 and 13 show the distributions of the classifier outputs for the high-sensitivity classifier ($n = 4$, $TPF_0 = 0.95$) and the LDA_{sfs} respectively. Using the LDA_{sfs}, the distribution of the malignant masses has a relatively long tail that overlaps with the distribution of the benign masses. With the high-sensitivity classifier, this tail seems to

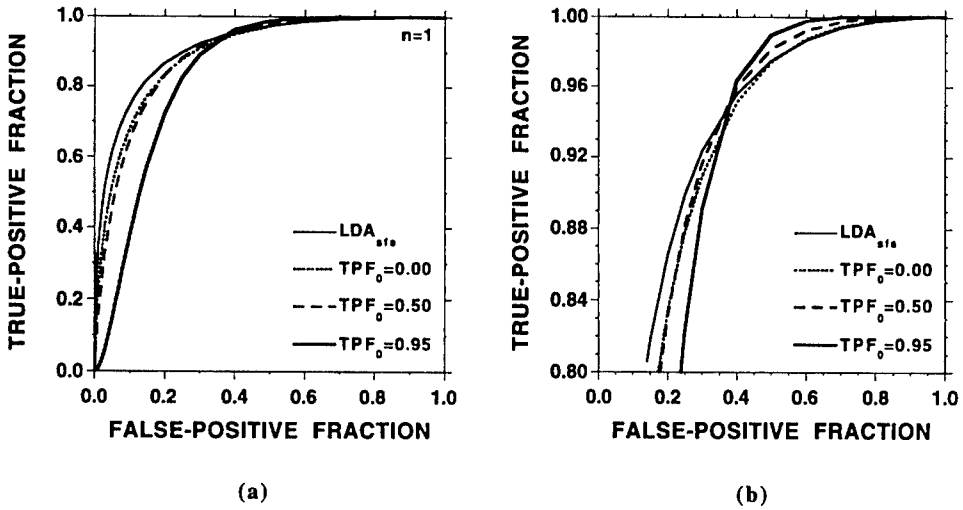


Figure 9. The ROC curves of the LDA_{sfs}, the ordinary GA-based classifier ($TPF_0 = 0$), and the GA-based high-sensitivity classifiers trained with $TPF_0 = 0.50$ and $TPF_0 = 0.95$ using power parameter $n = 1$: (a) the entire ROC curves, (b) enlargement of the curves for TPF > 0.8.

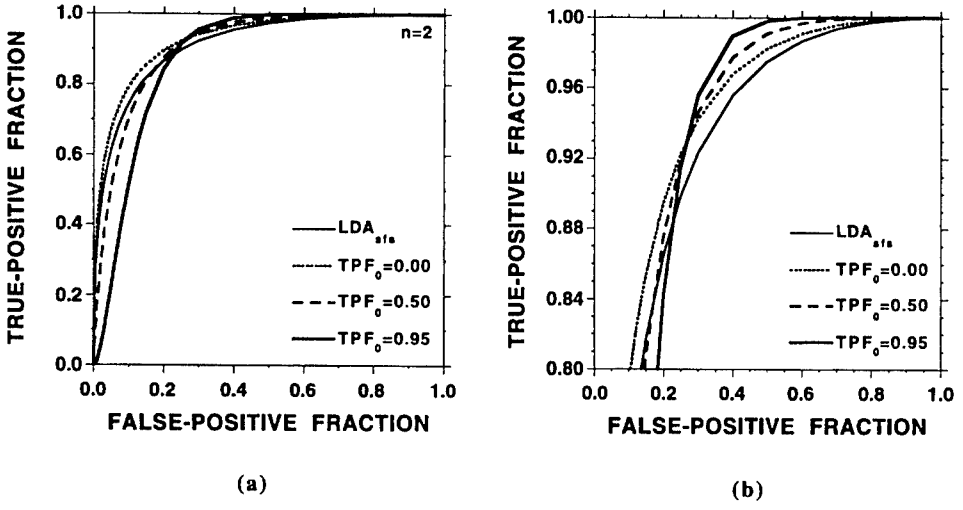


Figure 10. The ROC curves of the LDA_{sfs}, the ordinary GA-based classifier ($TPF_0 = 0$), and the GA-based high-sensitivity classifiers trained with $TPF_0 = 0.50$ and $TPF_0 = 0.95$ using power parameter $n = 2$: (a) the entire ROC curves, (b) enlargement of the curves for TPF > 0.8.

be shortened, so that more benign masses may be correctly diagnosed without missing malignancies. At 100% sensitivity, the specificity with the appropriate choice of the decision threshold was 61% and 34% for the high-sensitivity classifier and the LDA_{sfs} respectively.

4. Discussion

Figures 10 and 11 demonstrate that when the feature selection is performed with a properly designed fitness function in the GA, the designed classifier can be more effective than

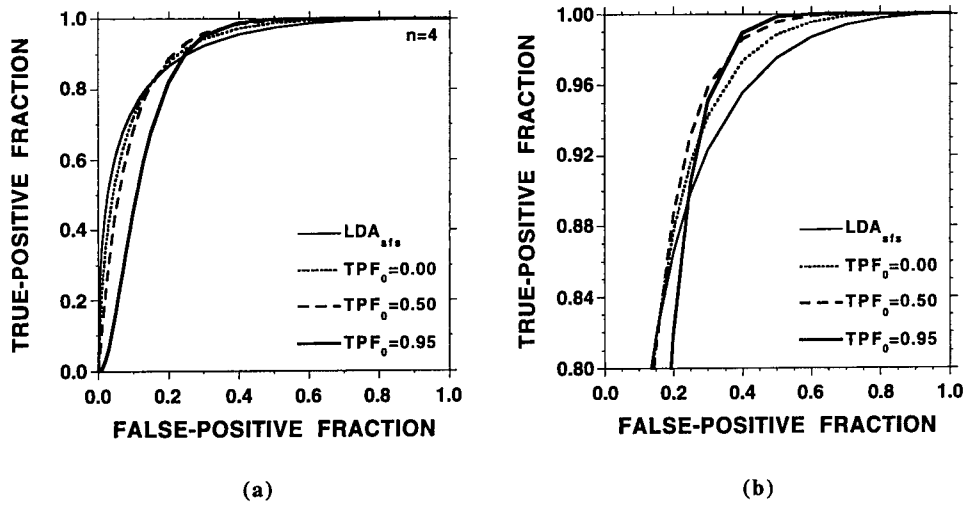


Figure 11. The ROC curves of the LDA_{sfs}, the ordinary GA-based classifier ($TPF_0 = 0$), and the GA-based high-sensitivity classifiers trained with $TPF_0 = 0.50$ and $TPF_0 = 0.95$ using power parameter $n = 4$: (a) the entire ROC curves, (b) enlargement of the curves for TPF > 0.8.

Table 3. The number of features, the area A_z under the ROC curve, the partial area above the true positive fraction of 0.5 ($A_{0.50}$), and that above 0.95 ($A_{0.95}$) for the GA parameters studied. For comparison purposes, the results with linear discriminant analysis are also included as the last row.

Power Parameter, n	TPF_0 value for GA training	Number of selected features	A_z	$A_{0.50}$	$A_{0.95}$
1	0	62	0.90 ± 0.02	0.81 ± 0.03	0.47 ± 0.07
1	0.5	61	0.89 ± 0.02	0.81 ± 0.03	0.51 ± 0.07
1	0.95	58	0.84 ± 0.02	0.76 ± 0.03	0.55 ± 0.05
2	0	60	0.93 ± 0.02	0.86 ± 0.03	0.51 ± 0.08
2	0.5	48	0.91 ± 0.02	0.85 ± 0.03	0.58 ± 0.07
2	0.95	50	0.88 ± 0.02	0.82 ± 0.03	0.63 ± 0.05
4	0	40	0.92 ± 0.02	0.85 ± 0.03	0.56 ± 0.07
4	0.5	39	0.91 ± 0.02	0.85 ± 0.03	0.62 ± 0.06
4	0.95	40	0.87 ± 0.02	0.81 ± 0.03	0.64 ± 0.05
Linear discriminant analysis		41	0.92 ± 0.02	0.83 ± 0.03	0.47 ± 0.07

LDA_{sfs} in the high-sensitivity region of the ROC curve. From table 3 it is observed that although the A_z value for the properly trained high-sensitivity classifier (e.g. $TPF_0 = 0.5$ or 0.95 and $n = 2$ or 4) may be less than that of the LDA_{sfs}, the partial area index $A_{0.95}$ is larger. The statistical analysis in this study showed that the difference between the properly designed high-sensitivity classifiers and the LDA_{sfs} at the high-sensitivity region of the ROC curve can be significant.

Comparing figure 9 with figures 10 and 11, it is observed that the selection of the power parameter n in GA training may be important. The classifiers designed with $n = 1$ did not exhibit a major advantage over the LDA_{sfs}, as also seen from table 3 and the statistical significance tests. From equation (3), it is seen that as the power parameter becomes larger, the difference in the fitness, and thus the probability of being chosen as

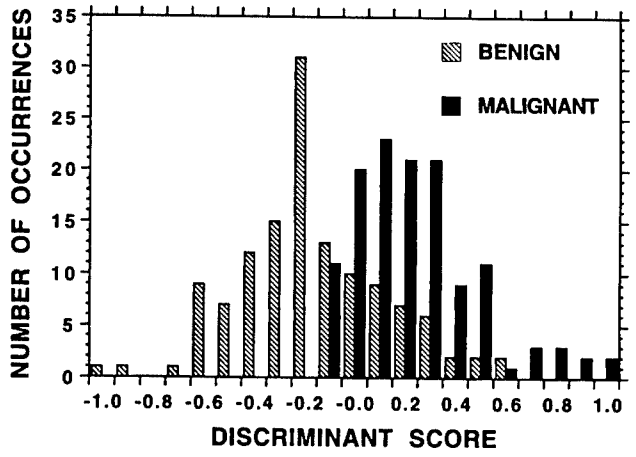


Figure 12. The distribution of the classifier output for the high-sensitivity classifier with $n = 4$, $TPF_0 = 0.95$. By setting an appropriate threshold on these classifier scores, 61% of masses could correctly be classified as benign without missing any malignancies in this study.

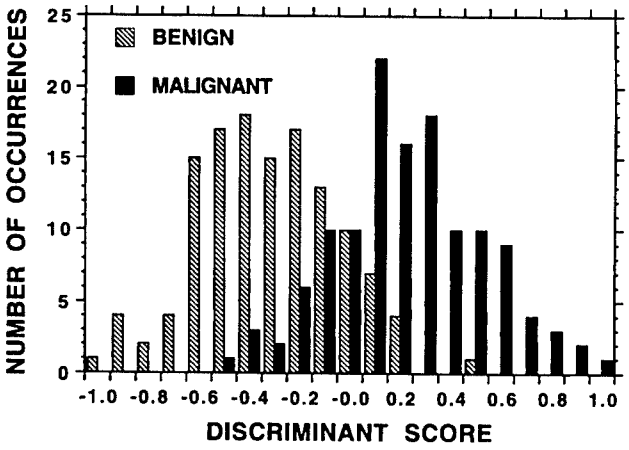


Figure 13. The distribution of the classifier output for LDA_{sfs}. By setting an appropriate threshold on these classifier scores, 34% of masses could be correctly classified as benign without missing any malignancies in this study.

parents, between the chromosomes are more amplified. Therefore, a larger value of n favours the reproduction of better chromosomes in a generation. Although it is desirable to favour the better chromosomes in any GA algorithm, too much emphasis on better chromosomes might suppress the chance of retaining segments of good genes in other chromosomes in the gene pool. This is best seen by letting n tend to infinity, and observing that only the best single chromosome will reproduce in this case, which reduces the GA to a random search algorithm. In our application, from table 3, it is observed that, for all three sensitivity thresholds ($TPF_0 = 0.95$, 0.50 and 0), the classifier trained with $n = 1$ has lower performance indices ($A_{0.95}$, $A_{0.50}$ and A_z) than its counterpart trained with $n = 2$ or $n = 4$. Although none of these differences reached statistical significance, the consistently poorer performance of the classifiers trained with $n = 1$ indicates that $n = 1$ may not be a good choice for GA training.

From figures 7 and 8 it is observed that the best fitness and the number of chromosomes did not change between iterations 140 and 200 for the high-sensitivity classifier with $n = 4$ and $TPF_0 = 0.95$. A similar trend was observed with the other values of n and TPF_0 investigated in this study. Therefore, 200 generations seems to be sufficient for the GA to complete its evolution in this application. In figure 8, the best A_z value was attained around the fiftieth generation, and the A_z value did not change considerably afterwards. However, the $A_{0.95}$ value increased until around 140 generations. This meant that the classification accuracy at high sensitivity continued to increase although the A_z value did not change, i.e. the shape of the ROC curve changed so that the specificity at the high-sensitivity region of the ROC curve increased, while the specificity at the low-sensitivity region of the ROC curve decreased.

Figures 9–11 and the statistical significance tests in section 3 show that although the GA-based high-sensitivity classifiers perform better than the ordinary GA-based classifiers at high sensitivity, the difference between the two classifiers is not statistically significant. Comparison of the LDA_{sfs} and the ordinary GA-based classifiers revealed that neither the difference between the A_z values, nor the difference between the $A_{0.95}$ values were statistically significant ($p > 0.3$). However, the difference between the $A_{0.95}$ values of the LDA_{sfs} and the GA-based high-sensitivity classifiers trained with power parameter $n = 2$ and $n = 4$ was statistically significant (two-tailed p-level <0.05), as described in section 3. Thus, it was necessary to use a high-sensitivity classifier in order to obtain statistically significant improvement over the LDA_{sfs}.

The GA-based high-sensitivity classifiers ($TPF_0 = 0.95$ and $TPF_0 = 0.5$) and the ordinary GA-based classifier ($TPF_0 = 0$) were designed to maximize the partial ROC areas above the chosen true-positive fraction thresholds. From table 3, it is observed that this goal is achieved for the GA-based classifiers with TPF_0 values of 0 and 0.95. For each n, the GA-based classifier with $TPF_0 = 0$ (ordinary GA-based classifier) yielded the highest A_z value, and the GA-based classifier with $TPF_0 = 0.95$ yielded the highest $A_{0.95}$ value among the classifiers. For the classifier with $TPF_0 = 0.5$, the $A_{0.50}$ value was larger than or equal to that of the other GA-based classifiers for $n = 1$ and $n = 4$. However, for $n = 2$, the ordinary GA-based classifier ($TPF_0 = 0$) had the highest $A_{0.50}$ value, although the difference was not statistically significant ($p > 0.3$). This result is not inconsistent with the GA principles or operation. Since the GA training is based on stochastic search, the GA tends to evolve towards the optimal solution, as evidenced by the comparison of the GA-based classifiers in table 3. However, the optimality of the solution is not guaranteed, and one may encounter situations that the design goal was not totally achieved, as evidenced by the fact that the ordinary GA-based classifier had the highest $A_{0.50}$ value for $n = 2$.

Given the probabilistic nature of GA-based feature selection, it is difficult to predict the conditions under which the GA may select a feature set that provides a better high-sensitivity classifier than LDA_{sfs}. Both our GA-based method and the stepwise feature selection algorithm were designed primarily to select features for classifying classes that have multivariate Gaussian distributions and equal covariance matrices. When these assumptions are not satisfied, the accuracy of feature selection will deteriorate to a different degree for both methods. One possible explanation for the relative success of the GA-based feature selection might be that our data violate the assumptions of multivariate normality and the equality of covariance matrices, and that the GA-based method is less sensitive to these violations.

In this study, our focus was to develop a methodology for the design of high-sensitivity classifiers for applications in CAD. For the specific application of discriminating malignant and benign breast lesions, our data set was limited and the features selected by the GA

may not be the optimal set of features for the general population. The same is true for the LDA_{sfs}. Considering that the data set contained only 255 masses, the number of features selected both by the GA and the LDA_{sfs} was large. As a result, if a classifier trained in this study is applied without modification to the population at large, the classification accuracy is likely to be poorer than that obtained in this paper. However, the methodology developed in this study is general. When a sufficiently large data set becomes available, the GA-based high-sensitivity feature selection algorithm can be reapplied, and a more robust feature set can be determined. The number of training cases required for generalizable classifier design and feature selection has been the subject of recent studies (Raudys and Jain 1991, Wagner *et al* 1997, Chan *et al* 1997b), and is currently under investigation.

An important consideration concerning the use of GAs for optimization is the speed of computation. Depending on the number of final features selected, the GA-based feature selection implemented in this study (340 features, 200 chromosomes, 200 generations and leave-one-case-out GA training) took between 24 and 60 h on an AlphaStation 500 (400 Mhz Alpha chip), whereas the stepwise feature selection performed on a PC compatible computer with a 90 MHz Pentium processor took less than 10 min. Therefore, GA-based feature selection implemented in this study may not be practical for studies where the feature selection has to be performed many times. The high-sensitivity classifier design method developed in this study may be more appropriate if the speed of computation is of secondary importance to the classification accuracy of the designed classifier. For example, the GA-based high-sensitivity classifier can be trained only once when a final set of features is desired for a large data set as discussed above.

5. Conclusion

We have developed a GA-based method to design a high-sensitivity classifier for CAD applications. The usefulness of the method was demonstrated by the problem of classifying masses on digitized mammograms. Texture features extracted from RBST images were used to distinguish malignant and benign masses. The accuracy of the high-sensitivity classifier was shown to be significantly higher than that of LDA_{sfs} above a true-positive fraction of 0.95. By using an appropriate decision threshold on the high-sensitivity classifier scores, 61% of the benign masses could correctly be identified without missing any malignant masses. The GA may therefore be a useful tool in the design of high-sensitivity classifiers for different classification problems in CAD or other applications.

Acknowledgments

This work is supported by a USPHS grant CA 48129, a Career Development Award (BS) from the USAMRMC (DAMD 17-96-1-6012), and a USAMRMC grant DAMD 17-96-1-6254. No official endorsement of any equipment or product of any companies mentioned in the publication should be inferred. The authors are grateful to Charles E Metz, PhD, for providing the LABROC and CLABROC programs.

References

Brill F, Brown D and Martin W 1992 Fast genetic selection of features for neural network classifiers *IEEE Trans. Neural Networks* **3** 324–8

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

Chan H-P, Sahiner B, Petrick N, Helvie M A, Lam K L, Adler D D and Goodsitt M M 1997a Computerized classification of malignant and benign microcalcifications on mammograms: texture analysis using an artificial neural network *Phys. Med. Biol.* **42** 549–67

Chan H-P, Sahiner B, Wagner R F, Petrick N and Mossoba J 1997b Effects of sample size on classifier design: quadratic and neural network classifiers *Proc. SPIE* **3034** 1102–13

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

Duda R O and Hart P E 1973 *Pattern Classification and Scene Analysis* (New York: Wiley)

Galloway M M 1975 Texture classification using grey level run lengths *Comput. Graphics Image Process.* **4** 172–9

Hall F M, Storella J M, Silverstone D Z and Wyshak G 1988 Nonpalpable breast lesions: recommendations for biopsy based on suspicion of carcinoma at mammography *Radiology* **167** 353–8

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

Hermann G, Janus C, Schwartz I S, Krivisky B, Bier S and Rabinowitz J G 1987 Nonpalpable breast lesions: accuracy of prebiopsy mammographic diagnosis *Radiology* **165** 323–6

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

Jacobson H G and Edeiken J 1990 Biopsy of occult breast lesions: analysis of 1261 abnormalities *J. Am. Med. Assoc.* **263** 2341–3

Jain A K 1989 *Fundamentals of Digital Image Processing* (New Jersey: Prentice-Hall)

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

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

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

McClish D K 1989 Analyzing a portion of the ROC curve *Med. Decision Making* **9** 190–5

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

Pohlman S, Powell K A, Obuchowski N A, Chilcote W A and Broniatowski S G 1996 Quantitative classification of breast tumors in digitized mammograms *Med. Phys.* **23** 1337–45

Rangayyan R M, El-Faramawy N, Desautels J E L and Alim O A 1996 Discrimination between benign and malignant breast tumors using a region-based measure of edge profile acutance *Digital Mammography '96* ed K Doi, M L Giger, R M Nishikawa and R A Schmidt (Amsterdam: Elsevier) pp 213–18

Raudys S J and Jain A K 1991 Small sample size effects in statistical pattern recognition: recommendations for practitioners *IEEE Trans. Pattern Anal. Machine Intell.* **13** 252–64

Sahiner B, Chan H-P, Petrick N, Goodsitt M M and Helvie M A 1997 Characterization of masses on mammograms: significance of the use of the rubber-band straightening transform *Proc. SPIE* **3034** 491–500

Sahiner B, Chan H-P, Petrick N, Helvie M A and Goodsitt M M 1998 Computerized characterization of masses on mammograms: the rubber band straightening transform and texture analysis *Med. Phys.* **25** 516–26

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

Sahiner B, Chan H-P, Petrick N, Wei D, Helvie M A, Adler D D and Goodsitt M M 1995 Classification of mass and normal breast tissue: an artificial neural network with morphological features *Proc. World Congress on Neural Networks* vol 2 (New Jersey: INNS Press) pp 876–9

——1996b Classification of mass and normal breast tissue: a convolution neural network classifier with spatial domain and texture images *IEEE Trans. Med. Imaging* **15** 598–610

——1996c Image feature selection by a genetic algorithm: application to classification of mass and normal breast tissue *Med. Phys.* **23** 1671–84

Wagner R F, Chan H-P, Mossoba J, Sahiner B and Petrick N 1997 Finite-sample effects and resampling plans: application to linear classifiers in computer-aided diagnosis *Proc. SPIE* **3034** 467–77

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

Weszka J S, Dyer C R and Rosenfeld A 1976 A comparative study of texture measures for terrain classification *IEEE Trans. Syst. Man Cybernetics* **6** 269–85

Computerized analysis of mammographic microcalcifications in morphological and texture feature spaces

Heang-Ping Chan[a] and Berkman Sahiner
Department of Radiology, University of Michigan, Ann Arbor, Michigan 48109

Kwok Leung Lam
Department of Radiation Oncology, University of Michigan, Ann Arbor, Michigan 48109

Nicholas Petrick, Mark A. Helvie, Mitchell M. Goodsitt, and Dorit D. Adler
Department of Radiology, University of Michigan, Ann Arbor, Michigan 48109

(Received 24 September 1997; accepted for publication 20 July 1998)

We are developing computerized feature extraction and classification methods to analyze malignant and benign microcalcifications on digitized mammograms. Morphological features that described the size, contrast, and shape of microcalcifications and their variations within a cluster were designed to characterize microcalcifications segmented from the mammographic background. Texture features were derived from the spatial gray-level dependence (SGLD) matrices constructed at multiple distances and directions from tissue regions containing microcalcifications. A genetic algorithm (GA) based feature selection technique was used to select the best feature subset from the multi-dimensional feature spaces. The GA-based method was compared to the commonly used feature selection method based on the stepwise linear discriminant analysis (LDA) procedure. Linear discriminant classifiers using the selected features as input predictor variables were formulated for the classification task. The discriminant scores output from the classifiers were analyzed by receiver operating characteristic (ROC) methodology and the classification accuracy was quantified by the area, A_z, under the ROC curve. We analyzed a data set of 145 mammographic microcalcification clusters in this study. It was found that the feature subsets selected by the GA-based method are comparable to or slightly better than those selected by the stepwise LDA method. The texture features ($A_z=0.84$) were more effective than morphological features ($A_z=0.79$) in distinguishing malignant and benign microcalcifications. The highest classification accuracy ($A_z=0.89$) was obtained in the combined texture and morphological feature space. The improvement was statistically significant in comparison to classification in either the morphological ($p=0.002$) or the texture ($p=0.04$) feature space alone. The classifier using the best feature subset from the combined feature space and an appropriate decision threshold could correctly identify 35% of the benign clusters without missing a malignant cluster. When the average discriminant score from all views of the same cluster was used for classification, the A_z value increased to 0.93 and the classifier could identify 50% of the benign clusters at 100% sensitivity for malignancy. Alternatively, if the minimum discriminant score from all views of the same cluster was used, the A_z value would be 0.90 and a specificity of 32% would be obtained at 100% sensitivity. The results of this study indicate the potential of using combined morphological and texture features for computer-aided classification of microcalcifications. © *1998 American Association of Physicists in Medicine.* [S0094-2405(98)00910-9]

Key words: computer-aided diagnosis, mammography, microcalcifications, genetic algorithm, linear discriminant analysis, ROC analysis

I. INTRODUCTION

Mammography is the most sensitive method for early detection of breast cancers. However, its specificity for differentiating malignant and benign lesions is relatively low. In the United States, the positive predictive value of mammography ranges from about 15% to 30%.[1,2] Various methods are being developed to improve the sensitivity and specificity of breast cancer detection.[3] Computer-aided diagnosis (CAD) is considered to be one of the promising approaches that may improve the efficacy of mammography.[4] Properly designed CAD algorithms can automatically detect suspicious lesions on a mammogram and alert the radiologist to these regions. They can also extract image features from regions of interest (ROIs) and estimate the likelihood of malignancy for a given lesion, thereby providing the radiologist with additional information for making diagnostic decisions.

There are two major approaches to the development of CAD schemes for classification of mammographic abnormalities. One approach uses computer vision techniques to extract image features from the digitized mammograms and classify the lesions based on the computer-extracted features. The computer-extracted features can include morphological features that are commonly used by radiologists for diagno-

 0094-2405/98/25(10)/2007/13/$10.00 © 1998 Am. Assoc. Phys. Med.

sis, as well as texture features that may not be readily perceived by human eyes. The computerized analysis may therefore increase the utilization of mammographic image information and improve the accuracy of differentiating malignant and benign lesions. The other approach uses radiologists' ratings of mammographic features or encodes the radiologists' readings with numerical values. The lesions are then classified based on these radiologist-extracted features. This approach assists radiologists by systematically extracting image features and by optimally merging the features with a statistical classifier to reach a diagnostic decision. Additional risk factors based on patient demographic information and medical or family histories may also be included as input in either approach.

A number of investigators have developed feature extraction and classification methods for characterization of mammographic masses or microcalcifications. Ackerman *et al.*[5] developed 4 measures of malignancy and classified lesions recorded on 120 digitized xeroradiographs by 3 decision methods. Kilday *et al.*[6] used 7 shape descriptors and patient age to classify 39 masses and could correctly classify 69% of the masses. Huo *et al.*[7] analyzed the spiculation of masses using a radial edge-gradient analysis technique and achieved an area, A_z, under the receiver operating characteristic (ROC) curve of 0.88 in a data set of 95 masses. Sahiner *et al.*[8,9] developed a rubber-band straightening image transformation technique to analyze the texture in the region surrounding a mass and obtained an A_z of 0.94 in a data set of 168 masses. Pohlman *et al.*[10] extracted 6 morphological descriptors to classify 47 masses and obtained A_z values ranging from 0.76 to 0.93. Wee *et al.*[11] analyzed 51 microcalcification clusters on specimen radiographs using the average gray level, contrast, and horizontal length of the microcalcifications and obtained 84% correct classification. Fox *et al.*[12] included cluster features in their classifier and obtained 67% correct classification in a data set of 100 clusters from specimen radiographs. Chan *et al.*[13–18] developed morphological and texture features and evaluated various feature classifiers for differentiation of malignant and benign microcalcifications. Shen *et al.*[19] used 3 shape features, compactness, moments, and Fourier descriptors to classify 143 individual microcalcifications with a nearest neighbor classifier and obtained 100% classification accuracy. Wu *et al.*[20] classified 80 pathologic specimens radiographs with a convolution neural network and obtained an A_z of 0.90. Jiang *et al.*[21] trained a neural network classifier to analyze 8 features extracted from microcalcification clusters and obtained an A_z of 0.92 in a data set of 53 patients. Thiele *et al.*[22] extracted texture and fractal features from the tissue region surrounding a microcalcification cluster for classification and achieved a sensitivity of 89% at a specificity of 83% for 54 clusters. Dhawan *et al.*[23] used features derived from first-order and second-order gray-level histogram statistics and obtained an A_z of 0.81 with a neural network classifier for a data set of 191 clusters.

Computerized classification of mammographic lesions using radiologist-extracted features has also been reported by a number of investigators. Ackerman *et al.*[24] estimated the probability of malignancy of mammographic lesions by analyzing 36 radiologist-extracted characteristics with an automatic clustering algorithm and obtained a specificity of 45% at a sensitivity of 100% in a data set of 102 cases. Gale *et al.*[25] analyzed 12 radiologist-extracted features of mammographic lesions with a computer algorithm and obtained a specificity of 88% at a sensitivity of 79% in a data base of 500 patients. Getty *et al.*[26] developed a computer classifier to enhance the differentiation of malignant and benign lesions by a radiologist during interpretation of xeromammograms. Using a similar approach, D'Orsi *et al.*[27] evaluated a computer aid and obtained an improvement of about 0.05 in sensitivity or specificity in mammographic reading. Wu *et al.*[28] trained a neural network to merge 14 radiologist-extracted features for classification of mammographic lesions and obtained an A_z of 0.89. Baker *et al.*[29] trained a neural network based on the lexicon of the Breast Imaging Recording and Data System of the American College of Radiology and found that the neural network could improve the positive predictive value from 35% to 61% in 206 lesions. Lo *et al.*[30] used a similar approach to predict breast cancer invasion and obtained an A_z of 0.91 for 96 lesions. Although the results of these studies varied over a wide range and the performances of the computer algorithms are expected to depend strongly on data set, they indicate the potential of using CAD techniques to improve the diagnostic accuracy of differentiating malignant and benign lesions.

In our early studies, we found that texture features extracted from spatial gray-level dependence (SGLD) matrices at multiple distances were useful for differentiating malignant and benign masses on mammograms. This may be attributed to the texture changes in the breast tissue due to a developing malignancy. The usefulness of SGLD texture measures in differentiating malignant and benign breast tissues was further demonstrated by analysis of mammographic microcalcifications.[17,18,31] In a preliminary study, we developed morphological features to describe the size, shape, and contrast of the individual microcalcifications and their variation within a cluster. We used these features to classify the microcalcifications and obtained moderate results.[13,15] In the present study, we expanded the data set and explored the feasibility of combining texture and morphological features for classification of microcalcifications. The classification accuracy in the combined feature space was compared with those obtained in the texture feature space or in the morphological feature space alone. We also studied the use of a genetic algorithm[32–34] (GA) to select a feature subset from the large-dimension feature spaces, and compared the classification results to those obtained from features selected with stepwise linear discriminant analysis (LDA).[35] Linear discriminant classifiers[36] were designed for the classification tasks. The performance of the classifiers was analyzed with ROC methodology[37] and the classification accuracy was quantified with the area, A_z, under the ROC curve.

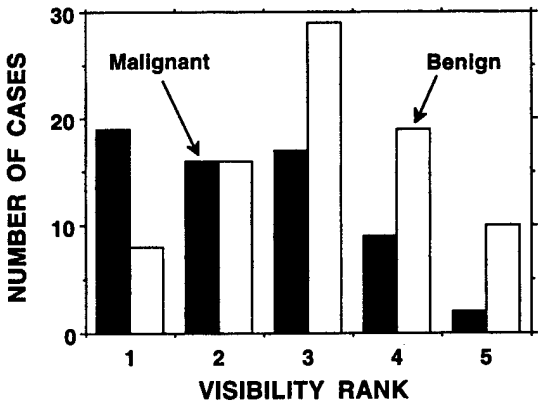


FIG. 1. Distribution of the visibility rankings of the 145 clusters of microcalcifications. Higher ranking corresponds to more subtle clusters.

II. MATERIALS AND METHODS

A. Data set

The data set for this study consisted of 145 clusters of microcalcifications from mammograms of 78 patients. The cases were selected from the patient files in the Department of Radiology at the University of Michigan. The only selection criterion was that it included a biopsy-proven microcalcification cluster. We kept the number of malignant and benign cases reasonably balanced so that 82 benign and 63 malignant clusters were included. All mammograms were acquired with a contact technique using mammography systems accredited by the American College of Radiology (ACR). The dedicated mammographic systems had molybdenum anode and molybdenum filter, 0.3 mm nominal focal spot, reciprocating grid, and Kodak MinR/MinR E screen-film systems with extended processing. A radiologist experienced in mammography ranked the visibility of each microcalcification cluster on a scale of 1 (obvious) to 5 (subtle), relative to the visibility range of microcalcification clusters encountered in clinical practice. The histogram of the visibility ranking of the 145 clusters is shown in Fig. 1. The histogram indicated the mix of subtle and obvious clusters included in the data set.

The selected mammograms were digitized with a laser scanner (Lumisys DIS-1000) at a pixel size of 0.035 mm $\times$0.035 mm and 12-bit gray levels. The digitizer has an optical density (O.D.) range of about 0 to 3.5. The O.D. on the film was digitized linearly to pixel value at a calibration of 0.001 O.D. unit/pixel value in the O.D. range of about 0 to 2.8. The digitizer deviated from a linear response at O.D. higher than 2.8.

B. Morphological feature space

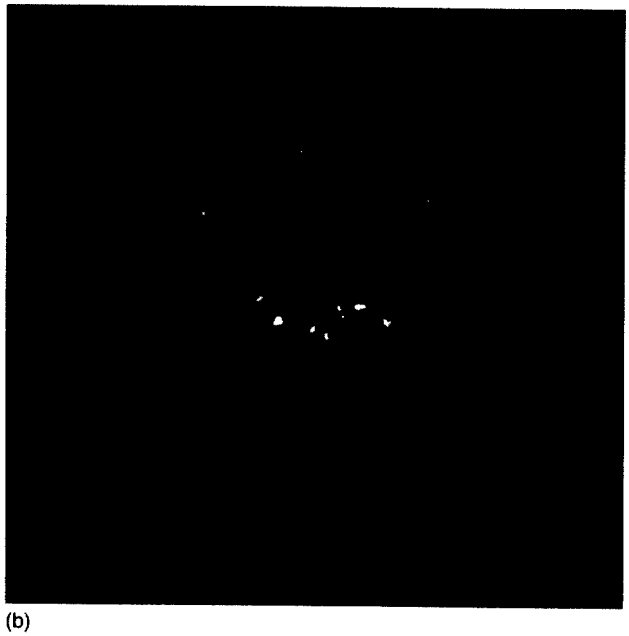
For the extraction of morphological features, the locations of the individual microcalcifications have to be known. We have developed an automated program for detection of individual microcalcifications.[38] However, the detection sensitivity is not 100% and the detected signals include false-positives. Furthermore, automated detection tends to have a higher likelihood of detecting obvious microcalcifications than subtle ones, which may bias the evaluation of the classification capability of the extracted features and the trained classifiers if microcalcifications detected by the automated program are used for classifier development. Since these variables are program dependent, we isolated the detection problem from the classification problem in this study by using manually identified true microcalcifications for the morphological feature analysis. The true microcalcifications were defined as those visible on the film mammograms with a magnifier. Magnification mammograms were used occasionally for verification when they were available, but in most cases only contact mammograms were used. At present, there is no other method that can more reliably identify individual microcalcifications on mammograms. Specimen radiographs can confirm the presence of the microcalcifications but the locations of the individual microcalcifications cannot be correlated with those on the mammograms because of the very different imaging geometry and techniques.

We have developed an automated signal extraction program to determine the size, contrast, signal-to-noise ratio (SNR), and shape of the microcalcifications from a mammogram based on the coordinate of each individual microcalcification. In a local region of 101×101 pixels centered at each signal site, the low frequency structured background is estimated by polynomial curve fitting in the horizontal and vertical directions and then averaging the fitted values obtained in the two directions at each pixel. This background estimation method is used because it can approximate the background more closely than two-dimensional surface fitting or the distance-weighted interpolation method (described below) used for texture feature extraction. The central $l\times l$ pixels that contain the signal are excluded from the curve fitting and noise estimation. The size l is chosen to be a constant of 15 pixels which is larger than the diameters of the microcalcifications of interest yet much smaller than the local region. The background pixel values in this $l\times l$ region are estimated from the fitted and smoothed background surface. The exclusion of the signal region is necessary so that the high contrast pixel values of the microcalcification will not affect the background estimation at the signal site. Other microcalcifications that may locate within the 101×101 pixel region are treated as background pixels because their effect on the estimated background levels at the signal site will be relatively small.

After subtraction of the structured background, the local root-mean-square (rms) noise is calculated. A gray-level threshold is determined as the product of the rms noise and an input SNR threshold. With a region growing technique, the signal region is then extracted as the connected pixels above the threshold around the manually identified signal location. A high threshold will result in extracting only the peak pixels of the microcalcification which may not represent its shape perceived on the mammogram. A low threshold will cause the microcalcification region to grow into the surrounding background pixels. Since there is no objective standard what the actual shape of a microcalcification is on a mammogram, the proper threshold to extract the signals was



(a)



(b)

FIG. 2. An example of a cluster of malignant microcalcifications in the data set: (a) the cluster with mammographic background, (b) the cluster after segmentation. Morphological features are extracted from the segmented microcalcifications.

determined by visually comparing the microcalcifications in the original image and the thresholded image of the microcalcifications superimposed on a background of constant pixel values. After an experienced radiologist compared a subset of randomly selected microcalcification clusters extracted at different thresholds, an SNR threshold of 2.0 was chosen for all cases. An example of a malignant cluster and the microcalcifications extracted at an SNR threshold of 2.0 is shown in Fig. 2.

The feature descriptors determined from the extracted microcalcifications are listed in Table I. The size of a microcalcification (SA) is estimated as the number of pixels in the signal region. The mean density (MD) is the average of the pixel values above the background level within the signal region. The second moments are calculated as

TABLE I. The 21 morphological features extracted from a microcalcification cluster.

	Average	Standard deviation	Coefficient of variation	Maximum
Area	AVSA	SDSA	CVSA	MXSA
Mean density	AVMD	SDMD	CVMD	MXMD
Eccentricity	AVEC	SDEC	CVEC	MXEC
Moment ratio	AVMR	SDMR	CVMR	MXMR
Axis ratio	AVAR	SDAR	CVAR	MXAR
No. of microcalcifications in cluster	NUMS			

$$M_{xx}=\sum_i g_i(x_i-M_x)^2/M_0, \tag{1}$$

$$M_{yy}=\sum_i g_i(y_i-M_y)^2/M_0, \tag{2}$$

$$M_{xy}=\sum_i g_i(x_i-M_x)(y_i-M_y)/M_0, \tag{3}$$

where g_i is the pixel value above the background, and (x_i,y_i) are the coordinates of the ith pixel. The moments M_0, M_x and M_y are defined as follows:

$$M_0=\sum_i g_i, \tag{4}$$

$$M_x=\sum_i g_i x_i/M_0, \tag{5}$$

$$M_y=\sum_i g_i y_i/M_0. \tag{6}$$

The summations are over all pixels within the signal region. The lengths of the major axis, $2a$, and the minor axis, $2b$, of the effective ellipse that characterizes the second moments are given by

$$2a=\sqrt{2[M_{xx}+M_{yy}+\sqrt{(M_{xx}-M_{yy})^2+4M_{xy}^2}]}, \tag{7}$$

$$2b=\sqrt{2[M_{xx}+M_{yy}-\sqrt{(M_{xx}-M_{yy})^2+4M_{xy}^2}]}. \tag{8}$$

The eccentricity (EC) of the effective ellipse can be derived from the major and minor axes as

$$\epsilon=\frac{\sqrt{a^2-b^2}}{a}. \tag{9}$$

The moment ratio (MR) is defined as the ratio of M_{xx} to M_{yy}, with the larger second moment in the denominator. The axis ratio (AR) is the ratio of the major axis to the minor axis of the effective eclipse.

To quantify the variation of the visibility and shape descriptors in a cluster, the maximum (MX), the average (AV) and the standard deviation (SD) of each feature for the individual microcalcifications in the cluster are calculated. The coefficient of variation (CV), which is the ratio of the SD to AV, is used as a descriptor of the variability of a certain

feature within a cluster. Twenty cluster features are therefore derived from the five features (size, mean density, moment ratio, axis ratio, and eccentricity) of the individual microcalcifications. Another feature describing the number of microcalcifications in a cluster (NUMS) is also added, resulting in a 21-dimensional morphological feature space.

C. Texture feature space

Our texture feature extraction method has been described in detail previously.[31] Briefly, texture features are extracted from a 1024×1024 pixel region of interest (ROI) that contains the cluster of microcalcifications. Most of the clusters in this data set can be contained within the ROI. For the few clusters that are substantially larger than a single ROI, additional ROIs containing the remaining parts of the cluster are extracted and processed in the same way as the other ROIs. The texture feature values extracted from the different ROIs of the same cluster are averaged and the average values are used as the feature values for that cluster.

For a given ROI, background correction is first performed to reduce the low frequency gray-level variation due to the density of the overlapping breast tissue and the x-ray exposure conditions. The gray level at a given pixel of the low frequency background is estimated as the average of the distance-weighted gray levels of four pixels at the intersections of the normals from the given pixel to the four edges of the ROI.[39] The estimated background image was subtracted from the original ROI to obtain a background-corrected image. An example of the background correction procedure is shown in Fig. 3.

As discussed in our previous study,[31] it was found that the texture features derived from the SGLD matrix of the ROI provided useful texture information for classification of microcalcification clusters. The SGLD matrix element, $p_{\theta,d}(i,j)$, is the joint probability of the occurrence of gray levels i and j for pixel pairs which are separated by a distance d and at a direction θ.[40] The SGLD matrices were constructed from the pixel pairs in a subregion of 512×512 pixels centered approximately at the center of the cluster in the background-corrected ROI so that any potential edge effects caused by background correction will not affect the texture extraction. We analyzed the texture features in four directions: $\theta = 0°$, 45°, 90°, and 135° at each pixel pair distance d. The pixel pair distance was varied from 4 to 40 pixels in increments of 4 pixels. Therefore, a total of 40 SGLD matrices were derived from each ROI. The SGLD matrix depends on the bin width (or gray-level interval) used in accumulating the histogram. Based on our previous study, a bin width of four gray levels was chosen for constructing the SGLD matrices. This is equivalent to reducing the gray-level resolution (or bit depth) of the 12-bit image to 10 bits by eliminating the 2 least significant bits.

From each of the SGLD matrices, we derived 13 texture measures including correlation, entropy, energy (angular second moment), inertia, inverse difference moment, sum average, sum entropy, sum variance, difference average, difference entropy, difference variance, information measure of correlation 1, and information measure of correlation 2. The formulation of these texture measures could be found in the literature.[31,40] As found in our previous study,[41] we did not observe a significant dependence of the discriminatory power of the texture features on the direction of the pixel pairs for mammographic textures. However, since the actual distance between the pixel pairs in the diagonal direction was a factor

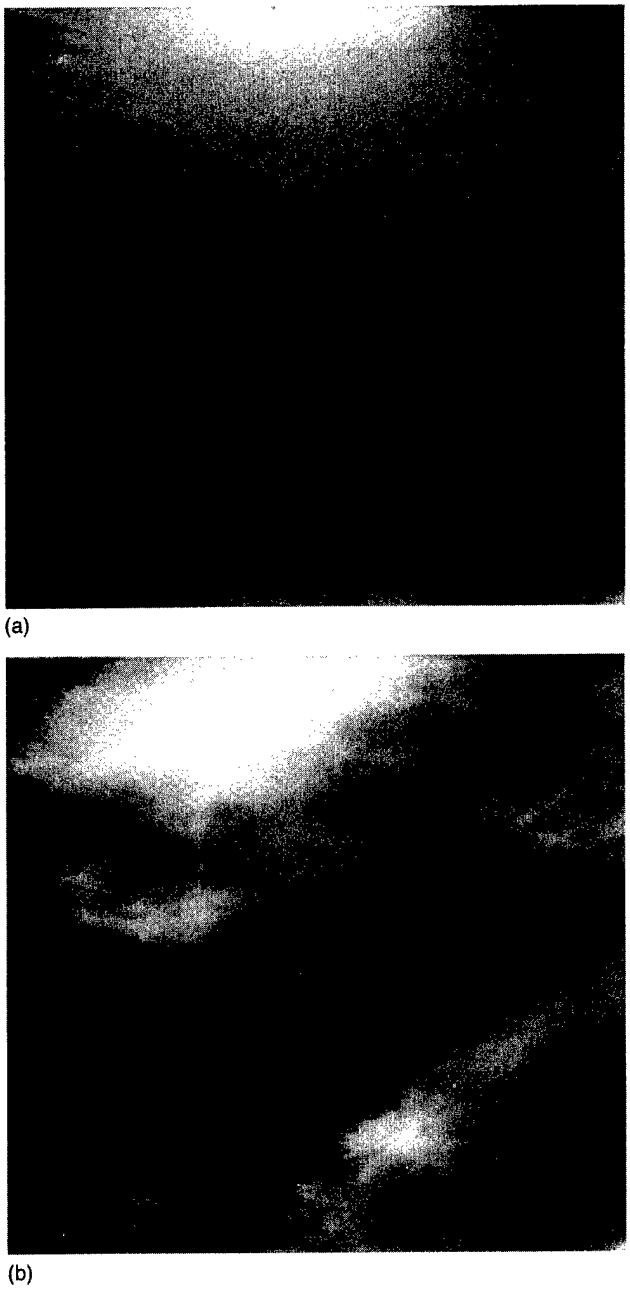


FIG. 3. An example of background correction for the ROIs before texture feature extraction. The ROI from the original image is shown in Fig. 2(a). (a) The estimated low frequency background gray level, and (b) the ROI after background correction. The background gray-level variation due to the varying x-ray penetration in the breast tissue is reduced. The contouring in the background image is a display artifact that does not exist in the calculated image file. For display purpose, the background-corrected ROI is contrast-enhanced to improve the visibility of the microcalcifications and the detailed structures.

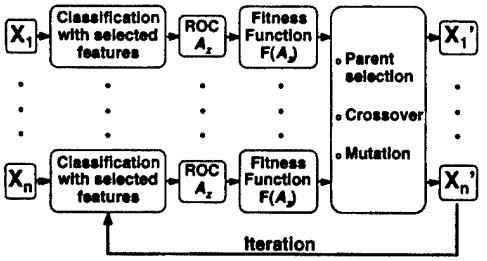


FIG. 4. A schematic diagram of the genetic algorithm designed for feature selection used in this study. $X_1,\ldots,X_n$ represents the set of parent chromosomes and $X_1',\ldots,X_n'$ represents the set of offspring chromosomes.

of $\sqrt{2}$ greater than that in the axial direction, we averaged the feature values in the axial directions (0° and 90°) and in the diagonal directions (45° and 135°) separately for each texture feature derived from the SGLD matrix at a given pixel pair distance. The average texture features at the ten pixel pair distances and two directions formed a 260-dimensional texture feature space.

D. Feature selection

Feature selection is one of the most important steps in classifier design because the presence of ineffective features often degrades the performance of a classifier on test samples. This is partly caused by the "curse of dimensionality" problem that the classifier is inadequately trained in a large-dimension feature space when only a finite number of training samples is available.[42–45] We compared two feature selection methods to extract useful features from the morphological, texture, and the combined feature spaces. One is a genetic algorithm approach, and the other is the commonly used stepwise linear discriminant analysis method.

1. Genetic algorithm for feature selection

The genetic algorithm (GA) methodology was first introduced by Holland in the early 1970s.[32,33] A GA solves an optimization problem based on the principles of natural selection. In natural selection, a population evolves by finding beneficial adaptations to a complex environment. The characteristics of a population are carried onto the next generation by its chromosomes. New characteristics are introduced into a chromosome by crossover and mutation. The probability of survival or reproduction of an individual depends more or less on its fitness to the environment. The population therefore evolves toward better-fit individuals.

The application of GA to feature selection has been described in the literature.[46,47] We have demonstrated previously that a GA could select effective features for classification of masses and normal breast tissue from a very large-dimension feature space.[34] The GA was adapted to the current problem for classification of malignant and benign microcalcifications. A brief outline is given as follows. Each feature in a given feature space is treated as a gene and is encoded by a binary digit (bit) in a chromosome. A "1" represents the presence of the feature and a "0" represents the absence of the feature. The number of genes (bits) on a chromosome is equal to the dimensionality (k) of the feature space, but only the features that are encoded as "1" are actually present in the subset of selected features. A chromosome therefore represents a possible solution to the feature selection problem.

The implementation of GA for feature selection is illustrated in the block diagram shown in Fig. 4. To allow for diversity, a large number, n, of chromosomes, $X_1,\ldots,X_n$, is chosen as the population. The number of chromosomes is kept constant in each generation. At the initiation of the GA, each bit on a chromosome is initialized randomly with a small but equal probability, P_{init}, to be "1." The selected feature subset on a chromosome is used as the input feature variables to a classifier, which was chosen to be the Fischer's linear discriminant in this study.

The available samples in the dataset are randomly partitioned into a training set and a test set. The training set is used to formulate a linear discriminant function with each of the selected feature subsets. The effectiveness of each of the linear discriminants for classification is evaluated with the test set. The classification accuracy is determined as the area, A_z, under the ROC curve. To reduce biases in the classifiers due to case selection, training and testing are performed a large number of times, each with a different random partitioning of the data set. In this study, we chose to partition the dataset 80 times and the 80 test A_z values were averaged and used for determination of the fitness of the chromosome.

The fitness function for the ith chromosome, $F(i)$, is formulated as

$$F(i)=\left[\frac{f(i)-f_{\min}}{f_{\max}-f_{\min}}\right]^2, \quad i=1,\ldots,n, \tag{10}$$

where

$$f(i)=\overline{A_z(i)}-\alpha N(i),$$

$\overline{A_z(i)}$ is the average test A_z for the ith chromosome over the 80 random partitions of the data set, $f_{\min}$ and $f_{\max}$ are the minimum and maximum $f(i)$ among the n chromosomes, $N(i)$ is the number of features in the ith chromosome, and α is a penalty factor, whose magnitude is less than $1/k$, to suppress chromosomes with a large number of selected features. The value of the fitness function $F(i)$ ranges from 0 to 1. The probability of the ith chromosome being selected as a parent, $P_s(i)$, is proportional to its fitness function:

$$P_s(i)=F(i)/\sum_{i=1}^{n} F(i), \quad i=1,\ldots,n. \tag{11}$$

A random sampling based on the probabilities, $P_s(i)$, will allow chromosomes with higher value of fitness to be selected more frequently.

For every pair of selected parent chromosomes, X_i and X_j, a random decision is made to determine if crossover should take place. A uniform random number in (0,1] is generated. If the random number is greater than P_c, the probability of crossover, then no crossover will occur; otherwise, a random crossover site is selected on the pair of chromosomes. Each chromosome is split into two strings at this site and one of the strings will be exchanged with the corre-

sponding string from the other chromosome. Crossover results in two new chromosomes of the same length.

After crossover, another chance of introducing new features is obtained by mutation. Mutation is applied to each gene on every chromosome. For each bit, a uniform random number in (0,1] is generated. If the random number is greater than P_m, the probability of mutation, then no mutation will occur; otherwise, the bit is complemented. The processes of parent selection, crossover, and mutation result in a new generation of n chromosomes, $X_1', \ldots, X_n'$, which will again be evaluated with the 80 training and test set partitions as described above. The chromosomes are allowed to evolve over a preselected number of generations. The best subset of features is chosen to be the chromosome that provides the highest average A_z during the evolution process.

In this study, 500 chromosomes were used in the population. Each chromosome has 281 gene locations. P_{init} was chosen to be 0.01 so that each chromosome started with two to three features on the average. We varied P_c from 0.7 to 0.9, P_m from 0.001 to 0.005, and α from 0 to 0.001. These ranges of parameters were chosen based on our previous experience with other feature selection problems using GA.[34]

2. Stepwise linear discriminant analysis

The stepwise linear discriminant analysis (LDA) is a commonly used method for selection of useful feature variables from a large feature space. Detailed descriptions of this method can be found in the literature.[35] The procedure is briefly outlined below. The stepwise LDA uses a forward selection and backward removal strategy. When a feature is entered into or removed from the model, its effect on the separation of the two classes can be analyzed by several criteria. We use the Wilks' lambda criterion which minimizes the ratio of the within-group sum of squares to the total sum of squares of the two class distributions; the significance of the change in the Wilks' lambda is estimated by F-statistics. In the forward selection step, the features are entered one at a time. The feature variable that causes the most significant change in the Wilks' lambda will be included in the feature set if its F value is greater than the F-to-enter (F_{in}) threshold. In the feature removal step, the features already in the model are eliminated one at a time. The feature variable that causes the least significant change in the Wilks' lambda will be excluded from the feature set if its F value is below the F-to-remove (F_{out}) threshold. The stepwise procedure terminates when the F values for all features not in the model are smaller than the F_{in} threshold and the F values for all features in the model are greater than the F_{out} threshold. The number of selected features will decrease if either the F_{in} threshold or the F_{out} threshold is increased. Therefore, the number of features to be selected can be adjusted by varying the F_{in} and F_{out} values.

E. Classifier

The training and testing procedure described above was used for the purpose of feature selection only. After the best subset of features as determined by either the GA or the stepwise LDA procedure was found, we performed the classification as follows.

The linear discriminant analysis[36] procedure in the SPSS software package[35] was used to classify the malignant and benign microcalcification clusters. We used a cross-validation resampling scheme for training and testing the classifier. The data set of 145 samples was randomly partitioned into a training set and a test set by an approximately 3:1 ratio. The partitioning was constrained so that ROIs from the same patient were always grouped into the same set. The training set was used to determine the coefficients (or weights) of the feature variables in the linear discriminant function. The performance of the trained classifier was evaluated with the test set. In order to reduce the effect of case selection, the random partitioning was performed 50 times. The results were then averaged over the 50 partitions.

The classification accuracy of the LDA was evaluated by ROC methodology. The output discriminant score from the LDA classifier was used as the decision variable in the ROC analysis. The LABROC program,[37] which assumes binormal distributions of the decision variable for the two classes and fits an ROC curve to the classifier output based on maximum-likelihood estimation, was used to estimate the ROC curve of the classifier. The ROC curve represents the relationship between the true-positive fraction (TPF) and the false-positive fraction (FPF) as the decision threshold varies. The area under the ROC curve and the standard deviation of the A_z were provided by the LABROC program for each partition of training and test sets. The average performance of the classifier was estimated as the average of the 50 test A_z values from the 50 random partitions.

To obtain a single distribution of the discriminant scores for the test samples, we performed a leave-one-case-out resampling scheme for training and testing the classifier. In this scheme, one of the 78 cases was left out at a time and the clusters from the other 77 cases were used for formulation of the linear discriminant function. The resulting LDA classifier was used to classify the clusters from the left-out case. The procedure was performed 78 times so that every case was left out once to be the test case. The test discriminant scores from all the clusters were accumulated in a distribution which was then analyzed by the LABROC program. Using the distributions of discriminant scores for the test samples from the leave-one-case-out resampling scheme, the CLABROC program could be used to test the statistical significance of the differences between ROC curves[48] obtained from different conditions. The two-tailed p value for the difference in the areas under the ROC curves was estimated.

III. RESULTS

The variations of best feature set size and classifier performance in terms of A_z with the GA parameters were tabulated in Table II(a)–(c) for the morphological, the texture, and the combined feature spaces, respectively. The number of generations that the chromosomes evolved was fixed at 75

TABLE II. Dependence of feature selection and classifier performance on GA parameters: (a) morphological feature space, (b) texture feature space, and (c) combined feature space. The number of generations that the GA evolved was fixed at 75. The best result for each feature space is identified with an asterisk.

(a)

P_c	P_m	α	No. of features	A_z (Training)	A_z (Test)
0.7	0.001	0	6	0.84	0.79
0.8			3	0.77	0.76
0.9			4	0.80	0.77
0.7	0.003		7	0.82	0.78
0.8			6	0.82	0.79
0.9			6	0.84	0.79
0.7	0.001	0.0005	3	0.77	0.76
0.8			4	0.80	0.77
0.9			3	0.77	0.76
0.7	0.003		6	0.84	0.79*
0.8			6	0.84	0.79
0.9			6	0.82	0.79
0.7	0.001	0.0010	3	0.77	0.76
0.8			4	0.80	0.77
0.9			3	0.77	0.76
0.7	0.003		6	0.84	0.79
0.8			7	0.84	0.79
0.9			4	0.80	0.77

(b)

P_c	P_m	α	No. of features	A_z (Training)	A_z (Test)
0.7	0.001	0	7	0.87	0.82
0.8			8	0.88	0.84
0.9			8	0.88	0.84
0.7	0.003		17	0.91	0.82
0.8			9	0.88	0.79
0.9			10	0.88	0.79
0.7	0.001	0.0005	9	0.88	0.85*
0.8			7	0.86	0.82
0.9			8	0.87	0.84
0.7	0.003		13	0.90	0.81
0.8			10	0.87	0.81
0.9			12	0.88	0.81
0.7	0.001	0.0010	7	0.87	0.83
0.8			9	0.88	0.83
0.9			8	0.88	0.83
0.7	0.003		10	0.88	0.83
0.8			21	0.94	0.82
0.9			12	0.88	0.80

(c)

P_c	P_m	α	No. of features	A_z (Training)	A_z (Test)
0.7	0.001	0	13	0.93	0.88
0.8			12	0.92	0.88
0.9			12	0.92	0.89
0.7	0.003		12	0.91	0.86
0.8			16	0.94	0.88
0.9			17	0.95	0.88
0.7	0.001	0.0003	12	0.92	0.87
0.8			12	0.92	0.86
0.9			12	0.93	0.88
0.7	0.003		13	0.93	0.87
0.8			13	0.93	0.88
0.9			12	0.94	0.89*
0.7	0.005		12	0.89	0.80
0.7	0.001	0.0010	11	0.92	0.87
0.8			10	0.91	0.87
0.9			11	0.91	0.86
0.7	0.003		10	0.91	0.86
0.8			14	0.93	0.87
0.9			13	0.92	0.87
0.7	0.005		11	0.89	0.81
0.8			12	0.88	0.82
0.9			12	0.89	0.81

TABLE III. Dependence of feature selection and classifier performance on F_{out} and F_{in} thresholds using stepwise linear discriminant analysis: (a) morphological feature space, (b) texture feature space, and (c) combined feature space. The best result for each feature space is identified with an asterisk. When the test A_z is comparable, the feature set with fewer number of features is considered to be better.

(a)

F_{out}	F_{in}	No. of features	A_z (Training)	A_z (Test)
2.7	3.8	2	0.76	0.76
1.7	2.8	4	0.79	0.76
1.7	1.8	6	0.83	0.79*
1.0	1.4			
1.0	1.2	7	0.84	0.79
0.8	1.0	9	0.85	0.79
0.6	0.8			
0.4	0.6	10	0.85	0.79
0.2	0.4	12	0.86	0.78
0.1	0.2			

(b)

F_{out}	F_{in}	No. of features	A_z (Training)	A_z (Test)
2.7	3.8	4	0.82	0.80
1.7	2.8			
1.0	1.4	8	0.88	0.83
1.0	1.2	10	0.89	0.82
0.8	1.0	11	0.89	0.83
0.6	0.8	14	0.91	0.85*
0.4	0.6	17	0.92	0.84
0.2	0.4	18	0.92	0.81
0.1	0.2	16	0.90	0.80

(c)

F_{out}	F_{in}	No. of features	A_z (Training)	A_z (Test)
3.0	3.2	6	0.84	0.80
2.9	3.2			
2.8	3.1			
2.0	3.1			
3.0	3.1	10	0.88	0.83
2.9	3.0			
2.7	2.8			
2.0	2.3	11	0.90	0.86
2.0	2.2			
1.9	2.0			
1.7	1.8			
1.3	1.5	14	0.92	0.86
1.0	1.2	19	0.95	0.86
1.0	1.1	23	0.96	0.87*
0.8	1.2	28	0.97	0.86

in these tables. The training and test A_z values were obtained from averaging results of the 50 partitions of the data sets using the selected feature sets.

The results of feature selection using the stepwise LDA procedure with a range of F_{in} and F_{out} thresholds were tabulated in Table III(a)–(c). The thresholds were varied so that the number of selected features varied over a wide range. Often different choices of F_{in} and F_{out} values could result in the same selected feature set as shown in the tables by the number of features in the set. The average A_z values obtained from the 50 partitions of the data set using the selected feature sets were listed. The best feature sets selected in the different feature spaces are shown in Table IV.

TABLE IV. The best feature sets selected by the GA and stepwise LDA methods (indicated by asterisk in Tables II and III) in the three feature spaces. The number of generations for chromosome evolution in the GA algorithm to reach the selected feature sets is listed. The abbreviations for the texture features are: correlation (CORE), energy (ENER), entropy (ENTR), difference average (DFAV), difference entropy (DFEN), difference variance (DFVR), inertia (INER), inverse difference moment (INVD), information measure of correlation 1 (ICO1), information measure of correlation 2 (ICO2), sum average (SMAV), sum entropy (SMEN), sum variance (SMVR). After an abbreviation, the letter "A" indicates diagonal features and the number indicates the pixel distance. The abbreviations for the morphological features can be found in Table I.

GA			Stepwise LDA		
Morphological generation 39	Texture generation 64	Combined generation 169	Morphological	Texture	Combined
CMVD	DFAVA_8	DFAVA_4	AVMD	DFAV_12	CORE_40
CVMR	DFEN_16	DFEN_28	CVMD	DFEN_4	COREA_16
CVSA	DFVRA_24	DFVRA_36	CVMR	DFEN_8	COREA_40
MXMR	DFVR_24	DFVR_12	CVSA	DFENA_12	DFAVA_8
MXSA	DFVR_4	DFVR_20	MXMR	DFENA_24	DFEN_4
SDMD	DFVR_8	ICO1A_20	MXSA	DFVR_24	DFEN_8
	ICO1A_12	ICO1A_32		DFVR_40	DFENA_36
	ICO2A_28	SMEN_16		ICO1_16	DFVR_20
	ICO2_40	SMEN_36		ICO1A_8	ICO1A_28
		AVAR		ICO2_40	ICO2_24
		CVMD		INER_8	ICO2_36
		CVSA		INVD_16	INER_12
		MXEC		INVD_4	INERA_16
		NUMS		INVDA_8	INVDA_36
		SDMD			SMEN_40
					SMENA_4
					AVAR
					CVMD
					CVSA
					MXAR
					MXEC
					NUMS
					SDMD

Table V compares the training and test A_z values from the best feature set in each feature space for the two feature selection methods. The GA parameters that selected the feature set with best classification performance in each feature space after 75 generations (Table II) were used to run the GA again for 500 generations. The A_z values obtained with the best GA selected feature sets after 75 generations are listed together with those obtained after 500 generations. The A_z values obtained with the leave-one-case-out scheme are also shown in Table V. The differences between the corresponding A_z values from the two resampling schemes are within 0.01. The two feature selection methods provided feature sets that had similar test A_z values in the morphological and texture feature spaces. In the combined feature space, there was a slight improvement in the test A_z value obtained with the GA selected features. Although the difference in the A_z

TABLE V. Classification accuracy of linear discriminant classifier in the different feature spaces using feature sets selected by the GA and the stepwise LDA procedure.

	Training A_z			Text A_z		
Feature selection	Morphological	Texture	Combined	Morphological	Texture	Combined
Cross-validation						
GA (75 generations)	0.84±0.04	0.88±0.03	0.94±0.02	0.79±0.07	0.85±0.07	0.89±0.05
GA (500 generations)	0.84±0.04	0.88±0.03	0.96±0.02	0.79±0.07	0.85±0.07	0.90±0.05
Stepwise LDA	0.83±0.04	0.91±0.03	0.96±0.02	0.79±0.07	0.85±0.06	0.87±0.06
Leave-one-case-out						
GA (75 generations)	0.83±0.03	0.88±0.03	0.94±0.02	0.79±0.04	0.84±0.03	0.89±0.03
GA (500 generations)	0.83±0.03	0.88±0.03	0.95±0.02	0.79±0.04	0.84±0.03	0.89±0.03
Stepwise LDA	0.83±0.03	0.91±0.02	0.96±0.02	0.79±0.04	0.85±0.03	0.87±0.03

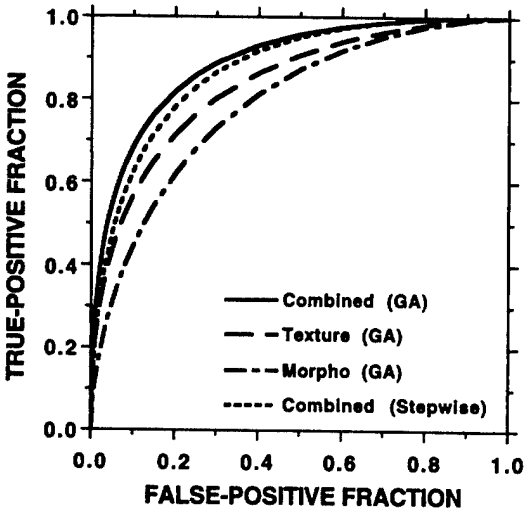


FIG. 5. Comparison of ROC curves of the LDA classifier performance using the best GA selected feature sets in the three feature spaces. In addition, the ROC curve obtained from the best feature set selected by the stepwise LDA procedure in the combined feature space is shown. The classification was performed with a leave-one-case-out resampling scheme.

values from the leave-one-case-out scheme between the two feature selection methods did not achieve statistical significance ($p=0.2$), as estimated by CLABROC, the differences in the paired A_z values from the 50 partitions demonstrated a consistent trend (40 out of 50 partitions) that the A_z from the GA selected features were higher than those obtained by the stepwise LDA. This trend was also observed in our previous study in which mass and normal tissue were classified.[34]

The ROC curves for the test samples using the feature sets selected by the GA were plotted in Fig. 5. The classification accuracy in the combined feature space was significantly higher than those in the morphological ($p=0.002$) or the texture feature space ($p=0.04$) alone. The ROC curve using the feature set selected by the stepwise procedure in the combined feature space was also plotted for comparison. The distribution of the discriminant scores for the test samples using the feature set selected by the GA in the combined feature space is shown in Fig. 6(a). If a decision threshold is chosen at 0.3, 29 of the 82 (35%) benign samples can be correctly classified without missing any malignant clusters.

Some of the 145 samples are different views of the same microcalcification clusters. In clinical practice, the decision regarding a cluster is based on information from all views. If it is desirable to provide the radiologist a single relative malignancy rating for each cluster, two possible strategies may be used to merge the scores from all views: the average score or the minimum score. The latter strategy corresponds to the use of the highest likelihood of malignancy score for the cluster. There were a total of 81 different clusters (44 benign and 37 malignant) from the 78 cases because 3 of the cases contained both a benign and a malignant cluster. The distributions of the average and the minimum discriminant scores of the 81 clusters in the combined feature space were plotted in Fig. 6(b) and Fig. 6(c), respectively. Using the average scores, ROC analysis provided test A_z values of 0.93 ± 0.03

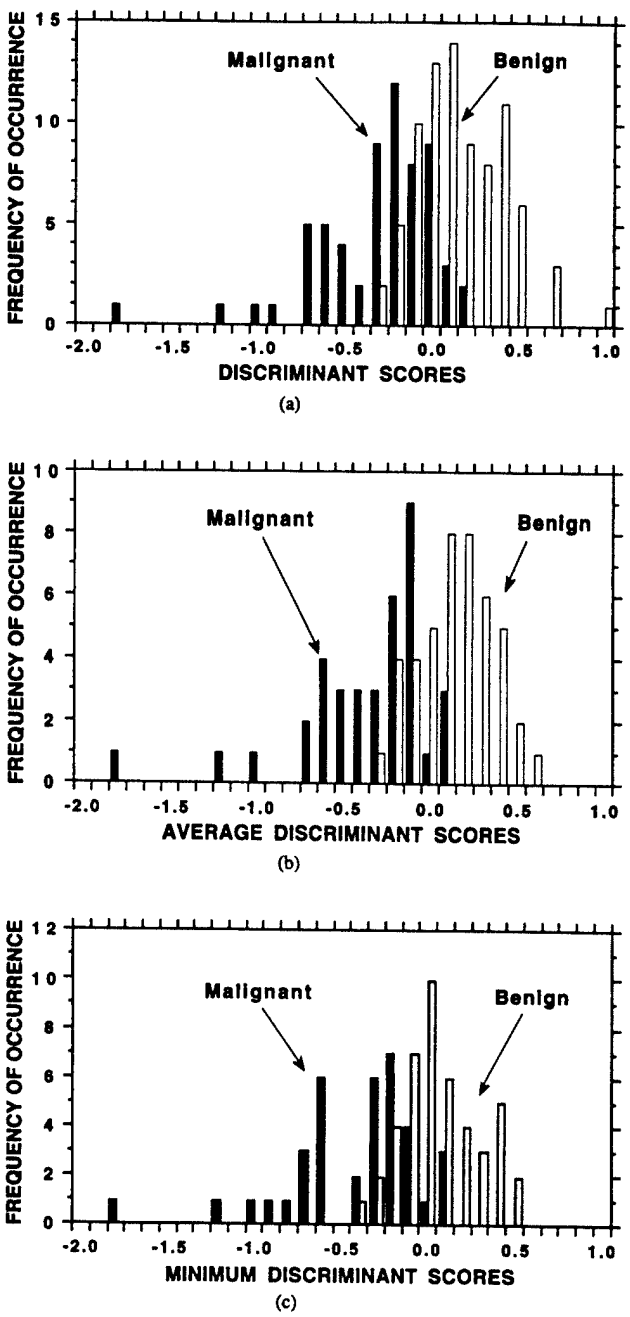


FIG. 6. Distribution of the discriminant scores for the test samples using the best GA selected feature set in the combined texture and morphological feature space. (a) Classification by samples from each film, (b) classification by cluster using the average scores, (c) classification by cluster using the minimum scores.

and 0.89 ± 0.04, respectively, for the GA selected and stepwise LDA selected feature sets. Using the minimum scores, the test A_z values were 0.90 ± 0.03 and 0.85 ± 0.04, respectively. The difference between the A_z values from the two feature selection methods did not achieve statistical significance in either case ($p=0.07$ and $p=0.09$, respectively). If a decision threshold is chosen at an average score of 0.2, 22 of the 44 (50%) benign clusters can be correctly identified with 100% correct classification of the malignant clusters. If a decision threshold is set at a minimum score of 0.2, 14 of the

44 (32%) benign clusters can be identified at 100% sensitivity.

IV. DISCUSSION

The Fischer's linear discriminant is the optimal classifier if the class distributions are multivariate normal with equal covariance matrices.[42] Even if these conditions are not satisfied, as in most classification tasks, the LDA may still be a preferred choice when the number of available training samples is small. Our previous investigation[43,45] of the dependence of classifier performance on design sample size indicated that, in general, the training performance (resubstitution) of a classifier is positively biased whereas the test performance (hold-out) is negatively biased by the sample size. The magnitudes of the biases increase when the dimensionality of the input feature space or the complexity of the classifier increases, or when the design sample size decreases. Therefore, the test performance of a linear classifier is generally better than that of a more complex classifier such as a neural network or a quadratic classifier when the training sample size is small. The training results should not be used for comparison of classifier performance because a classifier can often be overtrained and give a near-perfect classification on training samples while the generalization to any unknown test samples is poor. In this study, we evaluated the effectiveness of using the morphological and the texture features extracted from mammograms for classification of a microcalcification cluster. Although we expanded the data set from our previous study, the current data set was still relatively small. We therefore chose to use a linear discriminant classifier for this classification task. Stepwise feature selection or a GA was used to reduce the dimensionality of the feature space.

In the morphological feature space, the features related to three characteristics, mean density, the moment ratio, and the signal area, were chosen most often. The features related to axis ratio, eccentricity, and the number of microcalcifications in a cluster were chosen only when they were combined with texture features. These results indicate the usefulness of classification in multi-dimensional feature spaces. Some features that are not useful by themselves can become effective features when they are combined with other features. The results also indicate that all six characteristics of the microcalcifications designed for this task have some discriminatory power to distinguish malignant and benign microcalcifications. The morphological features are not as effective as the texture features. This is evident from the smaller A_z values in the morphological feature space. However, when the morphological feature space is combined with the texture feature space, the resulting feature set selected from the combined feature space can significantly improve the classification accuracy, in comparison with those from the individual feature spaces.

The SGLD texture features characterize the shape of the SGLD matrix and generally contain information about the image properties such as homogeneity, contrast, the presence of organized structures, as well as the complexity and gray-level transitions within the image.[40] As an example, the entropy feature measures the uniformity of the SGLD matrix. The entropy value is maximum when all the matrix elements are equal. The entropy value is small when large matrix elements concentrate in a small region of the SGLD matrix while the other matrix elements are relatively small. Therefore, large entropy represents a large but random variation of pixel values in an image without regular structures whereas small entropy represents an image with relatively uniform pixel values if the SGLD matrix peaks along the diagonal and an image with regular texture patterns if it peaks off the diagonal. The ambiguity may be resolved when the sum entropy and difference entropy measures are analyzed. Unlike morphological features, it is difficult, in general, to find the direct relationship between a texture measure and the structures seen on an image,[40] and often a combination of several texture measures extracted at different angles and pixel pair distances are required to describe a texture pattern. It may also be noted that some textures can only be described by second-order statistics and may not be distinguishable by human eyes. The feature selection methods are used to empirically find the combination of features that can most effectively distinguish the malignant and benign lesions.

From Table IV, it can be seen that many of the features in the best feature sets selected by the GA method and the stepwise LDA method are similar. In the morphological feature space, five of the six selected features are the same in the two feature sets. In the combined feature space, six morphological features (out of six and seven morphological features in the two sets, respectively) are the same. For the texture features, there are more variations in the features selected by the two methods. However, the differences are mainly in the pixel distances and the directions of the features, while the major types of the texture features are similar. For example, four types of texture features, energy, entropy, sum average, and sum variance were not selected in either the texture or the combined feature space by both methods. Another four types of texture features, difference average, difference entropy, difference variance, and information measure of correlation 1 were chosen in each case, and information measure of correlation 2 was chosen in three of the four cases. Inertia and inverse difference moment were selected by the stepwise LDA method in both the texture and the combined feature spaces. Sum entropy was selected by both methods in the combined feature space. These results indicate that some features are more effective than the others for distinguishing benign and malignant microcalcifications. The pixel distance and the direction of the texture features may be considered to be higher order effects that have less influence on the discriminatory ability of a given type of texture measure. The smaller differences in their discriminatory ability would subject them to greater variability of being chosen in the feature selection processes. It may also be noted that many of the features are highly correlated. The correlated features can be interchanged in a classifier model without a strong effect on its performance.

The GA solves an optimization problem based on a search guided by the fitness function. Ideally, the values for the P_m,

P_c, and α parameters chosen in the GA only affect the convergence rate but will eventually evolve to the same global maximum. However, when the dimensionality of the feature space is very large and the design samples are sparse, the GA often reaches local maxima corresponding to different feature sets, as can be seen in Table II. Similarly, the stepwise feature selection may reach a different local maximum and choose a feature set different from those chosen by the GA. The different feature sets may provide different or similar performance. The latter is often a result of the correlation among the features, as described above.

For the linear discriminant classifier, the stepwise LDA procedure can select near-optimal features for the classification task. We have shown that the GA could select a feature set comparable to or slightly better than that selected by the stepwise LDA. The number of generations that the GA had to evolve to reach the best selection increased with the dimensionality of the feature space as expected. However, even in a 281-dimensional feature space, it only took 169 generations to find a better feature set than that selected by stepwise LDA. Further search up to 500 generations did not find other feature combinations with better performance. Although the difference in A_z did not achieve statistical significance, probably due to the large standard deviation in A_z when the number of case samples in the ROC analysis was small, the improvements in A_z in this and our previous studies[34] indicate that the GA is a useful feature selection method for classifier design. One of the advantages of GA-based feature selection is that it can search for near-optimal feature sets for any types of linear or nonlinear classifiers, whereas the stepwise LDA procedure is more tailored to linear discriminant classifiers. Furthermore, the fitness function in the GA can be designed such that features with specific characteristics are favored. One of the applications in this direction is to select features to design a classifier with high sensitivity and high specificity for classification of malignant and benign lesions.[49,50] Although the GA requires much longer computation time than the stepwise LDA to search for the best feature set, the flexibility of the GA makes it an increasingly popular alternative for solving machine learning and optimization problems. Since feature selection is performed only during training of a classifier, the speed of a trained classifier for processing test cases is not affected by the choice of the feature selection method. Therefore, the longer computation time of GA is not a problem in practice if the GA can provide a better feature set for a given classification task.

V. CONCLUSIONS

In this study, we evaluated the effectiveness of morphological and texture features extracted from mammograms for classification of malignant and benign microcalcification clusters. We also compared a GA-based feature selection method and a stepwise feature selection procedure based on linear discriminant analysis. It was found that the best feature set was selected from the combined morphological and texture feature space by the GA-based method. A linear discriminant classifier using the best feature set and a properly chosen decision threshold could correctly identify 35% of the benign clusters without missing any malignant clusters. If the average discriminant score from all views of the same cluster was used for classification, the accuracy improved to 50% specificity at 100% sensitivity. Alternatively, if the minimum discriminant score from all views of the same cluster was used, the accuracy would be 32% specificity at 100% sensitivity. This information may be used to reduce unnecessary biopsies, thereby improving the positive predictive value of mammography. Although these results were obtained with a relatively small data set, they demonstrate the potential of using CAD techniques to analyze mammograms and to assist radiologists in making diagnostic decisions. Further studies will be conducted to evaluate the generalizability of our approach in large data sets.

ACKNOWLEDGMENTS

This work is supported by USPHS Grant No. CA 48129 and by U.S. Army Medical Research and Materiel Command Grant No. DAMD 17-96-1-6254. Berkman Sahiner is also supported by a Career Development Award by the U.S. Army Medical Research and Materiel Command (DAMD 17-96-1-6012). Nicholas Petrick is also supported by a grant from The Whitaker Foundation. The content of this publication does not necessarily reflect the position of the government and no official endorsement of any equipment and product of any companies mentioned in the publication should be inferred. The authors are grateful to Charles E. Metz, Ph.D. for use of the LABROC and CLABROC programs.

[a]Electronic mail: chanhp@umich.edu

[1]D. D. Adler and M. A. Helvie, "Mammographic biopsy recommendations," Current Opinion in Radiology **4**, 123–129 (1992).
[2]D. B. Kopans, "The positive predictive value of mammography," Am. J. Roentgenol. **158**, 521–526 (1991).
[3]M. Sabel and H. Aichinger, "Recent developments in breast imaging," Phys. Med. Biol. **41**, 315–368 (1996).
[4]F. Shtern, C. Stelling, B. Goldberg, and R. Hawkins, "Novel technologies in breast imaging: National Cancer Institute perspective," Society of Breast Imaging, Orlando, Florida, 153–156 (1995).
[5]L. V. Ackerman and E. E. Gose, "Breast lesion classification by computer and xeroradiograph," Cancer **30**, 1025–1035 (1972).
[6]J. Kilday, F. Palmieri, and M. D. Fox, "Classifying mammographic lesions using computerized image analysis," IEEE Trans. Med. Imaging **12**, 664–669 (1993).
[7]Z. Huo, M. L. Giger, C. J. Vyborny, U. Bick, P. Lu, D. E. Wolverton, and R. A. Schmidt, "Analysis of spiculation in the computerized classification of mammographic masses," Med. Phys. **22**, 1569–1579 (1995).
[8]B. Sahiner, H. P. Chan, N. Petrick, M. A. Helvie, D. D. Adler, and M. M. Goodsitt, "Classification of masses on mammograms using rubber-band straightening transform and feature analysis," Proc. SPIE **2710**, 44–50 (1996).
[9]B. Sahiner, H. P. Chan, N. Petrick, M. A. Helvie, and M. M. Goodsitt, "Computerized characterization of masses on mammograms: The rubber-band straightening transform and texture analysis," Med. Phys. **25**, 516–526 (1998).
[10]S. Pohlman, K. A. Powell, N. A. Obuchowshi, W. A. Chilote, and S. Grundfest-Broniatowski, "Quantitative classification of breast tumors in digitized mammograms," Med. Phys. **23**, 1337–1345 (1996).
[11]W. G. Wee, M. Moskowitz, N.-C. Chang, Y.-C. Ting, and S. Pemmeraju, "Evaluation of mammographic calcifications using a computer program," Radiology **116**, 717–720 (1975).

[12]S. H. Fox, U. M. Pujare, W. G. Wee, M. Moskowitz, and R. V. P. Hutter, "A computer analysis of mammographic microcalcifications: global approach.," Proceedings of the IEEE 5th International Conference on Pattern Recognition., IEEE, New York, 624–631 (1980).
[13]H. P. Chan, L. T. Niklason, D. M. Ikeda, and D. D. Adler, "Computer-aided diagnosis in mammography: Detection and characterization of microcalcifications," Med. Phys. **19**, 831 (1992).
[14]H. P. Chan, D. Wei, L. T. Niklason, M. A. Helvie, K. L. Lam, M. M. Goodsitt, and D. D. Adler, "Computer-aided classification of malignant/benign microcalcifications in mammography," Med. Phys. **21**, 875 (1994).
[15]H. P. Chan, D. Wei, K. L. Lam, S.-C. B. Lo, B. Sahiner, M. A. Helvie, and D. D. Adler, "Computerized detection and classification of microcalcifications on mammograms," Proc. SPIE **2434**, 612–620 (1995).
[16]H. P. Chan, B. Sahiner, K. L. Lam, D. Wei, M. A. Helvie, and D. D. Adler, "Classification of malignant and benign microcalcifications on mammograms using an artificial neural network," Proc. of World Congress on Neural Networks II, 889–892 (1995).
[17]H. P. Chan, D. Wei, K. L. Lam, B. Sahiner, M. A. Helvie, D. D. Adler, and M. M. Goodsitt, "Classification of malignant and benign microcalcifications by texture analysis," Med. Phys. **22**, 938 (1995).
[18]H. P. Chan, B. Sahiner, D. Wei, M. A. Helvie, D. D. Adler, and K. L. Lam, "Computer-aided diagnosis in mammography: Effect of feature classifier on characterization of microcalcifications," Radiology **197(P)**, 425 (1995).
[19]L. Shen, R. M. Rangayyan, and J. E. L. Desautels, "Application of shape analysis to mammographic calcifications," IEEE Trans. Med. Imaging **13**, 263–274 (1994).
[20]Y. Wu, M. T. Freedman, A. Hasegawa, R. A. Zuurbier, S. C. B. Lo, and S. K. Mun, "Classification of microcalcifications in radiographs of pathologic specimens for the diagnosis of breast cancer," Academic Radiology **2**, 199–204 (1995).
[21]Y. Jiang, R. M. Nishikawa, D. E. Wolverton, C. E. Metz, M. L. Giger, R. A. Schmidt, C. J. Vyborny, and K. Doi, "Malignant and benign clustered microcalcifications: automated feature analysis and classification," Radiology **198**, 671–678 (1996).
[22]D. L. Thiele, C. Kimme-Smith, T. D. Johnson, M. McCombs, and L. W. Bassett, "Using tissue texture surrounding calcification clusters to predict benign vs malignant outcomes," Med. Phys. **23**, 549–555 (1996).
[23]A. P. Dhawan, Y. Chitre, C. Kaiser-Bonasso, and M. Moskowitz, "Analysis of mammographic microcalcifications using gray-level image structure features," IEEE Trans. Med. Imaging **15**, 246–259 (1996).
[24]L. V. Ackerman, A. N. Mucciardi, E. E. Gose, and F. S. Alcorn, "Classification of benign and malignant breast tumors on the basis of 36 radiographic properties," Cancer **31**, 342 (1973).
[25]A. G. Gale, E. J. Roebuck, P. Riley, and B. S. Worthington, "Computer aids to mammographic diagnosis," Br. J. Radiol. **60**, 887–891 (1987).
[26]D. J. Getty, R. M. Pickett, C. J. D'Orsi, and J. A. Swets, "Enhanced interpretation of diagnostic images," Invest. Radiol. **23**, 240 (1988).
[27]C. J. D'Orsi, D. J. Getty, J. A. Swets, R. M. Pickett, S. E. Seltzer, and B. J. McNeil, "Reading and decision aids for improved accuracy and standardization of mammographic diagnosis," Radiology **184**, 619–622 (1992).
[28]Y. Wu, M. L. Giger, K. Doi, C. J. Vyborny, R. A. Schmidt, and C. E. Metz, "Artificial neural networks in mammography: Application to decision making in the diagnosis of breast cancer," Radiology **187**, 81–87 (1993).
[29]J. A. Baker, P. J. Kornguth, J. Y. Lo, M. E. Williford, and C. E. Floyd, "Breast cancer: Prediction with artificial neural network based on BI-RADS standardization lexicon," Radiology **196**, 817–822 (1995).
[30]J. Y. Lo, J. A. Baker, P. J. Kornguth, J. D. Iglehart, and C. E. Floyd, "Predicting breast cancer invasion with artificial neural networks on the basis of mammographic features," Radiology **203**, 159–163 (1997).
[31]H. P. Chan, B. Sahiner, N. Petrick, M. A. Helvie, K. L. Lam, D. D. Adler, and M. M. Goodsitt, "Computerized classification of malignant and benign microcalcifications on mammograms: texture analysis using an artificial neural network," Phys. Med. Biol. **42**, 549–567 (1997).
[32]J. H. Holland, *Adaptation in Natural and Artificial Systems* (University of Michigan Press, Ann Arbor, MI, 1975).
[33]D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning* (Addison-Wesley, New York, 1989).
[34]B. Sahiner, H. P. Chan, N. Petrick, D. Wei, M. A. Helvie, D. D. Adler, and M. M. Goodsitt, "Image feature selection by a genetic algorithm: Application to classification of mass and normal breast tissue on mammograms," Med. Phys. **23**, 1671–1684 (1996).
[35]M. J. Norusis, *SPSS for Windows Release 6 Professional Statistics* (SPSS Inc., Chicago, IL, 1993).
[36]P. A. Lachenbruch, *Discriminant Analysis* (Hafner, New York, 1975), Chaps. 1, 3.
[37]C. E. Metz, J. H. Shen, and B. A. Herman, "New methods for estimating a binormal ROC curve from continuously-distributed test results," Annual Meeting of the American Statistical Association, Anaheim, CA (1990).
[38]H. P. Chan, S. C. B. Lo, B. Sahiner, K. L. Lam, and M. A. Helvie, "Computer-aided detection of mammographic microcalcifications: Pattern recognition with an artificial neural network," Med. Phys. **22**, 1555–1567 (1995).
[39]B. Sahiner, H. P. Chan, N. Petrick, D. Wei, M. A. Helvie, D. D. Adler, and M. M. Goodsitt, "Classification of mass and normal breast tissue: A convolution neural network classifier with spatial domain and texture images," IEEE Trans. Med. Imaging **15**, 598–610 (1996).
[40]R. M. Haralick, K. Shanmugam, and I. Dinstein, "Texture features for image classification," IEEE Trans. Syst. Man Cybern. **SMC-3**, 610–621 (1973).
[41]H. P. Chan, D. Wei, M. A. Helvie, B. Sahiner, D. D. Adler, M. M. Goodsitt, and N. Petrick, "Computer-aided classification of mammographic masses and normal tissue: Linear discriminant analysis in texture feature space," Phys. Med. Biol. **40**, 857–876 (1995).
[42]K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. (Academic, New York, 1990), Chap. 3.
[43]H. P. Chan, B. Sahiner, R. F. Wagner, N. Petrick, and J. Mossoba, "Effects of sample size on classifier design: Quadratic and neural network classifiers," Proc. SPIE **3034**, 1102–1113 (1997).
[44]H. P. Chan, B. Sahiner, R. F. Wagner, and N. Petrick, "Classifier design for computer-aided diagnosis in mammography: Effects of finite sample size," Med. Phys. **24**, 1034–1035 (1997).
[45]R. F. Wagner, H. P. Chan, B. Sahiner, N. Petrick, and J. T. Mossoba, "Finite-sample effects and resampling plans: Applications to linear classifiers in computer-aided diagnosis," Proc. SPIE **3034**, 467–477 (1997).
[46]F. J. Ferri, P. Pudil, M. Hatef, and J. Kittler, "Comparative study of techniques for large-scale feature selection," Pattern Recognition in Practice IV, 403–413 (1994).
[47]W. Siedlecki and J. Sklansky, "A note on genetic algorithm for large-scale feature selection," Pattern Recogn. Lett. **10**, 335–347 (1989).
[48]C. E. Metz, P. L. Wang, and H. B. Kronman, "A new approach for testing the significance for differences between ROC curves measured from correlated data," in *Information Processing in Medical Imaging*, edited by F. Deconinck (The Hague, Martinus Nijhoff, 1984), pp. 432–445.
[49]B. Sahiner, H. P. Chan, N. Petrick, M. A. Helvie, D. D. Adler, and M. M. Goodsitt, "Classification of malignant and benign breast masses: Development of a high-sensitivity classifier using a genetic algorithm," Radiology **201**, 256–257 (1996).
[50]B. Sahiner, H. P. Chan, N. Petrick, M. A. Helvie, and M. M. Goodsitt, "Design of a high-sensitivity classifier based on genetic algorithm: Application to computer-aided diagnosis," Phys. Med. Biol. **43**, 2853–2871 (1998).

Computerized Radiographic Mass Detection—Part I: Lesion Site Selection by Morphological Enhancement and Contextual Segmentation

Huai Li, Yue Wang, K. J. Ray Liu*, Shih-Chung B. Lo, and Matthew T. Freedman

Abstract—**This paper presents a statistical model supported approach for enhanced segmentation and extraction of suspicious mass areas from mammographic images. With an appropriate statistical description of various discriminate characteristics of both true and false candidates from the localized areas, an improved mass detection may be achieved in computer-assisted diagnosis (CAD). In this study, one type of morphological operation is derived to enhance disease patterns of suspected masses by cleaning up unrelated background clutters, and a model-based image segmentation is performed to localize the suspected mass areas using stochastic relaxation labeling scheme. We discuss the importance of model selection when a finite generalized Gaussian mixture is employed, and use the information theoretic criteria to determine the optimal model structure and parameters. Examples are presented to show the effectiveness of the proposed methods on mass lesion enhancement and segmentation when applied to mammographical images. Experimental results demonstrate that the proposed method achieves a very satisfactory performance as a preprocessing procedure for mass detection in CAD.**

Index Terms—**Finite mixture, image enhancement, image segmentation, information criterion, morphological filtering, relaxation labeling.**

I. Introduction

In RECENT years, several computer-assisted diagnosis (CAD) schemes for mass detection and classification have been developed [1]–[13]. Though it may be difficult to compare the relative performance of these methods, because the reported performance strongly depends on the degree of subtlety of masses in the selected database, accurate selection of suspected masses is considered a critical and first step due to the variability of normal breast tissue and the lower contrast and ill-defined margins of masses [3], [6], and since no subtle masses should be missed before any further analysis.

A number of image processing techniques have been proposed to perform suspicious mass site selection. Kobatake *et al.* [1] proposed using a iris filter to detect tumors as suspicious regions with very weak contrast to their background. Sameti *et al.* [7] used fuzzy sets to partition the mammographic image data. Lau and Yin *et al.* independently proposed using bilateral-subtraction to determine possible mass locations [9], [13]. Some other investigators proposed using pixel-based feature segmentation of spiculated masses [4], [8]. Kegelmeyer has reported promising results for detecting spiculated tumors based on local edge characteristics and Laws texture features [8]. Karssemeijer *et al.* [4] proposed to identify stellate distortions by using the orientation map of line-like structures. Recently, Petrick *et al.* [6] proposed a two-stage adaptive density-weighted contrast enhancement filtering technique along with edge detection and morphological feature classification for automatic segmentation of potential masses. Kupinski and Giger [3] presented a radial gradient index-based algorithm and a probabilistic algorithm for seeded lesion segmentation.

Nevertheless, to our best knowledge, few work has been dedicated to improve the task of lesion site selection although it is indeed a very crucial step in CAD. Especially, few studies have used and justified model-based image processing techniques for unsupervised lesion site selection [11]. Zwiggelaar *et al.* developed a statistical model to describe and detect the abnormal pattern of linear structures of spiculated lesions [2]. In their work, the probability density function of the observation vectors for each class is assumed to be normal. we have experienced that the "normal" distribution for each class is nor true. Li *et al.* proposed using a Markov random field model to extract suspicious masses for mass detection [11]. In their study, most of model parameters were chosen empirically, and the mammogram was segmented into three regions (background, fat, and parenchymal or tumors).

Stochastic model-based image segmentation is a technique for partitioning an image into distinctive meaningful regions based on the statistical properties of both gray level and context images. A good segmentation result would depend on suitable model selection for a specific image modality [16], [17] where model selection refers to the determination of both the number of image regions and the local statistical distributions of each region. Furthermore, a segmentation result would be improved

Manuscript received February 3, 1997; revised January 9, 2001. This work was supported in part by the Department of Defense under Grants DAMD17-98-1-8045 and DAMD17-96-1-6254 through a subcontract from University of Michigan, Ann Arbor, and by the National Science Foundation (NSF) under NYI Award MIP-9457397. The Associate Editor responsible for coordinating the review of this paper and recommending its publication was M. Giger. *Asterisk indicates corresponding author.*

H. Li is with the Electrical Engineering Department and Institute for Systems Research, University of Maryland at College Park, College Park, MD 20742 USA. He is also with the Department of Radiology, Georgetown University Medical Center, Washington, DC 20007 USA.

Y. Wang is with the Department of Electrical Engineering and Computer Science, The Catholic University of America, Washington, DC 20064 USA. He is also with the Department of Radiology, Georgetown University Medical Center, Washington, DC 20007 USA.

*K. J. Ray Liu is with the Electrical Engineering Department and Institute for Systems Research University of Maryland at College Park, College Park, MD 20742 USA (e-mail: kjrliu@eng.umd.edu).

S.-C. B. Lo and M. T. Freedman are with the Department of Radiology, Georgetown University Medical Center, Washington, DC 20007 USA.

Publisher Item Identifier S 0278-0062(01)02831-2.

0278–0062/01$10.00 © 2001 IEEE

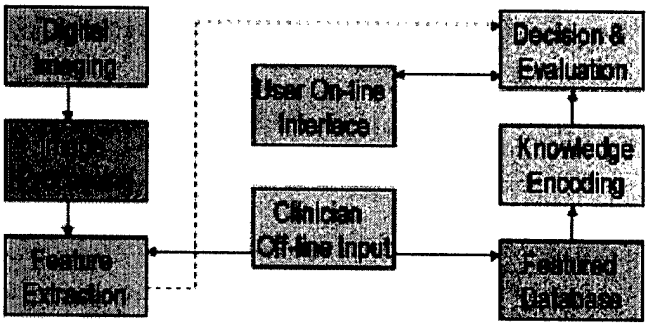


Fig. 1. Major components in CAD.

with preenhanced pattern of interest being segmented. The only assumption for suspected mass site selection is that suspected mass areas should be brighter than the surrounding breast tissues which is valid for most of the real cases. When some masses lie either within an inhomogeneous pattern of fibroglandular tissue or are partially or completely surrounded by fibroglandular tissue, enhancement of mass-related signals is important.

Fig. 1 shows a general block diagram of CAD systems. This paper focuses on "image processing" block, to just automatically pick up all possible lesion sites. We aim on two essential issues in the stochastic model-based image segmentation: enhancement and model selection. Based on the differential geometric characteristics of masses against the background tissues, we propose one type of morphological operation to enhance the mass patterns on mammograms. Then we employ a finite generalized Gaussian mixture (FGGM) distribution to model the histogram of the mammograms where the statistical properties of the pixel images are largely unknown and are to be incorporated. We incorporate the EM algorithm with two information theoretic criteria to determine the optimal number of image regions and the kernel shape in the FGGM model. Finally, we apply a contextual Bayesian relaxation labeling (CBRL) technique to perform the selection of suspected masses. The major differences of our work from the previous work [1]–[6], [8]–[13] are as follows.

1) We present a new algorithm of morphological filtering for image enhancement in which the combined operations are applied to the original gray tone image and the higher sensitive lesion site selection of the enhanced images are observed.
2) We justify and pilot test the FGGM distribution in modeling mammographic pixel images together with a model selection procedure based on the two information theoretic criteria. This allows an automatic identification of both the number (K) and kernel shape (α) of the distributions of tissue types.
3) We develop a new algorithm (CBRL) for segmenting mass areas where the comparable results are achieved as those using Markov random field model-based approaches while with much less computational complexity.

The presentation of this paper is organized as follows. In Section II, the proposed dual morphological operation enhancement technique is described in detail. The theory and algorithm on FGGM modeling, model selection, and parameter estimation are presented in Section III. This is followed by a discussion on the selection of suspicious masses using the CBRL approach. Evaluation results are given and discussed in Section IV. Finally, the paper is concluded by Section V.

II. Morphological Enhancement

One of the main difficulties in suspicious mass segmentation is that mammographic masses are often overlapped with dense breast tissues. Therefore, it is necessary to remove bright background caused by dense breast tissues while preserving the features and patterns related to the masses. For this purpose, background correction is an important step for mass segmentation. We propose a mass pattern-dependent background removal approach using morphological operations.

A. Morphological Filtering Theory

Morphological operations can be employed for many image processing purposes, including edge detection, region segmentation, and image enhancement. The beauty and simplicity of mathematical morphology approach come from the fact that a large class of filters can be represented as the combination of two simple operations: erosion and dilation. Let Z denote the set of integers and $f(i, j)$ denote a discrete image signal, where the domain set is given by $\{i, j\} \in N_1 \times N_2$, $N_1 \times N_2 \subset Z^2$ and the range set by $\{f\} \in N_3$, $N_3 \subset Z$. A structuring element B is a subset in Z^2 with a simple geometrical shape and size. Denote $B^s = \{-b : b \in B\}$ as the symmetric set of B and B_{t_1,t_2} as the translation of B by (t_1, t_2), where $(t_1, t_2) \in Z^2$. The erosion $f \ominus B^s$ and dilation $f \oplus B^s$ can be expressed as [19]

$$(f \ominus B^s)(i, j) = \min_{t_1,t_2 \in B_{i,j}} (f(t_1, t_2)) \quad (1)$$

$$(f \oplus B^s)(i, j) = \max_{t_1,t_2 \in B_{i,j}} (f(t_1, t_2)). \quad (2)$$

On the other hand, opening $f \circ B$ and closing $f \bullet B$ are defined as [19]

$$(f \circ B)(i, j) = ((f \ominus B^s) \oplus B)(i, j) \quad (3)$$

$$(f \bullet B)(i, j) = ((f \oplus B^s) \ominus B)(i, j). \quad (4)$$

A gray value image can be viewed as a two-dimensional surface in a three-dimensional space. Given an image, the opening operation removes the objects, which have size smaller than the structuring element, with positive intensity. Thus, with the specified structuring element, one can extract different image contexts by taking the difference between the original and opening processed image, which is known as "tophat" operation [19].

B. Morphological Enhancement Algorithms

Based on the properties of morphological filters, we designed one type of mass pattern-dependent enhancement approaches. The algorithm is implemented by dual morphological tophat operations following by a subtraction which is described as follows.

Step 1) The textures without the pattern information of interest are extracted by a tophat operation

$$r_1(i, j) = \max(0, [f(i, j) - (f \circ B_1)(i, j)]) \quad (5)$$

where $f(i, j)$ is the original image, and $r_1(i, j)$ is the residue image between the original image and the opening of the original image by a specified structuring element B_1. The size of B_1 should be chosen smaller than the size of masses.

Step 2) Let $r_2(i, j)$ be the mass pattern enhanced image by background correction, i.e., by the second tophat operation on $f(i, j)$

$$r_2(i, j) = \max(0, [f(i, j) - (f \circ B_2)(i, j)]) \quad (6)$$

where B_2 is a specified structuring element which has a larger size than masses.

Step 3) The enhanced image $f_1(i, j)$ can be derived as

$$f_1(i, j) = \max(0, [r_2(i, j) - r_1(i, j)]). \quad (7)$$

This operation is called "dual morphological operation." It can remove the background noise and the structure noise inside the suspected mass patterns. Fig. 2 shows the mass patch and the enhanced results of each step using the dual morphological operation. As we can see from Fig. 2, both background correction [Fig. 2(c)] and dual morphological operation [Fig. 2(d)] enhanced the mass pattern, but dual morphological operation removed more structural noise inside the mass region which in turn would improve the mass segmentation results.

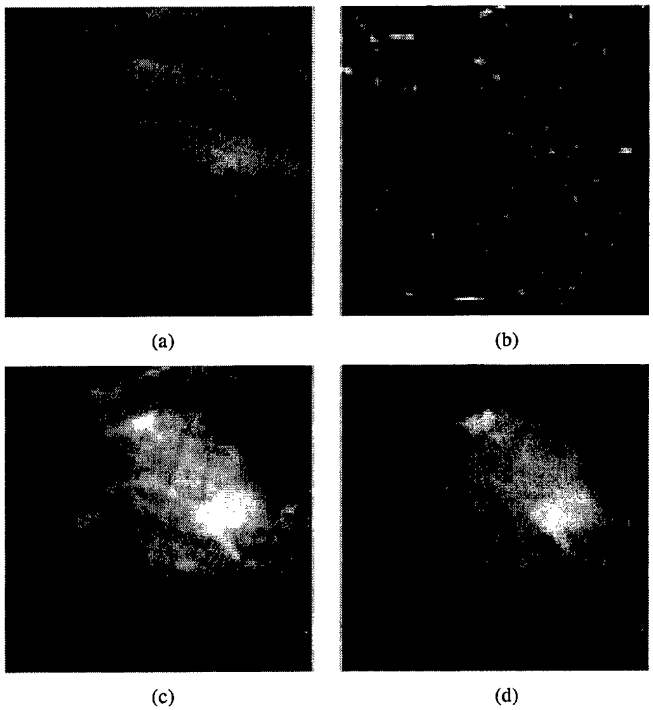


Fig. 2. Original and enhancement result of the mass patch using dual-morphological operation. (a) Original image block $f(i, j)$. (b) Textures $r_1(i, j)$. (c) Background correction result $r_2(i, j)$. (d) Enhanced result $f_1(i, j)$.

III. Model-Based Segmentation

A. Statistical Modeling

Given a digital image consisting of $N_1 \times N_2$ pixels, assume this image contains K regions. By randomly reordering all pixels in the underlying probability space, one can treat pixel labels as random variables and introduce a prior probability measure π_k. Then the FGGM probability density function (pdf) of gray level of each pixel is given by [17]

$$p(x_i) = \sum_{k=1}^{K} \pi_k p_k(x_i), \qquad i = 1, \ldots, N_1 N_2, \quad x_i = 0, 1, \ldots, L-1 \quad (8)$$

where x_i is the gray level of pixel i, and L is the number of gray levels. $p_k(x_i)$s are conditional region pdfs with the weighting factor π_k, satisfying $\pi_k > 0$, and $\sum_{k=1}^{K} \pi_k = 1$. The generalized Gaussian pdf given region k is defined by

$$p_k(x_i) = \frac{\alpha \beta_k}{2\Gamma(1/\alpha)} \exp[-|\beta_k(x_i - \mu_k)|^{\alpha}], \qquad \alpha > 0,$$
$$\beta_k = \frac{1}{\sigma_k} \left[\frac{\Gamma(3/\alpha)}{\Gamma(1/\alpha)}\right]^{1/2}. \quad (9)$$

where μ_k is the mean, $\Gamma(\cdot)$ is the Gamma function. β_k is a parameter related to the variance σ_k. It can be shown that when $\alpha = 2.0$, one has the Gaussian pdf; when $\alpha = 1.0$, one has the Laplacian pdf. When $\alpha \gg 1$, the distribution tends to a uniform pdf; when $\alpha < 1$, the pdf becomes sharp. Therefore, the generalized Gaussian model is a suitable model to fit the histogram distribution of those images whose statistical properties are unknown since the kernel shape can be controlled by selecting different α values.

The whole image can be well approximated by an independent and identically distributed random field $\mathbf{X}$. The corresponding joint pdf is

$$P(\mathbf{x}) = \prod_{i=1}^{N_1 N_2} \sum_{k=1}^{K} \pi_k p_k(x_i) \quad (10)$$

where $\mathbf{x} = [x_1, x_2, \ldots, x_{N_1 N_2}]$, and $\mathbf{x} \in \mathbf{X}$. $p_k(x_i)$ is given in (9). Based on the joint probability measure of pixel images, the likelihood function under FGGM modeling can be expressed as $\mathcal{L}(\mathbf{r}) = \prod_{i=1}^{N_1 N_2} p_{\mathbf{r}}(x_i)$ where $\mathbf{r} : \{K, \alpha, \pi_k, \mu_k, \sigma_k, k = 1, \ldots, K\}$ denotes the model parameter set.

B. Model Identification

With an appropriate system likelihood function, the objective of model identification is to estimate the model parameters by maximizing the likelihood function, or equivalently minimizing the relative entropy between the image histogram $p_{\mathbf{x}}(u)$ and the estimated pdf $p_{\mathbf{r}}(u)$, where u is the gray level. Based on the FGGM model, the EM algorithm is applied to estimate the model parameters. The EM algorithm is an iterative technique for maximum-likelihood (ML) estimation [20]. Recently, it has been used in many medical imaging applications [15]. Instead

of evaluating directly the value of ML, we use the global relative entropy (GRE) between the histogram and the estimated FGGM distribution to measure the performance of parameter estimation, given by

$$\mathrm{GRE}(p_{\mathbf{x}}\|p_{\mathbf{r}}) = \sum_{u} p_{\mathbf{x}}(u)\log\frac{p_{\mathbf{x}}(u)}{p_{\mathbf{r}}(u)}. \tag{11}$$

Motivated by the same spirit of conventional EM algorithm for finite normal mixtures (FNMs), we formulated the EM algorithm to estimate the parameter values of the FGGM. The algorithm is summarized as follows.

EM Algorithm:

1) For $\alpha = \alpha_{\min}, \ldots, \alpha_{\max}$
   - $m = 0$, given initialized $\mathbf{r}^{(0)}$
   - E-step: for $i = 1, \ldots, N_1N_2$, $k = 1, \ldots, K$, compute the probabilistic membership

$$z_{ik}^{(m)} = \frac{\pi_k^{(m)} p_k(x_i)}{\sum_{k=1}^{K} \pi_k^{(m)} p_k(x_i)}. \tag{12}$$

   - M-step: for $k = 1, \ldots, K$, compute the updated parameter estimates

$$\begin{cases} \pi_k^{(m+1)} = \dfrac{1}{N_1N_2}\displaystyle\sum_{i=1}^{N_1N_2} z_{ik}^{(m)} \\ \mu_k^{(m+1)} = \dfrac{1}{N_1N_2\pi_k^{(m+1)}}\displaystyle\sum_{i=1}^{N_1N_2} z_{ik}^{(m)} x_i \\ \sigma_k^{2(m+1)} = \dfrac{1}{N_1N_2\pi_k^{(m+1)}}\displaystyle\sum_{i=1}^{N_1N_2} z_{ik}^{(m)}(x_i - \mu_k^{(m+1)})^2. \end{cases} \tag{13}$$

   - When $|\mathrm{GRE}^{(m)}(p_{\mathbf{x}}\|p_{\mathbf{r}}) - \mathrm{GRE}^{(m+1)}(p_{\mathbf{x}}\|p_{\mathbf{r}})| \le \epsilon$ is satisfied, go to Step 2 Otherwise, $m = m+1$ and go to E-Step.
2) Compute GRE, and go to Step 1.
3) Choose the optimal $\hat{\mathbf{r}}$ which corresponds to the minimum GRE.

As we mentioned in Section I, the two important parameters in model selection are K and α. Determination of the region parameter K directly affects the quality of the resulting model parameter estimation and in turn, affects the result of segmentation. In this paper we propose an approach to determine the value of K based on two popular information theoretic criteria introduced by Akaike [23] and by Rissanen [24]. Akaike proposed to select the model that gives the minimum Akaike information criterion (AIC), defined by

$$\mathrm{AIC}(K) = -2\log(\mathcal{L}(\hat{\mathbf{r}}_{ML})) + 2K' \tag{14}$$

where $\hat{\mathbf{r}}_{ML}$ is the ML estimate of the model parameter set $\mathbf{r}$, and K' is the number of free adjustable parameters in the model [15], [23]. AIC criterion will select the correct number of the image regions K_0 when

$$K_0 = \arg\left\{\min_{1\le K\le K_{\mathrm{MAX}}} \mathrm{AIC}(K)\right\}. \tag{15}$$

Rissanen addressed the problem from a quite different point of view. Rissanen reformulated the problem explicitly as an information coding problem in which the best model fitness is measured such that it assigns high probabilities to the observed data while at the same time the model itself is not too complex to describe [24]. The model is selected by minimizing the total description length defined by minimum description length (MDL)

$$MDL(K) = -\log(\mathcal{L}(\hat{\mathbf{r}}_{ML})) + 0.5K'\log(N_1N_2). \tag{16}$$

Similarly, the correct number of the distinctive image regions K_0 will be estimated when

$$K_0 = \arg\left\{\min_{1\le K\le K_{\mathrm{MAX}}} MDL(K)\right\}. \tag{17}$$

C. Bayesian Relaxation Labeling

Once the FGGM model is given, a segmentation problem is the assignment of labels to each pixel in the image. A straightforward way is to label pixels into different regions by maximizing the individual likelihood function $p_k(x)$. This approach is called ML classifier, which is equivalent to a multiple thresholding method. Usually, this method may not achieve a good performance since there is lack of local neighborhood information to be included to make a good decision. CBRL algorithm [25] is one of the approaches, which can incorporate the local neighborhood information into labeling procedure and thus improve the segmentation performance. In this study, we developed the CBRL algorithm to perform/refine pixel labeling based on the localized FGGM model, which is defined as follows.

Let ∂i be the neighborhood of pixel i with an $m\times m$ template centered at pixel i. An indicator function is used to represent the local neighborhood constraints $R_{ij}(l_i, l_j) = I(l_i, l_j)$, where l_i and l_j are labels of pixels i and j, respectively. Note that pairs of labels are now either compatible or incompatible. Similar to reference [25], one can compute the frequency of neighbors of pixel i which has the same label values k as at pixel i

$$\pi_k^{(i)} = p(l_i = k|\mathbf{l}_{\partial i}) = \frac{1}{m^2-1}\sum_{j\in\partial i,\, j\ne i} I(k, l_j) \tag{18}$$

where $\mathbf{l}_{\partial i}$ denotes the labels of the neighbors of pixel i. Since $\pi_k^{(i)}$ is a conditional probability of a region, the localized FGGM pdf of gray level x_i at pixel i is given by

$$p(x_i|\mathbf{l}_{\partial i}) = \sum_{k=1}^{K} \pi_k^{(i)} p_k(x_i) \tag{19}$$

where $p_k(x_i)$ is given in (9). Assuming gray values of the image are conditional independent, the joint pdf of $\mathbf{x}$, given the context labels $\mathbf{l}$, is

$$P(\mathbf{x}|\mathbf{l}) = \prod_{i=1}^{N_1N_2}\sum_{k=1}^{K} \pi_k^{(i)} p_k(x_i) \tag{20}$$

where $\mathbf{l} = (l_i : i = 1, \ldots, N_1N_2)$.

It is known that CBRL algorithm can obtain a consistent labeling solution based on the localized FGGM model (19). Since

TABLE I
DISTRIBUTION OF THE EFFECTIVE SIZE OF THE 186 MASSES USED IN THIS STUDY. THE EFFECTIVE SIZE IS DEFINED AS THE SQUARE ROOT OF THE PRODUCT OF THE MAXIMUM AND MINIMUM DIAMETERS OF THE MASS

	$0-5mm$	$6-10mm$	$11-15mm$	$16-20mm$	$21-25mm$	$26-30mm$
#	3	55	78	29	17	4

l represents the labeled image, it is consistent if $S_i(l_i) \geq S_i(k)$, for all $k = 1, \ldots, K$ and for $i = 1, \ldots, N_1N_2$ [25], where

$$S_i(k) = \pi_k^{(i)} p_k(x_i). \tag{21}$$

Now we can define

$$A(\mathrm{l}) = \sum_{i=1}^{N_1N_2} \left(\sum_k I(l_i, k) S_i(k) \right) \tag{22}$$

as the average measure of local consistency, and

$$LC_i = \sum_k I(l_i, k) S_i(k), \qquad i = 1, \ldots, N_1N_2 \tag{23}$$

represents the local consistency based on l. The goal is to find a consistent labeling l which can maximize (22). In the real application, each local consistency measure LC_i can be maximized independently. In [25], it has been shown that when $R_{ij}(l_i, l_j) = R_{ji}(l_j, l_i)$, if $A(\mathrm{l})$ attains a local maximum at l, then l is a consistent labeling.

Based on the localized FGGM model, $l_i^{(0)}$ can be initialized by ML classifier

$$l_i^{(0)} = \arg\left\{ \max_k p_k(x_i) \right\}, \qquad k = 1, \ldots, K. \tag{24}$$

Then, the order of pixels is randomly permutated and each label l_i is updated to maximize LC_i, i.e., classify pixel i into kth region if

$$l_i = \arg\left\{ \max_k \pi_k^{(i)} p_k(x_i) \right\}, \qquad k = 1, \ldots, K \tag{25}$$

where $p_k(x_i)$ is given in (9), $\pi_k^{(i)}$ is given in (18). By considering (24) and (25),we developed a modified CBRL algorithm as follows.

CBRL Algorithm:

1) Given $\mathrm{l}^{(0)}$, $m = 0$
2) Update pixel labels
   - Randomly visit each pixel for $i = 1, \ldots, N_1N_2$
   - Update its label l_i according to

$$l_i^{(m)} = \arg\left\{ \max_k \pi_k^{(i)(m)} p_k(x_i) \right\}.$$

3) When

$$\frac{\sum (\mathrm{l}^{(m+1)} \oplus \mathrm{l}^{(m)})}{N_1N_2} \leq 1\%,$$

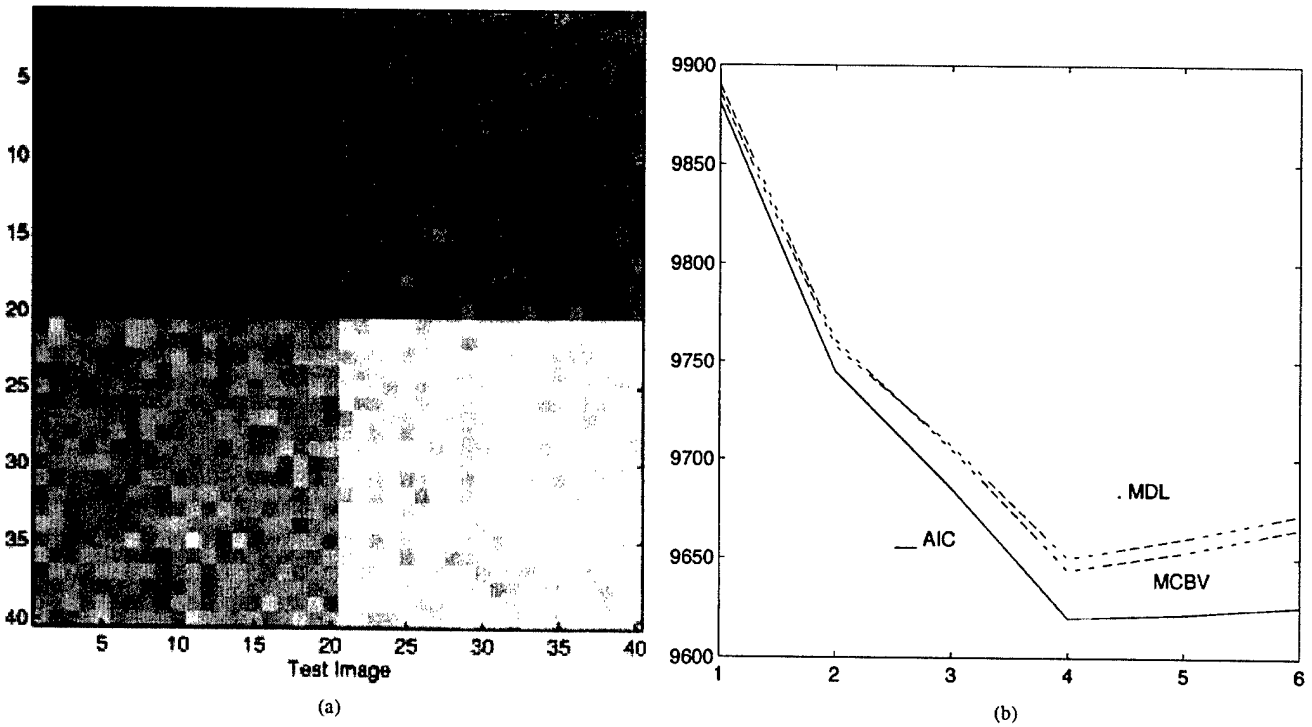
stop; otherwise, $m = m + 1$, and repeat Step 2.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

In this section, we present the results of using the morphological filtering and model-based segmentation approach we have introduced for enhancement and segmentation of suspicious masses in mammographic images. In addition to the qualitative assessment by the radiologists, we introduce several objective measures to assess the performance of the algorithms we have proposed for enhancement and segmentation.

A testing data set of 200 mammograms and two simulated tone images were used to test and evaluate the performance of the algorithms in this study. The mammograms were selected from the Mammographic Image Analysis Society (MIAS) database and the Brook Army Medical Center (BAMC) database created by the Department of Radiology at Georgetown University Medical Center. Of the 200 mammograms, 50 mammograms are normal, and each of the 150 abnormal mammograms contains at least one mass case of varying size, subtlety, and location. The areas of suspicious masses were identified by an expert radiologist based on visual criteria and biopsy proven results. The total data set includes 113 benign and 73 malignant masses. The distribution of the masses in terms of size is shown in Table I. The BAMC films were digitized with a laser film digitizer (Lumiscan 150) at a pixel size of 100 μm$\times$ 100 μm and 4096 gray levels (12 bits). Before the method was applied the digital mammograms were smoothed by averaging 4 $\times$ 4 pixels into one pixel. According to radiologists, the size of small masses is 3–15 mm in effective diameter. A 3-mm object in an original mammogram occupies 30 pixels in a digitized image with a 100-μm resolution. After reducing the image size by four times, the object will occupy the range of about seven to eight pixels. The object with the size of seven pixels is expected to be detectable by any computer algorithm. Therefore, the shrinking step is applicable for mass cases and can save computation time.

Experimental Evaluation of Morphological Enhancement: In order to justify the suitability of morphological structural elements, the geometric properties of the contexts and textures in mammograms were studied. The basic idea is to keep all mass-like objects within certain size range and remove all others by using the proposed morphological filters with specific structural elements. At the resolution of 400 μm, a disk with a diameter of seven pixels was chosen as the morphological structuring elements B_1 to extract textures in mammograms. Since the smallest masses have seven pixels in diameter with the resolution of 400 μm, this procedure would not destroy mass information. For the purpose of background correction, a disk with a diameter of 75 pixels was used as the morphological structuring element B_2. An object with a diameter of 75 pixels corresponds to 30 mm in the original mammogram. This indicates that all masses with sizes up to 30 mm can be enhanced by background correction. Masses larger than 30 mm are rare cases in the clinical setting. In the last stage of our approach, we applied morphological opening and closing filtering using a disk with a diameter of five to eliminate small objects which also contribute to texture noise.



(a) (b)

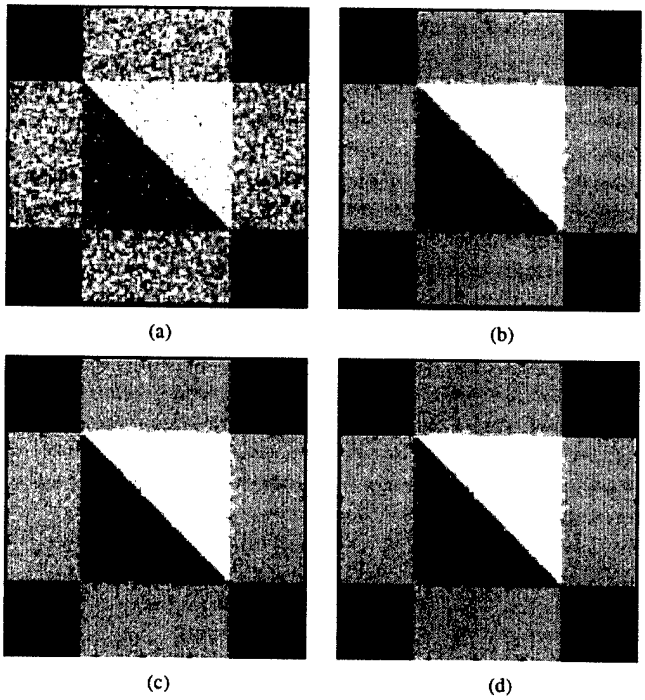
Fig. 3. (a) Original simulated test image for model selection ($k_0 = 4$, SNR = 10 dB) and (b) the AIC/MDL curves in model selection ($\sigma = 30$).

All testing mammograms were processed using the proposed enhancement approach with the suggested structuring element B_1 and B_2. Fig. 5 shows processed mammogram examples using the morphological enhancement. Compared the enhanced results [Fig. 5(b) and (d)] with the original mammograms [Fig. 5(a) and (c)], the proposed method not only enhanced all suspected mass patterns and reduced the texture noise, but also removed the background noise. In summary, the proposed morphological enhancement approach can enhance mass patterns and remove texture structure noises. For dense mammograms, such as the second example in Fig. 5(c) and (d), the mass is obscured by dense fibroglandular tissues, our experience shows applying the dual morphological operation to remove the fibroglandular tissue background is useful. In addition to the visual evaluation by the radiologist, we performed the segmentation to assess the effectiveness of the morphological filtering, based on the enhanced mammograms and the original mammograms.

Simulated Evaluation of Segmentation Algorithms: The performance of model selection using two frequently used methods, i.e., the AIC and MDL [22], were first tested and compared in the simulation study. The computer-generated data was made up of four overlapping normal components. Each component represents one local region. The value for each component were set to a constant value, the noise of normal distribution was then added to this simulation digital phantom. Three noise levels with different variance were set to keep the same signal-to-noise ratio (SNR), where SNR is defined by

$$\text{SNR} = 10 \log_{10} \frac{(\Delta\mu)^2}{\sigma^2} \tag{26}$$

where $\Delta\mu$ is the mean difference between regions, and σ^2 is the noise power. The original data for the simulation study are given in Fig. 3(a). The AIC and MDL curves, as functions of the number of local clusters K, are plotted in Fig. 3(b). According

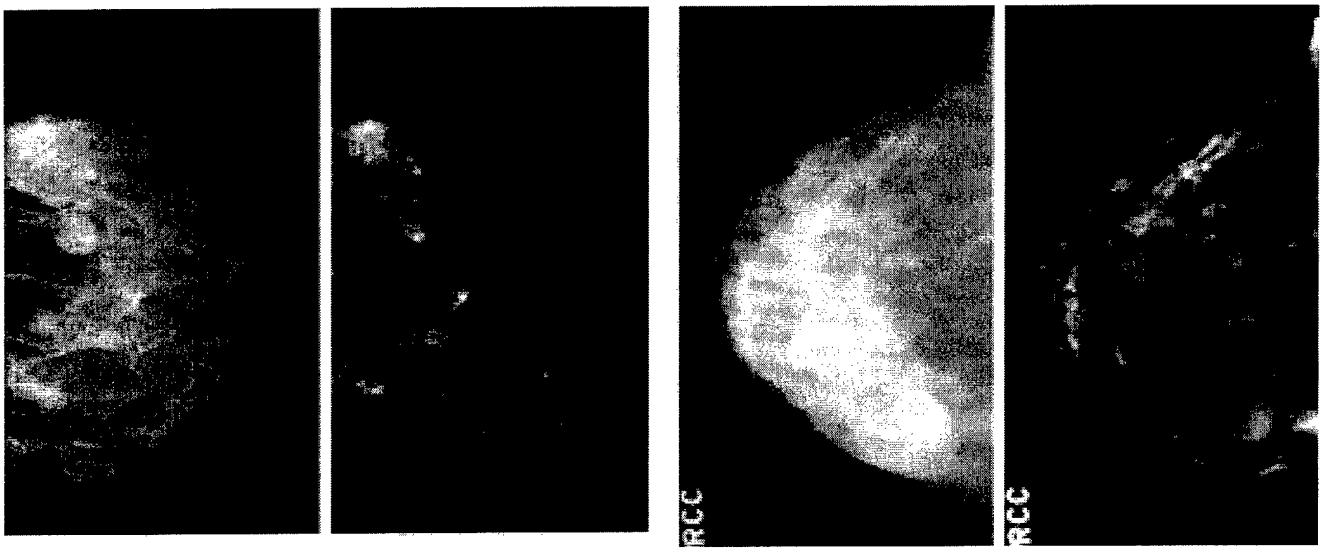


(a) (b) (c) (d)

Fig. 4. Image segmentation by CBRL on simulated image (with initialization by ML classification). (a) ML initialization. (b) First iteration in CBRL. (c) Second iteration in CBRL. (d) Third iteration in CBRL.

TABLE II
COMPARISON OF CBRL, ICM, AND MICM ALGORITHM: SIMULATED DATA

Item	CBRL Result	ICM Result	MICM Result
Classification Error	0.7935%	0.7508%	0.3113%



(a) (b) (c) (d)

Fig. 5. Examples of mass enhancement. (a) Original mammogram. (b) Enhanced mammogram. (c) and (d) Another original mammogram and its enhanced result.

to the information theoretic criteria, the minima of these curves indicate the correct number of the local regions. From this experimental figure, it is clear that the number of local regions suggested by these criteria are all correct.

For the validation of image segmentation using CBRL, we apply the algorithm first to a simulated image. We use ML classifier to initialize image segmentation, i.e., to initialize the quantified image by selecting the pixel label with largest likelihood at each node. The classification error after initialization is uniformly distributed over the spatial domain as shown in Fig. 4(a). Our experience suggested this to be a very suitable starting point for contextual relaxation labeling [21]. The CBRL is then performed to fine tune the image segmentation. It should be emphasized that the ground truth is known in this simulated experiment, the percentage of total classification error is used as the criterion for evaluating the performance of segmentation technique. In Fig. 4(a)–(d), the initial segmentation by the ML classification and the stepwise results of three iterations in the CBRL are presented. In this experiment, algorithm initialization results in an average classification error of 30%. It can be clearly seen that a dramatic improvement is obtained after several iterations of the CBRL by using local constraints determined by the context information. In addition, the convergence is fast as one can see, after the first iteration most of the misclassification are removed. We have also implemented two other independent and popular algorithms, namely, the iterated conditional mode (ICM) and the modified iterated conditional mode (MICM) algorithms, so as to assess the comparative performance of the segmentation results among different approaches [21], [22]. The only assumption being made by these three methods is the Markovian property of the context images which can be well justified by the underlying cell oncology and pathology. We have applied these three algorithms to the same testing image and the corresponding classification errors are presented in Table II. The final percentage of classification errors for Fig. 4(d) is 0.7935%. From this experimental comparison, it can be concluded that three algorithms achieved comparable segmentation accuracy and the result produced by the MICM algorithm is most superior, though in terms of computational complexity the CBRL algorithm is the least. It should be noticed that since in MICM algorithm an inhomogeneous configuration of the Markov random field is used, its superior performance is reasonable.

TABLE III
COMPUTED AICS FOR THE FGGM MODEL WITH DIFFERENT α

K	$\alpha = 1.0$	$\alpha = 2.0$	$\alpha = 3.0$	$\alpha = 4.0$
2	651250	650570	650600	650630
3	646220	644770	645280	646200
4	645760	644720	645260	646060
5	645760	644700	645120	646040
6	645740	644670	645110	645990
7	645640	644600	645090	645900
8	645550(min)	644570(min)	645030(min)	645850(min)
9	645580	644590	645080	645880
10	645620	644600	645100	645910

TABLE IV
COMPUTED MDLS FOR THE FGGM MODEL WITH DIFFERENT α

K	$\alpha = 1.0$	$\alpha = 2.0$	$\alpha = 3.0$	$\alpha = 4.0$
2	651270	650590	650630	650660
3	646260	644810	645360	646350
4	645860	644770	645280	646150
5	645850	644770	645280	646100
6	645790	644750	645150	646090
7	645720	644700	645120	645930
8	645680(min)	644690(min)	645100(min)	645900(min)
9	645710	644710	645140	645930
10	645790	644750	645180	645960

On Model-Based Segmentation—Real Case Study: In the real case study, we used two information criteria (AIC and MDL) to determine K. Tables III and IV shows the AIC and MDL values with different K and α of the FGGM model based on one original mammogram. As it can be seen from Tables III and IV, although with different α, all AIC and MDL values achieve the minimum when $K = 8$. It indicates that AIC and

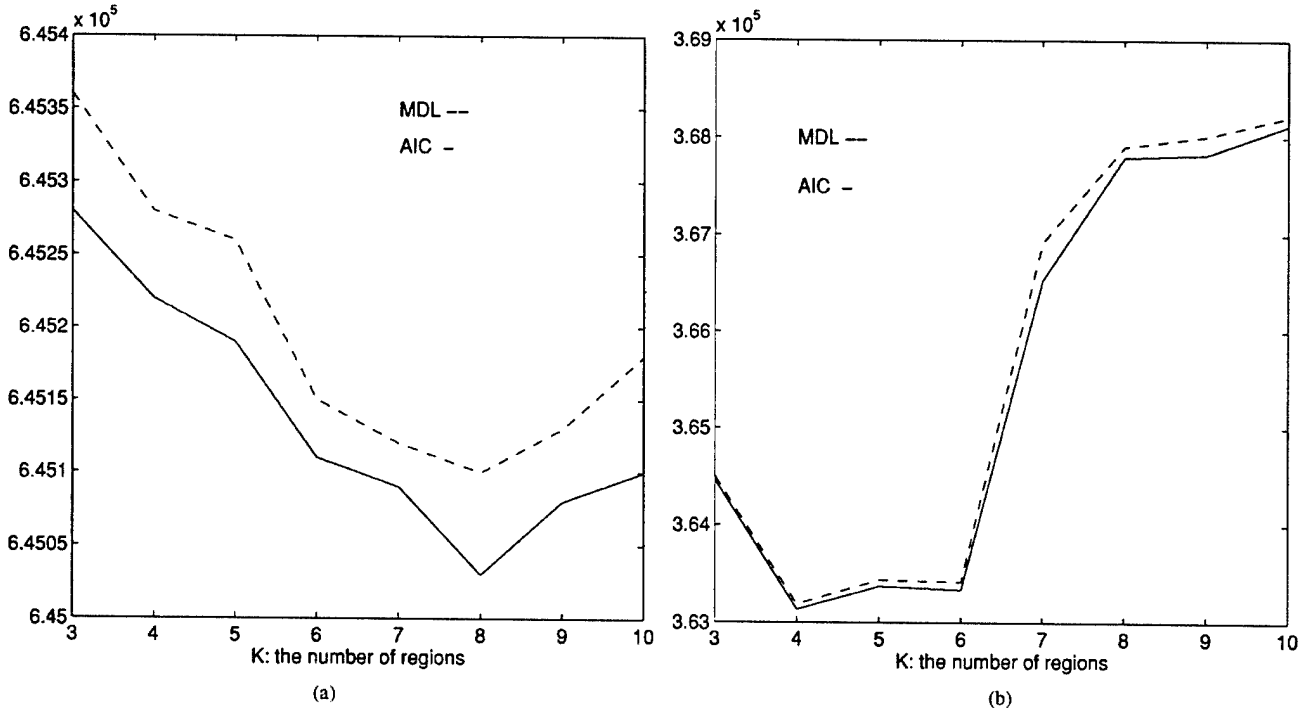


Fig. 6. AIC and MDL curves with different number of region K. (a) Result based on the original mammogram, the optimal $K = 8$. (b) Result based on the enhanced mammogram, the optimal $K = 4$.

MDL are relatively insensitive to the change of α. With this observation, we can decouple the relation between K and α and choose the appropriate value of one while fixing the value of another. Fig. 6(a) and (b) are two examples of AIC and MDL curves with different K and fixed $\alpha = 3.0$. Fig. 6(a) is based on the original mammogram and Fig. 6(b) is based on the enhanced mammogram. As we can see in Fig. 6(a), both criteria achieved the minimum when $K = 8$. It should be noticed that though no ground truth is available in this case, our extensive numerical experiments have shown a very consistent performance of the model selection procedure and all the conclusions were strongly supported by the previous independent work reported by [14]. Fig. 6(b) indicates that $K = 4$is the appropriate choice for the mammogram enhanced by dual morphological operation. This is believed to be reasonable since the number of regions decrease after background correction.

We fixed $K = 8$, and changed the value of α for estimating the FGGM model parameters using the proposed EM algorithm with the original mammogram The GRE value between the histogram and the estimated FGGM distribution was used as a measure of the estimation bias. We found that GRE achieved a minimum distance when the FGGM parameter $\alpha = 3.0$ as shown in Fig. 7. The similar result was shown when we applied the EM algorithm to the enhanced mammogram with $K = 4$. This indicated that the FGGM model might be better than the FNM model ($\alpha = 2.0$) in modeling mammographic images when the true statistical properties of mammograms are generally unknown, though the FNM has been most often chosen in many previous work [15].

After the determination of all model parameters, every pixel of the image was labeled to a different region (from 1 to K) based on the CBRL algorithm. We then selected the brightest region, which corresponding to label K, plus a criterion of closed isolated area, as the candidate region of suspicious masses. According to the visual inspections by the radiologists, when we use $K - 1$ instead of K, the results are over-segmented. For the case of using $K + 1$, the results are under-segmented. In order to quantify the performance differences between the different segmentation methods, several groups have suggested that the segmentation results may be compared against radiologists' outlines of the lesions [3]. Though the proposed comparison measures are quantitative, the performance measures are still qualitative, since the reference base (e.g., gold standard by the radiologists) is qualitative, subjective, and imperfect. Therefore, in this model-supported approach, in addition to the visual inspections by the radiologists, we have also introduced an objective measure, the GRE between the histogram of the pixel images $p_{\mathbf{x}}(u)$ and the FGGM of the *segmented* image $p_{\mathbf{x},\mathbf{l}}(u)$ to assess the performance of the segmentation, defined by

$$\mathrm{GRE}(p_{\mathbf{x}}(u)\|p_{\mathbf{x},\mathbf{l}}(u)) = \sum_{u} p_{\mathbf{x}}(u) \log \frac{p_{\mathbf{x}}(u)}{p_{\mathbf{x},\mathbf{l}}(u))} \qquad (27)$$

where $\mathbf{l}$ is the context image estimated by the segmentation algorithm. Considering that the ergodic theorem is the most fundamental principle in the detection and estimation theory, it is believed that when a good segmentation is achieved, the distance between the $p_{\mathbf{x}}(u)$ and $p_{\mathbf{x},\mathbf{l}}(u)$ should be minimized and

TABLE V
COMPARISON OF SEGMANTATION ERROR RESULTING FROM NONCONTEXTUAL AND CONTEXTUAL METHODS

Method	Soft Classification	Bayesian Classification	CBRL
GRE Value	0.0067	0.4406	0.1578

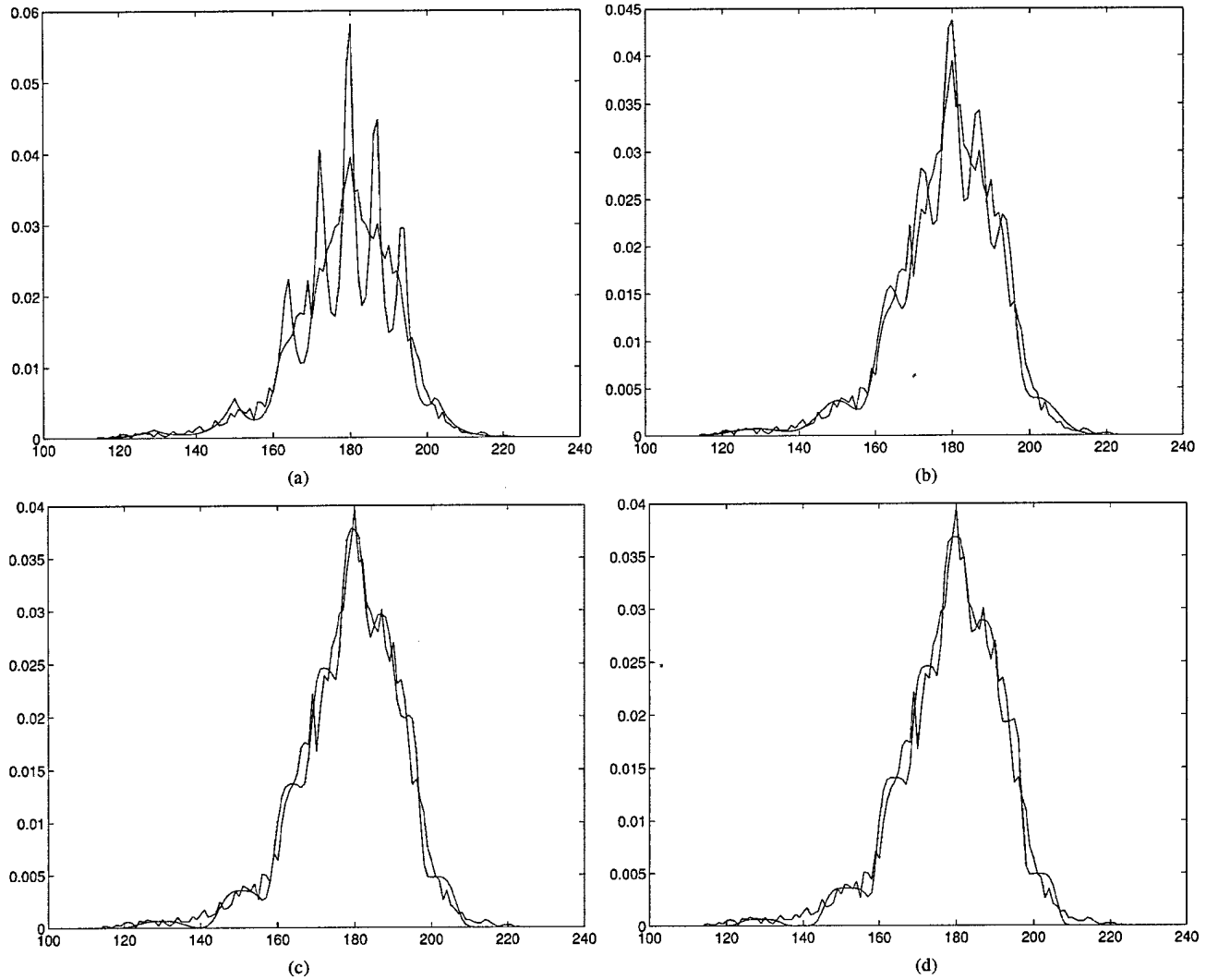


Fig. 7. Comparison of learning curves and histogram of the original mammogram with different α, $k = 8$. The optimal $\alpha = 3.0$. (a) $\alpha = 1.0$, GRE = 0.0783. (b) (a) $\alpha = 2.0$, GRE = 0.0369. (c) $\alpha = 3.0$, GRE = 0.0251. (d) $\alpha = 4.0$, GRE = 0.0282.

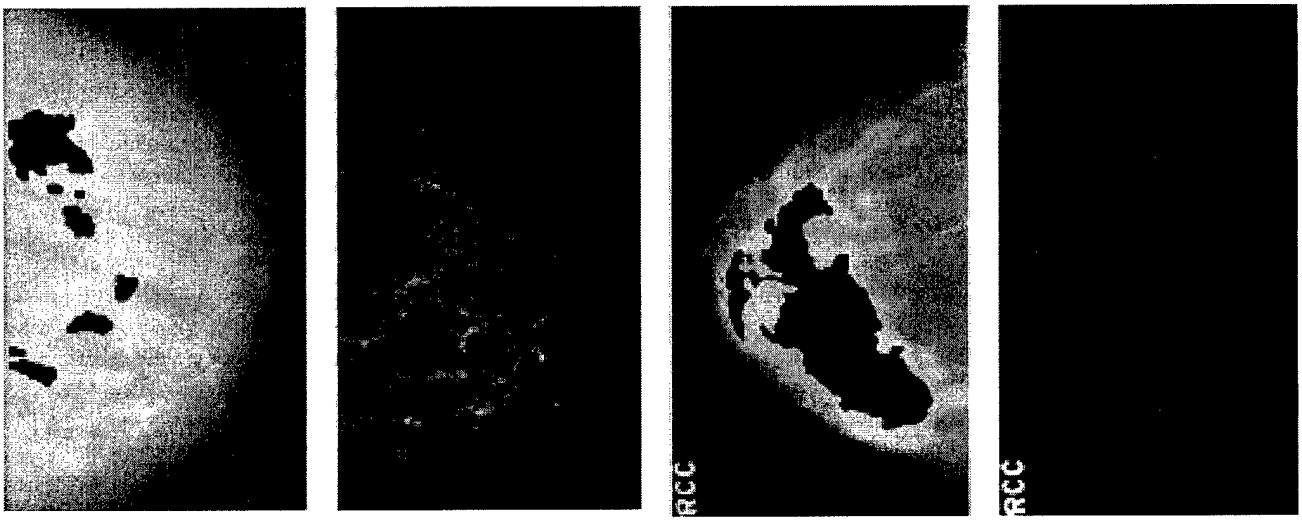
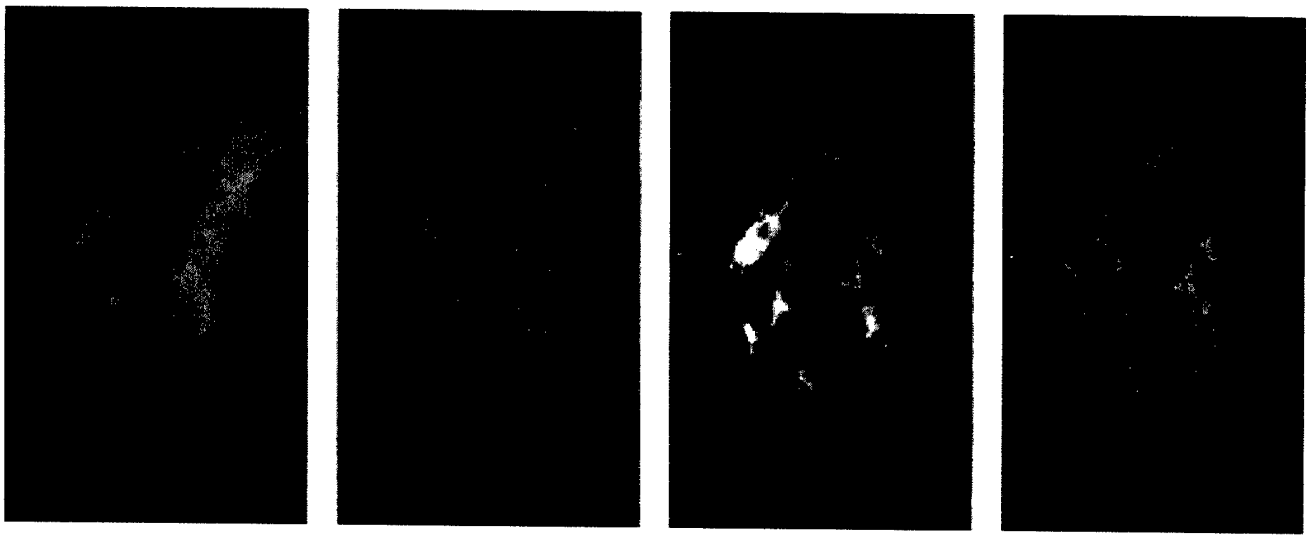
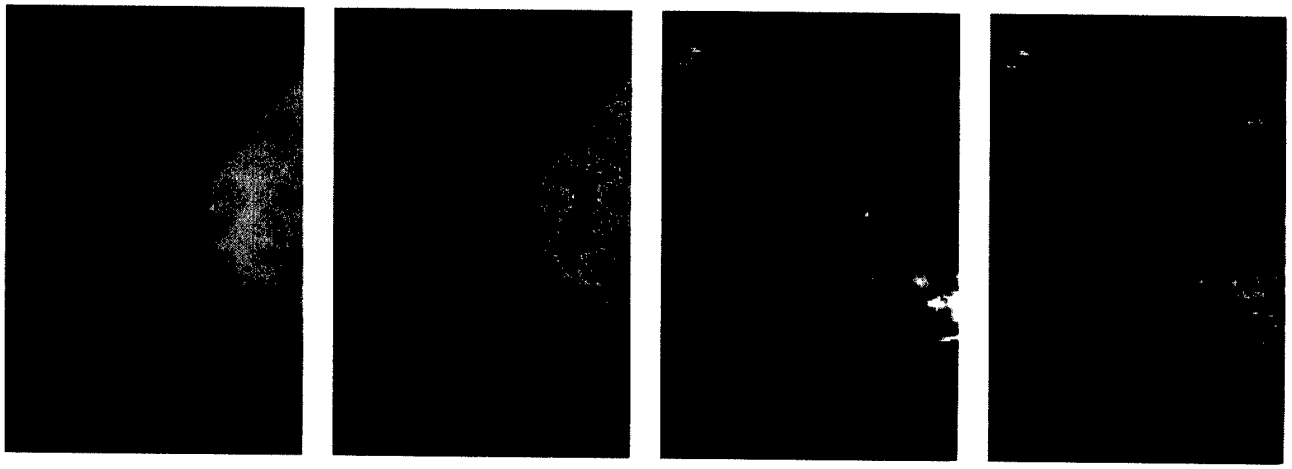


Fig. 8. Suspected mass segmentation results based on the original mammogram. (b) Result based on the enhanced mammogram, $K = 4$, $\alpha = 3.0$. (c) and (d) Results based on another original mammogram and its enhanced image.



(a) (b) (c) (d)

Fig. 9. Examples of normal mixed fatty and glandular mammogram. (a) Original mammogram. (b) Segmentation result based on the original mammogram. (c) Enhanced mammogram. (d) Result based on the enhanced mammogram, $k = 4$, $\alpha = 3.0$.



(a) (b) (c) (d)

Fig. 10. Examples of normal dense mammogram. (a) Original mammogram. (b) Segmentation result based on the original mammogram. (c) Enhanced mammogram. (d) Result based on the enhanced mammogram, $k = 4$, $\alpha = 3.0$.

this measure links the image text and its sample averages. Our experience has suggested that this post-segmentation measure may be a suitable objective criterion for evaluating the quality of image segmentation in a fully unsupervised situation [22], [26]–[28]. Table V shows our evaluation data from three different segmentation methods when applied to the real images.

Performance of Combined Morphological Filtering and Model-Based Segmentation using a Larger Database: The proposed segmentation method was used to extract suspicious mass regions from the 200 testing mammograms. Without enhancement, a total of 1142 potential mass regions were isolated including 114 of the 186 true masses. With enhancement, a total of 3143 potential mass regions were extracted including 181 of the 186 true masses. The results demonstrated that more true masses were picked up after enhancement although more false cases were also included. The undetected areas mainly occurred at the lower intensity side of the shaded objects or obscured by fibroglandular tissues that, however, were extracted on morphological enhanced mammograms. In addition, when the margins of masses are ill defined, only parts of suspicious masses were extracted from the original mammograms. For the purpose of "lesion site selection," we believe that the sensitivity should be the sole criterion for the performance evaluation of the method. We have 181/186 versus 114/186. Our method is unsupervised and automatic and does not involve any detection effort at this moment. To our best knowledge, there is no objective criterion available for the evaluation of image enhancement performance before a detection effort is involved. We only claimed that the enhancement step is important and effective with respect to the purpose of "lesion site selection."

Fig. 8 demonstrates some segmentation results based on the original and enhanced mammograms. We compared the segmentation results based on the enhanced mammogram ($K = 4$, and $\alpha = 3.0$) with those based on the original mammogram ($K = 8$, and $\alpha = 3.0$) as shown in Fig. 8. Comparing the results in Fig. 8(b) with those in Fig. 8(a), we can see that after

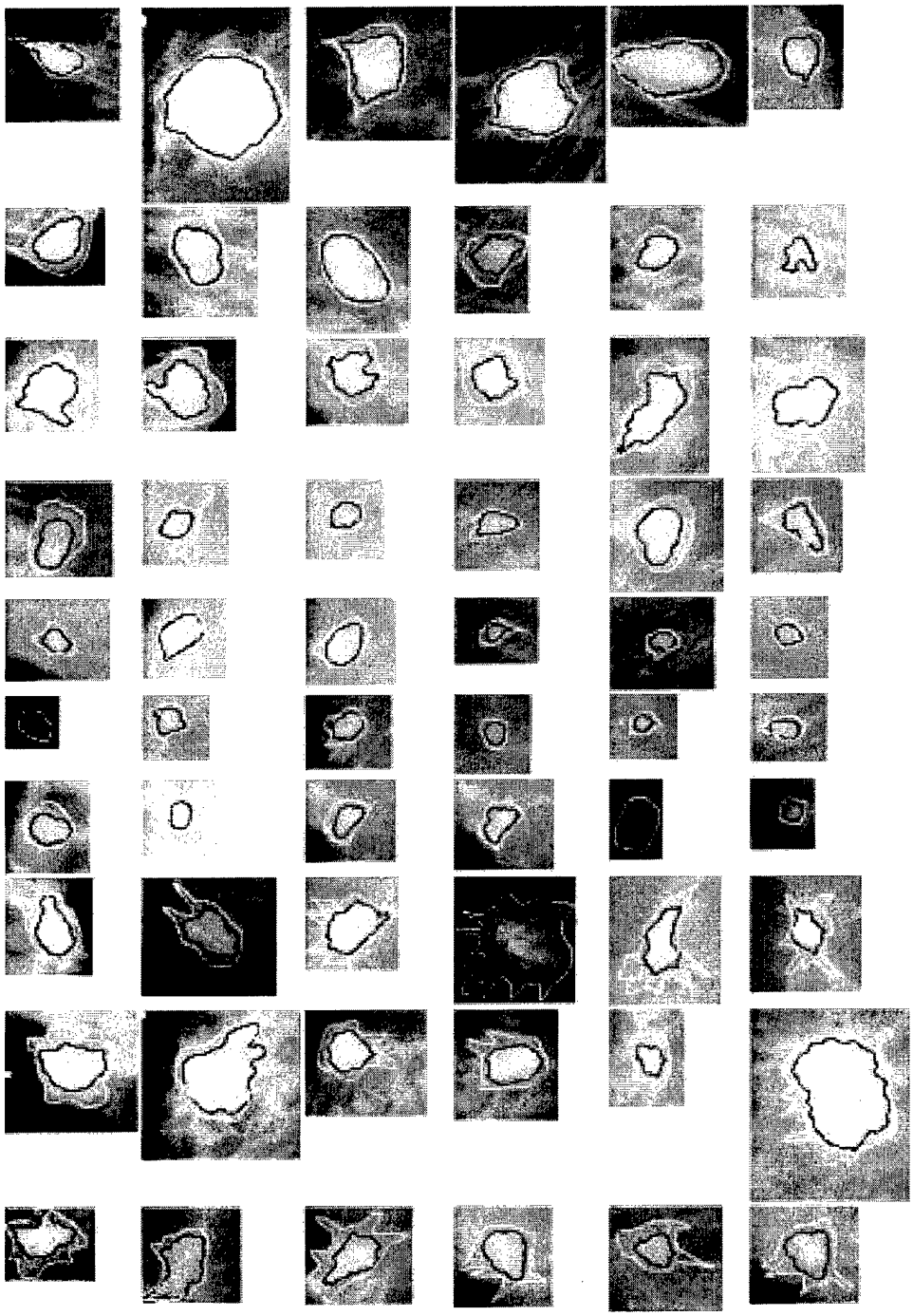


Fig. 11. Comparison results of segmantation based on the enhanced mammograms. Black outlines denote the computer-segmented result. White outlines denote ther radiologist-segmanted results.

enhancement, a more accurate region was detected for the suspected mass which has ill-defined margin. Getting an accurate suspected region is a crucial issue since geometric features are extracted based on suspected regions and these features are very important for further true mass detection. In addition, we observed that one suspected mass was missed in Fig. 8(a) but was detected in Fig. 8(b). As we have mentioned in Section I, none of the suspected masses should be missed in the segmentation step. Fig. 8(c) and (d) demonstrate the segmentation of a suspected mass that lies in dense breast tissue. As shown in Fig. 8(c), the whole fibroglandular tissue area was segmented when based on the original mammogram. After enhancement, the suspected region was segmented exactly as shown in Fig. 8(d).

We have also included the segmentation results on the normal mammograms. Fig. 9 demonstrate the segmentation results based on the original and enhanced mixed fatty and glandular mammograms. Fig. 10 demonstrate the segmentation results based on the original and enhanced dense mammograms. We

would like to emphasize that the objective of this paper is to provide a segmentation technique which can enhance and extract potential mass site from the background so that the characterization of the related mass pattern can be accurately extracted in terms of focused feature selection and analysis. The method of course will produce many mass-like areas, but it will be a plausible outcome since the accurate description of nonmass cases characterized by mass-like sites will benefit the follow-on detection step where the performance of the classifier depends on an accurate separation of mass and nonmass in the featured spaces. The details will be described in [29].

For the purpose of evaluating the performance of the segmentation method, we used both simulated studies and expert visual inspection to validate the methods and results. The radiologist has concluded that the lesion characteristics after the proposed enhancement have been better displayed and all possible lesion areas have been successfully identified. In addition to the visual inspection, we have measured the overlap between the computer-segmented and the radiologist segmented mass regions to evaluate our method. Fig. 11 shows the comparison results of segmentation based on the enhanced mammograms. Fig. 11 includes 60 benign and malignant mass patches which were cut from the whole mammograms after the segmentation. The white outline was drawn by the radiologist while the black outline was produced by the computer and was superimposed upon the original image. As we can see from Fig. 11, for most of cases, the ratio of mutual overlap area of the radiologist segmented mass region and the computer-segmented mass region to the radiologist segmented mass area is large than 50%. In addition, even the poorest result picked the true lesion in the correct location and depicted the characteristics of the mass reasonably. It is important to understand that "lesion area segmentation" is not our objective, so there is no "best" or "worst" segmentation results. Our objective is "lesion site selection" with a possible highest sensitivity through a global unsupervised enhancement and segmentation scheme.

V. Conclusion

In this paper, we propose a combined method of using morphological operations, a FGGM modeling, and a CBRL to enhance and segment various breast tissue textures and suspicious mass lesions from mammographic images. This phase is a crucial step in mass detection for an improved CAD. We emphasized the importance of model selection which includes the selection of the number of image regions K and the selection of FGGM kernel shape controlled by α. The experimental results indicate that the suspected masse sites selection can be affected by different K and α. We proposed the EM algorithm together with the information theoretic criteria to determine the optimal K and α. With optimal K and α, the segmentation results can be significantly improved. We also showed that with the proposed pattern-dependent enhancement algorithm using morphological operations, the subtle masses can be segmented more accurately than those when the original image is used for extraction without enhancement. To summarize, the morphological filtering enhancement combined with the stochastic model-based segmentation is an effective way to extract mammographic suspicious patterns of interest, and thereby may facilitate the overall performance of mammographic CAD of breast cancer.

Acknowledgment

The authors would like to thank Z. Gu of the Lombardi Cancer Center and I. Sesterhenn of the Armed Forces Institute of Pathology for their scientific input on the knowledge of cell oncology and pathology, and R. Shah MD, Director of Breast Imaging, BAMC for his evaluation of cases to our database.

References

[1] H. Kobatake, M. Murakami, H. Takeo, and S. Nawano, "Computerized detection of malignant tumors on digital mammograms," *IEEE Trans. Med. Imag.*, vol. 18, pp. 369–378, May 1999.

[2] R. Zwiggelaar, T. C. Parr, J. E. Schumm, I. W. Hutt, C. J. Taylor, S. M. Astley, and C. R. M. Boggis, "Model-based detection of spiculated lesions in mammograms," *Med. Image Anal.*, vol. 3, no. 1, pp. 39–62, 1999.

[3] M. A. Kupinski and M. L. Giger, "Automated seeded lesion segmentation on digital mammograms," *IEEE Trans. Med. Imag.*, vol. 17, pp. 510–517, Aug. 1998.

[4] N. Karssemeijer and G. M. te Brake, "Detection of stellate distortions in mammogram," *IEEE Trans. Med. Imag.*, vol. 15, pp. 611–619, Oct 1996.

[5] W. K. Zouras, M. L. Giger, P. Lu, D. E. Wolverton, C. J. Vyborny, and K. Doi, "Investigation of a temporal subtraction scheme for computerized detection of breast masses in mammograms," *Excerpta Medica*, vol. 1119, pp. 411–415, 1996.

[6] N. Petrick, H. P. Chan, B. Sahiner, and D. Wei, "An adaptive density-weighted contrast enhancement filter for mammographic breast mass detection," *IEEE Trans. Med. Imag.*, vol. 15, no. 1, pp. 59–67, 1996.

[7] M. Sameti and R. K. Ward, "A fussy segmentation algorithm for mammogram patttion," in *Digital Mammgraphy*. ser. International Congress Series, K. Doi, Ed. Amsterdam, The Netherlands: Elsevier, 1996, pp. 471–474.

[8] W. P. Kegelmeyer Jr., J. M. Pruneda, P. D. Bourland, A. Hillis, M. W. Riggs, and M. L. Nipper, "Computer-aided mammographic screening for spiculated lesions," *Radiology*, vol. 191, pp. 331–337, 1994.

[9] F. F. Yin, M. L. Giger, C. J. Vyborny, K. Doi, and R. A. Schmidt, "Comparison of bilateral-subtraction and single-image processing techniques in the computerized detection of mammographic masses," *Investigat. Radiol.*, vol. 28, no. 6, pp. 473–481, 1993.

[10] B. Zheng, Y. H. Chang, and D. Gur, "Computerized detection of masses in digitized mammograms using single-image segmentation and a multilayer topographic feature analysis," *Acad. Radiol.*, vol. 2, pp. 959–966, 1995.

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

[12] M. L. Giger, C. J. Vyborny, and R. A. Schmidt, "Computerized characterization of mammographic masses: Analysis of spiculation," *Cancer Lett.*, vol. 77, pp. 201–211, 1994.

[13] T. K. Lau and W. F. Bischof, "Automated detection of breast tumors using the asymmetry approach," *Comput. Biomed. Res.*, vol. 24, no. 9, pp. 1501–1513, 1995.

[14] M. J. Bianchi, A. Rios, and M. Kabuka, "An algorithm for detection of masses, skin contours, and enhancement of microcalcifications in mammograms," in *Proc. , Symp. Computer Assisted Radiology*, Winston-Salem, NC, June 1994, pp. 57–64.

[15] T. Lei and W. Sewchand, "Statistical approach to x-ray CT imaging and its application in image analysis–Part II: A new stochastic model-based image segmentation technique for x-ray CT image," *IEEE Trans. Med. Imag.*, vol. 11, pp. 62–69, Feb. 1992.

[16] Y. Wang, T. Adali, and S.-C. B. Lo, "Automatic threshold selection using histogram quantization," *SPIE J. Biomedical Optics*, vol. 2, no. 2, pp. 211–217, April 1997.

[17] J. Zhang and J. W. Modestino, "A model-fitting approach to cluster validation with application to stochastic model-based image segmentation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 12, pp. 1009–1017, Oct. 1990.

[18] H. Li, K. J. R. Liu, Y. Wang, and S. C. Lo, "Morphological filtering and stochastic modeling-based segmentation of masses on mammographic images," in *Proc. IEEE Nuclear Science Symp. Medical Imaging Conf.*, 1996, pp. 1792–1796.

[19] J. Serra, *Image Analysis and Mathematical Morphology*. London, U. K.: Academic, 1982.

[20] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc. Ser. B*, vol. 39, pp. 1–38, 1977.

[21] Y. Wang, T. Adali, C. M. Lau, and S. Y. Kung, "Quantitative analysis of MR brain image sequences by adaptive self-organizing finite mixtures," *J. VLSI Signal Processing*, vol. 18, no. 3, pp. 219–240, 1998.

[22] Y. Wang, T. Adali, S. Y. Kung, and Z. Szabo, "Quantification and segmentation of brain tissues from MR images: A probabilistic neural network approach," *IEEE Trans. Image Processing*, vol. 7, pp. 1165–1181, Aug. 1998.

[23] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Automat. Contr.*, vol. 19, no. 6, pp. 716–723, 1974.

[24] J. Rissanen, "Modeling by shortest data description," *Automat.*, vol. 14, pp. 465–471, 1978.

[25] R. A. Hummel and S. W. Zucker, "On the foundations of relaxation labeling processes," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 5, pp. 267–286, Mar. 1983.

[26] A. Hoover, G. J. Baptoste, X. Jiang, P. J. Flynn, H. Bunke, D. B. Goldgof, K. Bowyer, D. W. Eggert, A. Fitzgibbon, and R. B. Fisher, "An experimental comparison of range image segmentation algorithms," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 18, pp. 673–688, July 1996.

[27] Y. J. Zhang, "A survey on evaluation methods for image segmentation," *Pattern Recogn.*, vol. 29, no. 8, pp. 1335–1346, 1996.

[28] A. M. Bensaid, L. O. Hall, J. C. Bezdek, L. P. Clarke, M. L. Silbiger, J. A. Arrington, and R. F. Murtagh, "Validity-guided clustering with applications to image segmentation," *IEEE Trans. Fuzzy Syst.*, vol. 4, pp. 112–123, May 1996.

[29] H. Li, Y. Wang, K. J. R. Liu, S.-C. B. Lo, and M. T. Freedman, "Computerized Radiographic Mass Detection—Part II: Decision Support by Featured Database Visualization and Modular Neural Networks," *IEEE Trans. Med. Imag.*, vol. 20, no. 4, pp. 302–313, Apr. 2001.

Computerized Radiographic Mass Detection—Part II: Decision Support by Featured Database Visualization and Modular Neural Networks

Huai Li, Yue Wang, K. J. Ray Liu*, Shih-Chung B. Lo, and Matthew T. Freedman

Abstract—**Based on the enhanced segmentation of suspicious mass areas, further development of computer-assisted mass detection may be decomposed into three distinctive machine learning tasks: 1) construction of the featured knowledge database; 2) mapping of the classified and/or unclassified data points in the database; and 3) development of an intelligent user interface. A decision support system may then be constructed as a complementary machine observer that should enhance the radiologists performance in mass detection. We adopt a mathematical feature extraction procedure to construct the featured knowledge database from all the suspicious mass sites localized by the enhanced segmentation. The optimal mapping of the data points is then obtained by learning the generalized normal mixtures and decision boundaries, where a is developed to carry out both soft and hard clustering. A visual explanation of the decision making is further invented as a decision support, based on an interactive visualization hierarchy through the probabilistic principal component projections of the knowledge database and the localized optimal displays of the retrieved raw data. A prototype system is developed and pilot tested to demonstrate the applicability of this framework to mammographic mass detection.**

Index Terms—**Feature extraction, knowledge database, mass detection, neural network, visual explanation.**

I. Introduction

IN ORDER to improve mass lesion detection and classification in clinical screening and/or diagnosis of breast cancers, many sophisticated computer-assisted diagnosis (CAD) systems have been recently developed [1]–[10]. Although the clinical roles of the CAD systems may still be debatable, the fundamental role should be complementary to the radiologists' clinical duties, where the pathways of achieving ultimate performance enhancement taken by the machine observer and human observer may not necessarily be close. For example, CAD systems may attack the tasks that the radiologists cannot perform well or find difficult to perform. Because of generally larger size and complex appearance of masses, especially the existence of spicules in malignant lesions, as compared with microcalcifications, feature-based approaches are largely adopted in many CAD systems [1]–[4], [6], [7]. Kegelmeyer has first reported promising results for detecting spiculated tumors based on local edge characteristics and Laws texture features [7]. Zwiggelaar *et al.*developed a statistical model to describe and detect the abnormal pattern of linear structures of spiculated lesions [1]. Karssemeijer *et al.* [2] proposed to identify stellate distortions by using the orientation map of line-like structures. Petrick *et al.* presented to reduce the false positive detection by combining the breast tissue composition information [4]. Zhang *et al.* used the Hough spectrum to detect spiculated lesions [6].

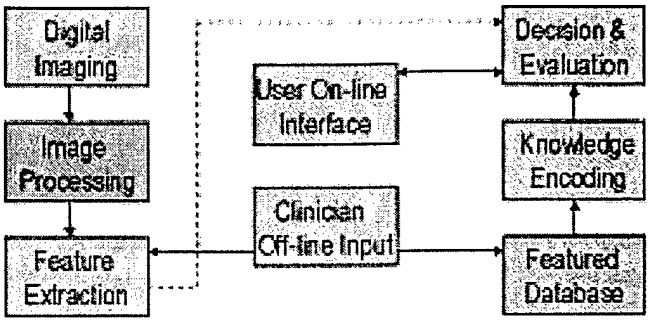


Fig. 1. Major components in CAD.

Although many previously proposed approaches have led to impressive results [1]–[5], [7], several fundamental issues remain unresolved in the application of CAD systems. Fig. 1 shows a general block diagram of CAD systems. Previous research has demonstrated that: 1) breast cancer is missed on mammograms in part because the optical density and contrast of the cancer is not optimal for human observer; 2) computer-based detection appears to be more affected by different criteria than human perception; 3) the challenges and pathways to the human or machine observers may be quite different, and 4) decision making by the CAD systems are largely not transparent to the user. For example, the training cases contributing to the database are often selected by the human observer while the featured knowledge database is constructed through mathematical pathways of feature extraction. The mismatch

Manuscript received February 3, 1997; revised January 9, 2001. This work was supported in part by the Department of Defense under Grants DAMD17-98-1-8045 and DAMD17-96-1-6254 through a subcontract from University of Michigan, Ann Arbor, and by the National Science Foundation (NSF) under NYI Award MIP-9457397. The Associate Editor responsible for coordinating the review of this paper and recommending its publication was M. Giger. *Asterisk indicates corresponding author.*

H. Li is with the Electrical Engineering Department and Institute for Systems Research, University of Maryland at College Park, College Park, MD 20742 USA. He is also with the Department of Radiology, Georgetown University Medical Center, Washington, DC 20007 USA.

Y. Wang is with the Department of Electrical Engineering and Computer Science, The Catholic University of America, Washington, DC 20064 USA. He is also with the Department of Radiology, Georgetown University Medical Center, Washington, DC 20007 USA.

*K. J. Ray Liu is with the Electrical Engineering Department and Institute for Systems Research University of Maryland at College Park, College Park, MD 20742 USA (e-mail: kjrliu@eng.umd.edu).

S.-C. B. Lo and M. T. Freedman are with the Department of Radiology, Georgetown University Medical Center, Washington, DC 20007 USA.

Publisher Item Identifier S 0278-0062(01)02830-0.

0278–0062/01$10.00 © 2001 IEEE

between the human supervised case selection in training and the machine dominant mass candidates selection in testing may exist. Second, the featured knowledge database is often high-dimensional with complex internal structures. Imposing a heuristically designed neural network for learning from the training data set may prevent a correct identification of the intrinsic data structure and an accurate estimation of the class boundaries. There may also exist the mismatch between the data structure and classifier architecture or between the class boundaries and decision boundaries. Furthermore, since the machine observer and human observer may not detect the same set of masses, the "black box" nature of most CAD systems to the clinical users will prevent a natural on-line integration of human intelligence and further upgrade of a CAD system. An interactive user interface should be considered to leverage the complementary roles of the CAD in the clinical practice.

As a step toward improving the performance of a CAD system, we have put considerable efforts to conduct various studies and develop reliable image enhancement and lesion selection techniques. The methods and results have been reported in [24], where the purposes of the research were to localize the potential mass sites and help accurate feature extraction. This paper addresses the further development of computer-assisted mass detection based on the 1) construction of the featured knowledge database; 2) mapping of the classified and/or unclassified data points in the database; and 3) development of an intelligent user interface (IUI). The clinical goal is to eliminate the false positive sites that correspond to normal dense tissues with *mass-like* appearances through featured discrimination. We adopt a mathematical feature extraction procedure to construct the featured knowledge database from all the suspicious mass sites localized by the enhanced segmentation. The optimal mapping of the data points is then obtained by learning the generalized normal mixtures and decision boundaries, where a probabilistic modular neural network (PMNN) is developed to carry out both soft and hard clustering. A visual explanation of the decision making is further invented as a decision support tool, based on an interactive visualization hierarchy through the probabilistic principal component projections of the knowledge database and the localized optimal displays of the retrieved raw data. The motivation of this work comes from the following considerations. First, though both human and machine observers use the same set of raw data in the diagnostic stage, the construction of the knowledge database for training machine classifiers and that accomplished by human brains are indeed different. Thus, the knowledge database should be established with both machine and expert organized representative cases. Second, a quantitative understanding of the knowledge database used by the machine observer should be acquired to logically compare and/or predict the performance of CAD systems with respect to the human observers without possible under- or over-estimation, and to optimize the feature extraction and design of the machine learner for best final performance. Finally, since the human and machine observers indeed take different learning and intelligence pathways, an IUI should be developed to visually (e.g., transparently) explain the entire internal decision making process of the CAD system to the human observer to enhance the clinical decision when facing either consistent or conflicting opinions.

The major differences between our work and the previous work [1]–[10] are as follows.

1) We construct a knowledge database by combining both expert and machine selected cases where the assignment of class memberships (e.g., mass and nonmass classes) is supervised by the radiologists or pathological report *after* all the cases are collected.
2) We impose a model identification procedure to determine the optimal number and kernel shape of the local clusters within each of the two classes in a high-dimensional feature space. The model is then estimated using the expectation–maximization (EM) algorithm and information theory.
3) We develop a PMNN, which is considered as a nonlinear classifier, to carry out the mapping function of the knowledge database. In the knowledge database, the decision likelihood boundaries and the class prior probabilities are determined in a separate fashion, and the structure of PMNN is optimized by adapting to the database structure.
4) We derive a probabilistic principal component projection scheme to reduce the dimensionality of the feature space for natural human perception. The scheme leads to a hierarchical visualization algorithm allowing the complete data set to be analyzed at the top level, with best separated clusters and subclusters of data points analyzed at deeper levels.

The framework of the proposed method for mass detection is illustrated in Fig. 2. A detailed description of this paper is organized as follows. In Section II, the procedure of the knowledge database construction is described. The data mapping process for decision making is presented in Section III. Section IV presents the design of the IUI for the CAD systems. Finally, major results and discussions are summarized in Section V.

II. Knowledge Database Construction

Given the available information contained in the raw data of mass sites and in order to establish machine intelligence carried out by various machine observers, a knowledge database may be constructed in a multidimensional feature space. It should be emphasized however that the knowledge acquired by the human brain uses much more sophisticated processes than the artificial systems. Though feature extraction has been a key step in most pattern analysis tasks, the mathematical procedures are often done intuitively and heuristically. The general guidelines are:

1) *Discrimination:* Features of patterns in different classes should have significantly different values.
2) *Reliability:* Features should have similar values for the patterns of the same class.
3) *Independence:* Features should not be strongly correlated to each other.
4) *Optimality:* Some redundant features should be deleted. A small number of features is preferred for reducing the complexity of the classifier.

Many useful image features have been suggested previously by both image processing and pattern analysis communities [11]–[13]. These features can be divided into three categories, namely, intensity features, geometric features, and texture

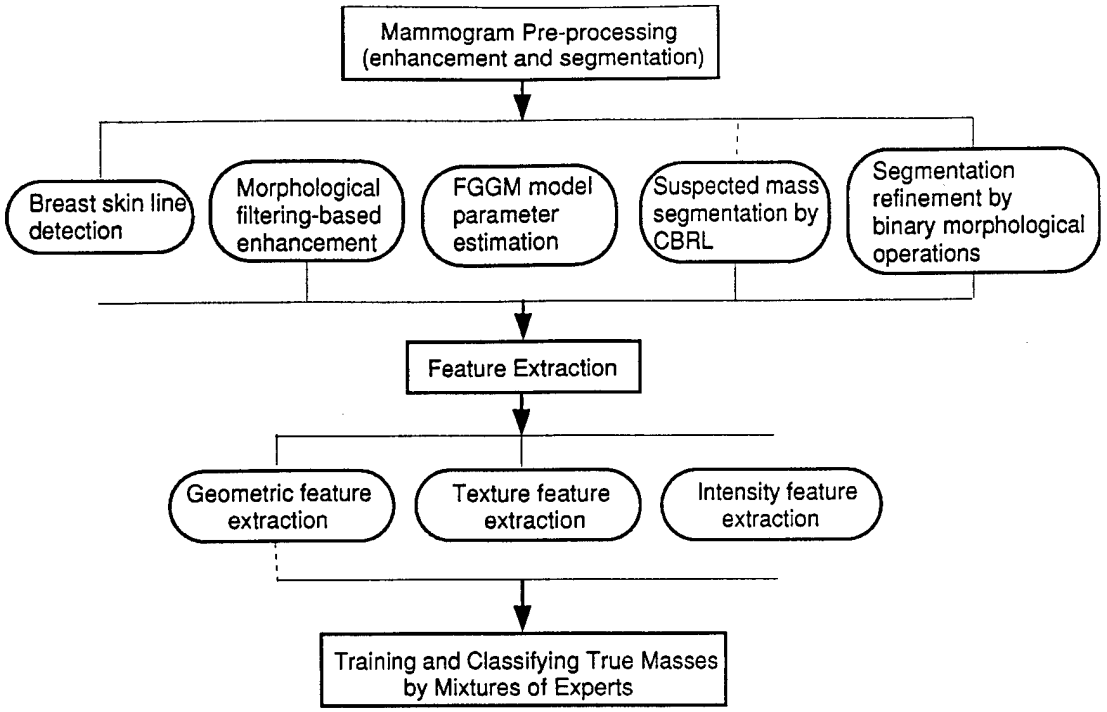


Fig. 2. The flow diagram of mass detection in digital mammograms.

features, whose values are calculated from the pixel matrices of the regions of interest (ROIs). Though these features are mathematically well defined, they may not be complete since they cannot capture all of the capable aspects of human perception nature. Thus, in this study, we have included several additional expert-suggested features to reflect the radiologists' experience. The typical features are summarized in Table I, where Fig. 3 shows the raw image of corresponding featured sites.

The joint histogram of the feature point distribution extracted from true and false mass regions are investigated, and the features that can better separate the true and false mass regions are selected for further study. Our experience has suggested that three features, i.e., the site area, two measured compactness (circularity), and difference entropy, were having better discrimination and reliability properties. Their definitions are given as follows.

1) *Compactness 1*

$$C_1 = \frac{A_1}{A} \tag{1}$$

where A is the area of the actual suspected region, and A_1 is the area of the overlapped region of A and the effective circle A_c, which is defined as the circle whose area is equal to A and is centered about the corresponding centroid of A.

2) *Compactness 2*

$$C_2 = \frac{P}{4\pi A} \tag{2}$$

where P is the boundary perimeter, and A is the area of region.

TABLE I
THE SUMMARY OF MATHEMATICAL FEATURES

Feature Sub-Space	Features
A. Intensity Features	1. contrast measure of ROIs; 2. standard derivation inside ROIs; 3. mean gradient of ROIs boundary
B. Geometric Features	1. area measure; 2. circularity measure; 3. deviation of the normalized radial length; 4. boundary roughness;
C. Texture Features	1. energy measure; 2. correlation of co-occurrence matrix; 3. inertia of co-occurrence matrix; 4. entropy of co-occurrence matrix; 5. inverse difference moment; 6. sum average; 7. sum entropy; 8. difference entropy; 9. fractal dimension of surface of ROI;

3) *Difference Entropy*

$$DH_{d,\theta} = -\sum_{k=0}^{L-1} p_{x-y}(k) \log p_{x-y}(k) \tag{3}$$

where

$$p_{x-y}(k) = \sum_{i=0}^{L-1} \sum_{j=0}^{L-1} p_{d,\theta}(i,j), \quad |i-j| = k. \tag{4}$$

Several important observations are worth reiteration:

1) The knowledge database that will be used by the CAD system are constructed from the cases selected by both lesion localization procedure and human expert's experience. This joint set provides more complete knowledge to

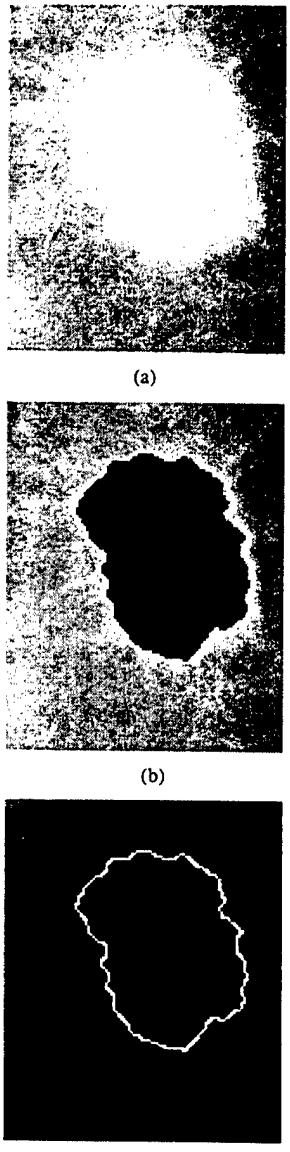


Fig. 3. One example of mass segmentation and boundary extraction. (a) Mass patch; (b) segmentation; (c) boundary extraction.

the machine observer. In particular, during the interactive decision making, CAD system can still provide opinion when the cases are missed by the localization procedure but presented to the system by the radiologists.

2) The knowledge database is defined quantitatively in a high dimensional feature space. It provides not only the knowledge for training the machine observer, but also an objective base for evaluating the quality of feature extraction or network's learning capability, and the on-line visual explanation possibility.
3) The assignment of the cases' class memberships (e.g., mass and nonmass classes) is supervised by the radiologists or pathological reports. A complete knowledge database includes three subsets: raw data of mass-like sites, corresponding feature points, and class membership labels.

III. Data Mapping For Decision Making

The decision making support by a CAD system addresses the problem of mapping a knowledge database, given a finite set of data examples. The mapping function can therefore be interpreted as a quantitative representation of the knowledge about the mass lesions contained in the database [14]. Instead of mapping the whole data set using a single complex network, it is more practical to design a set of simple class subnets with local mixture clusters, each one of which represents a specific region of the knowledge space. Inspired by the principle of *divide-and-conquer* in applied statistics, PMNN has become increasingly popular in machine learning research [14], [15], [19]–[22]. In this section, we present its applications to the problem of mapping from databases in mass detection, with a constructive criterion for designing the network architecture and the learning algorithm that are governed by information theory [25].

A. Statistical Modeling

The quantitative mapping of a database may be decomposed into three distinctive learning tasks: the detection of the structure of each class model with local mixture clusters; the estimation of the data distributions for each induced cluster inside each class; and the classification of the data into classes that realizes the data memberships. Recently, there has been considerable success in using finite mixture distributions data mapping [15], [17], [18], [20]. Assume that the data points $\vec{x}_i$ in a multidimensional database come from M classes $\{\vec{\omega}_1, \ldots, \vec{\omega}_r, \ldots, \vec{\omega}_M\}$, and each class contains K_r clusters $\{\vec{\theta}_1, \ldots, \vec{\theta}_k, \ldots, \vec{\theta}_{K_r}\}$, where $\vec{\omega}_r$ is the model parameter vector of class r, and $\vec{\theta}_k$ is the kernel parameter vector of cluster k within class r. The class conditional probability measure for any data point inside the class r, i.e., the standard finite mixture distribution (SFMD), can be obtained as a sum of the following general form:

$$f(\vec{u}|\vec{\omega}_r) = \sum_{k=1}^{K_r} \pi_k g(\vec{u}|\vec{\theta}_k) \tag{5}$$

where $\pi_k = P(\vec{\theta}_k|\vec{\omega}_r)$ with a summation equal to one, and $g(\vec{u}|\vec{\theta}_k)$ is the kernel function of the local cluster distribution. For the model of global class distributions, we denote the Bayesian prior for each class by $P(\vec{\omega}_r)$. Then the sufficient statistics according to the Bayes' rule, are the posterior probability $P(\vec{\omega}_r|\vec{x}_i)$ given a particular observation $\vec{x}_i$

$$P(\vec{\omega}_r|\vec{x}_i) = \frac{P(\vec{\omega}_r)f(\vec{x}_i|\vec{\omega}_r)}{p(\vec{x}_i)} \tag{6}$$

where $p(\vec{x}_i) = \sum_{r=1}^{M} P(\vec{\omega}_r)f(\vec{x}_i|\vec{\omega}_r)$.

B. Class Distribution Learning

Class distribution learning addresses the combined estimation of regional parameters $(\pi_k, \vec{\theta}_k)$ and detection of the structural parameter K_r and the kernel shape of $g(\cdot)$ in (5) based on the observations $\mathbf{x}_r$. One natural criterion used for learning the optimal parameter values is to minimize the distance between the SFMD, denoted by $f_r(\vec{u})$, and the class data histogram, denoted by $f_{\mathbf{x}_r}(\vec{u})$ [17]. In this paper, we use relative entropy (Kullback–Leibler distance), suggested by information theory

[25], as the distance measure (for simplicity we use $f_r(\vec{u})$ to denote $f(\vec{u}|\vec{\omega}_r)$ in our formulation), given by

$$D(f_{\mathbf{x}_r}\|f_r) = \sum_{\vec{u}} f_{\mathbf{x}_r}(\vec{u}) \log \frac{f_{\mathbf{x}_r}(\vec{u})}{f(\vec{u}|\vec{\omega}_r)}. \tag{7}$$

We have previously shown that when relative entropy is used as a distance measure, the distance minimization method is equivalent to the soft-split classification-based method under the criterion of maximum likelihood (ML) [23].

Another important issue concerning unsupervised distribution learning is the detection of the structural parameters of the class distribution, called model selection [15]. The objective here is to propose a systematic strategy for determining the optimal number and kernel shape of local clusters, when the prior knowledge is not available. This is indeed the case when the structure of the mass lesion patterns for a particular type of cancer may be arbitrarily complex, so correct identification of the database structure is very important. Thus, it will be desirable to have a neural network structure that is adaptive, in the sense that the number and kernel shape of local clusters are not fixed beforehand. In this paper, we applied two popular information theoretic criteria, i.e., the Akaike information criterion and minimum description length to guide the model selection procedure [24].

As the counterpart for adaptive model selection, there are many numerical techniques to perform ML estimation of cluster parameters [17]. For example, EM algorithm first calculates the posterior Bayesian probabilities of the data through the observations and the current parameter estimates (E-step) and then updates parameter estimates using generalized mean ergodic theorems (M-step). The procedure cycles back and forth between these two steps. The successive iterations increase the likelihood of the model parameters. The scheme provides winner-takes-in probability (Bayesian "soft") splits of the data, hence allowing the data to contribute simultaneously to multiple clusters. For the sake of simplicity, we assume the kernel shape of local clusters to be a multidimensional Gaussian with mean $\vec{\mu}_{kr}$ and variance Γ_{kr}. We summarize the EM algorithm as follows.

1) **E-Step:** for training sample $\vec{x}^{(t)}, t = 1, \ldots, N$, compute the probabilistic membership

$$h_{kr}^{(m)}(t) = \frac{\pi_{kr}^{(m)} p_k^{(m)}(\vec{x}^{(t)}|\vec{\omega}_r)}{\sum_{k=1}^{K_r} \pi_{kr}^{(m)} p_k^{(m)}(\vec{x}^{(t)}|\vec{\omega}_r)}. \tag{8}$$

2) **M-Step:** compute the updated parameter estimates

$$\pi_{kr}^{(m+1)} = \frac{1}{N} \sum_{t=1}^{N} h_{kr}^{(m)}(t) \tag{9}$$

$$\vec{\mu}_{kr}^{(m+1)} = \frac{1}{N\pi_{kr}^{(m+1)}} \sum_{t=1}^{N} h_{kr}^{(m)}(t)\vec{x}^{(t)} \tag{10}$$

$$\Gamma_{kr}^{(m+1)} = \frac{1}{N\pi_{kr}^{(m+1)}} \sum_{t=1}^{N} h_{kr}^{(m)}(t) \left[\vec{x}^{(t)} - \vec{\mu}_{kr}^{(m+1)}\right] \times \left[\vec{x}^{(t)} - \vec{\mu}_{kr}^{(m+1)}\right]^T. \tag{11}$$

C. Decision Boundary Learning

The objective of data classification is to realize the class membership l_{ir} for each data points based on the observation $\vec{x}_i$ and the class statistics $\{P(\vec{\omega}_r), f(\vec{u}|\vec{\omega}_r)\}$. It is well known that the optimal data classifier is the Bayes classifier since it can achieve the minimum rate of classification error [26]. Measuring the average classification error by the mean squared error E, many previous researchers have shown that minimizing E by adjusting the parameters of class statistics is equivalent to directly approximating the posterior class probabilities when dealing with the two class problem [13], [26]. In general, for the multiple class problem the optimal Bayes classifier (minimum average error) classifies input patterns based on their posterior probabilities: input $\vec{x}_i$ is classified to class $\vec{\omega}_r$ if

$$P(\vec{\omega}_r|\vec{x}_i) > P(\vec{\omega}_j|\vec{x}_i) \tag{12}$$

for all $j \neq r$. It should be noted that in the formulation of classifier design, the optimal criterion used for the future data classification has been intuitively and directly applied to the learning of class statistics from the training data set.

Direct learning of posterior probability is a complex task. Great effort has been made in designing the classifier as an estimator of the posterior class probability [19]. By closely investigating the global class distribution modeling, we found that the classifier design for data classification can be dramatically simplified at the learning stage. Revisit (6), since the class prior probability $P(\vec{\omega}_r)$ is a known parameter when a supervised learning is applied, the posterior class probability $P(\vec{\omega}_r|\vec{x}_i)$ can be obtained without any further effort. Thus, by conditioning $P(\vec{\omega}_r)$, the problem is formulated as a supervised classification learning of the class conditional likelihood density $f(\vec{u}|\vec{\omega}_r)$. Thus, an efficient supervised algorithm to learn the class conditional likelihood densities called the "decision-based learning" [21] is adopted in this paper. The decision-based learning algorithm uses the *misclassified* data to adjust the density functions $f(\vec{u}|\vec{\omega}_r)$, which are initially obtained using the unsupervised learning scheme described previously, so that the minimum classification error can be achieved. Define the rth class discriminant function $\phi_r(\vec{x}_i, \mathbf{w})$ to be $P(\vec{\omega}_r)f(\vec{x}_i|\vec{\omega}_r)$. Given a set of training patterns $\mathbf{X} = \{\vec{x}_i; i = 1, 2, \ldots, M\}$. The set $\mathbf{X}$ is further divided into the "positive training set" $\mathbf{X}^+ = \{\vec{x}_i; \vec{x}_i \in \vec{\omega}_r, i = 1, 2, \ldots, N\}$ and the "negative training set" $\mathbf{X}^- = \{\vec{x}_i; \vec{x}_i \notin \vec{\omega}_r, i = N+1, N+2, \ldots, M\}$. If the misclassified training pattern is from positive training set, reinforced learning will be applied. If the training pattern belongs to the negative training set, we anti-reinforce the learning, i.e., pull the kernels away from the problematic regions. The boundary refinement is summarized as follows:

Reinforced Learning: $\mathbf{w}^{(j+1)} = \mathbf{w}^{(j)} + \eta l'(d(t))\nabla\phi(\mathbf{x}(t), \mathbf{w})$

Antireinforced Learning: $\mathbf{w}^{(j+1)} = \mathbf{w}^{(j)} - \eta l'(d(t))\nabla\phi(\mathbf{x}(t), \mathbf{w})$ (13)

PMNN is a probabilistic modular network designed especially for data classification where a Bayesian decomposition of the learning process provides a unique opportunity to optimize

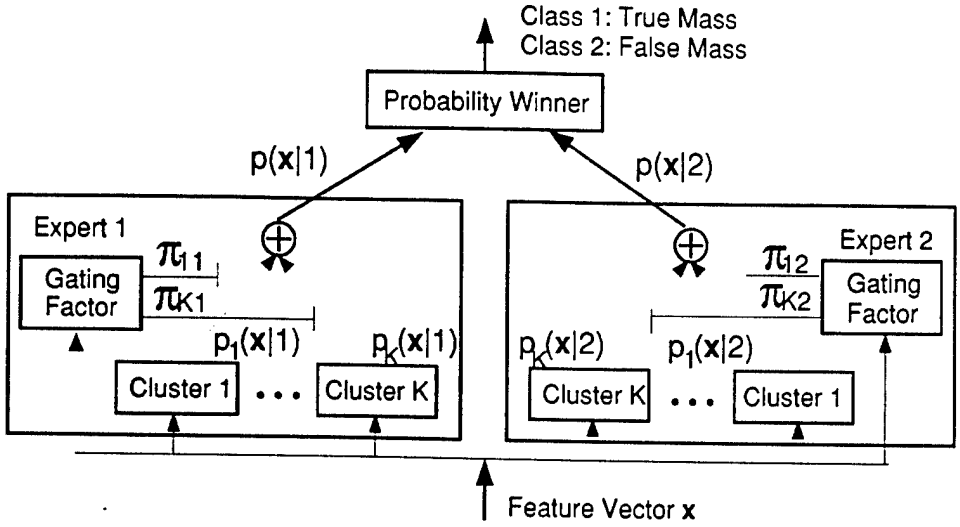


Fig. 4. The structure of the PMNN.

the structure of training scheme [14], [22]. Since the information about class population is, in general, physically uncorrelated with the conditional features about the individual class, a decoupled two-step training, in terms of both network structure and learning rule, makes much more sense than that in the conventional posterior-type neural networks, i.e., the conditional likelihood of each class and the class Bayesian prior should be adjusted separately in the classification spaces. Thus, PMNN consists of several disjoint subnets and a winner-takes-all network. The subnet outputs of the PMNN are designed to model the likelihood functions (likelihood-type network) which are first estimated from equally presented class samples, and the final decision boundaries are determined simply weighting the likelihood by the class populations. For a M-classification problem, PMNN contains M different class subnets, each of which represents one data class in the database. Within each subnet, several neurons (or clusters) are applied in order to handle problems which have complicated decision boundaries. The outputs of class subnets are fed into a winner-take-all network. The winner-take-all network categorizes the input pattern to the data class whose subnet produces the highest output value.

The structure of the PMNN used in this study is shown in Fig. 4. The PMNN consists of two subnets. Within each subnet, there are several neurons (or clusters). The outputs of class subnets are fed into a probability winner processor, which categorizes the input pattern to the data class whose subnet produces the highest probability value. The training scheme of the PMNN is based on the unsupervised learning. Each subnet is trained individually, and no mutual information across the classes may be utilized. In our study, one modular expert is trained to detect true masses, and the other is trained to detect false masses. After training, the feature vectors extracted from ROIsub are entered to this network to classify true or false masses. In both training and testing processes, we assume that the feature vectors $\vec{x}_i$ in class r ($r = 1, \ldots, M$) is a mixture of multidimensional Gaussian distributions, i.e.,

$$f(\vec{x}_i|\vec{\omega}_r) = \sum_{k=1}^{K_r} \pi_{kr} p_k(\vec{x}_i|\vec{\omega}_r) \tag{14}$$

where $\sum_{k=1}^{K_r} \pi_{kr} = 1$ and $p_k(\vec{\omega}_r) = N(\vec{\mu}_{kr}, \Gamma_{kr})$ is a multidimensional Gaussian distribution within cluster k of class r.

IV. Interactive Visual Explanation

In order to improve the utility of the CAD systems in clinical practice, an IUI is highly desired. Different from many previously proposed approaches, we have organized our database from both mathematical-localized and radiologist-selected mass-like cases, and formed the featured knowledge database based on both mathematical-based and radiologist-selected image features. This off-line effort should enhance the performance of the machine observer through better quality of training set and optimal design of neural network architecture. Our experience has suggested, however, that further improvement of CAD systems requires on-line natural integration of human intelligence with the computer' output, since human perception has and can play an important role in the clinical decision making. In this research, we have pilot developed an IUI where the major functions include: 1) interactive visual explanation of the CAD decision making process; 2) on-line retrieval of the optimally displayed raw data and/or similar cases; and 3) supervised upgrade of the knowledge database by radiologist-driven input of the "unseen" and/or "typical" cases. Our preliminary studies have shown that the visual presentation of both raw data and CAD results to radiologists may provide visual cues for improved decision making.

As a step toward understanding the complex information from data and relationships, structural and discriminative knowledge reveals insight that may prove useful in data mining. Hierarchical minimax entropy modeling and probabilistic principal component projection are proposed for data explanation, which is both statistically principled and visually effective at revealing all of the interesting aspects of the data set. The methods involve multiple use of standard finite normal mixture models and probabilistic principal component projections. The strategy is that the top-level model and projection should explain the entire data set, best revealing the presence of clusters and relationships, while lower-level models and

projections should display internal structure within individual clusters, such as the presence of subclusters and attribute trends, which might not be apparent in the higher-level models and projections. With many complementary mixture models and visualization projections, each level will be relatively simple while the complete hierarchy maintains overall flexibility yet still conveys considerable structural information. In particular, a probabilistic principal component neural network is developed to generate optimal projections, leading to a hierarchical visualization algorithm. This algorithm allows the complete data set to be analyzed at the top level, with best separated subclusters of data points analyzed at deeper levels.

Research evidence suggests that for analysis of complex and high-dimensional data sets, structure decomposition and dimensionality reduction are the natural strategies in which the model-based approach and visual explanation have proven to be powerful and widely-applicable [27]. However, there is a trade-off between maximizing (structure decomposition) and minimizing (dimensionality reduction) the entropy of the system. In this research, a minimax entropy approach is adopted through the use of progressive model identification and principal component projection. The complete visual explanation hierarchy is generated by performing principal projection (dimensionality reduction) and model identification (structure decomposition) in two iterative steps using information theoretic criteria, EM algorithm, and probabilistic principal component analysis (PCA). Hierarchical probabilistic principal component visualization involves: 1) evaluation of posterior probabilities for mixture data set; 2) estimation of multiple principal component axes from probabilistic data set; and 3) generation of a complete hierarchy of visual projections.

Suppose the data space is d-dimensional with coordinates $y_1, \ldots, y_d$ and the data set consists of a set of d-dimensional vectors $\{\mathbf{t}_i\}$ where $i = 1, \ldots, N$. Now consider a three-dimensional (3-D) latent space $\mathbf{x} = (x_1, x_2, x_3)^T$ together with a linear function which maps the latent space to the data space by $\mathbf{y} = \mathbf{W}\mathbf{x}+\mathbf{b}$ where $\mathbf{W}$ is a $d\times 3$ matrix and $\mathbf{b}$ is a d-dimensional mean vector. If we introduce a probability distribution $p(\mathbf{x})$ over the latent space given by a Gaussian estimated from the latent variables $\{\mathbf{x}_i\}$, then a similar full-dimensional Gaussian distribution in data space can be defined by convolving this distribution with a general diagonal Gaussian conditional probability distribution $p(\mathbf{t}|\mathbf{x}, \Lambda_d)$ in data space where Λ_d is the covariance matrix, resulting in a final form of

$$p(\mathbf{t}) = \int p(\mathbf{t}|\mathbf{x})p(\mathbf{x})d\mathbf{x} \tag{15}$$

where the log likelihood function for this model is given by $L = \sum_i \log p(\mathbf{t}_i)$. Suppose $\mathbf{W}$ is determined by the PCA, ML can be used to fit the model to the data and hence determine values for the parameters $\mathbf{b}$ and Λ_d [27]. Using a soft clustering of the data set and multiple PCAsub corresponding to the clusters, a *mixture* of latent models takes the form of $p(\mathbf{t}) = \sum_{k=1}^{K_0} \pi_k p(\mathbf{t}|k)$ where K_0 is the number of components in the mixture, and the parameters π_k are the prior probabilities corresponding to the components $p(\mathbf{t}|k)$. Each component is an independent latent model with PCA projection $\mathbf{W}_k$ and parameters $\mathbf{b}_k$ and Λ_{dk}. This procedure can be further extended to a hierarchical mixture model formulated by $p(\mathbf{t}) = \sum_{k=1}^{K_0} \pi_k \sum_j \pi_{j|k} p(\mathbf{t}|k, j)$

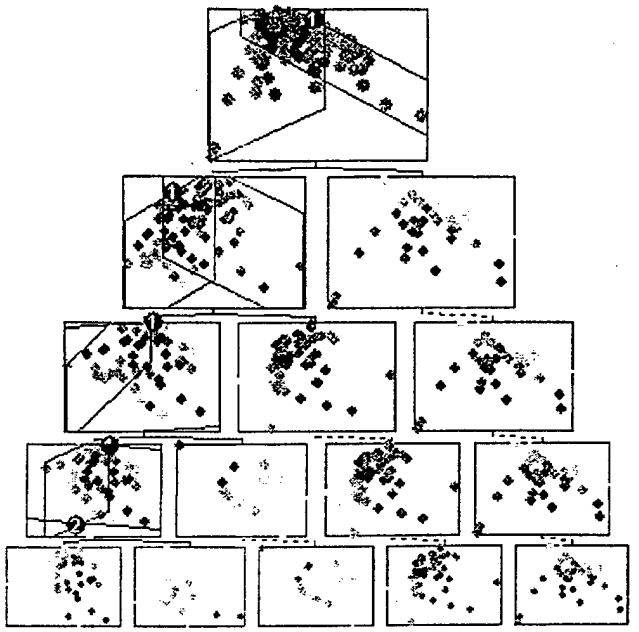


Fig. 5. The hierarchical view of computed features for mass and nonmass samples (Database A, see Table II).

where $p(\mathbf{t}|k, j)$ again represent independent latent models [27]. With a soft partitioning of the data set via EM algorithm, data points will effectively belong to more than one cluster at any given level. This step is automatically available in our approach since the estimation of parent latent model involves the calculation of posterior probabilities denoted by z_{ik}. Thus, the effective input values are $z_{ik}\mathbf{x}_i$ for an independent visualization space k, corresponding to the visualization space k in the hierarchy. It should be emphasized that *probabilistic* means both neural network based learning and posterior probability weighted inputs. Further projections can again be performed by using the effective input values $z_{ik}z_{j|k}\mathbf{t}_i$ for the visualization subspace j. Fig. 5 shows the hierarchical view of computed features for mass and nonmass samples. In Fig. 5, a hierarchical visualization view of a high dimensional feature data set was generated using hierarchical data visualization algorithm. One hundred and 25 real cases were involved, among them 75 are mass sites, 50 are nonmass sites. Nine features were computed on 125 cases. The dimension of the resulted feature data set became 125 $\times$ 9 (Database A, see Table II). Hierarchical visualization tool enables the visualization of high dimensional data set through dimension reduction and data modeling so that data distribution features of the data set can be well recognized. For instance, the clusters and subclusters of mass and nonmass data points and the boundaries of the clusters can be revealed for further research purpose.

In the use of a hierarchical minimax entropy mixture model, an interactive visualization environment is required to enable a flexible computerized experiment such that a human-database interaction can be performed effectively. We have developed an interactive environment for visualizing five-dimensional (5-D) data sets, based on state-of-the-art computer graphics toolkits such as object-oriented OpenGL and OpenInventor. With a sophisticated set of various kinds of simulated lights, color

TABLE II
THE SUMMARY OF EXPERIMENTAL DATABASES

Database	Descriptions
A	Nine features extracted from 75 mass sites and 50 non-mass sites. Used for visualizing hierarchically projected high dimensional feature space. Result is presented in Figure 5.
B	A simulated two-dimensional feature space. Used to show the effect of model selection on decision boundary estimation. Result is shown in Figure 6.
C	ORL standard database. Used to show the improvement of PMNN with decision-based learning. Result is discussed in the text.
D	The training data set consisting of 50 mammograms, with 50 true mass sites and 50 false mass sites. Three most discriminatory features are extracted. Used for both PMNN training and visualization. Result is given in Figure 7.
E	The testing data set consisting of 46 mammograms, with 23 normal cases and 23 biopsy proven mass cases with each of them having at least one true mass site. Three most discriminatory features, the same as database D, are extracted. Used to test the overall performance of our CAD system prototype where the mass candidates were selected using the method reported in Part I, automatically. Result is shown in Figure 8 and also discussed in the text.

texturing editors, and 3-D manipulator and viewers (we have integrated 3-D mouse and stereo glass units into our existing system), our system allows one to examine the volumetric data sets with any viewpoint and dynamically walk through its internal structures to better understand the spatial relationships among clusters and decision surfaces present. One of the most important features in our approach is to attach the decision surface to the 3-D probability cloud in support of decision making, and to link each data point in the visualization space to its raw data so that the user can on-line retrieve the corresponding raw data such as an original image for interim decision making.

V. EXPERIMENTAL RESULTS AND DISCUSSIONS

In this section, we present the experimental results using the information theoretic criteria and PMNNs to generate the mapping function of the featured database, and the preliminary results using the hierarchical minimax entropy projections to conduct visual explanation of the decision making. For the validation of the database mapping using the proposed algorithms, global relative entropy (GRE) value between the (SFMD)and the joint histogram is used as an objective measure to evaluate the fitness of the mapping function. A summary of the databases we used in our study is presented in Table II.

As we have discussed in Sections III and IV, model selection is the first and a very important learning task in mapping a database and the objective of the procedure is to determine both the number and the kernel shape of local clusters in each class. This procedure is used not only in the data mapping for decision making but also in the structure decomposition for hierarchical visual explanation. Our experience has suggested that an incorrect model selection will affect the performance of data-classification based decision making. For the sake of simplicity, we discuss this conclusion in the following 2-D example. Let us form a simulated featured database with two major features that well characterize the two targeted classes, as it shown in Fig. 6 (Database B, see Table II). The ground truth is that class 1 contains only one local cluster while class 2 contains two local clusters. With a model selection procedure using the proposed criteria, the intrinsic data structure was correctly identified. According to the principle of designing the optimal structure of PMNN and visual explanation hierarchy, the result of these criteria also determines the most appropriate number of mixture components in the corresponding PMNN and projected cluster decomposition. Two PMNN with different architecture orders were designed and trained to determine the classification boundaries between the two classes. The classification results are shown in Fig. 6(a) and (b). The result in Fig. 6(a) is with the right cluster number in Class 2, while the result in Fig. 6(b) is with the wrong cluster number in Class 2. From this simple experiment, we have shown that the decision boundary with the right cluster number may be much more accurate than that with heuristically determined cluster number, since the decision boundary between class 1 and class 2 will be determined by four cross points in the first case while in the second case the decision boundary will be determined by only two cross points. It should be emphasized that the error of data classification is theoretically controlled by the accuracy in estimating the decision boundaries between classes, and the quality of the boundary estimates is indeed dependent upon the correct structure of the class likelihood function.

As we have discussed before, although the knowledge database contains both machine-localized and human-selected cases, in clinical settings "unseen" and/or subtle cases contribute the major false positives. We have also pilot tested the PMNN method to the so-called "$M + 1$ classes" problem, in which the disease pattern under testing could be either from one of the M classes, or from some other unknown classes (the "unknown" class or the "intruder" class). Note that the unknown class probability is often very hard to estimate because of the lack of sufficient training samples (for example, in the mass detection problem, the unknown classes include the ROIsub over the normal tissues). In our experiment, PMNN uses different decision rule from that of the "M classes" problem: pattern $\vec{x}_i$ belongs to class r if both of the following conditions are true: a) $\phi(\vec{\omega}_r, \vec{x}_i) > \phi(\vec{\omega}_j, \vec{x}_i)$, $\forall j \neq r$, and b) $\phi(\vec{\omega}_r, \vec{x}_i) > T$. T is a threshold obtained by decision-based

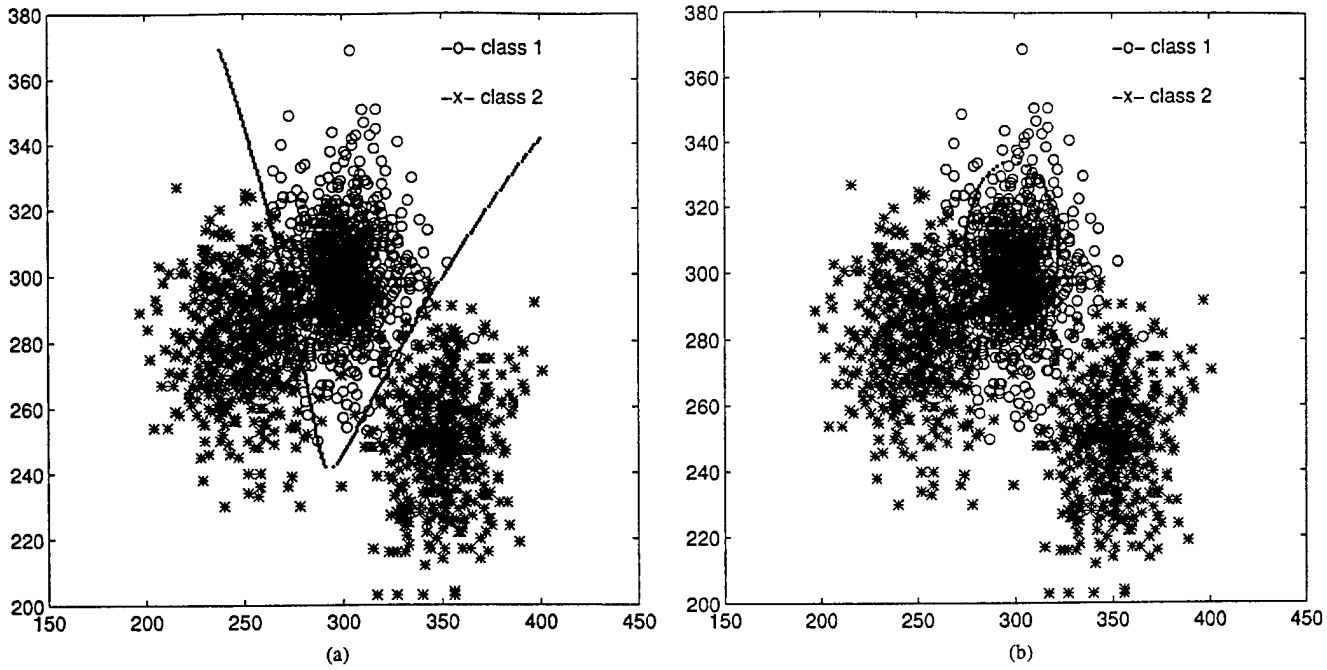


Fig. 6. The classification examples with a two-dimensional (2-D) simulated database (Dabase B, see Table II). (a) Class 2 contains two local clusters. (b) Class 2 contains one local cluster.

learning. Otherwise pattern $\vec{x}_i$ belongs to the unknown class. We observed consistent and significant improvement in classification results compared with the pure Bayesian decision. Using the ORL (Olivetti Research Laboratory, Cambridge, U.K.) standard database (Database C, see Table II), our experience has shown an increase of correct detection rate from 70% to 90% [14].

In the third experiment, we use the proposed classifier to distinguish true masses from false masses based on the features extracted from the suspected regions. The objective is to reduce the number of suspicious regions and identify the true masses. 150 mammograms, each of them contains at least one mass case of varying size and location, were selected in our study. The areas of suspicious masses were identified following the proposed procedure with biopsy proven results. Fifty mammograms with biopsy proven masses were selected from the 150 mammograms for training (Database D, see Table II). The mammogram set used for testing contained 46 single-view mammograms: 23 normal cases and 23 with biopsy proven masses (Database E, see Table II) which were also selected from the 150 mammograms. All mammograms were digitized with an image resolution of 100 μm $\times$ 100 μm/pixel by the laser film digitizer (Model: Lumiscan 150). The image sizes are 1792 $\times$ 2560 $\times$ 12 bpp. For this study, we shrunk the digital mammograms with the resolution of 400 μm by averaging 4 $\times$ 4 pixels into one pixel. According to radiologists, the size of the small masses is 3–15 mm. The middle size of masses is 15–30 mm. The large size of masses is 30–50 mm, which are rare in mammograms. A 3-mm object in an original mammogram occupies 30 pixels in a digitized image with a 100-μm resolution. After reducing the image size by four times, the object will occupy the range of about seven to eight pixels. The object with the size of seven pixels is expected to be detectable by any computer algorithm. Therefore, the shrinking step is applicable for mass cases and can save computation time.

After the segmentation, the area index feature was first used to eliminate the nonmass regions. In our study, we set $A_1 = 7 \times 7$ pixels and $A_2 = 75 \times 75$ pixels as the thresholds. A_1 corresponds to the smallest size of masses (3 mm), and an object with a area of 75 $\times$ 75 pixels corresponds to 30 mm in the original mammogram. This indicates that the scheme can detect all masses with sizes up to 30 mm. Masses larger than 30 mm are rare cases in the clinical setting. When the segmented region satisfied the condition $A_1 \leq A \leq A_2$, the region was considered to be suspicious for mass. For the purpose of representative demonstration, we have selected a 3-D feature space consisting of compactness I, compactness II, and difference entropy. According to our investigation, these three features have the better separation (discrimination) between the true and false mass classes. It should be noticed that the feature vector can easily extend to higher dimensionality. A training feature vector set was constructed from 50 true mass ROIsub and 50 false mass ROIsub (Database D, see Table II). The training set was used to train two modular probabilistic decision-based neural networks separately. In addition to the decision boundaries recommended by the computer algorithms, a visual explanation interface has also been integrated with 3-D to 2-D hierarchical projections. Fig. 7(a) shows the database map projection with compactness definition I and difference entropy. Fig. 7(b) shows the database map projection with compactness definition II and difference entropy. Our experience has suggested that the recognition rate with compactness I are more reliable than that with compactness II. In order to have more accurate texture information, the computation of the second-order joint probability matrix $p_{d,\theta}(i, j)$ is only based on the segmented region of the original mammogram. For the shrunk mammograms, we found that

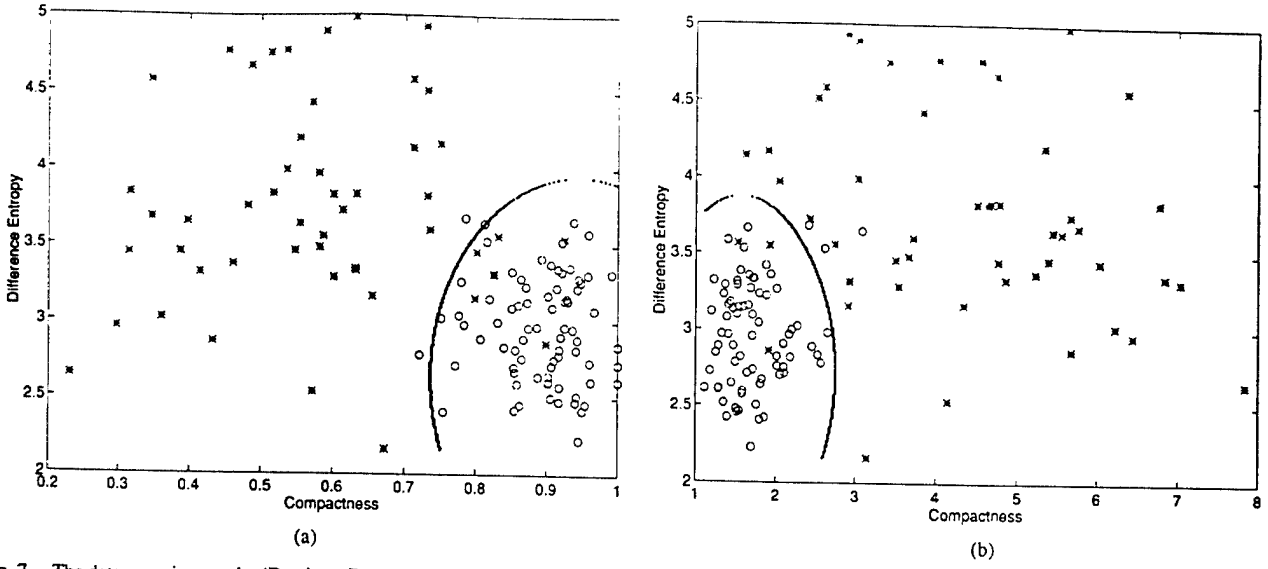


Fig. 7. The data mapping results (Database D, see Table II). -o- denotes true mass cases; -*- denotes false mass cases. (a) The mapping using compactness I. (b) The mapping using compactness II.

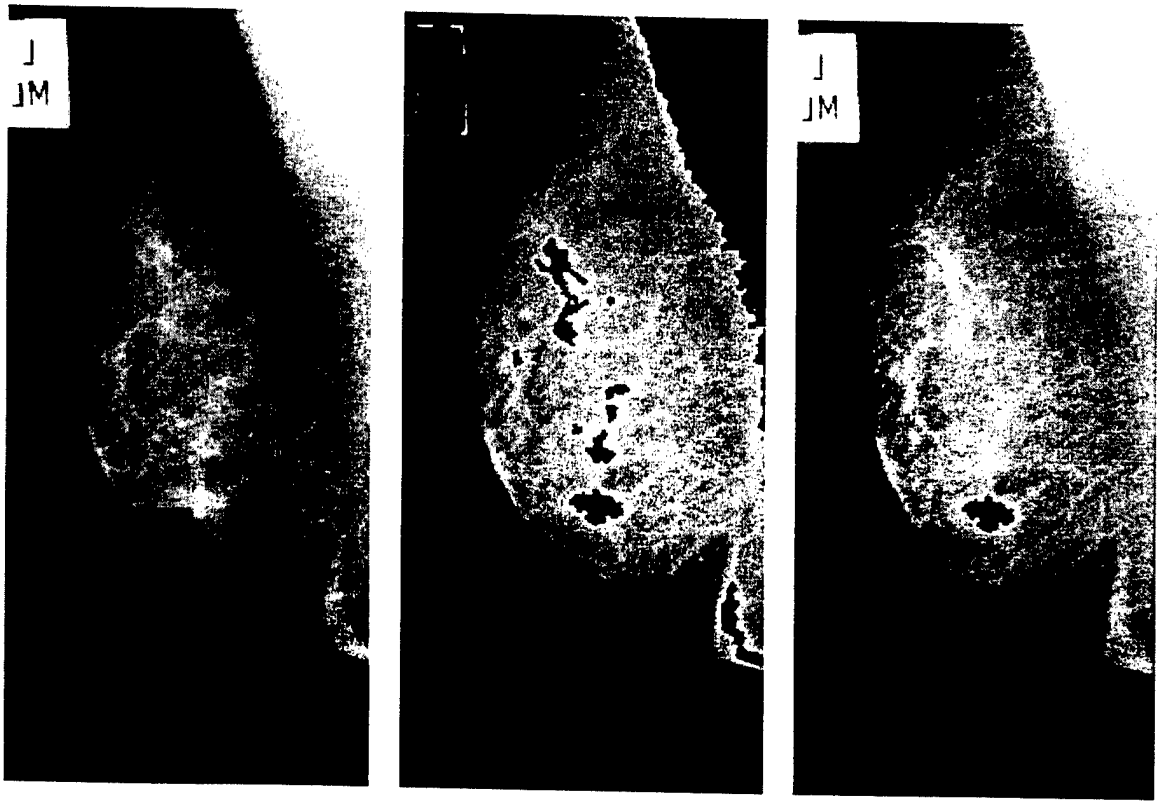


Fig. 8. One example of the mass detection using the proposed approach (Database E, see Table II).

the difference entropy had better discrimination with $d = 1$. The difference entropy used in this study was the average of values at $\theta = 0°, 45°, 90°$, and $135°$.

We have conducted a preliminary study to evaluate the performance of the algorithms in real case detection, in which 6–15 suspected masses/mammogram were detected and required further clinical decision making. We found that the proposed classifier can reduce the number of suspicious masses with a sensitivity of 84% at 1.6 false positive findings/mammogram based on the testing data set containing 46 mammograms (23 of them have biopsy proven masses) (Database E, see Table II). Fig. 8 shows a representative mass detection result on one mammogram with a stellate mass. After the enhancement, ten regions with brightest intensity were segmented. Using the area criterion, too large and too small regions were eliminated first and the rest regions were submitted to the PMNN for further evaluation. The results indicated that the stellate mass lesion was correctly detected.

For further evaluation, receiver operating characteristic (ROC) method may be employed. However, we do not feel

ROC analysis will provide really a better evaluation but an alternative method to this case. First, most ROC analysis reported by others were based on different database thus are not comparable since ROC results are highly data-dependent. Second, ROC analysis only indicate an "overall" performance with limitations at least in twofold: it is for multithreshold thus the corresponding system may not be optimal to a particular application where only one threshold is needed; and it cannot provide a mathematically traceable feedback to improve the performance of the system or the one component in the system. Third, currently used FROC analysis package imposes several assumptions on the distributions of the cases which are invalid in most applications and particularly untrue in our situation. For example, our assumptions about the data distributions is SFNM that is clearly different from the restricted conditions imposed by the application of existing FROC analysis algorithm. In our approach, a quantitative mapping of the knowledge database is performed with hierarchical SFMD modeling and should be perfectly (at least in the theoretical sense) carried out by the corresponding PMNN classifier. In other words, optimal decision making should have already been achieved according to the Bayesian rule. It is reasonable to acknowledge that in order to compare the overall performance with the other systems, an ROC study may be further conducted. We are currently working on developing a new generation of FROC analysis package with a caution to remove the forementioned problems.

Another important consideration with the present approach is the measure of quality in visual explanation [29]. This is not a glamorous area, but progress in this area is eminently critical to the future success of visual exploration [28]. What is the correct matrix for a direct projection of a particular multimodal data set? How effective was a particular visualization tool? Did the user come to the correct conclusion? It may be agreeable that the benchmark criteria in visual exploration are very different and difficult [28]. As shared by Bishop and Tipping [27], we believe that in data visualization there is no objective measure of quality, and so it is difficult to quantify the merit of a particular data visualization technique, and the effectiveness of such a techniques is often highly data-dependent. The possible alternative is to perform a rigorous psychological evaluation using simple and controlled environment, or to invite domain experts to direct evaluate the efficacy of the algorithm for a specified task. For example, we can compare the domain expert's performances with and without the system aid. In that case, the ROC method may be used to evaluate the performance of our algorithm when used by the radiologists. While the optimality of these new techniques is often highly data-dependent, we would expect the hierarchical visualization model to be a very effective tool for the data visualization and exploration in many applications.

In summary, we employed a mathematical feature extraction procedure to construct the featured knowledge database from all the suspicious mass sites localized by the enhanced segmentation. The optimal mapping of the data points was then obtained by learning the generalized normal mixtures and decision boundaries. A visual explanation of the decision making was further invented as a decision support, based on an interactive visualization hierarchy through the probabilistic principal component projections of the knowledge database and the localized optimal displays of the retrieved raw data. A prototype system was developed and pilot tested to demonstrate the applicability of this framework to mammographic mass detection.

ACKNOWLEDGMENT

The authors would like to thank R. F. Wagner of the Food and Drug Administration and S.-Y. Kung of the Princeton University for their valuable scientific input.

REFERENCES

[1] R. Zwiggelaar, T. C. Parr, J. E. Schumm, I. W. Hutt, C. J. Taylor, S. M. Astley, and C. R. M. Boggis, "Model-based detection of spiculated lesions in mammograms," *Med. Image Anal.*, vol. 3, no. 1, pp. 39–62, 1999.

[2] N. Karssemeijer and G. M. te Brake, "Detection of stellate distortions in mammogram," *IEEE Trans. Med. Imag.*, vol. 15, pp. 611–619, Oct. 1996.

[3] L. Miller and N. Ramsey, "The detection of malignant masses by nonlinear multiscale analysis," *Excerpta Medica*, vol. 1119, pp. 335–340, 1996.

[4] N. Petrick, H. P. Chan, B. Sahiner, M. A. Helvie, M. M. Goodsitt, and D. D. Adler, "Computer-aided breast mass detection: False positive reduction using breast tissue composition," *Excerpta Medica*, vol. 1119, pp. 373–378, 1996.

[5] W. K. Zouras, M. L. Giger, P. Lu, D. E. Wolverton, C. J. Vyborny, and K. Doi, "Investigation of a temporal subtraction scheme for computerized detection of breast masses in mammograms," *Excerpta Medica*, vol. 1119, pp. 411–415, 1996.

[6] M. Zhang, M. L. Giger, C. J. Vyborny, and K. Doi, "Mammographic texture analysis for the detection of spiculated lesions," *Excerpta Medica*, vol. 1119, pp. 347–351, 1996.

[7] W. P. Kegelmeyer Jr., J. M. Pruneda, P. D. Bourland, A. Hillis, M. W. Riggs, and M. L. Nipper, "Computer-aided mammographic screening for spiculated lesions," *Radiology*, vol. 191, pp. 331–337, 1994.

[8] R. N. Strickland, "Tumor detection in nonstationary backgrounds," *IEEE Trans. Med. Imag.*, vol. 13, pp. 491–499, June 1994.

[9] H. P. Chan, D. Wei, M. A. Helvie, B. Sahiner, D. D. Alder, M. M. Goodsitt, and N. Petrick, "Computer-aided classification of mammographic masses and normal tissue: Linear discriminant analysis in texture feature space," *Phys. Med. Biol.*, vol. 40, pp. 857–876, 1995.

[10] M. L. Giger, C. J. Vyborny, and R. A. Schmidt, "Computerized characterization of mammographic masses: Analysis of spiculation," *Cancer Lett.*, vol. 77, pp. 201–211, 1994.

[11] A. K. Jain, *Fundamentals of Digital Image Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1989.

[12] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-3, no. 6, pp. 610–621, Nov. 1973.

[13] R. Schalkoff, *Pattern Recognition: Statistical, Structural, and Neural Approaches*. New York: Wiley, 1992.

[14] Y. Wang, S. H. Lin, H. Li, and S. Y. Kung, "Data mapping by probabilistic modular networks and information theoretic criteria," *IEEE Trans. Signal Processing*, vol. 46, pp. 3378–3397, Dec. 1998.

[15] L. Perlovsky and M. McManus, "Maximum likelihood neural networks for sensor fusion and adaptive classification," *Neural Networks*, vol. 4, pp. 89–102, 1991.

[16] H. Gish, "A probabilistic approach to the understanding and training of neural network classifiers," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, 1990, pp. 1361–1364.

[17] D. M. Titterington, A. F. M. Smith, and U. E. Markov, *Statistical Analysis of Finite Mixture Distributions*. New York: Wiley, 1985.

[18] C. E. Priebe, "Adaptive mixtures," *J. Amer. Stat. Assoc.*, vol. 89, no. 427, pp. 910–912, 1994.

[19] S. Haykin, *Neural Networks: A Comprehensive Foundation*. New York: MacMillan College, 1994.

[20] M. I. Jordan and R. A. Jacobs, "Hierarchical mixture of experts and the EM algorithm," *Neural Computation*, vol. 6, pp. 181–214, 1994.

[21] S. Y. Kung and J. S. Taur, "Decision-based neural networks with signal/image classification applications," *IEEE Trans. Neural Networks*, vol. 1, pp. 170–181, Jan. 1995.

[22] S. H. Lin, S. Y. Kung, and L. J. Lin, "Face recognition/detection by probabilistic decision-based neural network," *IEEE Trans. Neural Networks (Special issue on Artificial Neural Networks and Pattern Recognition)*, vol. 8, Jan. 1997.

[23] Y. Wang, L. Luo, H. Li, and M. T. Freedman, "Hierarchical minimax entropy modeling and probabilistic principal component visualization for data explanation and exploration," presented at the SPIE Medical Imaging Conf., San Diego, CA, Feb. 20–26, 1999.

[24] H. Li, Y. Wang, K. J. R. Liu, S.-C. B. Lo, and M. T. Freedman, "Computerized Radiographic Mass Detection—Part I: Lesion Site Selection by Morphological Enhancement and Contextual Segmentation," *IEEE Trans. Med. Imag.*, vol. 20, no. 4, pp. 289–301, Apr. 2001.

[25] T. W. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[26] H. V. Poor, *An Introduction to Signal Detection and Estimation*. Berlin, Germany: Springer–Verlag, 1988.

[27] C. M. Bishop and M. E. Tipping, "A hierarchical latent variable model for data visualization," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, pp. 281–293, Mar. 1998.

[28] G. M. Nielson, "Challenges in visualization research," *IEEE Trans. Visual. Comput. Graphics*, vol. 2, pp. 97–99, 1996.

[29] E. R. Tufte, *Visual Explanation: Images and Quantities, Evidence and Narrative*. Cheshire, U.K.: Graphics, 1996.

Journal of VLSI Signal Processing 18, 263–274 (1998)
© 1998 Kluwer Academic Publishers. Manufactured in The Netherlands.

Application of Artificial Neural Networks to Medical Image Pattern Recognition: Detection of Clustered Microcalcifications on Mammograms and Lung Cancer on Chest Radiographs

SHIH-CHUNG B. LO, JYH-SHYAN J. LIN*, MATTHEW T. FREEDMAN AND SEONG K. MUN
ISIS Center, Department of Radiology, Georgetown University Medical Center, 2115 Wisconsin Ave., Suite 603, Washington, DC 20007

Received September 30, 1996; Revised April 25, 1997

Abstract. Three neural network models were employed to evaluate their performances in the recognition of medical image patterns associated with lung cancer and breast cancer in radiography. The first method was a pattern match neural network. The second was a conventional backpropagation neural network. The third method was a backpropagation trained neocognitron in which the signal propagation is operated with the convolution calculation from one layer to the next. In the convolution neural network (CNN) experiment, several output association methods and trainer imposed driving functions in conjunction with the convolution neural network are proposed for general medical image pattern recognition. An unconventional method of applying rotation and shift invariance is also used to enhance the performance of the neural nets.

We have tested these methods for the detection of microcalcifications on mammograms and lung nodules on chest radiographs. Pre-scan methods were previously described in our early publications. The artificial neural networks act as final detection classifiers to determine if a disease pattern is presented on the suspected image area. We found that the convolution neural network, which internally performs feature extraction and classification, achieves the best performance among the three neural network models. These results show that some processing associated with disease feature extraction is a necessary step before a classifier can make an accurate determination.

1. Introduction

Clinical studies in the use of chest radiographs for the detection of lung nodules including those reported by Stitik [1] and Heelan [2] have demonstrated that even highly skilled and highly motivated radiologists, task-directed to detect any finding of suspicion for a pulmonary nodule, and working with high quality chest radiographs, still fail to detect more than 30 percent of the lung cancers that can be detected retrospectively. In the series reported by Stitik, many of the missed lesions would be classified as T1NxMx lesions, the stage of non-small cell lung cancer that C. Mountain indicates has the best prognosis (42%, 5 year survival) [3]. This is the stage (nodules 0.3–2 cm in diameter, separate from the hilum) of lung cancer that a computer-assisted diagnostic program should tackle. Figure 1 shows a chest radiograph containing a nodule overlapped by a rib. This is a rather typical case, because 40% of the lungs are covered by posterior ribs or rib crossings.

Although mammography has a high sensitivity for detection of breast cancers when compared to other diagnostic modalities, studies indicate that radiologists do not detect all carcinomas that are visible in retrospective analyses of the images [4–6]. These missed detections are often a result of the very subtle nature of the radiographic findings. However, many missed diagnoses can be attributed to human factors such as subjective or varying decision criteria, distraction by other image features, or simple oversight [7, 8]. Early breast cancers are often characterized by masses and clustered microcalcifications [9]. It has been reported that

Current address: Deus Technologies, Inc., Rockville, MD.

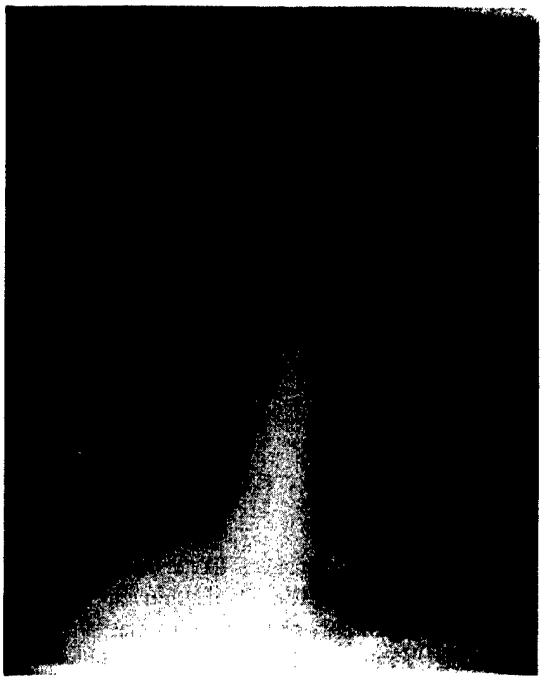


Figure 1. A chest radiograph showing a nodule overlapped on a rib.

between 40% and 50% of breast carcinomas detected radiographically demonstrate masses on mammograms [10, 11]; 30–50% of breast carcinomas presented as microcalcifications, and 60–80% of breast carcinomas reveal microcalcifications upon histologic examinations [11–13]. Breast cancer patterns associated with masses will be discussed in our future papers. Breast cancer associated clustered microcalcifications are one of two disease objects studied in this paper. Typically, the sizes of microcalcifications vary from 0.16 mm to 1.0 mm. Figure 2 shows a mammogram containing clustered microcalcifications which are surrounded by dense glandular tissues.

Various computer-based image perception techniques have been proposed for the detection of disease patterns [14, 15]. With each of these methods there is a trade-off between increased sensitivity and decreased specificity. In general, by setting less stringent criteria on computer algorithms, the sensitivity of the detecting programs can be increased. However, when using any of these methods to detect subtle diseases, we must use additional methods to decrease the number of false positives. For this reason, several investigators have attempted to use various advanced image processing and artificial classifiers to improve disease detection [16–18].

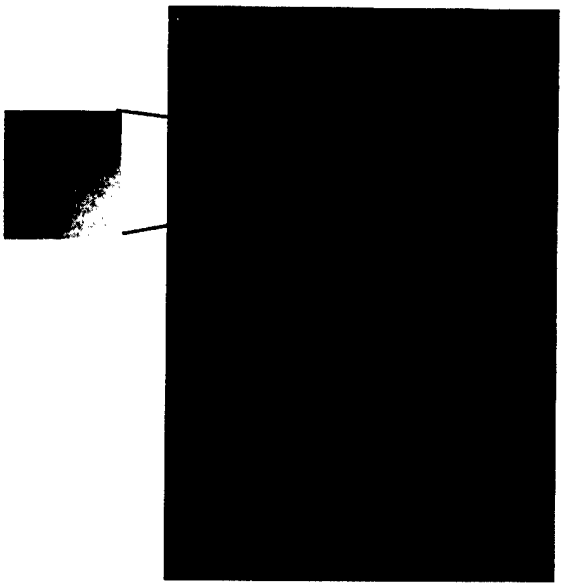


Figure 2. A mammogram showing clustered microcalcifications. The cluster area in original size is enhanced by a local histogram equalization process for display purposes.

Many artificial neural network models have recently been applied to diagnostic imaging research [19, 20]. The main tasks of these research efforts are aimed at assisting radiologists either in the accuracy improvement of quantitative measures or in the improvement of sensitivity and specificity for a disease detection. In diagnostic imaging, the neural network techniques incorporated with image processing methods have become a major research trend in the field of computer-aided diagnosis. Medical diagnoses involve very sophisticated decision-making processes. We, therefore, limited our studies to the recognition of specific disease patterns. In this paper, we will also discuss characteristics of some disease patterns in clinical images and their implications on the neural network classifications.

2. Materials and Research Objectives

2.1. Disease Patterns on Projection X-Ray Images

Projection radiographs shown on films are generated by the transmission of X-ray beams through a patient. The resulting X-rays, of varying intensity, form a radiographic image. For many years, this technique has been used as a diagnostic procedure for screening or primary examination of a disease associated with physical tissue

changes. The major drawback of projection radiography is that X-ray beams project the original anatomical three-dimensional objects onto a two-dimensional image. In other words, each pixel intensity on the image represents a total X-ray attenuation integrated from a line passing through the patient. Bone and soft tissue, and abnormal changes of tissue can be distinguished from one another in an X-ray image because they attenuate X-rays differently. However, subtle abnormalities superimposed on various normal tissues and bones are difficult to discern. The degree of sophistication in recognition of disease patterns in these images, which requires professional training, differs significantly from that of the character recognition or other image pattern recognition. The degree of difficulty is not easy to measure. Qualitatively speaking, the ratio of signal and structure noise in the task of disease pattern recognition can be very small. Consider a local suspected area that may or may not contain a disease pattern, $s(x, y) \approx d(x, y) \in P_d$, where $d(x, y)$ represents an image patch that has been proven to be a disease pattern. The collected set of these proven patches is called P_d. This local area often contains some background information resulting from normal tissues, $b(x, y) \in B$. The total intensity function denoted as $f(x, y)$ is given by

$$f(x, y) = s(x, y) + b(x, y). \tag{1}$$

In general, four situations are possible in a suspected area:

(a) $s(x, y) \gg b(x, y)$ (i.e., high signal to background ratio) representing obvious true cases;
(b) $s(x, y) \ll b(x, y)$ representing subtle cases,
(c) $s(x, y) = 0$ and $b(x, y)$ is similar to one of $d(x, y)$, where $d \in P_d$, and
(d) $s(x, y) = 0$ and $b(x, y)$ is not similar to any disease patterns, representing obvious false cases.

Most cases falling in situation (b) result in true-negatives. Cases associated with situation (c) may produce a false-positive by a classifier.

Pattern match and backpropagation, two commonly used pattern classifiers, were employed to compare the performance in the detection of clustered microcalcifications selected from mammograms and the detection of lung nodules extracted from chest radiographs. Regions of interest (ROI), formatted at $32 \times 32 \times 12$ bit, normal or abnormal, were extracted by the corresponding methods previously described [16, 17]. Both geometrical pattern and relative intensity of a local area on a radiographic image are important information in a radiographic reading. The background trend of each ROI was removed to eliminate low frequency variation [16]. However, the background structures (i.e., radiographic image of bone on chest image, vessels, and large soft tissue differences) remained in each ROI. No normalization procedure was taken, because normalization can mix a disease pattern with a non-disease pattern. For example, (a) small nodules and end-on vessels and (b) microcalcifications and film defects basically differ only in contrast. They would not be distinguishable if the feature of contrast is normalized in the pre-processing. Since many disease patterns are superimposed on background structures, we have not experienced a successful unsupervised training technique with our database. Three supervised training methods, however, achieved some success and are discussed in the following sections.

2.2. *Disease Pattern Characteristics of Microcalcifications on Mammograms and Lung Nodules on Chest Radiographs*

In general, the larger the nodule the higher the contrast of the nodule profile on the radiograph. Small rounded objects possessing high contrast are most likely end-on vessels. In addition, the size of end-on vessels is inversely proportional to their distance from the center of the heart. This is because anatomical distribution of larger arteries and vessels are closer to the heart. Clinical instruction indicates that faint tails of the vessel turned in a horizontal direction may be observable. A rib crossing, which sometimes look like an opaque round object, can also produce a false-positive detection. See Figure 3 for examples of end-on vessels, rib crossings, and true nodules.

On the other hand, the gray value differences (i.e., contrast) between the peak of microcalcifications and local background tissue are somewhat proportional to the size of the calcifications on mammograms. Film defects, caused by scratches of screen/film system or cold spots of film emulsion, are high contrast bright spots. The contrast of film defects is independent of size. Several image blocks shown in Figure 4 demonstrate the difference between microcalcifications and film defects. All image blocks were randomly selected from our database and processed by a histogram expansion for display purposes. It is essential to use a sufficiently small digitization to preserve the disease pattern. Potential problems of using a large digitization spot for

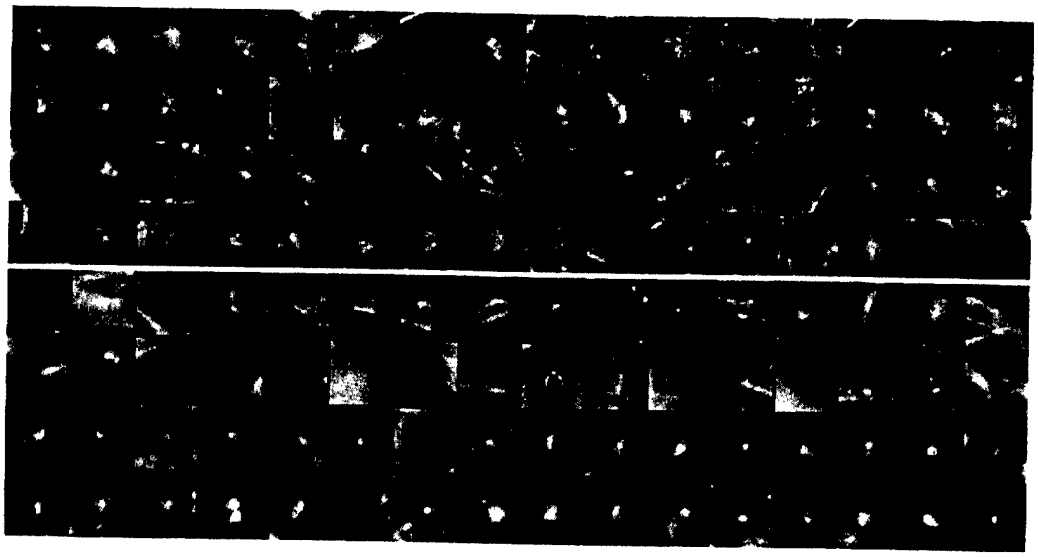


Figure 3. The upper 4 rows show 64 nodule blocks sampled from the database. Each image block on rows 5 and 6 contain no nodule but a lung or rib structure. Each image block on the bottom two rows contains an end-on vessel.

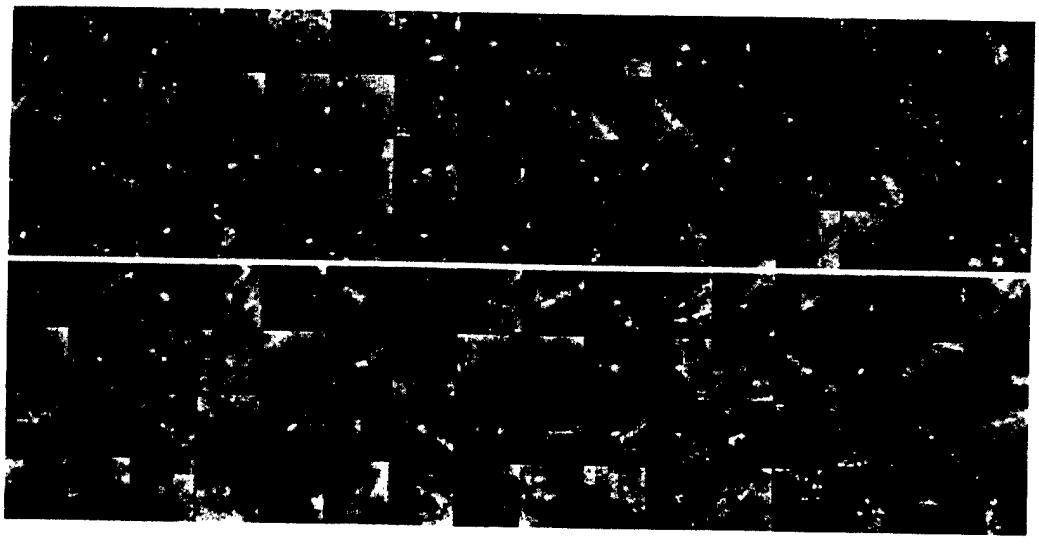


Figure 4. Each image block, extracted from mammogram, on the upper 4 rows contains at least a calcification. Each image block on the bottom 4 rows contains at least a local maximum value of gray scale (bright spot) which is not a calcification. Each block, at matrix elements (1, 4), (5, 4), (7, 4), (9, 4), and (2, 6).contains a bright spot due to a film defect.

acquiring mammographic images are: (a) the edge of a small film defect can be blurred and (b) very small microcalcifications are not actually digitized. These problems are less pronounced with a digitization spot size of 0.1 mm which was the specification of the Lumysis laser film scanner (Lumiscan Model 150).

Chest images were digitized and reformatted (shrunk by using pixel averaging) with a matrix size of $512 \times 625 \times 12$ bits per image and each pixel represents a 0.7 mm $\times$ 0.7 mm square area. Mammograms were digitized with a computer format of $2048 \times 2500 \times 12$ bits per image and each pixel represents 0.1 mm $\times$ 0.1 mm square area. The suspected microcalcification patches shown in Figure 4 are for display purposes. In the study of microcalcification detection, only the central region of 16×16 pixels (i.e., 1.6 mm $\times$ 1.6 mm) was used as input for the performance evaluation of the three neural network systems.

3. Comparative Studies Using Neural Networks

3.1. *Associated Memory Based Pattern Match Neural Networks for Disease Detection*

A classifier takes a feature vector and produces a classification. The core portion of the pattern match classifier searches for the most similar pattern in the memory. If no pattern match is found in the memory, a new pattern is created and stored for that particular classification in the memory during the training. Several neural networks belong to this type of pattern match: (a) adaptive resonance theory (ART) and its extensions (i.e., ART-2 [21], ARTMAP [22], etc.), (b) category learning originated by Reilly et al., 1992, known as RCE method [23], and (c) Dynamic Stable Associate Learning (DYSTAL) [24–26].

We used the processed image block (i.e., patch) as the input feature vector. Many feature vectors of this kind may be "contaminated" by original background structures, which are difficult to discern as disease patterns or background patterns. The authors are aware that it is important to extract features representing various aspects of disease patterns prior to the classification task. However, our goal was to compare which method better distinguishes disease patterns from non-disease image patterns using the image patches as input data. Since DYSTAL was originally designed to use image data as input for a classification task, it was selected as one of the methods for the study.

In DYSTAL, there are three rules for aggregating the input feature vector and propagating the signals:

(a) the aggregation rules are based on the correlation between the input feature vectors and learned patterns (the correlation measures the similarity between the inputs and learned patterns),
(b) the propagation rule depends on the maximum number of these resulting similarity values, and
(c) the learning rule permits the system to maintain learning patterns as needed.

The similarity measure is defined as the correlation of a learning pattern and the input feature vector [26]

$$S^j = CC(P^j, I) = \frac{\sum_i \left(P_i^j - \bar{P}^j\right) \times (I_i - \bar{I})}{\sqrt{\left(\sum_i \left(P_i^j - \bar{P}^j\right)^2 \times \sum_i (I_i - \bar{I})^2\right)}}, \quad (2)$$

where P_i^j is the value of the ith element of the jth patch vector, $\bar{P}^j$ is the mean value of the elements of the patch P^j, I_i is the value of the ith element of the input feature vector, and $\bar{I}$ is the mean value of the elements of the inputs. This similarity measure uses the cosine of the angle between the two vectors I and P^j in the n dimensional hyper-space, where n is also the number of patch elements.

The DYSTAL also uses the winner-take-all approach of propagating maximum simliarity. If the maximum similarity is lower than a pre-defined value, the new feature vector, will be stored as a newly learned pattern in the memory. The learned pattern is then assigned to an associated class which is either a true or a false disease case.

3.2. *Convolution Neural Network for Disease Pattern Recognition*

The connection between nodes in the conventional backpropagation neural network (BPNN) uniformly spreads from a front layer to a back layer [27]. However, it is known that the neighborhood correlation is usually higher than that of the long distance correlation between two pixels on an image. It is conceivable that features associated with nearby pixels should be emphasized. In neural network terms, the local signal interactions rather than non-local interactions shall be established to instruct the neural network learning. A convolution neural network (CNN), whose nets are locally formed, is selected as one of the classification methods in the experiment. The structure of the CNN is a simplified version of the neocognitron [28, 29]. We used only a 2 hidden-layer structure and eliminated all the complex-cell layers. Nets between two adjacent layers were selectively interconnected across groups. We modified the neocognitron network structure and used a convolution constrained backpropagation method for the training. This modification is necessary because (a) the original neocognitron is designed for a binary image, (b) the original 9 hidden-layer structure is very computationally intensive for an iterative training method such as the BPNN, (c) a one or two hidden-layer structure is considered adequate for relatively simple image patterns such as lung nodules and microcalcifications. Figure 5 shows the fundamental structure of this neural network.

In the CNN signal processing, each group in the receiving layer gets signals from a group of weights (e.g., kernels). For the forward signal propagation, the resultant of the weighting factors of the kernel convoluting the element values of the front layer is collected

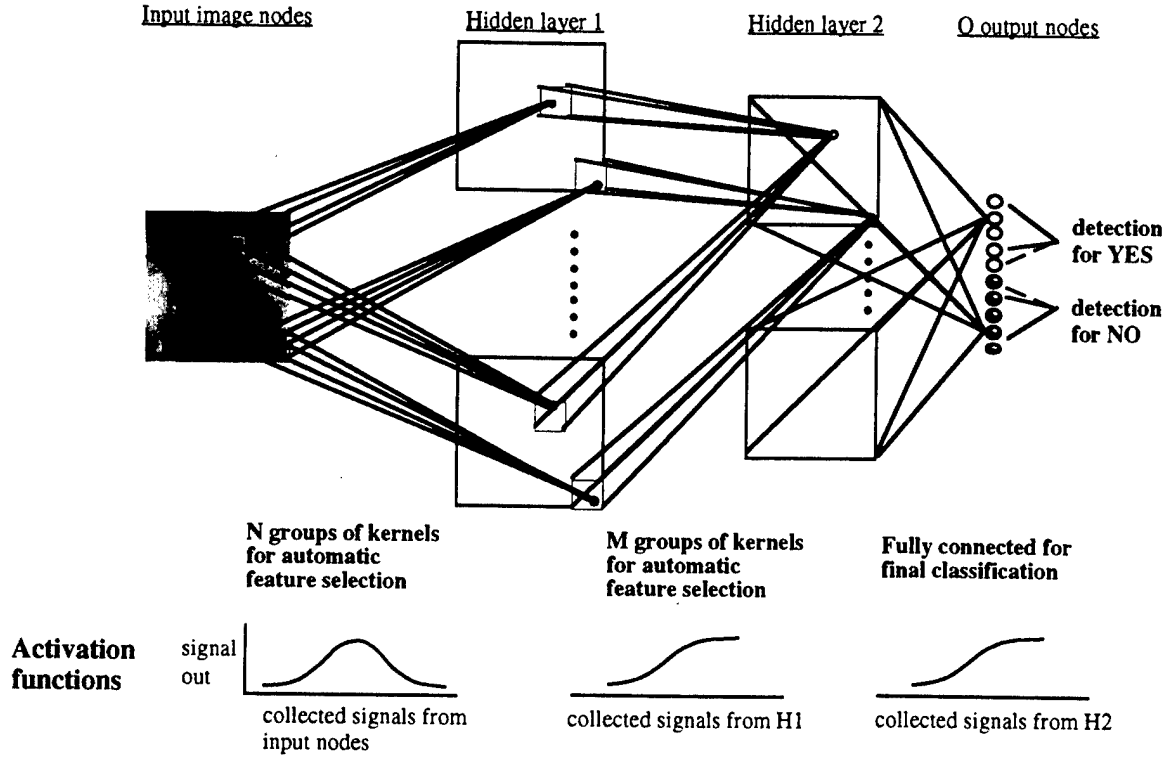


Figure 5. Artificial convolution neural network for disease pattern recognition.

onto the corresponding matrix elements of the receiving layer. This operation accounts for the major difference between the convolution type neural network and regular fully connected neural network. In the lung nodule study, we used an image patch size of 32×32 (i.e., 21.4 mm $\times$ 21.4 mm) with a convolution kernel size of 7×7. In the study of microcalcification detection, the central region of 16×16 pixels (i.e., 1.6 mm $\times$ 1.6 mm) of the original image patch size of 32×32 with a convolution kernel of 5×5 was used. The choices of using 7×7 and 5×5 convolution kernels were based on extensive studies [30, 31] in lung nodule and microcalcification cases, respectively. One reason for using much smaller size kernel is that microcalcifications are very tiny compared to observable lung nodules. In addition, small kernels are appropriate for small objects for evaluating the difference between true and false microcalcifications. Each hidden layer consists of 10 groups. The output layer has 10 nodes (2 categories) which were fully connected to the second hidden layer.

3.3. *Training of Neural Networks*

3.3.1. Classification Invariance of Matrix Operations. In general, medical image patterns possess either a circular symmetric shape (e.g., nodules) or appear as small objects with a variety of geometric patterns (e.g., calcifications). In such cases, image pattern recognition does not call on top-down or left-right geometry as classification criteria. Therefore, we can take advantage of this characteristic as an invariance. In other words, we can rotate and/or shift the input vector two-dimensionally and maintain the same output assignments for the training. This method may have two effects on the neural network: (i) to instruct the neural network that the rotation and shift of the input vector would receive the same classification result; and (ii) to increase the total number of training samples which is expected to enhance the performance of the neural network. We only rotated each suspected image block 8 times for input to test our hypothesis. Four of the rotations are: 0°, 90°, 180°, 270°. In addition, we also flipped over (left-right) the original image matrix and used the same rotations again to obtain 4 additional rotations.

3.3.2. Modification of Backpropagation Training for the CNN. As indicated in Section 2.2, a high signal of a feature can result from a negative object such as higher contrasts in end-on vessels than those in nodules

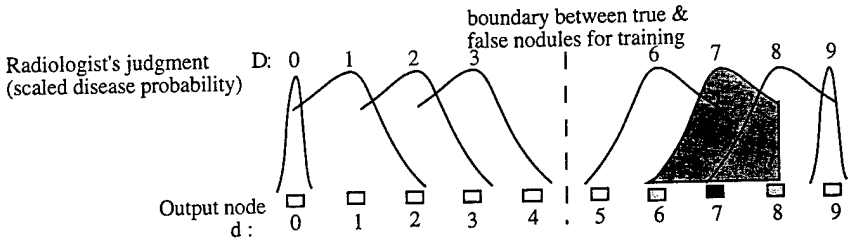


Figure 6. Fuzzy output association is constructed by a Gaussian and a trainer imposed repulsive function. Note that only one curve is used for a training case associated with an output target node (e.g., node 7 represents the activated fuzzy function).

and higher peak values in film defects than those of microcalcifications. Therefore, we used a Gaussian-like activation function for the cumulated signal propagation between input layer and the first hidden layer. The purpose of this activation function is to treat both low and high cumulated signals as false features that would eventually facilitate the classification process in the following layers. This Gaussian-like activation function would also be appropriate for the BPNN using an image block as the vector described in Section 3.5. In the conventional BPNN, fully connected rather than locally connected networks were implemented.

We used the sigmoid activation function for the forward signal propagation for all layers other than the first hidden layer and applied backpropagation training for the adjustment of weights between any two adjacent layers. The main difference between conventional weights and kernel weights is that the former are independent and the latter are constrained by grouping. By looking at the CNN processing, one may find that signals are filtered and modulated as in a circuit system. Signal propagation from one layer to the next is composed of: (a) an adaptive convolution combiner and (b) activation functions (Gaussian-like—Eq. (3)—and sigmoid—Eq. (4) functions for the first hidden layer and for other layers, respectively. See Fig. 5) which are given below:

$$S_x(i, j; n) = \frac{4 \times \exp\left\{-\sum_{u,v,m \Leftrightarrow n}[k_x(u, v; n) \times S_{x-1}(i-u, j-v; m)]\right\}}{1 + \exp\left\{-\sum_{u,v,m \Leftrightarrow n}[k_x(u, v; n) \times S_{x-1}(i-u, j-v; m)]\right\}} \quad (3)$$

and

$$S_x(i, j; n) = \frac{1}{1 + \exp\left\{-\sum_{u,v,m \Leftrightarrow n}[k_x(u, v; n) \times S_{x-1}(i-u, j-v; m)]\right\}} \quad (4)$$

where $S_x((i, j); n)$ represents the signal at node (i, j), nth group, and x layer. $k_x((u, v); n)$ denotes a weighting factor value at net (u, v), nth group, and connecting from $x-1$ to x layer. $m \Leftrightarrow n$ represents those in group m that connect to group n.

3.3.3. Backpropagation Neural Network Trained by Radiologists. We modeled radiologists' diagnostic rating (i.e., the probability of a disease existing in a suspected area) and incorporated it into the neural network training. In fact, when a radiologist determines a specific probability of a disease pattern in an image area based on his/her training and experience, this probability would be accompanied with a variation (or a standard deviation). An asymmetric output association distribution is shown in Figure 6. The use of asymmetric fuzzy assignment attempted to direct non-disease cases toward low value nodes and to push disease cases toward high value nodes. With this fuzzy assignment for the output nodes in the training, the relation between adjacent nodes was established. This supervised training can be generally applied to any situation where an association of outputs is necessary.

3.4. Classification of Output Values in the Testing

Corresponding to the grading system arranged in the training, a polarized (linearly weighted) function is given as an indication. In practice, we can define a normalized disease detection index (NDDI) for the judgment of a suspected area:

$$\text{NDDI} = \frac{\sum_{n \in \text{true nodes}}\left[O_n \times \left(n - n_0 + \frac{1}{2}\right)\right]}{\sum_{n=0}^{N-1}[O_n] \times \frac{N-1}{2}}, \quad (5)$$

where n denotes the node in the output layer, n_0 is the node number of the least likely true node, O_n is the output value at node n, and N is the total number of output nodes. Hence, a nodule detection index of 0 indicates a definite non-nodule and a nodule detection index of 1 or greater implies a definite nodule case determined by the neural network. The calculated

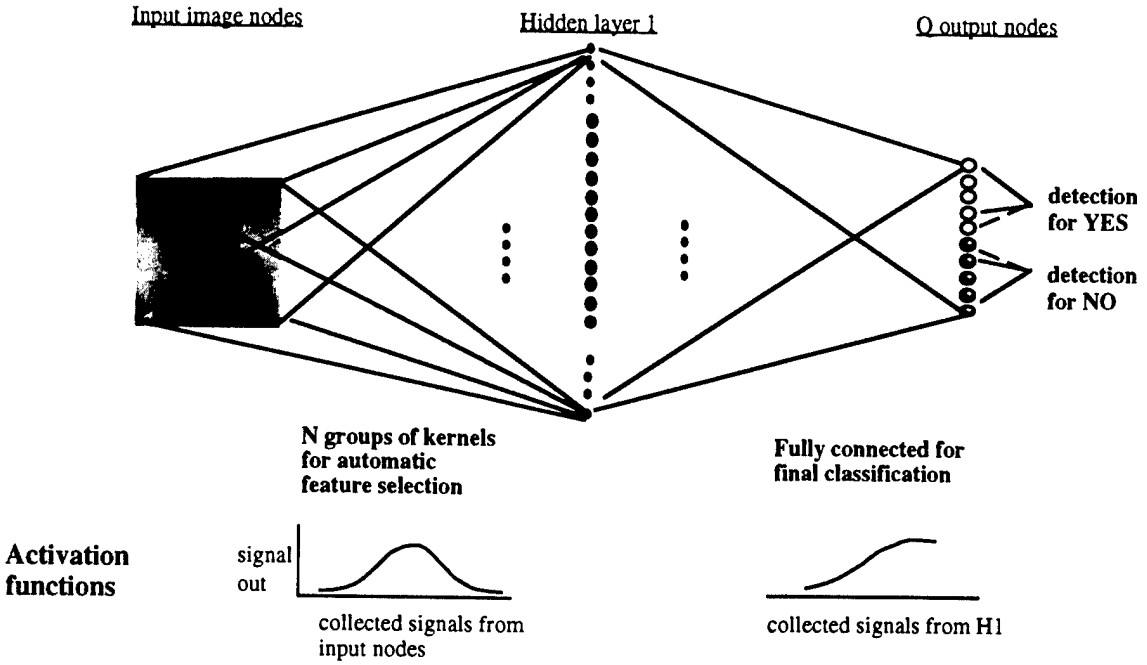


Figure 7. Artificial backpropagation neural network with fully connected nodes for disease pattern recognition.

NDDIs were evaluated by the receiver operating characteristic (ROC) analysis to measure the performance of the neural network. In general, A_z, representing the area underneath the ROC curve, is an index which signifies the performance of a system. The ROC curve is formed with the true-positive rates versus the false-positive rates of a system. We also used a performance measure—relative detection accuracy—converted from the curve to compare the results of the neural network systems.

3.5. *Backpropagation Neural Network Technique for Disease Pattern Recognition*

We have also investigated the performance of the conventional backpropagation (BP) neural network with (BP/1H) and without (BP/0H) a hidden layer. In other words, the background reduced image pixel values were used as input signals for the input layer. We expected that the hidden layer would serve as a feature extractor. The same training and testing data sets, which again were "contaminated" (i.e., very strong background sructures such as rib ovelapping a nodule) and used in the pattern match neural network, were entered into the BP neural network. Basically, we set up the experiment as described in Section 3.2 with one exception: the fully connected neural nets are used to compare the effectiveness of the neural network architecture designs. The structure of the BPNN with one hidden layer is shown in Figure 7. We tried several arrangements for the number of nodes used in the hidden layer. Our experiment indicated that approximately 200 nodes and 450 nodes in the hidden layer would be appropriate for nodule and microcalcifcation studies, respectively. These numbers may be altered according to the size of the image block used.

4. Results

4.1. *Detection of Clustered Microcalcifications*

After the pre-scan process by the computer program, 38 digital mammograms provided 220 true and 1132 false subtle microcalcifications. For the neural network studies, we divided the mammograms into two sets: (A_m) 19 images (containing 108 true and 583 false image blocks) and (B_m) another set of 19 images (containing 112 true and 549 false image blocks). We did not ask radiologists to rate image blocks in the training set. Therefore, only 2 output nodes with 8 rotated input patches were used. Neither output association nor a trainer imposed function was employed. We found that the use of a small image block of 16×16 resulted in

Table 1. Performance of neural networks in the detection of clustered microcalcifications using group A_m as training set and group B_m as testing set.

Neural networks	DYSTAL	BP/0H	BP/1H	CNN
A_z (Area under the ROC curve)	0.78	0.75	0.86	0.97
Detection accuracy				
(% true-positive detection) / (# false-positive per image)	70 / 4.3	70 / 4.5	75 / 3.5	90 / 0.5

Table 2. Performance of neural networks in the detection of clustered microcalcifications using group B_m as training set and group A_m as testing set.

Neural networks	DYSTAL	BP/0H	BP/1H	CNN
A_z (Area under the ROC curve)	0.76	0.77	0.84	0.97
Detection accuracy				
(% true-positive detection) / (# false-positive per image)	70 / 4.3	70 / 4.2	75 / 3.7	90 / 0.5

the best performance in the detection of single microcalcification [31].

Tables 1 and 2 show the performance resulting from the three neural network systems. DYSTAL and BP/0H, acting as classifiers, receive the lowest performance. The best performance index (A_z) was 0.90 when the determination was based on individual microcalcifications and was improved to 0.97 when the determination was based on the clustered microcalcifications using CNN. In the latter evaluation, suspected clusters including 1 or 2 spots within a 1 cm area were rejected and the average NDDI taken from the clustered spots was used for the ROC evaluation. This is because the detection of clustered microcalcifications is more clinically significant than individual calcifications, since the clustered microcalcifications (3 or more) are a strong indication of breast carcinoma in radiological diagnosis. The comparative study was based on detection strategies: (i) to first detect suspected individual microcalcifications and (ii) then to cluster them as group when possible; otherwise rejected the detection.

4.2. *Detection of Lung Nodules*

The first group (A_1) of image blocks were extracted from 31 chest radiographs containing multiple nodules. A senior radiologist selected 91 true nodules and 247 non-nodules areas. The second group (B_1) was collected from 31 images containing 95 nodules and 258 non-nodules and was confirmed by biopsy or by follow-up showing growth of the nodule. The pre-scan process was performed first to locate the center of the high intensity island and isolate the image block for training. For the training, each original and its 7 "brother" image blocks shared the same score vector (probability of a disease and output association) pre-determined by the radiologist. During the training, the original and its 7 "brother" image blocks were entered as a group in the same sequence. Tables 3 and 4 show the performance of using different neural network techniques and corresponding enhancement methods (i.e., fuzzy output training).

Table 3. Performance of neural networks in the detection of lung nodules using group A_1 as training set and group B_1 as testing set.

Neural networks	DYSTAL	BP/0H	BP/1H	CNN	CNN/FUZZY
A_z (Area under the ROC curve)	0.56	0.58	0.68	0.82	0.89
Detection accuracy					
(% true-positive detection) / (# false-positive per image)	60 / 7	60 / 6.6	70 / 5	80 / 4	80 / 2.5

Table 4. Performance of neural networks in the detection of lung nodules using group B_1 as training set and group A_1 as testing set.

Neural networks	DYSTAL	BP/0H	BP/1H	CNN	CNN/FUZZY
A_z (Area under the ROC curve)	0.57	0.61	0.70	0.83	0.88
Detection accuracy					
(% true-positive detection) / (# false-positive per image)	60 / 7	65 / 6.5	70 / 4.8	80 / 4	80 / 2.5

These comparison studies of both diseases imply that pattern classifiers such as DYSTAL and BP/0H cannot function alone to analyze image blocks (patches) with substantial background structures. Once the feature extraction procedure was added, the performance of the neural network increased as evident in the results of BP/1H, CNN, and CNN/FUZZY in Tables 1–4. We also learned that the convolution for two-dimensional feature extraction and fuzzy training guided by radiologists' determination were successful methods to improve the disease detection. With the neural network used in these studies, we could not isolate which procedure, the feature extraction or the final classification, was improved by the CNN training.

5. Discussion and Conclusions

Medical image pattern recognition using extracted features for input has been proposed in the detection of disease patterns [20]. Since only a small number of inputs are used, the computational time can be much less than that of the CNN for the training. As long as the features of a disease pattern are well-defined and can be quantified as values or vectors, many neural network techniques should be able to classify them. On the other hand, the CNN can internally extract features of disease patterns and is capable of distinguishing non-disease from disease patterns. A potential advantage of using the CNN is that once trained kernels are analyzed, feature extraction can be specifically defined not only by the users' experience but also by the confirmation of the CNN.

In this study, we have utilized the CNN in conjunction with several effective training methods: (i) providing a radiologists' rating scale for the training of neural nets, (ii) introducing the neural network with the classification invariance of input matrix operations, (iii) using output association functions to fuzzify the radiologists' determination and to establish the relationship between adjacent output nodes, and (iv) rendering trainer imposed functions to enhance the performance of the neural network. We found that the performance of the CNN in detecting both diseases improved significantly by administering these training methods.

Considering the convolution operation as a feature extraction processing from the input layer to the hidden layer in the CNN, we found that feature extraction is an important procedure to assist the classifier (e.g., conventional BP) in performing the recognition task. Pattern classifiers, including those newly developed neural networks, would not be able to distinguish "highly contaminated" feature vectors. Approximately 40% of lung nodules are superimposed on a posterior rib with various of orientations. However, less than 10% of microcalcifications in our database are obstructed by other abrupt breast tissues. In addition, the probability of having end-on vessels, which resemble nodules, on a chest radiograph is higher than that of film defects, which resemble microcalcifications. In other words, chest radiographs contain much more background structures than mammograms do. These background structures will contaminate the feature vector and lead to degradation of the machine observers' performance. Clinical studies also indicated that highly experienced human observers can detect only 68% of lung nodules [1] and 95% of clustered microcalcifications [5].

Acknowledgments

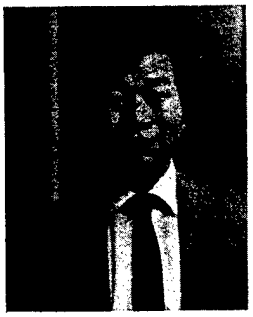
This work is supported in part by a University of Michigan Subgrant of the US Army Grant and American Cancer Society Research Award (RPG-95-034-03-EDT). The content of this information does not necessarily reflect the position or the policy of the government or American Cancer Society.

A portion of the database for the studies of microcalcifications was supplied by Dr. Heang-Ping Chan of University of Michigan, Ann Arbor. The LABROC program was written by Dr. Charles Metz and his colleagues at The University of Chicago. The authors are also grateful to Ms. Susan Kirby for her editorial assistance.

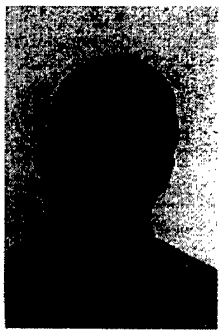
References

1. F.P. Stitik, M.S. Tockman, and N.F. Khouri, "Chest radiology," in *Screening for Cancer*, A.B. Miller (Ed.), Academic Press, New York, pp. 163–191, 1985.
2. R.T. Heelan, B.J. Flechinger, M.R. Melamed et al., "Non small cell lung cancer: Results of the New York screening program," *Radiology*, Vol. 151, pp. 289–293, 1984.
3. C.F. Montain, "Value of the new TNM staging system for lung cancer," *5th World Conference on Lung Cancer*, CHEST, Vol. 96, No. 1, pp. 47s–49s, 1989.
4. P.J. Haug, I.M. Tocino, P.D. Clayton, and T.L. Bair, "Automated management of screening and diagnostic mammography," *Radiology*, Vol. 164, p. 747, 1987.
5. C.J. Baines, A.B. Miller, C. Wall et al., "Sensitivity and specificity of first screen mammography in the Canadian national breast screening study: A preliminary report from five centers," *Radiology*, Vol. 160, p. 295, 1986.

6. L.W. Bassett, D.H. Bunnell, R. Jahanshahi, R.H. Gold, R.D. Arndt, and J. Linsman, "Breast cancer detection: One versus two views," *Radiology*, Vol. 165, p. 95, 1987.
7. J.E. Martin, M. Moskowitz, and J.R. Milbrath, "Breast cancer missed by mammography," *AJR*, Vol. 132, p. 737, 1979.
8. L. Kalisher, "Factors influencing false negative rates in xeromammography," *Radiology*, Vol. 133, p. 297, 1979.
9. L. Tabar and P.B. Dean, *Teaching Atlas of Mammography*, 2nd edition, Thieme, NY, 1985.
10. F.M. Hall, J.M. Storella, D.Z. Silverstone, and G. Wyshak, "Nonpalpable breast lesions: Recommendations for biopsy based on suspicion of carcinoma at mammography," *Radiology*, Vol. 167, p. 353, 1988.
11. E.A. Sickles, "Mammographic detectability of breast microcalcifications," *AJR*, Vol. 139, p. 913, 1982.
12. E.A. Sickles, "Mammographic features of 300 consecutive nonpalpable breast cancers," *AJR*, Vol. 146, p. 661, 1986.
13. W.A. Murphy and K. DeSchryver-Kecskemeti, "Isolated clustered microcalcifications in the breast: Radiologic-pathologic correlation," *Radiology*, Vol. 127, p. 335, 1978.
14. K. Doi, "Feasibility of computer-aided diagnosis in digital radiography," *Japanese Journal of Radiological Technology*, Vol. 45, pp. 653–663, 1989.
15. K. Doi, M.L. Giger, H. MacMahon et al., Potential usefulness of real-time computer output to radiologists' interpretations. Scientific Exhibit, Space 10-001, Presented at Raiological Society of North America, 1992 Chicago Ill.
16. H.P. Chan, K. Doi, and S. Galhotra, "Image feature analysis and computer-aided diagnosis in digital radiography: 1. Automated detection of microcalcifications in mammography," *Medical Physics*, Vol. 14, pp. 538–548, 1987.
17. M.L. Giger, K. Doi, and H. MacMahon, "Image feature analysis and computer-aided diagnosis in digital radiography: 3. Automated detection of nodules in peripheral lung field," *Medical Physics*, Vol. 15, pp. 158–166, 1988.
18. M.L. Giger, N. Ahn, K. Doi, H. MacMahon, and C.E. Metz, "Computerized detection of pulmonary nodules in digital chest images: Use of morphological filters in reducing false-positive detections," *Medical Physics*, Vol. 17, pp. 861–865, 1990.
19. S.-C.B. Lo, M.T. Freedman, J. Lin, and S.K. Mun, "Automatic lung nodule detection using profile matching and back-propagation neural network techniques," *J. Digital Imaging*, Vol. 6, No. 1, pp. 48–54, 1993.
20. Y. Wu, K. Doi, M.L. Giger, and R.M. Nishikawa, "Computerized detection of clustered microcalcification in digital mammograms: Applications of artificial neural networks," *Medical Physics*, Vol. 19, pp. 555–560, 1992.
21. G.A. Carpenter, S. Grossberg, and J.H. Reynolds, "ARTMAP: Supervised real-time learning and classification of nonstationary data by a self-organizing neural network," *Neural Networks*, Vol. 4, pp. 565–588, 1991.
22. G.A. Carpenter, S. Grossberg, and D.B. Rosen, "ART-2 A: An adaptive resonance algorithm for rapid category learning and recognition," *Neural Networks*, Vol. 4, pp. 565–588, 1991.
23. D.L. Reilly, L.N. Cooper, and C. Elbaum, "A neural model for category learning," *Biol. Cybern.*, Vol. 45, pp. 35–41, 1982.
24. D.L. Alkon, K.T. Blackwell, G.S. Barbor, A.K. Rigle, and T.P. Vogl, "Pattern-recognition by an artificial network derived from biologic neuronal systems," *Biol. Cybern.*, Vol. 62, p. 363, 1990.
25. J.M. Irvine, K.T. Blackwell, D.L. Alkon, and T.P. Vogl, "Angular separation in neural networks," *J. Artificial networks*, Vol. 1, No. 1, pp. 167–180, 1994.
26. K.T. Blackwell, T.P. Vogl, D.S. Hyman, G.S. Barbour, and D.L. Alkon, "A new approach to hand-written character recognition," *Pattern Recognition*, Vol. 25, pp. 655–666, 1991.
27. D.E. Rumelhart and J.L. McClelland, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, MIT Press, 1986.
28. K. Fukushima, S. Miyake, and T. Ito, "Neocognitron: A neural network model for a mechanism of visual pattern recognition," *IEEE Trans. on Systems, Man, and Cyber.*, Vol. 13, No. 5, pp. 826–834, 1983.
29. K. Fukushima and N. Wake, "Handwritten alphanumeric character recognition by the neocognitron," *IEEE Trans. on Neural Networks*, Vol. 2, pp. 355–365, 1991.
30. S.-C.B. Lo, H.P. Chan, J.S. Lin, H. Li, M.T. Freedman, and S.K. Mun, "Artificial convolution neural network for medical image pattern recognition," *Neural Networks*, Vol. 8, Nos. 7/8, pp. 1201–1214, 1995.
31. J.S. Lin, "Convolution neural network architecture with application for lung nodule detection in digital chest radiography," Ph.D. Dissertation 1994, Department of Electrical Engineering, University of Maryland, College Park, Maryland.

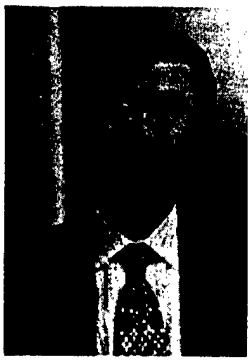


Shih-Chung B. Lo received his B.S. in Physics from National Cheng-Kung University, Taiwan, R.O.C. in 1975 and his Ph.D. in Medical Physics from UCLA in 1986. He was an MRI system engineer at Phillips Medical Systems, Shelton, Connecticut for about two years. He joined the faculty at Georgetown University in 1987 where he is currently an Associate Professor of Radiology. He has published numerous papers on computer-aided diagnosis, radiological image compression, computed tomography and digital radiography. He was the program chairman of The Workshop of Computer-Aided Diagnosis in Medical Imaging 1994 and Biomedical Application Co-Chairman of The World Congress of Neural Networks 1995. Since 1995, he has been a program committee member of the Image Processing Session of the Medical Imaging Symposium sponsored by SPIE. His current research interests are in computer-aided diagnosis, medical image processing, computed tomography, and magnetic resonance functional imaging.



Jyh-Shyan Lin received his B.S. degree in Electrical Engineering from Tankang University in 1983 and his M.S. degree in Communication Engineering in 1995 from National Chiao-Tung University, both in Taiwan, R.O.C. He received his Ph.D. degree in Electrical Engineering from the University of Maryland at College Park in 1994. He was a research assistant from 1992–1994 and became a research associate from 1994 to 1996 at the Center for Imaging Science and Information Systems (ISIS), Radiology Department, Georgetown University Medical Center. Later he served as a research scientist at Neuromedical Systems, Inc., Suffern, New York. Currently he is a senior scientist at Deus Technologies, Inc., Rockville, Maryland. His research interests include artificial neural networks, image processing, and pattern recognition.

Matthew T. Freedman received his A.B. in General Science in 1963 from the University of Rochester, Rochester, New York and received his doctorate in medicine in 1967 from the State University of New York, Brooklyn. Currently, he is Associate Professor of Radiology at the Georgetown University Medical Center, Washington, DC. He is a general radiologist with clinical and research activities in chest, musculoskeletal, and breast imaging. He is Director of Mammography Research and Clinical Director of the Center for Imaging Science and Information Systems. He has been involved in research in digital radiography, image processing and computer aided diagnosis since 1991.



Seong K. Mun received his B.S. in Physics in 1969 from University of California, Riverside, and his Ph.D. in Physics in 1979 from the State University of New York, Albany. He was an Assistant Professor of Radiation Medicine at Georgetown University Hospital from 1981 to 1983. He became Director of Imaging Physics, Department of Radiology, Georgetown University Hospital from 1982 to 1983. From 1983 to 1984, he was Managing NMR Physicist at Columbia University, Neurological Institute of New York, Columbia Presbyterian Hospital. Currently, he is Professor of Radiology at the Center for Imaging Sciences and Information Systems, Georgetown University Medical Center. He was a co-organizer of the National Forum: Telemedicine On-Line Today and the Strategic Defense Initiative Technology Application Symposium. He is the founder of the International Image Management and Communication Systems Conference; President of the Board of Scientific Counselors, National Library of Medicine, NIH; and is the Editor of *Medical Advances through Technology*.

Predictive Decomposition as a Framework in Dyadic Transforms

Shih-Chung B. Lo[1]*, *Member, IEEE,* Jianhua Xuan[2], *Member, IEEE,* and Huai Li[3], *Member, IEEE*

Abstract - A generalized decomposition method (H+PC) based on Haar transform has been derived. This general form can exactly describe dyadic transforms. Another general form (B+PC), which is a subset of the doublet system, based on the binomial filter can describe triplet-type decompositions including whole point symmetric biorthogonal transformations. Both systems can be unified by the delta function basis decomposition system (D+PC). In this paper, we found that these three bases and their expansions using predictive approximation form the dyadic decomposition family. Wavelet and integer wavelet based decomposition methods can also be included in this unified framework. This framework clearly bridges the relationship among various types of dyadic transforms. We show that many dyadic decompositions can be directly computed from their basis. A model of adaptive decomposition is also presented. Theoretical development and detailed computational methods are given. The property of low entropy in the decomposed data sequence is used as a major criterion for comparing various dyadic transforms. For the readers' convenience, the nomenclature of the symbols used in this paper is appended.

Index Terms - Dyadic Decomposition, Data Compression, Haar Transform, and Binomial Transform.

I. Introduction

In the past two decades, the applications of sub-band and wavelet decomposition methods for data compression have been extensively discussed in the literature [1,2]. The development of coding methods based on spatial-temporal correlation for multi-resolution decomposition pyramid has made those compactly supported transforms effective for image data compression [3,4]. Recently, we found that significantly different compression efficiencies were obtained while decomposing an image pattern with different orthogonal wavelets [5]. We believe that it is essential to investigate the relationship among dyadic decomposition methods prior to further analysis of the data characteristics using different decomposition methods. Through this study, we found that a predictive approximation [6,7] can be added onto basic decomposition methods such as Haar and binomial transforms and form generalized transformation systems. For each pair of mother scale and wavelet functions, there exists a set of weights to convert the transform results obtained from the base transformation into a wavelet transformation. With constraints imposed on the filter bank, the generalized system is perfectly reconstructable (PR) and is capable of producing low entropy. Both characteristics are basic requirements of the decomposition in a meaningful data compression scheme. Due to the broad area of signal decomposition, we chose to focus on discrete dyadic decompositions in this paper.

II. Haar Transform and Its Variants

The discrete Haar transform, which is formed by a doublet pair (1,1) and (1,-1), is one of the simplest and reversible transforms. For a given data sequence X: $(x_i,\ i=0,..N-1)$, the discrete Haar transformation splits the data sequence into two sequences by down-sampling:

$l_n = (x_{2n+1} + x_{2n})/\sqrt{2}$, n = 0,1,...,(N/2)-1 ...(1)

and $t_n = (x_{2n+1} - x_{2n})/\sqrt{2}$, n = 0,1,...,(N/2)-1. ...(2)

The reconstruction of the pair elements (x_{n+1}, x_n) possesses identical forms of the above two processes through up-sampling.

The binomial transform is another basic transform which is composed of two Haar bases and form a triplet pair (1,2,1) (i.e., $(1,1)\otimes(1,1)$) for low-pass and (1,-2,1) (i.e., $(1,-1)\otimes(1,-1)$) for high-pass. Using the dyadic decomposition format, we define these two pairs of bases (four basic components) as:

$L_H = \bar{X}\otimes(1,1)\downarrow_2$ $\quad H_H = \bar{X}\otimes(1,-1)\downarrow_2$

$L_B = \bar{X}\otimes(1,2,1)\downarrow_2$ $\quad H_B = \bar{X}\otimes(-1,2,-1)\downarrow_2$

In the following sections, we will show that dyadic decomposition methods can be extended from these two pairs of bases through various combinations. Strictly speaking, the Haar transform is the most fundamental form since the triplet pair can be obtained by convolving two Haar pairs.

III. High-Pass Decomposition Based on Haar Transform (H+P and S+P Transforms)

For an integer data sequence, discrete Haar transform can be approximated by Sequential (S) transform [8]. Basically, S transform computes (a) the average of the two adjacent elements of the integer data sequence as the low-pass component and (b) the difference as the high-pass component. More specifically, the former is the truncated integer of the average value. Hence Eqs. (1) and (2) can be rewritten as:

$b_n = \lfloor (x_{2n} + x_{2n+1})/2 \rfloor$ ...(3)

and $d_n = x_{2n+1} - x_{2n}$, ...(4)

where $\lfloor . \rfloor$ stands for a truncation operation that turns a real number into an integer. The corresponding inverse operations are:

$x_{2n+1} = b_n + \lfloor (d_n + 1)/2 \rfloor$ ...(5)

and $x_{2n} = x_{2n+1} - d_n$. ...(6)

III.A. Prediction in the high-pass component through Haar transform

Said and Pearlman added a predictive term, e_n, onto Eq. (4) attempting to further reduce the first-order entropy in the high-pass process [6, 7]. In the context of reversible operation, the prediction in the high-pass component can be further generalized (Ge_n).

$\hat{d}_n = (x_{2n+1} - x_{2n}) + \lfloor Ge_n + 1/2 \rfloor = d_n + \lfloor Ge_n + 1/2 \rfloor$ ...(7)

Eqs. (3) and (7) create a general form for S+P (i.e., S plus prediction) transforms. Eq. (8) is the non-truncation version of Eq. (7) for H+P (i.e., Haar plus prediction) transforms:

$$\hat{t}_n = (x_{2n+1} - x_{2n}) + Ge_n = \tilde{t}_n + Ge_n, \quad ...(8)$$

where $\tilde{t}_n$ is the same as d_n. Its corresponding low-pass counterpart is $\tilde{l}_n$ which is the non-truncation value of b_n. The estimation term, Ge_n, which is a component in the new difference value, $\hat{t}_n$, can be expressed by the decomposed values of its neighboring elements using polynomial terms

$$Ge_n = \sum_r \left(\sum_{i=-n}^{N/2-1-n} {}_r\alpha_i (\tilde{l}_{n+i})^r + \sum_{j=-n}^{-1} {}_r\beta_j (\tilde{t}_{n+j})^r \right) \quad ...(9)$$

Eq. (9) does not guarantee that an arbitrary set of ${}_r\alpha_i$ and ${}_r\beta_j$ can produce low first-order entropy. In practice, only certain sets of ${}_r\alpha_i$ and ${}_r\beta_j$ can produce low global entropy. This equation indicates that the predictive value of $\hat{t}_n$ can be composed by low-pass and high-pass components of discrete Haar transform. Usually, it only takes a few neighboring elements to compute the predictive value. Typically, the corresponding ${}_r\alpha_i$ and ${}_r\beta_j$ are small weights in order to produce a low entropy. In [6], Said and Pearlman suggested use the linear terms only (i.e., ${}_r\alpha_i = 0$ and ${}_r\beta_j = 0$ for $r \neq 1$):

$$Ge_n = e_n = \sum_{i=-n}^{N/2-1-n} \alpha_i \tilde{l}_{n+i} + \sum_{j=-n}^{-1} \beta_j \tilde{t}_{n+j} \quad ...(10)$$

They also gave several examples for S+P transform [6,7], which empirically produce low entropy, as shown in Table I.

Table I. Examples of Contribution Weights Suggested by Said and Pearlman.

Examples of Prediction	Contribution terms				
	α_1	α_0	α_{-1}	α_{-2}	β_{-1}
A	0	1/4	0	-1/4	0
B	0	1/4	-1/8	-3/8	1/4
C	-1/16	5/16	1/4	-1/2	3/8

Remark: The contribution terms given by [6] are scaled differently from Eq. (10). β_0=1 for all S+P cases.

Providing the property of perfect reconstruction, Eq. (8) is the high-pass wing of a generalized dyadic transform. Its reconstruction process can be obtained by reversing the computing order in the inverse H+P transform:

$$\tilde{t}_n = \hat{t}_n - Ge_n = \hat{t}_n - \left(\sum_r \left(\sum_{i=-n}^{N/2-1-n} {}_r\alpha_i (\tilde{l}_{n+i})^r + \sum_{j=-n}^{-1} {}_r\beta_j (\tilde{t}_{n+j})^r \right) \right) \quad ...(11)$$

and its counterpart in the inverse S+P transform is:

$$d_n = \hat{d}_n - \lfloor Ge_n + 1/2 \rfloor =$$

$$\hat{d}_n - \left\lfloor \sum_r \left(\sum_{i=-n}^{N/2-1-n} {}_r\alpha_i (b_{n+i})^r + \sum_{j=-n}^{-1} {}_r\beta_j (d_{n+j})^r \right) + 1/2 \right\rfloor \quad ...(12)$$

The average values, $\tilde{l}_{n+i}$, are always available during the reconstruction of $\tilde{t}_n$. However, only those $\tilde{t}_{n+j}$, where j = -n, -n+1, ... 0, are available while computing reconstruction values from low to high indices. The inverse S+P transform shares the same context.

For data compression, particularly for lossless compression, a minimum requirement for data decomposition is that decomposition operation must be reversible. Since Haar and S bases are reversible, Eqs. (8) and (9) provide a new dimension for generalization of the system. In the following section, we show this generalization approach and attempt to link it with the two-band and orthogonal wavelet decompositions. Furthermore, Eq. (8) as a generalized high-pass form of Haar based dyadic decompositions implies that implementation of switching different H+P transforms can be performed. The use of different sets of S+P transforms on the same data sequence is also allowed with Eq. (7). This is because we can convert decomposed values between two transformation systems using two sets of α_i and β_j while operating different transforms on different characteristics of data segments. We will further discuss this application in Section VII.

III.B. High-pass decomposition in S+P and H+P transforms

High-pass coefficients in a generalized system are obtained by adding a predictive term in Haar or S transform. Even beginning with an integer data sequence, the high-pass coefficients of S+P are different from those obtained from H+P. Nevertheless, the decomposed data are highly correlated. The decomposed data in the two low-pass domains are slightly different in the truncated lowest bit. They also differ in the scaling factor of $1/\sqrt{2}$ (i.e., $b_n \approx \tilde{l}_n = l_n/\sqrt{2}$). In high-pass domains, however, the scaling factor is the main difference between the two systems (i.e., $d_n = \tilde{t}_n = \sqrt{2}t_n$). It is seen that the decomposed coefficients in the low-pass and high-pass domains are scaled independently. The use of different scaling factors is mainly due to the requirement of integer computing in S family systems.

In the following subsection, we use predictive terms (i.e., Eqs. (7) and (8)) to expand high-pass processes as a part of generalization in dyadic decomposition. The low–pass processes remain unchanged so that the computation in H+P and its approximation in S+P follow Eqs. (1) and (3), respectively. The reconstruction process can be arranged by inverse sequential operation or designing an orthogonal based low-pass to satisfy the condition $\sum_u h_u h_{u-2l} = \delta_{l0}$ for $\forall l$ [9].

III.C. Reformulating m-tap orthogonal filter coefficients, (h_u, u=0,1,...m-1) in wavelet transform

Since l_n and t_n are two decomposed components of Haar transform, Eqs. (1) and (8) representing the two wings of H+P transform can be computed through scaled Haar bases ($\tilde{l}_n$ and $\tilde{t}_n$), we can employ scaled Haar bases to

develop S+P and H+P transform systems. As an example, the high-pass process of an m-tap orthogonal wavelet filtering can be made equivalent to that of an H+P transform (i.e., Eq. (8)). Providing an m-tap mother scale function (h_u, $u=0,1,..,m-1$), a unique set of α_i and β_j can be found to match the wavelet function (g_v, $v=0,1,..,m-1$ for $g_v=(-1)^{m-1-v}h_{m-1-v}$). By omitting the convolution operations with the data vector $\vec{x}$ for the both sides of the equation, we let

$$\begin{aligned}&(.....)\times\alpha_i\\&+(...\ \tfrac{1}{2},\ \tfrac{1}{2},\ 0,\ 0)\times\alpha_{-1}\\&+(...\ 0,\ 0,\ \tfrac{1}{2},\ \tfrac{1}{2})\times\alpha_0\\&+(.....)\times\beta_j\\&+(...-1,\ 1,\ 0,\ 0)\times\beta_{-1}\\&+(...\ 0,\ 0,-1,\ 1)\times\beta_0\\&=(...-h_3,h_2,-h_1,h_0)\times C\end{aligned} \qquad ...(13)$$

where C is the offset scaling factor mentioned earlier. Based on Eqs. (8) and (10), $b_n=(...\ 0,\ 0,\ \frac{1}{2},\ \frac{1}{2})\otimes\vec{X}$, $b_{n-1}=(...\ \frac{1}{2},\ \frac{1}{2},\ 0,\ 0)\otimes\vec{X}$, $\tilde{t}_n=(...\ 0,\ 0,-1,\ 1)\otimes\vec{X}$, $\tilde{t}_{n-1}=(...-1,\ 1,\ 0,\ 0)\otimes\vec{X}$, and so on; where $\otimes$ represents the convolution. In addition, β_0 should be unity. The rest of the contribution weights (i.e., α_i and β_j) as well as C can be solved in terms of the filter coefficients (i.e., h_u) of an m-tap transformation. Specifically:

$$C=2/(h_0+h_1),\ \alpha_0=2(h_0-h_1)/(h_0+h_1) \text{ and } \beta_0=1,$$
$$\alpha_i=2(h_{2i}-h_{2i+1})/(h_0+h_1) \text{ and}$$
$$\beta_j=(h_{2j}+h_{2j+1})/(h_0+h_1) \qquad ...(14)$$

This solution indicates that Eq. (8) is the high-pass wing of the generalized decomposition form that covers high-pass of the orthogonal wavelet transforms. The computed values using Eq. (8) with weights given by Eq. (14) are exactly the same as the high-pass coefficients through wavelet transform multiplied by the constant C. Depending upon the low-pass component, the inverse transformation can be made through inverse sequential process or its corresponding wavelet-based synthetic process. The synthetic filter coefficients are usually rearranged through the decomposition filter coefficients. Without a corresponding decomposition in the low-pass wing, the use of inverse sequential process is necessary for perfect reconstruction in S+P and H+P systems. In fact, inverse sequential operation is the only way for an S+P system to be perfectly reconstructable. A similar algorithm development can be derived to associate doublet-type two-band filter with H+P systems.

From Eq. (14), we can easily find that $\sum_i\alpha_i=C\sum_u(-1)^u h_u$. Since the property of zero-mean of high-pass filtering is maintained (i.e., $\sum_u(-1)^u h_u=0$) in an orthogonal wavelet or a two-band system, $\sum_i\alpha_i$ must vanish to match the case [9]. There are two physical meanings associated with this situation: (a) the sum of contribution weights is 0 and (b) the contribution value by the neighbors can be reformatted by difference values of the average values. It is seen that the prediction can be made not only by the difference values of the adjacent elements but also from the difference values of the average values (i.e., $\tilde{l}_n$ or b_n). This certainly is an effective strategy for making a good prediction through a decomposition method.

IV. High-Pass in Binomial Based Decomposition (B+P and S_b+P Transforms)

Instead of operating two adjacent elements of the data sequence in Haar transform, binomial family systems possessing symmetric filters operate odd number of adjacent elements for each set of convolution computation. The binomial basis is formed by a triplet pair: (1,2,1) and (-1,2,-1) mentioned earlier (Section II). In this paper, the binomial filter based transform using S transform operations is called S_b transform. The transform operations, corresponding to Eqs. (3) and (4), for the average and difference values are given below:

$$b'_n=\lfloor(x_{2n-1}+2x_{2n}+x_{2n+1})/4\rfloor=$$
$$\lfloor(x_{2n}+\lfloor(x_{2n-1}+x_{2n+1})/2\rfloor)/2\rfloor,\ n=1,2,...,(N/2)-1, \qquad ...(15)$$
$$d'_n=2x_{2n}-x_{2n-1}-x_{2n+1},\ n=1,2,...,(N/2)-1. \qquad ...(16)$$

In fact, d'_n is a composed difference value. In practice, b'_0 follows Eq. (3), and d'_0 follows Eq. (4) multiplied by 2. In this paper, we use l'_n and t'_n for the corresponding decomposed low-pass and high-pass values of binomial transform and use $\tilde{l}'_n$ and $\tilde{t}'_n$ for the non-truncation values of b'_n and d'_n, respectively. The inverse operations of the S_b transform are given below:

$$x_{2n}=b'_n+\lfloor(d'_n+3)/4\rfloor \qquad ...(17)$$
$$\text{and } x_{2n+1}=2x_{2n}-d'_n-x_{2n-1}. \qquad ...(18)$$

where x_{2n-1} is obtained from the last computation (i.e., $2x_{2n-2}-d'_{n-1}-x_{2n-3}$) during the reconstruction process.

IV.A. Prediction in the high-pass process of binomial decomposition - S_b+P and B+P [10,11,12]

There is a subset of binomial based filters whose filter bank posses a property of $(-1)^k k_{2n+1}=(-1)^k k_{-2n+1}$ This subset system can be described by B+P transform. The reconstruction process can be arranged by inverse sequential operation or designing a biorthogonal based low-pass to satisfy the condition $\sum_n\tilde{k}_n k_{n-2l}=\delta_{l0}$ *for* $\forall l$ [11]. This subset includes whole point symmetric (WPS) filters (i.e., $k_{2u-n}=k_n$ for a constant $u\in Z$ and $\forall n$). Half point symmetric (HPS) filters (i.e., $k_{2u+1-n}=k_n$ for a constant $u\in Z$ and $\forall n$) and other non-binominal filters cannot be directly expanded from the triplet system.

We can add an estimation term onto Eq. (16), as turning Eq (2) to Eq. (7) in the doublet system, to further reduce the first-order entropy:

$$\hat{d}'_n = (2x_{2n} - x_{2n-1} - x_{2n+1}) + \lfloor Ge'_n + 1/2 \rfloor = d'_n + \lfloor Ge'_n + 1/2 \rfloor \quad ...(19)$$

The estimation term, Ge'_n, in the high-pass of S_b+P transform, $\hat{d}'_n$, can be expressed by its neighbors and associated values in the decomposed data sequence. In fact, Ge'_n shares the same form of Ge_n given by Eq. (9). The reconstruction process for d'_n shares the same form of d_n as shown in Eq. (12). However, the new expressions of Ge'_n and d'_n contain the coefficients of S_b+P decomposition rather than the coefficients of S+P decomposition. The non-truncation process of the above derivations (i.e., Eqs. (15)-(19)) is called B+P transform herein.

The decomposed coefficients in the low-pass process are multiplied by a scaling a factor of $1/\sqrt{2}$ (i.e., $b'_n \approx \tilde{l}'_n = l'_n/\sqrt{2}$). In high-pass domains, the scaling factor is the main difference between these two binomial systems (i.e., $d'_n = \tilde{t}'_n = 2\sqrt{2}t'_n$). Again, due to the requirement of integer computing in the S_b family, their scaling factors are different in these two versions of binomial decompositions.

Eq. (15) and its non-truncation version represent the low-pass decompositions for S_b+P and B+P transforms, respectively. The general form of the high-pass component in a binomial system follows the non-truncation version of Eq. (19).

IV.B. Reformulating m'-tap biorthogonal wavelet coefficients, (k_u , u=-m,...0,...m, and m= (m'-1)/2)

Based upon the symmetric indices used in general binomial filter, we turn Eq. (9) into a locally symmetric operation to facilitate the study between the predictive values and triplet-type filter coefficients. A generalized predictive expression for the B+P transform is given by:

$$\hat{t}'_n = \tilde{t}'_n + Ge'_n \quad ...(20)$$

where $Ge'_n = \sum_r \left(\sum_{i=-\overline{m}}^{\overline{m}} {}_r\alpha'_i (\tilde{l}'_{n-\overline{m}+i})^r + \sum_{j=-\overline{m}}^{\overline{m}-1} {}_r\beta'_j (\tilde{t}'_{n-\overline{m}+j})^r \right)$ and $\overline{m} = (m+1)/2$. A reduced polynomial expression using the first-order terms is

$$e'_n = \sum_{i=-\overline{m}}^{\overline{m}} \alpha'_i \tilde{l}'_{n-\overline{m}+i} + \sum_{j=-\overline{m}}^{\overline{m}-1} \beta'_j \tilde{t}'_{n-\overline{m}+j} \quad ...(21)$$

For S_b+P transform, b'_n and d'_n would replace $\tilde{l}'_n$ and $\tilde{t}'_n$, respectively. The low-pass wing of a B+P transform is the same as the binomial low-pass component that is the non-truncation version of Eq. (14). Providing an m'-tap filter coefficients, a set of α'_i and β'_j can be found so that the high-pass process of a B+P transform can be made equivalent to those of an m'-tap biorthogonal wavelet. This statement can be proven as shown below. We arrange the filter length (m' or m'+2) and length of contribution weights (m), so that $m = (m'-1)/2$ if mod(m'+1, 4)=0; otherwise $m=(m'+1)/2$. In the latter case, we add two terms $k_m = k_{-m} = 0$, so that the m'-tap becomes an m'+2-tap filter in the following derivation.

$$\begin{aligned}
&(\tfrac{1}{4}, \tfrac{1}{2}, \tfrac{1}{4}, 0,\ldots\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0\ldots) \times \alpha'_{-\overline{m}} \\
&\quad\ldots \\
&+(\quad\ldots\ \tfrac{1}{4}, \tfrac{1}{2}, \tfrac{1}{4},\ 0,\ 0,\ 0,\ 0\ldots) \times \alpha'_{-1} \\
&+(\quad\ldots\ 0,\ 0,\ \tfrac{1}{4}, \tfrac{1}{2}, \tfrac{1}{4},\ 0,\ 0\ldots) \times \alpha'_{0} \\
&+(\quad\ldots\ 0,\ 0,\ 0,\ 0,\ \tfrac{1}{4}, \tfrac{1}{2}, \tfrac{1}{4}\ldots) \times \alpha'_{1} \\
&\quad\ldots \\
&+(\quad\ldots\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ldots 0,\ \tfrac{1}{4}, \tfrac{1}{2}, \tfrac{1}{4}) \times \alpha'_{\overline{m}} \\
&+(-1,\ 2,-1,\ 0,\ldots\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0\ldots) \times \beta'_{-\overline{m}} \\
&\quad\ldots \\
&+(\quad\ldots -1,\ 2,-1,\ 0,\ 0,\ 0,\ 0\ldots) \times \beta'_{-1} \\
&+(\quad\ldots\ 0,\ 0,-1,\ 2,-1,\ 0,\ 0\ldots) \times \beta'_{0} \\
&+(\quad\ldots\ 0,\ 0,\ 0,\ 0,-1,\ 2,-1\ldots) \times \beta'_{1} \\
&\quad\ldots \\
&+(\quad\ldots\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ldots 0,-1,\ 2,-1) \times \beta'_{\overline{m}} \\
&= (-k_{-m}, \ldots\ldots\ldots -k_{-3}, k_{-2}, -k_{-1}, k_0, -k_1, k_2, -k_3, \ldots\ldots\ldots\ldots -k_m) \times C'
\end{aligned} \quad ...(22)$$

where C' is the scaling factor mentioned earlier and $C' = 4\beta'_0 \Big/ \left[h_0 + \sum_{u=-\overline{m}}^{\overline{m}} (-1)^u h_{2u+1} \right]$. Let β'_0 be unity for the central term of second part of the summation representing $\tilde{t}'_n$. The solution would be $\alpha'_0 = 2k_0 C'_0 - 4$, where $C'_0 = 4 \Big/ \left[h_0 + \sum_{u=-\overline{m}}^{\overline{m}} (-1)^u h_{2u+1} \right]$.

The other contribution weights are

$$\alpha'_{i+1} = [(k_{2i} - 2k_{2i+1} + k_{2i+2})C'_0] - \alpha'_i \quad ...(23)$$

and

$$\beta'_{j+1} = [(k_{2j} + 2k_{2j+1} + k_{2j+2})C'_0/4] - \beta'_j.$$

Since C' serves as a scaling factor between two decomposition systems, one can use it to adjust the range of decomposed coefficients for different applications. From Eq. (23), we can obtain that $\sum_i \alpha'_i = 2C' \sum_u (-1)^u k_u$. Since $\sum_u (-1)^u k_u = 0$ as the property of zero-mean filtering in a biorthogonal wavelet system, $\sum_i \alpha'_i$ vanishes [10]. The above formulation does not meet the PR criterion through the sequential decomposition and reconstruction embedded in Eqs. (19) and (20). With a corresponding low-pass component (see section VI), the inverse transformation of the B+P system is perfectly reconstructable through its biorthogonal synthetic process. The length of synthetic filter and its coefficients usually differ from those of the decomposition (analytic) filter.

If we let $\beta'_{\bar{m}}$ be unity for the last term of the summation representing $\tilde{t}'_n$. The remaining weights (i.e., α'_i and β'_j) as well as the C' value can be solved in terms of the filter coefficients (i.e., k_u) of the m'-tap triplet-based transformation. A set the predictive weights can be obtained:

$$C'_{\bar{m}} = 4/(k_m + k_{m-1}),$$
$$\alpha'_{\bar{m}} = (k_{m-1} - k_m)C'_{\bar{m}} \text{ and } \beta'_{\bar{m}} = 1,$$
$$\alpha'_i = [(k_{2i} - 2k_{2i+1} + k_{2i+2})C'_{\bar{m}}] - \alpha'_{i+1} \quad ...(24)$$
and
$$\beta'_j = [(k_{2j} + 2k_{2j+1} + k_{2j+2})C'_{\bar{m}}/4] - \beta'_{j+1}.$$

With this decomposition, the sequential operation is reversible. Unfortunately, the above formulation would produce high number coefficients particularly in a long filter system. Because biorthogonal wavelets possess a property of $|k_{\pm u}| > |k_{\pm u \pm p}|$ for $u \times p > 0$. In [12], $|k_0/k_{\pm m}|$ ranges from 6 for 5-tap to 90 for 13-tap while designing effective biorthogonal filter coefficients.

Based on Eq. (23), an approximation can be made by folding the filter coefficients for sequential reconstruction required by the S_b family. We assume that the data sequence is symmetric as a mirror for each convolution operation. This assumption is false in reality. However, it still can produce a good prediction for Eq. (20) as an approximation. With this assumption, the property of perfection reconstruction resumes in Eq. (19) through the sequential operation. Folding the filter coefficient is also equivalent to turning the triplet system into doublet system so that $K_{-u} = 2k_{-u}$ and $K_u = 0$ for u=1,.. m and $K_0 = k_0$. This approximation also alters a non-causal filtering process onto a causal process that is required in an S_b based transform for the purposes of perfect reconstruction. The contribution weights with negative indices in Eq. (24) become

$$C'_0 = 4/k_0,$$
$$\alpha'_0 = k_0 C'_0/16 \text{ and } \beta'_0 = 1,$$
$$\alpha'_{-i-1} = [(K_{-2i} - 2K_{-2i-1} + K_{-2i+2})C'_0] - \alpha'_{-i} \quad ...(25)$$
and
$$\beta'_{-j-1} = [(K_{-2j} + 2K_{-2j-1} + K_{-2j-2})C'_0/4] - \beta'_{-j}.$$

Other weights with positive indices vanish (i.e., $\alpha'_i = \beta'_j = 0$ for i>0 or j>0).

The above three equations, each as a high-pass component of the system, serve different purposes: Eq. (23) is used in B+P (and B+PC to be discussed) transform; Eq. (24) can be used for S_b+P and B+P transforms; and Eq. (25) is an approximation version for the S_b+P and B+P transforms. Strictly speaking, only Eq. (23) shows the relationship between the high-pass components of B+P and the biorthogonal wavelet systems. Eqs. (24) and (25) are used for the purposes of sequential decomposition with reversible process. In a generalized B+P transformation (i.e., Eq. (20) or (21)), the PR property may or may not exist due to the non-causal characteristics of the filters.

V. Haar Plus Prediction and Composite (H+PC Transforms)

In Section III, we showed that the high-pass process of an orthogonal wavelet transform can be computed through Haar basis (i.e., Eqs. (13) and (14)). The high-pass process actually computes extended difference values summarized in Eq. (8). With a computer implementation, each transform coefficient must be stored in a real number. Therefore, the truncation and approximated rational values used in Section III should be abandoned when a prediction is used to compute the decomposed coefficients of an orthogonal wavelet transform. Similar to Eq. (14), the low-pass process of an orthogonal dyadic wavelet transform can be computed through Haar bases and becomes H+PC that stands for generalized prediction including both the high-pass and low-pass processes. The low-pass component in H+PC is:

$$\hat{l}_n = \tilde{l}_n + a_n = \tilde{l}_n + \left(\sum_i \gamma_i \tilde{l}_{n+i} + \sum_j \lambda_j \tilde{t}_{n+j} \right) \quad ...(26)$$

where a_n is the added composite value. The use of polynomial terms for the generalization of composite value would increase the complexity of the system. The PR property of each system would require further investigation. Additionally, its application is not observed in the low-pass component. To show that the low-pass of orthogonal wavelets is a subset of H+PC systems, we let

$$\begin{aligned}
&(.....)\times\gamma_i \\
&+(\ 0,\ 0,\ \tfrac{1}{2},\ \tfrac{1}{2}...)\times\gamma_1 \\
&+(\ \tfrac{1}{2},\ \tfrac{1}{2},\ 0,\ 0...)\times\gamma_0 \\
&+(.....)\times\lambda_j \\
&+(\ 0,\ 0,-1,\ 1...)\times\lambda_1 \\
&+(-1,\ 1,\ 0,\ 0...)\times\lambda_0 \\
&=\overline{(h_0, h_1, h_2, h_3...)}
\end{aligned} \quad ...(27)$$

The solution for the above equation is straightforward:

$$\gamma_i = (h_{2i} + h_{2i+1}) \quad \text{and} \quad \lambda_j = (h_{2j+1} - h_{2j})/2 \quad \text{for i\&j} = 0,...,m/2. \quad ...(28)$$

It is seen that the main contribution to the low-pass filtering comes from $(\ \tfrac{1}{2},\ \tfrac{1}{2},\ 0,\ 0...)\times\gamma_0$. When it convolves with the data sequence, the result is $l_0\gamma_0/\sqrt{2}$. Since neither l_n nor b_n values are maintained in a computer implementation, the inverse sequential reconstruction using Eq. (11) is no longer a valid method. Eqs. (14) and (28) show methods to convert decomposed Haar transform coefficients (scaled) to transform coefficients of a PR two-band transform [2] (including dyadic orthogonal wavelets). The inverse transformation of the transformed coefficients should follow its corresponding dyadic inverse transform operation (e.g., inverse wavelet transform). In contrast to the high-pass process, the summation of weights contributed by difference values is 0 (i.e., $\sum_j \lambda_j = \frac{1}{2}\sum_u (-1)^u h_u$). In other words, the low-pass coefficients can be made not only by the average values at adjacent pixels but also can be contributed from the composed difference values (i.e., t_n or

d_n). For the purpose of entropy reduction, these additional contributions may not be necessary. However, they are embedded in the low-pass process of wavelet transformation and dyadic decomposition for the purposes of perfect reconstruction using convolution operation. Since the compression relies on a good prediction in high-pass domain, usually the low-pass domain is further decomposed into a multi-level dyadic decomposition to obtain a greater number of high frequency elements.

From the above derivations, we find that the two-band, H+P, and S+P transforms are special cases of the H+PC transform. Figures 1, 2, and 3 summarize the forward and inverse processes in the three transformation systems and the relationships among them. These three figures show that the decomposition procedure can be computed through Haar transform. The decomposed coefficients in discrete Haar transform can then be converted into a doublet-type dyadic transform through a corresponding prediction method discussed above.

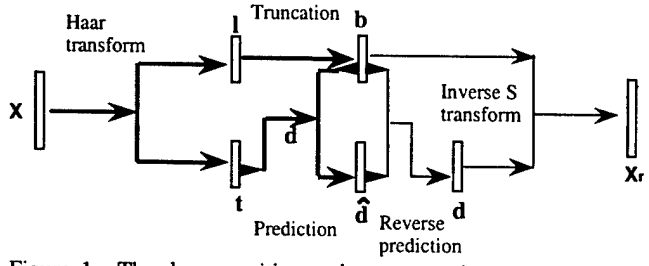


Figure 1. The decomposition and reconstruction processes of an S+P transform [6], which is a truncation version of an H+P transform. Bold and plain lines represent the forward and inverse transforms, respectively. The jointed plane line indicates a composition of two sources of data. ▶ stands for a convolution process with a segment of the data.

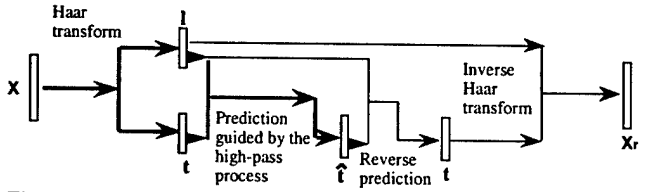


Figure 2. The decomposition and reconstruction processes of an H+P transform.

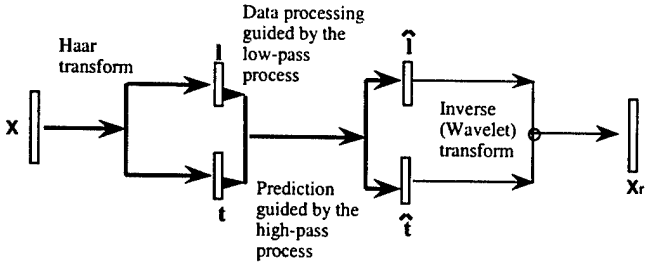


Figure 3. The decomposition and reconstruction processes of a dyadic PR transform [2] (e.g., orthogonal wavelet). The decomposition can be generated from an H+PC transform.

VI. Binomial Plus Prediction and Composite (B+PC Transforms)

In Section IV, Eq. (23) shows that the high-pass process of a biorthogonal wavelet transform can be computed through a binomial basis decomposition (i.e., B+P transform). Similar to Eq. (28), the low-pass coefficients can be converted from the decomposed coefficients of the binomial filtering to B+PC transform. By replacing α'_i with γ'_i and β'_j and λ'_j in Eq. (22), the solution for the weights in terms of the filter coefficients (k_u) are:

$$\gamma'_{\lfloor m/2 \rfloor} = (k_{m-1} + 2k_m) \quad and \quad \lambda'_{\lfloor m/2 \rfloor} = (k_{m-1} - 2k_m)/4,$$

$$\begin{cases} \gamma'_{-i} = \gamma'_i = (2k_{2i} + k_{2i+1} + k_{2i-1}) - \gamma'_{i+1} \\ \lambda'_{-j} = \lambda'_j = (k_{2j+1} - 2k_{2j} + k_{2j-1})/4 - \lambda'_{j+1} \end{cases}$$

for $i\&j=0,...,(m-2)/2$. ...(29)

For most binomial-based decomposition systems, the main contribution to the low-pass filtering comes from the central filtering component: $(...0,\ 0,\ \frac{1}{4},\ \frac{1}{2},\ \frac{1}{4},\ 0,\ 0...) \times \gamma'_0$. Again, the inverse transformation using the transformed coefficients should follow its corresponding dyadic inverse transform operation (e.g., inverse biorthogonal wavelet). In addition, the summation of weights contributed by difference values is 0 (i.e., $\sum_j \lambda'_j = \frac{1}{4}\sum_u (-1)^u k_u = 0$) in the low-pass process.

Similar to the doublet system, an additional composite term in the low-pass wing of the B+P is added to form B+PC. Eqs. (24) and (29) indicate that a unique set of $\alpha'_i, \beta'_j, \gamma'_i$ and λ'_j can be found in the B+PC system to match a binomial-based wavelet or a triplet two-band system.

The relationships between binomial, B+PC, B+P, S_b+P, triplet two-band and binomial-based wavelet (i.e., wavelets that satisfy $(-1)^n k_{2n+1} = (-1)^n k_{-2n+1}$ including WPS-biorthogonal transforms) are exactly the same as their counterparts developed through Haar transform. In other words, similar system processing diagrams, as those shown in Figures 1, 2, and 3, can be applied to their binomial versions by replacing H with B, S with S_b, doublet with triplet, and wavelet with binomial-based wavelet processes, respectively.

VII. Switching Transformation Kernels in Dyadic Decomposition - An Adaptive Approach

Since a data string can contain varieties of data patterns at different data segments, one may wish to use different decomposition kernels for the treatment. This can be performed by using an S+P and/or an S_b+P transform to another one (so as to H+P and B+P decompositions) because the starting and ending processes of a data segment using either method are systematically the same. However, the transition from the first to the second transform is the subject of the algorithm modification. In addition, it is necessary to record the starting and ending points of a specific S+P or S_b+P transform. This overhead can be reduced if the corresponding decomposed data possess a marker and applied transforms are pre-registered. The following algorithms show application of an S+P transform on a data segment: $(x_i,\ i = 0,..M-1)$, and an S_b+P

transform on the following data segment $(x_i,\ i = M,..N-1)$. For the first segment decomposition using an S+P transform:

$$b_n = \lfloor (x_{2n} + x_{2n+1})/2 \rfloor \quad \text{for n=0,1...,(M/2)-1,} \quad ...(30)$$

$$\begin{cases} d_n = x_{2n+1} - x_{2n} & for\ n=0,..m/2\ [build\text{-}up\ term] \\ \hat{d}_n = x_{2n+1} - x_{2n} + e_n & for\ n=(m/2)+1,...(M/2)-1\ , \end{cases} \quad ...(31)$$

where m is the total length of contribution weights in $\hat{d}_n$. For the second segment decomposition using an S_b+P transform:

$$b'_n = \lfloor (x_{2n-1} + 2x_{2n} + x_{2n+1})/4 \rfloor \quad \text{for n = (M/2),...,(N/2)-1,} \quad ...(32)$$

$$\begin{cases} d'_n = 2x_{2n} - x_{2n-1} - x_{2n+1} & for=(M/2),..(M+m'-1)/2\ [build\text{-}up\ term] \\ \hat{d}'_n = 2x_{2n} - x_{2n-1} - x_{2n+1} + e'_n & for=((M+m'-1)/2)+1,..(N/2)-1\ , \end{cases} \quad ...(33)$$

where m' is the length of high-pass kernel in $\hat{d}'_n$. The reconstruction for both transforms are almost independent except x_{M-1} is shared. Care must be paid on the overlapped convolution elements. In case that a process starts from an S+P to another S+P transform (or from an S_b+P to another S_b+P transform), the build-up coefficients can be ignored in the second transform. The above derivations are also applicable to the corresponding H+P and B+P transforms by turning all the truncation operations off and using real computation.

In digital implementation, however, a sequentially decomposed data sequence with multiple wavelets or sub-band using exactly the same length of the data space for perfect reconstruction is also solvable by treating each segment independently. This can be done by using mirrored data extension for each segment for each convolution based computation. This algorithm, however, does not seem as naturally performed as those derived in Eqs. (30)-(33). If the application does not require the same total length of the data sequence, additional data space (i.e., the length of each kernel) can be provided to accommodate the joint data between two decomposition processes.

VIII. Results of Predictive and Composite Terms Converting from Known Filters

VIII.A. Filters in the Doublet System

By replacing filter coefficients (h_u) of Daubechies wavelets [9] into Eq. (14), α_i and β_j values can be obtained. The second column of Table II shows nine sets of α_i and β_j values as H+P transforms corresponding to the high-pass processes of Daubechies wavelets. Using the same filter coefficients for Eq. (28), the right column on Table II shows nine sets of γ_i and λ_j associated with the low-pass components of H+PC transforms. Substituting the values of C, α_i and β_j into Eq. (8) and the values of γ_i and λ_j into Eq. (26), it would produce exactly the same high-pass and low-pass components, respectively, as Daubechies wavelets would. In S+P transforms, however, approximation is made by (i) downward truncation, (ii) set γ_i and λ_j to zero and C to unity, and (iii) use of approximated but sufficiently accurate rational values of α_i and β_j for the purposes of fast computation.

VIII.B. Filters in the Triplet System

There are several known biorthogonal filters and 2-band filters proposed in the literature. Typically, the reported filters are quadrapole. The analytic filters are composed of (k_u) for low-pass and $((-1)^{|u|}\tilde{k}_u)$ for high-pass. The synthetic filters are composed of $(\tilde{k}_u)$ for low-pass and $((-1)^{|u|}k_u)$ for high-pass. By replacing filter coefficients (k_u) into Eq. (23), the corresponding composite weights (i.e., γ'_i and λ'_j values) can be obtained. The predictive weights (α'_i and β'_j values) can be obtained by using the synthetic filter coefficients $(\tilde{k}_u)$ for Eq. (29). Tables III-VII shows five sets of original filters and their conversions in predictive and composite weights associated with the B+PC transforms. Table VIII shows the composite and predictive weights of a PR quadrature mirror filter proposed in [18].

IX. A Unified Perspective of Dyadic Decomposition - Summary

The split pair (i.e., (1,0) and (0,1)) of delta functions, called singlet basis system, has the lowest basis of dyadic decomposition. The corresponding decomposition forms are:

D+PC (delta+ prediction & composite) Transforms — D+P Transforms

first-path: $\hat{l}''_n = x_{2n} + a''(x_{2n\pm i}) \approx x_{2n}$ ...(34)

and

second-path: $\hat{t}''_n = x_{2n+1} + e''(x_{2n+1\pm i}) = x_{2n+1} + e''(x_{2n+1\pm i})$ ...(35)

where i = -p,..., 0, ..., q; a'' and e'' are the added process and prediction terms, respectively. The well-known DPCM is a special case of the D+PC transform by giving $a''(.) = -x_{2n-1}$ and $e''(.) = -x_{2n}$. In many applications, spline interpolation methods (called S_d+P transform in this paper) are used for the prediction of e''. Using high-order polynomial terms to model the data sequential can be accomplished by an iterative search using orthogonal least square (OLS) method [19] or other modeling techniques [20]. Although it takes substantially modeling effort for each image pattern, the polynomial prediction is one of interesting data decomposition schemes in the recent publications [21]. The non-linear polynomial terms can also be added onto H+PC and B+PC systems as shown in Eqs. (10) and (21), their applications in data decomposition should be of interested to the investigators in the field.

By comparing three generalized expressions (i.e., Eqs. (10), (20) and (35)), we find that both doublet and triplet systems can be formed by the generalized singlet decomposition system. Although the doublets and triplets seem to function independently, they share exactly the same decomposition principles. We can integrate major dyadic decomposition methods through a unified framework.

Table II. The Weights of Predictive and Composite Terms for the Daubechies Filter Coefficients

Names	C	Indices	α	β	Indices	γ	λ
D4	1.515749527851	0	-0.535898384862	1.000000000000	0	1.319479216883	-0.176776695297
		-1	0.535898384862	0.071796769725	1	0.094734345491	0.176776695297
D6	1.755060181656	0	-0.832286317816	1.000000000000	0	1.139562062261	-0.237110478180
		-1	1.044065157711	0.285080113551	1	0.324866482108	0.297444261064
		-2	-0.211778839890	-0.044065157715	2	-0.050214982000	-0.060333782882
D8	2.115899710319	0	-1.025087303111	1.000000000000	0	0.945224383862	-0.242234378622
		-1	1.394091283712	0.637834792253	1	0.602896998513	0.329432268674
		-2	-0.461004174828	-0.165244816522	2	-0.156193429883	-0.108938096777
		-3	0.092000194228	0.023577057747	3	0.022285609882	0.021740206726
D10	2.618035204426	0	-1.161692571582	1.000000000000	0	0.763931667771	-0.221863435912
		-1	1.533855467064	1.129337492784	1	0.862736674339	0.292940191269
		-2	-0.549918340455	-0.359377373963	2	-0.274539756651	-0.105025008740
		-3	0.219425342839	0.093372230314	3	0.071330003627	0.041906492027
		-4	-0.041669897860	-0.012101902702	4	-0.009245026714	-0.007958238642
D12	3.299433666451	0	-1.263957432420	1.000000000000	0	0.606164633748	-0.191541573524
		-1	1.438168880348	1.759232063963	1	1.066384259730	0.217941778156
		-2	-0.318388177157	-0.587351260219	2	-0.356031561532	-0.048248913199
		-3	0.230890210880	0.206254974567	3	0.125024471117	0.034989370028
		-4	-0.106030209385	-0.051187739089	4	-0.031028197117	-0.016067940760
		-5	0.019316727734	0.006103880398	5	0.003699956426	0.002927279298
D14	4.215928263960	-0	-1.343562649551	1.000000000000	0	0.474391373567	-0.159343632698
		-1	1.093400166579	2.527268506668	1	1.198914378251	0.129674901720
		-2	0.337823095148	-0.775608936892	2	-0.367942188923	0.040065090533
		-3	-0.039222424363	0.320245765170	3	0.151921828418	-0.004651694942
		-4	0.230208564525	0.045227203738	4	0.021455395304	0.027302239283
		-5	0.051103039635	0.027362589736	5	0.012980576529	0.006060710291
		-6	-0.009086819972	-0.003052177979	6	-0.001447926904	-0.001077677252
D16	5.445326519367	0	-1.407375942321	1.000000000000	0	0.367287433157	-0.129227874336
		-1	0.491582583521	3.433238673897	1	1.260985419951	0.045138026321
		-2	1.460362721375	-0.816376007316	2	-0.299844648218	0.134093218853
		-3	-0.698498943693	0.351822304627	3	0.129219911194	-0.064137471023
		-4	0.145493422954	-0.167328226846	4	-0.061457554933	0.013359476464
		-5	0.028505924229	0.061878299970	5	0.022727121964	0.002617466935
		-6	-0.024387508070	-0.014326908277	6	-0.005262093366	-0.002239306310
		-7	0.004317742010	0.001519171558	7	0.000557972622	0.000396463095
D18	7.094396788531	0	-1.459719865014	1.000000000000	0	0.281912621977	-0.102878363625
		-1	-0.372207203729	4.476958828200	1	1.262111201741	-0.026232477180
		-2	3.025555678304	-0.567822747103	2	-0.160076399454	0.213235583552
		-3	-1.740833956566	0.183390036787	3	0.051699966115	-0.122690766280
		-4	0.697794301300	-0.130916974640	4	-0.036907147582	0.049179255270
		-5	-0.156862185552	0.080211411183	5	0.022612609239	-0.011055357504
		-6	-0.003133602695	-0.031941487319	6	-0.009004708440	-0.000220850538
		-7	0.011473492089	0.007371194069	7	0.002078032647	0.000808630560
		-8	-0.002066672340	-0.000754190669	8	-0.000212615869	-0.000145655255
D20	9.308956243593	0	-1.503459195971	1.000000000000	0	0.214846857979	-0.080753371089
		-1	-1.501142274445	5.658263936561	1	1.215660228386	-0.080628925261
		-2	4.943230475727	0.145805806185	2	0.031325919334	0.265509383994
		-3	-3.009730910234	-0.319189839154	3	-0.068576934041	-0.161657807356
		-4	1.530871927260	0.100830971613	4	0.021663217438	0.082225755885
		-5	-0.583394250868	0.017478204114	5	0.003755137237	-0.031335105440
		-6	0.133487910271	-0.033170705790	6	-0.007126621916	0.007169864525
		-7	-0.005557945353	0.015768241034	7	0.003387757042	-0.000298526774
		-8	-0.005300425106	-0.003734397410	8	-0.000802323550	-0.000284694920
		-9	0.000994688719	0.000373868474	9	0.000080324467	0.000053426437

Remark1: The rational values for each set of α_i *and* β_j for the implementation of S+P transform can be adjusted based on the dynamic range of the applied data sequence if the approximation raises a concern. For data decomposition, however, any set of these weights would produce a PR result anyway.

Remark 2: The C values do not apply to γ_i *and* λ_j, which have been discussed in Section V.

Figure 4 shows the unified perspective of dyadic decomposition subsystems. This diagram also summarizes the dyadic decomposition systems discussed in this paper. The relationships among major decomposition methods, which are of interest to many investigators in current data compression research, are shown in Table IX.

In this paper, we attempt to use prediction as a key expansion to generalize decomposition of a data sequence. Data decomposition through a prediction could produce low entropy and result in high compression efficiency. As far as data compression is concerned, all dyadic decomposition methods can be unified by the following three statements:

(a) The singlet (i.e., (1,0) and (0,1)), Haar (i.e., (1,1) and (1,-1) as a pair of doublets), and binomial bases (i.e., (1,2,1) and (-1,2,-1) as a pair of triplets) are filter elements in the dyadic decomposition family.
(b) The low-pass coefficients are weighted average values of the neighbor pixel values and their composed values.
(c) The high-pass coefficients are the weighted difference values of the neighbor pixel values and their composed values.

Table III. A Spline Filter on Table I of [13].

u,i,j	0	± 1	± 2	± 3	± 4
k_u	0.994369	0.419845	-0.176777	-0.066291	0.033146
γ'_i	1.966641	-0.309359	0.033146		
λ'_j	0.005524	-0.011049	0.008286		
$\tilde{k}_u$	0.707107	0.353553			
α'_i	0.000000				
β'_j	1.000000				

$C_0' = 2.828427$

Table IV. A Spline Variant Filter Proposed in [14] and on Table II of [13].

u,i,j	0	± 1	± 2	± 3	± 4
k_u	0.852699	0.377403	-0.110624	-0.023849	0.037829
γ'_i	1.655203	-0.158323	0.037829		
λ'_j	0.012549	-0.015731	0.009457		
$\tilde{k}_u$	0.788485	0.418092	-0.040690	-0.064539	
α'_i	-0.403201	0.201599			
β'_j	1.000000	-0.096803			

$C_0' = 2.28949$

Table V. A Laplacian Pyramid Filter Proposed in [15] and on Table III of [13].

u,i,j	0	± 1	± 2	± 3
k_u	0.848528	0.353553	0.070711	
γ'_i	1.555635	-0.070711		
λ'_j	0.035355	-0.017678		
$\tilde{k}_u$	0.858630	0.368706	0.075761	015152
α'_i	0.223602	-0.111801		
β'_j	1.000000	-0.065217		

$C_0' = 2.459502$

Table VI. A 13-11-Tap Biorthogonal Filter [16].

u,i,j	0	± 1	± 2	± 3	± 4	± 5	± 6
k_u	0.767245	0.383269	-0.068878	-0.033474	0.047281	0.003759	-0.008473
γ'_i	1.608249	-0.143345	0.054799	-0.008473			
λ'_j	-0.018440	-0.001397	0.009941	-0.002118			
$\tilde{k}_u$	0.832847	0.448109	-0.069164	-0.108737	0.006292	0.014182	
α'_i	-0.626277	0.357840	-0.044704				
β'_j	1.000000	-0.159502	0.017548				

$C_0' = 2.025415$

Table VII. A 5-3-Tap Biorthogonal Filter [17, 16].

u, i, j	0	± 1	± 2
k_u	1.060660	0.353553	-0.176777
γ'_i	1.767767	-0.176777	
λ'_j	-0.044194	-0.088388	
$\tilde{k}_u$	0.707107	0.353553	
α'_i	0		
β'_j	1.000000		

$C_0' = 2.828427$

Table VIII. A 9-tap Quadrature Mirror Filter (QMF) [18]

u, i, j	0	± 1	± 2	± 3	± 4
k_u	0.798430	0.413948	-0.073882	-0.060394	0.028220
γ'_i	1.747113	-0.194670	0.028220		
λ'_j	-0.037563	0.011727	0.007055		
$\tilde{k}_u$	0.798430	0.413948	-0.073882	-0.060394	0.028220
α'_i	-0.344004	0.107392	0.064610		
β'_j	1.000000	-0.111424	0.016153		

$C_0' = 2.289491$

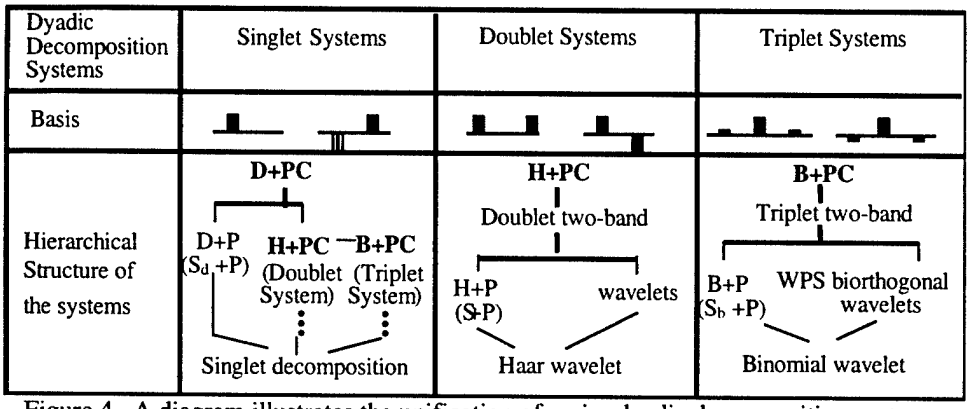


Figure 4. A diagram illustrates the unification of major dyadic decomposition systems.

Table IX. Relationships among Dyadic Transforms in Data Decomposition

Dyadic Decomposition Methods	Remarks
H+PC transform	Generalized form based on doublets. Not all of them are PR.
B+PC transform	Generalized form based on triplets. Not all of them are PR.
H+P transform	Special cases of H+PC. An adaptive transform can be implemented and co-exist with B+P. They are PR.
B+P transform	Special cases of B+PC. An adaptive transform can be implemented and co-exist with H+P. Causal and some non-causal cases are PR.
Discrete orthogonal wavelet transform	Special cases of H+PC transform. High-pass can be exactly described by H+P Predictive and composite terms of Daubechies wavelets are shown in Table II.
Discrete biorthogonal wavelet transform	Special cases of B+PC transform. High-pass can be exactly described by a B+P transform.
Two-band decomposition	Special cases of H+PC or B+PC transform.
S+P transform	Special cases of H+P transform. An adaptive transform can be implemented and co-exist with S_b+P. Rational computation is always PR.
S_b+P transform	Special cases of B+P transform. An adaptive transform can be implemented and co-exist with S+P. Only causal filtering cases are PR.

X. Conclusions and Discussion

Although our study aims at decomposition methods for data compression, base transformations and their extensions including wavelet transforms as well as their integer versions are members of the same family. We also showed that a transform in the dyadic family can be computationally constructed by its base transform. As examples, we showed analytical and synthetic wavelets can be computed through either Haar or binomial bases. This approach, which is extensively described in the paper, can be deemed as a part of element theory for the dyadic transformations as a whole.

The three transforms (i.e., S+P, S_b+P, and S_d+P) based on the sequential operation are appropriate for digital data processing. Since they use rational numbers in the computer

implementation, the computational speed in forward and inverse transforms can be greatly increased over convolution-based transforms using real number for implementation. This is particularly noticeable in data compression. In a lossy compression scheme, a quantization process would be added prior to the coding. The advantage of data accuracy using real number implementation would be diminished. Except the triplet system, S family transforms would perform much more effectively than real number computation methods. They are perfectly reconstructable and can provide entropy as low as their counterparts using real number computation. Additionally, their adaptive implementations are readily available.

As a part of results of this paper, Table II shows all Daubechies wavelets can be reformulated by the linear prediction terms (i.e., $\alpha_i, \beta_j, \gamma_i$ and λ_j). When approximating the high-pass process of the decomposition using S+P system, γ_i and λ_j set to zero and C sets to unity. In addition, α_i and β_j can be approximated with rational numbers for fast computation.

Having drawn the unified perspective, one shall be able to explore an optimal dyadic decomposition (or wavelet) methods for a defined data pattern. Specifically, we can search for a set of solutions for the predictive contribution weights to minimize the entropy of $\hat{t}_n$ or $\hat{t}'_n$. Using H+GP as example, a minimum entropy can be searched in a 2r-space for a given image pattern. It is observed that the number polynomial order (r) and the kernel length are in single digits, a full search using Eqs. (1) and (9) through the hierarchical decomposition tree is practically possible. Since the low-resolution data sequence possesses fewer structures, the use of high-order polynomial terms for modeling optimal decomposition in high-level tree may not be necessary. This modeling method shares the same spirit of [5] and can be a highly effective method in searching for appropriate decomposition kernels. In case of linear system, the contribution weights can be associated with a compactly supported wavelet, massive data patterns in detail structures can be documented if we find that the result obtained from a kernel is significantly different from others [22]. These applications can be of greatly useful in various fields of digital signal processing.

Acknowledgments

This work was supported by an NIH grant No. RO1CA79139 and US Army Research Grant No. DAMD17-96-1-6254 (through a subgrant from University of Michigan, Ann Arbor). The content of this paper does not necessarily reflect the position or the policy of the government, and no official endorsement should be inferred. The authors are grateful to Ms. Lisa Kinnard for her editorial assistance.

References

1. A. N. Akansu and M. J. T. Smith, *Subband and Wavelet Transforms: Design and Applications*, Kluwer Academic Publishers, Norwell MA, 1996.
2. A. N. Akansu and R. A. Haddad, *Multiresolution Signal Decomposition*, Academic Press, Inc., San Diego, CA, 1992.
3. J. M. Shapiro, "Embedded image coding using zerotrees of wavelet coefficients," IEEE Trans. Image Processing, vol. 41, No. 12, pp. 3445-3462, Dec. 1993.
4. Z. Xiong, K. Ramchandran, M. T. Orchard, and K. Asai, "Wavelet packets-based image coding using joint space-frequency quantization," In Proc. IEEE Int. Conf. on Image Proc. Austin, TX, pp. 324-328, 1994.
5. S-C. B. Lo, H. Li, J. Wang, M. T. Freedman, and S. K. Mun, "On optimization of orthonormal wavelet decomposition: Implications of Data Accuracy, Feature Preservation, and Compression Effects," SPIE Vol. 2707, Med. Imag., pp.201-214, 1996.
6. A. Said, and W. A. Pearlman, "Reversible image compression via multiresolution representation and predictive coding," Proc. SPIE Conf. Visual Communications and Image Processing '93, Proc. SPIE 2094, pp. 664-674. Cambridge, MA, Nov. 1993.
7. A. Said, and W. A. Pearlman, " A new, fast, and efficient image codec based on set partitioning in hierarchical trees," IEEE Trans. Circuits Syst. Video Technol., vol. 6, pp. 243-250, 1996.
8. V. K. Heer and H-E. Reinfelder, "A comparison of reversible methods for data compression," Proc. SPIE vol. 1233, Med. Imag., IV, pp. 354-365, 1990.
9. I. Daubechies, "Orthonormal bases of compactly supported wavelets", Comm. on Pure Appl. Math., vol. XLI, pp. 909-996, 1988.
10. M. Vetterli and C. Herley, "Wavelets and filter banks: Theory and Design," IEEE Trans. Signal Processing, Vol. 40, No. 9, pp. 2207-2232, 1992.
11. A. Cohen, I. Daubechies, and J. C. Feauveau, "Biorthogonal bases of compactly supported wavelets," Comm. on Pure Appl. Math., vol. 45, pp. 485-560, 1992.
12. R. R. Coifman and M. V. Wickerhauser, "Entropy-based algorithms for best basis selection," IEEE Trans. Inform. Theory, vol. 38, pp. 713-718, 1992.
13. M. Antonini, M. Barlaud, P. Mathieu, I. Daubechies, "Image coding using wavelet transform," IEEE Trans. on Image Proc., vol. 1, pp. 205-220, 1992.
14. A. I. Cohen, I. Daubechies, and J. C. Feauveau "Biorthogonal bases of compactly supported wavelets," Commun. Pure Apply Math., vol. 45, pp. 485-560, 1992.
15. P. Burt and E. Adelson, "The Laplacian pyramid as a compact image code," IEEE Trans. On Commun., vol. 31, pp. 482-540, 1983.
16. J. D. Villasenor, B. Belzer, and J. Liao, "Wavelet filter evaluation for image compression," IEEE Trans. on Image Proc., vol. 4-8, pp. 1053-1060. 1995.
17. E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J. Heeger, "Shiftable multiscale transforms," IEEE Tran. on Information Theory, vol. 38, pp. 587-607, 1992.
18. T. Senoo and B. Girod, "Vector quantization for entropy coding of image subbands," IEEE Tran. on Imag. Proc. vol. 1, No. 4, pp. 526-532, Oct. 1992.
19. M. J. Korenberg and L. D. Paarman, "Orthogonal approaches to time-series analysis and system

identification," IEEE Signal Processing Mag., pp. 29-43, July 1991.

20. R. Gruter, J. M. Vesin, E. Pruvot, C. Seydoux, and M. Fromer, "Linearity assessment of heart rate time series," In Proc., 18th Annu. Int. Conf., IEEE Engineering in Medicine and Biology (EMBS96), Amsterdam, The Netherlands, Nov. 1996.
21. R. Gruter, O. Egger, J. M. Vesin, M. Kunt, "Rank-order polynomial subband decomposition for medical image compression," IEEE Trans. On Med. Imag., vol. 19 No. 10, pp. 1044-1052, Nov., 2000.
22. S-C. B. Lo, H. Li, M. T. Freedmam, "On optimization of wavelet decomposition for image compression and feature preservation," Submitted to IEEE Trans. on Med. Imag.

Nomenclature:

Symbols of the three main decomposition systems

singlet	doublet	triplet	
D	H	B	abbreviation of the three bases, namely: delta, Haar, and binomial systems.
l"	l	l'	low-pass components of the three bases.
t"	t	t'	high-pass components of the three bases.
S_d	S	S_b	integer versions of the three basis systems using rational number computation.
b"	b	b'	integer versions of l", l, and l', respectively.
d"	d	d'	integer versions of t", t, and t', respectively.
D+P	H+P	B+P	three extended systems with prediction in the high-pass process.
Ge"	Ge	Ge'	added general prediction with polynomial terms in the high-pass process.
e"	e	e'	added linear prediction terms in the high-pass process.
$\tilde{l}''$	$\tilde{l}$	$\tilde{l}'$	scaled low-pass components of the three bases.
$\tilde{t}''$	$\tilde{t}$	$\tilde{t}'$	scaled high-pass components of the three bases.
$\hat{t}''$	$\hat{t}$	$\hat{t}'$	high-pass components of the three generalized prediction systems.
	${}_r\alpha$	${}_r\alpha'$	weights of the low-pass component for the prediction in r^{th} polynomial term.
	${}_r\beta$	${}_r\beta'$	weights of the high-pass component for the prediction in r^{th} polynomial term.
	C	C'	scaling factors between H+P and Haar and between B+P and binomial.
S_d+P	S+P	S_b+P	three sequential type (S) transforms with prediction.
$\hat{d}''$	$\hat{d}$	$\hat{d}'$	high-pass components of the three sequential type transforms.
D+PC	H+PC	B+PC	dyadic systems with generalized high-pass and low-pass processes.
a''	a	a'	added composite terms in the low-pass process.
$\hat{l}''$	$\hat{l}$	$\hat{l}'$	low-pass components of the three generalized systems.
	γ	γ'	weights of the low-pass components in the composite term.
	λ	λ'	weights of the high-pass components in the composite term.

Filter coefficients

h_u : orthogonal wavelet coefficients.

k_u, $\tilde{k}_u$: analytic and synthetic filter coefficients of biorthogonal wavelets

K_u : altered coefficients in biorthogonal wavelet for the approximation of causal filter in B+P.

A Multiple Circular Paths Convolution Neural Network System for Detection of Mammographic Masses

Shih-Chung B. Lo[1], Huai Li[2], Yue Wang[4], Matthew T. Freedman[1], and Seong K. Mun[1]

[1]ISIS Center, Radiology Department, Georgetown University Medical Center, Washington, DC.
[2]Odyssey Technologies Inc., Jessup, Maryland.
[4]Department of Electrical Engineering and Computer Sciences, Catholic University of America, Washington, DC.

(Submitted to IEEE Medical Imaging for Review)

Abstract A multiple circular path convolution neural network (MCPCNN) architecture specifically designed for the analysis of tumor and tumor-like structures has been constructed. We first divided each suspected tumor area into sectors and computed the defined mass features for each sector independently. These sector features were used on the input layer and are coordinated by convolution kernels of different sizes that propagated signals to the second layer in the neural network system. The convolution kernels were trained, as required, by presenting the training cases to the neural network.

In this study, randomly selected mammograms were processed by a dual morphological enhancement technique. Radiodense areas were isolated and were delineated using a region growing algorithm. The boundary of each region of interest was then divided into 36 sectors using 36 equi-angular dividers radiated from the center of the region. A total of 144 BI-RAD based features (i.e., 4 features per sector for 36 sectors) were computed as input values for the evaluation of this newly invented neural network system. The overall performance was 0.78-0.80 for the areas (Az) under the ROC curves using the conventional feed-forward neural network in the detection of mammographic masses. The performance was markedly improved with Az values ranging from 0.84 to 0.89 using the MCPCNN. This paper does not intend to claim a highest mass detection system. Instead it reports a potentially better neural network structure for analyzing a set of the mass features defined by an investigator.

Key words: Mammography masses, computer-aided diagnosis, neural network, convolution neural network, sector features, and BI-RAD.

1. Introduction

It is known that effective treatment of breast cancer calls for early detection of cancerous lesions (e.g., clustered microcalcifications and masses associated with malignant cellular processes) [1-3]. Breast masses appear as areas of increased density on mammograms. It is particularly difficult for radiologists to detect and analyze a suspected area where a mass is overlapped with dense breast tissue. These masses are more readily seen as time progresses, but the further the tumor has progressed, the lower the possibility of a successful treatment. Therefore, increasing the chances of early breast cancer detection in improving today's clinical system is of vital importance in breast cancer diagnosis.

Several research groups have developed computer algorithms for automated detection of mammographic masses [4-8]. Investigators also attempted to classify the malignant or benign nature of the detected tumors [9-12]. The results of these detection programs indicate that a high true-positive (TP) rate can be obtained at the expense of 2 or 3 false-positive (FP) detections per mammogram. Mammographically, a multiplicity (more than two) of similar benign-appearing breast lesions argues strongly for benignity [13-16] and, indeed, the more masses that are identified, the less chance that they represent cancer [17]. If the computer indicates multiple suspicious locations on a mammogram, the radiologist has to seek out one mass that possesses mammographic features, which are different from the others. The significant lesion may be missed due to the multiplicity of possible lesions. We therefore believe that a more useful and fundamental approach to computer-aided diagnosis (CADx) of masses is to devise computer programs to analyze features of a suspected area [18,19] and to provide feature measures and estimates of the likelihood of malignancy by making comparisons within a digital mammographic database. The computer therefore serves as a second opinion and also provides a reproducible and an objective evaluation of the mass. With this aid, the radiologist may also increase his/her sensitivity by lowering the threshold of suspicion, while maintaining the overall specificity and reading efficiency.

2. Clinical Background of Breast Lesions and Technical Approach in Mass Detection

2.1 Description of clinical background

Most commonly, breast cancer presents itself as a mass. The same lesion shows a somewhat different picture from one projection to the other. Difficulties in masses also vary with the underlying breast parenchyma. In the fatty breast, masses are generally easy to detect. In the dense breast, mass detection is more difficult and auxiliary signs aid this detection. When the breast contains one mass, the decision process is based on its size, shape, and margins. When there are several masses, one looks at each, trying to determine whether any has features to suggest cancer. Furthermore, one looks to see if any mass is different in appearance from the others. Multiple small, well-defined, similar masses that present themselves bilaterally are all likely to be benign. Large, poorly-defined, spiculated and unusually radiodense masses are extremely likely to be malignant. In this study, we used several computational features (see section 3.2) highly associated with four major features of breast masses routinely used in clinical reading:

Density - Malignant lesions tend to have greater radiographic density due to high attenuation and less compressibility of cancer than normal tissue. Radiolucent lesions are typically benign and the diagnosis can be made from the mammogram.

Size - If the lesion has morphological features suggesting malignancy, it should be considered suspicious regardless of the size. Isolated masses with non-cystic densities greater than 8 mm in diameter can be malignant. In general, the larger a lesion, the more suspicious it is.

Shape - The more irregular the shape of a lesion, the more likely the possibility of malignancy. Lesions tend to be round, ovoid and/or lobulated. Small and frequent lobulations are suspicious. Lesions in the lateral aspect of the breast near the edge of the parenchyma with a reniform shape and a hilar indentation or notch usually represent a benign intramammary lymph node. Breast carcinoma hidden in the dense tissues can cause parenchymal retraction, which possess different shapes.

Margins -The margins of the lesion should be carefully evaluated for areas of spiculation, stellate patterns or ill-defined regions. Most breast cancers have ill-defined margins secondary to tumor infiltration and associated fibrosis. The appearance of spiculations and a more diffuse stellate pattern are almost pathognomonic for cancer. Lesions with sharply defined margins have a high likelihood of being benign; however, up to 7% of malignant lesions can be well circumscribed.

These are known clinical features and have been adapted in "Breast Imaging - Reporting and Data System" (BI-RAD) [20] of the American College of Radiology (ACR). Figures 1(A) and 1(B) show two breast images containing masses. In Figure 1(A), a malignant mass is superimposed on the dense glandular tissue. However, its spiculated nature makes it easily identifiable. In Figure 1(B), another malignant mass is located on the fatty background but is associated with a large body of glandular tissue. This mass is not easily detectable by the computer because its density is lower than the neighboring glandular tissue. Furthermore, one end of the mass is fully connected with this tissue.

2.2. Technical approach for detection of mammographic masses

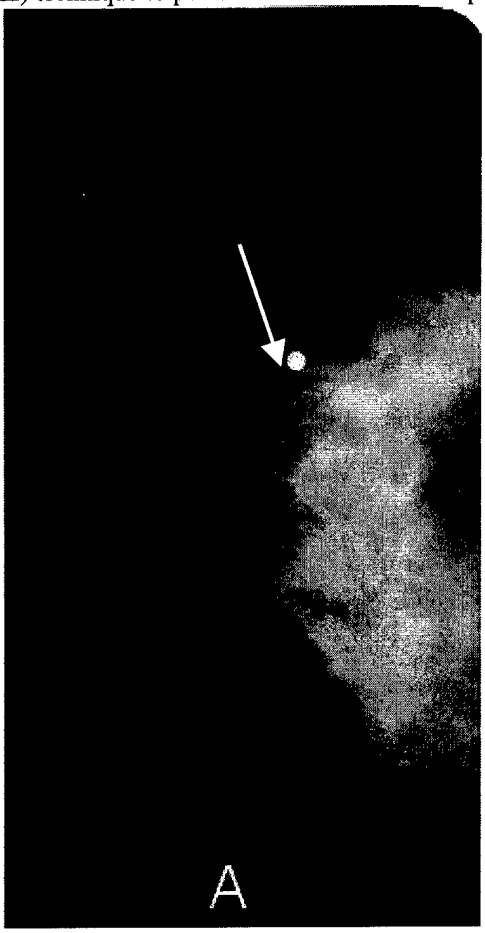
In this study, our goal was to detect clinically suspicious lesions. The differentiation of benign and malignant status of the mammographic masses can be extended from this study model and will be reported in our future work. The study was conducted with the following steps: (1) use background correction method and morphological operations to extract radio-opaque areas, (2) delineate the boundary of the areas, (3) compute the features and texture of the masses with emphasis on the boundary, and (4) design training strategy using neural networks as classifiers for the recognition of mass features. The overall detection scheme of the study framework is shown in Figure 2.

3. Development of Technical Methods

3.1. Preprocessing and Extraction of Suspicious Masses

In automatic mass detection, accurate selection of suspected masses is considered a critical first step due to the variability of normal breast tissue and the lower contrast and ill-defined margins of masses. In our previous study [18], we aimed to improve the task of lesion site selection using model-based image processing techniques for unsupervised lesion site selection. We focused on two essential issues in the stochastic model-based image segmentation: enhancement and model selection. Based on the differential geometric characteristics of masses against the background tissues, we proposed one type of morphological operation to enhance the mass patterns on mammograms by removing high intensity background caused by breast tissues while keeping mass-signals [18]. Then we employed a finite generalized Gaussian mixture (FGGM) distribution to model the histogram of the mammograms where the statistical properties of the pixel images are largely unknown and are to be incorporated. We incorporate the EM algorithm with two information theoretic criteria to determine the optimal number of image regions and the kernel shape in the FGGM model. Finally, we applied a contextual Bayesian relaxation labeling

(CBRL) technique to perform the selection of suspected masses.



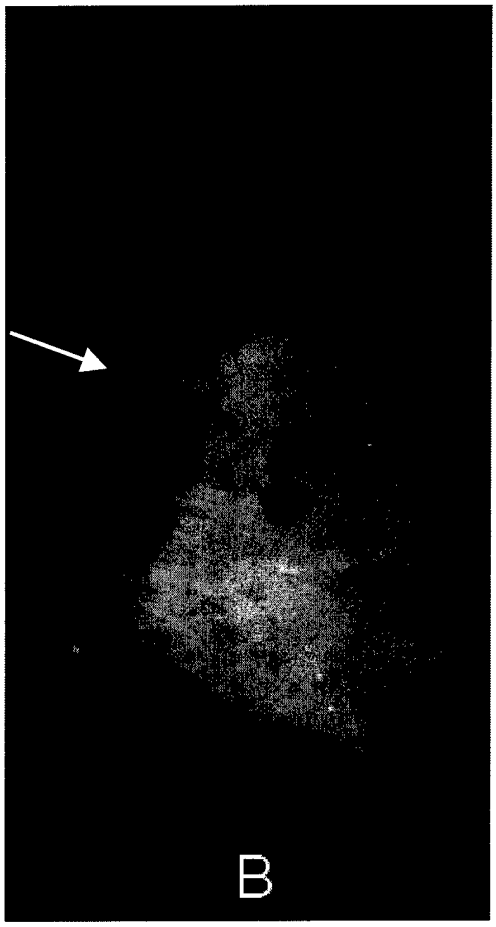


Figure 1. (A) Dense breast containing a malignant mass. (B) Fatty and glandular breast containing a malignant mass

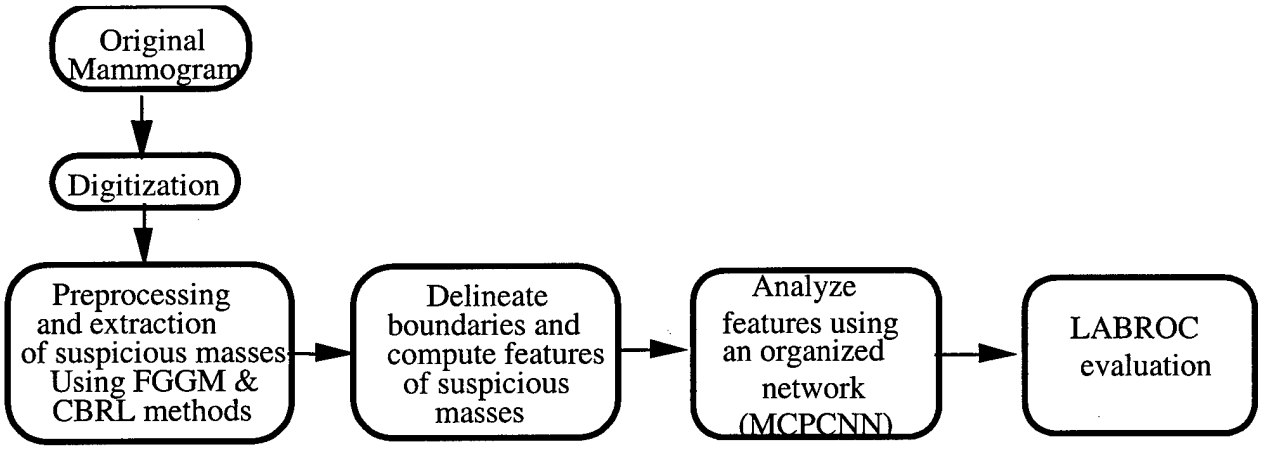


Figure 2. A system flow chart for the detection of masses in this study.

We consistently processed the mammograms using this very effective pre-screening segmentation method. In that study [18], the FGGM method isolated 1,142 potential masses including 114 of the 186 true masses in 200

mammograms. The mammograms were collected from the Mammographic Image Analysis Society (MIAS) database and Brook Army Medical Center (BAMC) database. After morphological enhancement, 3,143 potential masses were extracted using the FGGM technique. Of them, 181 were masses; however, 5 masses were not extracted. The results demonstrated that more true masses were picked up after enhancement although more false cases were also included. The undetected areas mainly occurred at the lower intensity side of the shaded objects or more obscured by fibroglandular tissues that, however, were extracted on morphological enhanced mammograms. Additionally, when the margins of masses are ill defined, only parts of suspicious masses were extracted from the original mammograms. We therefore decided to use the proposed morphological operation as a preprocessing step for the image enhancement prior to a segmentation method for the extraction of potential masses on the mammograms.

Based on the CBRL segmented region of interest, we employed a region growing method using a 4-neighbors connection method assisted with a template masking operation to fill unconnected holes in the ROI:

$$IF \quad f(x-a, y-b) > V \; and \; f(x,y) \in S, \quad then \quad f(x-a, y-b) \in S \qquad ...(1)$$

$$IF \quad f(x-d, y-d) \in S, \quad then \quad f(x-t, y-s) \in S \qquad for \; t \le d \; and \; s \le d \qquad ...(2)$$

where V denotes the threshold value of the originally CBRL segmented ROI, S represents the set of growing region, and [*a, b*] is a set of four conditions (i.e., [1, 0], [-1, 0], [0, 1], and [0, -1]) for the four neighboring pixels. In eq. (2), *d* is the size of template. In practice, we found that *d* should be set at 5 pixels to fill the holes without disrupting the boundary.

3.2. Feature extraction of the masses

Feature extraction methods play an essential role in many pattern recognition tasks. Once the features associated with an image pattern are extracted accurately, they can be used to distinguish one class of patterns from the others. Recently, many investigators have found that the multilayer perceptron (MLP) neural network using the error backpropagation training technique is a very powerful tool to serve as a classifier [22, 23]. In fact, the use of MLP neural network system for classification of disease patterns has been widely applied in the field of computer-aided diagnosis [24-28].

The success of using a classifier for a pattern recognition task would rely on two factors: (a) selected features that could describe a discrepancy between image patterns and (b) accuracy of the feature computation. Should either one fail, no analyzer or classifier would be able to achieve an expected performance. By analyzing many clinical samples of various sizes of masses, we found that the peripheral portion of the mass plays an important role for mammographers to make a diagnosis. The mammographer usually evaluates the surrounding background of a radiodense area when a region is suspected.

We used the CBRL segmented ROI to compute the center. Since the segmented ROIs were somewhat smaller than the mammographer's delineation and on the denser region of the suspected patch, the computed centers were quite close to the visual center. We then divided the boundary of the ROI into 36 sectors (i.e., 10^{O} per sector) using 36 equi-angular dividers radiated from the center of the ROI. The following features were computed within each 10^{O} sector of the region:

(a) "*l*" - the length from the center of the ROI to the boundary segment of the sector.

(b) "*a*" - the cos(θ) (where θ is the normal angle of the boundary)

(c) "*g*" - the average gradient of gray value on the segment along the radial direction (i.e., $g = \sum_{i=1}^{N} \frac{g_i}{N}$) where N is the number of pixels of *i* along the radial direction from *l*/3 inside the boundary to the boundary (see the left *l*/3 line segment, Figure 3). Technically speaking, this set of gradient values may also serve as a fuzzy system on the input layer in the neural network (to be described in Section 3.3).

(d) "*c*" - the gray value difference (i.e., contrast) along the radial direction. Specifically, $c = \sum_{i=1}^{P} \frac{h_i}{P} - \sum_{o=1}^{P} \frac{b_o}{P}$ where h_i (or b_o) represents a pixel value along the radial direction. The position *l*/3 inside the boundary is the center of pixels h_i (*i=1,2,3,... P*) and position *l*/3 outside the boundary is the center of pixels b_O (*o=1,2,3,... P*), and *P* is the number of pixels equivalent to a segment of *l*/6 and was used for averaging (see Figure 3).

Hence, a total of 144 computed features (4 features/sector for 36 sectors) were used as input values for the

classification of the ROI. The relationship between the computed features and BI-RADS descriptors are discussed below:

(1) ROI Size -
The size of ROI is provided by the 36 "*l*" values.

(2) ROI Shape (round, oval, lobulated, or irregular) -
The 36 "*l*" and 36 "*a*" values can describe the shape of the ROI.

(3) ROI Margin (circumscribed, microlobulated, obscured, ill-defined, or spiculate) -
The 36 "*g*" and 36 "*l*" values can describe the ROI margin.

(4) ROI Density (fat-containing, low density, isodense, or highly dense) –

The 36 "*c*" and 36 "*g*" values can be used to describe the density distribution of the ROI.

In short, the selected features are greatly associated with the main mass descriptors indicated in the BI-RADS. The reason for using 36 values for each nominated feature is four-fold: (a) mass boundary varies, it is difficult to describe an image pattern using a single value; (b) due to the general shape of the masses, the features of masses can be easily analyzed by the polar coordinate system; (c) in case some features are inaccurately computed in several directions due to the structure noises, such as the breast slender lines, there may still exist a sufficient number of correct features; (d) generally more accurate results can be produced by using subdivided parameters rather than using global parameters in a pattern recognition task when the parameters are barely discernable and sample sizes are sufficiently large. Other computational features (e.g., difference entropy [19] and other higher order features) are eligible but require further investigation.

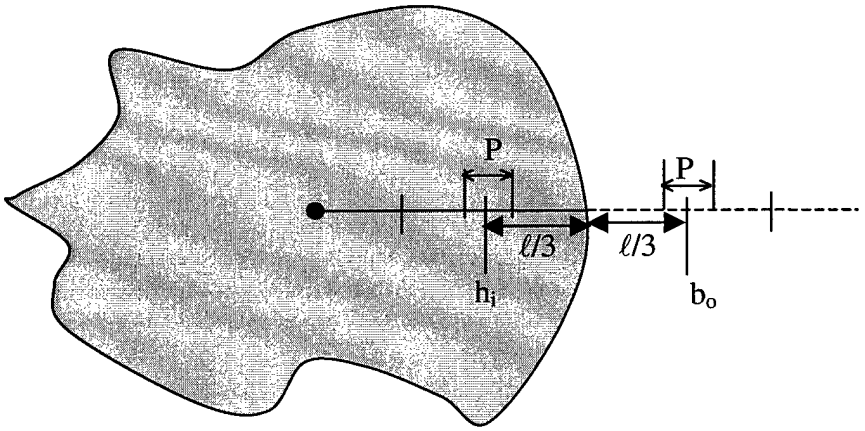


Figure 3. A suspicious mass is delineated and shown as the shaded region. Contrast is computed by subtracting the averaged background pixel value (i.e., b_o, o=1,2,..P) from the averaged foreground value (i.e., h_i, i=1,2,..P).

3.3. The neural network structure specifically designed for the extracted boundary features

(A) Multiple paths with circular networking to instruct the neural network in analyzing sector features

This paper focuses on neural network design and arrangement of features for effective pattern recognition of ROIs. We designed several neural network connections between the input and the first hidden layers as shown in Figure 4. In this neural network system, the first layer also functions as a correlation layer that transforms and encodes the signals from input nodes into correlation features for further neural network process. Figure 4(A), (B), and (C) illustrate the full connection, a self correlation (SC) network, and a neighborhood correlation (NC) network, respectively. Network connections with multiple sectors (i.e., 20^o, 30^o, 40^o, and 50^o of the neighborhood correlation) are grouped separately as independent NC paths. In the following study, we used four SC paths for a single sector and thirteen NC paths for four types of multi-sectors. The method of using the multiple correlation connections was motivated by our research experience in two-dimensional convolution neural network (2-D CNN) where we found that more than 10 multiple convolution kernels in the CNN were necessary in the detection of lung nodules and microcalcifications [25].

Compared to 2-D CNN systems, the computation required in the 1-D CNN (e.g., 144 input features) is relatively small. The combination of the networking paths described earlier for MCPCNN was implemented using C

programming language. The internal computation algorithm used in the MCPCNN shares the same convolution process as that in the 2-D CNN [25]. Rotation invariance and flip invariance for training the 1-D convolution kernels in the MCPCNN were employed.

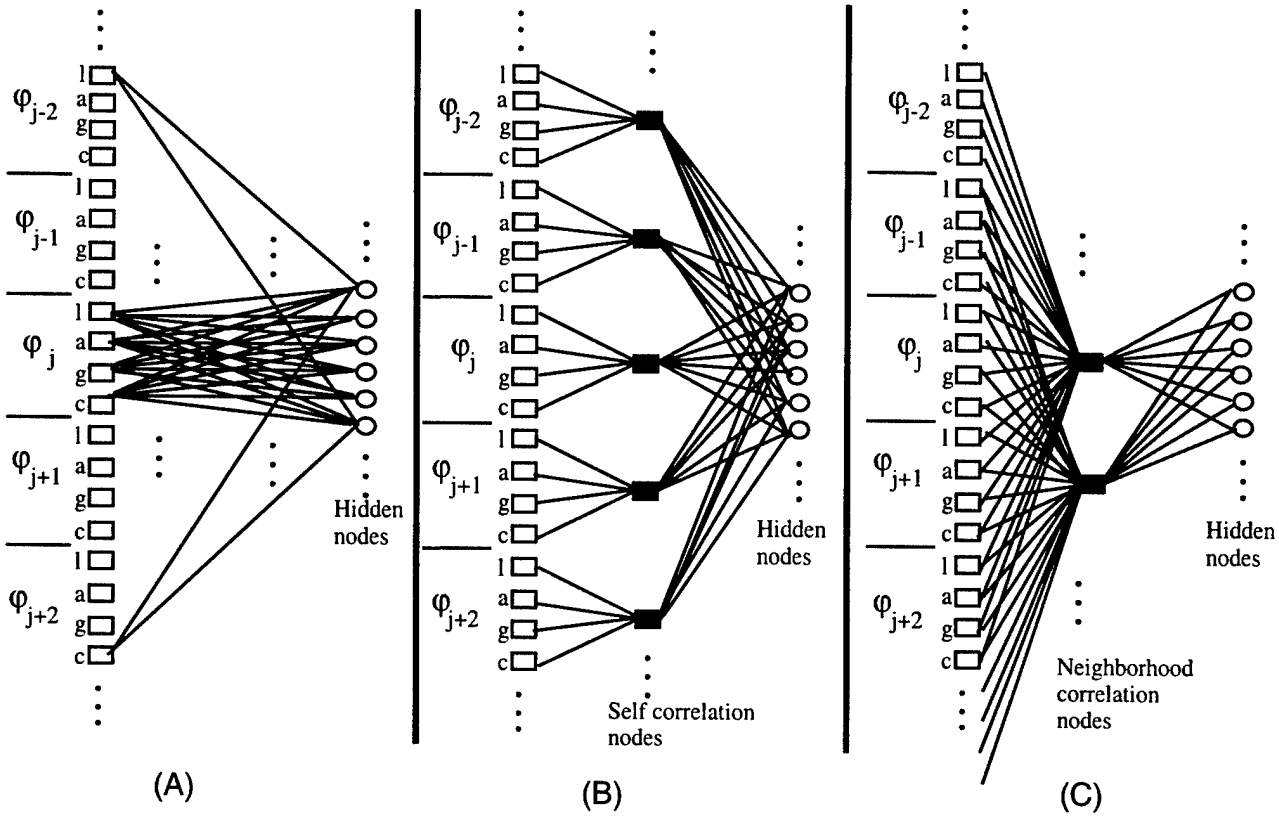


Figure 4. Three types of network paths connecting the input and the hidden layers in the MCPCNN:
(A) Full connection.
(B) A self correlation (SC) path; each node on the layer connects to a single set of the features (l,a,g,c) for the fan-in and fully connects to the hidden nodes for fan-out.
(C) A neighborhood correlation (NC) path; each node on the layer connects to the input nodes of adjacent sectors for the fan-in and fully connects to the hidden nodes for fan-out.
The fan-in nets emphasizing self correlation in (B) and neighborhood correlation in (C) represent convolution weights (i.e., the same type of sectors possess the same set of weighting factors).

The fully connected neural network is a conventional feed-forward MLP neural network. The signals of the fully connected neural network join the other network processes (i.e., SC paths and NC paths) at the single node of the output layer. The signal received at the output node is scaled between 0 and 1. During the training, 0 and 1 were assigned at the output node to perform backpropagation computation for a non-mass and a mass, respectively. The backpropagation is computed in such a way that the computed incremental errors (see equations (9) and (10) are retraced into every independent network path. Excluding the output layer, the SC and NC signals are independently arranged and are processed through two types of one-dimensional convolution processes in the forward propagation. The learning algorithms for all three types of circular network paths are based on the backpropagation training method.

Let $V^0(n, s)$ represents an input signal at the node n and sector s. The signal processed through an SC path and to be received at each node, n, on the first hidden layer is

$$N^1_{SC}(n) = \sum_i V^0(n', s') \otimes W_{i,sc}(n'; n), \tag{3}$$

where $\otimes$ stands for convolution operation, $W_{i,sc}(n';n)$ is the weight factor connected to node n from node n' through a self correlation path, i, regardless of the sector. The signal processed through an NC path and to be received at each node, n, on the first hidden layer is

$$N_{NC}^{1}(n) = \sum_{j} V^{0}(n', s') \otimes W_{j,nc}(n'; n). \quad (4)$$

where $W_{j,nc}(n',s';n)$ is the weight factor connected from node n' sector s' through a neighborhood correlation path, j, and $(s1, s2)$ is a range of sectors for the neighborhood correlation. Nevertheless, the signals processed through an NC path and to be received at each node, n, on the first hidden layer is

$$V^{1}(n) = S\left(N^{1}(n)\right) \text{ and } N^{1}(n) = N_{F}^{1}(n) + N_{SC}^{1}(n) + N_{NC}^{1}(n) + b^{0}(n), \quad (5)$$

where $N_{F}^{1}(n)$ is the processed signal contribute by full connection, $b^{0}(n)$ represents the bias at node n and $S(z)$ is a sigmoid function given by

$$S(z) = \frac{1}{1 + \exp(-z)}, \quad (6)$$

The sigmoid function would produce modulated values ranging from 0 to 1. The signals on other hidden layers in each path are processed the same as a conventional fully connected neural network. Other than the first hidden layer, the receiving signals at a hidden layer, l, collected from the previous hidden layer, $l-1$, are given by,

$$V^{l}(n) = S\left(N^{l}(n)\right) = S\left(\sum_{n'} V^{l-1}(n') \cdot W^{l-1}(n'; n)\right), \quad (7)$$

where n' and n denote the nodes at layers $l-1$ and l, respectively.

Let the t-th change of the weight be $\Delta W^{l}_{p}(n',s';n)$ and the t-th change of the bias be $\Delta b^{l}(t)$. The error function is defined as

$$E = \frac{1}{2}(T - O)^{2} \quad (8)$$

where T and O denote the target output value and the actual output value, respectively when the input values $V^{0}(n',s')$, are entered in the network. In this model, the error backpropagation algorithm, which updates the kernel weights, is given below:

$$\Delta W_{p}^{l}[t+1] = \eta\left(\sum_{n}\sum_{s} \delta_{p}^{l+1}(n', s'; n, s) \cdot V_{p}^{l+1}(n, s)\right) + \alpha \Delta W_{p}^{l}[t] \quad (9)$$

$$\Delta b_{p}^{l}[t+1] = \eta \sum_{n}\sum_{s} \delta_{p}^{l+1}(n', s'; n, s) + \alpha \Delta b_{p}^{l}[t] \quad (10)$$

$$\delta_{p}^{l}(n', s'; n, s) = S'\left(N_{p}^{l}(n', s')\right)\left(\sum_{n}\sum_{s} \delta_{p}^{l+1}(n, s) \cdot W_{p}^{l+1}(n, s)\right) \quad (11)$$

In the case of the last layer,

$$\delta^{L}(n) = S'\left(N^{L}(n)\right)\left(T(n) - O(n)\right) \quad (12)$$

where $S'(z)$, η, α, and T denote the derivative of $S(z)$, the learning rate, the weighting factor contributed by the momentum term, and the desired output image, respectively. Furthermore, s or s' = 1 and p=1 when $l \neq 0$.

During the training, we added an isotropic constraint to the weights of the 1-D convolution kernels so that

$W^{0}_{q}(n, -s) = W^{0}_{q}(n, s)$ (13)

where q is not the fully connected path. These additional constraints are used to induce the kernels functioning as correlation processing filters and could facilitate the algorithm in searching for an appropriate filter.

(B) Resampling the training set through utilization of rotation and flip invariance of the features

In this neural network model, there are no starting and ending sectors. The forward and backpropagation computation can start from any sector. Considering a flipped patch, the characteristics of mass feature should remain the same. To take advantage of this flip invariance, the same numerical target value can be assigned at the output node for the flipped image patch in order to double the amount of cases during training.

Since we designed a 10^{o} increment for each rotation, each SC or NC path would process through 36 times using the defined features for each image patch. To simplify this network computation, we shifted one small sector (4 nodes) on the input layer at a time to conduct the circular convolution process with the SC and NC kernels in the

following experiments. By reversing the sequence of the sector, one can train the flipped version of the suspicious masses. Hence, using the properties of the rotation invariance and flip invariance for the neural network training literally increases the number of the training set by a factor of 72.

In summary, we have developed a complete detection procedure for the automatic recognition of mammographic masses including background adjustment, contrast enhancement, ROI segmentation, feature extraction, and MCPCNN system with a training method. Figure 5 shows a flow diagram for the essential sections of the computational procedures.

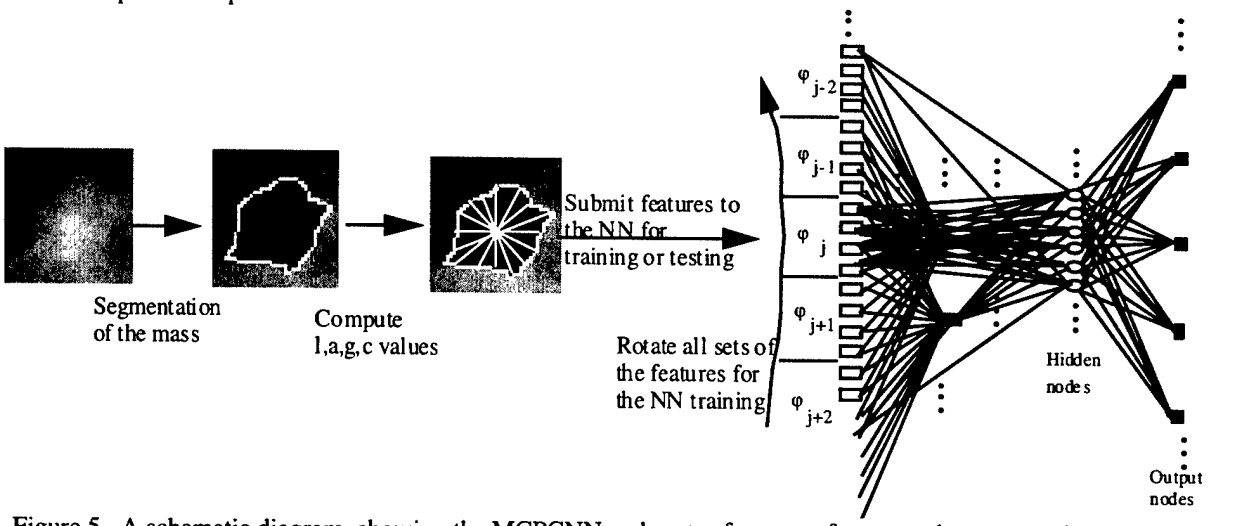


Figure 5. A schematic diagram, showing the MCPCNN and sector features of masses, that was used in the following study.

4. Experiments and Results

The 200 mammograms were selected from the Mammographic Image Analysis Society (MIAS) database and the Brook Army Medical Center (BAMC) database created by the Department of Radiology at Georgetown University Medical Center. Of the 200 mammograms, 50 mammograms are normal, and each of the 150 abnormal mammograms contains at least one mass case of varying size, subtlety, and location. Both the Cranio-Caudal (CC) and Medio-Lateral Oblique (MLO) projection views were used. The films were digitized with a computer format of 2048×2500×12 bits (for an 8"×11" area where each image pixel represents 100 μm square). Ninety-one mammograms, either a CC or an LMO view film, were selected from 91 patient film jackets. No two mammograms were selected from the same patient. All the digitized mammograms were miniaturized to 512×625×12 bits using 4×4 pixel averaging before the method was applied. According to radiologists, the size of small masses is 3-15 mm in effective diameter. A 3 mm object in an original mammogram occupies 30 pixels in a digitized image with a 100 μm resolution. After reducing the image size by four times, the object will occupy the range of about 7-8 pixels. The object with the size of 7 pixels is expected to be detectable by any computer algorithm. After pre-processing and an object screening based on the circularity test and the size test (between 3mm and 30mm), a total of 125 suspicious areas were selected from the testing mammograms (91 cases) for this study. Specifically, the screening procedure of reducing false-positives involves two steps (a) image patches with circularity less than 0.25 or diameter greater than 30mm were eliminated (b) using probability modular neural network (PMNN) to rule out the majority of false-positives. Of the 125 suspicious areas, 75 ROIs contained masses based on corresponding biopsy reports with one experienced radiologist reading. This set of ROIs was used in [19] and discussed in Figure 6 and Table II of [19].

4.1. Experiment 1

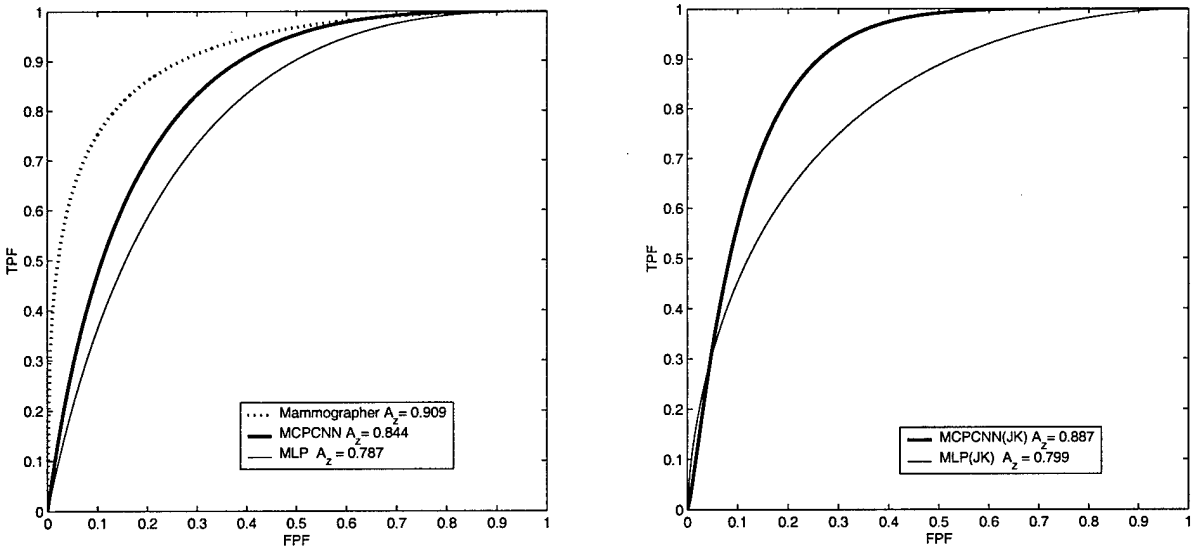
We randomly selected 54 computer-segmented ROIs where 30 patches were matched with the radiologist's identification and 24 were not. This database was used to train two neural network systems: (1) a conventional 3-layer neural network and (2) the proposed MCPCNN training method using the same neural network learning algorithm. The structure of the MCPCNN was described earlier. In the study, we used one fully connected path, four SC paths, four NC paths covering 2 sectors, four NC paths covering 3 sectors, three NC paths covering 4 sectors, and two NC paths covering 5 sectors in the first step network connection for the MCPCNN. All paths in the neural network have their hidden layers. Only one hidden layer per path was used. Both neural network systems were trained by the error backpropagation algorithm by feeding the features from the input layer and registering the

corresponding target value at the output node. Completion of the training was determined by the mean square error (i.e., $\sum_{i=i}^{N}(T_i - O_i)^2 \Big/ N$, where N is number of samples) when it was approximately reduced to $3x10^{-5}$. Once the training of the neural networks was completed, we then used the remaining 71 computer segmented ROIs for the testing. Forty-five out of 71 ROIs were masses and 26 ROIs were not. Neither the images nor their corresponding patients in the testing set could be found in the training set. The neural network output values were fed into the LABROC4 program [29] for the performance evaluation. The results indicated that the areas (Az) under the receiving operator characteristic (ROC) curves were 0.7869±0.0536 and 0.8443±0.0457 using the conventional neural network (MLP) and the MCPCNN, respectively. The ROC curves of these two neural network systems are shown in Figure 6(A). The Az value was 0.7869 ±0.0536 when using the MLP method with 125 hidden nodes. The performance of the MLP remains about the same at 0.7809 ±0.0551 of Az using the same neural network parameters but with 30 hidden nodes.

We also invited another senior mammographer to conduct an observer study using the ROC study protocol. The mammographer was asked to rate each patch using a numerical scale ranging 0-10 for its likelihood of being a breast mass. The image patches were displayed on a SUN monitor (Model: GDM-20D10). The image size shown on the monitor was reduced to approximately 7"x9" as compared to the original film size (8"x10"). These 71 numbers were also fed into the LABROC4 program. The Az of the mammographer's performance on this set of test cases was 0.909 ±0.0340. The corresponding ROC curve is also shown in Figure 6(A).

4.2. Experiment 2

We also conducted a leave-one-case-out experiment (i.e., Jackknife procedure) using the same database. In this experiment, we used those image patches extracted from 91 mammograms (one mammogram per 91 case) for the training and used the image patches (most of them are single) extracted from the remaining one mammogram as test objects. The procedure was repeated 91 times to allow every ROI extracted from each mammogram to be tested in the experiment. For each individual ROI, the computed features were identical to those used in Experiment 1. Again, the training was stopped when the mean square error value approximately equal to $3x10^{-5}$. Both neural network systems were independently trained and evaluated with the same procedure. The results indicated that the Az values were 0.7985 ±0.0394 and 0.8866 ±0.0289 using the conventional neural network (MLP) and the MCPCNN, respectively. The performance of the MLP decreased to an Az of 0.7608 ±0.0429 using the same neural network parameters but with 30 hidden nodes. Figure 6(B) shows the ROC curves of these two neural network systems using the leave-one-of-out procedure [30] in the experiment.



(A) (B)

Figure 6. The ROC curves obtained from corresponding experiments.

(A) The left figure shows that the performance of MCPCNN training method is superior to that of the conventional MLP method. The highest curve is the ROC performance of the senior mammographer.

(B) The right figure shows that the ROC results were increased using the leave-one-case-out procedure in both

neural network systems. The MCPCNN still showed higher performance than conventional MLP method.

We also used CLABROC program [31] to analyze the ROC data and compare the ROC results. The results and their statistical significances using two tailed p value of 0.05 as the threshold are shown in Table I. The radiologist's performance is greater than conventional neural network system with a p value of 0.0447 in the first experiment. The MCPCNN was also proven to be superior to the MLP with a statistically significant result ($p = 0.0241$).

Table I. ROC Performance of the Test Methods in Distinguishing True and False Masses

	Comparative Analyses of Methods	Az of Method (1)	Az of Method (2)	*P* Values	Statistical Significance
Experiment 1	(1) Radiologist vs. (2) MCPCNN	0.909 ± 0.0340	0.8443 ± 0.0457	0.1855	No
	(1) Radiologist vs. (2) MLP	0.909 ± 0.0340	0.7869 ± 0.0536	0.0447	**Yes**
	(1) MCPCNN vs. (2) MLP	0.8443±0.0457	0.7869 ± 0.0536	0.1344	No
Experiment 2	(1) MCPCNN vs. (2) MLP	0.8866 ±0.0289	0.7985 ± 0.0394	0.0241	**Yes**

5. Discussion

It is known in the field of artificial intelligence that the key factors in pattern recognition are: (1) effective methods in the extraction of features and (2) classification methods for the extracted features. In this study, we showed that the training method designed to guide the analyzer is also an important factor for a pattern recognition task. Though this finding is not new, the research of developing training methods for various pattern recognition tasks has not been established in the field of medical imaging. Our studies demonstrated that with proper network connections and task-oriented guidance, organized features would assist the neural network in performing the task.

Technically speaking, a feed-forward MLP neural network provides an integrated process for classification and sometimes for feature extraction. The output values of the hidden nodes can be interpreted as a reorganized set of features presented to the output layer for classification. The drawback of the MLP is, the user has a very little control and little understanding about the network learning. The MCPCNN is a network design that partially remedies these issues and is applicable for any pattern recognition task associated with ROIs. The MCPCNN (a member of the CNN family) possesses shared weights in the hidden layer(s) that act as filter kernels for extracting correlated features. With a higher resolution mammogram, a finer sector ($<10^{O}$) would be preferred for the analysis mass, especially for the study of classification of masses. During forward and backpropagation training, the kernels would comply with both signals from input and output layers for all training cases, so as to maximize the classification performance. One reason that we do not recommend using 2DCNN for the detection of masses is the sizes of masses vary. It would require a large fixed size to cover the maximum mass size when using the 2DCNN. The varieties of mass shapes and potential long spiculated patterns make the use of the 2DCNN not practical. Since the MCPCNN processes the features computed from sectors, it does not limit the sizes of its ROIs. Best of all, the MCPCNN also has the ability to classify partially obscured masses. The 2DCNN, however, would be more appropriate for the detection of microcalcifications and small lung nodules.

As far as the research in the detection of masses is concerned, we have shown that use of MCPCNN with sector features is an effective approach. Since the MCPCNN coordinates the input data and performs correlation between features of adjacent sectors in the first stage of data processing, the internal neural network learning algorithm can be changed if a learning algorithm is found to be more effective. In fact, the MCPCNN is a technique that can effectively classify features arranged in the polar coordinate system. A technique using the rubber band straightening transformation, independently developed by Sahnier et al. [12], for the detection of masses also employs a similar concept in extracting feature and/or texture in the polar coordinate system. We believe that integration of features and texture values computed at small sectors will be the research trend in mass detection and tumor classification.

6. Conclusions

In the clinical course of detecting masses, mammographers usually evaluate the surrounding background of a radiodense area when an ROI is suspected. In this study, we simulated this fundamental concept with a neural network system (i.e., MCPCNN). In order for the MCPCNN to function, boundary features of the suspicious region in each radial sector were computed. We found that the MCPCNN is capable of analyzing correlated features within the sector and between adjacent sectors, which led to an improvement in detecting mammographic masses.

Through this study, we found that the selected features are somewhat effective in the detection of masses. These features were "computationally translated" from the qualitative descriptors of BI-RAD. These features can be extended for the improvement of the mass detection, but this task is beyond the scope of this paper. With the preliminary studies shown above, we found the MCPCNN coupling with the proposed training method produced greater results than the conventional neural network. We found that the performances of both neural network systems were improved in Experiment 2. This may have occurred due to the number of training samples that was increased from 54 to 124. In Experiment 2, the Az value was improved by 0.043 using the MCPCNN, which was higher than the Az difference of 0.012 obtained by the conventional training method. The results implied that the MCPCNN learned more effectively than the conventional neural network when the number of training cases was increased. With the use of a larger database and advanced texture features proposed by others, it is expected that the performance of MCPCNN should be significantly improved. This paper does not intend to claim the best mass detection system, in comparison to similar systems; but rather its goal is to report a potentially better neural network structure for analyzing a set of mass features.

Acknowledgments

This work was supported by US Army Grant No. DAMD17-96-1-6254. The content of this paper does not necessarily reflect the position or policy of the government. A part of the database, used in the study, was provided by Dr. Robert Shah of Brooke Army Medical Center. The LABROC4 and CLABROC programs were written by Dr. C.E. Metz and his colleagues at the University of Chicago.

References

1. L. Nystrom, L. E. Rutqvist, S. Wall, A. Lindgren, M. Lindqvist, S. Ryden, et. al., "Breast cancer screening with mammography: Overview of Swedish randomized trials," Lancet, vol. 341, pp. 973-978, 1993.
2. S. Shapiro, "Screening- Assessment of current studies," Cancer, vol. 74, pp. 231-238, 1994.
3. L. Tabar, G. Fagerberg, S. Duffy, N. E. Day, A. Gad, and O. Grontoft, "Update of the Swedish two-country program of mammographic screening for breast cancer," Radiology Clinics of North America: Breast Imaging - Current Status and Future Directions, vol. 30, pp. 187-210, 1992.
4. D. Brzakovic, X. M. Luo, and P. Brzakovic, "An approach to automated detection of tumors in mammograms," IEEE Trans. Med. Img., vol. 9, pp. 233, 1990.
6. B. Sahiner, H. P. Chan, N. Petrick, D. Wei, M. A. Helvie, D. D. Adler, and M. M. Goodsitt, "Classification of mass and normal breast tissues: A convolution neural network classifier with spatial domain and texture images," IEEE Trans. on Med. Imag., vol. 15, pp. 598-610, 1996.
6. M. A. Kupinski and M. L. Giger, "Automated Seeded Lesion Segmentation on Digital Mammograms," IEEE Trans. Med. Imag., vol. 17, No. 4, pp. 510-517, 1998.
7. D. Wei, H. P. Chan, M. A. Helvie, B. Sahiner, N. Petrick, D. D. Adler, and M. M. Goodsitt, "Classification of mass and normal breast tissue on digital mammograms: Multiresolution texture analysis," Med. Phy., vol. 25, no. 4, pp. 516-526, 1998.
8. R. Zwiggelaar, T. C. Parr, J. E. Schumm, I. W. Hutt, C. J. Taylor, S. M. Astley, and C. R. M. Boggis, "Model-based Detection of Spiculated Lesions in Mammograms," Medical Image Analysis, Vol. 3, No. 1, pp. 39-62, 1999.
9. L. Hadjiiski, B. Sahiner, H. P. Chan, N. Petrick, and M. A. Helvie, "Classification of malignant and benign masses based on hybrid ART2LDA approach," IEEE Trans. on Med. Imag., vol. 18, No. 12, pp. 1178-1187, 1999.
10. H. Kobatake, M. Murakami, H. Takeo, and S. Nawano, "Computerized Detection of Malignant Tumors on Digital Mammograms," IEEE Trans. on Med. Imag., Vol. 18, No. 5, pp. 369-378, 1999.
11. N. Petrick, H. P. Chan, D. Wei, B. Sahiner, M. A. Helvie, D. D. Adler, "Automated Detection of Breast Masses on Mammograms Using Adaptive Contrast Enhancement and Texture Classification", Med. Phy., vol. 23(10), pp. 1685-1696, 1996.
12. B. Sahiner, H. P. Chan, N. Petrick, M. A. Helvie, and M. M. Goodsitt, "Computerized characterization of masses on mammograms: The rubber band straightening transform and textures analysis," Med. Phy., vol. 25(4), pp. 516-526, 1998.
13. D. D. Adler, *Breast masses: differential diagnosis. In Feig SA, ed. ARRS categorical course syllabus on breast imaging.* Reston, VA: American Roentgen Ray Society, p. 31, 1988.
14. M. J. Homer, "Imaging features and management of characteristically benign and probably benign breast lesions," Radiol. Clin. North Am., vol. 25, p. 939, 1987.

15. S. Pohlman, K.A. Powell, N. A. Obuchowski, W.A. Chilcote, S. Grundfest-Broniatowski, "Quantitative Classification of Breast Tumors in Digitized Mammograms", Med. Phys. vol. 23(8), pp. 1337-1345, 1996.
16. M. Moskowitz, *Circumscribed lesions of the breast. In Moskowitz M, ed. Diagnostic categorical course in breast imaging.* Oak Brook, Ill: Radiol. Soc. of North Am., 1986, p. 31.
17. E. A. Sickles, *The rule of multiplicity and the developing density sign. In: Feig SA, ed. ARRS categorical course syllabus on breast imaging.* Reston, VA: Am. Roent. Ray Soc., 1988, p. 177.
18. H. Li, Y. Wang Y, K-J. R. Liu, S-C. B. Lo, and M. T. Freedman, 'Computerized Radiographic Mass Detection-Part I: Lesion Site Selection by Morphological Enhancement and Contextual Segmentation," IEEE Trans. on Med. Imag., 2001, pp. 289-301.
19. H. Li, Y. Wang Y, K-J. R. Liu, S-C. B. Lo, and M. T. Freedman, 'Computerized Radiographic Mass Detection-Part II: Decision Support by Featured Database Visualization and Modular Neural Networks," IEEE Trans. on Med. Imag., 2001, pp. 302-313.
20. *Breast Imaging – Reporting and Data System*, American College of Radiology, Reston, Virginia, 1993.
21. J. Serra, *Image Analysis and Mathematical Morphology*, London, U.K. Academic, 1982.
22. S. Haykin, *Neural Networks: A Comprehensive Foundation,* 2nd ed., Prentice Hall, Inc., Upper Saddle River, New Jersey, 1999.
23. D. E. Rumelhart, G. E. Hinton, and R. J. Williams, *Learning internal representation by error propagation," ch. 8 of Parallel Distributed Processing: Explorations in the Microstructure of Cognition, vol. 1: Foundation,* D.E. Rumelhart & J.L. McClelland Eds., Cambridge, MA: M.I.T. Press, 1986, pp. 318-362.
24. S-C. B. Lo, S. L. Lou, J. S. Lin, M. T. Freedman, M.V. Chien, and S. K. Mun, "Artificial convolution neural network techniques and applications to lung nodule detection," IEEE Trans. Med. Imag., vol. 14, No. 4, pp. 711-718, 1995.
25. S-C. B. Lo, H. P. Chan, J. S. Lin, H. Li, M. T. Freedman, and S. K. Mun, "Artificial convolution neural network for medical image pattern recognition," Neural Networks, 1995, Vol. 8, No. 7/8, pp. 1201-1214, 1995.
26. H. P. Chan, S-C. B. Lo, B. Sahiner, K. L. Lam, and M. A. Helvie, "Computer-aided diagnosis of mammographic microcalcifications: Pattern recognition with an artificial neural network," Med. Phy., vol. 24, No. 10, pp. 1555-1567, 1995.
27. Y. Wu, K. Doi, M. L. Giger, & R. M. Nishikawa, "Computerized detection of clustered microcalcifications in digital mammograms: Applications of artificial neural networks," Med. Phy., vol. 19, pp. 555-560, 1992.
28. Y. Wu, M. T. Freedman, S-C. B. Lo, R. A. Zuurbier, A. Hasegawa, and S. K. Mun, "Classification of microcalcifications in radiographs of pathological specimen for the diagnosis of breast cancer," Acad. Radiol., pp. 199-204, 1995.
29. C. E. Metz, B. A. Herman, and J. H. Shen, "Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data," Stat Med., vol. 17, pp. 1033-1053, 1998.
30. K. Fukunaga and R. R. Hayes, "Effects of sample size in classifier design," IEEE Trans. Pattern Anal. Machine Intell., PAMI-11, pp. 873-885, 1989.
31. C. E. Metz, P.-L. Wang and H. B. Kronman, "A new approach for testing the significance of differences between ROC curves measured from correlated data," In Information Processing In Medical Imaging (F. Deconinck, Eds.). Martinus Nijhoff: The Hague, pp. 432-445, 1984.

Optimization of Wavelet Decomposition for Image Compression and Feature Preservation

Shih-Chung B. Lo[1], Huai Li[2], Matthew T. Freedman[1]

[1]Center of Imaging Science and Information Systems, Radiology Department, Georgetown University Medical Center, Washington, D.C. 20007

[2]Odyssey Technologies Inc., Jessup, Maryland

Li was with the ISIS Center, Radiology Department, Georgetown University Medical Center where he worked with the other authors to conduct this study.

(Paper submitted to IEEE Transaction on Medical Imaging for review)

Abstract A neural network based framework has been developed to search for an optimal wavelet kernel that is tailored for a specific image processing task. In this paper, a linear convolution neural network was used to seek a wavelet that minimizes errors and maximizes compression efficiency for an image or a defined image pattern such as microcalcifications on mammograms. We have used this method to evaluate the performance of tap-4 wavelets on mammograms, CTs, MRIs, and Lena images. We found that Daubechies wavelet or those wavelets possessing similar filtering characteristics produces the highest compression efficiency with the smallest mean-square-error. However, Haar wavelet produces the best results on sharp edges and low-noise smooth areas. We also found that a special wavelet, whose low-pass filter coefficients are (0.32252136, 0.85258927, 0.38458542, -0.14548269), can greatly preserve the microcalcification features in peak signal-to-noise ratio, contrast, and figure of merit during a course of compression. Explanations of the experimental results are provided by reviewing the spectrum of wavelet filters. This newly developed optimization method can be generalized to other image analysis applications where a wavelet decomposition is employed.

Keywords: Optimization of wavelet, neural network, wavelet decomposition, image feature restoration, and image compression.

I. Introduction

In the field of transform coding, discrete cosine transform (DCT) based decomposition methods were developed extensively in the 1970's and 1980's. Most of these techniques are associated with block DCT [1]-[4]. However, several investigators have indicated that the use of full-frame DCT [5]-[7] can produce high compression efficiency with high data fidelity and without blocky artifact. This method is particularly appropriate for high-resolution large-sized images. Recently, sub-band and wavelet transformations have been widely used in image compression research [8]-[10]. Unlike DCT, wavelet transform coefficients are partially localized in both spatial and frequency domains, and form a multiscale representation of the image. In addition, wavelet transform coefficients possess orientation specificity. Since wavelet transform has these attractive features, many efforts have been made to effectively encode transformed coefficients [11]-[13]. As a result, it has been shown that wavelet transform, in many situations, obtained significantly greater results than those obtained from block DCT techniques in a course of image compression.

The wavelet decomposition is determined by one mother wavelet function and its dilation and shift versions. Since the mother wavelet functions are not unique, different wavelet bases can produce different wavelet coefficients. Investigators often have a difficulty in selecting an optimal wavelet for a specific image procession application. Many technical issues relating to this area remain unsolved. The choice of optimal wavelet depends on different criteria in various applications. The early work on selecting wavelet basis can be found in [14]-[16]. In [16], Tewfik et. al. proposed a method for selecting a wavelet for signal representation based on minimizing an upper bound of the L^2 norm of error in approximating the signal up to a desired scale. In [14], Coifman and Wickerhauser developed an entropy-based algorithm for choosing the best wavelet basis that can achieve higher compression ratios with a generalization of wavelet packets, at the expense of increased processing time. Villasenor, et al. derived a wavelet filter evaluation metric according to the filter impulse response and step response in addition to regularity. Based on this metric, some of the best filters suitable for image compression were selected from a biorthogonal wavelet filter bank [15].

In the field of image compression, one major criticism in the evaluation of reconstructed images is that investigators usually present a global measure of a large image rather than a local quantification of a specific image pattern. The former only provides overall performance of a compression technique on an image. In many applications, particularly in medical imaging, a local image pattern can be of major concern. In such a case, the performance at the region of interest (ROI) should be weighted much higher than that of other places. Our research goal is to investigate which wavelet filter performs the best compression result for a given image pattern. In our experiment, we isolated various types of ROIs on several medical imaging modalities for the evaluation of data fidelity and compression ratio using wavelet decomposition techniques.

On the selection of an optimal wavelet, we propose to use a linear convolution neural network system that possesses an embedded wavelet operation. Through a controlled backpropagation algorithm, the neural network is capable of searching for an optimal wavelet that minimizes quantization errors and at the same time produces the highest compression efficiency. This newly developed convolution neural network method can also be extended to evaluate various wavelets in preserving defined image features.

The rest of this paper is organized as follows. In Section II, discrete wavelet transform is briefly reviewed and a wavelet based convolution neural network is described. In addition, migration from a wavelet kernel to another, which is embedded in the searching method of the neural network system, is also presented. Section III describes the experimental method used to evaluate the proposed approach. The results are given in Section IV. Section V discusses the results of wavelet search with the image patterns and characteristics of optimized wavelets. Section VI summarizes the technical achievements and their implications in the field of medical image compression.

II. Algorithm Development

II.A. Two-Dimensional Wavelet Decomposition

Following Mallat's 2-D wavelet analysis [9], the two-dimensional scaling function is composed of two one-dimensional scaling functions in both directions if they are separable:

$$\phi(x,y) = \phi(x)\phi(y) \quad ...(1)$$

where $\phi(x)$ is a scaling function. The associated two-dimensional wavelets are defined as

$$\psi^H(x,y) = \phi(x)\,\psi(y) \quad ...(2)$$

$$\psi^V(x,y) = \psi(x)\phi(y) \quad ...(3)$$

$$\psi^D(x,y) = \psi(x)\psi(y) \quad ...(4)$$

where $\psi(x)$ is the 1-D wavelet corresponding to the 1-D scaling function. Using the sub-band coding algorithm, the wavelet transform (2-D DWT) of a matrix has four parts:

$$W_{LL}(f(x,y)) = \sum_{u,v}[(f(x,y)h(u-2x,0))h(0,v-2y)] = \sum_{u,v}[f(x,y)h_{LL}(u-2x,v-2y)] \quad ...(5)$$

$$W_{LH}(f(x,y)) = \sum_{u,v}[(f(x,y)h(u-2x,0))g(0,v-2y)] = \sum_{u,v}[f(x,y)h_{LH}(u-2x,v-2y)] \quad ...(6)$$

$$W_{HL}(f(x,y)) = \sum_{u,v}[(f(x,y)g(u-2x,0))h(0,v-2y)] = \sum_{u,v}[f(x,y)h_{HL}(u-2x,v-2y)] \quad ...(7)$$

$$W_{HH}(f(x,y)) = \sum_{u,v}[(f(x,y)g(u-2x,0))g(0,v-2y)] = \sum_{u,v}[f(x,y)h_{HH}(u-2x,v-2y)] \quad ...(8)$$

where h and g functions are the low-pass and high-pass filters of the sub-band decomposition with condition

$$g(u) = (-1)^u h(1-u). \quad ...(9)$$

The low-pass filter, h, also must satisfy two criteria to construct the orthogonal basis of compactly supported wavelets [8], [9]. For simplicity, we also use g_u and h_u to replace $g(u)$ and $h(u)$, respectively, in this paper.

(a) $$\left[\sum_u h_{2u}\right] - \sqrt{2}/2 = \left[\sum_u h_{2u+1}\right] - \sqrt{2}/2 = 0; \quad ...(10)$$

(b) should be orthogonal; this means that

$$\left[\sum_u h_u \times h_{u+2n}\right] - \delta_{u,u+2n} = 0 \quad ...(11)$$

where $\delta_{i,j}$ is Dirac delta function and n is an integer. However, high degree of regularity and high degree of vanish moments

were not imposed in this study. Because these are very strong constraints from the compression point of view. Typically, those filters performing perfect reconstruction are illegible in a data compression scheme. With a single low-pass filter system, the criterion of perfect reconstruction is the same as that in (11). Because the synthesis filter is the reflection function of the analysis filter. For simplicity, we will use a single low-pass filter to perform wavelet transform in the following algorithm development.

The 2-D filters at the second forms of (5)-(8) are the vector products of h and/or g filters. The relationship between high-pass and low-pass filters makes the unification of the four sets of decomposition possible as shown in Section II. D. According to wavelet theory, it is known that given a set of h, one can calculate the Fourier transform of the scaling and wavelet functions as follows:

$$\Phi(w) = H_0(e^{iw/2})\Phi(w/2) \quad ...(12)$$

$$\Psi(w) = H_1(e^{iw/2})\Phi(w/2) \quad ...(13)$$

where H_0 and H_1 are Fourier transforms of h and g filters, respectively. Hence, both the scaling and wavelet functions can be obtained through infinite recursion by using (12) and (13), respectively.

II.B .Construction of a Neural Network using Wavelet Decomposition

The artificial neural network described in this paper is based on the linear convolution process which is used in sub-band and wavelet decomposition techniques. Each wavelet processing in the neural network performs exactly the same as the conventional wavelet transform given in (5)-(8). Our approach is to use the searching capability of the neural network to obtain the most suitable wavelet kernel through an image compression scheme [17]. In this paper, one major research task is to minimize error and simultaneously achieve the highest compression efficiency during the course of compression and decompression processes. In order to match the sub-band decomposition, several characteristics of the neural network must be established: (a) no hidden but one output layer is used, (b) local connection through convolution process rather than fully connected nets is employed, and (c) the convolution process must be reversible (wavelet kernels are used in this paper). In order to study the data fidelity, we add a quantization in the compression process. Therefore, the image cannot be fully reconstructed after the decompression process. The differences between original and reconstructed images are not due to the inverse transformation but because of the inaccuracy of the quantized transform coefficients. Figure 1 shows the structure of the neural network using quantized transform coefficients as the targets.

Minimization of quantization errors was not the only issue in our technical consideration. The method to minimize the entropy must also be taken into account for the optimization in a course of data compression. We combined both issues by multiplying the mean-square-error function with an imposed entropy reduction function. The cost (error) function for searching the optimal wavelet kernel in the neural network becomes

$$Ef(i,j) = Z(QT(i,j)) \times [T(i,j) - QT(i,j)]^2/2 \quad ...(14)$$

where $QT(i,j)$ is the quantized transform coefficient at pixel (i,j) and $Z(QT(i,j))$, which is the entropy reduction function for a set of quantization coefficients, is given below:

$$Z(QT(i,j)) = \begin{cases} 0 & for \quad QT(i,j) = 0 \\ 1 & for \quad |QT(i,j)| = 1 \\ F(n,q) & for \quad |QT(i,j)| = n. \end{cases} \quad ...(15)$$

$F(n,q)$, which is a ramp function, is a function of quantization factor, q, and is somewhat inversely proportional to the quantized integer n. The value of the ramp function should always be smaller than 1.

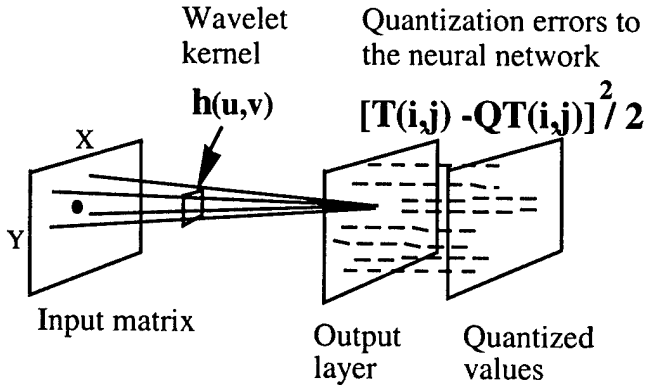


Figure 1. A wavelet-based neural network system that was used to seek an optimal wavelet for minimization of the quantization errors. *T(i,j)* and *QT(i,j)* denote transform and quantized coefficients in the high-frequency domain, respectively.

The reason to design the entropy reduction function for a fixed quantizer, q, using (15) is three-fold: (a) since most low value coefficients ($-0.5q < T(i,j) < 0.5q$) are associated with noise when q is not a very large value, there is no need to backpropagate errors from the output node possessing quantized value 0 in the neural net; (b) the more the small quantized values there are, the lower the assemble entropy; and (c) the probability to turn a high quantized value into a quantized value is very low, therefore errors backpropagated from high quantized value should be less emphasized as compared to low quantized value (e.g., 1, 2, or so). When q is very small, the quantization error is in the range of global image noise. In this case, the neural network would rely on the guidance of Z function to search for a wavelet filter that produces more low transform values. The success of this cost (i.e., error) function design is depicted in our experiment shown in Section IV.

Based on the neural network shown in Figure 1, we can seek an optimal convolution kernel. The specific searching algorithm is given in Section II.C. Section II.D shows a method to conduct orthogonal wavelet decomposition without using the high-pass filter. Hence, the low-pass filter is the only kernel to process the signals through 4 channels using two-dimensional (2-D) wavelet decomposition. In practice, the kernel directly suggested by the neural network in each epoch may not be a wavelet kernel. Section II.E provides algorithms that modify the kernel to fulfill the requirements of wavelet kernel. Through this process, we can find a wavelet that produces the lowest quantization errors with the lowest entropy of the quantized transform coefficients.

II.C. Signal Propagation through Convolution Process and Searching Methods in the Neural Network

The signal propagation from input layer to output layer involving convolution computation is given below [18]:

$$T_c(i,j) = K_c(i,j) \otimes S(i,j) \qquad ...(16)$$

where $\otimes$ represents convolution, $S(i,j)$ is the original image, subscript c denotes the channel number, and $K_c(i,j)$ is the convolution kernel for channel c. For the wavelet decomposition, the relationship between $K_c(i,j)$ and the wavelet filters (i.e., h and g filters) will be given in Sections II.C. and II.D.

Since we treat the wavelet transform as a locally connected neural network, the well-known backpropagation (BP) method can be used as a searching process by altering the 2–D convolution kernels in each epoch [19]. For each composed signal, a linear function instead of a typical sigmoid function in the neural network system is used in this process. The updated kernel suggested by backpropagation in the neural network is given by

$$K_c(u,v)[t+1] = K_c(u,v)[t] + \eta \sum_{i,j} \mu(i,j) S(u-i, v-j) + \alpha \Delta K_c(u,v)[t] \qquad ...(17)$$

where t is the iteration number during the searching, α is the gain for the momentum term received in the previous learning loop, η is the gain for the current weight changes, and μ is the weight-update function which is given by

$$\mu(i,j) = \frac{\partial Ef}{\partial K_c(u,v)} . \qquad ...(18)$$

Since a 2-D DWT is composed of a 4-channel wavelet decomposition, there are 4 associated convolution kernels to be updated simultaneously. The association between low-pass and high-pass filters, as shown in (9), is a necessary constraint of compact support in orthogonal wavelets. In order to preserve this property, we rearrange the decomposition method; hence, only a single kernel is needed to perform the 2-D DWT as demonstrated in the next subsection.

II.D. Unification of the Four Channels Decomposition in 2-D DWT

Using (6) as an example to rewrite the decomposition equation by replacing the g with the h filter, we have:

$$W_{LH}(f(x,y)) = \sum_{u,v}\left[(f(x,y)h(u-2x,0))(-1)^v h(0,2y+1-v)\right] \quad ...(19)$$

or

$$W_{LH}(f(x,y)) = \sum_{u,v}\left[(((-1)^v f(x,-y))h(u-2x,0))h(0,v-2y)\right]$$

$$= \sum_{u,v}\left[(((-1)^v f(x,-y))h_{LL}(u-2x,v-2y)\right] = \sum_{u,v}\left[f_{LH}(x,y)h_{LL}(u-2x,v-2y)\right]. \quad ...(20)$$

Converting (7) and (8) to use the 2-D low-pass filter as the kernel is a matter of changing the orientation from y- to x-direction. These conversions also indicate that one can use a single 2-D filter to compute the four quadrants of the 2-D wavelet transform by flipping the matrix position in x- and/or y-directions and alternating the sign of the flipped matrix corresponding to the directions.

The alternated sign of the source vector makes the convolution operation unconventional. A precalculation method, that involves a cross product of two vectors, can be employed: flipping the data sequence of an image is the first vector and the second vector is fixed and composed of +1 and -1. An example of 1-D precalculation steps for tap-6 kernel prior to the convolution operation is given below:

Original data sequence: $a_1, a_2, a_3, a_4, a_5, a_6$

Flipped data sequence: $a_6, a_5, a_4, a_3, a_2, a_1$

Resultant data sequence: $a_6, -a_5, a_4, -a_3, a_2, -a_1$

In the case of 2-D, three matrices associated with horizontal, vertical, and diagonal decomposition for the second matrix in precalculation are shown in Figure 2. With this precalculation (or cross product of two matrices), only the low-pass filter $h_u h_v$ (h_u in 1-D) is needed for the final wavelet transform operation.

$$\begin{bmatrix} + & + & + & + & + & + \\ - & - & - & - & - & - \\ + & + & + & + & + & + \\ - & - & - & - & - & - \\ + & + & + & + & + & + \\ - & - & - & - & - & - \end{bmatrix} \quad \begin{bmatrix} + & - & + & - & + & - \\ + & - & + & - & + & - \\ + & - & + & - & + & - \\ + & - & + & - & + & - \\ + & - & + & - & + & - \\ + & - & + & - & + & - \end{bmatrix} \quad \begin{bmatrix} + & - & + & - & + & - \\ - & + & - & + & - & + \\ + & - & + & - & + & - \\ - & + & - & + & - & + \\ + & - & + & - & + & - \\ - & + & - & + & - & + \end{bmatrix}$$

Vertical operator *Horizontal operator* *Diagonal operator*

Figure 2. Three matrices used for the cross product precalculation.

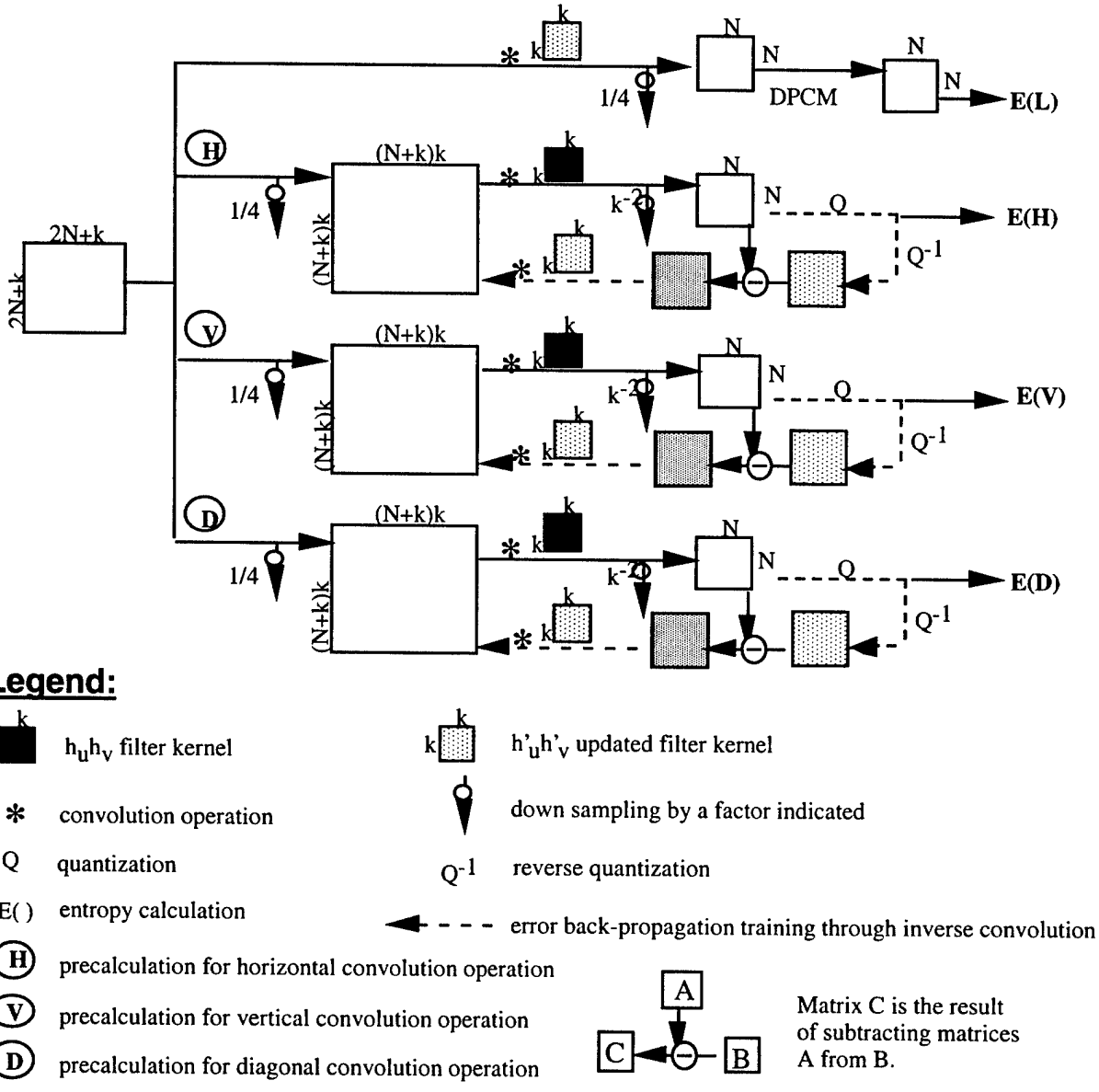
Nevertheless the resultant matrix of this precalculation or the cross product of two matrices must be held in the computer memory to facilitate the computation for forward convolution and the corresponding backpropagation. After precalculation, the size of the intermediate images is ($k/2 \times k/2$) times the original image size. The factor of 1/2× 1/2 is due to the 1/2 down sampling two-dimensionally in a conventional forward wavelet transform. The largest three blocks shown in Figure 3 are the intermediate images $S_O(xk/2, yk/2)$.

The perfect reconstruction criterion of the new filter may not be self-sustained with each updated version. However, some small modification is possible to make the final version of h_u, if the conditions of being a wavelet filter set are to be fully met. Based on each precalculated image $S_O(xk/2,yk/2)$ described earlier, (17) can be rewritten for updating 2-D wavelet kernel

$$K(u,v)[t+1] = K(u,v)[t] + \eta\sum_{i,j}\mu(i,j)S_0(u-xk/2, v-yk/2) + \alpha\Delta K(u,v)[t] \quad ...(21)$$

where index $i = 0,1,...(k-1)^2$ corresponds to the sub-image of S_O matched to the kernel size. (21) represents the updated

kernel suggested by the backprogapation, these values require a conversion to a new wavelet kernel $h'_u h'_v$. Assuming the wavelet filter is a 2-D vector (i.e., $h_u h_v = h_v h_u = h_{LL}$, where $u\&v$ = 0,1,2, ... k-1), then only k free parameters ought to be updated for each wavelet process. A solution to satisfy the wavelet constraints and to make $h'_u h'_v$ approximately equal to $K'(u,v)$ is given in Section II.E.



Legend:

- k (k) $h_u h_v$ filter kernel
- k (k) $h'_u h'_v$ updated filter kernel
- ✻ convolution operation
- down sampling by a factor indicated
- Q quantization
- Q^{-1} reverse quantization
- E( ) entropy calculation
- ◄- - - error back-propagation training through inverse convolution
- (H) precalculation for horizontal convolution operation
- (V) precalculation for vertical convolution operation
- (D) precalculation for diagonal convolution operation

Matrix C is the result of subtracting matrices A from B.

Figure 3. A proposed wavelet searching scheme based on a grouped kernel backpropagation neural network to obtain an optimal kernel for image compression.

II.E. Converting the Kernel Suggested by the Neural Network to Fulfill Requirements of a Wavelet Filter

As indicated in (21), the updated weights, $K(u,v)[t+1]$ or $K'(u,v)$ of the kernel suggested by the BP at $t+1$ searching iteration are independent. One must realize that each epoch in the neural network searching is only a suggestion or approximation that the changes of weights may produce a lower value for the defined error function, *Ef*. To properly use this suggestion for making a new wavelet kernel, let's assume that there exists a set of h'_u so that the updated 2-D version of the wavelet filter is very close to $K'(u,v)$. A function based on the square difference is used in the derivation

$$f(h'_u) = \sum_{u,v} (h'_u h'_v - K(u,v))^2 . \quad ...(22)$$

Here we intend to minimize the function, *f*, subject to the constraint equations. Lagrangian multiplier method can be employed to solve this problem by combining *f* and constraint equations:

$$df(h'_u) + \sum_p \lambda_p dC_p(h'_u) = 0 \quad ...(23)$$

where *d* represents the differentiation operation of a function and λ_p is the Lagrangian multiplier for the corresponding constraint equation, $C_p(h'_u) = 0$, referred to (10) and (11). Using this approach we can obtain a set of h'_u while *f* is also minimized.

III. Materials and Experimental Methods

A database consisting of 45 mammograms was used to conduct the study. Of these mammograms, 38 contain biopsy proven clustered microcalcifications. A total of 220 microcalcifications were embedded in 41 clusters. All 45 mammograms were digitized by a LumyScan (model 150) film digitizer (Lumysis Sunnyvale, CA) with spot size of 0.1 mm. Each patch of 32× 32 pixels (i.e., an area of 3.2 mm× 3.2 mm) with its center at the peak value was isolated for the study of quantization impact on microcalcifications. Note that typical sizes of microcalcifications range from 0.2 mm to 1.0 mm. It is important to isolate a specific image pattern, otherwise the neural network searching would be out of focus and could lead to a failure study. The processes of searching optimal wavelet kernels for original mammograms and microcalcification patches were conducted as separate studies. Each image was decomposed by 3-level wavelet transform. Quantization values were *q*, *q*/2, and *q*/4 for decomposition of high frequency coefficients on levels 1, 2, and 3, respectively. For each searching epoch, the mean-square-error (MSE) between the original and decompressed images as well as %zeros (i.e., number of zeros / total number of pixels) were computed. Since %zeros generally contributes the most important factor to gain a high compression, it was used as a coarse index for the evaluation of compression efficiency for each epoch.

In order to demonstrate each wavelet performance, we sorted the first coefficient h_0 of the low-pass filter associated with the mother scale function as the horizontal scale because the searching epoch could not represent the wavelet being used as shown in Figure 6. All h_0 values were greater than -0.1464466094 and smaller than 0.85255533905. The corresponding h_1 values are greater than 0.35355339 and smaller than 0.85255533905. Those h_1 values, which are greater than -0.1464466094 and smaller than 0.35355339, have corresponding conjugate values in the former set and can be ignored.

We have also performed the same study using the isolated 220 microcalcification patches. The 2-D profiles of microcalcificiations and their nearby areas (i.e., the areas that are not included in the microcalcification profile but within the isolated patch 32× 32 pixels) were evaluated separately during this course of the neural network search. In addition, features of the microcalcifications were computed to evaluate their changes. These features of microcalcifications are:

(a) the peak value, P;

(b) the contrast, C = P-m_b ;

where m_b is the average background value which is the immediate boundary of the microcalcification profile;

(c) the signal-to-noise-ratio, PSNR = C/ σ_b ;

where σ_b stands for the standard deviation of the background; and

(d) the area, A, occupied by the 2-D microcalcification profile.

IV. Results

In the neural network searching process, the MSE was not the only factor to be concerned; the entropy reduction function was another factor that drove the neural network to perform a search. In the first neural network experiment, we found that the MSE changes very little with a low quantization factor (q=16). The neural network process in searching for the next wavelet kernel was random and no minimum of MSE could be found in the mammogram study. However, the %zeros changed which led the neural network to converge at the maximum value of %zeros. In the microcalcification study, we found that %zeros does not change much until $h_0 > 0.6$. Figure 4 shows examples of selected cluster microcalcifications. Figure 5 shows the original learning steps. The figure indicates that MSEs moved toward smaller values using the proposed neural network searching mechanism. The h_0 values of searching epochs in Figure 5 were sorted in ascending order, and the MSEs as well as %zeros are replotted in Figure 6 which shows that Daubechies wavelet performs the lowest MSE. More specifically, microcalcification profiles suffered higher MSEs than their background areas as indicated in Figure 7.

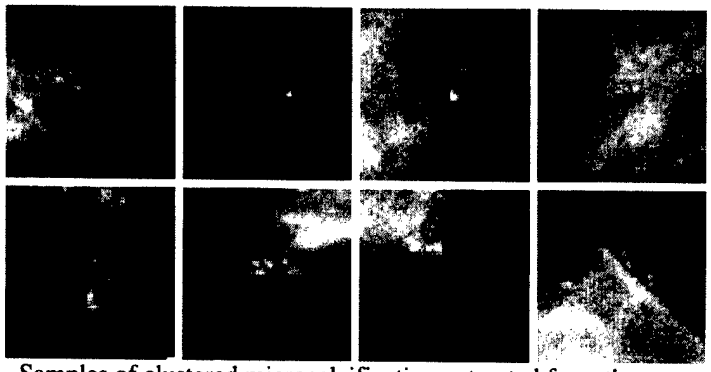


Figure 4. Samples of clustered microcalcification extracted from the mammograms.

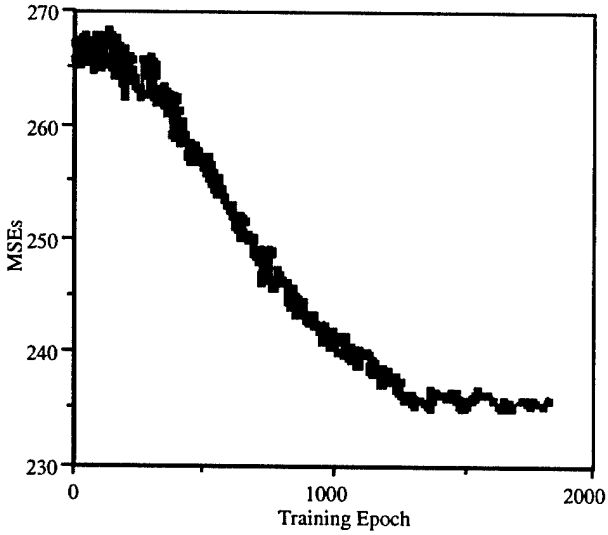


Figure 5. MSEs were decreased during the neural network search on 220 microcalcifications (q=64).

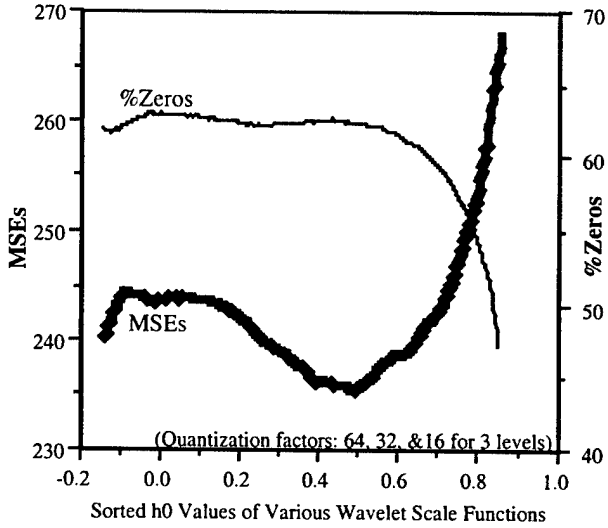


Figure 6. Decomposition performance of wavelets on 220 microcalcifications (q=64).

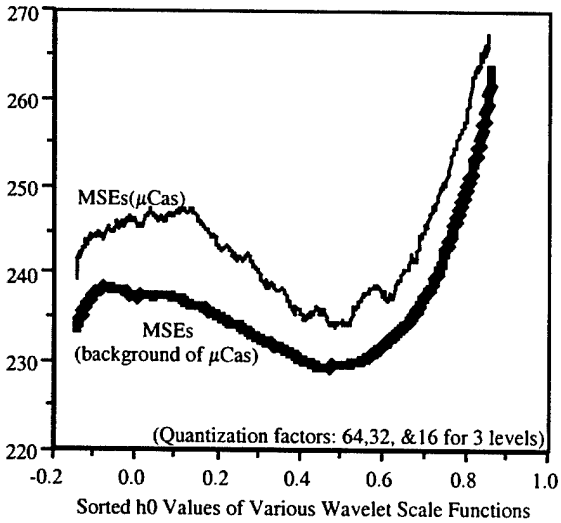


Figure 7. Decomposition performance of wavelets on 220 microcalcification profiles and background (q=64).

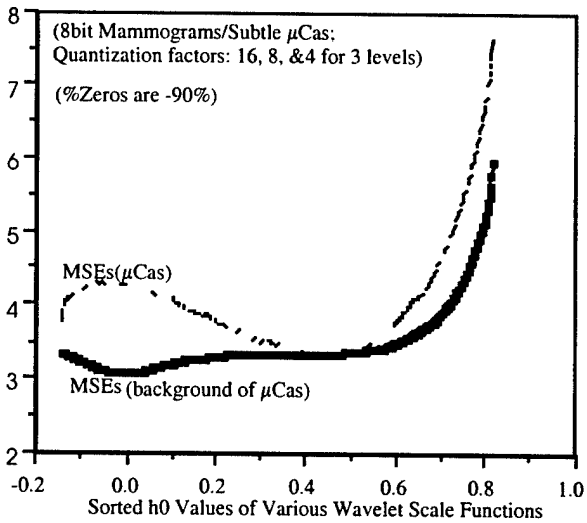


Figure 8. Decomposition performance of wavelets on 220 microcalcification profiles and background (8-bit, q=16).

These results were altered when a very large quantization factor was used. In Figure 8, all the digital values in

microcalcification patches were rounded-off to 8-bit prior to the study that assumed digitized mammograms containing about 4-bit of noise [20]. Although the largest quantization factor was 16 for 8-bit mammograms, the effective quantization factor was equivalent to ≈ 256 in 12-bit mammograms. Figure 8 shows that Haar wavelet ($h0 = 0.0$) performs the highest and the lowest MSEs for 2-D microcalcification profiles and their backgrounds, respectively. However, Daubechies wavelet performs in an opposite way. This is probably because Haar wavelet can produce a lower entropy in low-noise smooth areas.

The results of the microcalcification evaluation study based on quantized wavelet coefficients are shown in Figures 9-12. In fact, the evaluation was performed with an identical experimental condition as that in Figure 8. However, microcalcification features were measured instead of MSEs and %zeros. Note that the percent numbers decreases in peak values, contrast, and SNR were shown in negative values. In other words, the lower the percent number decrease value is, the more microcalcifications involving negative changes. The figure of merit (FOM) for each measure was a composed value given by

$$FOM = (\,\%\text{ No. decrease} \times \%\text{ decrease} + \%\text{No. increase} \times \%\text{ increase}) \times 100. \qquad ...(23)$$

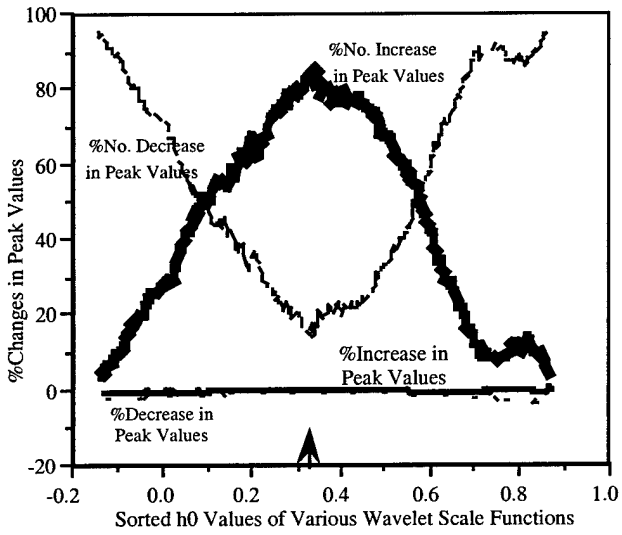


Figure 9. Peak value changes due to quantization effects on wavelet domain for microcalcifications.

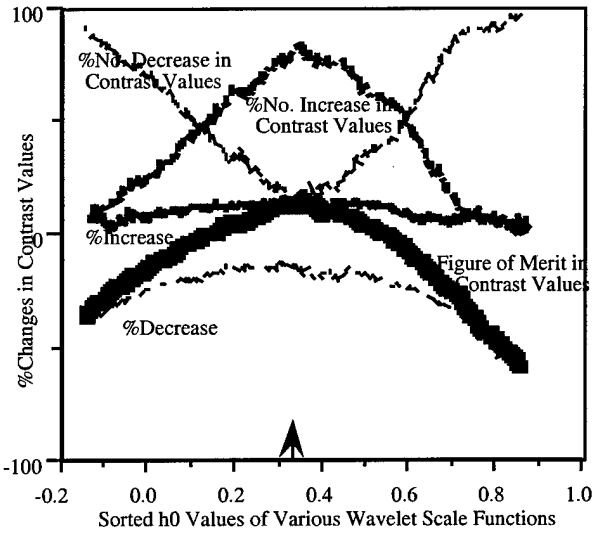


Figure 10. Contrast changes due to quantization effects on wavelet domain for microcalcifications.

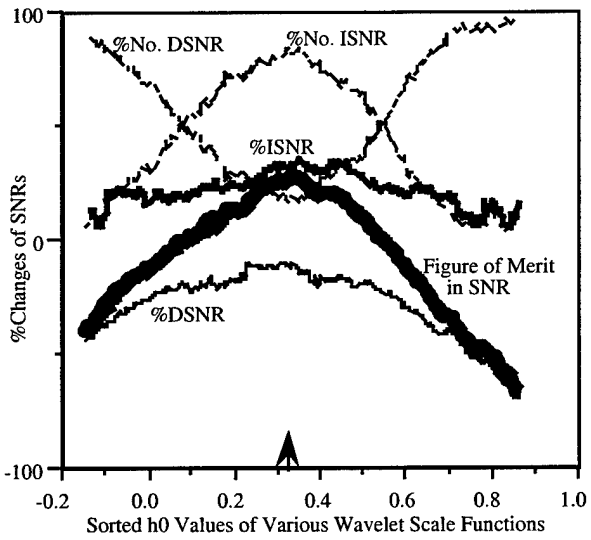


Figure 11. PSNR changes due to quantization effects on wavelet domain of microcalcifications.

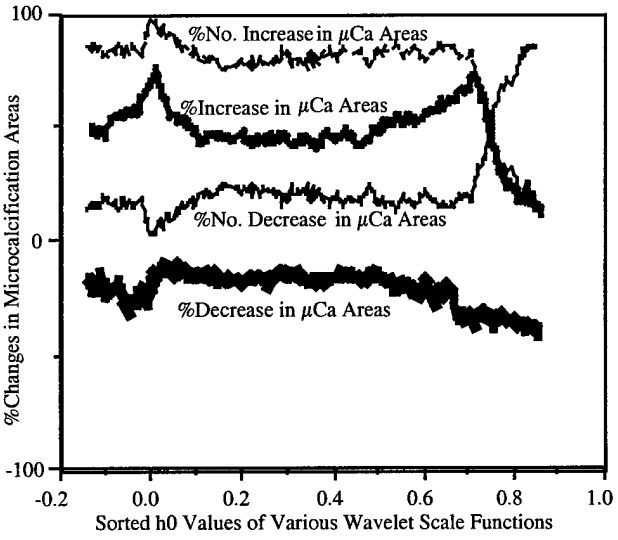


Figure 12. Percent changes in microcalcification profiles due to quantization effects on wavelet domain.

As indicated in Figure 9, the peak values had a very little changes. However, the percent in number increasing in peak values, contrast values, and SNRs of microcalcifications had approximately the same distribution in Figures 9, 10, and 11. The highest FOMs in all three measures occurred at the wavelet with the low-pass filter coefficients: (0.32252136, 0.85258927, 0.38458542, -0.14548269) which is marked with an arrow sign on the sort h_0 axis of the Figures. We call this wavelet a microcalcification friendly wavelet or μCaF wavelet for short. Figure 12 shows minor percent area changes of microcalcification profiles from 0.2 to 0.6 of h_0 values. These effects were not observed when a low quantization factor was used.

We also test the algorithm on other images. Figures 13 and 14 show the curves of MSEs and %zeros against the sorted h_0 values the Lena image and mammograms, respectively. In both figures, Daubechies (h_0 = 0.48296291) and its nearby wavelets produce the highest %zeros implying the largest compression ratio for mammograms and the lowest MSE for the Lena and mammograms. In these studies, we also found that when a larger quantization factor (q=64) was used, the MSE seemed to function in the neural network search. When a small quantization factor was used, the quantization errors were somewhat random; hence the neural network might not be properly functioned.

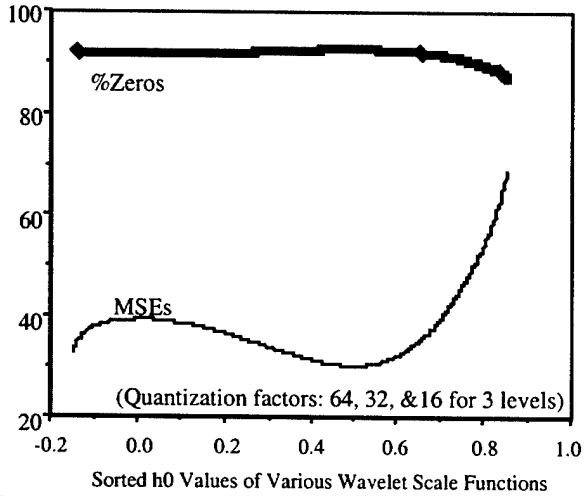


Figure 13. Decomposition performance of wavelets on the Lena image (q=64).

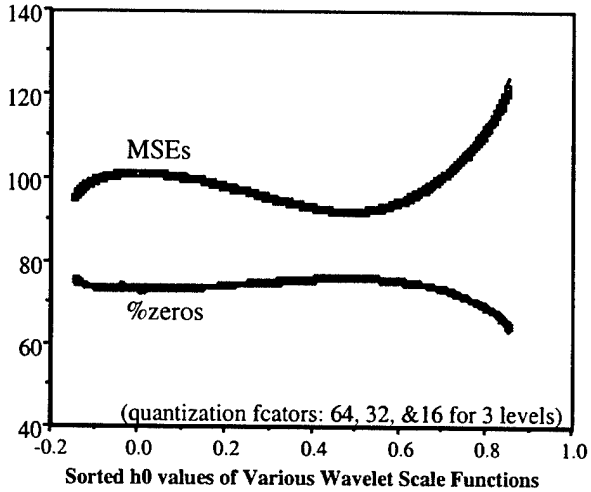


Figure 14. Decomposition performance of wavelets on mammograms (q=64).

V. Discussion

Having observed the above results, it is technically interesting to review the spectrum of wavelets mentioned. The low-pass h and the high-pass g filters of the wavelets are shown in Figures 15 and 16, respectively. Note that the μCaF wavelet is the same wavelet marked on the horizontal axis in Figures 9, 10, and 11. Attention should be paid to the g filters since they are responsible for decomposing high frequency coefficients regardless quantization. Essentially, the g filter performs calculation involving the positive weight multiplied by the center pixel value plus the adjacent pixel values on the two sides multiplied by the negative weights of the g filter. Daubechies wavelet has quite balanced negative terms at the two sides of the positive weight and the sum of negative weights is negatively equal to the positive weight. The latter is a constraint in all wavelet filters anyway. In addition, the absolute value of g_1 ($=-h_2$) or g_2 ($=h_1$) should be reasonably large, which would maintain the low-pass and the high-pass characteristics for h and g filters, respectively. In fact, those wavelets near Daubechies wavelet including the μCaF wavelet possess this property. From the signal processing point of view, these balanced weights in a filter are very important characteristics to create low entropy values for general textures. We suspect that this property may have something to do with so called "high regularity" in the wavelet theory.

In short, we found that the main reason that a wavelet filter can produce a low entropy for a set of data is because the weight sum of the g filter is zero. For a general data sequence, the g filter can perform even better when

(a) the absolute value of g_1 ($=-h_2$) or g_2 ($=h_1$) is much larger than that of other weights.

(b) the opposite signed weights are evenly distributed at the two sides of g_1 or g_2.

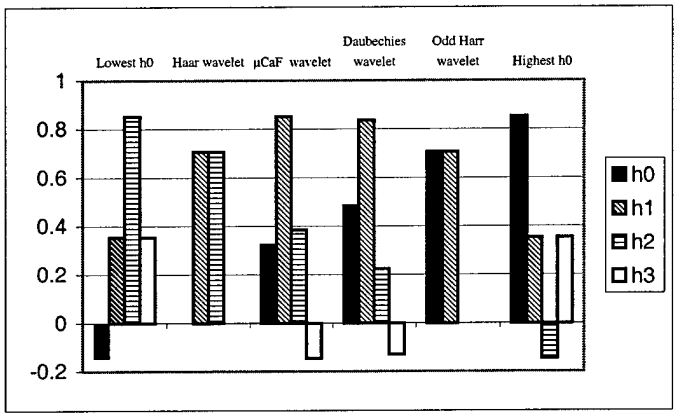


Figure 15. Low-pass filters of several interesting wavelets.

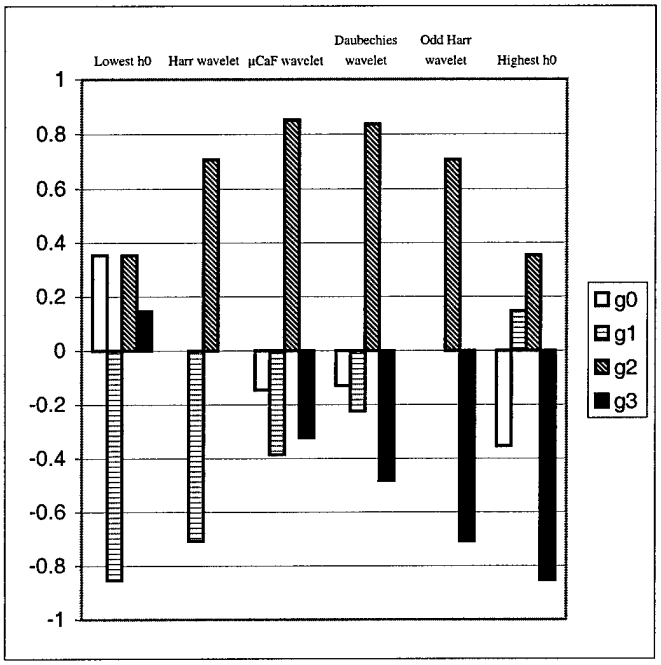


Figure 16. High-pass filters of the same wavelets.

For low-noise smooth signals, Haar wavelet may slightly outperform the others. For sharp edges, Haar wavelet would greatly outperform the others, as depicted in Figure 17 where only bones as well as edges between bones and soft tissues isolated on computed tomography (CT) images were the subjects for the evaluation. One of the tested CT head images was shown in Figure 18.

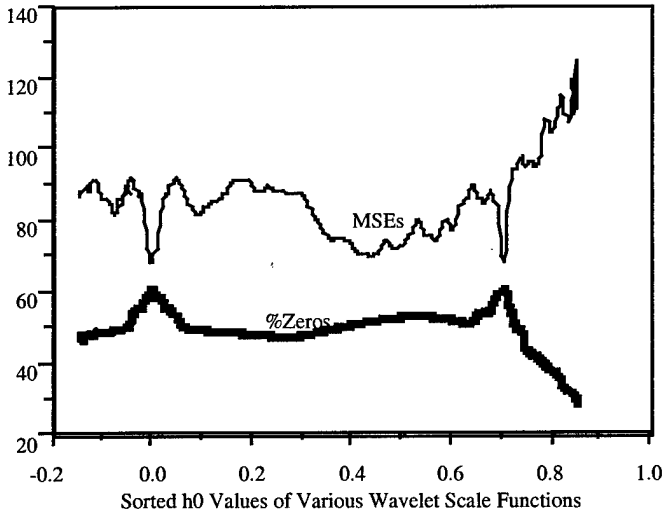


Figure 17. Decomposition performance of wavelets on CT head bones and bone edges (q=64).

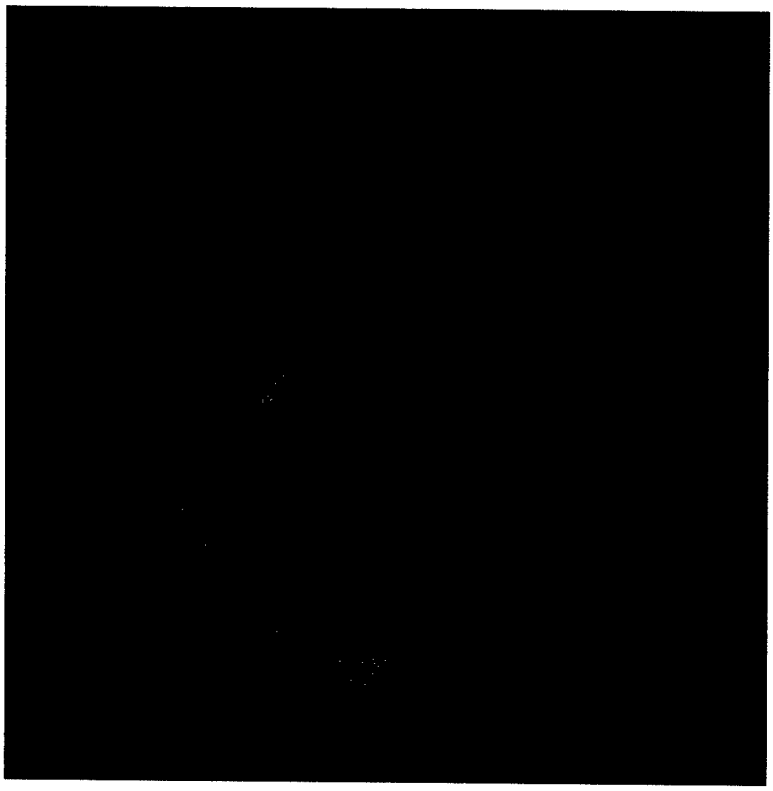


Figure 18. A CT head image

It was interested in finding out that the μCaF wavelet with (h_u, u=0..3) = (0.32252136, 0.85258927, 0.38458542, -0.14548269) results in the highest fidelity of features. Figure 9 provides evidences showing MSEs of 2-D microcalcification profiles and background gradually merge from Haar to Daubechies wavelets. Since contrast and PSNR values are computed using the peak and background values of the microcalcifications, the optimization of these measures should occur somewhere between Haar and Daubechies wavelets.

In the field of compression, it is known that the higher the compression ratio is, the higher the error that will be generated in the decompressed image. However, through these studies we discovered a new phenomenon associated with these two main quantitative measures in compression. We found that higher compression coincided with less error in all the studies (see Figures 4, 5, 7, & 16) using a fixed quantizer. This may be because high compression is associated with low entropy, which means that the data contains more low values and less variation between the originally transformed and quantized coefficients. This phenomenon happens only when the quantization factor is fixed. We would like to call to the reader's attention regarding the link between this phenomenon and the designed error function that comprises MSE and entropy reduction terms for searching an optional wavelet in the convolution neural network. With this concurrent trend (i.e., less error is associated with low entropy using a fixed quantizer), the neural network seems to be effectively operated in this searching task. Otherwise they would have functioned as competing factors and would have made the task difficult during the neural network search.

Although we have shown the general framework of a wavelet filter search using a neural network algorithm, only tap-4 wavelets were employed in our experiment. It seems that the above findings can be generalized for high order wavelets because the g filter is the key operator for the wavelet decomposition. The distribution of weights for high order wavelets should be maintained as discussed above in order to obtain a low entropy. We will continue to investigate the performance of dual low-pass filter wavelets where both an odd and even number of weights are used. We predict that high performance wavelets in compression and data accuracy should possess balanced distribution of weights in the g filter of wavelets [10, 15].

In our previous papers, we indicated that wavelet (both single and dual low-pass filter systems) decomposition might be appropriate for low-resolution small images such as the Lena image, CTs and MRIs. For high-resolution large images such as digitized chest radiographs and mammograms, we found that the full-frame DCT performed with the highest compression efficiency [21]. This is because the DCT can pack highly correlated image information in a small frequency area. The DWT, however, requires many levels in decomposition to achieve a high compression ratio. The data inaccuracy would propagate from high level wavelet domains to low level and to the reconstructed image.

VI. Conclusions

A neural network based method has been developed to search for optimal wavelet kernels which can produce the most favorable set of transform coefficients to preserve data accuracy and/or defined image features during the compression. In this paper, our technical achievements are: (a) development of a unified method to facilitate multi-channel wavelet decomposition; (b) designing a cost error function consisting of MSE and imposed entropy reduction function to seek an optional wavelet kernel in the convolution neural network; and (c) converting a neural network suggested kernel into a filter constrained by the wavelet requirements.

In all medical image modalities we have tested so far (including mammography, CT, MRI), Daubechies wavelet or its nearby wavelets generally performs slightly greater compression results than those of other wavelets based on the measure of mean-square-error. With a large quantization factor, Haar wavelet produces the lowest and highest MSEs for the background and microcalcification profile areas, respectively. However, Daubechies wavelet produces an opposite result. In addition, we found that the μCaF wavelet (i.e., 0.32252136, 0.85258927, 0.38458542, -0.14548269), possesses the highest feature preservation capability in microcalcification peak, contrast, and PSNR. Through this study, we also found that Haar wavelet sometimes produced a dramatic result for high contrast edges. In addition, optimization usually occurs on a band of wavelets not at a single wavelet.

We, therefore, conclude that Daubechies wavelet (and its nearby wavelets) is generally applicable for image compression. However, Haar wavelet is suitable for low-noise smooth areas and sharp edges. For a specific image pattern such as microcalcifications on mammograms, one might find that a wavelet filter can best preserve the features.

By reviewing the g filters of various wavelets, we found those optimal wavelets for general image texture have some things in common: they possess balanced negative terms at the two sides of the positive weight and the absolute value of g_1 or g_2 is much larger than that of the other weights.

Acknowledgments

This work was supported by an NIH/NCI Grant No. RO1CA79139 and an Army grant No. DAMD17-96-1-6254. The content of this paper does not necessarily reflect the position or the policy of the government, and no official endorsement should be inferred.

References

[1] W. K. Pratt, *Digital Image Processing.* New York: John Wiley & Sons, 1978.

[2] A. K. Jain, "Image data compression: A review," *Proc. IEEE*, Vol. 69, pp. 349-389, 1981.

[3] A. Rosenfeld and A. C. Kak, *Digital Picture Processing.* Academic Press, 1982.

[4] "Initial draft for adaptive discrete cosine transform technique for still picture data compression Standard," ISO/IEC JTC1/SC2/WG8 N800, JPEG Technical Specification, Revision 8, Aug. 1990.

[5] S. C. Lo and H. K. Huang, "Compression of radiological images with matrix sizes 512, 1024, and 2048," *Radiology*, pp. 519-525, 1986.

[6] H. K. Huang, S. C. Lo , B. K. Ho and S. Lou, "Radiological image compression using error-free an irreversible compression and reconstruction in 2-dimensional direct cosine transform coding techniques," *J. Optical Society of America,* pp. 984-992, 1987.

[7] S. C. Lo, B. H. Krasner, S. K. Mun, and S. C. Horii, "The full-frame entropy encoding for radiological image compression," *SPIE Proc. Medical Imaging V,* Vol. 1444, pp. 265-277, 1991.

[8] I. Daubechies, "Orthonormal based of compactly supported wavelets", *Comm. on Pure and Appl. Math.,* Vol. XLI, pp. 909-996, 1988.

[9] S. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Trans. Pat. Anal. Mach. Intel.,* Vol. 11, No. 7, pp. 674-693, 1989.

[10] M. Antonini, M. Barlaud, P. Mathieu, and I. Daubechies, "Image coding using wavelet transform," *IEEE Trans. Image Proc.,* Vol. 1, No. 2, pp. 205 - 220, 1992.

[11] T. Senoo and B Girod, "Vector quantization for entropy coding of image subbands," IEEE Trans. Image Processing, vol. 1, pp. 526-533, 1992.

[12] J. Shapiro, "An embedded hierarchical image coder using zerotrees of wavelet coefficients," In Proc. IEEE Data Compression Conf. '93, Mar. 1993, pp. 214-223.

[13] Z. Xiong, K. Ramchandran, M. T. Orchard, and K. Asai, "Wavelet packets-based image coding using joint space-frequency quantization," *In Proc. IEEE Int. Conf. on Image Proc.* Austin, TX, pp. 324-328, 1994.

[14] R. R. Coifman and M. V. Wickerhauser, "Entropy-based algorithms for best basis selection," IEEE Trans. Inform. Theory, vol. 38, pp. 713-718, 1992.

[15] J. D. Villasenor, B Belzer, and J. Liao, "Wavelet filter evaluation for image compression," IEEE Trans. on Image

Proc. vol. 4-8, pp. 1053-1060. 1995.

[16] A. H. Tewfik, D. Sinha, and P. Jorgensen, "On the optimal choice of a wavelet for signal representation," IEEE Trans. Inform, Theory, vol. 38, pp. 747-765, 1992.

[17] S. C. Lo, H. Li , J. S. Lin, A. Hasegawa, Y. C. Wu, M. T. Freedman, and S. K. Mun, "Artificial convolution neural network with wavelet kernel for disease pattern recognition," *SPIE Proc., Medical Imaging,* Vol, 2434, pp. 579-588, 1995.

[18] Y. LeCun, I. Guyon, L.D. Jackel, D. Henderson, B. Boser, R.E. Howard, J.S. Denker, W. Hubbad, and H.P. Graf, "Handwritten digital recognition: Applications of neural network chips and automatic learning," IEEE Comm. Magazine, pp. 41-46, Nov. 1989.

[19] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representation by error propagation," In D.E. Rumelhart & J.L. McClelland, & the PDP Research Group (Eds.), *Parallel Distributed Processing,* vol. 1, Cambridge, MA: MIT Press, 1986, pp. 318-362.

[20] S. C. Lo, B. H. Krasner, and S. K. Mun, "Noise impact on error-free image compression," *IEEE Trans. Medical Imaging,* Vol. 9, No. 2, pp. 202-206, 1990.

[21] S. C. Lo, H. Li, B. H. Krasner, M. T. Freedman, and S. K. Mun, "Large-frame compression using DCT and wavelet transform techniques," *SPIE Proc., Med. Imag.,* Vol. 2431, pp. 195-202, 1995.

Wavelet-Based Artificial Convolution Neural Network For Image Pattern Recognition: Applications to Microcalcification Detection on Mammograms

Shih-Chung B. Lo and Matthew T. Freedman

Center of Imaging Science and Information Systems, Georgetown University Medical Center, Washington, D.C. 20007

(Submitted to IEEE Medical Imaging for Review)

Abstract A wavelet based two-dimensional convolution neural network (WBCNN) has been developed for image pattern recognition. The structure of the convolution is based on the neocognitron. Nets between two adjacent layers in the feature selection level of the neural network are selectively interconnected across groups. Each group in the receiving layer receives signals from a group of weights (i.e., kernels). For the forward signal propagation, the product obtained from the kernel convoluting the front layer is collected onto the corresponding matrix element of the receiving layer. Since isolated patterns processed by internal filtering and classification layers built-in the neural network structure, the image patterns are expected to be more recognizable. The WBCNN was trained by a grouped process of backpropagation. In the WBCNN, we forced each updated convolution kernel to be orthonormal. Therefore, features (transformed coefficients) selected on the transform domain are linearly independent. Hence, the fully connected layers in the classification level of the CNN can perform more effectively.

The applications of the CNN for pattern recognition have been very successful. In this initial study, we used only a two-level structure and eliminated all complex-cell layers to evaluate the effects of wavelet kernel process. Although, we only limit improvement on the ROC performance using the WBCNN in the mammographic microcalcification studies, this method can assist us in the analysis of the trained kernels and expected to lead to the optimization of feature extraction in a course of pattern recognition.

Key words: Artificial neural network, wavelet decomposition, detection of microcalcifications, and image pattern recognition.

I. Introduction

Currently, the scope of research activities involving wavelet transform [1], [2] and artificial neural networks[3], [4] extends to a broad variety of fields. Successful applications have been reported in many areas, however, not much work has been done to adapt the strength of the two techniques in advancing the technology. In this study, we modify the two-dimensional convolution processing of a newly developed neural network to adapt the kernels with wavelet basis. The reasons for using wavelet kernels for the convolution process in the neural network are: (a) extracted features are linearly independent with wavelet decomposition, (b) many choices of wavelet bases allow the optimization of the system, and (c) the capability to perform multiresolution analysis.

One of the major criticisms of using a conventional backpropagation neural network (BPNN) for image recognition is that the neural nets are fully and uniformly connected from one node of the upper layer to all nodes in the lower layer. However, there is no guarantee that the adjacent pixel information in the image is more weighted than non-local pixel information during the training. On the other hand, the convolution neural network (CNN), which is a simplified version of vision-type neural network [5], has been inherently designed to perform local feature extraction. In this paper, we chose the CNN as the fundamental architecture of signal propagation platform and intend to explore advanced image pattern recognition algorithm. We follow a two-step approach of computer search [6]-[9]: (a) preliminary search for extracting suspected disease areas, and (b) examination of the suspected areas for final classification. The second step is the principal subject of this paper.

II. Algorithm Development

II.A. Review of Convolution Neural Network

Compare to other artificial neural networks, the neocognitron [5] possesses a network structure most similar to human vision. The neocognitron is composed of an input layer (retina) and four levels of grouped neurons. Each level consists of two layers. The front layer of each level is called a simple-cell layer. The second layer is a complex-cell layer. Each layer is composed of many neurons which are collected into several groups. Nets from simple-cell layer to complex-cell layer do not interconnect between groups. Nets from complex-cell layer to the next simple-cell layer are selectively interconnected between groups. All the nets are organized with a 2-dimensional convolution kernel. In the

training, the weights are adjusted by relatively complicated rules (functions) and inhibitory and excitatory theories.

The Convolution Neural Network for Image Pattern Recognition [11-13]

In the convolution neural network, weighting factors are shared and are formed as a kernel for each group in a collection layer. The convolution is processed between the weighting kernel and the front layer. This accounts for major differences between the vision type neural network and the regular fully connected neural network.

Besides the nets between the last hidden layer and output layer, Figure 1 shows that convolution processes are operated from an image block (e.g. $X^{\times}Y$ pixels) with a convolution kernel (e.g., size of $k^{\times}k$). The resultant data can also be organized as 2-D feature maps. Depending upon the number of independent kernel used (e.g., N kernels), we will receive M groups of 2-D feature maps in the next hidden layer. All nodes (e.g., O_n) on the output layer are fully connected to the last hidden layer. This neocognitron variant can be interpreted as following: (a) the convolution processes are designed to perform automatic feature extraction onto feature maps in the hidden layers. The fully connected networks, which process final signals, are used to make final classification as a regular neural network.

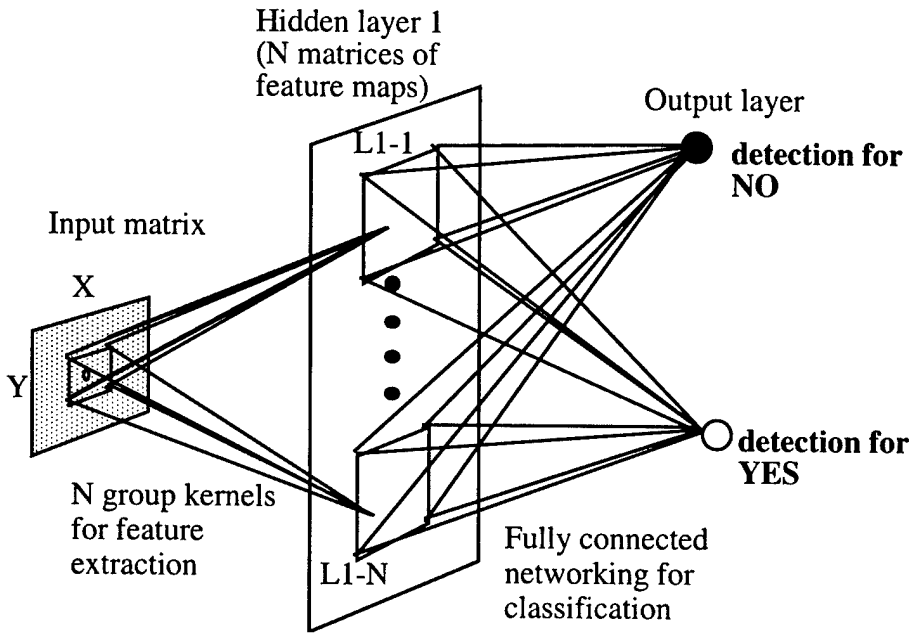


Figure 1. A simplified convolution neural network.

Signal Propagation and Training of the CNN

The signal propagation and backpropagation for fully connected networking follow standard BPNN algorithm [3]. However, the signal propagation from input layer to feature maps involving convolution computation is given below:

$$S_x((i,j);n)=\frac{1}{1+\exp\left\{-\sum_m\left[K((u,v);n,m)\otimes S_{x-1}((i,j);m)\right]\right\}} \quad ...(1)$$

or

$$S_x((i,j);n)=\frac{1}{1+\exp\left\{-\sum_{u,v,m}\left[K((u,v);n,m)\times S_{x-1}((i-u,j-v);m)\right]\right\}} \quad ...(2)$$

where $S_X((i,j);\ n)$ represents the signal at node $(i,\ j)$, nth group, and x layer. $K_X((u,v);\ n)$ denotes a weighting factor value at net $(u,\ v)$, nth group, and connecting from $x\text{-}1$ to x layer. $m\Leftrightarrow n$ represents the connection between groups m in layer $x\text{-}1$ and n in layer x.

Similar to a fully connected networking in a backpropagation neural network, the iterative version of kernel weights

is:

$$K_x((u,v);n)[t+1] = K_x((u,v);n)[t] + \eta \sum_{i,j} \{\delta_x((i,j);n) S_{x-1}((i-u, j-v);m)\} + \alpha \Delta K_x((u,v);n)[t] \quad ...(3)$$

where t is the iteration number during the training, α is the gain for the momentum term received in the last learning loop, η is the gain for the current weight changes, and δ is the weight-update function which is given as

$$\delta_x((i,j);n) = S_x((i,j);n)[1 - S_x((i,j);n)] Q_x((i,j);n) \quad ...(4)$$

where $Q_x((i,j);n) = \sum_{i,j,m \leftrightarrow n} K_{x+1}((u,v);m) \times \delta_{x+1}((i+u, j+v);m)$

Classification of Output Values in the Testing

Corresponding to the grading system arranged in the training, a polarized (linearly weighted) function is given as an indication. With this we can define a normalized object detection index (NODI) for the judgment of a suspected area:

$$NODI = \frac{\sum_{n=N/2}^{N-1} [O_n \times (n - (N-1)/2)]}{\sum_{n=0}^{N-1} [O_n] \times (N-1)/2} \quad ...(5)$$

where n denotes the node in the output layer, O_n is the output value at node n, and N is the total number of output nodes. Hence an object detection index of 0 or near 0 indicates a definite non-object and an objection detection index of 1 or greater implies a definite case with the judgment of the neural network. The reason for the weighting is that the score line is centered at (N-1)/2 (i.e., 0.5 for 2 nodes in the output layer) and polarization of true and false depends on the position of the nodes.

After receiving NODI value from each suspected area, we use a computer program (LABROC) [14], based on receiver operating characteristic (ROC) [15] analysis to evaluate the performance of the neural networks. The area under the curve referred to as A_Z, can be read as a performance index of the system using ROC analysis. In general, the higher the A_Z is, the better the performance.

II.B. Wavelet Kernels for CNN

Two-Dimensional Wavelet Transform

In the process of a two-dimensional wavelet decomposition, horizontal (x-) and vertical (y-) directions are considered preferential. Following Mallat's 2-D wavelet analysis [16], the two-dimensional scaling function is composed of two one-dimensional scaling functions in both directions:

$$\phi(x,y) = \phi(x)\phi(y) \quad ...(6)$$

where $\phi(x)$ is a scaling function. The associated two-dimensional wavelets are defined as

$$\psi^H(x,y) = \phi(x)\psi(y) \quad ...(7)$$

$$\psi^V(x,y) = \psi(x)\phi(y) \quad ...(8)$$

$$\psi^D(x,y) = \psi(x)\psi(y) \quad ...(9)$$

where $\psi(x)$ is the 1-D wavelet corresponding to the 1-D scaling function. Using the sub-band coding algorithm, the wavelet transform (2-D DWT) of a matrix has four parts:

$$W_{LL}(f(n,m)) = \sum_{u,v} [(f(n,m)h(u-2n,0))h(0,v-2m)] = \sum_{u,v} [f(n,m)h_{LL}(u-2n,v-2m)] \quad ...(10)$$

$$W_{LH}(f(n,m)) = \sum_{u,v}[(f(n,m)h(u-2n,0))g(0,v-2m)]$$
$$= \sum_{u,v}[f(n,m)h_{LH}(u-2n,v-2m)] \quad ...(11)$$

$$W_{HL}(f(n,m)) = \sum_{u,v}[(f(n,m)g(u-2n,0))h(0,v-2m)]$$
$$= \sum_{u,v}[f(n,m)h_{HL}(u-2n,v-2m)] \quad ...(12)$$

$$W_{HH}(f(n,m)) = \sum_{u,v}[(f(n,m)g(u-2n,0))g(0,v-2m)]$$
$$= \sum_{u,v}[f(n,m)h_{HH}(u-2n,v-2m)] \quad ...(13)$$

where h and g functions are the low and high pass filters of the subband decomposition with condition $g(u) = (-1)^u h(1-u)$. The low pass filter, h, also must satisfy three criteria to construct the orthonormal basis of compactly supported wavelets [1], [2]: (a) $\sum_u h(2u) = \sum_u h(2u+1) = \sqrt{2}/2$; (b) should be orthonormal; and (c) have a certain degree of regularity. The 2-D filters at the second forms of the above four equations are the vector products of h and/or g filters. The relationship between high pass and low pass filters make the unification of the above four sets of decomposition possible.

According to the wavelet theory, it is known that given a set of h, one can calculate the Fourier transform of the scaling and wavelet functions as follows:

$$\Phi(w) = H_0(e^{iw/2})\Phi(w/2) \quad ...(14)$$

$$\Psi(w) = H_1(e^{iw/2})\Phi(w/2) \quad ...(15)$$

where H_0 and H_1 are Fourier transforms of h and g filters, respectively. Hence, both the scaling and wavelet functions can be obtained through infinite recursion by using Eqs. (14) and (15), respectively.

Using the Low Pass Filter for the Four Channels Decomposition of 2-D DWT

Using Eq. (10) as an example to rewrite the decomposition equation by replacing g with h filter, we have:

$$W_{LH}(f(n,m)) = \sum_{u,v}\left[(f(n,m)h(u-2n,0))(-1)^l h(0,2m+1-v)\right] \quad ...(16)$$

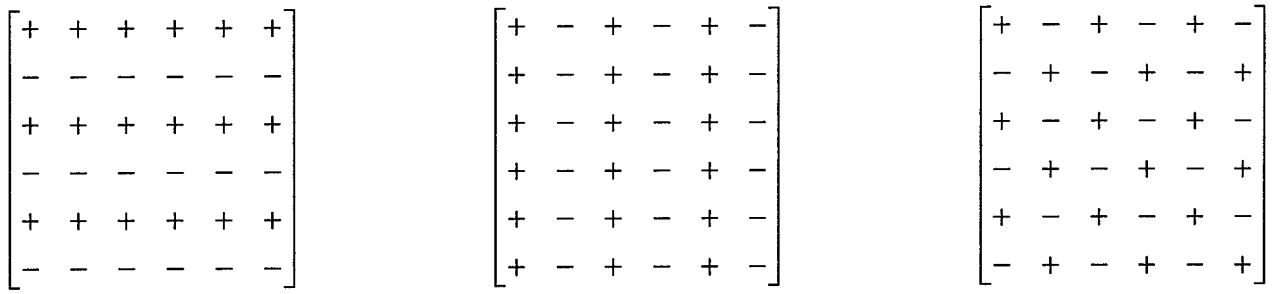
or

$$W_{LH}(f(n,m)) = \sum_{u,v}\left[(((-1)^v f(n,-m))h(u-2n,0))h(0,v-2m)\right]$$
$$= \sum_{u,v}\left[(((-1)^v f(n,-m))h_{LL}(u-2n,v-2m)\right] \quad ...(17)$$
$$= \sum_{k,l}\left[f_{LH}(n,m)h_{LL}(u-2n,v-2m)\right]$$

Converting Eq. (12) to use the 2-D low pass filter as the kernel is a matter of changing the orientation from y- to x-direction (or combining both directions for Eq. (13)). These conversions also indicate that one can use a single 2-D filter to compute the four quadrants of the 2-D wavelet transform by flipping the matrix position in x- and/or y-direction(s) and alternating the sign of the flipped matrix corresponding to the direction(s).

The alternated sign of the source matrix makes the convolution operation unconventional. We have developed a precalculation method that involves a cross product of two matrices: the flipped version of the original image is the first matrix, and the associated second matrix shown in Figure 2 is composed of +1 and -1. However, the resultant matrix of this precalculation (or cross product of two matrices) must be held in the computer memory to facilitate the computation for forward convolution and the corresponding backpropagation. After precalculation, the size of the intermediate image

is ($k/2\times k/2$) times the original image size. The factor of 1/2× 1/2 is due to the 1/2 down sampling two-dimensionally in a conventional forward wavelet transform.



Vertical operator *Horizontal operator* *Diagonal operator*

Figure 2. Three matrices used for the cross product precalculation.

The CNN Convolution Process with Wavelet Kernels

Since we can combine all four convolution operations by using only one kernel, the wavelet convolution operation can be adapted by the CNN convolution processing described earlier. This decomposition platform is particularly convenient for the CNN backpropagation training. Figure 3 shows the block diagram in four sections of the wavelet decomposition processes for the forward and backpropagation calculation.

The above convolution processing using a wavelet kernel would only replace one out of N feature maps in the hidden layer of Figure 1. To replace all N feature maps, a total of $4N$ channels with N independent wavelet kernels is required. In Figure 3, the updated filter kernel, $h_u'h_v'$, does not guarantee holding the criteria to serve as a low pass filter for a wavelet transform.

The 2-D composed low pass filters h_{LL} in the WBCNN serves the same role as the kernels K in the conventional CNN. To satisfy the criteria of a wavelet transform, the updated low pass filters would require the following conditions to be fulfilled:

(A) using known scale function of wavelet kernels h_u for the initialization of CNN kernels;

(B) the old and new kernel are constrained by : (i) $\sum_u h_{2u}[t] = \sum_u h_{2u+1}[t] = \sqrt{2}/2$ and (ii) $\sum_u h_u[t] \times h_{u+2n}[t] = \delta_{u,u+2n}$ where δ is the Derac delta function and n is an integer. These two constraints ensures the orthogonal property of h_u and g_u filters;

(C) computing new scaling and wavelet functions using the recursive algorithm as indicated by Eqs. (14) and (15) to ensure their existence, otherwise, the new h_u ' in (B) must be modified.

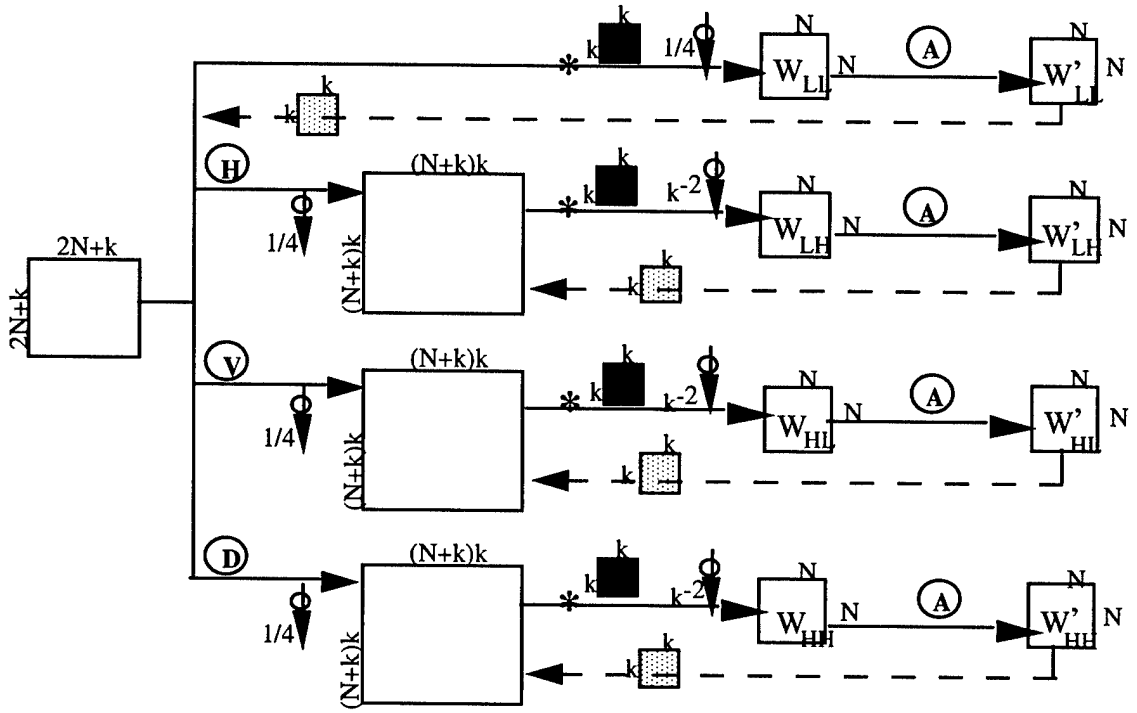
One of the original criteria regarding the so-called "high degree of regularity" was not enforced in the algorithm. The orthonormality of h_u filter may not be self-sustained with each updated version. However, some small modification is possible to make the final version of h_u orthonormal, if the conditions of being a wavelet filter set must be fully met. Nevertheless, the CNN trained filters, which are adaptive versions of wavelet kernels, may have already served as optimal feature extractors in the applications of image pattern recognition.

Using Eq. (10) to update the h_{LL}, which is the low pass filter in 2-D, it would require complicate algorithm development to satisfy the criteria required by the wavelet theory. However, Based on the unified kernel technique with precalculated image $S_O(xk/2,yk/2)$ described earlier, Eq. (2) can be rewritten for updating 2-D wavelet kernel.

$$K_{u,v}[t+1] = K_{u,v}[t] + \eta \sum_i \delta(i) S_0(xk/2+u, yk/2+v) + \alpha \Delta K_{u,v}[t] \qquad ...(18)$$

where index i = 0,1,...$(k\text{-}1)^2$ corresponds to the sub-image of S_O matched to the kernel size. Eq. (18) represents the updated kernel suggested by the BP, these values require a conversion to a new wavelet kernel $h'_u h'_v$. Assuming the wavelet filter is a scale vector (i.e., $h_u h_v = h_v h_u = h_{LL}$, where $u\&v$ = 0,1,2, ... k-1), then only k free parameters ought to be trained for a set of wavelet transform. A solution to satisfy condition (B) and to make $h'_u h'_v$ approximately equal to $K_{u,v}$ is given in Appendix.

Since the decomposed feature maps on the low-low sub-channel have different image characteristics from the others, we free the kernel so that the low-low channel is not constrained by the other three channels. In this way, each set of decomposition has two kernels, one operates on the low-low channel, the other one operates on the remaining three channels. Without the separation of kernels, we found that the neural network had a difficulty in reaching a convergence in the training. In fact, when we used completely uncorrelated wavelet kernels the results were improved as depicted in the experiment.



Definition of symbols

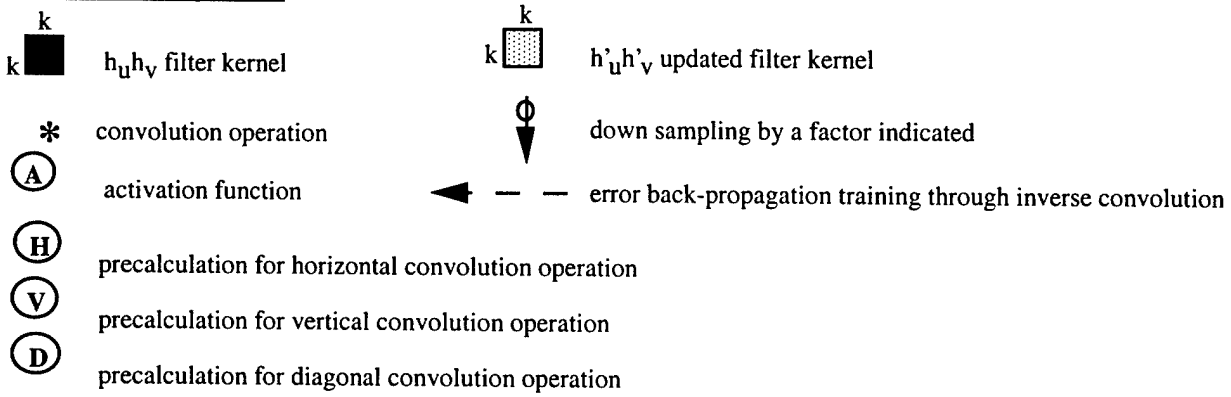


Figure 3. Signal propagation block diagram for a section of the convolution operation using a wavelet kernel in the CNN/WK architecture.

II.C. Classification Invariance of Matrix Operations

It is a good strategy to use invariance or variance characteristics of the system for training of a neural network. In many situations, the orientation of an image pattern does not used as a clinical indication to associate with a disease. Hence, we can take advantage of this characteristic as an invariance. In practice, one can rotate and/or to shift the input matrix and maintain the same output assignments for the training. This method may cause two effects to the neural network: (a) to instruct the neural network that the rotation and shift of input vector would receive the same classification result; and (b) to increase total number of the training samples which is expected to enhance the performance of the neural network. From the point of view of intermediated matrices in the CNN, the convolution

operation maintains the geometrical correspondence of feature maps and the original matrix. In other words, the features maps can be rotated and/or shifted using the same convolution operator on the matrix before or after applying a geometrical function as shown in eqs. (19) and (20).

$$S_x\left((i,j)_\theta;n\right)=\frac{1}{1+\exp\left\{-\sum_m\left[K((u,v);n,m)\otimes S_{x-1}\left((i,j)_\theta;m\right)\right]\right\}} \quad ...(19)$$

or

$$S_x\left((i,j)_\theta;n\right)=S_x\left((i\cos\theta+j\sin\theta,-i\sin\theta+j\cos\theta);n\right). \quad ...(20)$$

In practice, equation (18) is much computationally efficient. Besides, these two equations have different meanings in terms of neural network training. In equation (19), the neural network treats each rotated matrix as a separated matrix. Therefore, the backpropagation computation is always in effects for each epoch. However, the backpropagation would only be taken once for all corresponding rotated/shifted feature maps. Hence, the convolution kernel (or wavelet kernel when wavelet operation is used) receives much more training at the former method than that at the latter.

Rotation may require interpolation which would slightly alter the pixel values and should be acceptable for the input of the CNN. However, the use of shifting can be complicated, because it involves (a) how important the center information for disease patterns are in the neural network learning; and (b) how much shifting can be used without sacrificing critical portions of image information.

III. Application To The Detection Of Microcalcifications

III.A. The wavelet-based convolution neural network

In this initial study, we used only a one-hidden layer structure and eliminated all the complex-cell layers. The hidden layer is composed of two dozen groups of feature maps. We basically replace the convolution kernel with wavelet constrained kernel in Figure 1 in the study. The wavelet kernels to be trained are responsible for the feature extraction from input matrix. The fully connected nets are trained in the same time as the kernels between input and the first hidden layers and are responsible for merging extracted features for classification. Although two output nodes are shown in Figure 1 for generalization of using the system particularly when a fuzzy output training is used. For the crispy training (e.g., (0,1) and (1,0) for false and true cases, respective), only one node is necessary. Because the other node will respond conjugatively. In this study, the crispy training method was adapted and only one node was used.

An image block of $16^{\times}16$ pixels (i.e., $1.7^{\times}1.7$ mm^2) with a convolution kernel size of $6^{\times}6$, which was suggested by the previous study for the detection of microcalcifications [11], was used in this study. The second layer consists of different number of subimages for three different experiments. Each group has $12^{\times}12$ pixels formatted in a square array. The output layer has 2 nodes (groups) which fully connect to the second layer.

In an earlier study, we found that 2-hidden layer architecture of the CNN outperformed than one-hidden layer CNN [10][11]. In this paper, we concentrated on the convolution operation of the CNN. Therefore, only one single hidden layer consisting of groups of feature maps will be used to simplify the study.

III.B. The Experiment for the Detection of Microcalcifications on Mammograms

We have evaluated the CNN and CNN/WK algorithms in the detection of subtle microcalcifications. A total of 68 mammograms (only 38 of them consists of subtle microcalcifications) were digitized by a laser scanner with a pixel size of 0.105 mm. The initial search prior to the final interpretation by the neural network follows the basic scheme which uses background removal and signal extraction methods to pre-scan the mammograms and to extract all possible suspected areas [7], [17], [18]. After the pre-scan process by the computer program, the 68 digital mammograms provide 265 true and 1,821 false subtle microcalcifications. Figure 4 shows some of the suspected regions which may or may not contain microcalcifications. In this study, grouped jack-knife experiment was performed [19]. The training set consists of 15 normal and 19 abnormal cases randomly selected from the database and the testing set consists of the residual 15 normal and 19 abnormal cases. A total of 10 combinations of training and test was performed.

The image blocks of suspected calcification were automatically extracted and were centered. Prior to the CNN process, the background of all the image blocks were removed using a three-level wavelet high-pass filtering technique. Specifically, after extracting each suspected region from the original digital mammogram, a three-level wavelet transform suggested by Daubechies [1] was used and only the lowest frequency was eliminated prior to reconstructing the image block. The high-pass filtered image blocks were used as the input of the CNN. The kernels of CNN/WK were initialized with Daubechies' 8-tap, 6-tap of a composed trigonometrical function [20], or Haar's 4-tap filters. Each

filter was used repetitively with additional 7 times for 48 kernels (i.e., 24 for low and 24 for the 3 high pass filters which produce 96 subimages) and additional 14 times for 120 kernels (i.e., 60 for low and 60 for 3 high pass filters which produce 240 subimages) in the CNN/WK studies. The same filters were also used for 120 fully uncorrelated kernels which possesses only 120 subimages.

The average NODIs of eight rotated image versions were used to evaluate the performance of the neural networks using the LABROC program. One must realize that the detection of clustered microcalcifications is clinically more significant than that of individual calcifications, since the clustered microcalcifications (three or more) are a strong indication to breast carcinoma in radiological diagnosis. Once NODIs were collected, the clinical criterion was added. Hence, the computer program rejected suspected clusters containing only one or two calcifications and calculated the average NODI among the clustered calcifications for the ROC evaluation. The clustering procedure was done by grouping the detected microcalcifications in a $1cm^2$ region of the mammogram. Five ROC curves with different CNN kernels are shown in Figure 5. The syntax of "nK" and "nWK" represent "n" groups of non-constraint kernels and wavelet constraint kernels used in CNN and CNN/WK experiments, respectively. The A_Zs of original CNN and newly developed CNN/WK were 0.91 and 0.83, respectively, by using 24 initial kernels. However, the results of A_Zs were 0.89 and 0.90 with the CNN and CNN/WK respectively, by using 60 initial kernels. Both experiments were using the same wavelet kernel for high frequency components. When all wavelet kernels were uncorrelated, the results were further improved to Az = 0.93 using 120 uncorrelated kernels (CNN/120UWK, 30 initial kernels).

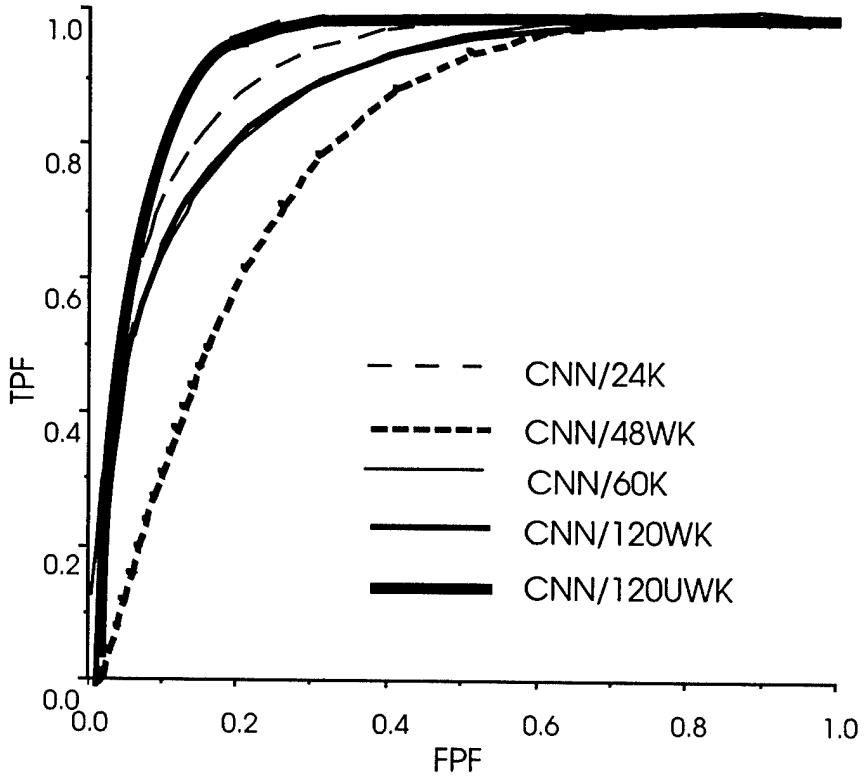


Figure 5. Four ROC curves represent different performance of convolution neural networks in the detection of clustered microcalcifications. Note that CNN/nK and CNN/2nWK have compatible number of hidden nodes and nets in the CNN.
(a) CNN/24K: $A_Z = 0.91$; (b) CNN/48WK: $A_Z = 0.83$; (c) CNN/60K: $A_Z = 0.89$;
(d) CNN/120WK: $A_Z = 0.90$; and (e)CNN/120UWK: $A_Z = 0.93$.

IV. Conclusions

In this experiment, the A_Z of the CNN/24WK was 0.83 which was lower than 0.91 of the CNN/24K. However, the A_Z was greatly improved to 0.93 when 80 uncorrelated wavelet kernels was used. This maybe because only an average of approximate eight free parameters were available in each kernel of the CNN/WK. On the other hand, the CNN had 6× 6 (or 36) free parameters in each kernel. We found that 48 groups for CNN/WK (which has compatible number of hidden nodes as CNN/24K) were not sufficient to extract necessary features for classification. When 120 kernels of

kernels were used, the CNN/WK would have sufficient free parameters in which lead to a higher ROC performance. We further discovered that it is important to uncorrelated the kernels to increase the number of free parameters which seems to be an essential factor to improve the performance of the convolution neural network.

Acknowledgments

This work was supported by a US Army Research Grant #DAMD17-93-J-3007. The content of this paper does not necessarily reflect the position or the policy of the government, and no official endorsement should be inferred. LABROC program was written by C. Metz and his colleagues at the University of Chicago. The authors are also grateful to Ms. Susan Kirby for her editorial assistance.

Appendix

As indicated in Eq. (18), the updated weights, $K_{u,v}[t+1]$ or $K'_{u,v}$, of the kernel suggested by the BP at $t+1$ training iteration are independent. To properly use this suggestion for making a new wavelet kernel, let's assume that there exists a set of h'_u so that both summations of even elements as well as odd elements of the new filter vanish when subtracting $\sqrt{2}/2$ from each of them.

$$g_1(h'_u) = \sum_u h'_{2u} - \sqrt{2}/2 = 0; \quad \text{...(A1)}$$

$$g_2(h'_u) = \sum_u h'_{2u+1} - \sqrt{2}/2 = 0; \quad \text{...(A2)}$$

$$g_p(h'_u) = \sum_u h'_u \times h'_{u+2n} = \delta_{u,u+2n}\ ; \text{ where p = 3,4,... k-3.} \quad \text{...(A3)}$$

In order to update each 2-D filter very close to $K'_{u,v}$, a function based on the square difference is used for the derivation

$$f(h'_u) = \sum_{u,v} (h'_u h'_v - K'_{u,v})^2 . \quad \text{...(A4)}$$

Here we intend to minimize function, f, subject to the constraints indicated in Eqs. (A1), (A2), and (A3). Lagrangian multiplier method can be employed to solve this problem by adding df and multiples of dg_1, and dg_2 to obtain

$$df(h'_u) + \sum_q \lambda_q dg_q(h'_u) = 0 \quad \text{...(A5)}$$

where d represents the differentiation operation of a function; λ_q Lagrangian multipliers. The partial differentiation form of (A5) is given below

$$\frac{\partial f(h'_u) + \sum_q \lambda_q \partial g_q(h'_u)}{\partial (h'_u)} = 0 \quad \text{for u=0,1,2,... k-1.} \quad \text{...(A6)}$$

Eqs. (A5), (A7), (A1), and (A2) represents a set of $k+2$ equations which can solve $2k+3$ unknowns, λ_q, and h_u. In this case, all λ_q need not be determined.

References

[1] I. Daubechies, "Orthonormal based of compactly supported wavelets", Comm. on Pure and Appl. Math., Vol. XLI, pp. 909-996, 1988.

[2] S. Mallat, "A theory For multiresolution signal decomposition: The wavelet representation", IEEE Trans. Pat. Anal. Mach. Intel., Vol. 11-7, pp. 674-693, 1989.

[3] D.E. Rumelhart, G.E, Hinton, and R. J. Williams, "Learning internal representations by error propagation, in Rumelhart & McClelland (Eds.), *Parallel distributed processing:* Explorations in the microstructure of cognition. Vol. 1: Foundation. MIT Press 1986.

[4] DARPA, DARPA Neural network Study, 1988, Washington, AFCEA Press.

[5] K. Fukushima, S. Miyake, and T. Ito, "Neocognitron: A neural network model for a mechanism of visual pattern recognition. IEEE Trans. on Systems, Man, and Cybernetics, vol. 13-5, pp. 826-834, 1983.

[6] K. Doi, "Feasibility of computer-aided diagnosis in digital radiography," Jap J Radio Tech. 45:653-663, 1989.

[7] K. Doi , M. L. Giger ML, H. MacMahon, et al. "Clinical radiology and computer-aided diagnosis: Potential partner in medical diagnosis?" RSNA Scientific Exhibit, Space 129, 1990.

[8] H.P. Chan, K. Doi, S. Galhotra, C. J. Vyborny, H. MacMahon, P. M. Jokich, "Image feature analysis and computer-aided diagnosis in digital radiography. 1. Automated detection of microcalcifications in

mammography", Med. Phys., vol. 14, pp. 538, 1987.

[9] M. L. Giger, N. Ahn, K. Doi, H. MacMahon, and C. E. Metz, "Computerized detection of pulmonary nodules in digital chest images: Use of morphological filters in reducing false-positive detections," Med Phys Vol. 17-5, pp. 861-865, 1990.

[10] S. C. Lo, J.S . Lin, M. T. Freedman, and S. K. Mun, "Computer-assisted diagnosis of lung nodule detection using artificial convolution neural network," SPIE Proc. Medical Imaging, Vol. 1898, pp. 1993.

[11] S. C. Lo, H. P. Chan, J.S . Lin, M. T. Freedman, and S. K. Mun, "Artificial convolution neural network for medical image pattern recognition," Neural Networks, vol. 8-7, pp.1201-1214, 1995.

[12] H. P. Chan, S. C. Lo, B. Sahiner, K. L., Lam, and M. A., Helvie, "Computer-Aided Detection of Mammographic Microcalcifications," Medical Physics, Vol . 22-10, pp. 1555-1567, 1995.

[13] S. C. Lo, S. L. Lou, J. S. Lin, M. T. Freedman, M. V. Chien, and S. K. Mun, "Artificial Convolution Neural Network Techniques and Applications for Lung Nodule Detection," IEEE Trans. on Med. Img., Vol. 14-4, pp. 711-718, 1995.

[14] C. E. Metz, B. A. Herman, and J. H. Shen, "Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data," Stat Med., vol. 17, pp. 1033-1053, 1998.

[15] J. A. Swets and P. M. Pickett, *Evaluation of Diagnostic Systems*, Academic Press, New York, 1982.

[16] S. Mallat, "Multifrequency channel decompositions of images and wavelet models", IEEE Trans. Acoustics, Speech, Sig. Proc., Vol. 37-12, pp. 2091-2110, 1989.

[17] H. P. Chan, K. Doi, C. J. Vyborny, et al, "Improvement in radiologists' detection of clustered microcalcifications on mammograms: The potential of computer-aided diagnosis," Invest Radio vol. 25, pp. 1102-1110, 1990.

[18] A. Hasegawa, Y. C. Wu, M. T. Freedman, S. K. Mun, "Adaptive-size neural-network-based computer-aided diagnosis of microcalcifications," SPIE Proc. Medical Imaging, Vol. 2434, pp. 557-562, 1995.

[19] K. Fukunaga and R. R. Hayes, "Effects of sample size in classifier design," IEEE Trans. Pattern Anal. Machine Intell., vol. 11, pp.873-885, 1989.

[20] M. A. Cody, "The fast wavelet transform," Dr. Dobb's Journal, pp. 16-28, April 1992.

Classification of false positive findings on computer aided detection of breast microcalcifications

Matthew Freedman[1], Shih-Chung B. Lo, Dorothy Steller Artz, Ivan Lau, Seong Ki Mun.

ISIS Center, Department of Radiology, Georgetown University Medical Center, Washington DC 20007

Introduction

False positive detections of microcalcifications by computer aided diagnosis (CADx) systems are a distraction to the radiologist and raise questions as to the eventual clinical utility of computer aided diagnosis systems. We have carefully analyzed the mammographic findings that appear in the locations of CADx detections and have counted and classified them.

Methods

Two different series were run representing two different settings of the CADx algorithm. In Series 1, 200 mammogram images were analyzed. In Series 2, 95 mammogram images were analyzed. The settings for the algorithm were changed between Series 1 and 2. In Series 1, the parameters were set to detect a minimum of three suspected microcalcification foci with an average CNN output value of 0.7. In Series 2, we set the algorithm to detect a minimum of four suspected foci of microcalcification with an average CNN output value of 0.8 as the threshold.

Abnormalities seen at the sites of CADx localization were classified as representing artifacts, true positive findings and false negative findings. We included normal non-calcified punctate anatomic structures as artifacts in this analysis.

The CADx program was run prior to the mammograms being interpreted by the radiologist. Cases were selected from the clinical cases of the breast cancer screening service. Case selection required that each patient have both a current and prior study and to have images of both breasts. Once these criteria were met, the cases were assigned or not assigned to the CADx group by selecting every other case. Cases were digitized at 100 microns using a Lumiscan 150 film scanner (Lumisys, Sunnyvale, CA). They were then processed by the CADx program and the results returned later that day to the radiologist for assessment. The radiologist, who by then had interpreted the mammograms for the official clinical report, proceeded to review the CADx findings and classify any identified abnormalities based on examination of the original mammography film with a 2 X and 5 X magnifying lens. Only one indeterminate cluster of microcalcifications was detected by the CADx program that had not been detected by (in this case) either of the two radiologists who had interpreted the study. The cluster was stable and had been missed by both the radiologist initially interpreting the older study and the radiologist interpreting the newer study. Because it was stable, no additional evaluation was done of this cluster.

In many of the sites identified by the CADx algorithm, there was more than one finding that could have resulted in the CADx detection. We chose to code these findings separately. Because of this, there are many more false positive detections indicated than the number of false positives per image would suggest. We cannot assess the effect of multiple artifacts or combinations of true microcalcifications and artifacts in the performance of the CADx program, and so we chose to record all findings. In assessing the number of false positive detections per image, we looked at each site that was recorded. If at least one microcalcification was present along with the non-calcium structures, we graded that as a true detection for calcifications in determining the number of false positives per image. We did separate calculations for the number of false positives per image using the criteria of 1 or more and 2 or more microcalcifications in the identified field.

[1] Suite 603, 2115 Wisconsin Avenue NW, Washington, DC 20007

Findings

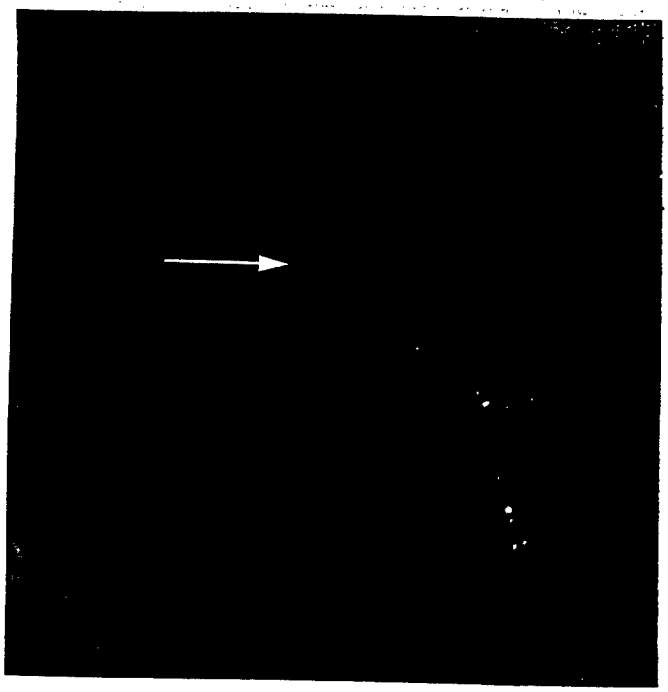
True negatives, true positives, false negatives, false positives

True negatives

In Series 1, 44% of the mammogram films had no CADx detections and no clusters of calcifications were seen when the radiologist re-assessed the film.. In Series 2, 31% of the mammogram films had no CADx detections and no clusters of calcifications on film re-assessment.

True positives

True positives as defined in this study, were detections with one or more small benign calcifications or indeterminate microcalcifications. In this series, the true positive detection rate was 86% in Series 1 and 94% in Series 2 when measured against a single radiologists interpretation of the mammographic images with the CADx output and when using the presence of at least one microcalcification as a true positive detection. Overall, because we recorded separately each finding in a location identified by the CADx program, 29% of the details found in regions identified by the CADx program in Series 1 and 27 % of the detections in Series two were true positives. Vascular calcifications were considered to be false positives. (Figure 1 demonstrates a true positive detection)



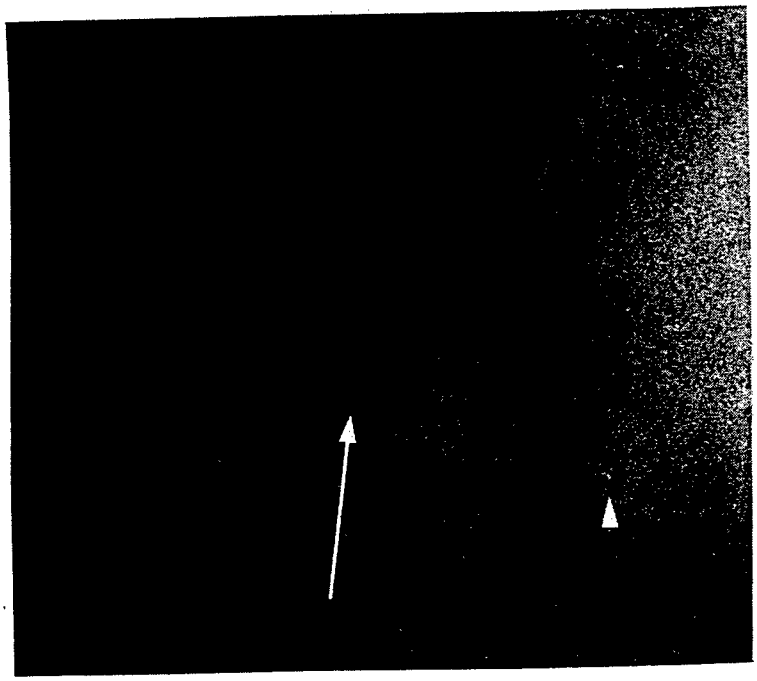
When tested previously with a proven set of cases, the CADx algorithm performance was 87% true positive detection rate at 0.5 false positive clusters per image. (Lo SCB, Chan HP, Lin JS, et al. Artificial convolution neural network for medical image pattern recognition. Neural Networks. 1995. 8:1201-1214.)

True positive and true negative findings combined

If one combines the true negatives and true positive cases, 73% of the mammogram films in Series 1 and 58% of the films in Series 2 were correctly classified.

False negatives

False negative detections were defined as cases in which a benign or indeterminate cluster of microcalcifications were present on the mammogram film, but was not detected by the CADx algorithm. False negative results were seen in 8% of films in Series 1 and 3% of films in Series 2. (Figure 2 demonstrates a false negative cluster of microcalcifications (arrow) next to a true positive calcification cluster (arrowhead).)



False positive detections

False positive detections accounted for 71% of the details recorded in Series 1 and 73% of the details in Series 2. As previously stated a false positive location could have multiple details within it that could explain the detection and each was recorded separately.

False positive detections per image

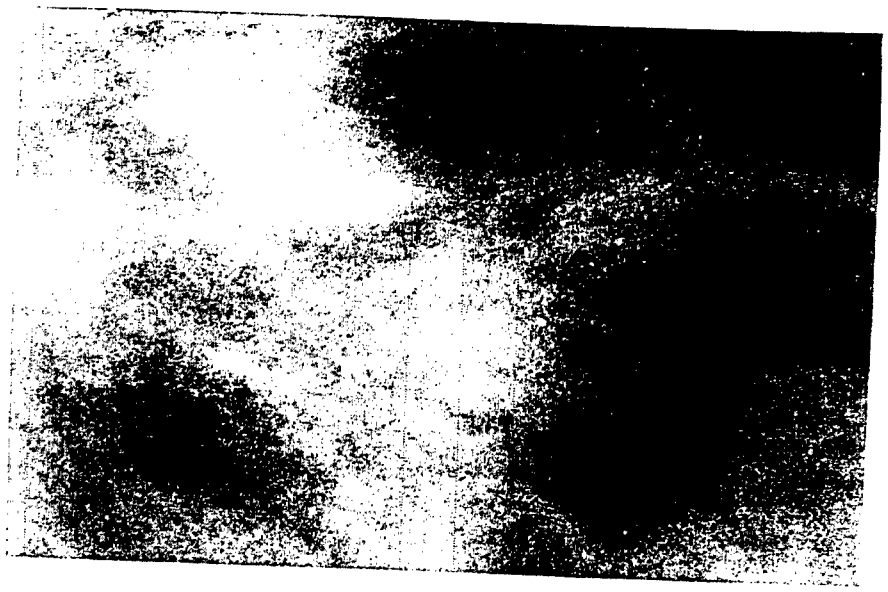
In recording the number of artifacts, we recorded separately each of the types of artifacts found in any CADx defined area of abnormality. In assessing the number of false positives per image, we accepted any CADx identified location as being a true positive if one or more true microcalcifications were present at the same site. A false positive was a location indicated by the CADx program in which calcifications were not present. Using these criteria, in Series 1 there were 0.7 false positive detections per image and in Series 2 there were 0.9 false positive detections per image.

If one uses the criteria of two or more calcifications for a true positive detection, in Series 1, the false positive rate was 0.8 per image and for Series 2, the rate was 1.0 false positive detections per image.

Detailed analysis of false positive detections

Developer artifact

Developer artifact accounted for 2% of detections in Series 2. This is demonstrated in Figure 3. Developer sediment defects are small punctate details often with a halo around them.



Punctate densities

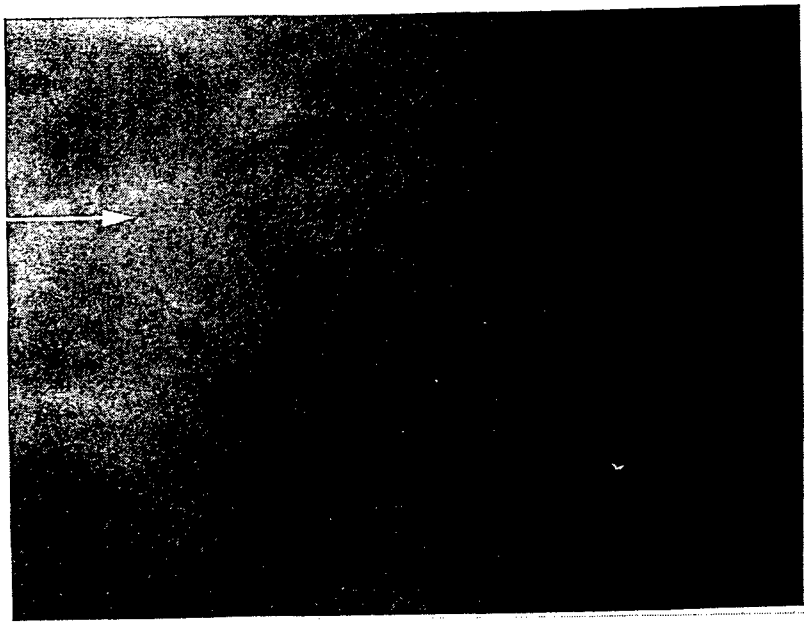
Punctate densities are small point like objects seen in many mammograms. Their exact nature is uncertain, but they are small and slightly radiodense like faint calcifications. They may represent dilated terminal glandular elements or dilated ducts. They have the size and distribution of small benign calcifications, but have a lesser radiodensity. In some cases we could not determine whether small calcifications were present within them. They accounted for 21% of the detections in Series 1 and 37% in Series 2. They are demonstrated in Figure 4 with arrows pointing to some of them. Punctate densities are different from film granularity, but because of the magnification used for this prints, film granularity is seen. The algorithm did not seem to be detecting regions of film granularity as these were widely present and the algorithm only selected a few regions where other findings were present as well.





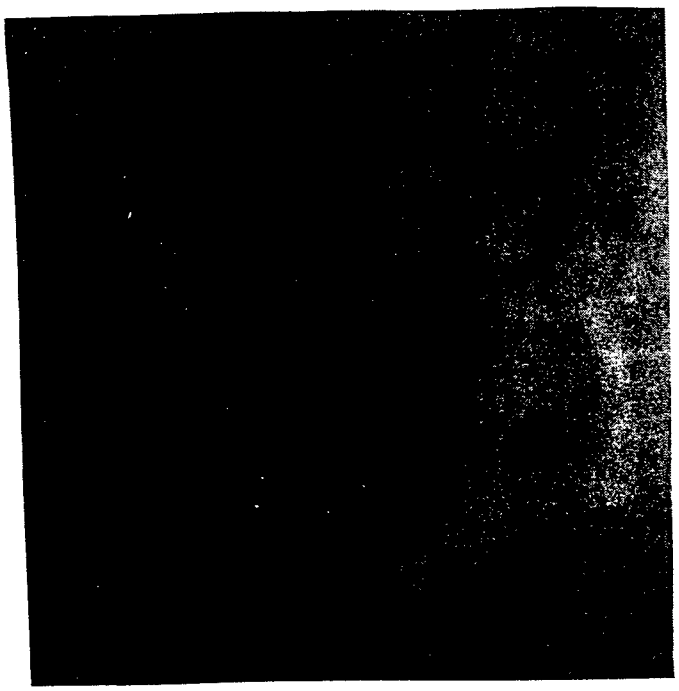
Grid artifacts

Short parallel lines with more punctate regions of increased optical density are seen in some regions of false positive detections. They are considered to represent grid artifact from failure of rapid enough movement of the grid. They were more common in the older mammograms (1994-5) than in the 1996 mammograms. They accounted for 13% of the false positive detections in Series 1 and 5% in Series 2. They are demonstrated in Figure 5 and localized with arrows.



Film defects

Film defects (processor pick defects) are commonly seen in mammogram films. They were 28% of the false positive detections in Series 1 and 25% of those in Series 2. They are demonstrated in Figure 6.



Fibers

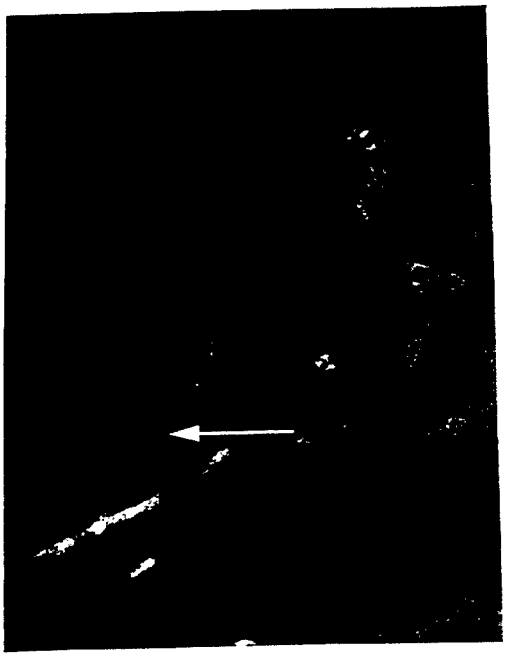
Fibers are short lines, either straight or curved, thought to represent the fibrous structure or ductal structure of the breast. They are similar to the punctate densities and there exact nature is uncertain. They accounted for 6% of the false positive detections in Series 1 and 1% of those in Series 2. They are demonstrated in Figure 7 and localized with arrows.



Vascular calcification

Vascular calcification accounted for 4% of the false positive detections in Series 1 and 2% of the false positive detections in Series 2. They are demonstrated in Figure 8 (arrow). The algorithm did not detect long linear calcifications, but detected vascular calcifications when they were punctate.





Deodorant

Small white flecks of calcium like material were seen in the axilla of some patients. The frequency of this was not recorded, but the algorithm did identify this finding. This is demonstrated in figure 9 (arrows).



Discussion

False positive detections by computer aided diagnosis programs are a considered to be undesirable. There is a tradeoff between the true positive detection rate and the false positive detection rate. Our goal in this project was to determine by careful analysis the several causes of false positive detections in our algorithm. The results show that image quality of the original mammogram is an important determinate of the number of false positives. Image quality defects accounted for 41% of the false positive detections in Series 1 and 32% in Series 2. These false positive findings should decrease with careful attention to film quality. Most of the remaining false positive detections were caused by normal punctate structures in the normal breast. These punctate defects resemble the appearance of very small calcification clusters, but have a lower radiodensity. Their size and density appear to represent a continuum of increasing size and density that appears to merge with true (benign) microcalcification appearance. We believe these represent dilated glands or ducts. In some cases we were uncertain whether or not calcifications in dilated ducts were present or if we were seeing the dilated duct itself. Previous reports on the accuracy of CADx programs for microcalcifications have indicated that vascular calcifications and film emulsion defects were important contributors to CADx false positives. In this study, we have identified several additional causes of CADx false positive detections.

Conclusion

False positive detections in computer aided microcalcification programs are not random responses of the computer algorithm to unknown features. Better understanding of their causes should promote algorithm modification. Since the computer algorithm is, in general, responding to true punctate or short linear findings that resemble microcalcifications, this suggests that computer aided systems will function best with high quality artifact free films and that computer detection systems may need to be combined with improved classification systems to decrease the number of false positive detections.

Acknowledgment: This work was supported by the US Army Grant #DAMD17-96-1-6254. The content of this report does not necessarily reflect the position or policy of the US Government and no official endorsement should be inferred.

Detection of Mammographic Masses Using Sector Features With A Multiple Circular Path Neural Network

Shih-Chung B. Lo[a], Huai Li[b], Akira Hasegawa[c], Yue J. Wang[a,d], Matthew T. Freedman[a], and Seong K. Mun[a]

[a]ISIS Center, Radiology Department, Georgetown University Medical Center, Washington, DC.
[b]Odyssey Technologies Inc., Jessup, Maryland.
[c]Tokyo Institute of Technology, Tokyo, Japan.
[d]Department of Electrical Engineering and Computer Sciences, Catholic University of America, Washington, DC.

Abstract

In the clinical course of detecting masses, mammographers usually evaluate the surrounding background of a radiodense when breast cancer is suspected. In this study, we adapted this fundamental concept and computed features of the suspicious region in radial sections. These features were then arranged by circular convolution processes within a neural network, which led to an improvement in detecting mammographic masses.

In this experiment, randomly selected mammograms were processed by morphological enhancement techniques. Radiodense areas were isolated and were delineated using the region growing algorithm with a valley blocking technique. The boundary of each region of interest was then divided into 36 sectors using 36 equi-angle dividers radiated from the center of the area. Four features at each section were computed: (1) the radius, (2) the normal angle of the boundary, (3) the average gradient along the radial direction, and (4) the gray value difference (i.e., contrast) along the radial direction. Hence, 144 computed features (i.e., 4 features per sector for 36 sectors) were used as input values for the newly designed multiple circular path neural network (MCPNN). The neural network is constructed to emphasize on the correlation information associated with the feature interactions within the angle and between adjacent angles.

We have tested this approach on our research database consisting of 91 mammograms. The over-all performance in the detection of masses was 0.78-0.80 for the areas (Az) under the ROC curves using the conventional neural network. However, the performance was improved to Az values of 0.84-0.89 using the multiple circular path neural network.

1. Introduction

It is well known that effective treatment of breast cancer calls for early detection of cancerous lesions (e.g., clustered microcalcifications and masses associated with malignant cellular processes)[1,2,3]. Breast masses appear as areas of increased density on mammograms. It is particularly difficult for radiologists to detect and analyze a suspected area where a mass is overlapped with dense breast tissue. These masses are more readily seen as time progresses, but the further the tumor has progressed, the lower the possibility of a successful treatment. Therefore, increasing the chances of early breast cancer detection in improving today's clinical system is of vital importance to breast cancer patients.

Several research groups have developed computer algorithms for automated detection of mammographic masses[4,5,6,7,8]. At least one of these groups has also attempted to classify the malignant or benign nature of the detected tumors[5]. The results of these detection programs indicate that a high true-positive (TP) rate can be obtained at the expense of 2 or 3 false-positive (FP) detections per mammogram. This FP rate is unacceptably high for mass detection in clinical practice. Mammographically, a multiplicity (more than two) of similar benign-appearing breast lesions argues strongly for benignity[9, 10,11,12] and, indeed, the more masses that are identified, the less chance that they represent cancer[12]. If the computer indicates multiple detections on each mammogram, the radiologist has to seek out the one mass that has mammographic features that differ from the others. The significant lesion may be missed due to the multiplicity of possible lesions. We therefore believe that a more useful and fundamental approach to CADx of masses is to devise computer programs to analyze features of a suspected mass, which are detected by the

radiologist, and provide feature measures and estimates of the likelihood of malignancy by comparing the computer's database. The computer therefore serves as a second opinion and also provides a reproducible and objective evaluation of the mass. With this aid, the radiologist may also increase his/her sensitivity by lowering the threshold of suspicion, while maintaining the overall specificity and reading efficiency.

2. Clinical Background of Breast Lesions and Technical Planning in Mass Detection

2.1 Brief Description of Clinical Background

Most commonly, breast cancer presents as a mass. The same lesion shows a somewhat different picture from one projection to the other. Difficulties in mass detection also vary with the underlying breast parenchyma. In the fatty breast, masses are generally easy to detect. With the dense breast, mass detection is more difficult and auxiliary signs aid this detection. Breasts can contain one, several, or many masses. When there is one mass, the decision process is based on its size, shape, and margins. The larger the mass is and the less well defined its margins, the greater the chance of cancer. When there are several masses, one looks at each, trying to determine whether any has features to suggest cancer (poorly defined, spiculate, unusually radiodense for size) and one also looks to see whether any mass is different in appearance from the others. Multiple small, well-defined, similar masses presenting bilaterally are all likely to be benign. The greater the asymmetry, size, lack of circularity, edge unsharpness, and radiodensity, the more suspicious.

Clinical features of breast masses are further discussed below:

Density - Malignant lesions tend to have greater radiographic density due to high attenuation and less compressibility of cancer than normal tissue. Radiolucent lesions are typically benign and the diagnosis can be made from the mammogram.

Size - If the lesion has morphological features suggesting malignancy, it should be considered suspicious regardless of the size. Isolated masses with non-cystic densities greater than 8 mm in diameter can be malignant. In general, the larger a lesion, the more suspicious it is.

Shape - The more irregular the shape of a lesion, the more likely the possibility of malignancy. Lesions tend to be round, ovoid and/or lobulated. Small and frequent lobulations are suspicious. Lesions in the lateral aspect of the breast near the edge of the parenchyma with a reniform shape and a hilar indentation or notch usually represent a benign intramammary lymph node. Breast carcinoma hidden in the dense tissues can cause parenchymal retraction, which possess different shapes.

Margins -The margins of the lesion should be carefully evaluated for areas of spiculation, stellate patterns or ill-defined regions. Most breast cancers have ill-defined margins secondary to tumor infiltration and associated fibrosis. The appearance of spiculations and a more diffuse stellate pattern are almost pathognomonic for cancer. Lesions with sharply defined margins have a high likelihood of being benign; however, up to 7% of malignant lesions can be well circumscribed.

Recently these clinical features have been adapted in a standard of the American College of Radiology (ACR). This diagnosis standard is known as "Breast Imaging - Reporting and Data System" (BI-RAD)[13].

2.2. Technical Planning

In this study, our goal was to extract clinically suspicious lesions. The differentiation of benign and malignant status was beyond the scope of this work. Hence, we will only provide methods in extracting potential lesions from glandular tissue in the following sections. (Note that lesions can be overlapped with dense breast parenchyma.) The study was conducted with the following steps: (1) use background correction method and morphological operations to extract radio-opaque areas, (2) delineate the boundary of the areas, (3) compute the features and texture of the masses with emphasis on the boundary, (4) design and plan training strategy using a neural network as classifier for the recognition of mass features. An overall detection scheme of our proposed framework is shown in Figure 1.

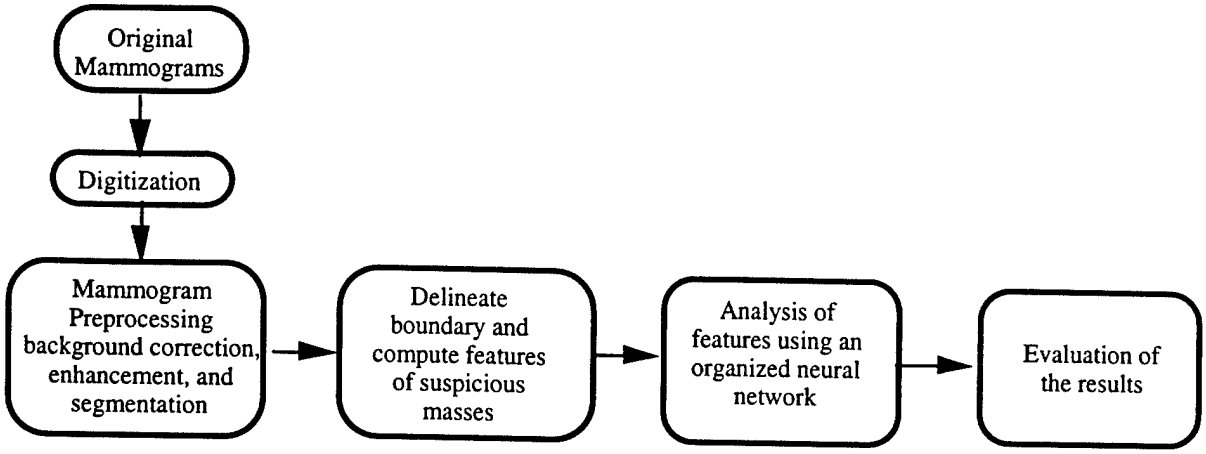


Figure 1. A flowchart for the detection of masses in this study.

3. Development of Technical Methods

3.1. Preprocessing for Image Consistency and Mass Enhancement Using Morphological Operations

One of the main difficulties in automatic mass-detection is that mammographic masses are often overlapped with breast tissues. In such cases, it is necessary to remove bright background caused by breast tissues but to keep mass-signals. For this purpose, background correction is an indispensable technique for mass detection.

The theory of mathematical morphology is powerful in analyzing and describing geometrical relations. Essentially it is a formalization of intuitive concepts such as size or shape. The two basic morphological operations are "erosion" and "dilation," which are consistently defined for binary and gray-scale images. Using these two basic operations, two other basic and important operators, "opening" and "closing", can be defined as follows:

opening: $$\mathbf{X}_{\mathbf{B}} \equiv (\mathbf{X} \ominus \boldsymbol{B}) \oplus \boldsymbol{B}, \qquad \text{...(1)}$$

closing: $$\boldsymbol{X}^{\boldsymbol{B}} \equiv (\boldsymbol{X} \oplus \boldsymbol{B}) \ominus \boldsymbol{B}, \qquad \text{...(2)}$$

where $\boldsymbol{X}$ indicates the original image, $\boldsymbol{B}$ represents the structuring element, and $\oplus$ and $\ominus$ indicate the operations "dilation" and "erosion," respectively. Based on the "opening" operation, we have developed an operation for background correction. The operation is represented by

$$\boldsymbol{X} - \boldsymbol{X}_{\boldsymbol{B}} = \boldsymbol{X} - (\boldsymbol{X} \ominus \boldsymbol{B}) \oplus \boldsymbol{B}. \qquad \text{...(3)}$$

This equation represents the subtraction of the image processed by the operator "opening" from the original image.

Figure 2 shows the effect of the operation represented by Eq. (3): (a) illustrates a structuring element, (b) shows the original signal (gray line) and the processed signal (black line) by "opening", and (c) denotes the final output signal of the operation indicated by Eq. (3). (c) is the subtraction of the black line signal from the gray line signal in (b). Note that the detected peak signals were not affected by the operation. Hence the mass signals detected by the operation retain their original shapes.

As can be seen in this graph, the size of the detected peak significantly depends on the size of the structuring element. All peaks, which are smaller than the structuring element, can be detected. In our mass detection process, a 52 pixel-diameter structuring element will be used to detect masses whose sizes are less than 52 pixels in diameter. An object with a diameter of 52 pixels in a 512×625 pixel reduced image occupies 250 pixels in its original digitized image, and its real size is expected to be about 2.5 cm.

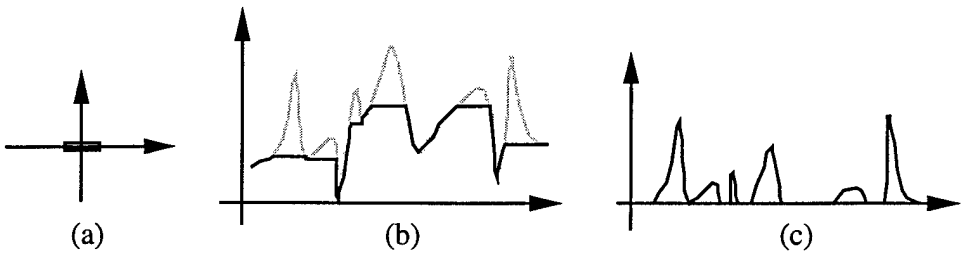


Figure 2. Effect of operation in Eq. (3): (a) structuring element, (b) original signal (gray line) and signal after opening (black line), and (c) output signal of operation in Eq. (3).

3.2. Feature Extraction of Masses

Feature extraction methods have played essential roles in many pattern recognition tasks. Once the features associated with an image pattern are extracted accurately, they can be used to distinguish one class of patterns from the others. Recently, many investigators have found that the multilayer perceptron neural network using the error back propagation training technique is a very powerful tool to serve as an analyzer (or classifier). Recently, the back propagation neural network (BPNN) for classification of features has widely been used in the field of computer-aided diagnosis[14,15,16,17,18].

The success of using an analyzer for a pattern recognition task would rely on two issues: (a) selected features that could describe discrepancy between patterns and (b) accuracy of the feature computation. Should either one fail, no analyzer or classifier would be able to achieve the expected performance. By analyzing many clinical samples of various sizes of masses, we found that the peripheral portion of the mass plays an important role for mammographers to make a diagnosis. The mammographer usually evaluates the surrounding background of a radiodense area when breast cancer is suspected.

We, therefore, performed boundary detection of the suspected masses on the morphologically enhanced mammogram. A region growing with valley blocking technique was employed to delineate all the suspected areas. Then, the boundary was divided into 36 sectors (i.e., 10^{o} per sector) using 36 equi-angle dividers radiated from the center of suspicious area. The following features were computed within each 10^{o} sector of the area:

(a) "l" - the length from the center of mass to the shortest boundary segment.
(b) "a" - the normal angle of the boundary segment (or the value of cos(a)).
(c) "g" - the average gradient of gray value on the segment along the radial direction. Technically speaking, this set of gradient values may also serve as a fuzzy system for the input layer in the neural network to be described.
(d) "c" - the gray value difference (i.e., contrast) along the radial direction. (averaged gray value (h_i) calculated from the mass area located at "l"/3 inside the boundary and the average background value (b_o) calculated from the peripheral area near "l"/3 outside of the suspicious area).

Hence, a total of 144 computed features (4 features/sector for 36 sectors) can be used as input values for the analysis of suspicious areas. The relationship between the computed features and BI-RADS descriptors are discussed below:

(1) Mass Size -
The 36 "l" values would provide sufficient data for the neural network to determine the size.
(2) Mass Shape (round, oval, lobulated, or irregular) -
The 36 "l" and 36 "a" values could approximate the shape of a mass.
(3) Mass Margin (circumscribed, microlobulated, obscured, ill-defined, or spiculate) -
The 36 "g" and 36 "l" values should be able to describe the characteristics of the mass margin.
(4) Mass Density (fat-containing, low density, isodense, or highly dense) -
The 36 "c" and 36 "g" values would be able to describe the density of the mass.

In short, the selected features are greatly matched with the BI-RADS descriptors. The reason for using 36 values for each nominated feature is four-fold: (a) mass boundary varies, it is difficult to describe an image pattern using a single value; (b) due to the general shape of the masses, the features of masses can be easily analyzed by the polar coordinate system; (c) in case some features are inaccurately computed in several directions due to the structure noises, such as the breast slender lines, there may still exist a sufficient number of correct features; (d) generally more accurate results can be produced by using subdivided parameters rather than using global parameters in a pattern recognition task. Other computational features (e.g., difference entropy[8] and other higher order features) are eligible but require further investigation.

3.3. The Neural Network Structures Specifically Designed for the Extracted Boundary Features

(A) Multiple path with circular networking to instruct the neural network in analyzing sector features

We designed several neural network connections between the input and the first hidden layers as shown in Figure 3. Figure 3(a), (b), and (c) illustrate the full connection, a self correlation (SC) networking, and a neighborhood correlation (NC) networking, respectively. Note that the input and hidden nodes should be completely matched when combining more than one path in the study. In this case, the correlation layers only function as branch connections between input and hidden layers. When using NC paths, networking engagement within multiple sectors (e.g., 20^{o}, 30^{o}, 40^{o}, and 50^{o} of the neighborhood correlation) can be grouped. The method of using the multiple correlation connections was motivated by our two-dimensional convolution neural network (2-D CNN) research experience where we found that more than 10 multiple convolution kernels were necessary to archive an outstanding neural network performance in the detection of lung nodules and microcalcifications[15].

Compared to 2-D CNN systems, the required computation using 1-D input features (i.e., 144) is relatively small. The combination of the networking paths described earlier for MCPNN was implemented using C programming language. The internal computation algorithm used in the MCPNN shares the same convolution process as that in the 2-D CNN[15]. One additional training method using flipping invariance was employed and is described as follows.

(B) Training methods and the utilization of characteristics of flipping invariance of the features

Because we used the circular paths, there were no starting and ending sectors. The forward and back propagation computation can be started from any sector. Since the mass characteristics of the flipped patch remained the same, we flipped each patch in the training set and kept the same numerical value for the target output. Since we designed a 10^{o} increment for each rotation, each SC or NC networking would need to process through 36 times for the computed feature set for each patch. To simplify this network computation, we shifted one small set (4 nodes) on the input layer a time to conduct the circular convolution process with the SC and NC kernels. By reversing the sequence of the sector, we can train the flipped version of the suspicious masses. Hence, the characteristics of flipping invariance literally increase the number of the training set by a factor of 2. The flipping procedure was also used for the BPNN experiment described below.

3.4. Summary of Feature Extraction Methods and the MCPNN

We have described our approach on the feature extraction, the design of MCPNN, and its corresponding training method. Figure 4 shows a flow diagram of the proposed method. Since the MCP only alters the input data connection from the input to the first hidden layer, any learning algorithm can be applied within the neural network. For simplicity, we used the back propagation algorithm for both the conventional and proposed neural network systems in the following experiments.

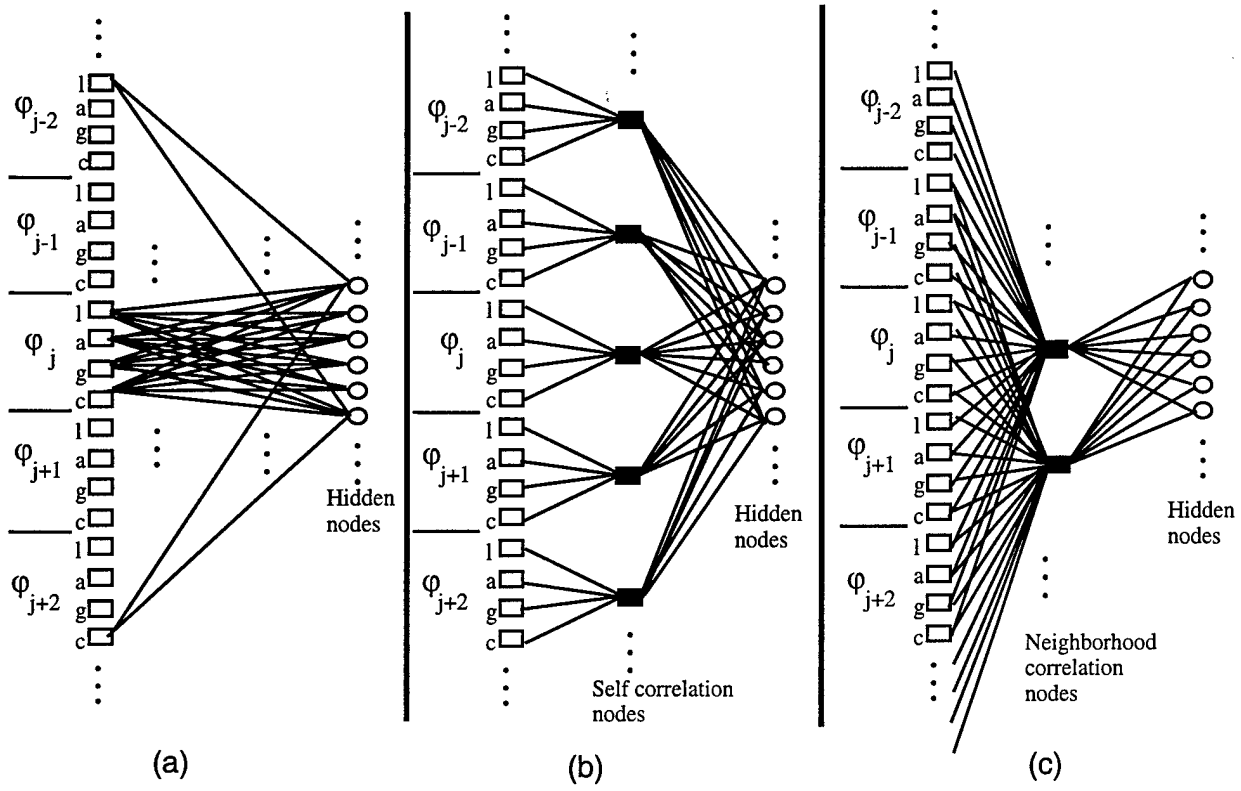


Figure 3. Three types of network paths connecting the input and the hidden layers:
(a) Full connection.
(b) A self correlation (SC) path; each node on the layer connects to a single set of the features (l,a,g,c) for the fan-in and fully connects to the hidden nodes for fan-out.
(c) A neighborhood correlation (NC) path; each node on the layer connects to five adjacent sets of the features for the fan-in and fully connects to the hidden nodes for fan-out.
Note that the fan-in nets emphasizing self correlation in (b) and neighborhood correlation in (c) represent convolution weights (i.e., the same type of sectors possess the same set of weighting factors).

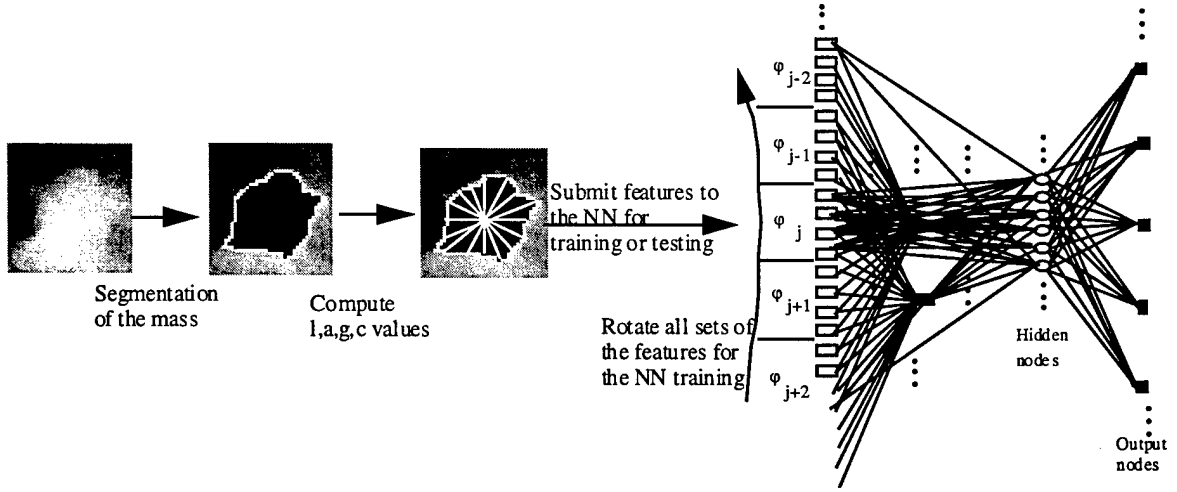


Figure 4. A flow chart, involving the MCPNN and sector features of masses, was used in the following study.

4. Experiments and Results

We selected 91 mammograms and digitized each mammogram with a computer format of 2048× 2500×12 bits (for an 8"×11" area where each image pixel represents 100 μm square). No two mammograms were selected from the same patient film jacket. All the digitized mammograms were miniaturized to 512×625×12 bits using 4×4 pixel averaging and were processed by the above methods to perform mass detection. Based on the corresponding biopsy reports, one experienced radiologist read all 91 mammograms and identified 75 areas containing masses. (Note that the reports recorded the malignancy of the biopsy specimens. The radiologist only used them as reference for the identification of masses.) Through the pre-process and the first step screen based on the circularity test, a total of 125 suspicious areas were extracted from the 91 digitized mammograms.

4.1. Experiment 1

We randomly selected 54 computer-segmented areas where 30 patches were matched with the radiologist's identification and 24 were not. This database was used to train two neural network systems: (1) a conventional 3-layer BP neural network (with 125 nodes in the hidden layer) and (2) the proposed MCP training method using the same neural network learning algorithm. The structure of the MCPNN was described earlier. However, we used one fully connected path, four SC paths, four 20° NC paths, four 30° NC paths, three 40° NC paths, and two 50° NC paths in the first step network connection for the MCPNN. Both neural network systems were trained by the error back propagation algorithm by feeding the features from the input layer and registering the corresponding target value at the output side. Once the training of the neural networks was complete, we then used the remaining 71 computer segmented areas for the testing. None of the images and their corresponding patients in the testing set could be found in the training set. The neural network output values were fed into the LABROC program[19] for the performance evaluation. The results indicated that the areas (Az) under the receiving operator characteristic (ROC) curves were 0.781 and 0.844 using the conventional BPNN and the MCPNN, respectively. The ROC curves of these two neural network training methods are shown in Figure 5(A). We also invited another senior mammographer to conduct an ROC observer study. The mammographer was asked to rate each patch using a numerical scale ranging 0-10 for its likelihood of being a mass. These 71 numbers were also fed into the LABROC program. The mammographer's performance in Az on this set of test cases was 0.909. The corresponding ROC curve is also shown in Figure 5(A).

4.2. Experiment 2

We also conducted a leave-one-case-out experiment using the same database. In this experiment, we used those patches extracted from 90 mammograms for the training and used the patches (most of them are single) extracted from the remaining one mammogram as test objects. The procedure was repeated 91 times to allow every suspicious patch from each mammogram to be tested in the experiment. For each individual suspicious area, the computed features were identical to those used in Experiment 1. Again, both neural network systems were independently evaluated with the same procedure. The results indicated that the Az values were 0.799 and 0.887 using the conventional back propagation neural network and the MCPNN, respectively. Figure 5(B) shows the ROC curves of these two neural network systems using the leave-one-of-out procedure in the experiment.

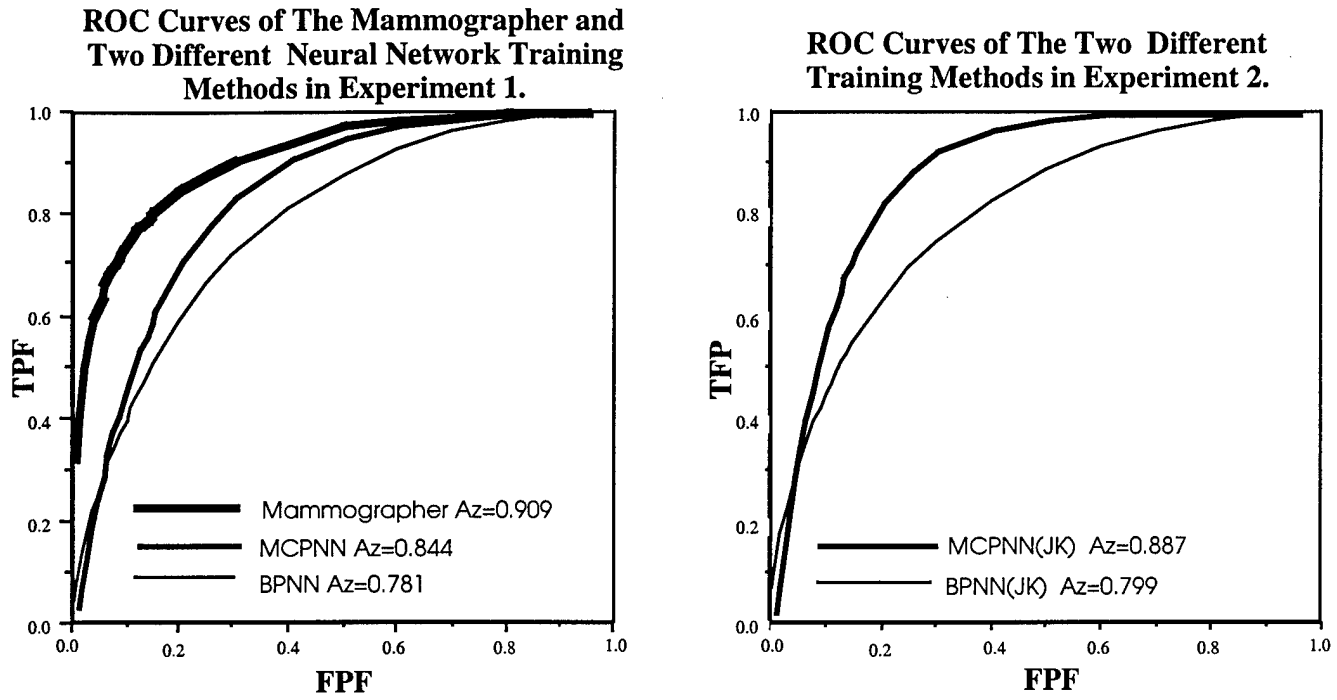


Figure 5. The ROC curves obtained from corresponding experiments.

(A) The left figure shows that the performance of MCPNN training method is superior to that of the conventional input method. The highest curve is the ROC performance of the senior mammographer.

(B) The right figure shows similar results with a higher performance using the leave-one-case-out procedure as described in Experiment 2.

5. Conclusions and Discussion

Through this study, we found that the selected features are somewhat effective in the detection of masses. **These features were "computationally translated" from the qualitative descriptors of BI-RAD. Another uniqueness of this study was on the test of our newly developed MCP training method.** In Experiment 1, we found that the performances of both neural network systems were increased. This might be due to the increased number of cases (from 54 to 124) in the training set. In Experiment 2, the Az value was improved by 0.043 using the MCPNN training method that was higher than Az difference of 0.018 obtained by the conventional training method. The results implied that the MCPNN learned more effectively than the conventional BP when the number of training cases was increased.

It is known in the field of artificial intelligence that the key factors in pattern recognition are: (1) effective methods in the extraction of features and (2) analytic methods (e.g., back propagation neural network) for the extracted features. In this study, we showed that the training method designed to guide the analyzer is also an important factor to a success of a pattern recognition task. Though this finding is not new, the trend of developing training methods for various pattern recognition tasks was not established in the field of pattern recognition. In this work, we demonstrated that organized features with proper network connection and task-oriented guidance would assist the neural network in performing the task.

As far as the research in recognition of masses is concerned, we believe that main concept of using sectors is an effective approach. **Note that any features arranged in the polar coordinate system can be trained by the MCP method.** Since the MCP only coordinates the input data, the internal neural network learning algorithm can be changed to other learning algorithms. A technique using the rubber band straightening transformation, independently developed by Sahnier[20], for the detection of masses also employs a similar concept in extracting feature and/or texture in the polar coordinate. We believe that integration of effective feature and texture values computed at small sectors will be the research trend in mass detection.

Since the mass can be overlapped with glandular tissues, a significant part of the mass may be obscured and is unrecoverable by digital image processing techniques. By reviewing those failure cases, we found that substantial false-negative cases were in this category. However, these cases were correctly identified by the radiologists. This implies that we need to find a way to train the neural network to recognize those cases with sufficient sectors showing signs of masses. Further research based on this pilot study is planned and the results will be reported shortly by the authors.

Acknowledgments

This work was supported by US Army Grant No. DAMD17-96-1-6254 (through a sub-grant from the University of Michigan, Ann Arbor). The content of this paper does not necessarily reflect the position or policy of the government. A part of the database, used in the study, was provided by Dr. Robert Shah of Brooke Army Medical Center. The LABROC program was written by Dr. C.E. Metz and his colleagues at the University of Chicago. The authors are also grateful to Ms. Susan Kirby for her editorial assistance.

References

1. Nystrom L, Rutqvist LE, Wall S, Lindgren A, Lindqvist M, Ryden S, et. al.: Breast cancer screening with mammography: Overview of Swedish randomized trials, Lancet, 1993, vol. 341, pp. 973-978.
2. Shapiro S: Screening- Assessment of current studies, Cancer, 1994, vol. 74, pp.231-238.
3. Tabar L, Fagerberg G, Duffy S, Day NE, Gad A, and Grontoft O: Update of the Swedish two-contry program of mammographic screening for breast cancer, Radiology Clinics of North America: Breast Imaging - Current Status and Future Directions, 1992, vol. 30, pp. 187-210.
4. Lai SM, Li X, Bischof WF: On techniques for detecting circumscribed masses in mammograms, IEEE Trans Med Imaging 1989;8:377.
5. Brzakovic D, Luo XM, Brzakovic P, An approach to automated detection of tumors in mammograms. IEEE Trans Med Imaging 1990;9:233.
6. Yin FF, Giger ML, Doi K, Vyborny CJ, Schmidt RA, Metz CE: Computerized detection and characterization system for use in mammographic screening programs. Radiology 1990; 177(P):245.
7. Sahiner B, Chan HP, Petrick N, Wei D, Helvie MA, Adler DD, and Goodsitt MM: Classification of mass and normal breast tissues: A convolution neural network classifier with spatial domain and texture images, IEEE Trans. on Med. Img., 1996, vol. 15, pp. 598-610.
8. Li H, Lo SC, Wang Y, Freedman MT, and Mun SK: Mammographic mass Detection by Stochastic Modeling and a Multi-Module Neural Network, SPIE Proc. Med Img (in Image processing), 1997, vol. 3034, pp. 480-490.
9. Adler DD: Breast masses: differential diagnosis. In: Feig SA, ed. ARRS categorical course syllabus on breast imaging. Reston, VA: American Roentgen Ray Society, 1988:31.
10. Homer MJ: Imaging features and management of characteristically benign and probably benign breast lesions. Radiol Clin North Am 1987;25:939.
11. Moskowitz M: Circumscribed lesions of the breast. In: Moskowitz M, ed. Diagnostic categorical course in breast imaging. Oak Brook, Ill: Radiological Society of North American, 1986:31.
12. Sickles EA: The rule of multiplicity and the developing density sign. In: Feig SA, ed. ARRS categorical course syllabus on breast imaging. Reston, VA: American Roentgen Ray Society, 1988:177.
13. Breast Imaging – Reporting and Data System, American College of Radiology, Reston, Virginia, 1993.
14. Lo SC, Freedman MT, Lin J, and Mun SK: Automatic lung nodule detection using profile matching and backpropagation neural network techniques, J. of Digital Imaging, vol. 6(1), 1993, pp. 48-54.
15. Lo SC, Chan HP, Lin JS, Li H, Freedman MT, and Mun SK: Artificial convolution neural network for medical image pattern recognition, Neural Networks, 1995, Vol. 8, No. 7/8, pp. 1201-1214. (1)
16. Chan, HP, Lo SC, Sahiner B, Lam KL, and Helvie MA: Computer-aided diagnosis of mammographic microcalcifications: pattern recognition with an artificial neural network, Medical Physics, 1995, vol. 24, No. 10, pp. 1555-1567.
17. Wu Y, Doi K, Giger ML, & Nishikawa RM: Computerized detection of clustered microcalcifications in digital mammograms: Applications of artificial neural networks, Med Phy, vol. 19, 1992, pp. 555-560.

Integer Wavelet Compression Guided by a Computer-Aided Detection System in Mammography

Shih-Chung B. Lo*[a], Erini Makariou[a], Andrzej Delegacz[a], Heang-Ping Chan[b], Donald D. Dorfman[c], Matthew T. Freedman[a], and Kevin Berbaum[c]

[a]ISIS Center, Georgetown University Medical Center, Washington, DC
[b]Radiology Department, the University of Michigan, Ann Arbor, Michigan
[c]Radiology Department, the University of Iowa, Iowa City, Iowa

ABSTRACT

Since an image data compression technique is usually associated with a low-pass filter, the unsharpness of calcifications and edges are of clinical concerns in mammography. The same effect may turn film defects into calcification-like spots and could produce false-positive detection by the radiologist. In this study, we employed a highly sensitive calcification detection system to guide an S+P integer wavelet compression, so that the data fidelity of calcifications or unknown spots are fully preserved. The prediction component of the S+P decomposition is based on Daubechies'D8.

Our results indicated that the modified CAD program detected an average of 1,193 potential calcifications on CC view mammograms and an average of 948 potential calcifications on MLO view mammograms, respectively. Compressed data rates between 0.1 to 0.43 bit/pixel were studied. The compressed images were evaluated by subjective comparison studies. The results indicated that no difference could be observed between the original and the 0.43 bit rate decompressed images. The radiologist identifies 20% of the compressed images at 0.1 bit rate suffering from minor blurry artifacts and 6% of the compressed images possessing greater edge sharpness. Without a lossless compression for microcalcifications, the radiologist identified 20% of the microcalcifications on the compressed mammograms at 0.1 bit rate suffering from minor compression artifacts.

Keywords: Compression, computer-aided diagnosis, wavelet, mammography, microcalcifications, just noticeable difference

1. INTRODUCTION

The recent advancements in high-speed digital computers, networking, as well as the gradual acceptance of high-resolution digital radiographic systems have revived the interest in the development of digital radiography including mammography for routine clinical use. Currently, it is possible to obtain a digital mammogram having high spatial resolution by digitizing screen-film images with a laser digitizer[1, 2, 3] or directly digital systems[4, 5].

The research and development of teleradiology and telemammography systems has progressed through many technical and clinical endeavors [6, 7, 8]. The clinical utilization of teleradiology systems is not known with regard to workloads, reliability, and clinical protocols. The selection of efficient and cost-effective wide-area networks for various applications is presently more an art than a science. In this area, two technical problems remain: (a) no model exists by which radiologists can apply the experience of others to design and implement a teleradiology system; (b) teleradiology systems have not been studied for use in research and education.

Since a large computer space (10 to 40 Mbytes) is required to store a mammogram, it takes a long time for an economical channel to transmit the image. It would take ≈5 hours to transmit an uncompressed mammogram and about 1 hour to transmit the breast area with losslessly compressed data. If an Ethernet is used, the transmission can be increased by a factor of 15 which is still too slow for clinical use. In this study, we used a combined technique that integrates an integer wavelet compression technique and a highly sensitive computer-aided detection (CAD) process to compress digital mammography. The decompressed mammograms would possess error-free at all small bright spot including calcifications.

* lo@isis.imac.georgetown.edu

2. CAD-GUIDED COMPRESSION SCHEME FOR DIGITAL MAMMOGRAPHY

We randomly selected 100 mammograms from our clinical database for this study; of which, 50 were CC view and another 50 were MLO view mammograms. Each of these mammograms contains isolated and/or clustered microcalcifications. Each mammogram was digitized by a Lumisys laser scanner (LumiScan Model 150) at 100 microns per pixel so that each digitized mammogram takes 1,792×2,560×16 bits of a computer space. However, only 12 out of 16 bits were used to store the digital data for each pixel.

As a preprocessing step, the boundary of the breast on the mammogram was first delineated. The compression method and the CAD detection program only applied to the area within the breast boundary. The CAD system developed by the research group at the ISIS Center of Georgetown University Medical Center was used in conjunction with the compression scheme. The CAD system was modified from an existing CAD program to identify calcifications and local maximum on the mammogram. The existing CAD system consists of six major components[9, 10]: (1) delineation of breast area on mammogram, (2) high-pass enhanced filter for reduction of breast parenchyma and enhancement of calcifications, (3) extraction of the local maximum intensity as suspected calcification, (4) computing features (e.g., size, shape, contrast, etc.) on the original mammogram, (5) applying convolution neural network for the recognition of the calcifications using Gaussian as the activation function, and (6) clustering the suspected spots. The first three steps of this detection scheme were adapted to form a highly sensitive CAD system. For each detection, a region of interest (ROI) containing 10×10 pixels centered at the detected spot was the subject for lossless compression separated from the overall compression of the mammogram.

We used an integer wavelet transform (S+P algorithm)[11] to decompose the whole mammogram followed by a linear quantization process and arithmetic coding to encode the quantized wavelet coefficients. The prediction component of the integer wavelet is an approximated version of Daubechies' D8[12]. The suspected calcification spots identified by the CAD system were compressed by the same wavelet without the quantization. Figure 1 illustrates this compression scheme.

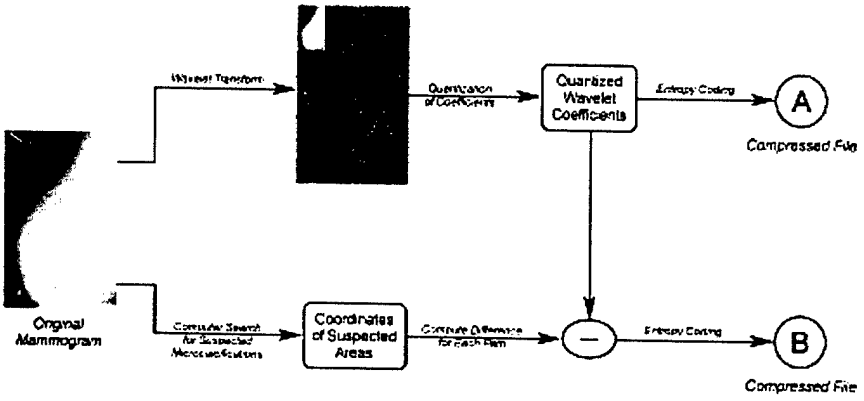


Figure 1: A CAD guided compression scheme based on integer wavelet decomposition.

3. DESCRIPTION OF OBSERVER PERFORMANCE STUDIES

We asked the study radiologist to read a set of images at a time. Each set of images is a pair of the original and one of the three compression modes. The three compression modes are: (i) 0.3 bit/pixel in file A with lossless for suspected calcifications in file B, (ii) 0.1 bit/pixel in file A with lossless for suspected calcifications in file B, and (iii) 0.1 bit/pixel in file A only. A questionnaire consisting of four sections of quality measures was used for each comparison of a pair of images.

Each set of decompressed and original images were displayed on a Compaq computer monitor. The orders on right and left as a pair of images were randomly assigned. Three basic image functions (i.e., window/level, pan, and zoom) were provided for the radiologist to adjust viewing parameters. The radiologist was asked to rate image quality in four sections: (1) calcification observability, (2) edge sharpness, (3) overall image quality, and (4) noise appearance. A four-section questionnaire for each pair of images was used as shown in Figure 2.

Letters "L" and "R" indicate that the left and right sides rank higher for the measured quality, respectively. A non-zero score indicates that one side of the image has either slightly (for L1 or R1), or moderately (for L2 or R2), or significantly (for L3 or R3) better quality or less noise than the other side. A score of "0" indicates that the pair of images has identical image quality or noise appearance. If there is a noticeable difference between images that are scored "0" on the measured quality, the radiologist would check the "yes" box below the "0" score. If a noticeable difference is identified, the radiologist would check one of the lowest boxes that indicates the favored side of the image.

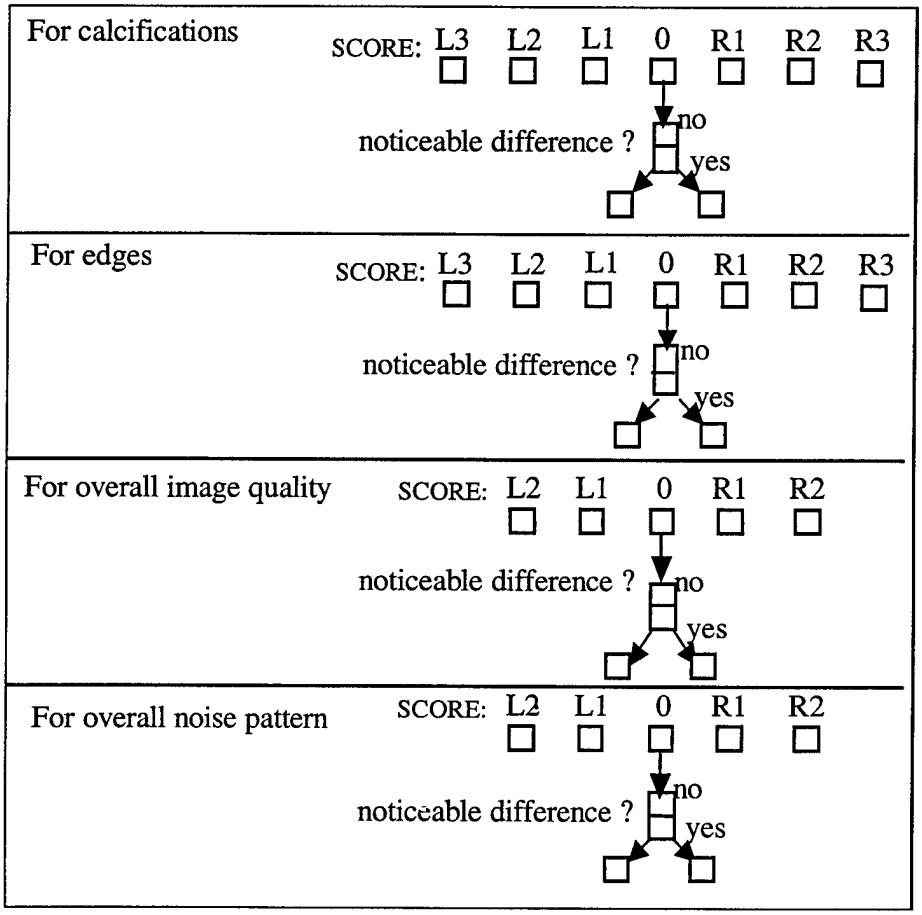


Figure 2: A questionnaire for qualitative measures for a pair of images.

4. RESULTS

In this study, the modified CAD program detected an average of 1,193 bright spots in CC view mammograms and an average of 948 bright spots in MLO view mammograms, respectively. Figures 3 and 4 show two sample mammograms, their compressed counterparts, and the subtracted images amplified by 100 times.

The average compression ratios and computed mean-square-errors are shown in Table I. Table II illustrates the results of the radiologist's qualitative measures by comparing the original and compressed image pairs.

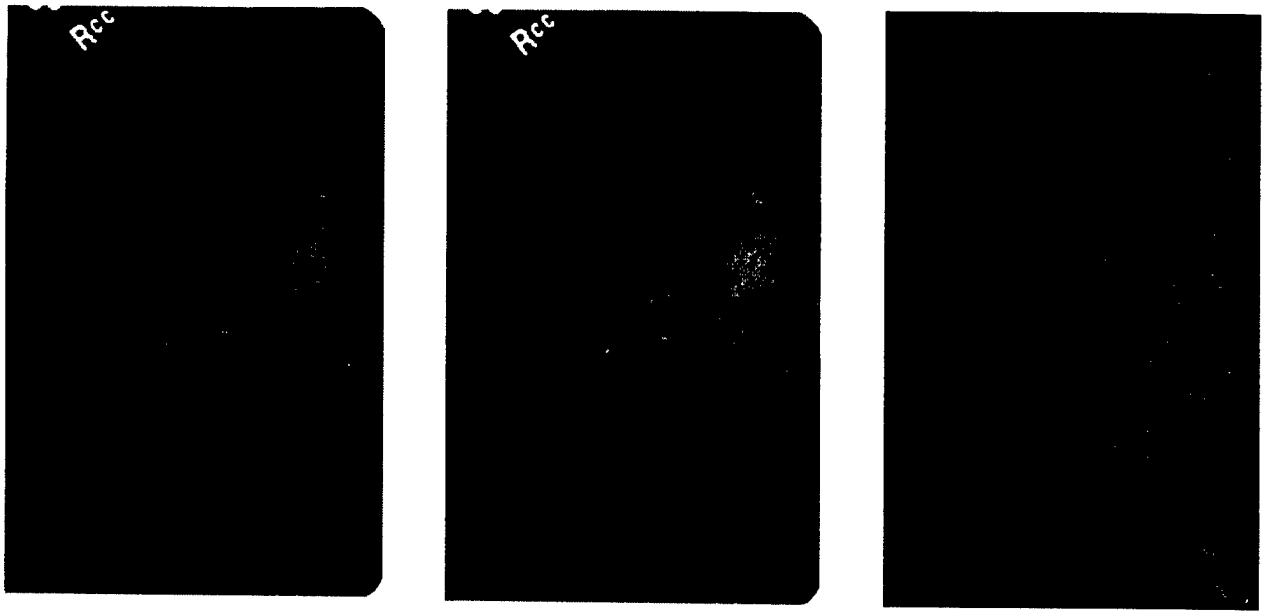


Figure 3: (A) A CC view mammogram, (B) its compressed image at 0.4 bit/pixel, and (C) an enhanced subtraction image resulting from (A)-(B). The uniform squares in (C) result form the lossless compression at the CAD detected areas.

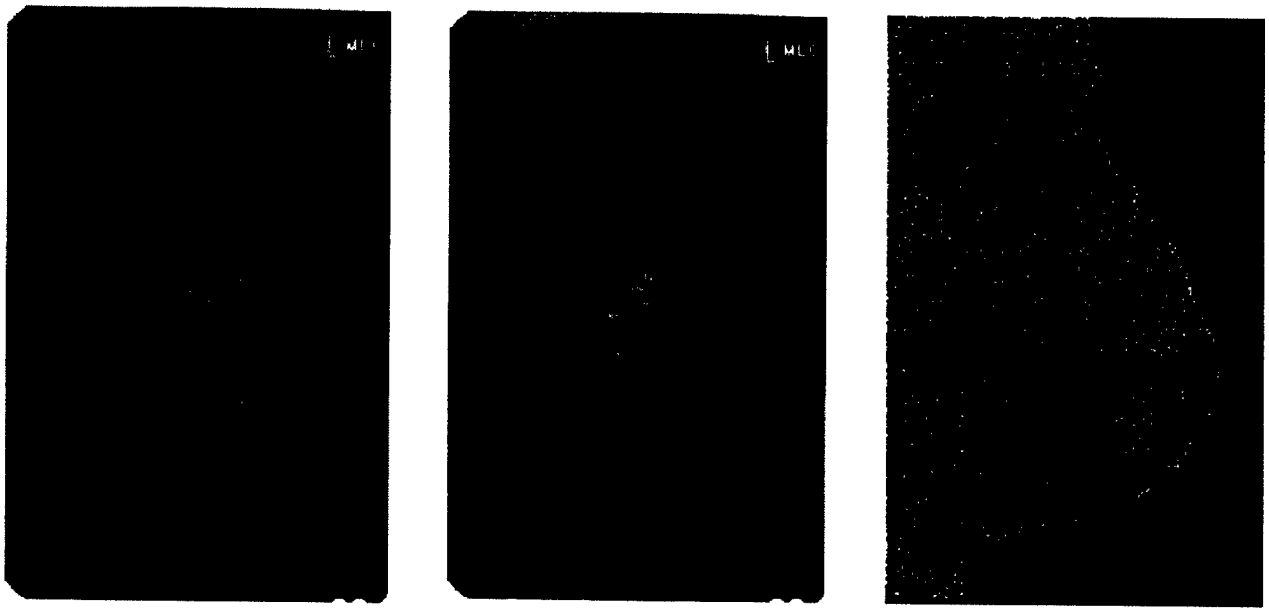


Figure 4: (A) A MLO view mammogram, (B) its compressed image at 0.41 bit/pixel, and (C) an enhanced subtraction image resulting from (A)-(B). The uniform squares in (C) result form the lossless compression at the CAD detected areas.

Table II shows that no difference could be observed between the original and the 0.43 bit rate decompressed images. In fact, it is interesting that the radiologist seemed slightly in favor of the appearances of microcalcifications and edges in the compressed mammograms. The radiologist identified 20% of the compressed images at 0.1 bit rate suffering from minor blurry artifacts and 6% of the compressed images possessing greater edge sharpness. Without using lossless compression for microcalcifications, the radiologist could identify 20% of the less sharp microcalcifications on the compressed mammograms

at 0.1 bit rate. The radiologist also identified that 18% and 6% of the compressed images at 0.1 bit rate possess degraded overall image quality and higher image noise, respectively. Degradation of image quality in compressed images at 0.1 bit rate is highly associated with unsharpness of microcalcifications and edges. The image quality degradation at 0.1 bit rate is also correlated the size of breast area. It is estimated that if the size of breast takes more than one half of the whole mammogram, degraded image quality and edge unsharpness would be observable by the radiologist.

Table I. Compression Ratios and Mean-Square-Errors of the Three Compression Modes.

Mode	A	B	C
Procedure	0.3 bit/pixel + lossless for spots	0.1 bit/pixel + lossless for spots	0.1 bit/pixel
Average Bit Rate	0.43 bit/pixel	0.23 bit/pixel	0.1 bit/pixel
Compression Ratio	27:1	52:1	120:1
Mean Square Error (Standard Devation)	50.73 (36.81)	102.72 (62.48)	105.63 (63.97)

Table II. Qualitative Measures by Comparing the Paired Images (Original and Compressed).

Measurement Category	Micro-Calcifications			Edge Sharpness			Overall Image Quality			Overall Noise Pattern		
	Type A	Type B	Type C	Type A	Type B	Type C	Type A	Type B	Type C	Type A	Type B	Type C
Original Worse Than Compressed	7	4	1	6	2	3	1	0	2	1	0	0
of which:												
- same, but in favor of compressed	3	4	1	5	2	0	0	0	2	1	0	0
- slightly worse	4	0	0	1	0	3	1	0	0	0	0	0
- moderately worse	0	0	0	0	0	0	0	0	0	0	0	0
Original Better Than Compressed	7	5	10	4	11	15	1	7	7	0	2	3
of which:												
- same, but in favor of original	6	1	4	4	6	9	1	3	2	0	2	0
- slightly better	1	4	5	0	5	5	0	4	4	0	0	3
- moderately better	0	0	1	0	0	1	0	0	1	0	0	0
No Difference	36	16	14	40	12	7	48	18	16	49	23	22

Type A - Compression with preservation of suspicious calcifications; Compression rate: 0.43 bit/pixel (0.3+0.13); Total 50 Cases
Type B - Compression with preservation of suspicious calcifications; Compression rate: 0.23 bit/pixel (0.1+0.13); Total 25 Cases
Type C - Global compression; Compression rate: 0.1 bit/pixel; Total 25 Cases

5. CONCLUSIONS AND DISCUSSION

In this study, we show that an advanced image compression method can be integrated with a computer detection technique. Since the computer detection technique can identify potential clinical ROIs, it is appropriate for the compression program treat these ROIs with a different compression strategy. This method is not designed to optimize the compression. It rather is a realistic approach for the clinical usage of the compression technique in digitized or digital mammography. Using this integrated approach, we found that mammograms compressed at 0.43 bit/pixel (i.e., 37:1) contain the same visual quality as the original mammograms. This was confirmed by the above study using just noticeable difference perception study.

ACKNOWLEDGEMENTS

The U.S. Army Medical Research and Materiel Command under DAMD17-96-1-6254 supported this work.

REFERENCES

1. RE. Alvarez, Eds. K. Doi, L Lanzl, PJP.Lin, *Recent Developments in Digital Imaging,* AAPM
2. WR. Brody, *Digital Radiography*, Raven Press, NY, 1984.
3. SC. Lo, P. Butson, JS. Lin, A. Hasegawa, and SK. Mun, "Performance Characteristics of Ultra High-Resolution CCD Film Scanners," *SPIE Proc. Med. Imaging* 1995,
4. M. Sonoda, M. Takano, J. Miyahara, and H. Kato, "Computed Radiography Utilizing Scanning Laser Stimulated Luminance," *Radiology* 1983; 148: 833-838.
5. M. Ishida, H. Kato, K. Doi, and PH. Frank, "Development of a New Digital Radiographic Image Processing System," *Proc SPIE.* vol. 347, 1982, p. 42.
6. RJ. Steckel, "Daily X-Ray Rounds in a Large Teaching Hospital Using High-Resolution Closed-Circuit Television," *Radiology* 1972;105-321.
7. DV. Jelaso, G. Southwonh, LH. Purcell, "Telephone Transmission of Radiographic Images," *Radiology* 1978, vol. 127, pp. 147-149.
8. NJ. Kagetso, DRP. Zulauf, RC. Ablow, "Clinical Trial of Digital Teleradiology in the Practice of Emergency Room Radiology," *Radiology* 1987, vol 165, 551-554.
9. SC. Lo, JS. Lin, H. Li, A. Hasegawa, MT. Freedman, and SK. Mum, "Detection of Subtle Clustered Microcalcifications Using Fuzzy Modeling and Convolution Neural Network," *SPIE Proceedings, Medical Imaging on Image Processing* 1996, vol. 2710, pp. 8-15.
10. SC. Lo, HP. Chan, JS. Lin, H. Li , MT. Freedman, and SK. Mun, "Artificial Convolution Neural Network for Medical Image Pattern Recognition," *Neural Networks* 1995, Vol. 8, No. 7/8, pp. 1201-1214.
11. A. Said, and WA. Pearlman , "A New, Fast, and Efficient Image Codec Based on Set Partitioning in Hierarchical Trees," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 6, 1996, pp. 243-250.
12. SC. Lo, J. Xuan, H. Li, Y. Wang, MT. Freedman, and SK. Mun, "Dyadic Decomposition: A Unified Perspective on Predictive, Subband, and Wavelet Transforms," *SPIE Medical Imaging on Image Processing* 1997, vol. 3031, pp. 286-301.