AD_____

Grant Number DAMD17-94-J-4456

TITLE: Computer-Aided Mammography Using Automated Feature Extraction for the Detection and Diagnosis of Breast Cancer

PRINCIPAL INVESTIGATOR: Joseph Y. Lo, Ph.D.

CONTRACTING ORGANIZATION: Duke University Medical Center
Durham, North Carolina 27710

REPORT DATE: October 1997

TYPE OF REPORT: Final

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

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	October 1997	Final (15 Sep 94 - 14 Sep 97)

4. TITLE AND SUBTITLE
Computer-Aided Mammography Using Automated Feature Extraction for the Detection and Diagnosis of Breast Cancer

5. FUNDING NUMBERS
DAMD17-94-J-4456

6. AUTHOR(S)
Joseph Y. Lo, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Duke University Medical Center
Durham, North Carolina 27710

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

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

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200*

This project explored computer-aided diagnosis (CADx) of breast cancer using artificial neural network (ANN) computer models. The most important accomplishment was the development of the first CADx system that merged radiologist-extracted features to predict not only malignancy but also invasion among breast lesions. We also explored many other projects which contributed to the success of the main project. We identified an optimal subset of input features to the predictive model while maintaining diagnostic accuracy. We also evaluated the feasibility of excluding history findings which may substantially reduce data collection effort and model complexity. We investigated the underlying characteristics of the inputs using self-organizing map ANNs. We also evaluated the behavior of the computer model by analyzing the error surfaces in weight space. In on-going work, we developed image processing techniques for the extraction of mass margins and the detection of microcalcification clusters.

Together, all these projects contributed toward the development of a unified computer-aided diagnosis system for mammography. By providing information which was previously available only through biopsy, the system may assist in surgical planning for patients with breast lesions, and may reduce the cost and morbidity of "unnecessary" surgical biopsies.

14. SUBJECT TERMS	15. NUMBER OF PAGES
Breast Cancer	26
	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

FOREWORD

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

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

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

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

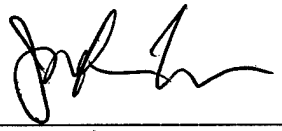
____ In conducting research using animals, the investigator(s) adhered to the "Guide for the Care and Use of Laboratory Animals," prepared by the Committee on Care and Use of Laboratory Animals of the Institute of Laboratory Resources, National Research Council (NIH Publication No. 86-23, Revised 1985).

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

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

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

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



10/15/97

PI - Signature Date

Table of Contents

1. Introduction

1.1 Purpose and Significance

The purpose of this project was to improve the diagnosis and treatment of breast cancer by reducing the cost and morbidity of unnecessary biopsies. Although mammography is very sensitive, there are a large number of false-positive biopsies. Of women with radiographically-suspicious, nonpalpable lesions who are sent to biopsy, only 15 to 34% actually have a malignancy by histologic diagnosis [1, 2]. The cost of these unnecessary biopsies is a major obstacle to widespread acceptance of mammographic screening [3]. It has been shown that surgeon's fees and biopsy costs account for over half the cost of detecting small breast cancers in a screening population [4]. Preventing unnecessary biopsies is therefore one of the most important ways to improve the efficacy of mammographic screening. Many previous reports have discussed the need to reduce the number of benign biopsies [5, 6].

To address this problem, we proposed a computer-aided diagnosis (CADx) system for mammography. The project as originally proposed involved three major components:

(1) develop artificial neural network (ANN) decision models to predict breast cancer based upon mammographic findings extracted manually by radiologists,

(2) investigate image processing techniques to extract comparable features automatically from digitized mammograms,

(3) evaluate the clinical efficacy of the combined system by feeding the computer-extracted findings into the decision model based on radiologist findings.

The system would perform automated feature extraction from mammograms using artificial neural network (ANN) and other image processing techniques, then predict the outcome of biopsy (benign vs. malignant). The intent is to identify probably benign lesions for which biopsies may be spared.

1.2 Technical Objectives

The original technical objectives are listed below. Several important changes were required during the course of the project, as described in detail in the next section.

(1) ***Identify an optimal subset of features that would provide adequate diagnostic performance.***
   1a. Retrain the features-to-diagnosis ANN using sub-groups of features. The goal is to maintain the sensitivity of the original network while keeping specificity reasonably high.
   1b. Encode the multiple-value features into binary "sub-features", then repeat step 1a to reduce the number of sub-features. The sub-features will be easier to extract by automated schemes.

(2) ***Investigate conventional and ANN methods for extracting the optimal subset of features directly from mammograms.***
   2a. Implement established techniques which have demonstrated promise for extracting features belonging to our reduced feature set.
   2b. Investigate several ANN techniques for feature extraction, focusing on features which may be difficult to classify by conventional techniques in step 2a. For both 2a and 2b, evaluate these techniques by comparing the extracted features against radiologists' findings.

(3) ***Evaluate the automated CAD system clinically.***
   3a. Implement the CAD system by feeding the best feature extraction techniques from step 2 into the best features-to-diagnosis ANN from step 1, and compare the resulting diagnosis against the biopsy result.
   3b. Evaluate the accuracy of the CAD system retrospectively by using patient records from our computerized mammography database.

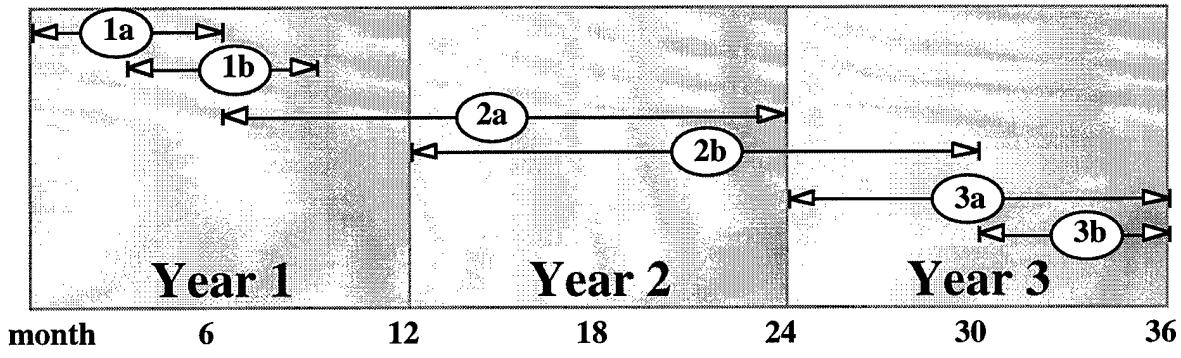


Figure 1. Time line for proposal project period.

1.3 Focus of Research

During the course of this project, it has become necessary to refine the focus of the research, increasing the emphasis on using radiologist-extracted findings rather than computer-processed features. As originally proposed, the project assumed that the predictive model based on radiologist-extracted findings was nearly finalized, and it only remained to identify the findings most important in diagnosing benign vs. malignant breast lesions. It was further assumed that a wide variety of image processing techniques could be used to extract these same findings, thereby completely automating the diagnostic process. As the project unfolded, it became apparent that these assumptions were not entirely correct. The image processing aspect of the project presented fundamental challenges which seemed to exceed the scope of a three-year postdoctoral fellowship project. In particular:

(a) There remained a substantial amount of work in developing and optimizing predictive models based on radiologist-extracted findings. This work has taken most of the effort during this project and also resulted in an unexpected level of success. Specific accomplishments in this area are summarized in the next section.

(b) It required considerable effort to establish a database consisting of digitized mammograms and the corresponding "truth files" indicating the location and gold standard diagnosis of breast lesions. Due to the relatively new nature of CADx research in mammography, these challenges were not well understood at the time of the original proposal. Since then, several projects of much greater magnitude have been funded at other institutions to establish such databases.

(c) It was originally proposed that a comprehensive list of image processing techniques would be investigated to detect or characterize mammographic features. In retrospect, re-implementing these established techniques would be tantamount to building an image processing library, a non-innovative ordeal which would still require many person-years of effort.

(d) Research with mass cases during the second budget period demonstrated that radiologists do very well with these cases, with ROC areas exceeding 0.9. This makes it exceedingly challenging to develop comparable CADx techniques, and also reduces the clinical significance of such techniques.

(e) Early in the course of the project, a study at another institution demonstrated that the detection of microcalcifications may require digitization resolution of 35 μm per pixel or better [7]. The characterization or diagnosis of these microcalcifications would likely require even greater resolution. These technical requirements far exceed the 100 μm/pixel digitizer available at this institution.

Accordingly during the second year, we refocused this project to emphasize the development of predictive models using radiologist-extracted findings. This proved to be much more fruitful, resulting in many unique and interesting studies in CADx of breast cancer.

1.4 Overview of Accomplishments

The following is a summary of the main accomplishments achieved during the three year course of this project.

During the first budget period, our institute adopted the BI-RADS (Breast Imaging Reporting and Data System) lexicon to improve consistency and accuracy of data reporting and also to allow our techniques to be used in all institutions that adopt this standardized system [8, 9]. We developed new ANNs which performed comparably to the expert mammographers who extracted the findings in the first place. Addressing specific aim 1a, we then developed an empirical technique for identifying an optimal subset of input features to the ANN. In the first annual report, we discussed why it was no longer necessary to pursue specific aim 1b.

During the second budget period, we accomplished four studies. In accordance with specific aim 2a, we improved the performance of the ANN using the optimized subset of findings by incorporating radiologist impression as an additional input finding. We also described a semi-automated technique using classic image processing techniques to extract and characterize the boundary of breast masses, and developed ANNs using those boundary findings. Addressing specific aim 2b, we explored the use of ANN techniques for computer-aided diagnosis of breast cancer. We studied the underlying behavior of these networks by examining their error surfaces in weight space. As discussed in the second annual report, we shifted the focus of the project to explore the highly promising results from using radiologist-extracted findings. Specifically, we developed an ANN to predict whether breast carcinomas were in situ or invasive, based on BI-RADS findings and age alone.

During the third and final budget period, we addressed specific aim 3 by refining and evaluating our CADx techniques clinically. In the process, we revisited several of the previous specific aims due to newer data or insights into the problem. To provide the largest possible database for system evaluation, we continued and expanded our data collection efforts, resulting in 500 patient cases, each characterized with BI-RADS mammographic findings, history features, and gold standard pathologic diagnosis from excisional biopsy. Pertaining to aim 1a, we investigated the feasibility of eliminating all history findings except patient age, thus reducing the data collection effort and ANN complexity. Regarding aim 2a, we explored a new approach for the detection of microcalcification clusters by using a combination of median filtering, bimodality segmentation, and grouping techniques. Addressing aim 2b, we investigated novel self-organizing map ANNs to analyze the BI-RADS feature data. Finally, we expanded the invasion-predicting ANN to handle all patients, including those with benign lesions, in an effort to develop a system capable of addressing all diagnostic biopsy cases.

These studies will be described in more detail in the Body of this report below.

2. Body

In the following sections, we will describe all the major studies undertaken during the three-year course of this project.

2.1 Data preparation

For the development of CADx systems, it is of vital importance to develop a good database of patient cases which provide the examples for supervised training and evaluation of the ANNs. Inconsistent or incomplete data can substantially decrease the ANN's accuracy and ability to generalize. Before and during the first budget period, we collected the initial database of 206 breast lesions. For all patients, needle localization and excisional biopsy were completed and histologic results were available. Ten mammographic findings were encoded using the BI-RADS lexicon, an improvement on our own, more subjective lexicon described in the original proposal. During the second and third budget periods, we continued to expand the number of cases to 500, thus providing a more representative sample of patients. In addition, we also expanded the number of findings to 18 by adding patient history features to all 500 cases. Details of the data collection and encoding process were described in the first annual report.

2.2 Initial malignancy-predicting ANN using 206 cases

The architecture and training algorithms for the three-layer, feed-forward, backpropagation ANNs were described in detail in the first annual report. The initial malignancy-predicting ANN was developed using the 206 cases using all ten mammographic features and patient age. This network performed with A_z of 0.84 ± 0.03, which was not significantly different from the expert radiologists' A_z of 0.85 ± 0.03 (2-tailed p-value = 0.54). When all the other history findings were included in the ANN inputs, performance was improved to A_z of 0.89 ± 0.02 but this was still not significantly better than the radiologists (p=0.29) [10]. This suggested the usefulness of collecting more cases, which can help to improve the statistical significance of these results.

2.3 Optimized reduction of input features

We next sought to identify the minimal subset of features which would still yield accurate diagnostic performance. There were several motivating reasons for doing so. Fewer features would reduce the data-entry effort of radiologists, which in turn makes it more likely that they would incorporate the ANN into their standard reading process. Fewer inputs should also permit reducing the number of hidden nodes and hence the number of ANN weights, thus ameliorating the problem of overconditioning due to insufficient training cases [11].

We developed a procedure for the ranking and elimination of individual input findings. Each feature was excluded and a new ANN was retrained using all the other

features and its performance was measured by A_z. The assumption was that the exclusion of an important feature would degrade performance more than the exclusion of an unimportant feature. Once ranked, the input features were discarded in order from least to most important in a manner analogous to backwards discrimination analysis, reducing the number of features to ten, nine, eight, and so on. Each simplified network was retrained and re-tested with the round robin process as before, and its performance was compared to that of the expert radiologists.

A six-feature network emerged as the best compromise between minimizing features and maximizing performance. Its ROC area of 0.86 ± 0.03 was not significantly different than that of the expert radiologists ($p = 0.34$). This work demonstrated an empirical technique for identifying an optimal subset of input features for a complex, nonlinear classification system.

2.4 Incorporating radiologist impression as an input feature

CADx work traditionally follow one of two paradigms, either pitting ANN output vs. radiologist diagnosis, or encouraging the radiologist to incorporate the ANN output into her final diagnosis. This study explored a novel option in which the ANN considers the radiologist diagnosis as an input finding along with BI-RADS findings. Since the radiologist impression is based on the human expert's consideration of the mammograms, clinical findings, and general experience, we hypothesized that it may provide important diagnostic information for the ANN.

Details of this study were described in the second annual report. In brief, an ANN was developed to predict the outcome of biopsy using only 4 features: mass margin, calcification description, age, and the radiologist impression. This network's A_z of 0.89 was barely significantly better than the radiologists' performance ($p=0.07$). Note that previously the ANN with several additional BI-RADS findings but not the radiologist impression failed to achieve statistical significance, with much larger p-values.

2.5 Extraction of mass margin by image processing

This study sought to replace the very important finding of mass margin with equivalent computer-extracted features. For this study, 41 mammograms were digitized to 100 micron per pixel resolution, and a 512 by 512 pixel region of interest (ROI) centered at the mass was extracted. The background was fitted to a second-order polynomial and subtracted, and the ROI was further median filtered to reduce noise. Starting from the center of the mass, possible mass boundaries were identified using a combination of region growing and global thresholding techniques. The results from each iteration were displayed to the user as a pseudocolor image. The user manually selected the color-coded threshold that most closely approximated the mass boundary.

Given the mass boundary, the irregularity and circularity were calculated. An ANN using both features and the patient age performed with A_z of 0.89 ± 0.06, but this was much worse than an ANN based on age and the radiologist-extracted mass margin,

which performed with A_z of 0.96 ± 0.03. In fact, for this limited sample of mass cases, the radiologists' impressions distinguished benign from malignant masses perfectly with A_z of 1.0. Over larger sample sizes, radiologists still performed with A_z over 0.9.

Unlike our previous studies, these ANNs did not match or outperform the radiologists that they were intended to assist. Since our expert radiologists already diagnosed masses with very high accuracy, there is in fact very little room for improvement. Preliminary additional efforts using fractal dimension analysis as an alternative mass margin extraction technique and increasing the number of cases to 100 failed to yield significantly better results.

2.6 Error surfaces of simplified ANN.

This study was a theoretical investigation into the underlying behavior of an actual ANN medical application. A single-layer perceptron was developed to predict whether masses were benign or malignant, based only on the patient age and the mass margin finding characterized by radiologists. This very simplified network was characterized by only two weights and a bias value. Three-dimensional error surfaces were visualized by calculating performance in terms of mean squared error (MSE) and ROC area index (A_z) each as a function of the three parameters. The comparison between the two different performance measures was of particular importance because at that time, most CADx techniques were designed to optimize MSE yet were evaluated in terms of A_z. This study was in fact one of the first to recognize this distinction.

From 266 randomly selected patients who underwent biopsy, 138 cases had masses. Performance was evaluated by doing a grid search over a range of weights and plotting the testing MSE and A_z against all three combinations of two weights at a time. We noted striking differences in both the optimal weight values and the underlying error surfaces for the two performance measures. Fortunately, A_z performance was a plateau over a wide range of weights, so the optimal MSE solution corresponded to A_z that was very good albeit not globally maximized. We attribute this again to the mass cases being relatively "easy" to classify, and in future work we will repeat this study but using the calcification cases which are much more challenging. Nevertheless this study demonstrated that it is important to be aware that optimizing ANNs by MSE may not necessarily result in optimization of the A_z. Several studies are also under way at this institution and elsewhere to directly optimize the A_z by using genetic algorithm techniques.

2.7 Predicting invasion of breast cancers

Up to this point, all CADx research in mammography dealt with the detection of breast lesions or the diagnosis of these lesions as benign vs. malignant. As many as 80% of biopsied malignancies are invasive [12]. Traditionally, these patients require a diagnostic excisional biopsy/lumpectomy, followed by the second, therapeutic surgical procedure of mastectomy and/or axillary dissection. For these invasive cancers, stereotaxic biopsy has also been proposed to provide histologic diagnosis in lieu of excisional biopsy, so that the patients may undergo a single-stage therapeutic surgical procedure for the mastectomy and/or axillary dissection [13, 14].

We hypothesized that it was possible to distinguish between in situ vs. invasive carcinoma based on mammographic and history findings only. If patients with invasive breast cancer can be diagnosed in this manner, excisional biopsy may be obviated and histologic confirmation may be obtained instead via stereotaxic needle core biopsy. The patients may then undergo the single-stage therapeutic surgery.

We focused on the 96 biopsy-proven malignancies from the 266 cases overall available at that time. A network was developed using 9 BI-RADS findings and patient age. It distinguished between invasive and in situ cancers surprisingly well with A_z of 0.91 ± 0.03. More importantly, it was possible to set a threshold to the ANN outputs for these malignant cases such that all in situ cancers were below that threshold and thus correctly classified (100% specificity), while 48 of 68 invasive cancers were above the threshold and thus also correctly classified (71% sensitivity). In other words, this ANN had the potential to identify a large majority of invasive cancers for the single-stage therapeutic procedure, while correctly rejecting all in situ cancers. This study was the first to develop a multivariate predictive model using readily available medical findings, i.e., BI-RADS mammographic features and patient age, to accurately classify invasion among breast cancers.

In the following sections, the results from the third and final budget period will be summarized. Since this is the first and only report of these results, they will be examined in much greater depth than the preceding studies.

2.8 Expanded database of 500 patient cases

Due to concerns of statistical significance on several studies within this project, we continued to collect patient data throughout the course of the project. We also expanded the scope of the database by adding history findings to the BI-RADS mammographic findings, thus providing the computer decision models with all the readily available medical information. Radiologists still have access to several sources of additional information, namely other mammographic views, prior films, and clinical findings.

By the end of the project, our patient database has reached 500 lesions from 478 consecutive women with nonpalpable breast lesions who underwent excisional biopsy. Furthermore, the data collection process has now become integrated into the standard

reading procedure such that all patients undergoing breast biopsy in the future will automatically be recorded. This high-quality database was used in several of the studies during the third budget period of this project, and it will facilitate many future breast cancer research projects at this as well as other institutions.

Another important benefit of having this many patient cases is the ability to switch from the round robin sampling technique to k-fold crossvalidation technique. Under this new scheme, the data are divided into k equal portions. A network is trained using k-1 portions and tested on the excluded portion, and the process is then repeated until every portion has been excluded once for testing. This approach is computationally many orders of magnitude faster than round robin testing, which is equivalent to crossvalidation with k equal to the number of cases.

2.9 Reducing number of inputs by eliminating history findings

In previous work we attempted to identify the most important findings contributing toward the diagnosis of breast cancer. We considered only BI-RADS findings and the patient age and utilized a relatively smaller database of 266 cases only. Later in the project, we realized there was relatively less value in eliminating a few BI-RADS findings, but it would be useful to be able to exclude all history findings except age, thus reducing the number of findings from 18 to 11 and avoiding an entire category of inputs which are more subjective, prone to error, and difficult to collect.

All 500 available biopsied cases were used for model training and testing. Three different ANN models were developed: ANN-10 using only the ten BI-RADS mammographic findings, ANN-11 using the ten BI-RADS findings and just patient age, and ANN-18 using all available BI-RADS and history findings. The networks were compared to each other and the original radiologist's impression using several clinically relevant performance measures: the receiver operating characteristic (ROC) area index, A_Z, over the testing cases; specificity given perfect or near-perfect sensitivity of 95%, 98%, and 100%; and PPV of biopsy at perfect sensitivity. While the ROC area measures the performance over the entire range of sensitivities and specificities, the only operating points which are clinically relevant correspond to the specificity at near-perfect sensitivity. This measure of performance shows how many benign biopsies may be obviated given the requirement that few if any cancers are missed. To compare all four decision models against each other, statistical significance of differences in both A_Z and specificity at 98% sensitivity were evaluated using the CLABROC program (courtesy of Dr. Charles Metz, the University of Chicago).

The performance of the four decision models described above are listed in Table 1. The radiologists initially recommended biopsies for all 500 cases at the time of mammography, so by definition their sensitivity among these cases was 100% (all cancers diagnosed) while their specificity was 0% (no biopsies spared). The PPV of biopsy according to the radiologist impression was 35%, corresponding to the prevalence of cancer in this data set, which was very typical of diagnostic mammography at this institution.

Table 1. Performance of decision models

	ROC area Az	specificity at 100% sensitivity	specificity at 98% sensitivity	specificity at 95% sensitivity	PPV
Radiologist impression	0.82 ± 0.02	0%	12%	37%	35%
ANN-10 mammo only	0.84 ± 0.02	6%	39%	53%	36%
ANN-11 mammo+age	0.86 ± 0.02	30%	42%	51%	43%
ANN-18 mammo+Hx	0.87 ± 0.02	22%	41%	52%	41%

The network using all available findings, ANN-18, performed with the highest ROC area of 0.87±0.02. At 95% sensitivity, the network's specificity was 52%. In other words, for a threshold which correctly diagnosed 95% of all cancers (168 of 174), over half of all benign biopsies were identified (168 of 326) and may have potentially been obviated. Higher sensitivities could also be attained with some loss in specificity by adjusting the threshold applied to the decision model's outputs. For example, when the sensitivity was raised from 95% to 98% (missing only 3 out of 174 cancers), the specificity dropped to 41% (obviating 133 of 326 benign biopsies). At perfect 100% sensitivity where no cancers were missed, the specificity was reduced to 22% corresponding to obviating 72 of 326 benign biopsies.

For most of the performance measures shown in Table 1, the ANN-18 network (using mammographic and history findings) was comparable to the simpler ANN-11 network (using mammographic findings and the single history finding of patient age). There was usually a large decrease in performance with the ANN-10 network based on mammographic findings only. All three networks out-performed the expert radiologists who originally extracted the findings used by the networks.

To demonstrate the statistical significance of these differences in performance among the decision models, two performance measures were analyzed in depth: the ROC area A_z in Table 2 and the specificity at 98% sensitivity in Table 3. Each table consists of a confusion matrix showing the two-tailed p-value for all possible comparisons between the four decision models, with significant differences ($p < 0.05$) emphasized in boldface.

Table 2. Statistical comparison of ROC area for all 4 decision models

	Radiologist impression 0.82±0.02	ANN-10 mammo only 0.84±0.02	ANN-11 mammo+age 0.86±0.02	ANN-18 mammo+Hx 0.87±0.02
Radiologist impression 0.82±0.02		0.459	0.079	**0.042**
ANN-10 mammo only 0.84±0.02			**0.028**	**0.020**
ANN-11 mammo+age 0.86±0.02				0.324
ANN-18 mammo+Hx 0.87±0.02				

Table 3. Statistical comparison of specificity at 98% sensitivity for all 4 decision models

	Radiologist impression 12%	ANN-10 mammo only 39%	ANN-11 mammo+age 42%	ANN-18 mammo+Hx 41%
Radiologist impression 12%		**0.015**	**0.001**	**0.001**
ANN-10 mammo only 39%			0.130	**0.043**
ANN-11 mammo+age 42%				0.410
ANN-18 mammo+Hx 41%				

In both Tables 2 and 3, the first row compares all three networks to the radiologist impression. The full-featured ANN-18 was the only network that significantly outperformed the radiologists ($p=0.042$) in A_z. When comparing specificity at 98% sensitivity, however, all three networks were significantly better than the radiologist's impression ($p=0.015$, 0.001, and 0.001 for ANN-10, ANN-11, and ANN-18, respectively). The second row compares the basic ANN-10 model using mammographic findings only to the history-enhanced networks: ANN-11 with age and ANN-18 with all

history findings. The addition of the single feature of patient age to get ANN-11 significantly improved A_Z (p=0.028) but not specificity (p=0.130). Adding all history findings to get ANN-18 did significantly improve both A_Z and specificity (p=0.020 and 0.043, respectively). Finally, the single value in the last row compares ANN-11 based on mammographic findings plus age to ANN-18 with all the other history findings. The inclusion of the seven history findings other than age did not produce a significant difference in either A_Z or specificity (p=0.324 and 0.410, respectively).

These experiments sought the answers to two simple but important questions. Given the existing models using all available mammographic and history findings, can one eliminate all history findings but the patient age and still maintain diagnostic performance? Conversely, given a model based on mammographic information only, can including the single additional finding of patient age significantly improve performance?

Addressing the first question, the results revealed that it was indeed possible to maintain performance while eliminating all history findings except for patient age. When the ANN-18 with all mammographic and history findings was compared against the ANN-11 with mammographic findings and just patient age, the two models performed nearly identically over all the measures considered, including A_Z, specificity over a range of perfect or near-perfect sensitivities, and PPV (see last two rows in Table 1). In particular, for the two most clinically relevant and often reported performance measures of A_Z and specificity at 98% sensitivity, the differences between the two models were not statistically significant.

These results have important implications for the use of history findings in decision models to predict breast cancer from mammograms. Experiences at this institution confirmed that history findings are laborious to collect and can be quite subjective or unreliable. The patient age on the other hand is readily and accurately available. The current study suggested that, for the purpose of developing malignancy-predicting models based primarily upon mammographic findings, it is possible to replace a full list of history findings with just the patient age. This can substantially decrease the effort of collecting patient data, making these decision models much easier to develop and to implement clinically.

As for the second question, the contribution of the age finding itself was demonstrated by comparisons between the ANN-11 and ANN-10 networks, which used mammographic findings with or without age, respectively. The addition of age to mammographic findings significantly improved A_Z from 0.84 to 0.86 (p = 0.028). The age feature also increased the specificity at 98% sensitivity from 39% for ANN-10 to 42% for ANN-11, although this difference was not statistically significant (p = 0.13).

These results confirm the importance of the patient age in predicting whether a lesion is benign or malignant. This conclusion is based on 500 cases and is consistent with the previous study using 266 cases described in a previous section [15]. These conclusions may even generalize to other malignancy-predicting techniques which utilize mammographic information only. In other words, the simple inclusion of patient

age would require minimal additional effort and model complexity, but may benefit decision models such as those based upon features extracted automatically from mammograms by image processing techniques.

The conclusions from this study should be qualified with an important caveat. Although the current sample set of 500 cases was larger than those in previous reports, there is still no assurance that this database is representative of the entire patient population at this institution much less elsewhere. To address this important concern, on-going studies will analyze many more patients from this institution as well as several others nationwide. This additional data should help improve the statistical significance of results shown in Tables 1 and 2, and may help to address the inconsistencies between the different performance measures.

2.10 Detection of microcalcification clusters

In the introduction to this report, we cited a recently published report which suggested that 35 μm per pixel digitizer resolution is required in order to detect microcalcifications effectively [7]. The characterization of such microcalcifications would require even better resolution since it would take at least several pixels per object to describe its shape and margin. Since the digitizer available at this institution is limited to 100 μm per pixel resolution, we surmised that the detection and characterization of microcalcifications as originally proposed was well beyond our technical abilities.

We did commence a pilot study, however, to explore some new techniques for the detection of microcalcification clusters. The rationale was that for detection of clusters, it was not necessary to detect every single calcification. In addition, if these results proved promising, we might then attempt to secure high-resolution images via collaborative efforts with other institutions.

We randomly selected 20 mammograms with calcifications for this study. A 512x512 ROI was manually extracted which contained the calcifications. The background trend in the ROIs were removed using an unsharp masking filter. Noise and artifacts were further removed using an 11x11 median filter whose mask size was designed to be larger than most calcifications. The ROI was then segmented using a bimodality technique. For a small scrolling window of 16x16 pixels, the histogram of pixel values was assumed to be comprised of two gaussian distributions, representing the calcifications vs. the background. This bimodality technique produces a binary segmentation of the calcifications from the background. The calcifications are then identified and clustered automatically using the inter-object distances, and clusters with too few calcifications are discarded. The result is a collection of calcification clusters.

Preliminary results over these 20 sample ROIs indicate 100% sensitivity (no missed clusters) with typically only 1-2 false positives per image. We are enthusiastic about the sensitivity of this technique, which successfully detected many extremely subtle microcalcifications which were barely visible in the original films. These preliminary results appeared to be comparable to detection techniques developed at

other institutions. Much work remains to be performed, however. We are developing an automatic scoring program to verify the true and false positive rates of detection. We have also acquired a database of 100 images with truth files courtesy of Dr. Maria Kallergi of the University of South Florida. These images will be used as an independent testing set. In the future, if higher resolution images become available, it may be possible to not only detect but also characterize these clusters. Although this work could not be completed during the proposed budget period, this project has laid the foundation for many years of promising research.

2.11 Self-organizing map ANNs to analyze input findings

Almost all previous studies in computer-aided diagnosis of mammography were based on the classic feedforward, backpropagation network architecture. The purpose of this study is to analyze mammographic findings using self-organizing map (SOM) artificial neural networks. Using two findings of patient age and mass margin extracted by radiologists, self-organizing maps were developed to analyze both the distribution and topology of the input findings.

Self-organizing maps use unsupervised learning to classify patterns [16, 17]. The neurons of a self-organizing map are distributed and ordered corresponding to the location as well as frequency of the input pattern vectors. Such networks may yield important insights into the breast cancer diagnostic process.

For this study, 266 cases were used. The networks were implemented using MATLAB (The MathWorks, Inc., Natick, MA). Results were visualized as two-dimensional feature maps corresponding to the two input findings of mass margin (on vertical axis) and age (on horizontal axis). Each neuron's weight vector is shown as a data point, while neighboring neurons are indicated by connecting lines.

Figure 1 below shows a 1-dimensional layer of 5 neurons which were trained to produce a self-organizing map of the input findings. Note the approximately linear behavior which then tapers off for higher values of age, suggesting the correlation between increasing age and mass margin values with increasing likelihood of cancer.

Figure 2. Feature map for 1-D self-organizing map with 5 neurons.

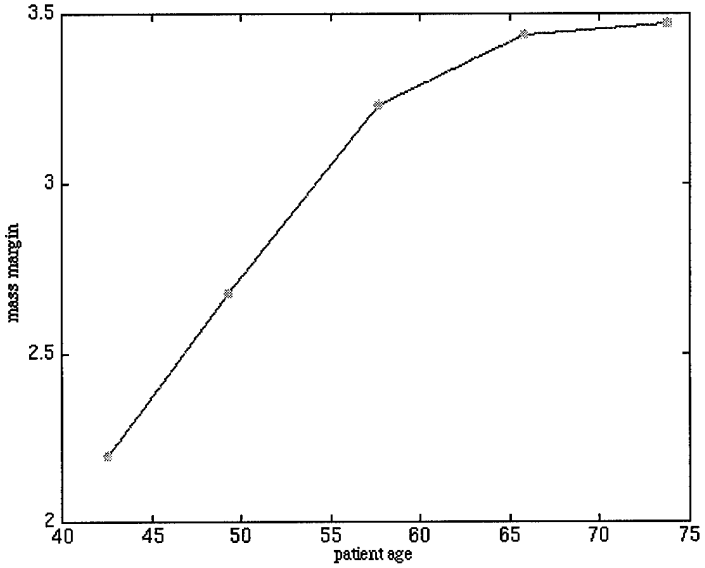
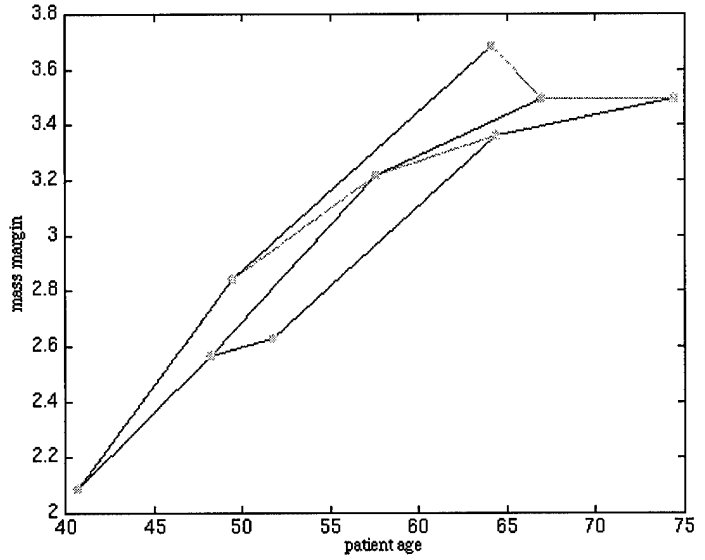


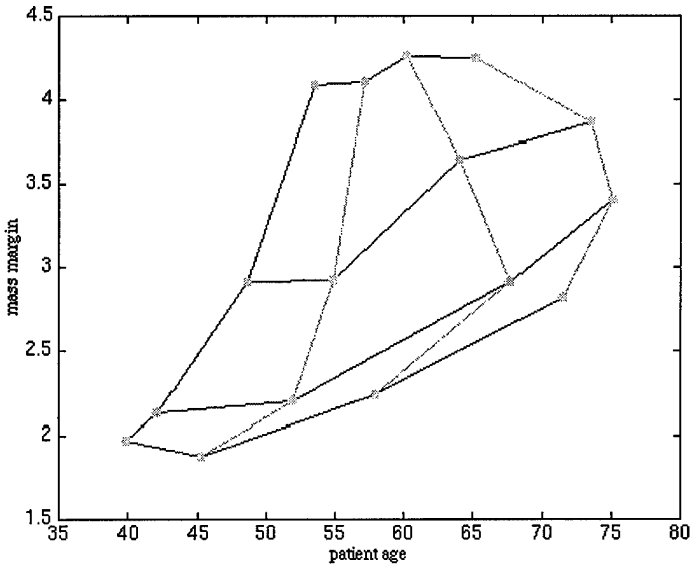
Figure 2 below shows a 2-dimensional self-organizing map with 9 neurons in a 3x3 array. This network spanned a similar amount of the input space as the simpler 5-neuron network, with a larger number of neurons densely covering the older patients with higher mass margin values in the upper right quadrant of the graph.

Figure 3. Feature map for 2-D 3x3 self-organizing map.



Finally, Figure 3 below shows a 2-dimensional, 4x4 self-organizing map. The increase in number of neurons resulted in a network which spanned a much greater range over the input space with much more evenly distributed neurons.

Figure 1. Feature map for 2-D 4x4 self-organizing map.



This study demonstrated the use of self-organizing maps to learn the distribution and topology of mammographic findings data. In each case, the neuron vectors were able to group the inputs into different classes using unsupervised training. Increasing the dimensionality and number of neurons resulted in clusters which covered wider spans of the input space and did so more evenly.

To arrive at their diagnosis, radiologists have to consider a large number of variables, most of which have non-binary continuous values or may take on one of many different discrete values. By analyzing mammographic findings using self-organizing maps, it may be possible to elucidate underlying trends and patterns within the data. It may also lead to methods for grouping or encoding the input data which may reduce the dimensionality of the problem, thereby facilitating the development of supervised training networks.

2.12 Revised invasion-predicting ANN for all patients

We described previously an ANN to distinguish between in situ vs. invasive carcinomas. That network was trained and tested only on biopsy-proven cancers. The current study expands on the previous work by attempting to predict both breast lesion malignancy and invasion among all diagnostic mammography patients. The computer model consisted of multi-stage artificial neural networks (ANNs) which merged features from the standardized BI-RADS mammography lexicon and patient history in order to predict biopsy outcome.

This study utilized all 500 cases available to date from our patient database. Two backpropagation ANNs using these findings were cascaded to predict first benign vs. malignant lesions, and then for the latter category, in situ vs. invasive cancer. The first-stage ANN considered all 500 cases with the goal of eliminating all probably benign

lesions, thus obviating many benign biopsies. As with our previous malignancy-predicting ANNs, this was achieved by selecting a low threshold over the ANN outputs such that the network performed with perfect sensitivity. All cases with outputs below this threshold were actually benign, and these patients may be closely followed up rather than sent to biopsy. The cases exceeding the threshold were comprised of all the cancers as well as some benign cases. These indeterminate cases were referred to the second-stage ANN, which was designed to identify the opposite extreme, those cases which were probably invasive cancers. For these lesions classified as probably invasive cancer, excisional biopsy may be avoided by obtaining histologic confirmation via stereotaxic needle core biopsy, and the patients may then undergo a single-stage therapeutic surgery for mastectomy or axillary dissection.

Both ANNs were trained and tested with k-fold crossvalidation sampling (k=10 and 11, respectively). The first-stage ANN distinguished benign from malignant lesions with ROC area of 0.86 ± 0.02, improving positive predictive value of biopsy from 35% to 47% and sparing 134 of 326 (41%) of benign biopsies, while missing only 3 malignancies (98% sensitivity). The remaining 363 indeterminate patients were referred to the second-stage ANN, which identified 40 of 120 (33%) invasive cancers among these patients, with ROC area of 0.82 ± 0.02. Together, the two networks have the potential to obviate 174 out of the 500 biopsies.

This study was particularly unique and important, and represented the culmination of the three years of research effort in this project. It was the first unified computer model that predicted breast lesion malignancy as well as invasion, based upon mammographic features and patient history. This work will be presented at the Radiological Society of North America 1997 annual conference in November [18].

3. Conclusions

This postdoctoral fellowship project has explored computer-aided diagnosis (CADx) of breast cancer using artificial neural network (ANN) computer models. The most important accomplishment was the development of the first CADx system that merged radiologist-extracted features to predict not only malignancy but also invasion among breast lesions.

We also explored many other projects which contributed to the success of the main project above. We identified an optimal subset of input features to the predictive model while maintaining diagnostic accuracy. We also evaluated the feasibility of excluding history findings which may substantially reduce data collection effort and model complexity. We investigated the underlying characteristics of the inputs using self-organizing map ANNs. We also evaluated the behavior of the computer model by analyzing the error surfaces in weight space. In on-going work, we developed image processing techniques for the extraction of mass margins and the detection of microcalcification clusters.

Together, all these projects contribute toward the development of a unified computer-aided diagnosis system for mammography. By providing information which was previously available only through biopsy, the system may assist in surgical planning for patients with breast lesions, and may reduce the cost and morbidity of "unnecessary" surgical biopsies. These implications will become increasingly important as mammography screening becomes increasingly widespread.

This postdoctoral fellowship has also been very fruitful for the PI. During the course of this project, we produced 17 publications and abstracts (listed in the following section). In addition, this project has directly led to the recent funding of two major grants, a Whitaker Foundation fellowship and an NIH FIRST Award, both commencing during 1998. Although the current project has only funded the PI over the past three years, the upcoming projects will fund a wide variety of personnel including graduate students, post-doctoral fellows, and full faculty members. In conclusion, the current project has helped the PI to become an independent researcher in this field, and the work achieved herein will continue to foster breast cancer research for many years to come.

4. Publications / Abstracts Resulting from this Project

1. Floyd CE, Jr, **Lo JY**, Yun AJ, Sullivan DC, and Kornguth PJ. Prediction of breast cancer malignancy using an artificial neural network. Cancer 1994; 74: 2944-2948.
2. **Lo JY**, Floyd CE, Jr, and Tourassi GD. *Artificial neural networks for diagnosis in radiology.* in *Computer-Aided Diagnosis Workshop.* 1994. Georgetown University Medical Center, Washington, DC:
3. **Lo JY**, Grisson AT, Floyd CE, Jr, and Kornguth PJ. *Computer-aided diagnosis of mammograms using an artificial neural network: merging of standardized input features from the ACR lexicon.* in *SPIE Medical Imaging 1995: Image Processing.* 1995.
4. Baker JA, Kornguth PJ, **Lo JY**, and Floyd CE, Jr. Artificial neural network: improving the quality of breast biopsy recommendations. Radiology 1995; 198: 131-135.
5. Baker JA, Kornguth PJ, **Lo JY**, Williford ME, and Floyd CE, Jr. Breast cancer: prediction with artificial neural network based on BI-RADS standardized lexicon. Radiology 1995; 196: 817-822.
6. **Lo JY**, Baker JA, Kornguth PJ, and Floyd CE, Jr. Computer-aided diagnosis of breast cancer: artificial neural network approach for optimized merging of mammographic features. Academic Radiology 1995; 2: 841-850.
7. **Lo JY**, Baker JA, Kornguth PJ, and Floyd CE, Jr. *Computer-aided diagnosis of mammography: Artificial neural networks for optimized merging of standardized BIRADS features.* in *World Congress on Neural Networks 95 (International Neural Network Society Annual Meeting).* 1995. Washington, D.C.:
8. **Lo JY**, Baydush AH, Baker JA, Kornguth PJ, and Floyd CE, Jr. Computer-aided diagnosis of breast mass malignancy using automated feature extraction and artificial neural networks. Radiology 1995; 197(P): 425.
9. **Lo JY**, Baker JA, Kornguth PJ, and Floyd CE, Jr. Application of artificial neural networks to the interpretation of mammograms based on the radiologist impression and optimized BI-RADS™ image features. Radiology 1995; 197(P): 242.
10. Floyd CE, Jr, **Lo JY**, Baker JA, and Kornguth PJ. Interactive computer-aided diagnosis of breast cancer. Radiology 1995; 197(P): 533.
11. **Lo JY**, Kim J, Baker JA, and Floyd CE, Jr. *Computer-aided diagnosis of mammography using an artificial neural network: Predicting the invasiveness of breast cancers from image features.* in *SPIE Medical Imaging 1996: Image Processing.* 1996.
12. **Lo JY** and Floyd CE, Jr. *Analysis of error surfaces of neural network applied to computer-aided diagnosis in mammography.* in *World Congress on Neural Networks '96 (International Neural Network Society 1996 Annual Meeting).* 1996. San Diego, CA: Lawrence Erlbaum Associates, Inc.

13. **Lo JY**, Baker JA, and Floyd CE, Jr. Artificial neural networks for the prediction of breast cancer invasiveness by using Breast Imaging and Reporting Data System mammography lexicon. Radiology 1996; 201(P): 370.

14. Floyd CE, Jr, **Lo JY**, Tourassi GD, Baker JA, Vitittoe NF, and Vargas-Voracek R. Computer aided diagnosis in thoracic and mammographic radiology. Medical Imaging Technology 1996; 14: 629-634.

15. **Lo JY**, Baker JA, Kornguth PJ, Iglehart JD, and Floyd CE, Jr. Predicting breast cancer invasion with artificial neural networks on the basis of mammographic features. Radiology 1997; 203: 159-163.

16. **Lo JY** and Floyd CE, Jr. *Self-organizing maps for analyzing mammographic findings.* in *IEEE International Conference on Neural Networks.* 1997. Houston, TX: IEEE.

17. **Lo JY**, Baker JA, Frederick ED, Kornguth PJ, and Floyd CE, Jr. Predicting breast lesion malignancy and invasion using the BI-RADS mammography lexicon. Radiology 1997; (RSNA 97 proceedings, in press).

5. References Cited in this Report

1. Kopans DB. The positive predictive value of mammography. American Journal of Roentgenology 1992; 158: 521-526.
2. Knutzen AM and Gisvold JJ. Likelihood of malignant disease for various categories of mammographically detected, nonpalpable breast lesions. Mayo Clinic Proceedings 1993; 68: 454-460.
3. Bassett LW, Bunnell DH, Cerny JA, and Gold RH. Screening mammography: referral practices of Los Angeles physicians. American Journal of Roentgenology 1986; 147: 689-92.
4. Cyrlak D. Induced costs of low-cost screening mammography. Radiology 1988; 168: 661-3.
5. Varas X, Leborgne F, and Leborgne JH. Nonpalpable, probably benign lesions: role of follow-up mammography. Radiology 1992; 184: 409-14.
6. Sickles EA. Periodic mammographic follow-up of probably benign lesions: results in 3,184 consecutive cases. Radiology 1991; 179: 463-8.
7. Chan H-P, Niklason LT, Ikeda DM, and Lam KL. Digitization requirements in mammography: Effects on computer-aided detection of microcalcifications. Medical Physics 1994; 21: 1203-11.
8. Kopans DB. Standardized mammography reporting. Radiologic Clinics of North America 1992; 30: 257-264.
9. D'Orsi CJ and Kopans DB. Mammographic feature analysis. Seminars in Roentgenology 1993; 28: 204-230.
10. Baker JA, Kornguth PJ, Lo JY, Williford ME, and Floyd CE, Jr. Breast cancer: prediction with artificial neural network based on BI-RADS standardized lexicon. Radiology 1995; 196: 817-822.
11. Lippman RP. An introduction to computing with neural nets. IEEE ASSP Magazine 1987; : 4-22.
12. Ciatto S, Cataliotti L, and Distante V. Nonpalpable lesions detected with mammography: review of 512 consecutive cases. Radiology 1987; 165: 99-102.
13. Jackman RJ, Nowels KW, Shepard MJ, Finkelstein SI, and Marzoni FA, Jr. Stereotaxic large-core needle biopsy of 450 nonpalpable breast lesions with surgical correlation in lesions with cancer or atypical hyperplasia. Radiology 1994; 193: 91-5.
14. Liberman L, Dershaw DD, Rosen PP, Cohen MA, Hann LE, and Abramson AF. Stereotaxic core biopsy of impalpable spiculated breast masses. AJR Am J Roentgenol 1995; 165: 551-4.

15. Lo JY, Baker JA, Kornguth PJ, and Floyd CE, Jr. Computer-aided diagnosis of breast cancer: artificial neural network approach for optimized merging of mammographic features. Academic Radiology 1995; 2: 841-850.

16. Kohonen T. *Automatic formation of topological feature maps in a self-organizing system.* in *2nd Scandinavian Conference on Image Analysis.* 1981. Espoo:

17. Kohonen T. The self-organizing map. IEEE 1990; 78: 1464-1480.

18. Lo JY, Baker JA, Frederick ED, Kornguth PJ, and Floyd CE, Jr. Predicting breast lesion malignancy and invasion using the BI-RADS mammography lexicon. Radiology 1997; (RSNA 97 proceedings, in press).