MODULAR MACHINE LEARNING METHODS FOR COMPUTER-AIDED DIAGNOSIS OF BREAST CANCER

by

Mia Kathleen Markey

Department of Biomedical Engineering
Duke University

Date: June 27, 2002

Approved:

Joseph Y. Lo, Ph.D., supervisor

Jay A. Baker, M.D.

Carey E Floyd

Carey E. Floyd, Jr., Ph.D.

Georgia D. Tourassi, Ph.D.

Gregg E. Trahey, Ph.D.

Dissertation submitted in partial fulfillment of
the requirements for the degree of Doctor
of Philosophy in the Department of
Biomedical Engineering in the Graduate School
of Duke University
2002

UMI Number: 3077145

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

ABSTRACT

MODULAR MACHINE LEARNING METHODS FOR COMPUTER-AIDED DIAGNOSIS OF BREAST CANCER

by

Mia Kathleen Markey

Department of Biomedical Engineering
Duke University

Date: June 27, 2002

Approved:

Joseph Y. Lo, Ph.D., supervisor

Jay A. Baker m,

Jay A. Baker, M.D.

Carey E. Floyd, Jr., Ph.D.

Georgia D. Tourassi, Ph.D.

Gregg E. Trahey, Ph.D.

An abstract of a dissertation submitted in partial
fulfillment of the requirements for the degree
of Doctor of Philosophy in the Department of
Biomedical Engineering in the Graduate School of
Duke University
2002

Abstract

The purpose of this study was to improve breast cancer diagnosis by reducing the number of benign biopsies performed. To this end, we investigated modular and ensemble systems of machine learning methods for computer-aided diagnosis (CAD) of breast cancer. A modular system partitions the input space into smaller domains, each of which is handled by a local model. An ensemble system uses multiple models for the same cases and combines the models' predictions.

Five supervised machine learning techniques (LDA, SVM, BP-ANN, CBR, CART) were trained to predict the biopsy outcome from mammographic findings (BI-RADS™) and patient age based on a database of 2258 cases mixed from multiple institutions. The generalization of the models was tested on second set of 2177 cases. Clusters were identified in the database using *a priori* knowledge and unsupervised learning methods (agglomerative hierarchical clustering followed by K-Means, SOM, AutoClass). The performance of the global models over the clusters was examined and local models were trained for clusters.

While some local models were superior to some global models, we were unable to build a modular CAD system that was better than the global BP-ANN model. The ensemble systems based on simplistic combination schemes did not result in significant improvements and more complicated combination schemes were found to be unduly optimistic. One of the most striking results of this dissertation was that CAD systems trained on a mixture of lesion types performed much better on masses than on calcifications. Our study of the institutional effects suggests that models built on cases mixed between institutions may overcome some of the weaknesses of models built on cases from a single institution. It was suggestive that each of the unsupervised methods identified a cluster of younger women with well-circumscribed or obscured, oval-shaped masses that accounted for the majority of the BP-ANN's recommendations for follow up. From the cluster analysis and the CART models, we determined a simple diagnostic rule that performed comparably to the global BP-ANN. Approximately 98% sensitivity could be maintained while providing approximately 26% specificity. This should be compared to the clinical status quo of 100% sensitivity and 0% specificity on this database of indeterminate cases already referred to biopsy.

Acknowledgements

First and foremost, I would like to thank my advisor, Joseph Y. Lo. I will be very proud if I provide my own first student with anything close to the quality and depth of mentoring that Jo has given me.

Many thanks are due to the other members of my committee: Jay Baker, Carey Floyd, Gina Tourassi, and Gregg Trahey. I don't know if it takes a village to raise a child, but it certainly takes a committee to shepherd a graduate student along. Carey and Gina have often acted as additional advisors above and beyond the traditional level for committee members, for which I am very grateful.

I have received considerable support from all the members of the Digital Imaging Research Division, but I would especially like to thank Neal and Fredder for getting me off to a good start and Rene, Marios, and Brian for keeping me on that path. Additional thanks to Käthe and Beth for keeping me supplied with data and to Brian, Gina, and Anya for scientific programming.

I would like to thank my family for being there for me. I can't begin to thank my husband, Eric Stuckey, enough, but I suppose that I should spare posterity all the mushy stuff. Thanks to Henry for providing motivation to "stay on target". My siblings (Matt, Mo, Lydia, Michelle, Molly, & Margie) have all stood by me, but Michelle and Molly deserve extra recognition for experiencing the trials and tribulations of graduate school along with me. I give special thanks to my parents, John and Dolores Markey. I would like to reassure Dragon Lady that graduate school is a damn sight better than squeezing the shit out of a dead turkey. Quasimodo should note that no experiments involving plastic butter dishes and woodstoves were performed as part of this dissertation.

List of Tables

List of Figures

List of Abbreviations

ACR - American College of Radiology

BI-RADS™ - breast imaging reporting and data system

BP-ANN – back-propagation artificial neural network

AUC – non-parametric estimate of area under the ROC curve

A_Z - semi-parametric estimate of area under the ROC curve

CAD – computer-aided diagnosis

CART – classification and regression tree

CBR – case-based reasoning

DDSM – Digital Database for Screening Mammography

Duke – Duke University Medical Center

FDA – Food and Drug Administration, US Department of Health and Human Services

FPF – false positive fraction, 1 – specificity, probability of false alarm (P_{FA})

LDA – linear discriminant analysis

MLO - mediolateral oblique

partial AUC – normalized, non-parametric estimate of partial area under the ROC curve (TPF 0.9 - 1.0)

partial AUC index - normalized, semi-parametric estimate of partial area under the ROC curve (TPF 0.9 - 1.0)

ROC – receiver operating characteristic

SOM – self-organizing map

SVM – support vector machine

TPF – true positive fraction, sensitivity, probability of detection (P_D)

UPenn – University of Pennsylvania Medical Center

1 Background

1.1 Breast Cancer and Mammography

Among American women, breast cancer is the most common cancer, excluding skin cancers, and is the second leading cause of cancer deaths, after lung cancer [1, 2]. It is estimated that 203,500 American women will be diagnosed with invasive breast cancer (plus an additional 54,300 *in situ*) and 39,600 will die of breast cancer in 2002 [2]. Women in the United States have about a 1 in 8 lifetime risk of developing invasive breast cancer [3, 4]. Mammographic screening has been shown to reduce the mortality of breast cancer by as much as 30% [5, 6]. However, mammography has a low positive predictive value (PPV). Only about 10-34% of the women who undergo biopsies for pathological diagnosis of breast cancer are found to have malignancies [7]. Our goal of the application of computer-aided diagnosis to mammography is to reduce the false positive rate. Avoiding benign biopsies spares women unnecessary discomfort, anxiety, and expense. Moreover, the cost of benign biopsies is the major induced cost of mammographic screening [8].

The American College of Radiology (ACR) has defined a standard system for the reporting of mammographic findings [9]. The ACR Breast Imaging Reporting and Data System (BI-RADS™) is a lexicon for the description of mammographic lesions, mostly in terms of categorical features [10]. The BI-RADS™ lexicon includes such morphological features as the description of the margin of a mass and the distribution of calcifications. It has been demonstrated that the BI-RADS™ final assessment rating is an indicator of the likelihood of malignancy [11-13] and that different values of the BI-RADS™ features are associated with different odds of malignancy [11]. Previous work in our laboratory has established the utility of BI-RADS™ features as inputs to predictive computer models [14-18]. While some inter- and intra-observer variability in the use of the BI-RADS™ lexicon has been observed [15, 19, 20], BI-RADS™ is an important part of mammography standardization and it is expected that more consistency will be seen as radiologists gain familiarity with the lexicon. Moreover, our research group has shown that a computer model based on radiologists' differing descriptions of lesions can make consistent, accurate predictions of malignancy [21]. For this study, mammographic lesions were summarized according to the BI-RADS™ lexicon (Section 1.4).

In addition to mammographic findings, patient history descriptors are also related to breast cancer status. The single most important risk factor is age. Increasing age is associated with increasing risk of breast cancer; a 60 year old white American woman has a fourteen fold increase in her chances of developing breast cancer relative to a 30 year old white American woman [6]. Previous work in our laboratory has established the utility of age [16, 18, 22] as an input for predictive computer models. In agreement with the epidemiological data, there is some evidence that age is a particularly valuable input in our predictive models [23]. Thus patient age was also used in this study (Section 1.4).

1.2 Computer-Aided Diagnosis

Computer aided diagnosis (CAD) of breast cancer is the application of computational techniques to the problem of interpreting breast images, usually mammograms [24-28]. There are two major topics in breast cancer CAD: detection of mammographic lesions and diagnosis of cancer from identified lesions. In the detection task, the goal is to assist a radiologist in the identification, and often the localization, of lesion-containing regions of mammograms. While CAD for mammographic detection is still an active area of research, there are currently three vendors with FDA approved commercial systems: R2 Technology (Los Altos, CA), CADx Medical Systems (Laval, Quebec), and Howtek (Hudson, NH). In the diagnosis task, the goal is to assist a radiologist in determining whether an identified breast lesion is an indication of the presence of cancer. This study focused on the diagnosis of breast lesions that were identified by radiologists as suspicious enough to warrant biopsy. In other words, these cases are generally considered indeterminate and more challenging, and any reduction in the number of benign biopsies would represent an improvement over the status quo, in which all such cases were referred to biopsy. Currently, there aren't any FDA approved CAD systems to aid in the classification of breast lesions as benign or malignant. It is important to keep the legal climate in mind when discussing breast cancer CAD systems. The Physician Insurers Association of America reports that breast cancer is the most common and second most expensive condition resulting in claims against physicians [29].

Breast imaging CAD systems generally have two major components: (1) a feature extraction algorithm and (2) a decision algorithm. An image of a lesion in a diagnosis task, or potential lesion in a detection task, must be summarized by a set of numerical features that serve as the inputs to a decision algorithm. Numerical features can be encoded from radiologists' observations about breast lesions using a lexicon such as BI-RADS™ [10]. Alternatively, numerical measures, such as texture [30] or morphological [31, 32] features, can be calculated directly from the image. The decision algorithms used were typically developed in statistics or machine learning, a subfield of artificial intelligence that is focused on the development of algorithms that enable computers to learn from experience. Many types of decision algorithms have been employed in breast cancer CAD, including linear discriminant analysis [33], genetic algorithms [34, 35], rule-based systems [36], neural networks [37-40], and case-based reasoning [18].

1.2.1 Modular and Ensemble Breast Cancer CAD Systems

The focus of this study was to investigate the utility of using combinations of multiple machine learning algorithms in a modular or ensemble breast cancer CAD system to reduce the number of benign biopsies performed. A modular system uses multiple classifiers to solve a classification problem by partitioning the input space into smaller domains, each of which is handled by a local model [41]. The local models can be thought of as experts for a particular kind of case. The idea behind such a "divide-and-conquer" approach is to break the problem down into smaller, simpler problems that will be easier to solve. An ensemble system uses multiple classifiers to solve a classification problem by training multiple models for the same cases and then combining models' predictions [41]. The idea behind such an approach is that "two heads are better than one".

Modular and ensemble systems have been previously applied in breast cancer CAD. Simple ensembles of classifiers using voting or averaging to combine their predictions have shown promise in computer-aided detection of breast masses [42-44]. Zheng *et al.* employed a modular scheme, in which the data were partitioned by a difficulty measure, for computer-aided detection of breast masses with encouraging results [45]. Zheng *et al.* also investigated a promising ensemble of modular models, formed by taking the average of the predictions from modular models in which the data were partitioned using three features [46]. Huo *et al.* described a modular system, in larger (a few thousand cases) than those typically used in breast cancer CAD research (a few hundred cases) [32, 49-51].

1.3.1 Receiver Operating Characteristic Analysis

Receiver Operating Characteristic (ROC) curves can be used to show the trade-off in sensitivity and specificity achievable by a classifier by varying the threshold on the output decision variable [52, 53]. Sensitivity, or the true positive fraction (TPF) or the probability of detection (P_D), is the fraction of positive cases that were classified correctly as positive. The specificity, or one minus the false positive fraction (FPF), is the fraction of negative cases that were correctly classified as negative. The false positive fraction is also known as the probability of false alarm (P_{FA}).

An ROC curve is generated by applying a threshold to the output of a classification scheme and then plotting the (FPF, TPF) pairs for each threshold. The performance of classification methods can be evaluated by directly comparing their ROC curves or by comparing indices calculated from their curves. In particular, the area under the ROC curve (AUC) is often used as a measure of classifier performance. Notice that the values for AUC range from 0.5 for chance to 1.0 for a perfect classifier. In evaluating models for diagnosing breast cancer, all sensitivities are not of equal interest. Only techniques that perform with very high sensitivity would be clinically acceptable since missing a cancer (false negative) is generally considered much worse that an unnecessary benign biopsy (false positive). Thus, the partial area under the curve (partial AUC) for the 90-100% sensitivity range is sometimes computed instead of the area under the full curve [54-56]. Notice that the partial AUC was normalized by dividing by the constant

$(1 - TPF_0)$, where $TPF_0 = 0.9$. Thus, the chance value is 0.05 while the value for a perfect system is 1.0.

Throughout this dissertation, the ROC curves were calculated non-parametrically (except as described in the error surface analysis Section 1.3.2). When semi-parametric fits were used, the area under the ROC curve was denoted A_Z and the partial area index was denoted partial AUC index. P-values and standard deviations on the AUC and partial AUC (trapezoid rule) were estimated by bootstrap sampling on the decision variable with 10,000 samples [57] (except as described in the error surface analysis Section 1.3.2). Non-parametric ROC analysis was performed using custom software written by members of our laboratory ("droc" and "bsp" programs, Brian Harrawood).

Results are also sometimes shown in terms of particular operating points. We chose to look at the specificity at 98% sensitivity. This sensitivity point was chosen in analogy with "probably benign breast lesions", which are a group of lesions that some advocate should be managed by short-term follow up rather than biopsy because the frequency of cancer among them is low (< 2%), the cancers are generally identified during the follow-up surveillance, and the cancers initially considered probably benign are still identified at an early stage [58-61]. This means that some radiologists would consider it acceptable to delay the diagnosis of a small percentage of breast cancers, thus the focus on 98% sensitivity. Notice that additional work would be required to determine if delaying the diagnosis of the 2% of malignancies misclassified by a CAD system would result in little change in outcome for the patient as is argued for probably benign lesions.

1.3.2 Perceptron Error Surface Analysis

1.3.2.1 Background

In recent years, many breast cancer CAD studies have focused on the use of artificial neural network (ANN) models. ANN models have been developed to predict malignancy among suspicious breast lesions based upon mammographic and history findings [14, 37-40]. Most networks for CAD are based on classic feed-forward, error-back-propagation paradigms, which are trained to minimize mean squared error (MSE) using a gradient descent technique. For a general discussion of such ANNs, please see Section 3.4. In "weight space," the ANN modifies a vector of weights, descending down a multi-dimensional error surface in search of the global minimum in MSE. Once trained, however, these ANNs are often evaluated according to other more clinically relevant measures of performance from receiver operating characteristic (ROC) analysis. Such measures include the ROC area index (A_Z) and the partial area index corresponding to the portion of the ROC curve in the high sensitivity range of 0.9 to 1.0 [54-56]. More information on the A_Z and partial AUC index measures is provided in the overview of ROC analysis in Section 1.3.1.

The relationship between these three performance measures is not well defined, but there is a generally unstated assumption that a classifier trained to optimize MSE will also tend to optimize other measures such as $A_Z$and partial AUC index. The validity of that assumption was questioned in recent studies. In one study, Kupinski *et al.* compared the performance of neural network models trained in the conventional manner (*i.e.*, minimize mean squared error) versus those trained by a niched Pareto multi-objective genetic algorithm (NP-GA) that simultaneously maximized sensitivity and specificity

[62]. Using simulated XOR (exclusive or) data, they found that the ROC curve generated by NP-GA training was superior to that resulting from conventional training for both a perceptron (logistic discriminant) and an artificial neural network. Kupinski *et al.* also compared the performance of a conventionally trained perceptron to a NP-GA trained perceptron for the task of breast mass detection [35]. They found that while there was no significant difference between the models in terms of A_Z, the NP-GA trained perceptron was significantly better in terms of the partial AUC index. In other words, the weights identified by minimizing the mean squared error were inferior to those identified by the NP-GA in terms of the model's performance at high sensitivities.

A related study demonstrated that different feature selection techniques might be preferred when partial AUC index is considered instead of A_Z. Sahiner *et al.* compared the performance of linear discriminant analysis classifiers using features selected by a linear discriminant analysis technique versus a genetic algorithm [34]. The former provided better A_Z but the latter had better partial AUC index.

All of the above studies examined the behavior of either linear or logistic discriminants. Although highly simplified compared to ANNs, these techniques are important for several reasons. First, their simplicity allows easy analysis of the relatively few parameters. For example, previous work at this institution presented a typical ANN for breast cancer CAD with 16 inputs and 10 hidden nodes, characterized by 180 weight parameters [23]. In comparison, the highly simplified perceptrons in this study were characterized by only four weights.

Secondly, several authors have reviewed recent studies where ANNs were applied to CAD problems, and suggested that a logistic model (such as a perceptron) would have likely provided similar performance while avoiding over-fitting problems [63, 64]. Indeed, many recent studies in the field of CAD have been based upon linear discriminant models [32, 65-67]. Any lessons learned from optimizing perceptrons would thus likely be useful to the field of CAD research.

The simple architecture of perceptrons was crucial to this study, which investigated the underlying behavior of these models by studying the error surfaces formed as a function of the parametric weights. In particular, the goal was to compare error surfaces resulting from measuring performance with MSE versus A_Z and partial AUC index.

1.3.2.2 Methods

1.3.2.2.1 Data Set

The data set for the error surface analysis consisted of 500 cases of non-palpable breast lesions from patients who had undergone excisional biopsy at Duke University Medical Center between 1991 and 1996 (see data collection form in Appendix 1). In other words, the data set consisted of a consecutive sample of actual clinical cases. Of these 500 lesions, 65% were found to be benign as a result of histopathologic diagnosis. The relatively low prevalence of disease in this data set is consistent with the literature concerning this diagnostic task [7, 68]. It is expected that models built on a clinically representative case mix will be better prepared to classify previously unseen clinical cases. The method of encoding the lesion descriptors has been previously described [23], and will only be summarized here. Expert radiologists retrospectively reviewed the patient films and recorded ten mammographic findings according to the Breast Imaging and Reporting Data System (BI-RADS™) lexicon [10], as well as other patient history data including the age. These findings were encoded into numeric values and used as input features in order to predict the known biopsy outcome of benign vs. malignant.

Please note that this preliminary study above was based on 500 cases from Duke University. This should be contrasted with the description of the much larger data set used throughout the remainder of the dissertation (Section 1.4).

1.3.2.2.2 Network Architecture

Even with the simplified architecture of a perceptron, it was still important to reduce the dimensionality of the input features in order to permit visualization and analysis. The number of inputs was therefore pruned to the three most important ones, based upon previous work in identifying the most important input findings for this diagnostic problem [23, 69]. The BI-RADS ™ findings used were mass margin and calcification morphology. In addition, a single patient history variable, age, was used. All features were scaled to the range of 0 to 1. This 3-input perceptron is shown in Figure 1-1. The perceptron had one weight per input (W1, W2, and W3) and a bias term (W4). The dot product of input vector and the weight vector is passed through a nonlinear activation function to produce the output. The inputs were the two BI-RADS™ findings, calcification morphology (weight W1) and mass margin (weight W2), and patient age (weight W3). The outputs of the perceptron range from 0, which indicates a benign lesion, to 1, which indicates a malignant lesion. Perceptron learning parameters were empirically optimized to minimize MSE: learning rate and momentum of 0.05 and 1000 iterations, with each iteration defined as a complete presentation of all training cases with weight adjustment after each case.

as "error surfaces." Notice that plotting the error surface is not an optimization technique, but instead is used to show general trends in the data. For a perceptron with only two weights, the error surface may be readily plotted in the "z" or third dimension. In the current study, however, two-dimensional slices of the error surface were plotted instead of attempting to visualize the four-dimensional error surface. In a slice, two of the weights were varied to produce the surface, while the other two weights were held constant. Figure 1-2 shows an example of an error surface slice. For simplicity, in the remainder of the error surface plots, the performance function was plotted as intensity as in Figure 1-3.

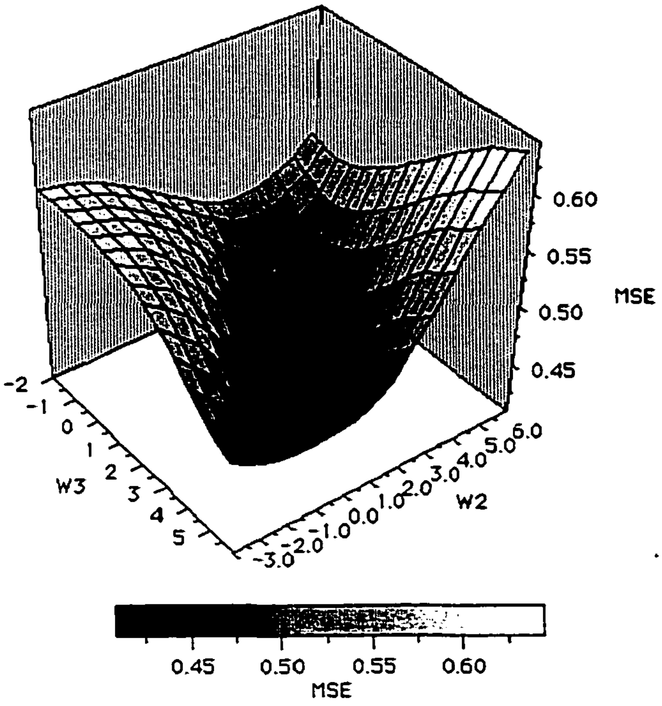


Figure 1-2. A MSE surface in weight space. The MSE is a function of the perceptron weights (W1, W2, W3, and W4). W1 and W4 were held constant.

To generate these slices, a grid search through weight space was performed. The perceptron with each combination of weights was applied to the data set. The MSE, ROC area (A_Z), or partial area index (partial AUC index) of each perceptron is indicated by intensity. More information on the A_Z and partial AUC index measures is provided in the overview of ROC analysis in Section 1.3.1. Note that while lower values for MSE indicate better performance, higher values for the performance measures A_Z and partial AUC index indicate better performance.

The ROC analysis was performed using LABROC4 software and the statistical comparisons were performed using CLABROC software, both provided by Charles Metz, Univ. of Chicago. Note that Metz provided our group with private versions of the software that he modified to calculate the partial AUC index as well as the A_Z. The software finds a maximum likelihood estimate of the area from a fit to the data. The estimates of significance include the contribution from correlation of the input data. Notice that this differs from non-parametric ROC calculations used throughout the rest of the dissertation. Please contrast this with the description provided in the overview of ROC analysis in Section 1.3.1.

The grid search over the weights was done in the vicinity of weights identified as optimal by training a perceptron to minimize the MSE of the data set. In other words, the training was used only to narrow down the reasonable range of weights over which the grid search was performed. With learning rate and momentum of 0.05 and 1000 iterations, the final weights were W1 = 1.65, W2 = 2.22, W3 = 2.56, and W4 = -3.21. In order to simplify the visualization further, the bias weight W4 was always fixed at that 'central' value. Each 2-D slice was generated by varying two of the feature weights while the bias and one remaining feature weight were held constant at the aforementioned 'central' values. The three combinations resulted in an "exploded box" showing the three-dimensional relationship between the three weights W1, W2, and W3. Each weight was varied approximately over the range of the central value +/- 150% of the central value. W1 was varied from –1.00 to 5.00. W2 was varied from –2.00 to 5.95. W3 was varied from –3.00 to 6.90.

1.3.2.3 Results

1.3.2.3.1 MSE vs. A_z

Figure 1-3 shows three two-dimensional slices through the MSE surface and Figure 1-4 shows three two-dimensional slices through the A_Z surface. Note that improved performance corresponds to minimizing MSE (darker grayscale value) but maximizing A_Z(brighter grayscale value). MSE is expected to range between 0 (perfect) and 0.5 (chance behavior), while A_Z ranges between 0.5 (chance) and 1 (perfect). While the MSE and A_Z surfaces are clearly not the same, the minimum observed on the MSE surface is in the same general location in weight space as the maximum observed on the A_Z surface. The best solution corresponding to the global minimum on the MSE surface, *i.e.* the central weights (W1 = 1.65, W2 = 2.22, W3 = 2.56, and W4 = -3.21), has MSE of 0.41 and A_Z of 0.80 ± 0.02. The best solution corresponding to the global maximum on the A_Z surface (W1 = 1.65, W2 = 1.90, W3 = 2.40, W4 = -3.21) has MSE of 0.41 and A_Z of 0.80 ± 0.02. The difference in the A_Z between the solutions was not statistically significant (two tail p = 0.14).

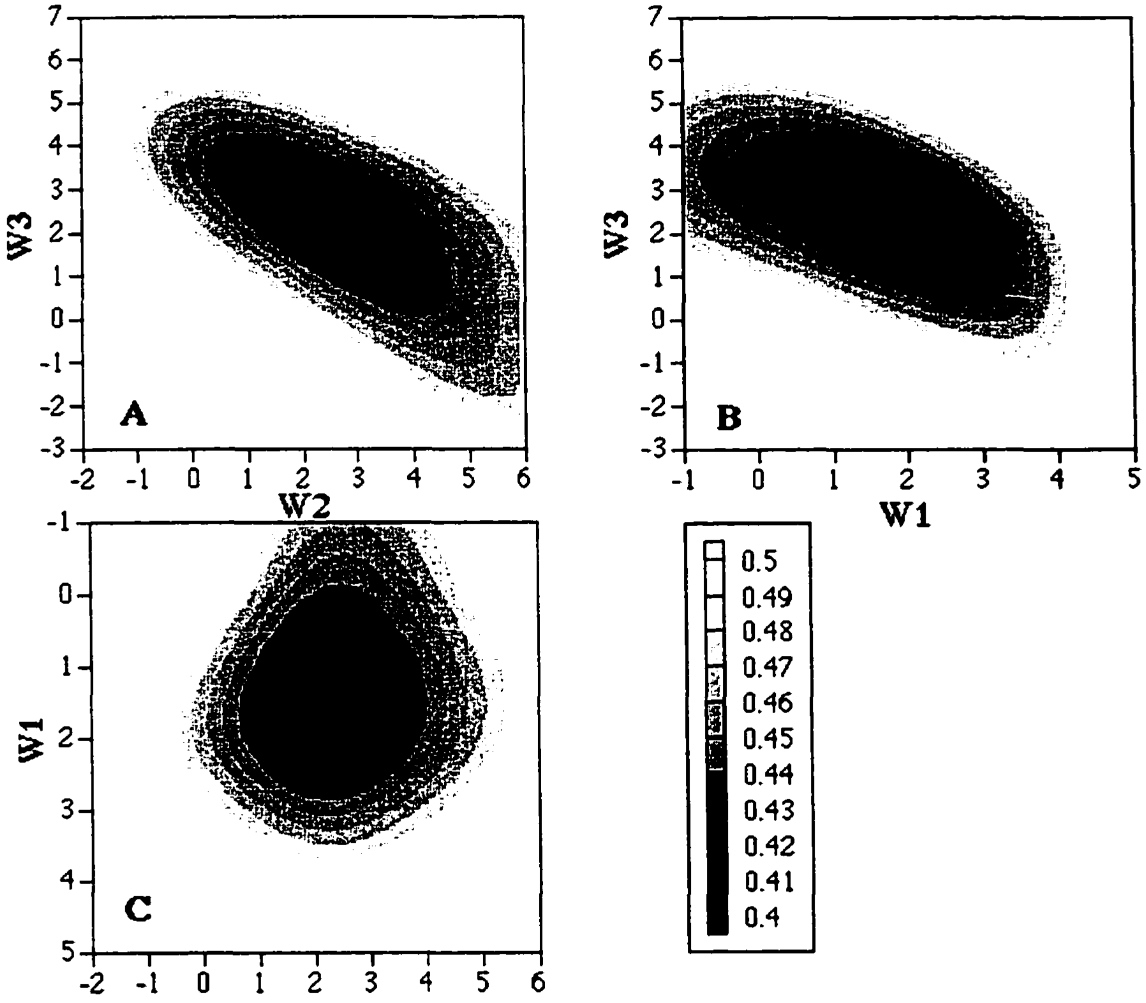


Figure 1-3. The MSE surface in weight space. The MSE is a function of the perceptron weights (W1, W2, W3, and W4). The MSE is shown as intensity. Darker gray indicates better performance. The slices through MSE surface are (A) W3 vs. W2 (B) W3 vs. W1 (C) W1 vs. W2. The subplots are arranged such that folding them into a box provides a way to visualize three of the weight dimensions.

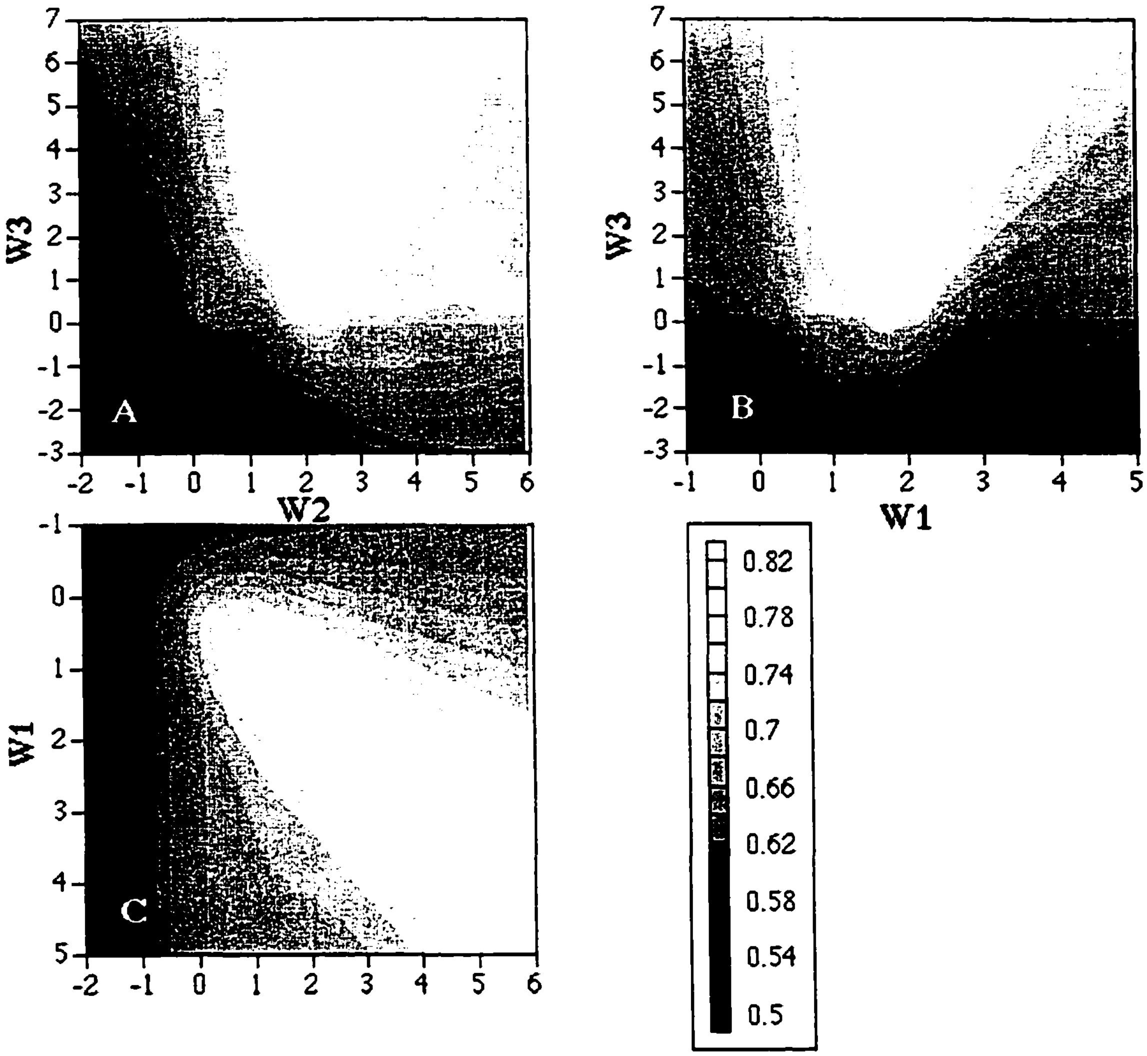


Figure 1-4. The A_z surface in weight space. The A_z is a function of the perceptron weights (W1, W2, W3, and W4). The A_z is shown as intensity. Lighter gray indicates better performance. The slices through the A_z surface are (A) W3 vs. W2 (B) W3 vs. W1 (C) W1 vs. W2.

1.3.2.3.2 MSE vs. partial AUC index

Figure 1-3 shows three two-dimensional slices through the MSE surface and Figure 1-5 shows three two-dimensional slices through the partial AUC index surface. There is less correspondence in the general appearance of the contours between the MSE and partial AUC index surfaces than was observed between MSE and A_z surfaces. The solution on the MSE surface, *i.e.* the central weights (W1 = 1.65, W2 = 2.22, W3 = 2.56, and W4 = -3.21) does not correspond to the best solution corresponding to a global maximum in the partial AUC index surface (W1 = 3.35, W2 = 2.22, W3 = 5.70, and W4 = -3.21). The solution on the MSE surface has MSE of 0.41 and partial AUC index of 0.24 ± 0.05. The solution on the partial AUC index surface has MSE of 0.58 and partial AUC index of 0.30 ± 0.04. The difference in partial AUC index between the solutions was statistically significant (two tail p = 0.006).

This same trend may be demonstrated by comparing a particular operating point, such as the specificity for 95% sensitivity. The best MSE solution resulted in a specificity of 25% while the best specificity solution resulted in a specificity of 31%. This difference in specificity at 95% sensitivity was again statistically significant (p = 0.002).

The difference in the solutions on the MSE and partial AUC index surfaces is illustrated by comparing the histograms of the outputs of the corresponding perceptrons (Figure 1-6). Since the partial AUC index measure describes the high sensitivity region of the ROC curve, the outputs of the perceptron with the highest partial AUC index tend to be higher than the outputs of the perceptron with the lowest MSE.

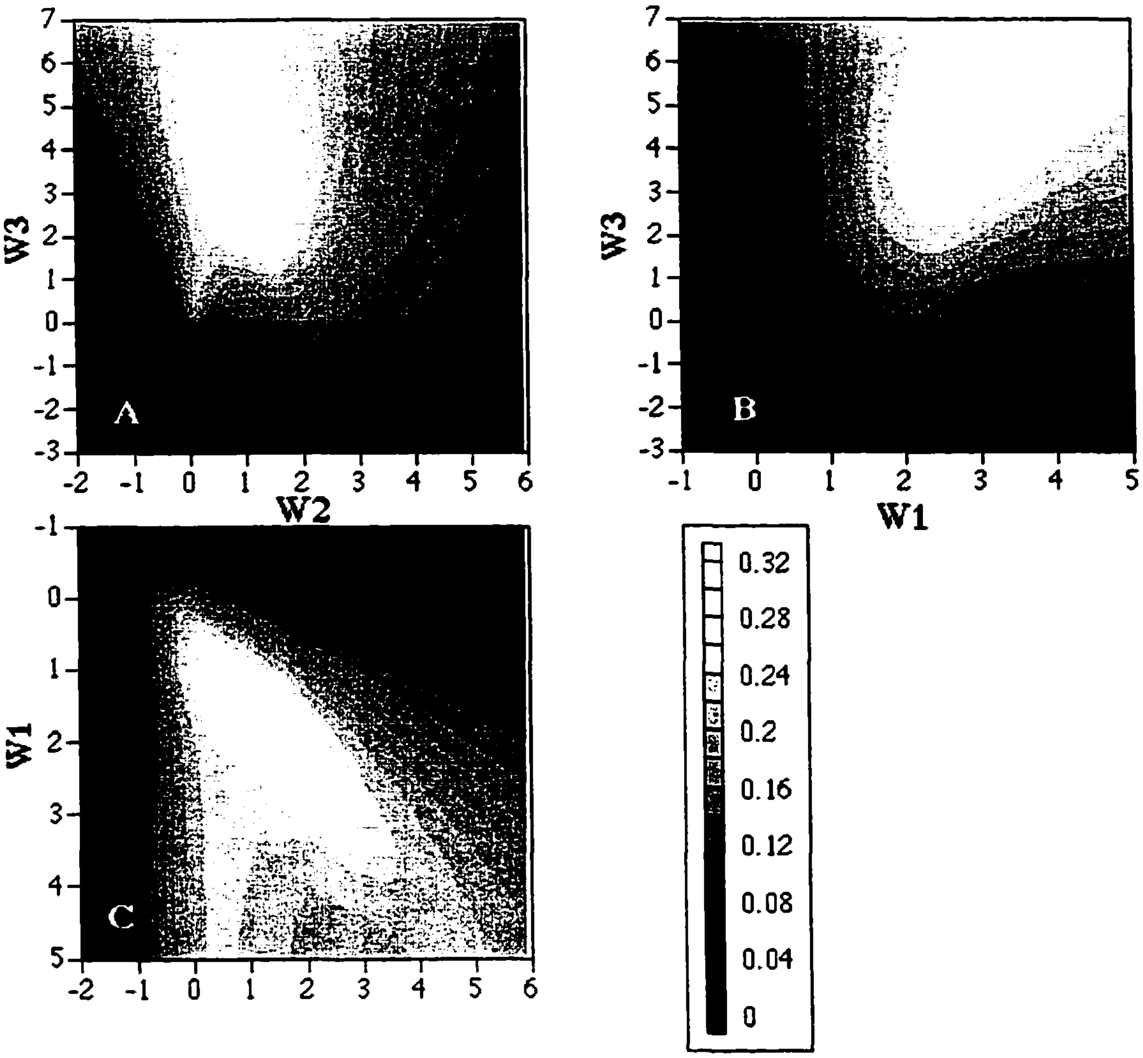


Figure 1-5. The partial AUC index surface in weight space. The partial AUC index is a function of the perceptron weights (W1, W2, W3, and W4). The partial AUC index is shown as intensity. Lighter gray indicates better performance. The slices through the partial AUC index surface are (A) W3 vs. W2 (B) W3 vs. W1 (C) W1 vs. W2.

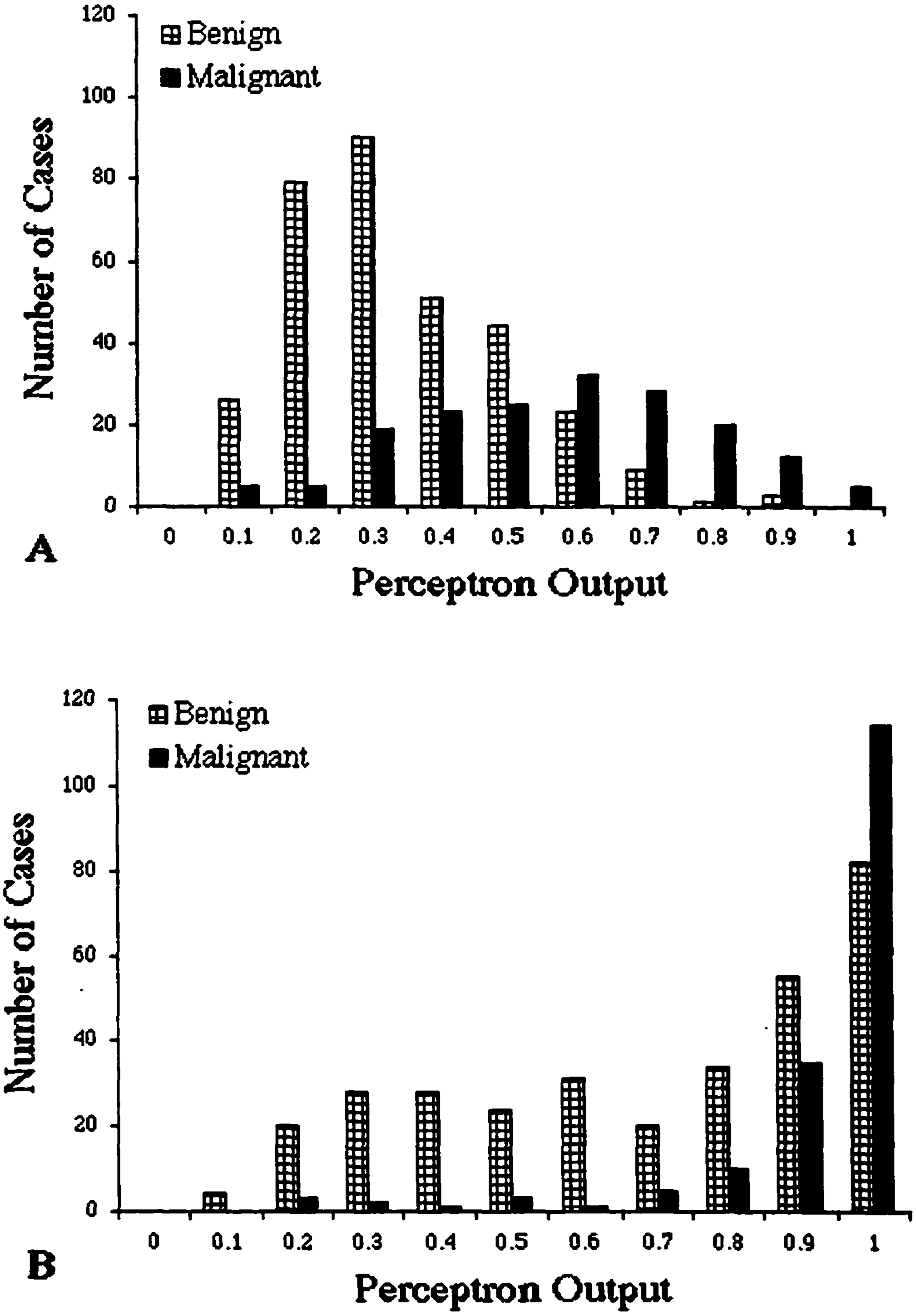


Figure 1-6. Histograms of the outputs of the perceptron for the weights that correspond to (A) the minimal MSE and (B) the maximal partial AUC index.

1.3.2.4 Discussion

The three metrics of performance studied here are important for different reasons. The MSE is the metric that many models including perceptrons and ANNs attempt to optimize directly, while the A_Z and partial AUC index have greater clinical significance. Consider the histograms (Figure 1-6) of network outputs of benign cases and malignant cases, where the network output of "0" indicates a benign lesion and "1" indicates a malignant lesion. MSE is a measure of the how close the distribution of benign cases is to a network output of "0" and how close the distribution of malignant cases is to "1". The area under the ROC curve is a measure of the overlap of the distributions. A training scheme that minimizes MSE, and so pulls the distributions to the edges, can also reduce the overlap of the distribution, and so increases A_Z. It should be noted, however, that the MSE can decrease without an accompanying change in A_Z, because each increment in A_Z can only result from the reversal of position for an adjacent pair of benign and malignant cases in the histogram. While a full convergence to MSE = 0 will also result in $A_Z = 1$, the latter can be achieved with any arbitrary MSE, as long as the two distributions do not overlap at all. In the current study, it was observed that the weights that minimized MSE also maximized A_Z.

In recent years, the sensitivity of breast cancer CAD techniques has been particularly emphasized, since there is a considerably greater cost in missing or delaying the diagnosis of an actual cancer (false negative) compared to referring a benign lesion for an unnecessary biopsy (false positive). For a range of sensitivities (*e.g.*, TPF_0 from 0.9 to 1), the partial AUC index can be thought of as an average specificity [56]. Unlike MSE and A_Z, partial AUC index is not symmetric in the sense that false negative and false positive cases do not contribute to the measure in the same way. In this work, the solution on the partial AUC index surface was found to not correspond well with the MSE solution. It should be noted that the differences in the weights that optimize MSE vs. partial AUC index may be due in part to biases inherent to the reduced amount of data that is associated with the high sensitivity region of the ROC curve.

If it is thought that A_Z is a suitable measure of performance of CAD systems for breast cancer, then this work can be interpreted as a reassurance that classifiers trained to minimize MSE may also maximize the measure of interest. This provides some justification for avoiding the task of attempting to directly optimize model performance according to A_Z. Note that optimizing for A_Z by gradient descent techniques is not straightforward since A_Z is not a continuous function.

However, if partial AUC index corresponding to a given high level of sensitivity is a better measure of the quality of CAD systems for breast cancer, then this work demonstrates that a classifier trained to minimize MSE may provide an inferior solution. Alternative methods of identifying good weights for a perceptron or multilayer network should be considered, such as evolutionary computing techniques that employ stochastic optimization.

1.3.2.5 Conclusion

In this dissertation, CAD models were evaluated using the ROC measures AUC and partial AUC. However, it should be noted that the CAD models were trained to optimize other performance measures. The perceptron example described above demonstrates that one should not assume that models trained to optimize non-ROC performance measures provide optimal solutions in terms of ROC performance measures.

Please note that the error surface analysis employed semi-parametric fits to the ROC curve while the results shown in later chapters are all based on non-parametric versions of the ROC curves. Please also note that the error surface analysis was based on 500 cases from Duke University. This should be contrasted with the description of the data set used throughout the remainder of the dissertation (Section 1.4).

1.3.3 Sampling

Two kinds of data sampling [70] were used in this study: bootstrap sampling [57] and k-fold cross-validation. As described in Section 1.3.1, bootstrap sampling was used to perform statistical tests on the ROC metrics. Cross-validation was used to the address the issue of model generalization.

Bootstrap sampling refers to sampling from the data many times (*e.g.*, 10000) with replacement. Bootstrap sampling was performed on the model outputs in order to estimate the standard deviation on ROC metrics such as the AUC. Notice that this is different from sampling on the model inputs in which a new model would have been built for each sample.

In k-fold cross-validation, the data are split into "k" non-overlapping sets. A model is trained on k-1 of the sets and tested on the held-out kth set. This is usually repeated until each of the k-sets has served as the held out set and the performance reported is the average performance over the k-sets. A special case of k-fold cross-validation is k = N, where N is the number of cases. In k = N, also called leave-one-out or round-robin sampling, a model is trained on N-1 cases and tested on the Nth cases and this is repeated until all the cases have been held out once. Although there are actually N separate models, the model outputs on the held-out cases were treated as if they came from a single model for purposes of ROC analysis.

Cross-validation was used in two ways in this study. First, the data set was randomly partitioned into two halves (Section 1.4). The first half was used for cluster analysis and model building. The second half was reserved for final validation of the results observed on the first half (Section 7). Second, in training models (Sections 3 and 5), round-robin sampling was used on the training half of the data. Notice that in the cluster analysis (Section 2) no additional sampling was performed; all of the cases in the training set were used.

Round-robin sampling alone is insufficient as the results can still be prone to bias. In particular, if the round-robin results are used to guide the selection of the parameters for a model (as they were in this study), then the round-robin results reported may be optimistic and performance may be lower when the model is tested on an independent evaluation set. This is why we chose to further verify our results using a held-out evaluation set. This is a more rigorous approach to the role of sampling in model evaluation than is frequently taken in the field of breast cancer computer-aided diagnosis.

1.4 Data Set

The data consisted of 4435 breast lesions pooled from three independent data sets (Duke, UPenn, DDSM). It is important to note that this represented the culmination of a decade-long data collection effort involving many members of the research group including faculty, students, and staff. This effort began prior to this dissertation project but was successfully concluded here. Each of the three component data sets were already among the largest known for this type of data, so the pooled data set is likely to be the largest available for some time.

For each lesion, the benign or malignant status from pathologic diagnosis was known. The overall malignancy fraction was 43%. The data were randomly partitioned into two sets. The training data set consisted of 2258 cases and the evaluation set consisted of 2177 cases. The training set was used for cluster analysis (Section 2) and for model building (Sections 3, 4, 5, 6). The evaluation set was used for final model validation (Section 7). The breakdown of the cases by training/evaluation set, institution, and malignancy is shown in Table 1-1 and Table 1-2 (see also Sections 2.2 and 7.1).

The first data set consisted of 1468 non-palpable, mammographically suspicious breast lesions that underwent biopsy (core or excisional) at Duke University Medical Center from 1990 to 2000 (see data collection form in Appendix 1). A total of 1530 cases were collected over several discontinuous time periods, but were collected consecutively within each time period. Of the 1530 cases, 61 were removed because it was not certain that they were non-palpable, leaving 1468 cases. Expert mammographers described each case using the Breast Imaging and Reporting Data System (BI-RADS™) lexicon [10]. The cases collected from 1990 to 1996 were read retrospectively and the cases collected from 1996 to 2000 were read prospectively. Each of the cases was read by one of 7 readers. When a lesion could be described by multiple descriptors (*e.g.*, pleomorphic and punctate), the mammographers were requested to report the descriptor that was most suspicious for malignancy (*e.g.*, pleomorphic).

Table 1-1. Institutional composition of the training and evaluation sets.

	Training Set	Evaluation Set	Total
Duke	751 (33%)	717 (33%)	1468 (33%)
UPenn	501 (22%)	487 (22%)	988 (22%)
DDSM	1006 (45%)	973 (45%)	1979 (45%)
Total	2258 (100%)	2177 (100%)	4435 (100%)

Table 1-2. Biopsy outcome composition of the training and evaluation sets.

	Training Set	Evaluation Set	Total
Benign	1276 (57%)	1273 (58%)	2549 (57%)
Malignant	982 (43%)	904 (42%)	1886 (43%)
Total	2258 (100%)	2177 (100%)	4435 (100%)

The second data set consisted of 988 mammographically suspicious breast lesions that underwent excisional biopsy at the University of Pennsylvania Medical Center from 1990 to 1997. The data collection procedures have been previously described [71]; in particular, we presume that the lesions in this data set were non-palpable, based upon the description of a data set of which this cases are a subset [12]. Each of the cases was read by one of 11 expert mammographers who described each case using the BI-RADS™ lexicon [10]. When a lesion could be described by multiple descriptors (*e.g.*, pleomorphic and punctate), the mammographers were requested to report·the descriptor that was most suspicious for malignancy (*e.g.*, pleomorphic).

The third data set consisted of 1979 biopsy-proven breast lesions from the Digital Database for Screening Mammography (see Appendix B) [72]. The DDSM contains screening mammograms obtained from 1988 to 1999 at Massachusetts General Hospital, Wake Forest University School of Medicine, Sacred Heart Hospital, and Washington University in St. Louis School of Medicine. The details of the case selection process were not clearly spelled out by Heath *et al.* [72], but since screening mammograms were used, presumably the lesions were non-palpable. A lesion was defined as any object recorded in a "*.overlay" file that had a "pathology" value. From benign volumes 1-14 and cancer volumes 1-15, there were 3693 overlay files from which 4029 lesions were extracted. Cases with a pathology value of "unproven" (37) or "benign, no call back" (72) were removed, leaving 3920 cases. Only the mediolateral oblique (MLO) views were used, resulting in 2001 cases. Two cases were identified as duplicates and were removed, leaving 1999 cases. The patient age information was extracted from the corresponding "*.ics" files. Twenty cases were removed due to problems with the age value (*e.g.*, age = -1005 or the same patient in a single study was reported to be different ages), leaving 1979 cases. Expert mammographers described each case using the BI-RADS™ lexicon [10]. Lesions that were described by multiple descriptors were encoded for our purposes using the descriptor that was most suspicious for malignancy (according to Table 1-3).

Specifically, the six BI-RADS™ features collected describe the mass margin, mass shape, calcification morphology, calcification distribution, associated, and special findings. In addition to the BI-RADS™ findings, the patient age was also collected resulting in a total of seven input findings describing each case. For cluster analysis and model building, missing values were encoded as zero. Over 4435 cases (1468 Duke, 988 UPenn, 1979 DDSM) and 7 features, a total of 31,045 values were collected. Of those 31,045 values, only 113 were missing (< 0.4%).

Each BI-RADS™ feature was encoded using uniformly scaled rank ordered categories. For example, when a mass is present for a case, the mass margin can take on one of five values: well circumscribed (1), microlobulated (2), obscured (3), ill-defined (4), or spiculated (5). The encoding of the BI-RADS™ findings (on the original scale) is shown in Table 1-3. Histograms of the features are shown in Appendix 3 (see also Section 2.2.3 with regards to patient age).

It is important to note that to the lesions in this database were non-palpable, to the best of our knowledge. In routine clinical practice, palpable lesions are usually not considered appropriate for short-term follow-up imaging. Thus, any CAD recommendations for follow-up made for the cases in this database do not represent ones that would have been discounted by the clinician purely on the basis of palpability.

Table 1-3. Encoding of the BI-RADS™ features.

Mass Margin	Mass Shape	Calcification Morphology	Calcification Distribution	Associated Findings	Special Findings
0 – no mass	0 – no mass	0 – no calcifications	0 – no calcifications	0 - none	0 – none
1 – well circumscribed	1 - round	1 – milk of calcium like	1 - diffuse	1 – skin lesion	1- intramam. lymph node
2 – microlobulated	2 - oval	2 - eggshell or rim	2 - regional	2 - hematoma	2 – asymmetric breast tissue
3 - obscured	3 - lobulated	3 - skin	3 - segmental	3 – post surgical scar	3 – focal asymm density
4 – ill-defined	4 - irregular	4 - vascular	4 - linear	4 - trabecular thickening	4 – tubular density
5 - spiculated		5 - spherical or lucent centered	5 - clustered	5 – skin thickening	
		6 - suture		6 – skin retraction	
		7 - coarse		7 – nipple retraction	
		8 – large rod-like		8 – axillary adenopathy	
		9 - round		9 – architectural distortion	
		10 - dystrophic			
		11 - punctate			
		12 - indistinct			
		13 - pleomorphic			
		14 – fine branching			

1.5 Summary

In Section 1 we provided an overview of breast cancer and mammography (Section 1.1) and the role that computer-aided diagnosis can play (Section 1.2). In particular, the purpose of this study was to investigate modular and ensemble CAD systems for reducing the number of benign biopsies performed (Section 1.2.1). We described the importance of ROC analysis (Section 1.3.1) and sampling (Section 1.3.3) in evaluating breast cancer CAD systems. In this study, considerable attention was paid to the issue of model generalization; thus, the data were partitioned into training and evaluation halves and round-robin sampling was used in building models on the training half. Moreover, we presented a case study of the relationship between performance metrics from ROC analysis and one commonly used in developing breast cancer CAD systems (Section 1.3.2). Finally, we supplied a detailed description of the data set used for the remainder of this dissertation (Section 1.4). It is worth noting that the database used in this study was very large and was comprised of cases mixed from multiple institutions. In Section 2, we explore methods of partitioning the data into groups, which is needed for developing a modular CAD system.

2 Cluster Analysis to Identify Groups in Breast Cancer CAD Database

2.1 Overview and Motivation

Some groups of interest are known to exist in this kind of data set, notably the groups of benign and malignant lesions. In fact, our primary goal is to develop methods to aid in predicting whether lesions are members of the benign group or the malignant group. In machine learning [73-75], this is referred to as a supervised learning task; we wish to learn how to predict something we generally don't know (malignancy status) from something we do know (lesion description, patient age) based on a set of labeled examples. A related problem is the unsupervised learning task in which we wish to learn something from a set of unlabeled examples. Generally the goal of unsupervised learning is cluster analysis, *i.e.*, to answer the question, what groups or clusters naturally exist in the data?

Why look for clusters when one is ultimately trying to solve a supervised problem? There are three motivations behind cluster analysis of a breast cancer computer-aided diagnosis database. First, cluster analysis can reveal trends in the data that were previously unknown (or under-appreciated) that may be valuable. In particular, the performance of a general model for predicting malignancy status can be evaluated in terms of its performance across the clusters, which represent subsets of patients defined by certain values of the input features. Second, cluster analysis can be used to directly predict the malignancy status. After groups are identified in the data, the prediction for a new case could be based on the biopsy outcome of the cases in the cluster to which the new case belongs (compare to Case-Based Reasoning, Section 3.5). Third, cluster analysis could serve as the first stage of a "divide-and-conquer" approach to breast cancer

CAD. To foreshadow, the third goal provided the greatest motivation for this work, though ultimately it was the first that netted the most interesting results.

The idea behind "divide-and-conquer" modular approaches is to break the problem down into smaller, simpler problems that will be easier to solve. A modular system uses multiple classifiers to solve a classification problem by partitioning the input space into smaller domains, each of which is handled by a local model [41]. The local models can be thought of as experts for a particular kind of case. Such approaches may be justified in light of recent results in this field (Section 1.2.1).

2.2 *A priori* Subsets

Modular breast cancer CAD systems based on *a priori* partitions of the data have shown promise in other studies [42-48]. Also, we already have significant knowledge about this problem. Thus, we first examined clusters based on *a priori* partitions of the data. Such *a priori* partitions can take advantage of the wealth of clinical knowledge or intuitively meaningful groupings of cases, but may be excessively biased if the knowledge is incomplete or incorrect.

In this data set, there are three *a priori* partitions of particular interest: institution (Section 2.2.1), lesion type (Section 2.2.2), and patient age (Section 2.2.3). In Section 4 we describe the performance of global CAD models over these partitions and in Section 5 we describe local models for these partitions that form modular systems.

2.2.1 Institution

Since the data were pooled from multiple institutions, we were interested in what differences exist in those sources of data (see Section 1.4). A significant concern in breast cancer CAD research is whether a model built on data from one institution will generalize well to data from another institution, thus eliminating the need for separate models for each institution.

Table 2-1 and Table 2-2 show the breakdown of the 2258 training cases by institution and biopsy outcome (see Section 1.4). Notice that 45% of the cases were from the DDSM set [72] and that the DDSM set was apparently designed to have approximately 50% prevalence, so it had a higher fraction of malignant cases than would be seen in a random sample of cases at this clinical decision point.

Table 2-1. Institutional composition of the training set.

Duke	751 (33%)
UPenn	501 (22%)
DDSM	1006 (45%)
Total	2258 (100%)

Table 2-2. Biopsy outcome and institutional composition of the training set. The fraction of cases that were benign vs. malignant was clearly dependent on the institution from which the cases were collected ($p < 0.01$, Chi-square test for independence).

	Benign	**Malignant**	**Total**
Duke	491 (65%)	260 (35%)	751 (100%)
UPenn	301 (60%)	200 (40%)	501 (100%)
DDSM	484 (48%)	522 (52%)	1006 (100%)
Total	1276 (57%)	982 (43%)	2258 (100%)

2.2.2 Lesion

Most breast biopsies are performed on lesions that present mammographically as either a mass or a cluster of microcalcifications [11]. CAD systems for detection generally perform better on calcifications than on masses, as shown in two recent review articles [25, 76] and a recent study of the ImageChecker® System from R2 Technology, Inc. (Sunnyvale, CA) [77]. CAD systems for diagnosis that are based on features automatically extracted from the images are typically designed for either masses or calcifications alone. We are unaware of any previous attempts to compare the performance on masses and calcifications within a single study. Given the differences in databases and CAD (diagnosis) techniques, it is not possible to directly compare the published performances on masses and calcifications in the literature. However, it is suggestive that classification studies on masses [47, 78] report performances that are better than those reported in studies on calcifications [31, 79]. CAD systems for diagnosis that are based on findings extracted by radiologists are often trained and evaluated over heterogeneous data sets including both masses and calcifications and the performances on masses and calcifications are not reported separately [14, 18, 37, 80]. Thus, the broad, *a priori* subsets of masses and calcifications are of particular interest.

We defined a "mass" to be any lesion for which Mass Margin > 0, Associated Findings = 0, and Special Findings = 0 (see feature encoding described in Table 1-3). Notice that this definition includes calcified masses. We defined a "calcification" as any lesion for which Calcification Morphology > 0, Mass Margin = 0, Associate Findings = 0, and Special Findings = 0.

Table 2-3 and Table 2-4 show the breakdown of the training set by lesion type, institution, and biopsy outcome (Section 1.4). Notice that the larger percentage of masses (50%) than calcifications (41%) over all was mostly due to the larger percentage of the masses in the DDSM set [72] as compared to the Duke and UPenn sets. Notice that the percent of the cases that were malignant (43%) was the same for masses and calcifications. In other words, the positive predictive value for the masses and calcifications was the same.

Table 2-3. Breakdown of the training set by lesion type and institution. The fraction of masses vs. calcifications was dependent on the institution from which the cases were collected (p < 0.01, Chi-square test for independence).

	Duke	UPenn	DDSM	Total
Mass	326 (43%)	250 (50%)	551 (55%)	1127 (50%)
Calcification	306 (41%)	236 (47%)	389 (39%)	931 (41%)
Other	119 (16%)	15 (3%)	66 (7%)	200 (9%)
Total	751 (100%)	501 (100%)	1006 (100%)	2258 (100%)

Table 2-4. Breakdown of the training set by lesion type and biopsy outcome. The fraction of cases that were benign vs. malignant was not dependent on whether the lesion was a mass or cluster of microcalcifications (p = 0.88, Chi-square test for independence).

	Benign	Malignant	Total
Mass	638 (57%)	489 (43%)	1127 (100%)
Calcification	531 (57%)	400 (43%)	931 (100%)
Other	107 (54%)	93 (47%)	200 (100%)
Total	1276 (57%)	982 (43%)	2258 (100%)

2.2.3 Patient Age

As discussed in Section 1.1, age is known to be an important risk factor. Thus, we investigated *a priori* subsets defined by patient age.

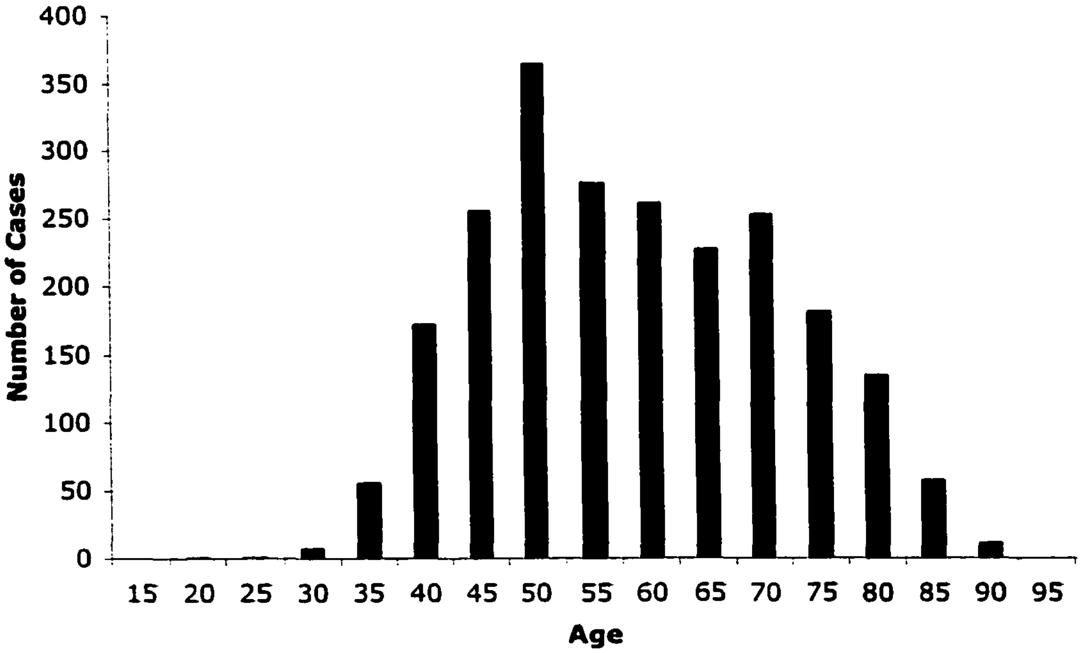


Figure 2-1. Distribution of patient age in the training set.

2.3 Cluster Profiling Methods to Aid in Interpreting Clusters

After clusters have been identified in a data set the natural next question is, what do they mean? The value of clusters in data analysis depends on our ability to summarize them and relate them to outcomes of interest. One approach that we have taken to this task to report a "profile" of each cluster. By a profile we mean a short description of what a "typical" case in the cluster is like. The profile consists of information about the typical values of the input features (BI-RADS™ and patient age). The biopsy outcome was not provided to the unsupervised machine learning techniques used to identify clusters. We also related the malignancy status of the lesions to the clusters (*e.g.*, compute the malignancy fraction for each cluster). Cluster profiles can be used to check that the clusters are consistent with what is already known and to extend our understanding of the data set and related clinical problem. The risk of any profiling technique is that in order to simplify the description of the clusters some information will be lost. It is difficult to know *a priori* what information will be important in interpreting the clusters. For this reason, it may be valuable to profile clusters by a variety of methods.

2.3.1 Mode

An obvious approach to developing cluster profiles is to compute summary statistics of the input features such as the mode or mean. The advantage of this approach is that there are a variety of summary statistics that are familiar and easy to calculate. One potential disadvantage is that such a simple implementation provides a statistic for each input feature, which might itself still be overwhelming if there is a large number of features. This was not a problem in this study as there were only seven features (see

Section 1.4). Another potential disadvantage is that computing summary statistics for each feature may ignore informative interactions between features.

We computed the mode for the BI-RADS™ features and the mean for the patients' age for each cluster. Since each of the BI-RADS™ features naturally includes "not present" as a value encoded as zero (Table 1-3), it is appropriate to introduce feature selection by eliminating features with mode of zero from the profile. Notice that one weakness of computing the mode is that it does not tell us how strongly a particular feature value dominated the others.

2.3.2 Constraint Satisfaction Neural Network

A Constraint Satisfaction Neural Network (CSNN) was also used to determine the profiles of the clusters [17, 81]. Custom software in the C language (written by Georgia D. Tourassi) was used to implement the CSNN and has been previously described by Tourassi, Markey, Lo, and Floyd [17]. Briefly, the CSNN is a Hopfield-type network of neurons arranged in a non-hierarchical way (Figure 2-2). There are symmetric, bidirectional weights between all pairs of neurons but there are no reflexive weights. The CSNN operates as a nonlinear, dynamic system that tries to reach a globally stable state by adjusting the activation levels of the neurons under the constraints imposed by the *a priori* fixed weight values. The Lypaponov energy function was used as a measure of the network stability. It was found that 1000 iterations were sufficient to achieve stability. The weights were predetermined using autoassociative back-propagation neural networks (auto-BP). In keeping with our previous work [17], the auto-BP networks were trained with a learning rate of 1.0 for 100 iterations and the root mean squared training error was approximately 0.1 (network outputs between 0 and 1).

For each cluster, a CSNN was used to generate a profile. Each category of the categorical BI-RADS™ features corresponded to a binary variable and associated neuron. For example, the mass margin with its five non-zero categories was represented by five separate neurons. Patient age was translated into a discrete variable with five levels (< 40 years, 40-50, 50-60, 60-70, > 70 years) [17]. An additional neuron was used to signify cluster membership. The activation level of the neuron indicating cluster membership was set and the other neurons were allowed to evolve until the network reached a stable state. The feature neurons that were activated defined the profile of the cluster (example shown in Figure 2-3). A profile is a list of feature values that succinctly summarizes the cluster and defines a "typical" case (*e.g.*, mass margin is well circumscribed, mass shape is round, and patient age is between 50 and 60 years). Notice that unlike common summary statistics, such as the cluster centroid, the CSNN profile implicitly includes feature selection; only features deemed relevant to the network for describing a cluster are included.

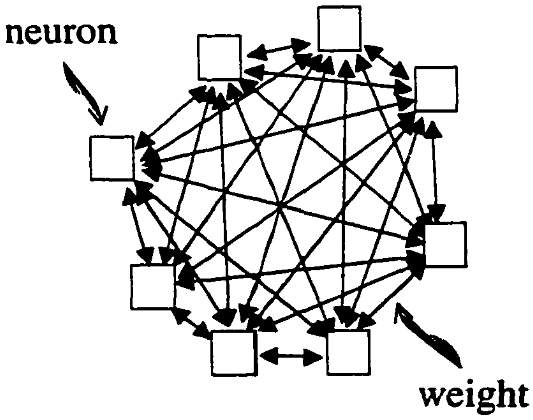


Figure 2-2. Schematic of the constraint satisfaction neural network (CSNN). Notice that the neurons are fully interconnected with no reflexive weights.

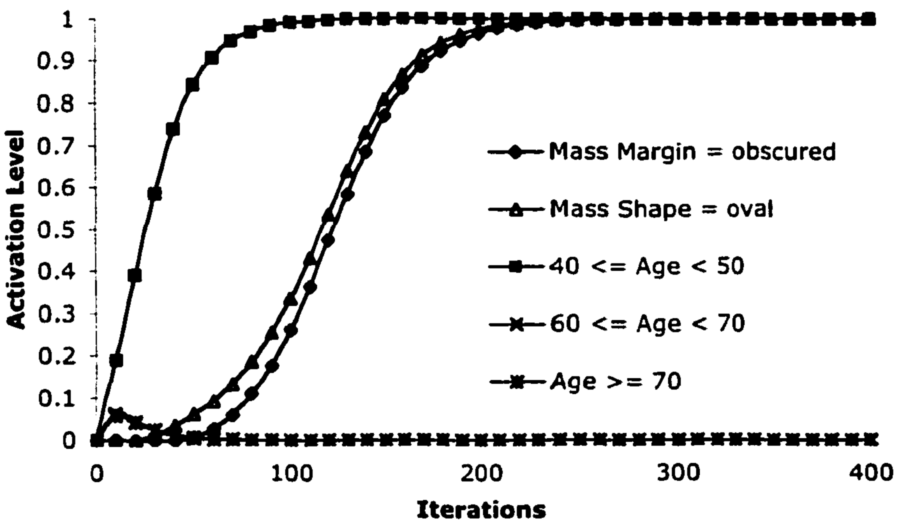


Figure 2-3. Activation levels as a function of the number of iterations for the neurons in the CSNN for cluster #6 identified by the SOM (Section 2.4.2). Most of the neurons never activate. The neurons corresponding to 60 ≤ age < 70 and age ≥ 70 fire briefly, but quickly die off. The final activation levels define the profile: mass margin = obscured, mass shape = oval, and 40 ≤ age < 50.

2.4 Unsupervised Learning Methods for Cluster Analysis

Machine learning is a subfield of artificial intelligence that is focused on the development of algorithms that enable computers to learn from experience. One way of conceptualizing the differences among machine learning algorithms is in terms of the way feedback is given regarding the method's performance. Techniques are described as supervised learning, reinforcement learning, or unsupervised learning [73, 75]. In supervised learning, the system is provided with examples and the correct response to those examples. An example of a supervised learning system is a classifier that modifies its internal parameters such that its predictions converge toward the known responses. Supervised learning techniques are appropriate when one has many examples of correct and incorrect pairings of inputs and outputs available for training. In reinforcement learning, the system is provided with examples and is given evaluation about its performance, but is not told the correct responses. Reinforcement learning methods are commonly used in "real time" learning environments such as in the training of an autonomous robot. In unsupervised learning, the system is provided with examples but is not given any information about the correct responses. Unsupervised approaches are used to answer the question, "What natural groupings exist in the examples given?"

Three unsupervised learning methods were used to identify clusters, or groups, in the breast cancer CAD database: agglomerative hierarchical clustering followed by K-Means (Section 2.4.1), Self-Organizing Map (Section 2.4.2), and AutoClass (Section 2.4.3).

2.4.1 Agglomerative Hierarchical Clustering and K-Means

Distance-based clustering is based on the assumption that similar cases are cases that are close to each other in the input feature space. Hierarchical or non-hierarchical methods can be used to group cases that are near each other into mutually exclusive clusters. Agglomerative hierarchical clustering begins with all cases as separate clusters and merges the closest clusters until some criterion is satisfied [82-84]. One weakness of agglomerative hierarchical clustering is that it can suffer from "chaining"; that is, which clusters are merged at step k depends on which ones were merged at step k-1 [84]. Non-hierarchical methods, such as K-Means [85, 86], assign and reassign cases to clusters until some criterion is satisfied. Notice that non-hierarchical methods require the user to specify initial clusters. Non-hierarchical methods perform poorly when random initial partitions are used but perform much better when an agglomerative hierarchical method is used to determine the initial clusters [84]. We used agglomerative hierarchical clustering to determine initial clusters for K-Means.

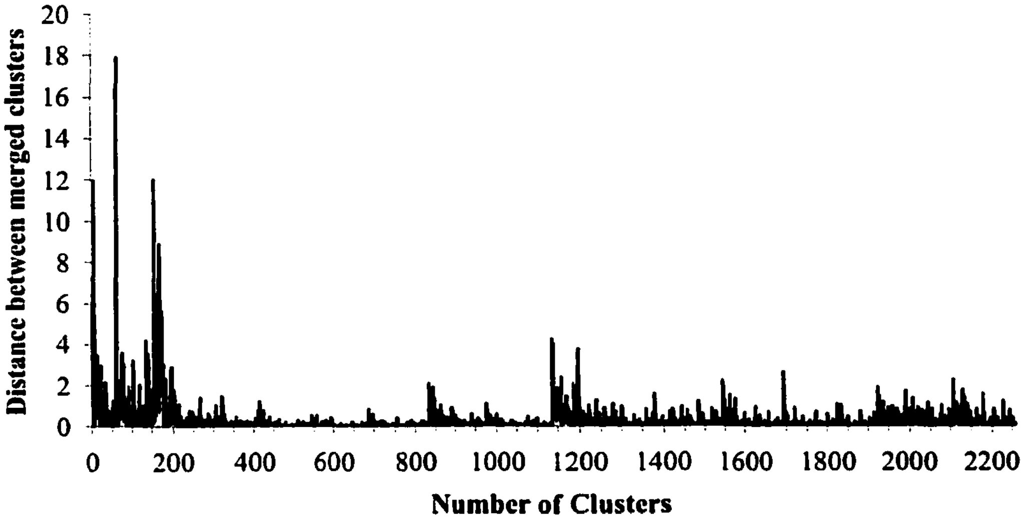
squares and exits when there is no further improvement in that criterion. The clusters from agglomerative hierarchical clustering were refined using K-Means by using the means of the clusters from hierarchical clustering as the initial centroids for K-Means.

2.4.1.2 Results

Figure 2-4 shows a plot of the distance between merged clusters versus the number of clusters from the agglomerative hierarchical clustering algorithm. We are interested in the smallest number of clusters for which very dissimilar clusters have not been wrongfully merged. Based on Figure 2-4, a cutoff of 220 was selected. However, 220 is far more clusters than was desired. Given that we are interested in clusters that could be used in the future for building submodels, it is preferable that the minimum number of cases per cluster be around 100 on average. An examination of the 220 clusters revealed that most of them (194) were very small (less than 20 case) and several were singletons (83) (Figure 2-5). Thus, only the means of the 26 clusters with at least 20 cases were initially used as starting centroids for K-Means. The K-Means algorithm failed to converge and indicated that there was an empty cluster. The centroid corresponding to the smallest cluster used from hierarchical clustering was removed and K-Means was applied again. This was repeated until the algorithm converged. In the end, the 10 largest clusters from hierarchical clustering were used as the starting centroids for K-Means.

Table 2-8 shows the summary information for the final 10 clusters that were identified by agglomerative hierarchical clustering and refined by K-Means. The percent of the cases that were malignant was quite different between the clusters and is also different from the value for the entire data set. Recall that this analysis was performed in an unsupervised fashion and that the clustering algorithms did not have access to the biopsy outcome for the cases. Table 2-8 also shows the mode profiles (Section 2.3.1) and the Constraint Satisfaction Neural Network profiles (CSNN, Section 2.3.2) for the clusters.

In examining the cluster profiles, several interesting results are apparent. First, some clusters appear to focus on recognized subtypes such as calcifications (A, B, C), masses (D, E, F), and architectural distortions (J). By inspection, we can also recognize that one of the smaller clusters (I) contains focal asymmetric densities and the other two (G, H) contain calcified masses. Moreover, the recognized subtypes are stratified across clusters by their mammographic descriptors and patient age. For example, while clusters A, B, and C all clearly include calcification cases, the women with lesions in cluster C are typically older than those clustered to A or B. Likewise, the distribution of the calcifications for lesions in cluster A was generally different than for lesions in clusters B and C.



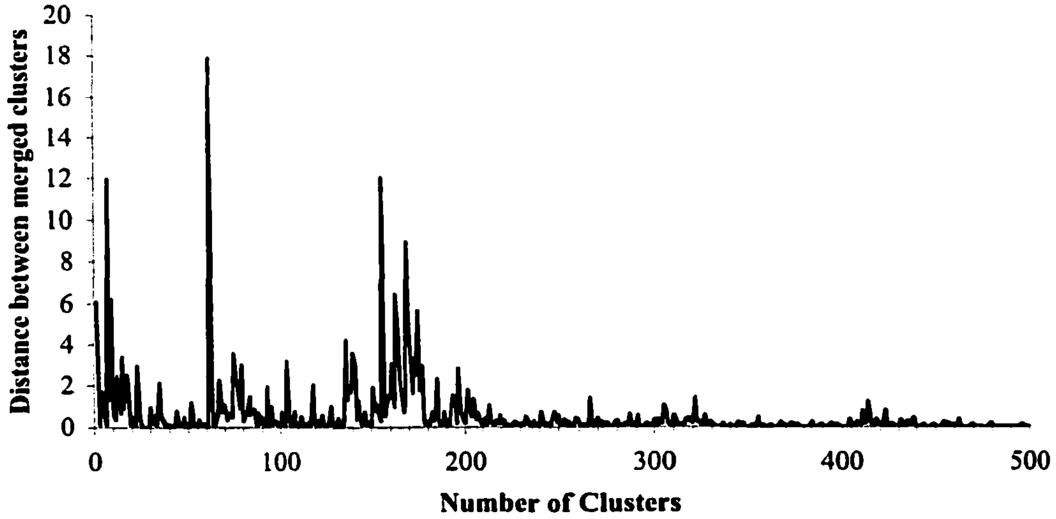


Figure 2-4. Distance between merged clusters as a function of the number of clusters in agglomerative hierarchical clustering. A cutoff of 220 was chosen for further analysis.

Table 2-8. The summary characteristics of the final 10 clusters that were identified by hierarchical clustering and refined by k-means. Notice that the percent of the cases that were malignant was quite different between the clusters and is also different from the value for the entire data set. Mode profiles (Section 2.3.1) and CSNN profiles (Section 2.3.2) are shown, except for the clusters with less than ~100 cases.

Cluster	Number of Cases	Percent Malignant	Mode profile	CSNN Profile
A	101	51%	segmented, pleomorphic calcifications mean age = 48 years	segmented, pleomorphic calcifications 40 ≤ age < 50
B	489	35%	clustered, pleomorphic calcifications mean age = 48	clustered, pleomorphic calcifications 40 ≤ age < 50
C	360	50%	clustered, pleomorphic calcifications mean age = 69	clustered, pleomorphic calcifications 60 ≤ age < 70
D	261	24%	well-circumscribed, oval mass mean age = 61	well-circumscribed, round mass 60 ≤ age < 70
E	426	19%	obscured, oval mass mean age = 43	obscured, oval mass 40 ≤ age < 50
F	398	78%	ill-defined, irregular mass mean age = 69	ill-defined, lobulated mass age ≥ 70
G	34	79%	-	-
H	27	48%	-	-
I	66	27%	-	-
J	96	69%	architectural distortion mean age = 58	architectural distortion 40 ≤ age < 50
All	2258	43%	-	-

2.4.2 Self-Organizing Map

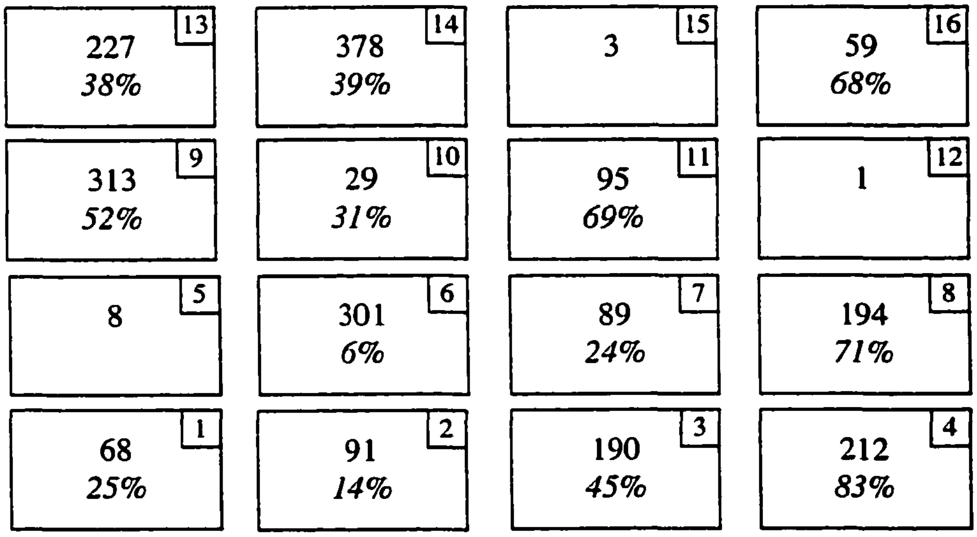
A self-organizing map relates similar cases (input vectors) to the same region of a map of neurons [88]. The distance between a case and a neuron is a measure of their similarity. After the most similar neuron is determined, that neuron and its neighbors are adjusted to have feature values closer to the matching case. The process is repeated until a stop criterion is satisfied. A cluster of cases is defined as the subset of cases that map to the same neuron.

2.4.2.1 Methods

The SOM was computed using the SOM toolbox in MATLAB® (The MathWorks Inc., Natick, MA). The basic SOM consisted of 16 neurons arranged in a single layer in a 2-D square grid of 4 by 4 neurons. For each case, the Euclidean distance between the case and each neuron was calculated based on the seven input features (see description of data set in Section 1.4). For input to the SOM, each feature was scaled by subtracting the mean and dividing by the standard deviation, resulting in each scaled feature having mean zero and standard deviation of one. After the most similar neuron was determined the neurons in its neighborhood were identified. The neighborhood of a neuron was defined as all the neurons within a given link distance of the matched neuron. The link distance is the number of links that must be taken to get from one neuron to another. All the neurons in the neighborhood were adjusted to have feature values closer to the current case. The amount that the neuron weights were adjusted was controlled by the learning rate. In the first phase, a relatively fast learning rate (0.9) that decreased over time (to 0.02) was used and the link distance threshold was varied from the maximum value to a specified low value (1.0). In the second phase, a slow learning rate (0.02), which further decreased over time, and a specified low link distance threshold (1.0) were used. The learning rates and distance threshold values used were the default values for the SOM toolbox.

2.4.2.2 Results

Figure 2-6 illustrates the arrangement of the neurons in the self-organizing map (SOM). The set of cases that were mapped to a neuron defined a cluster. Figure 2-6 shows the number of cases that were mapped to each neuron, *i.e.*, the number of cases in each cluster. The fraction of the cases in each cluster that were malignant is also shown in Figure 2-6 (bottom number in italics). The malignancy fraction is not shown for the clusters with fewer than 10 cases (#5, 12, and 15), on the assumption that no meaningful conclusions can be drawn from such a small number of cases. Recall that the SOM was not provided with the biopsy outcome information. The differences in the malignancy fraction are a reflection of differences in the BI-RADS™ features and patient age between the clusters. The overall malignancy fraction was 43%.



13 227 *38%*	14 378 *39%*	15 3	16 59 *68%*
9 313 *52%*	10 29 *31%*	11 95 *69%*	12 1
5 8	6 301 *6%*	7 89 *24%*	8 194 *71%*
1 68 *25%*	2 91 *14%*	3 190 *45%*	4 212 *83%*

Figure 2-6. Index of the neurons in the 4 x 4 map. Each neuron defined a cluster. The number of cases that were mapped to each neuron, *i.e.*, the number of cases in each cluster (normal type), and the fraction of the cases in each cluster that were malignant (*italics*) is shown. Malignancy fraction data not shown for the clusters with very few cases. Over all, 43% of the cases were malignant.

Figure 2-7, Figure 2-8, Figure 2-9, and Figure 2-10 show the effects that changing the SOM architecture have on the clusters identified. Alternative architectures allow one to vary the number of neurons as well as their topological layout, thus potentially allowing for variations in the complexity of the model. One alternative to a 4 x 4 SOM is a smaller but still square 3 x 3 SOM. In Figure 2-7, the clusters of the 3 x 3 and 4 x 4 SOMs are compared using a bubble plot. For each case, the neuron it mapped to was determined for each SOM. The number of cases for each pair of clusters between the two SOMs was plotted; the size of the circle indicates the number of cases. The more large bubbles that are present in such a plot, the more the SOMs agreed on the clustering of the cases. Similarly, Figure 2-8 shows the comparison with a 5 x 5 SOM. Linear trends (*i.e.*, bubbles lining up along the diagonals) indicate that the same cases are being mapped to the same region (*e.g.*, upper right-hand area) in the two SOMs. In addition to square topologies, other layouts were also investigated which utilized approximately the same number of neurons. Figure 2-9 shows the comparison to a 2 x 8 SOM and Figure 2-10 shows the comparison to a three-dimensional SOM of 2 x 3 x 3 neurons. Note that these two SOMs had approximately the same number of neurons as the 4 x 4 square SOM.

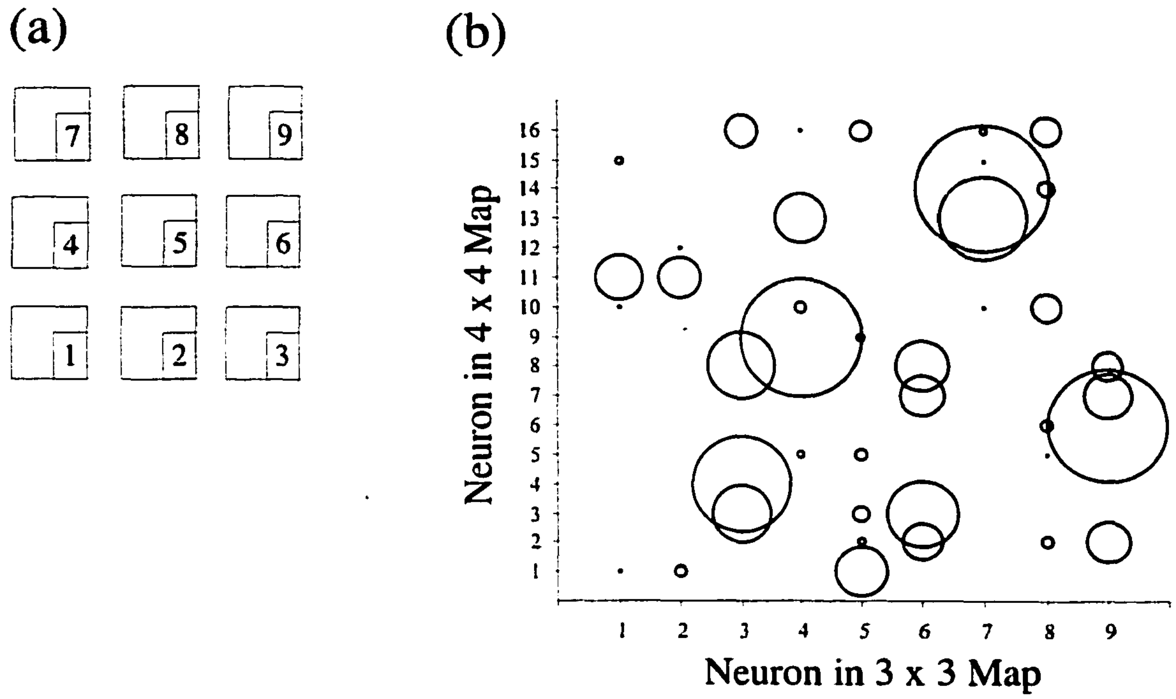


Figure 2-7. (a) The index of the neurons in the 3 x 3 map. (b) Comparison of the clusters identified by the 3 x 3 and 4 x 4 SOMs.

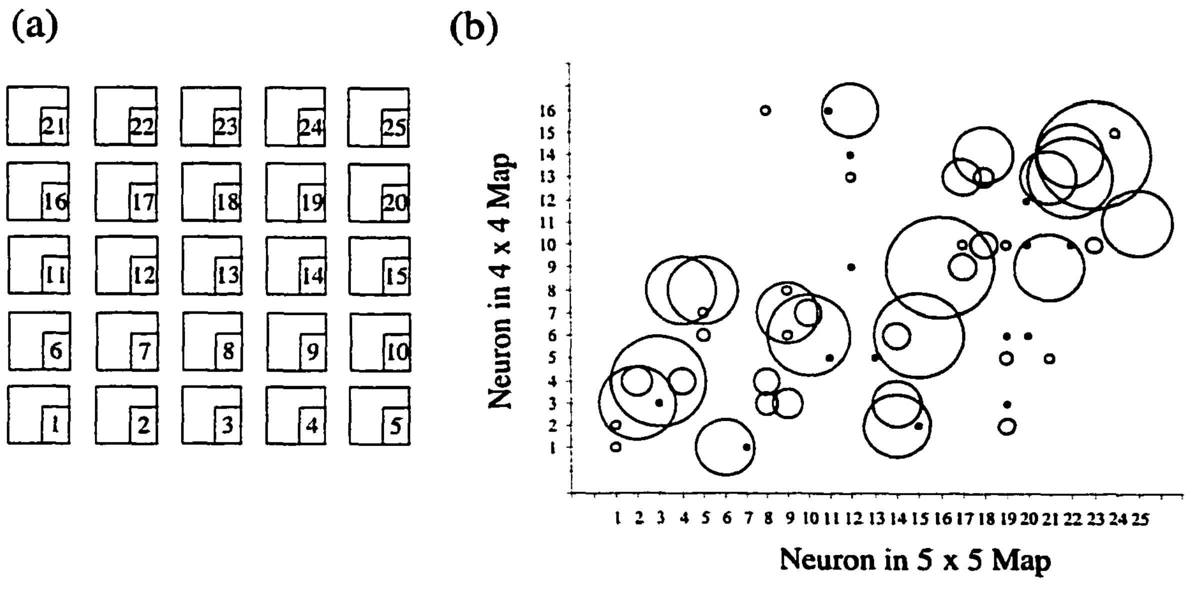


Figure 2-8. (a) Index of the neurons in the 5 x 5 map. (b) Comparison of the clusters identified by the 5 x 5 and 4 x 4 SOMs.

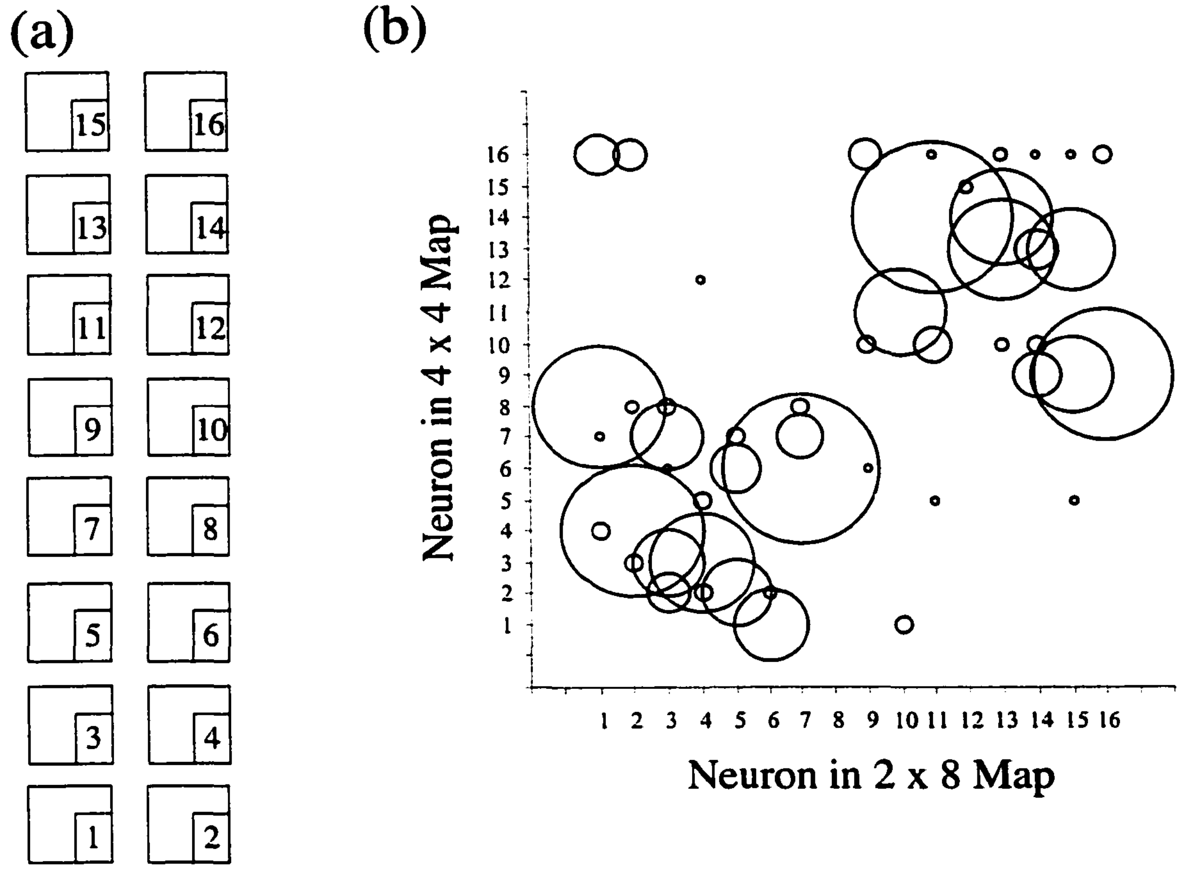


Figure 2-9. (a) Index of the neurons in the 2 x 8 map. (b) Comparison of the clusters identified by the 2 x 8 and 4 x 4 SOMs.

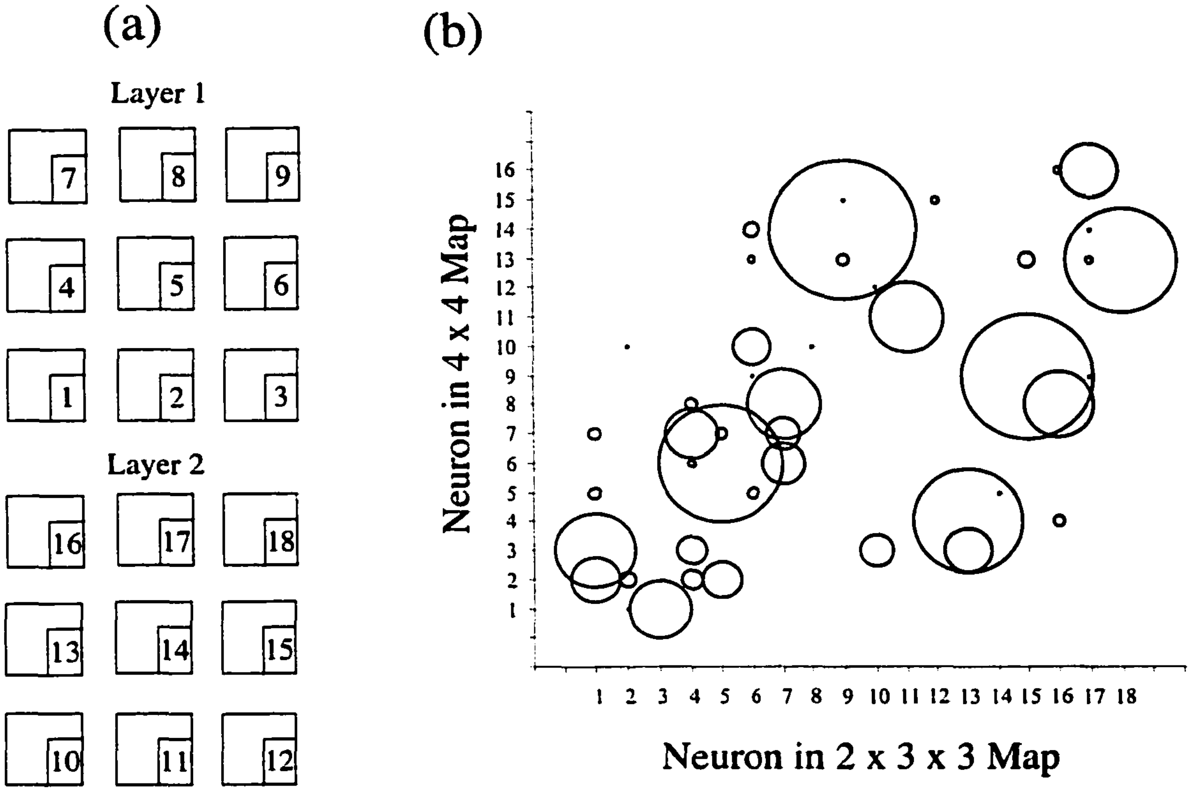


Figure 2-10. (a) Index of the neurons in the 2 x 3 x 3 map. (b) Comparison of the clusters identified by the 2 x 3 x 3 and 4 x 4 SOMs.

The SOM can be used to generate a malignancy prediction [89]. For each case, the prediction was the fraction of the cases that were malignant in the cluster that the case was mapped to by the SOM. Notice that using this approach limits the number of operating points on the non-parametric ROC curve to the number of clusters with unique malignancy fractions minus one (Figure 2-11). The performance of the back-propagation artificial neural network (BP-ANN, Section 3.4) is shown for comparison. The performance at the highest sensitivities was comparable. In particular, at 98% sensitivity the SOM operates with 0.26 ± 0.03 specificity and the BP-ANN operates with 0.25 ± 0.03 specificity (p = 0.93).

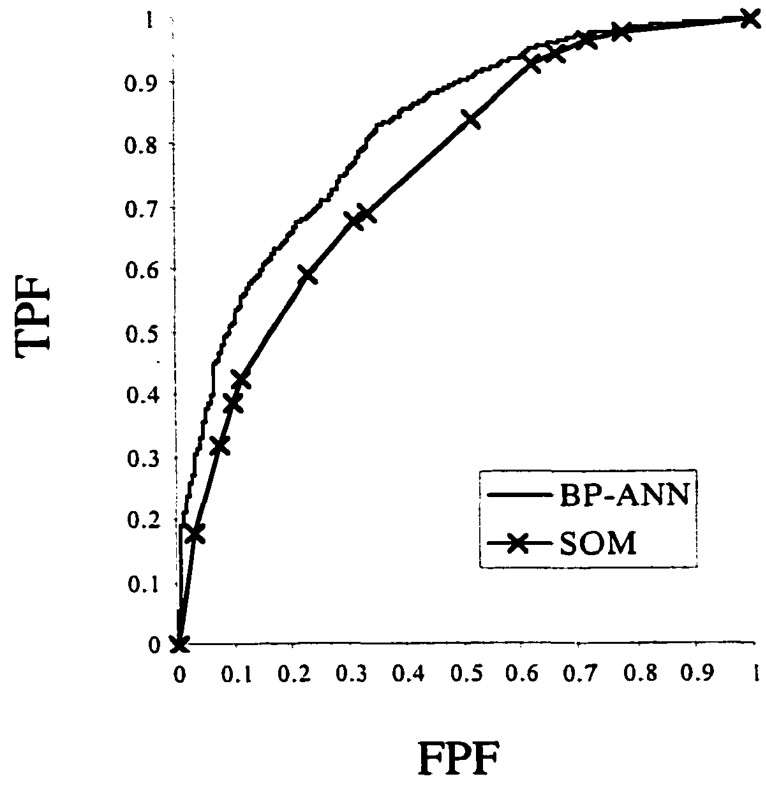


Figure 2-11. ROC curves for the SOM and BP-ANN. For each case, the prediction from the SOM was the fraction of the cases in the cluster it belonged to that were malignant. For the clusters with less than 50 cases, the over all malignancy fraction (0.43) was used.

For the 4 x 4 SOM, the cluster profiles generated by the constraint satisfaction neural network (CSNN, see Section 2.3.2) are shown in Figure 2-12. Each cell in the table represents the feature categories that were dominant or most strongly associated with the cases matching that cluster. Profiles were not computed for the clusters with very few cases. The mass cases are distributed over neurons #2, 3, 4, 6, 7, and 8. The profiles of neurons #9, 13, 14, and 16 indicate that those clusters contain microcalcifications. Neuron #1's profile indicates that that cluster is comprised of focal asymmetric densities. Note that the profile for neuron #10 includes only the age variable.

The profile for neuron #11 reveals that the lesions in that cluster are architectural distortions.

An alternative approach to generating cluster profiles is to compute summary statistics such as the feature mode (or mean for real-valued features such as age). Figure 2-13 shows the mode profiles (see Section 2.3.1) of the clusters identified by the 4 x 4 SOM. For the most part, there is considerable agreement between the CSNN and mode profiles. Most of the differences correspond to adjacent categories in the features (Table 1-3) where the CSNN has selected the second most prevalent value for the profile. However, using multiple methods to summarize the clusters may be beneficial. For example, the CSNN profile of neuron #16 (Figure 2-12) doesn't include any mass features yet the feature mode profile (Figure 2-13) shows that the mass features are usually non-zero. In fact, inspection of the cases in the cluster defined by neuron #16 reveals that they are calcified masses. Conversely, the CSNN profile for neuron #10 (Figure 2-12) includes only the age variable while the mode profile's (Figure 2-13) inclusion of values for the calcification variables may be misleading for this small cluster (N = 29) where there is little dominance by any single value.

13 clustered, pleomorphic calcifications 50 ≤ age < 60	14 clustered, pleomorphic calcifications 40 ≤ age < 50	15	16 clustered, pleomorphic calcifications 50 ≤ age < 60
9 clustered, pleomorphic calcifications 60 ≤ age < 70	10 40 ≤ age < 50	11 architectural distortion 40 ≤ age < 50	12
5	6 obscured, oval mass 40 ≤ age < 50	7 ill-defined, oval mass 50 ≤ age < 60	8 ill-defined, irregular mass 50 ≤ age < 60
1 focal asymmetric density 50 ≤ age < 60	2 well-circumscribed round mass 50 ≤ age < 60	3 obscured, oval mass 60 ≤ age < 70	4 ill-defined, lobulated mass age ≥ 70

Figure 2-12. CSNN profiles (Section 2.3.2) for the clusters identified by the 4 x 4 SOM. A cluster "profile" provides a description of a "typical" case in the cluster. Profiles were not computed for neurons #5, 12, and 15, which had very few cases mapped to them.

13 clustered, pleomorphic calcifications mean age = 56	14 clustered, pleomorphic calcifications mean age = 45	15	16 ill-defined, irregular mass clustered, pleomorphic calcifications mean age = 58
9 clustered, pleomorphic calcifications mean age = 70	10 regional, punctate calcifications mean age = 49	11 architectural distortion mean age = 59	12
5	6 well-circumscribed, oval mass mean age = 42	7 obscured, oval mass mean age = 52	8 ill-defined or spiculated, irregular mass mena age = 53
1 focal asymmetric density mean age = 58	2 well-circumscribed round mass mean age = 57	3 well-circumscribed, oval mass mean age = 70	4 ill-defined, irregular mass mean age = 73

Figure 2-13. Mode profiles (Section 2.3.1) for the clusters identified by the 4 x 4 SOM. A cluster "profile" provides a description of a "typical" case in the cluster. Profiles were not computed for neurons #5, 12, and 15, which had very few cases mapped to them.

2.4.2.3 Discussion

Neurons #5, #12, and #15 (Figure 2-6) correspond to clusters with very few cases. Inspection of the cases mapped to these clusters revealed that the cases are rare for this database. They included cases with findings that were seen with a very low prevalence in the set (*e.g.*, special finding of intramammary lymph node) or reflected incomplete or inconsistent data (*e.g.*, the calcification morphology was described but calcification distribution feature was not reported). Together these three clusters comprise only 0.5% of the cases. Therefore, no further analysis was performed on these clusters.

Considerable variability was seen in the fraction of the cases that were malignant from cluster to cluster. Several clusters had malignancy fractions that were notably different from the fraction of the entire data set (43%). One of the major goals of computer-aided diagnosis of breast cancer is to identify very likely benign cases as candidates for follow-up in lieu of biopsy, in order to reduce the number of benign biopsies. Therefore, the clusters with very low malignancy fractions (*e.g.*, neuron #6 with 6% malignant) are dominated by such very likely benign lesions and may be of particular interest for further studies. It is possible to use the clusters and their malignancy fractions directly as a tool for predicting biopsy outcome [89]. For each case, the prediction was the fraction of the cases that were malignant in the cluster that the case was mapped to by the SOM (Figure 2-11). For very high sensitivities, this prediction scheme (98% sensitivity, 0.26 ± 0.03 specificity) was competitive with the back-propagation artificial neural network (98% sensitivity, 0.25 ± 0.03 specificity, $p = 0.93$). The SOM prediction method in conjunction with the CSNN profiling method has the potential advantage that physicians may understand the intuition behind it better than they do the BP-ANN, which is often seen as a "black box". The SOM prediction method, similar to a case-based reasoning system, predicts the probability of malignancy of a new case by reporting the fraction of similar cases that were found to be malignant [18]. The SOM prediction method could also potentially be used in an ensemble of classifiers. If the outputs of two classifiers are not strongly correlated, it is possible that they could be combined to produce a classifier that is better than either of its component classifiers.

The effects of the changing the SOM architecture were investigated (Figure 2-7, Figure 2-8, Figure 2-9, and Figure 2-10). As indicated by the presence of large circles in the bubble plots, the SOMs with similar architectures showed substantial agreement in clustering the data. Moreover, the presence of linear trends in Figure 2-8, Figure 2-9, and Figure 2-10 suggest that similar SOM architectures result in similar geometric relationships between clusters. These data argue that the clustering is relatively insensitive to the SOM architecture for this problem.

Figure 2-12 lists the CSNN profiles (Section 2.3.2) for the clusters identified with the SOM. The successful separation of *a priori* known, coarse lesion types (masses, clustered microcalcifications, focal asymmetric densities, and architectural distortions) provided some quality assurance of the clustering. Clusters were further identified within the general group of mass lesions, reflecting different combinations of the mass margin, mass shape, and patient age variables. The cluster profiles that included calcification features showed stratification of the general group of calcification lesions only by patient age and not any of the calcification findings. Notice that while some features may not be considered useful by the CSNN for profiling individual clusters, it is possible that they could be useful to other summarizing techniques or to methods designed to describe the differences between clusters.

An alternative approach to characterizing the clusters is to calculate summary statistics for each of the features. Figure 2-13 shows the mode (Section 2.3.1) for each of the BI-RADS™ features and the mean of the patient age for each cluster. In general, there is good agreement in the cluster descriptions obtained from the mode and CSNN profiles.. However, they are not identical. The most notable differences are for neurons #10 and #16, which show the advantages and disadvantages respectively of the fact that the CSNN method inherently includes feature selection.

It may be easier to interpret a CSNN profile, with typically only a few dominant features per cluster, than to interpret as many summary values as there are input findings. Note as well that the CSNN takes into the account interdependencies between the features, while the summary statistics were based on each feature independently. CSNN profiles or summary statistics can be used to quickly sort through the results of a clustering technique, but additional characterization may be appropriate for clusters of particular interest.

2.4.3 AutoClass

AutoClass is a public-domain classification program (http://ic-www.arc.nasa.gov/ic/projects/bayes-group/autoclass/) based on the Bayesian solution to the finite mixture problem [90, 91]. Mixture models are based on the idea that the cases available are a sample from a mixture of distributions [82]. The probability that a case belongs to a certain group is estimated based on estimates of the parameters of the individual distributions in the mixture. With AutoClass, each case is not assigned to a class; a probability of membership for each class is returned.

AutoClass approaches the classification problem by breaking it into two parts: (1) determining the classification parameters for a given number of classes and (2) determining the number of classes. The posterior distribution of the classification parameters (class parameters of distributions in the mixture, class probabilities) is the product of the prior distribution of the parameters and the likelihood function, divided by a normalizing constant. The prior distribution describes our prior knowledge about the classification parameters, which for our purposes is an uninformative prior reflecting our lack of knowledge. The likelihood function describes the likelihood of observing a case (vector of features) given the number of classes, the class probabilities, and the class parameters. The normalizing constant is the integral of the non-normalized posterior distribution. Once the posterior distribution of the classification parameters is determined for all possible numbers of classes, the classification parameters are integrated out to give a posterior distribution for the number of classes.

2.4.3.1 Methods

The 2258 training cases were used on the original scale (Section 1.4). The six BI-RADS™ features were modeled as coming from a multinomial distribution. In other words, the ordering of the feature categories (Table 1-3) was not used. The patient age was modeled using a normal distribution. The patient age was defined to have a minimum value of zero and a relative error of 5%.

AutoClass is a statistically-based clustering method, unlike agglomerative hierarchical clustering followed by K-Means (Section 2.4.1) and the SOM (Section

2.4.2). One consequence of this is that instead of a hard clustering, each case is assigned a probability of being in a certain cluster, such that the probabilities across all clusters summed to100%. For this analysis, a case was considered to belong to the cluster for which its cluster membership probability was highest.

2.4.3.2 Results

AutoClass identified 5 clusters in the data (Table 2-9). Notice that the percent of the cases that were malignant varied notably between the clusters, even though the biopsy outcome was not provided to AutoClass. As with agglomerative hierarchical clustering followed by K-Means (Section 2.4.1) and the SOM (Section 2.4.2), the clusters focused on recognized subtypes such as calcifications (α), masses (β,γ), and calcified masses (ε). An interesting difference is that AutoClass did not stratify the calcifications across multiple clusters. The mode profile for cluster δ indicated zero for all of the BI-RADS™ features (no findings). Upon inspection of cluster δ, it was seen that 59 / 141 = 42% of the cases had Associated Findings, 53 / 141 = 38% of the cases had Special Findings, and 6 / 141 = 4% of the cases had both Associated and Special Findings. In other words, only 16% of the cases in cluster δ had neither Associated nor Special Findings.

While AutoClass provides the probability of cluster membership for the most likely cluster for each case, little variability was seen for this problem (Figure 2-14). In fact, 2079 / 2248 = 92% of the cases were assigned to their most probable cluster with a probability greater than 95%. However, even with such a limited range the probability of cluster membership may be informative. In particular, a threshold on the probability of cluster membership could be applied such that a case would only be considered a member of the cluster if the cluster membership probability was greater than 95%. For cluster β, this would result in a smaller cluster more homogeneous in malignancy (β', N = 544, 14% malignant). This suggests that in a cluster of mostly benign masses, some malignant masses were recognized as being less probable members of the cluster. Notice that since the average age of the cases in β with probability less than or equal to 95% (62 years) was higher than that of the cases in β with probability greater than 95% (51 years) that the average age would be reduced from 54 years for β to 51 years for β' (Figure 2-15). In other words, the less probable members of cluster β were more frequently malignant lesions in older women as compared to the more probable members of cluster β.

Table 2-9. Summary characteristics of the five clusters identified by AutoClass. The mode profiles are shown (Section 2.3.1).

Cluster	N	Percent Malignant	Mode Profile
α	961	43%	clustered, pleomorphic calcifications mean age = 56 years
β	685	21%	well-circumscribed, oval mass mean age = 54 years
γ	395	81%	spiculated, irregular mass mean age = 63 years
δ	141	43%	no findings mean age = 57 years
ε	76	63%	clustered, pleomorphic calcifications ill-defined, irregular mass mean age = 57

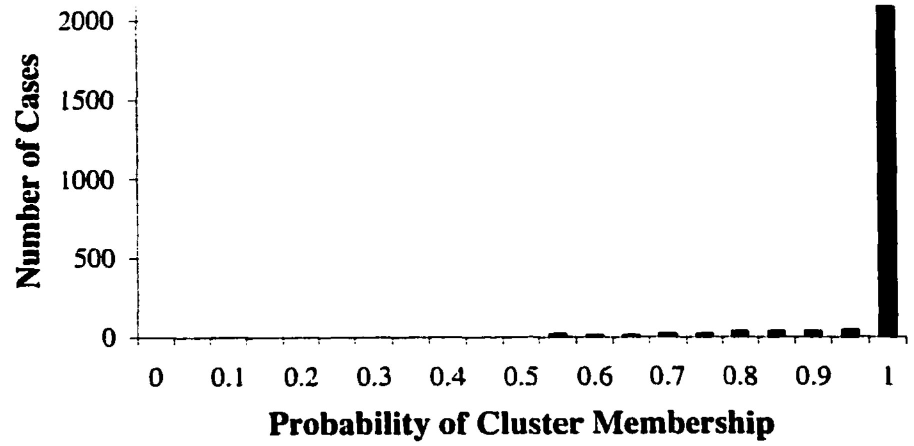


Figure 2-14. Distribution of the number of cases assigned to their most probable cluster with specified probability. Notice that the vast majority of cases were assigned to their most probable cluster with very high probability.

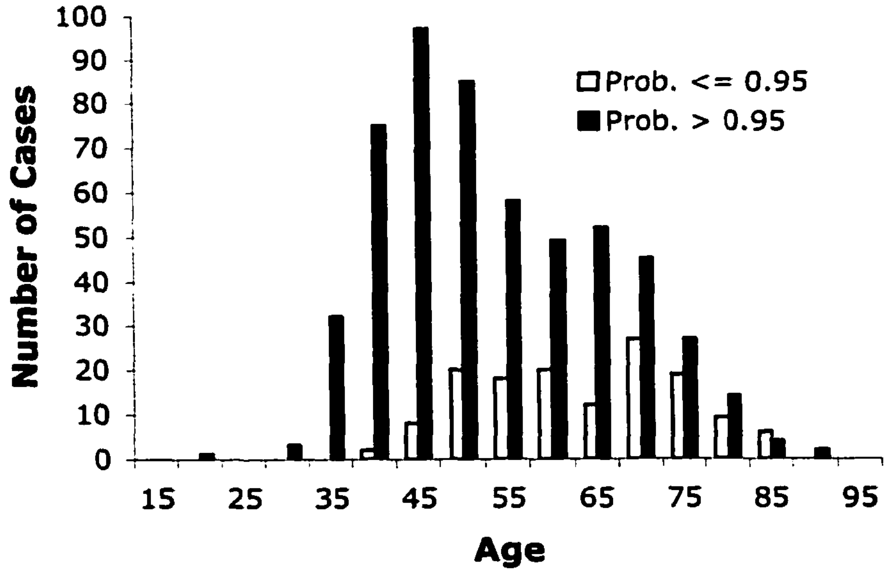


Figure 2-15. Age distribution for the cases in cluster β according to whether the probability of cluster membership was above or below 95%.

2.5 Comparison of Clustering Methods

Figure 2-16 shows a comparison of the clusters identified by the SOM (Section 2.4.2) and agglomerative hierarchical clustering followed by K-Means (Section 2.4.1). Notice that the clusters identified by agglomerative hierarchical clustering followed by

K-Means were manually sorted, so the linear trend in the plot should not be over-interpreted. The presence of large bubbles in the plot indicates that there is some agreement between the two clustering methods. This is not unexpected since the same measure of similarity (Euclidean distance) was used by both clustering methods.

Figure 2-17 shows a comparison of the clusters identified by the SOM (Section 2.4.2) and AutoClass (Section 2.4.3). The presence of large bubbles in the plot indicates that there is some agreement between the two clustering methods. The vertical pattern reflects the fact that fewer clusters were identified by AutoClass than were identified by the SOM.

Of particular interest is the fact that all three clustering methods identified a cluster of usually benign masses (Table 2-10). We will revisit these clusters in the analysis of the performance of global (Section 4) and local (Section 5) models across clusters. Notice that identification of a cluster with few malignancies is valuable from the point of view of using the clustering directly for identifying likely benign lesions to spare biopsy. However, such an extreme in the percentage of cases that are malignant is not a goal for the purpose of using cluster analysis as a front-end for a modular system. There were 260 cases (6% malignant) that were in cluster E *and* 6 *and* β.

Notice that the unsupervised methods for cluster analysis were all performed on the entire training set without providing the biopsy outcome. By comparison, the supervised methods for classification in the upcoming Section 3 were all performed using round-robin sampling (Section 1.3.3) with the biopsy outcome provided.

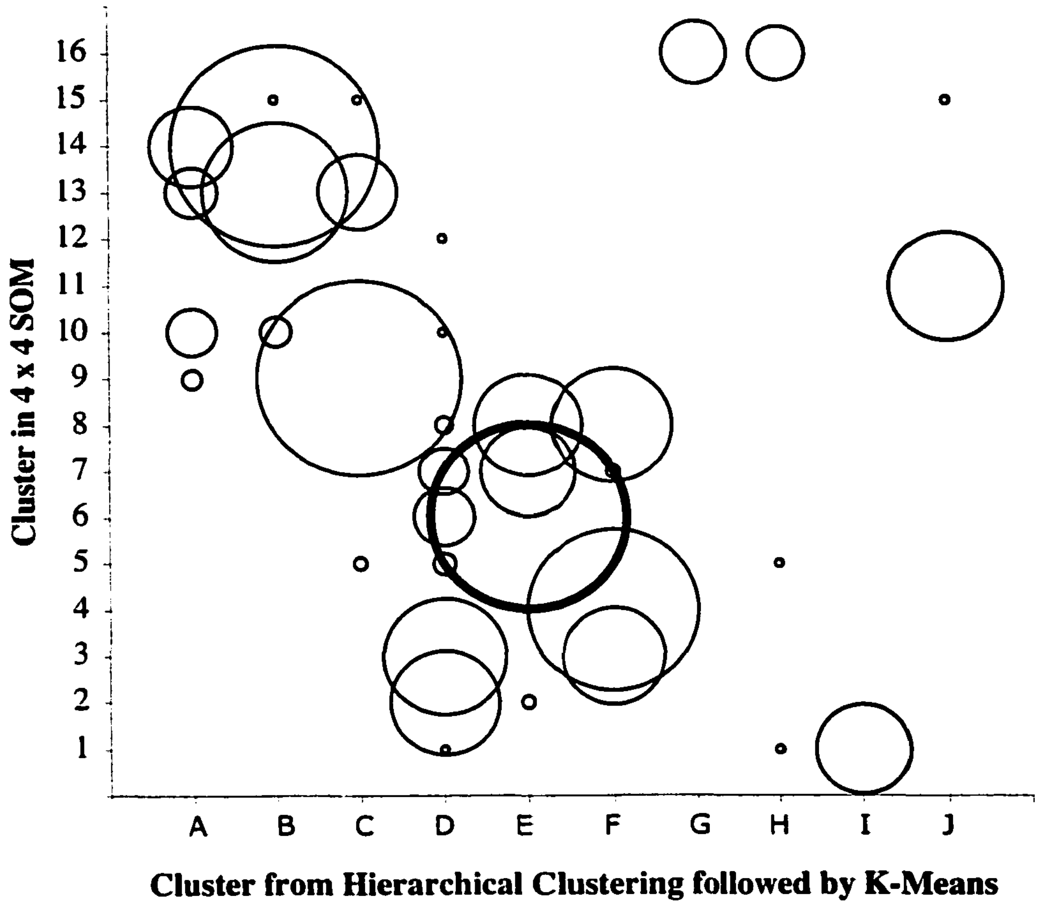


Figure 2-16. Comparison of the clusters identified by the SOM (Section 2.4.2) and agglomerative hierarchical clustering followed by K-Means (Section 2.4.1). The bubble indicating the number of cases in the intersection of clusters 6 and E is highlighted.

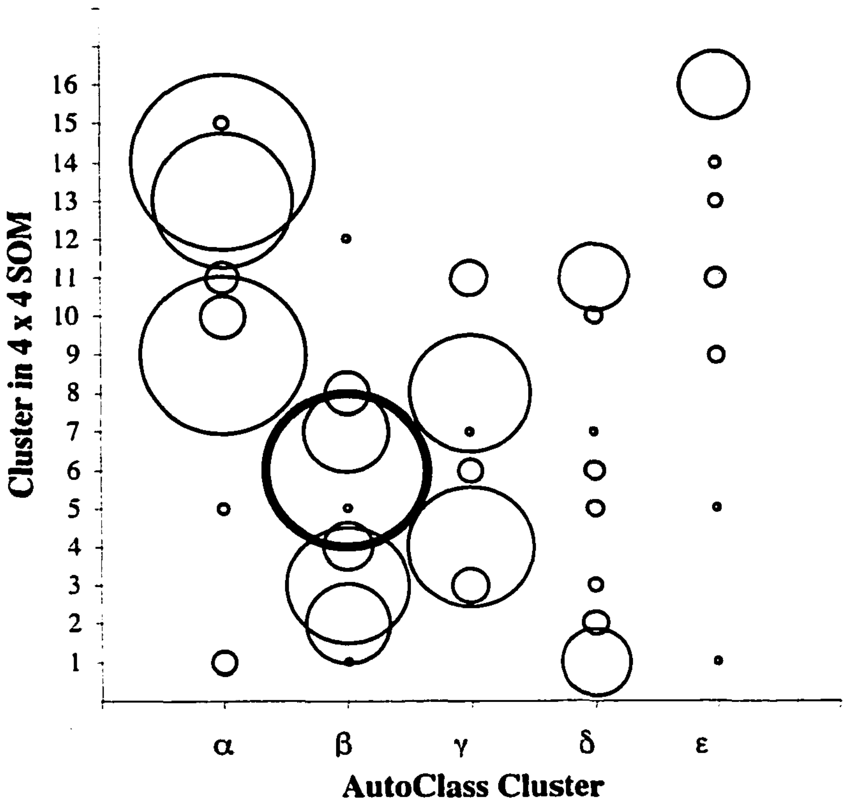


Figure 2-17. Comparison of the clusters identified by the SOM (Section 2.4.2) and AutoClass (Section 2.4.3). The bubble indicating the number of cases intersection of clusters 6 and β is highlighted.

Table 2-10. Agglomerative hierarchical clustering followed by K-Means (Section 2.4.1), SOM (Section 2.4.2), and AutoClass (Section 2.4.3) all identified a large cluster of usually benign masses.

Method	**Cluster**	**N**	**Percent Malignant**	**Mode Profile**	**CSNN Profile**
Aggl. Hierch, + K-Means	E	426	19%	obscured, oval mass mean age = 43	obscured, oval mass 40 ≤ age < 50
SOM	6	301	6%	well-circumscribed, oval mass mean age = 42	obscured, oval mass 40 ≤ age < 50
AutoClass	β	685	21%	well-circumscribed, oval mass mean age = 54 years	-

3 Global Models: Machine Learning Methods for Predicting Biopsy Outcome using the Training Set

3.1 Overview and Motivation

Several machine learning [73-75] methods were considered for the task of predicting the malignancy status from the BI-RADS™ features and patient age. These methods are all supervised learning techniques, as opposed to the unsupervised methods used for cluster analysis in Section 2. Thus, the biopsy outcome was provided to these methods while it had not been provided to the cluster analysis methods.

Why try several methods instead of just picking one? The problem is that there isn't a classification algorithm that is always superior to the alternative algorithms for all problems [74]. The nature of the problem (*e.g.*, how many training data are available) can suggest that certain approaches may be more fruitful than others, but there is no guarantee that any particular method will be the best. For this reason, we chose to investigate several methods for this task: Linear Discriminant Analysis (LDA), Support Vector Machines (SVM), Back-Propagation Artificial Neural Network (BP-ANN), Case-Based Reasoning (CBR), and Classification And Regression Trees (CART). However, a more detailed analysis was performed for BP-ANN and CBR since those models have been extensively applied to databases of BI-RADS™ features in our lab [14, 16, 18, 21, 23, 71, 92, 93].

An important characteristic of a classification algorithm is what kind of decision boundaries it can represent. In particular, some methods can only produce models with linear decision boundaries while others can produce models with either linear or non-linear decision boundaries. Linear decision boundaries can be thought of as those that are generalizations of a line in the input feature space. LDA and SVM are linear models while BP-ANN, CBR, and CART are non-linear models.

3.2 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is a classifier that forms a discriminant score (z, Equation 3-1) as a weighted (w_i) sum of the input variables (x_i)[74, 84]. The weights are determined by maximizing the ratio of the between-group sum of squares to the within-group sum of squares (Equation 3-2). The weights ($\hat{\mathbf{w}}$) that are the solution to this optimization problem are determined from the means of the input variables for the two classes ($\mathbf{m}_{benign}, \mathbf{m}_{malignant}$) and the covariance matrix ($\mathbf{S}$) of the input variables. Notice that the same covariance matrix is assumed for both classes.

Equation 3-1

$$z = \sum_i w_i x_i$$

Equation 3-2

$$\lambda = \frac{\sum_{j \in benign} (z_j - \mu_{malignant})^2 + \sum_{j \in malignant} (z_j - \mu_{benign})^2}{\sum_{j \in benign} (z_j - \mu_{benign})^2 + \sum_{j \in malignant} (z_j - \mu_{malignant})^2}$$

Equation 3-3

$$\hat{\mathbf{w}} = (\mathbf{m}_{benign} - \mathbf{m}_{malignant})'\mathbf{S}^{-1}$$

LDA can be applied in a stepwise manner to perform feature selection. The selection is based on Wilks' Lambda statistic (Equation 3-4), which is the ratio of the within-group sum of squares to the total sum of squares. In other words, the measure selects for features that minimize the within-group sum of squares (homogeneity) and maximizes the between-group sum of squares (separation). Notice that only a few of the possible combinations of features are considered and that this approach doesn't take into consideration relationships between variables that aren't in the model yet. Lack of inclusion of a feature in the model does not mean that the feature is unimportant; an important feature that is redundant with one already in the model would not be selected.

Equation 3-4

$$\Lambda = \frac{\sum_{j \in benign} (z_j - \mu_{benign})^2 + \sum_{j \in malignant} (z_j - \mu_{malignant})^2}{\sum_{j \in benign} (z_j - \mu_{malignant})^2 + \sum_{j \in malignant} (z_j - \mu_{benign})^2 + \sum_{j \in benign} (z_j - \mu_{benign})^2 + \sum_{j \in malignant} (z_j - \mu_{malignant})^2}$$

LDA is a popular model in breast cancer CAD [32, 65, 66] and has been previously applied to portions of this BI-RADS™ database [92]. Briefly, Markey *et al.* [92] used LDA to predict the biopsy outcome for 1453 cases from Duke University Medical Center with round-robin performance of $A_Z = 0.80 \pm 0.01$ and partial AUC index $= 0.28 \pm 0.03$.

3.2.1 Methods

LDA was implemented in SAS/STAT® (SAS Institute Inc., Cary, NC; "discrim" procedure). The LDA model predicted the biopsy outcome based on the seven input features. The 2258 training cases (see Section 1.4) were used to build the LDA model in a round-robin (leave-one-out) manner (see Section 1.3.3). The features were rescaled to 0 to 1 (by subtracting the minimum value and dividing by the maximum minus the minimum). The biopsy outcomes were provided as the model targets (supervised learning).

The SAS software was also used to perform stepwise LDA ("stepdisc" procedure, general description in Sharma 1996 [84]). The stepwise analysis iteratively adds or removes variables from the model. In other words, nested models are considered in which a larger model is compared to a simpler model that can be obtained by setting some of the parameters in the larger model to zero. The initial model was the null model. In each iteration Wilks' Lambda (Equation 3-4) was computed for individually adding one of the variables not currently in the model. The variable with the smallest Wilks' Lambda was added, provided the probability from the F-test was above the cutoff. In each iteration Wilks' Lambda was computed for individually removing the variables currently in the model. The variable with the largest Wilks' Lambda was added, provided the probability from the F-test was below the cutoff. The cutoff on the probability of the F-ratio was 0.05.

3.2.2 Results

The ROC curve for the global LDA models is shown in Figure 3-3 and the AUC and partial AUC values are shown in Table 3-1 (see Section 1.3.1 on ROC analysis). The results for the other global models are shown in the same figure and table and are compared in Section 3.7.

The stepwise LDA selected these variables in this order of decreasing significance: Age, Mass Margin, Calcification Morphology, Calcification Distribution, Associated Findings, and Mass Shape. The only feature not selected was Special Findings. Thus, redundant features are probably not a major problem with this data set.

3.3 Support Vector Machines

Support Vector Machines (SVM) is a supervised machine learning technique that identifies separating hyperplanes in kernel-induced feature spaces [94]. Our discussion of SVM follows that of Duda *et al.* [74] and Cristianni *et al* [94].

Instead of operating in the space of the original input features (x), a kernel ($K(x_1,x_2) = \langle \varphi(x_1) \bullet \varphi(x_2) \rangle$) is applied to map the input features to some higher dimensional space ($y = \varphi(x)$). Selection of the appropriate kernel function typically requires considerable knowledge about the problem. Without such prior knowledge, a variety of common kernels (*e.g.*, dot-product, Gaussian) can be investigated by trial and error. When the simple dot-product kernel is used, the method operates in the original input feature space (*e.g.*, $y = x$). The kernel selection dictates whether or not the SVM is a linear or non-linear classifier. It is important to recognize that using the dot-product kernel, as was done in this study, limits the SVM model to finding linear decision boundaries.

Ideally, the hyperplane identified is the one with maximal distance from the nearest training cases ("maximal margin"). A larger margin is expected to correspond to better classifier generalization. The training cases closest to the hyperplane are the most difficult to classify and are referred to as "support vectors". In training a SVM, the goal is to maximize Equation 3-5 subject to the constraints that $\sum_{i=1}^{N} z_i \alpha_i = 0$ and $\alpha_i \geq 0$. The α_i are the weights (in the dual formulation) and the z_i indicate to which class (±1, benign or malignant) each case belongs. A variety of algorithms have been applied to solving this optimization problem.

Equation 3-5

$$L(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{j,k=1}^{N} \alpha_j \alpha_k z_j z_k \langle y_j \bullet y_k \rangle$$

(Figure 3-1) [97-99]. The features describing a case are the inputs to the neurons at the front end of the network, and the classification or prediction for the case comes out of the neurons at the back end of the network. The output of each neuron in a BP-ANN was the result of an activation function ($y = 1/(1+e^{-x})$) applied to a weighted sum of the inputs to the neuron. The weights are the parameters adjusted as the network learns a given task. The ANN is feed-forward in the sense that each neuron in one layer feeds into each neuron in the next layer.

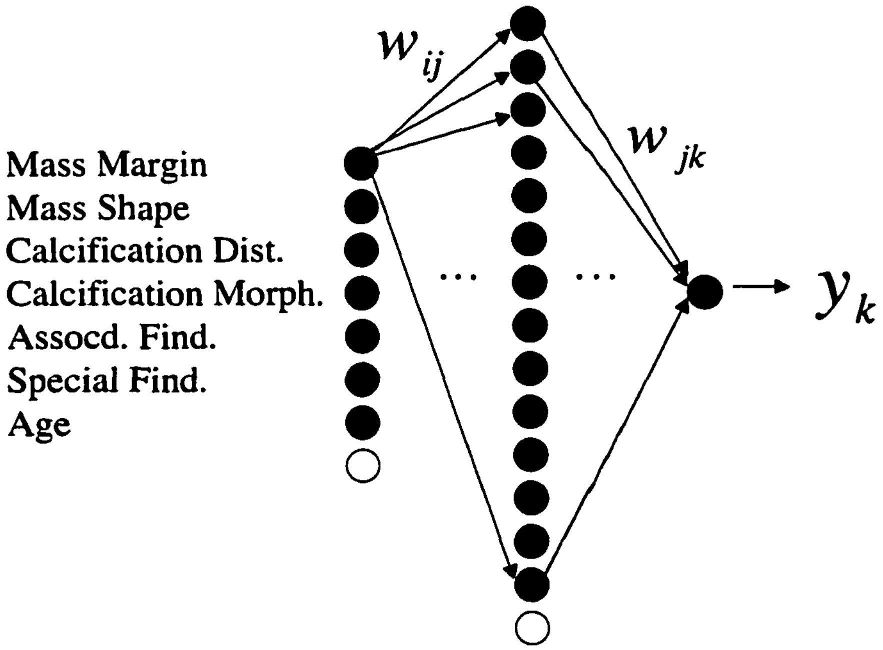


Figure 3-1. Illustration of the global BP-ANN. Only a small subset of the weights (w) are drawn; each node in the input layer is connected to each node in the hidden layer and each node in the hidden layer is connected to the output node. Bias terms are included and can be thought of as an extra neuron in each of the input and hidden layers whose input is always one.

The BP-ANN was trained to minimize the mean of the sum-of-squares error (MSE) using the back-propagation algorithm [97-99]. The MSE (Equation 3-6) is the squared difference between the network output ($y_i \in (0,1)$) and network target

($t_i \in \{0,1\}$), averaged over all of the cases (N, indexed by i). Some of the limitations of the sum-of-squares error for computer-aided diagnosis are discussed in Section 1.3.2.

Equation 3-6

$$MSE = \frac{\sum_i^N (y_i - t_i)^2}{N}$$

The back-propagation algorithm details how the error (Equation 3-6) should be propagated back through the network to adjust the weights (our description follows that of Mitchell [73]). At iteration n, the change in the weight (Δw_{ij}^n) from node i to node j depends on the change at iteration $n-1$ scaled by the momentum (α) and the product of the learning rate (η), the error term (δ_j), and the input (x_{ij}) from node i to node j (Equation 3-7). Separate learning rates (η) can be used for the different layers in the network. The error term (δ_j) depends on which layer the node j is in and is derived by taking the derivative of the error function with respect to the weights. For each node k in the output layer, δ_k was computed from the network target (t_k) and the node output (y_k) as shown in Equation 3-8. For each node j in the hidden layer, δ_j was computed from the node output (y_j), w_{jk}, and w_{jk} as shown in Equation 3-9.

Equation 3-7

$$\Delta w_{ij}^n = \eta \delta_j x_{ij} + \alpha \Delta w_{ij}^{n-1}$$

Equation 3-8

$$\delta_k = y_k(1 - y_k)(t_k - y_k)$$

Equation 3-9

$$\delta_j = y_j(1 - y_j) w_{jk} \delta_k$$

BP-ANN's are popular models in breast cancer CAD (some recent examples: [42, 46, 50, 100, 101]); in fact, LDA and BP-ANN are arguably the two most popular models in breast cancer CAD. BP-ANN's have been previously applied to portions of this BI-RADS™ database [14, 16, 21, 23, 71, 92]. Briefly, Markey *et al.* [92] used a BP-ANN to predict the biopsy outcome for 1453 cases from Duke University Medical Center with round-robin performance of $A_Z = 0.82 \pm 0.01$ and partial AUC index = 0.34 ± 0.03. Given the popularity of BP-ANN's and our laboratory's extensive experience with them, we treated the global BP-ANN model as our "gold standard" to which other models should be compared.

3.4.1 Methods

The BP-ANN had a single hidden layer and one output node indicating malignancy. Each neuron in the network used a logistic activation function ($y = 1/(1+e^{-x})$). The BP-ANN was trained to minimize the sum-of-squares error using the back-propagation algorithm [97-99]. A binary variable indicating benign or malignant was used as the network targets. The target values were clipped to 0.1 and 0.9 to ensure that the network weights remained finite (sigmoid units can't produce 0 or 1). The network weights were updated after the presentation of each case (stochastic gradient descent), which can help alleviate the problem of local minima. A momentum (α) term was used, which can also help the network escape local minima. The 2258 training cases (see Section 1.4) were presented to the network in a round-robin (leave-one-out) manner (see Section 1.3.3). Network training ended when the average testing error on the left-out cases began to increase (early-stopping) in order to avoid over-training. The network parameters were empirically optimized (learning rate (η), momentum (α), and number

3.5 Case-Based Reasoning

Case-Based Reasoning (CBR) is a machine learning technique in which past experience (cases) are used to generate solutions to the current problem [102]. In order to implement a CBR system there are two major design choices. First, how will the appropriate previous cases be identified? Second, how will the solutions to the previous cases be integrated to form a solution for the current problem?

The CBR system for breast cancer CAD based on portions of this BI-RADS™ database has been previously described [18, 93]. Briefly, Floyd *et al.* [18] used a CBR to predict biopsy outcome for 500 cases from Duke University Medical Center with round-robin performance of AUC = 0.83 and non-normalized partial AUC = 0.045.

3.5.1 Methods

The breast cancer CAD CBR system used a simple distance metric in the input feature space as the measure of similarity between the current case and the cases in the database. Based on previous experience, the Euclidean distance was used (Anya O. Bilska-Wolak, personal communication). The measure of similarity between cases i and j is shown in Equation 3-10, where k indexes the input features (x). The threshold on the distance measure was empirically optimized. By the threshold on the distance measure, we mean the cutoff on the similarity measure such that two cases are considered similar or not similar (*e.g.*, if $D \leq 0.31$, then the two cases are similar). Over the range of thresholds considered, the one that maximized the partial AUC was selected (see Section 1.3.1 on ROC analysis). The CBR inputs were rescaled to 0 to 1 (by subtracting the minimum value and dividing by the maximum minus the minimum). It should be noted that only a limited number of CBR models were considered and we do not claim that the one selected is globally optimal.

Equation 3-10

$$D = \sqrt{\sum_{k=1}^{N} (x_{ik} - x_{jk})^2}$$

The breast cancer CAD CBR system predicted the malignancy status of the current case as the fraction of the similar cases in the database that were malignant. Notice that this prediction can be viewed as an estimate of the *a posteriori* probability of malignancy, which is a monotonic function of the likelihood ratio. Also, the simple CAD CBR scheme used here can be thought of as a form of k-nearest-neighbor classification [74].

The CBR predictions of malignancy status were computed in a round-robin (leave-one-out) manner (see Section 1.3.3). The CBR analysis was performed using custom CBR software written in MATLAB® (The MathWorks Inc., Natick, MA) by members of our laboratory (Anya O. Bilska-Wolak).

3.5.2 Results

The general CBR had seven input features and used Euclidean distance as the similarity measure with a threshold of 0.31. The ROC curve is shown in Figure 3-3 and the AUC and partial AUC values are shown in Table 3-1 (see Section 1.3.1 on ROC analysis).

3.6 Classification And Regression Trees

Decision tree models classify data using a series of if-then rules depicted in a tree representation [73, 74]. There are a variety of algorithms for learning the tree from a data set (*e.g.*, ID3, C4.5, CART). The basis of these algorithms is the recursive partitioning of the data into more homogenous subsets. Classification And Regression Trees (CART) is one such algorithm that learns binary decision tree representations [103, 104]. In order to make a prediction for a test case, the if-then rules of a CART tree are followed to determine to which leaf the case maps. The model output is the fraction of the training cases at the leaf that were malignant.

CART uses the deviance, or likelihood statistic, to select the binary split on the input findings that increases the homogeneity of the resulting subsets [104]. The deviance (D) for each node in the tree was computed as the sum for all the cases at the node of the squared differences of the biopsy outcome ($y_i \in \{0,1\}$) from the mean biopsy outcome (μ) of the cases at the node (Equation 3-11). Notice that the deviance was zero if the cases at the node were homogeneous in biopsy outcome and that it increased as the heterogeneity of the node increased. All possible ways to split the cases at the node into two subsets based on the input features were considered. The split into right and left subsets that maximized the change in the deviance (Equation 3-12) was selected. The procedure was recursively repeated on the newly created right and left subsets.

Equation 3-11

$$D = \sum_i (y_i - \mu)^2$$

Equation 3-12

$$\Delta D = D - D_L - D_R$$

While decision trees have not been previously applied to a BI-RADS™ database, they have been used on other breast cancer CAD databases. Kegelmeyer *et al.* [105] used

CART for detecting masses in mammograms based on texture features. They reported a performance of 100% sensitivity and 82% specificity for 2-fold cross-validation on 85 cases. Kuo *et al.* [106] used C5.0 for classifying masses as benign or malignant based on texture features computed from ultrasound images. They reported a performance of 93% sensitivity and 97% specificity for training on 153 cases and testing on 90 cases.

3.6.1 Methods

The CART implementation in S-PLUS® (Insightful Corp., Seattle, WA) was used ("tree" function). The data were not rescaled; they were used on the original scale as described in Table 1-3. However, the BI-RADS™ features were treated as factor variables so the ordering of the values for each BI-RADS™ feature wasn't used. The data were recursively partitioned until they couldn't be separated further without producing sets of less than 25 cases. The CART model was trained in a round-robin (leave-one-out) manner (see Section 1.3.3). However, in order to display a single decision tree, a model was also built on all of the training cases.

3.6.2 Results

While the ROC performance measures were computed from the round-robin CART outputs, a CART model was also built on all of the training cases in order to have a single tree to display (Figure 3-2). Some of the decision rules have been indicated on the tree. For example, the first branch says, "if the Mass Margin is 0, 1, 2, or 3, follow the left branch", *i.e.*, all values but "spiculated", which corresponds to the highest risk of malignancy (see Table 1-3 for the encoding of the BI-RADS™ features). Notice the relationship between the labeled decision rules and those of the mass-specific local

CART model (Figure 5-3), which will be described later in Section 5.2.2, as well as to the profiles for clusters E, 6, and β (Table 2-10). There is a recurring theme of identifying lesions in younger women with relatively benign-seeming mass margins (especially well-circumscribed or obscured).

The ROC curve for the global CART model is shown in Figure 3-3 and the AUC and partial AUC values are shown in Table 3-1 (see Section 1.3.1 on ROC analysis).

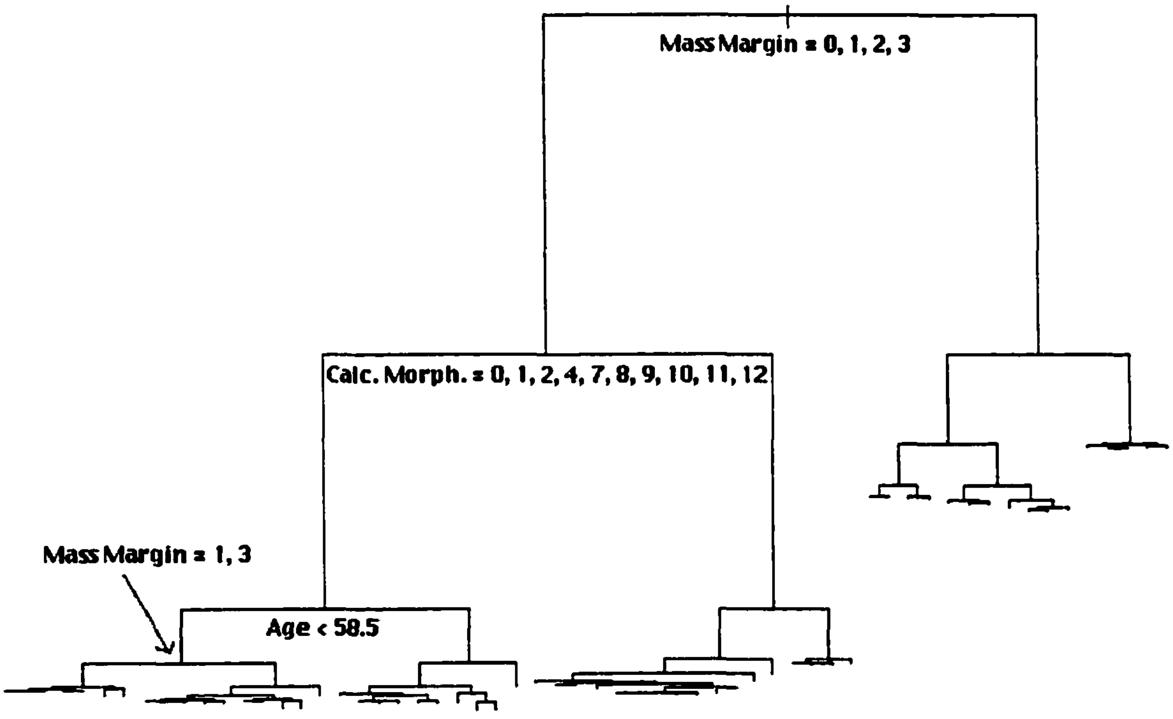


Figure 3-2. Global CART model trained on all of the training cases.

3.7 Summary

The ROC curves for the five global models are shown in Figure 3-3 (AUC) and Figure 3-4 (partial AUC). The results in terms of the AUC and partial AUC metrics are summarized in Table 3-1 and the statistical comparisons between all combinations of the five models are shown in Table 3-2 (AUC) and Table 3-3 (partial AUC). Recall that LDA and SVM are linear models while BP-ANN, CBR, and CART are capable of representing non-linear decision boundaries. Table 3-4 shows the performance of the global models for a threshold selected to give approximately 98% sensitivity. Notice that in general the non-linear models were superior to the linear models in the high sensitivity region (Figure 3-4 and Table 3-4). The five models were comparable when examined over the entire ROC curves (Figure 3-3).

LDA was comparable or superior to SVM, the other linear model considered. LDA was comparable or inferior to the non-linear models of CBR and CART and was inferior to the non-linear model BP-ANN.

SVM was comparable or inferior to LDA and CBR and was inferior to BP-ANN and CART. Over all, SVM was the worst of the global models investigated.

The BP-ANN, a non-linear model, was superior to both linear models considered (LDA, SVM). BP-ANN was superior to the non-linear model CART and was superior or comparable to the non-linear model CBR. Over all, BP-ANN was the best of the global models considered. However, it should be noted that greater effort was expended in optimizing the BP-ANN model than any of the other models.

The non-linear model CBR was superior or comparable to the linear models of LDA and SVM. The relative merit of the CBR and CART models was dependent on whether the entire ROC curve (CART better) or only the high sensitivity region (CBR better) was considered. Given the importance of maintaining high sensitivity for cancer diagnosis, the CBR was one of the better models but perhaps not as good as the BP-ANN, depending on the particular operating point considered. However, more time was spent optimizing the BP-ANN than the CBR.

The non-linear model CART was superior or comparable to the linear models of LDA and SVM. As mentioned above, the relative merit of the CBR and CART models was dependent on whether the entire ROC curve (CART better) or only the high sensitivity region (CBR better) was considered. Given the importance of maintaining high sensitivity for cancer diagnosis, CART was inferior to the other non-linear models of BP-ANN and CBR.

In Section 3, we investigated the performance of five global models across the entire training set. We observed that the non-linear global models (BP-ANN, CBR, CART) were consistently better than the linear global models (LDA, SVM). In our previous work with smaller data sets we had not been able to demonstrate the superiority of non-linear models over linear models for this task. Over all, the most promising models were found to be the BP-ANN and the CBR. In Section 4, we examine the performance of the global models (particularly BP-ANN and CBR) across the clusters in the training data that were described in Section 2.

Table 3-1. ROC performance of the round-robin results of the global models on the training data set. Non-parametric estimates of the ROC metrics are plus or minus the standard deviation estimated by bootstrap sampling on the model outputs (see Section 1.3.1).

Model	AUC	partial AUC
LDA	0.780 ± 0.010	0.261 ± 0.023
SVM	0.778 ± 0.010	0.237 ± 0.021
BP-ANN	0.820 ± 0.009	0.347 ± 0.022
CBR	0.788 ± 0.009	0.324 ± 0.019
CART	0.804 ± 0.009	0.286 ± 0.021

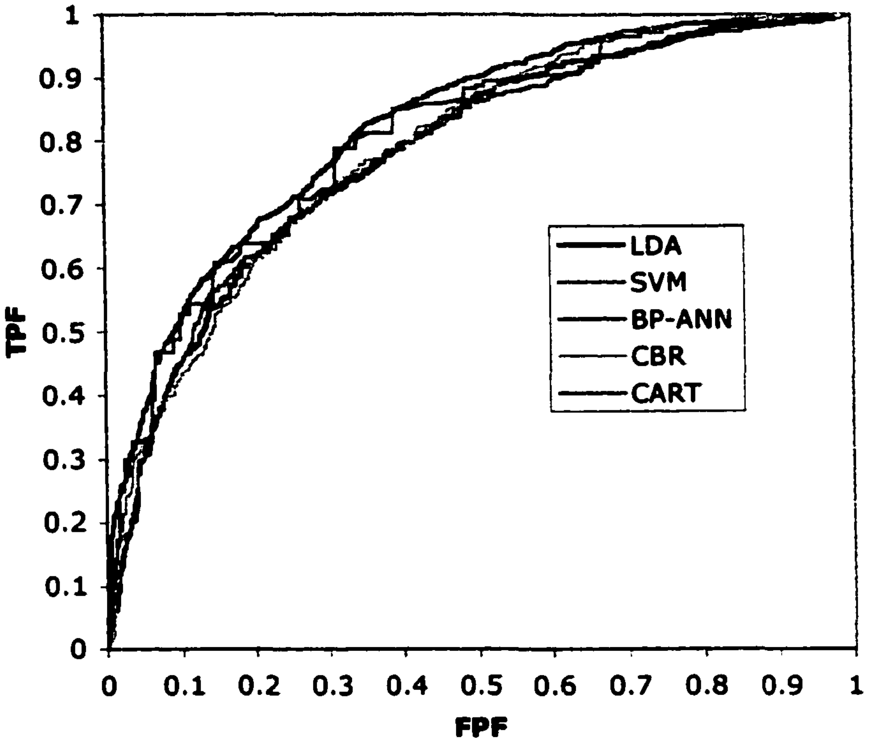


Figure 3-3. ROC curves of the round-robin trained global models on the training set.

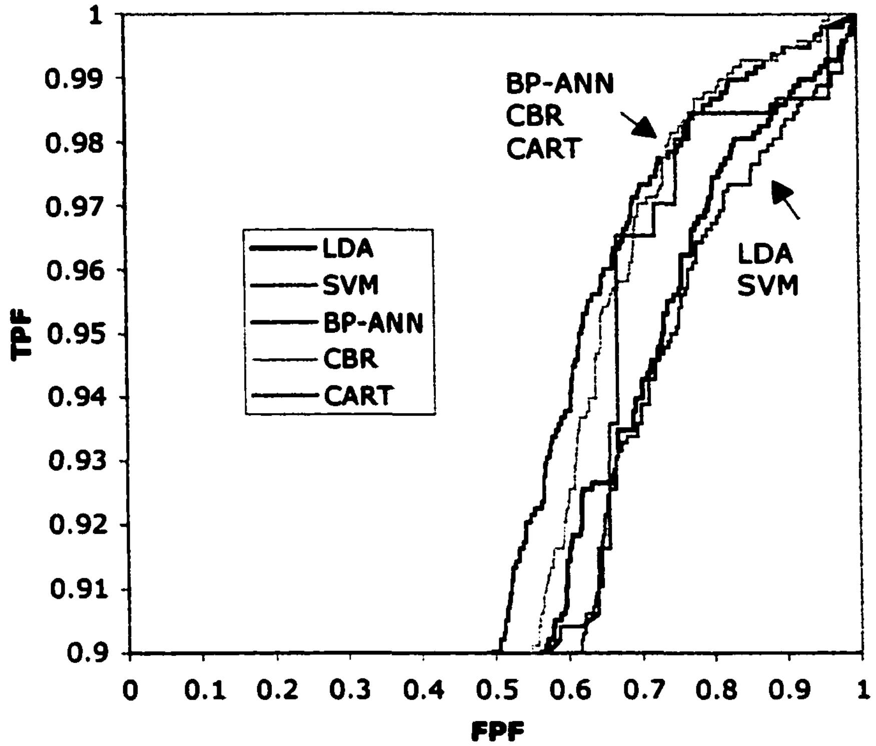


Figure 3-4. Close up of the high sensitivity region of the ROC curves for the round-robin trained global models on the training data. Notice that the non-linear models (BP-ANN, CBR, CART) were generally superior to the linear models (LDA, SVM) in this region.

Table 3-2. Comparison of the global models in terms of AUC for round-robin training on the 2258 training cases. Values shown are two-tailed p-values computed by bootstrap sampling on the decision variable (Section 1.3.1). Cells corresponding to symmetric comparisons are grayed out.

	LDA	SVM	ANN	CBR	CART
LDA					
SVM	0.39				
ANN	*< 0.01*	*< 0.01*			
CBR	0.24	0.12	*< 0.01*		
CART	*< 0.01*	*< 0.01*	*< 0.01*	*0.02*	

Table 3-3. Comparison of the global models in terms of partial AUC for round-robin training on the 2258 training cases. Values shown are two-tailed p-values computed by bootstrap sampling on the decision variable (Section 1.3.1). Cells corresponding to symmetric comparisons are grayed out.

	LDA	**SVM**	**ANN**	**CBR**	**CART**
LDA					
SVM	*< 0.01*				
ANN	*< 0.01*	*< 0.01*			
CBR	*< 0.01*	*< 0.01*	0.14		
CART	0.23	*0.02*	*< 0.01*	*0.01*	

Table 3-4. Performance of the global models for a threshold selected to give approximately 98% sensitivity. TP is the number of true positive classifications out of the 1276 actual positives. TN is the number of true negative classifications out of the 982 actual negatives. Notice that the non-linear models (BP-ANN, CBR, CART) performed with approximately twice the specificity and TN as the linear models (LDA, SVM) at this operating point.

	Threshold	**TP**	**Sensitivity**	**TN**	**Specificity**
LDA	0.1735	963	98.1%	179	14.0%
SVM	0.3685	964	98.2%	133	10.4%
ANN	0.1842	965	98.3%	303	23.8%
CBR	0.1333	964	98.2%	327	25.6%
CART	0.0698	963	98.1%	294	23.0%

4 Performance of Global Models on Clusters

4.1 Overview and Motivation

There are two motivations behind investigating the performance of a global model over the different partitions of the data into clusters that were described in Section 2. First, the performance over the clusters provides insight into the behavior of the global model. This insight could affect the ultimate clinical implementation of the model. For example, if a global model performs too poorly on a subset of cases, then one may choose to not apply the model to similar cases in a clinical setting. Similarly, identification of subsets on which the model performs poorly can drive the direction of future model development. Second, in order to assess the performance of a modular system in which separate models are used for each cluster the performance of the global model on the clusters is needed for comparison.

While five global models were considered in Section 3, we primarily focused on the performance of the global BP-ANN (Section 3.4) in this section. BP-ANN models have been used extensively on related BI-RADS™ databases and the global BP-ANN was arguably the best of the models we investigated. The global BP-ANN was significantly better in terms of the AUC than the other four global models (LDA, SVM, CBR, CART), significantly better than LDA, SVM, and CART in terms of the partial AUC, and not significantly different from CBR in terms of the partial AUC (Section 3.7).

Throughout this section, when we refer to the performance of the global model on a subset or cluster we mean the performance computed from the round-robin (Section 1.3.3) outputs. In other words, the round-robin outputs of the global model for the cases in the subset or cluster were used to compute the performance metric (*e.g.*, generate an ROC curve).

4.2 *A priori* Subsets

We investigated the performance of global models (Section 3) over the three *a priori* partitions discussed previously: institution (Section 2.2.1), lesion type (Section 2.2.2), and patient age (Section 2.2.3).

4.2.1 Institution

When the performance of the global BP-ANN trained on the cases mixed between the institutions (Section 3.4) was compared on the institution subsets (Section 2.2.1), none of the differences in AUC or partial AUC were significant (unpaired z-test, Table 4-1 and Table 4-2). Likewise, when the performance of the global CBR built on the cases mixed between the institutions (Section 3.5) was compared on the institution subsets, none of the differences in AUC or partial AUC were significant (Table 4-3 and Table 4-4). Thus, despite differences in cases collected at different institutions, CAD models trained on cases mixed between institutions may perform equally well on the different institutions in terms of the AUC and partial AUC.

However, the actual clinical implementation of a CAD model such as a BP-ANN would likely involve applying a threshold to the continuous model outputs in order to obtain a binary biopsy vs. follow up recommendation. Thus, the performance for a specific threshold intended to provide high sensitivity is more relevant than the AUC and partial AUC metrics. However, it should be recognized that the high sensitivity provided by a threshold is dependent on a small fraction of the cancers in the database (*e.g.*, the threshold for 98% sensitivity with 100 malignancies is defined by the model outputs of the 2 missed cancers). Consequently, a threshold selected to give a certain level of sensitivity on one data set may not provide the same sensitivity when applied to another data set.

We have previously shown that a BP-ANN threshold selected to give 98% sensitivity on cases from one institution (Duke) may not generalize to cases from another institution (UPenn) [71]. We show similar discrepancies in training and testing on different institutions in this dissertation (Section 5.2.1, Table 5-3). By comparison, in Table 4-5 we show that a threshold selected on the round-robin outputs from the mixed training set (Section 1.4) is appropriate for the three institution subsets in the training set.

In particular, the 2258 training cases were used to train a BP-ANN in a round-robin fashion. A threshold (0.1842) was determined that gave approximately 98% sensitivity on the round-robin outputs of all of the training cases, resulting in approximately 24% specificity. The round-robin outputs from the global model were then split out according to the institution subset to which each case belonged. The same threshold (0.1842) was applied to the three subsets of the round-robin outputs. The same threshold gave approximately 97-98% sensitivity and 23-24% specificity on each of the subsets (Table 4-5).

Table 4-1. ROC performance on the institution subsets for the global BP-ANN, round-robin trained on the training set of cases mixed between the institutions. The standard deviations were estimated by bootstrap sampling on the network outputs (see Section 1.3.1). In terms of the AUC, the performance was best on the Duke subset. In terms of the partial AUC, the performance was best on either the Duke or DDSM subsets. However, the pair-wise differences between the institution subsets were not significant (Table 4-2).

	AUC	**partial AUC**
Duke	0.821 ± 0.016	0.354 ± 0.042
UPenn	0.819 ± 0.019	0.291 ± 0.052
DDSM	0.808 ± 0.013	0.355 ± 0.029

Table 4-2. Statistical comparison of the AUC and partial AUC (unpaired z-test) for the global BP-ANN on the institution subsets (Table 4-1). The pair-wise differences between the institution subsets were not significant, indicating that the BP-ANN trained on the multi-institution set provides similar performance across the different institutions.

	AUC	**partial AUC**
Duke vs. UPenn	$p = 0.94$	$p = 0.35$
Duke vs. DDSM	$p = 0.53$	$p = 0.98$
UPenn vs. DDSM	$p = 0.63$	$p = 0.28$

Table 4-3. ROC performance on the institution subsets for the global CBR, round-robin predictions based on the training set of cases mixed between the institutions. The standard deviations were estimated by bootstrap sampling on the network outputs (see Section 1.3.1). In terms of the AUC, the performance was best on the Duke subset. In terms of the partial AUC, the performance was best on the DDSM subset. However, the pair-wise differences between the institution subsets were not significant (Table 4-4).

	AUC	**partial AUC**
Duke	0.792 ± 0.017	0.339 ± 0.033
UPenn	0.769 ± 0.021	0.271 ± 0.043
DDSM	0.785 ± 0.014	0.342 ± 0.027

Table 4-4. Statistical comparison of the AUC and partial AUC (unpaired z-test) for the global CBR on the institution subsets (Table 4-3). The pair-wise differences between the institution subsets were not significant, indicating that the CBR trained on the multi-institution set provides similar performance across the different institutions.

	AUC	**partial AUC**
Duke vs. UPenn	$p = 0.40$	$p = 0.21$
Duke vs. DDSM	$p = 0.75$	$p = 0.94$
UPenn vs. DDSM	$p = 0.53$	$p = 0.16$

Table 4-5. Generalization of threshold selected to give approximately 98% sensitivity on the global BP-ANN (Section 3.4) round-robin outputs to the institution subsets (Section 2.2.1) in the training set (Section 1.4).

Train	Test	Threshold	Sensitivity	Specificity
Mixed Train	Mixed Train	0.1842	965 / 982 = 98.3%	303 / 1276 = 23.8%
Mixed Train	Mixed Train: Duke	0.1842	256 / 260 = 98.5%	122 / 491 = 24.9%
Mixed Train	Mixed Train: UPenn	0.1842	193 / 200 = 96.5%	69 / 301 = 22.9%
Mixed Train	Mixed Train: DDSM	0.1842	516 / 522 = 98.9%	112 / 484 = 23.1%

4.2.2 Lesion

Global models (Section 3) trained on database containing a mixture of lesion types (Section 1.4) performed better on masses than calcifications. The difference in performance on masses and calcifications (Section 2.2.2) for the global LDA (Section 3.2) was significant (unpaired z-test, $p < 0.01$) in terms of both the AUC and the partial AUC (Table 4-6). Similarly, the difference in performance on masses and calcifications for the global BP-ANN (Section 3.4) was significant ($p < 0.01$, Table 4-7). The performance of the global CBR model (Section 3.5) was also significantly ($p < 0.01$) better on masses than on calcifications (Table 4-8). Likewise, the performance of the global CART model (Section 3.6) was significantly ($p < 0.01$) better on masses than on calcifications (Table 4-9). This is consistent with our previous comparisons of the mass and calcification subsets using a related dataset of cases collected at Duke [92, 93, 107].

Notice that the global BP-ANN model performed significantly better than the global CBR model on calcifications both in terms of AUC ($p < 0.01$) and partial AUC ($p < 0.01$). By comparison, the difference between the global BP-ANN and global CBR was not significant in terms of the partial AUC (Table 3-3).

Another way to study the difference in performance on masses and calcifications is to examine the effect of applying a threshold. For the global BP-ANN, a threshold of 0.1842 provided 98.3% sensitivity and 23.8% specificity on the entire set. That same threshold provided 97.3% sensitivity and 41.1% specificity on the masses but 99.8% sensitivity and 3.4% specificity on the calcifications. Clearly, the BP-ANN model was more specific for masses than for calcifications.

Table 4-6. ROC performance on the mass and calcification subsets for the global LDA, round-robin built on the training set of cases that included masses, calcifications, and other lesions. The standard deviations were estimated by bootstrap sampling on the network outputs (see Section 1.3.1).

	AUC	**partial AUC**
Mass	0.862 ± 0.011	0.468 ± 0.038
Calcification	0.666 ± 0.018	0.138 ± 0.023

Table 4-7. ROC performance on the mass and calcification subsets for the global BP-ANN, round-robin trained on the training set of cases that included masses, calcifications, and other lesions. The standard deviations were estimated by bootstrap sampling on the network outputs (see Section 1.3.1).

	AUC	**partial AUC**
Mass	0.885 ± 0.010	0.483 ± 0.036
Calcification	0.725 ± 0.017	0.183 ± 0.029

Table 4-8. ROC performance on the mass and calcification subsets for the global CBR, round-robin built on the training set of cases that included masses, calcifications, and other lesions. The standard deviations were estimated by bootstrap sampling on the network outputs (see Section 1.3.1).

	AUC	**partial AUC**
Mass	0.876 ± 0.010	0.463 ± 0.039
Calcification	0.641 ± 0.018	0.119 ± 0.021

Table 4-9. ROC performance on the mass and calcification subsets for the global CART, round-robin built on the training set of cases that included masses, calcifications, and other lesions. The standard deviations were estimated by bootstrap sampling on the network outputs (see Section 1.3.1).

	AUC	partial AUC
Mass	0.869 ± 0.011	0.415 ± 0.041
Calcification	0.719 ± 0.017	0.119 ± 0.028

4.2.3 Patient Age

The performance of the global BP-ANN (Section 3.4) was significantly (unpaired z-test, $p < 0.01$) better in terms of both AUC and partial AUC (Table 4-10) on the younger women as compared to the older women (Section 2.2.3). However, when the BP-ANN was later evaluated on the evaluation set (Section 1.4), the opposite trend was observed for AUC and no significant difference was observed for the partial AUC (Section 7.5.3). The apparent difference in performance on the age subsets was presumably due to sampling effects (Section 7.7.2).

Table 4-10. ROC performance on the subsets of younger and older women for the global BP-ANN, round-robin built on the training set of cases that included women of many ages. The standard deviations were estimated by bootstrap sampling on the network outputs (see Section 1.3.1).

	AUC	partial AUC
Age ≤ 55	0.826 ± 0.013	0.361 ± 0.042
Age > 55	0.775 ± 0.014	0.198 ± 0.028

4.3 Unsupervised Learning Methods for Cluster Analysis

We investigated the performance of global models (Section 3) on the clusters identified in the data by the three unsupervised learning methods previously described: agglomerative hierarchical clustering followed by K-Means (Section 2.4.1), Self-Organizing Map (Section 2.4.2), and AutoClass (Section 2.4.3).

mistakenly referred to follow up (11/17 = 65%) and the majority of the benign lesions that the BP-ANN would have correctly spared biopsy (218/303 = 72%) were in cluster E.

Notice that cluster E had the lowest percent of the cases that were malignant (19%) of the clusters identified by agglomerative hierarchical clustering followed by K-Means (Section 2.4.1). As described in Section 2 on cluster analysis, the SOM (Section 2.4.2) and AutoClass (Section 2.4.3) both identified a related cluster of frequently benign masses (Table 2-10).

Table 4-11. Performance of the global LDA (Section 3.2) and global BP-ANN (Section 3.4) models over the clusters identified by agglomerative hierarchical clustering followed by K-Means (Section 2.4.1).

Cluster	N	Percent Malignant	LDA AUC	LDA partial AUC	BP-ANN AUC	BP-ANN partial AUC
A	101	51%	0.6664	0.2998	0.7402	0.3179
B	489	35%	0.6331	0.1343	0.6775	0.1557
C	360	50%	0.6895	0.1012	0.7155	0.1244
D	261	24%	0.6700	0.0812	0.6809	0.2618
E	426	19%	0.8361	0.2777	0.8514	0.3196
F	398	78%	0.6892	0.1547	0.7613	0.2001
G	34	79%	0.8148	0.1429	0.8148	0.1799
H	27	48%	0.7418	0.2802	0.7473	0.3022
I	66	27%	0.6829	0.1968	0.6840	0.0556
J	96	69%	0.6773	0.1737	0.6348	0.1020
All	2258	43%	0.7802	0.2592	0.8204	0.3456

Table 4-12. For each cluster identified by agglomerative hierarchical clustering followed by K-Means (Section 2.4.1), the number of true negative classifications and the number of false negative classifications from the global BP-ANN (Section 3.4) are shown. There were 1276 actual negatives and 982 actual positives

Cluster	N	Percent Malignant	True Negatives	False Negatives
A	101	51%	2	0
B	489	35%	16	1
C	360	50%	1	0
D	261	24%	51	3
E	426	19%	218	11
F	398	78%	0	0
G	34	79%	0	0
H	27	48%	0	0
I	66	27%	14	2
J	96	69%	1	0

4.3.2 Self-Organizing Map

Recall that the Self-Organizing Map (SOM) was used to identify 16 clusters in the training data (Section 2.4.2). Table 4-13 lists how the global BP-ANN (Section 3.4) performed in terms of the AUC and partial AUC on the subsets identified by the SOM. In terms of partial AUC, the best performance was seen on cluster #4, though the performance on several clusters was similar. Cluster #4 was a group of older women with ill-defined, irregular or lobulated masses (Figure 2-12 and Figure 2-13).

Table 4-14 lists how the global BP-ANN performed in terms of the BP-ANN's recommendations for follow up instead of biopsy on the subsets identified by the SOM. A threshold was applied to the BP-ANN outputs such that the overall sensitivity was approximately 98% (965/982) with resulting specificity of approximately 24% (303/1276). In other words, 320 cases (303 actual negatives and 17 actual positives) fell below the threshold. These 320 cases that the BP-ANN would have recommended for follow up are shown in Table 4-14 according to which SOM cluster they belonged.

Notice that there is considerable variability in the performance on the clusters. Interestingly, the majority of the cancers that the BP-ANN would have incorrectly referred to follow up (11/17 = 65%) and the majority of the benign lesions that the BP-ANN would have correctly spared biopsy (242/303 = 80%) were in the cluster defined by neuron #6.

Notice that cluster #6 had the lowest percentage of the cases that were malignant (6%) of the clusters identified by the SOM. Agglomerative hierarchical clustering followed by K-Means (Section 2.4.1) and AutoClass (Section 2.4.3) both identified a related cluster (Table 2-10).

Table 4-13. Performance of the global BP-ANN (Section 3.4) over the clusters identified by the SOM (Section 2.4.2). AUC and partial AUC is not shown for clusters with less than 10 cases (#5, #12, #15).

Cluster	N	Percent Malignant	BP-ANN AUC	BP-ANN partial AUC
1	68	25%	0.6789	0.0533
2	91	14%	0.6203	0.0572
3	190	45%	0.6790	0.0566
4	212	83%	0.7395	0.2579
5	8	-	-	-
6	301	6%	0.7064	0.1422
7	89	24%	0.6576	0.1954
8	194	71%	0.7292	0.1048
9	313	52%	0.7261	0.1714
10	29	31%	-	-
11	95	69%	0.6243	0.0920
12	1	-	-	-
13	227	38%	0.7118	0.2266
14	378	39%	0.6928	0.1708
15	3	-	-	-
16	59	68%	0.8053	0.2105

Table 4-14. For each cluster identified by the SOM (Section 2.4.2), the number of true negative classifications and the number of false negative classifications from the global BP-ANN (Section 3.4) are shown. There were 1276 actual negatives and 982 actual positives.

Cluster	N	Percent Malignant	True Negatives	False Negatives
1	68	25%	15	2
2	91	14%	26	3
3	190	45%	0	0
4	212	83%	0	0
5	8	-	0	0
6	301	6%	242	11
7	89	24%	0	0
8	194	71%	0	0
9	313	52%	0	0
10	29	31%	4	0
11	95	69%	1	0
12	1	-	0	0
13	227	38%	2	0
14	378	39%	13	1
15	3	-	0	0
16	59	68%	0	0

4.3.3 AutoClass

Recall that AutoClass was used to identify 5 clusters in the training data (Section 2.4.3). The ROC performance of the global BP-ANN (Section 3.4) on the clusters identified by AutoClass (Section 2.4.3) is shown in Table 4-15. In terms of the AUC and partial AUC, the global BP-ANN performed best on clusters ε and β, though the performance on clusters γ and δ was fairly close. The cluster profiles (Section 2.3) indicated that cluster ε was a cluster of calcified masses (Table 2-9) and cluster β was a cluster of well-circumscribed masses (Table 2-9).

Table 4-16 lists how the global BP-ANN performed in terms of the BP-ANN's recommendations for follow up instead of biopsy on the subsets identified by AutoClass.

A threshold was applied to the BP-ANN outputs such that the overall sensitivity was approximately 98% (965/982) with resulting specificity of approximately 24% (303/1276). In other words, 320 cases (303 actual negatives and 17 actual positives) fell below the threshold. These 320 cases that the BP-ANN would have recommended for follow up are shown in Table 4-16 according to which AutoClass cluster they belonged. Notice that there is considerable variability in the performance on the clusters. Interestingly, the majority of the cancers that the BP-ANN would have incorrectly referred to follow up (13/17 = 77%) and the majority of the benign lesions that the BP-ANN would have correctly spared biopsy (265/303 = 88%) were in cluster β.

Notice that cluster β had the lowest PPV (21%) of the clusters identified by AutoClass. The SOM (Section 2.4.2) and agglomerative hierarchical clustering followed by K-Means (Section 2.4.1) both identified a related cluster (Table 2-10).

Table 4-15. Performance of the global BP-ANN (Section 3.4) on the clusters identified by AutoClass (Section 2.4.3) in terms of the AUC and partial AUC (Section 1.3.1).

Cluster	**N**	**Percent Malignant**	**BP-ANN AUC**	**BP-ANN partial AUC**
α	961	43%	0.7204	0.1692
β	685	21%	0.7977	0.3261
γ	395	81%	0.7708	0.2685
δ	141	43%	0.7335	0.2922
ε	76	63%	0.8519	0.3527

Table 4-16. For each cluster identified by AutoClass (Section 2.4.3), the number of true negative classifications and the number of false negative classifications from the global BP-ANN (Section 3.4) are shown. There were 1276 actual negatives and 982 actual positives

Cluster	N	Percent Malignant	True Negatives	False Negatives
α	961	43%	17	2
β	685	21%	265	13
γ	395	81%	0	0
δ	141	43%	20	2
ε	76	63%	1	0

4.4 Summary

We investigated the performance of global models over the three *a priori* partitions discussed previously: institution (Section 2.2.1), lesion type (Section 2.2.2), and patient age (Section 2.2.3). The study of the institutional effects suggests that models built on cases mixed between institutions may overcome some of the weaknesses of models built on cases from a single institution (Section 4.2.1). However, further cross-institutional studies of breast cancer CAD systems are needed. We found that CAD systems trained on a mixture of lesion types performed much better on masses than on calcifications (Section 4.2.2). We observed that the global BP-ANN performed better on the subset of younger women than on the subset of older women (Section 4.2.3). However, the age trend was reversed on the evaluation cases (Section 7.5.3) and we suspect that it was due to sampling effects (Section 7.7.2).

We investigated the performance of global models over the clusters identified by the three unsupervised learning techniques previously described: agglomerative hierarchical clustering followed by K-Means (Section 2.4.1), SOM (Section 2.4.2), and AutoClass (Section 2.4.3). Each method identified a single cluster that accounted for the majority of the cases that the BP-ANN would have recommended for follow up. The profiles of clusters identified indicated younger women with well-circumscribed or obscured, oval shaped masses (Table 2-10). Recall that the identification of likely benign cases that could be spared biopsy is the goal of such computer-aided diagnosis schemes. This suggests that cluster analysis and profiling techniques could be used to provide the physician with an alternative description of what the BP-ANN does for certain types of cases. In other words, the common feature descriptors of the related clusters identified by all of the clustering techniques may provide a way of justifying or explaining the behavior of the BP-ANN in recommending these cases for follow up. It also suggests the investigation of rule-based methods to identify relatively simple diagnostic criteria based upon those cluster profiles, such as the features that describe a very likely benign mass listed above, which might be applied to these cases to aid the radiologists in their decision making process (see Section 5.2.2).

In Section 4, we examined the performance of the global models (Section 3) developed on all the training cases over the clusters identified using *a priori* knowledge (Section 2.2) and unsupervised learning (Section 2.4). In Section 5, we will discuss the use of local models trained specifically for the different clusters in the data.

5 Modular Systems: Local Models for Predicting Biopsy Outcome using the Clusters Identified in the Training Set

5.1 Overview and Motivation

As discussed in Section 2.1, one of the motivations behind performing cluster analysis was the fact that it could serve as the first stage for a modular, "divide-and-conquer" approach. A modular system uses multiple classifiers to solve a classification problem by partitioning the input space into smaller domains, each of which is handled by a local model [41]. The idea behind such a "divide-and-conquer" approach is to break the problem down into smaller, simpler problems that will be easier to solve. Modular systems based on *a priori* subsets have shown promise in breast cancer CAD [45-48]. Thus, in this section we investigated the utility of building "local" models specifically for each of the clusters identified. The performance of each of those local models was compared to the performance of the global model. In particular, we routinely compared to the global BP-ANN (Section 3.4) since we have used BP-ANN models extensively in our laboratory [14, 16, 21, 23, 71, 92] and the overall performance of the global BP-ANN was generally better than that of the other global models (Section 3.7).

5.2 *A priori* Subsets

We investigated the performance of local models built specifically for the three *a priori* partitions discussed previously: institution (Section 2.2.1), lesion type (Section 2.2.2), and patient age (Section 2.2.3). We compared the performance of the local models for the subsets to the performance of the global model on the subsets (Section 4.2).

5.2.1 Institution

In the same manner as described for the global BP-ANN (Section 3.4), a BP-ANN was round-robin trained on the Duke subset of the training set (Section 2.2.1). In other words, the same program and criteria for parameter selection were used, but the Duke-specific BP-ANN was trained in a round-robin fashion on only the Duke cases while the global BP-ANN had been trained on a combination of Duke, UPenn, and DDSM cases. The Duke-specific BP-ANN had seven input nodes, a single hidden layer of 8 nodes, and a single output node. The first layer learning rate was 0.7, the second layer learning rate was 0.1, the momentum constant was 0.1, and the network was trained for 290 iterations. Likewise, a BP-ANN was built for the UPenn subset. The UPenn-specific BP-ANN had seven input nodes, a single hidden layer of 14 nodes, and a single output node. The first layer learning rate was 0.1, the second layer learning rate was 0.1, the momentum constant was 0.1, and the network was trained for 760 iterations. Finally, a BP-ANN was built for the DDSM subset. The DDSM-specific BP-ANN had seven input nodes, a single hidden layer of 38 nodes, and a single output node. The first layer learning rate was 0.5, the second layer learning rate was 0.1, the momentum constant was 0.1, and the network was trained for 610 iterations. The AUC and partial AUC for the institution-specific BP-ANNs on the institution subsets are shown in Table 5-1, which should be compared to the results of the global BP-ANN shown in Table 4-1. For example, the global BP-ANN on the Duke subset performed with AUC = 0.821 ± 0.016 and partial AUC = 0.354 ± 0.042. None of the differences in AUC or partial AUC were significantly different between the global and local models for Duke (p = 0.10, p = 0.83), UPenn (p = 0.22, p = 0.87), and DDSM (p = 0.44, p = 0.18). In other words, there was no advantage in terms of AUC or partial AUC in building institution-specific models rather than using the global model trained on the cases mixed between the institutions.

Table 5-1. ROC performance of the institution-specific local models trained in a round-robin fashion on the institution subsets in the training data. Standard deviations were computed by bootstrap sampling on the decision variable (Section 1.3.1).

Institution	**AUC**	**partial AUC**
Duke	0.808 ± 0.016	0.360 ± 0.038
UPenn	0.808 ± 0.020	0.286 ± 0.052
DDSM	0.803 ± 0.014	0.344 ± 0.030

On the other hand, our cross-institutional analysis demonstrated that it might be inadvisable to simply train a model on cases from one institution and apply it to cases from another institution (see also Lo, Markey, Baker, and Floyd [71]). Using the network parameters determined from round-robin training as described above, an institution-specific BP-ANN was trained on the training cases from one institution and tested on the training cases from another institution. Table 5-2 summarizes the cross-institutional performance in terms of the AUC and partial AUC. For each institution subset, we compared the performance of a model trained on one institution (*e.g.*, Duke) and tested on the current institution (*e.g.*, UPenn) to a model trained on another institution (*e.g.*, DDSM) and tested on the current institution (*e.g.*, UPenn). The differences in the AUC ($p = 0.42$) and partial AUC ($p = 0.62$) were not significant when the BP-ANN was trained on Duke vs. DDSM cases and tested on the UPenn cases. The differences in the AUC ($p = 0.02$) and partial AUC ($p < 0.01$) were significant when the BP-ANN was trained on UPenn vs. DDSM cases and tested on the Duke cases. The differences in the AUC ($p = 0.03$) and partial AUC ($p < 0.01$) were significant when the BP-ANN was trained on Duke vs. UPenn cases and tested on the DDSM cases. Thus, significant differences in the AUC and partial AUC were seen based on which institution was used to train and what institution was used to test the BP-ANN. More over, a threshold selected to give approximately 98% sensitivity on the round-robin outputs for the training institution often did not generalize when applied to the outputs on the testing institutions (Table 5-3). For example, a threshold on the local model trained on Duke cases selected to give 98% sensitivity on the Duke cases performed with only 95% sensitivity when the same model and threshold were applied to the UPenn cases. Such a drop in sensitivity would be clinically unacceptable. Recall that the results with the global BP-ANN suggested that a threshold might generalize for a BP-ANN trained on a mixture of cases from different institutions (Table 4-5).

Table 5-2. ROC performance for BP-ANN trained on cases from one institution and tested on cases from another institution (using training data, Section 1.4). Statistical comparisons were made for using the same institution data as the testing set and changing which institution data were used as the training set.

Train	Test	AUC	AUC p	partial AUC	partial AUC p
UPenn	Duke	0.7471	*p = 0.02*	0.1069	*p < 0.01*
DDSM	Duke	0.7795		0.2895	
Duke	UPenn	0.7903	p = 0.42	0.2618	p = 0.62
DDSM	UPenn	0.7976		0.2535	
Duke	DDSM	0.7950	*p = 0.03*	0.3419	*p < 0.01*
UPenn	DDSM	0.7669		0.2751	

Table 5-3. Sensitivity and specificity obtained when a threshold was applied to the BP-ANN output for a network trained on cases from one institution and tested on cases from another institution (using training data, Section 1.4). The threshold was selected to give approximately 98% sensitivity on the round-robin outputs on the training institution.

Train	Test	Threshold	Sensitivity	Specificity
Duke	UPenn	0.1769	94.5%	27.8%
Duke	DDSM	0.1769	98.5%	28.1%
UPenn	Duke	0.0875	90.8%	22.8%
UPenn	DDSM	0.0875	98.9%	9.7%
DDSM	Duke	0.2542	95.0%	32.6%
DDSM	UPenn	0.2542	94.0%	29.2%

5.2.2 Lesion

Since we observed a very consistent trend that global models performed better on masses than on calcifications (Section 4.2.2), particular attention was paid to building local models for partitions based on lesion type. Local BP-ANN, local CBR, and local CART models were built specifically for the mass and calcification lesions (Section 2.2.2).

For the mass cases, five input findings were used (Mass Margin, Mass Shape, Calcification Distribution, Calcification Morphology, and patient age) since the other two features were zero by definition (Associated Findings and Special Findings). For the calcification cases, three input findings were used (Calcification Distribution, Calcification Morphology, and patient age) since the other four were zero by definition (Mass Margin, Mass Shape, Associated Findings, and Special Findings). See Section 1.4 for a description of the available features.

In the same manner as described for the global BP-ANN (Section 3.4), local BP-ANNs were round-robin trained on the mass and calcification lesion subsets of the training set. The mass-specific BP-ANN had five input nodes, a single hidden layer of 4 nodes, and a single output node. The first layer learning rate was 0.8, the second layer learning rate was 0.1, the momentum constant was 0.1, and the network was trained for 60 iterations. The calcification-specific BP-ANN had three input nodes, a single hidden layer of 38 nodes, and a single outfput node. The first layer learning rate was 0.1, the second layer learning rate was 0.1, the momentum constant was 0.1, and the network was trained for 280 iterations. The ROC performance of the local BP-ANNs is summarized in Table 5-4 (compare to global BP-ANN in Table 4-7). The differences in performance between the global BP-ANN and local BP-ANN on the masses were not significant in terms of either the AUC (p = 0.20) or the partial AUC (p = 1.00). The differences in performance between the global BP-ANN and local BP-ANN on the calcifications were not significant in terms of either the AUC (p = 0.11) or the partial AUC (p = 0.74). Thus, no advantage in performance was seen for building a modular BP-ANN system for masses and calcifications.

Table 5-4. ROC performance of the local BP-ANNs on the mass and calcification subsets for which they were specifically trained. The standard deviations were estimated by bootstrap sampling on the network outputs (see Section 1.3.1).

	AUC	**partial AUC**
Mass	0.882 ± 0.010	0.484 ± 0.010
Calcification	0.731 ± 0.017	0.179 ± 0.031

In the same manner as described for the global CBR (Section 3.5), local CBRs were round-robin built on the mass and calcification lesion subsets of the training set. The mass-specific CBR used a threshold of 0.31 on the Euclidean distance. The calcification-specific CBR used a threshold of 0.09 on the Euclidean distance. The ROC performance of the local CBRs is summarized in Table 5-5 (compare to global CBR in Table 4-8) and the ROC curves are shown in Figure 5-1 and Figure 5-2. The difference in performance between the global CBR and local CBR on the masses was not significant in terms of the AUC (p = 0.10), but the local CBR performed significantly better than global CBR on masses in terms of the partial AUC (p < 0.01). The local CBR performed significantly better than the global CBR on the calcifications in terms of both the AUC (p = 0.01) and the partial AUC (p = 0.04). Interestingly, the correlation between the global

CBR and local CBR outputs on the calcification cases was only 0.54, which was much lower than what was seen with the BP-ANN models (0.96). Thus, there was an advantage in performance for building a modular CBR system for masses and calcifications. This would be an important finding if we were committed to using a CBR system (*e.g.*, because physicians may find CBR more intuitive than BP-ANN). However, the performance gains from the local CBR models did not bring them above the levels achieved by the global BP-ANN model. For the masses, the global ANN was borderline significantly better than the local CBR in AUC ($p = 0.05$) and there was no significant difference in terms of the partial AUC ($p = 0.64$). For calcifications, the global BP-ANN had a significantly better ($p < 0.01$) AUC than the local CBR and there was no significant difference in terms of the partial AUC ($p = 0.69$).

Table 5-5. ROC performance of the local CBRs on the mass and calcification subsets for which they were specifically trained. The standard deviations were estimated by bootstrap sampling on the network outputs (see Section 1.3.1).

	AUC	**partial AUC**
Mass	0.878 ± 0.010	0.477 ± 0.040
Calcification	0.685 ± 0.018	0.175 ± 0.026

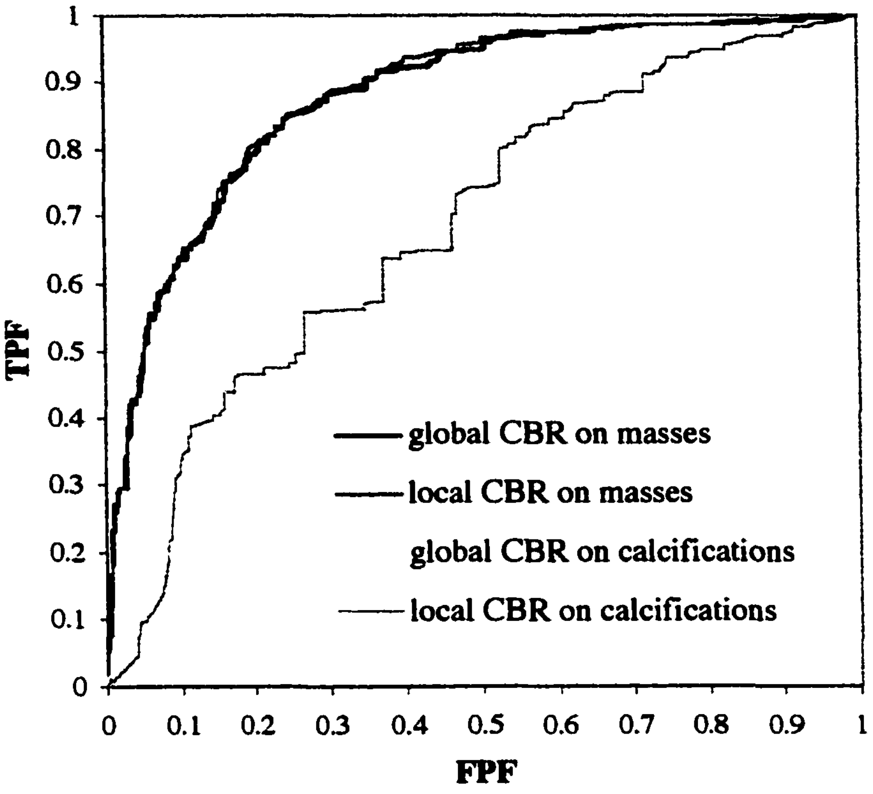


Figure 5-1. ROC curves for the global CBR and local CBRs on the mass and calcification subsets.

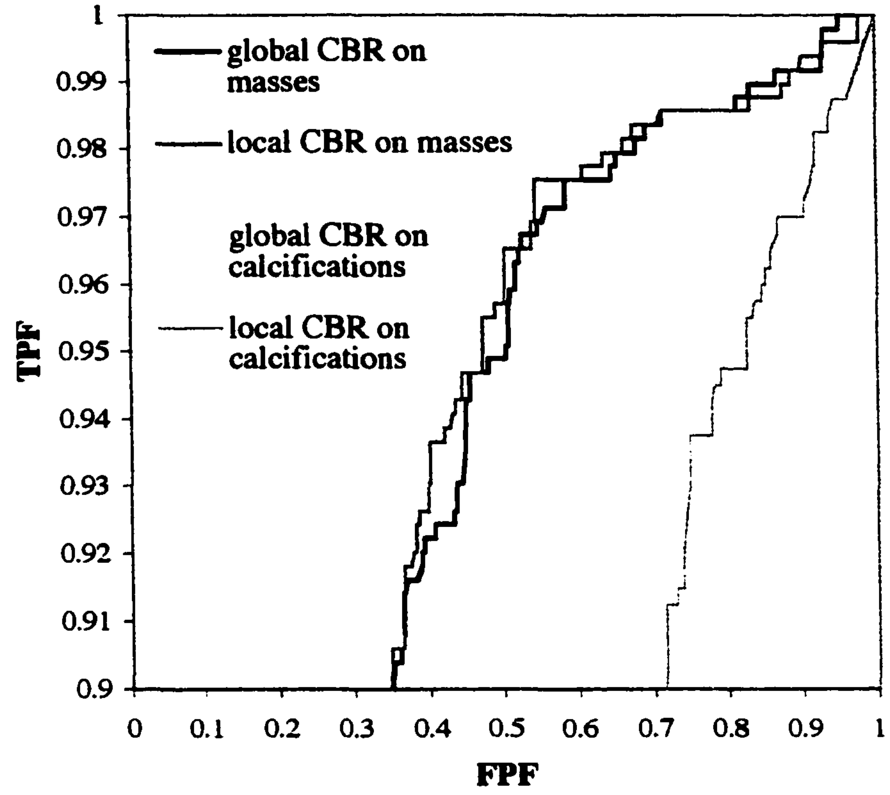


Figure 5-2. Close up of the high sensitivity regions of the ROC curves for the global CBR and local CBRs for the mass and calcification subsets.

In the same manner as described for the global CART (Section 3.6), local CARTs were round-robin trained on the mass and calcification lesion subsets of the training set. The ROC performance of the local CARTs is summarized in Table 5-6 (compare to the global CART in Table 4-9). The differences in performance between the global CART and local CART on the masses were not significant in terms of either the AUC (p = 0.75) or the partial AUC (p = 0.71). The local CART performed significantly worse than the global CART on the calcifications in terms of the AUC (p < 0.01), which was an interesting example of a local model not only not helping but actually making things worse. The difference in performance between the global CART and local CART on the calcifications was not significant in terms of the partial AUC (p = 0.15). Thus, no advantage in performance was seen for building a modular CART system for masses and calcifications.

Table 5-6. ROC performance of the local CARTs on the mass and calcification subsets for which they were specifically trained. The standard deviations were estimated by bootstrap sampling on the network outputs (see Section 1.3.1).

	AUC	**partial AUC**
Mass	0.870 ± 0.011	0.407 ± 0.047
Calcification	0.696 ± 0.018	0.087 ± 0.019

Figure 5-3 shows the mass-specific local CART model (trained on all the mass cases in the training set). Notice the relationship between the mass-specific local CART model and the global CART model (Figure 3-2) and the profiles of clusters E, 6, and β (Table 2-10). There is a consistent theme of grouping usually benign masses with well-circumscribed or obscured mass margin and patient age < 59 years.

The cluster analysis and CART models inspired us to test a very simple rule: if the Mass Margin was well-circumscribed or obscured and the age was less than 59 years and there were no calcifications, associated findings, or special findings, then don't biopsy, otherwise do biopsy. On the 2258 training cases, this rule gave 961 / 982 = 98% sensitivity and 336 / 1276 = 26% specificity. In other words, this rule performed comparably to the global BP-ANN with a threshold of 0.1842 (965 / 982 = 98% sensitivity, 303 / 1276 = 24% specificity).

There are several potential advantages of a simple rule over more complicated models. First, such a rule would be trivial to implement. Second, its simplicity makes it more understandable and thus clinicians may more readily accept it. Third, the transparency of the rule allows for more direct comparisons to clinically accepted criteria and guidelines. Comparison with current clinical criteria is an important area for future work.

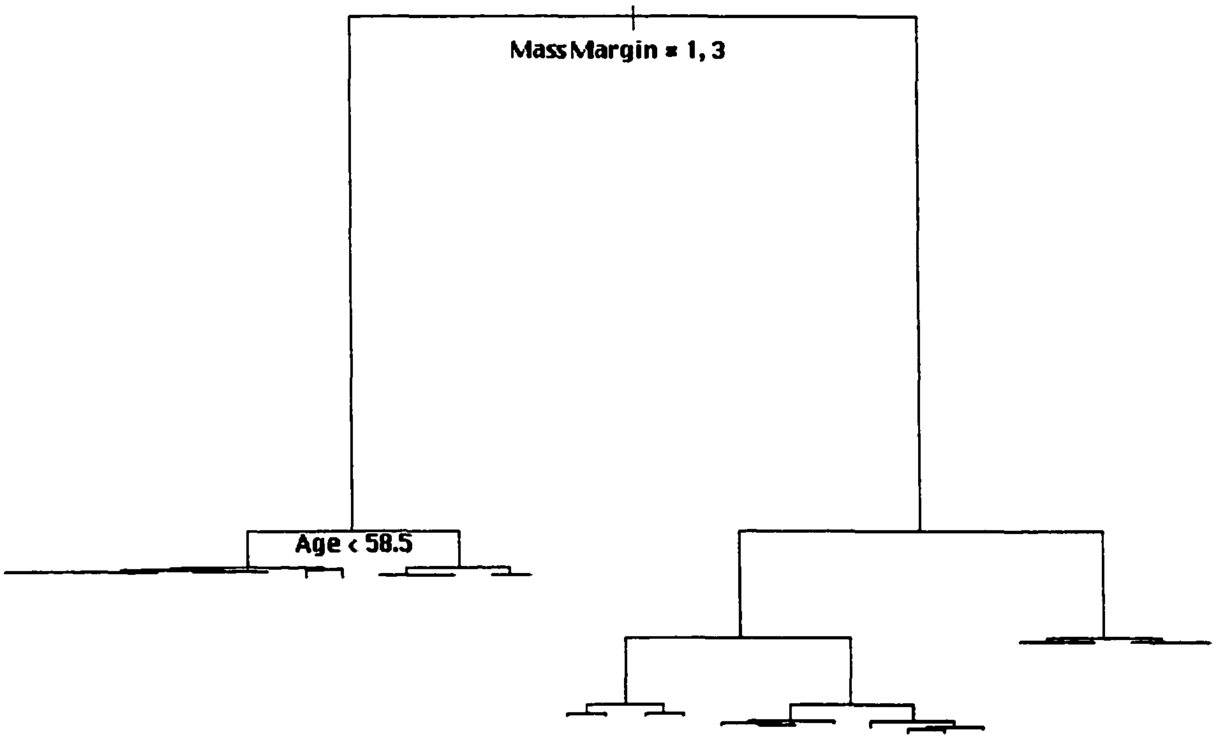


Figure 5-3. Local CART model for masses trained on all of the mass cases in the training set. Compare to the global CART model (Figure 3-2) and the profiles of clusters E, 6, and β (Table 2-10).

5.2.3 Patient Age

In the same manner as described for the global BP-ANN (Section 3.4), local BP-ANNs were round-robin trained on the subsets of younger (age ≤ 55) and older (age > 55) women in the training set. The younger-specific BP-ANN had 7 input nodes, a single hidden layer of 27 nodes, and a single output node. The first layer learning rate was 0.1, the second layer learning rate was 0.1, the momentum constant was 0.1, and the network was trained for 760 iterations. The older-specific BP-ANN had 7 input nodes, a single hidden layer of 4 nodes, and a single output node. The first layer learning rate was 0.1, the second layer learning rate was 0.1, the momentum constant was 0.1, and the network was trained for 800 iterations. The ROC performance of the local BP-ANNs is summarized in Table 5-7 (compare to global BP-ANN in Table 4-10). The differences in performance between the global BP-ANN and local BP-ANN on the younger women were not significant in terms of either the AUC (p = 0.07) or the partial AUC (p = 0.08). The difference in performance between the global BP-ANN and local BP-ANN on the older women was not significant in terms of the partial AUC (p =0.43) and the global BP-ANN was actually significantly better (p = 0.01) in terms of the AUC. Thus, no advantage in performance was seen for building a modular BP-ANN system for younger and older women. It is possible that a different choice of the age threshold (55 years) would give different results. Recall that there was concern that the observed difference between the older and younger women was an artifact of sampling (Section 4.2.3, Section 7.5.3, Section 7.7.2).

Table 5-7. ROC performance of the local BP-ANN models on the subsets of younger and older women in the training set.

	AUC	**partial AUC**
Age ≤ 55	0.818 ± 0.013	0.323 ± 0.050
Age > 55	0.761 ± 0.014	0.185 ± 0.029

5.3 Unsupervised Learning Methods for Cluster Analysis

In the previous section (Section 5.2), we evaluated local models built for the subsets defined by *a priori* partitions of the data. In a similar fashion, in this section we investigated the performance of local models built specifically for the clusters identified in the data by the three unsupervised learning methods previously described:

agglomerative hierarchical clustering followed by K-Means (Section 2.4.1), Self-Organizing Map (Section 2.4.2), and AutoClass (Section 2.4.3). We compared the performance of the cluster-specific local models to the global model on the clusters (Section 4.3).

5.3.1 Agglomerative Hierarchical Clustering and K-Means

Recall that we used agglomerative hierarchical clustering followed by K-Means (Section 2.4.1) to identify 10 clusters in the training data. In the same manner as described for the global LDA (Section 3.2), local LDAs were round-robin trained on the 7 clusters identified by agglomerative hierarchical clustering followed by K-Means (Section 2.4.1) that had approximately 100 or more cases (clusters A, B, C, D, E, F, and J). Likewise, local BP-ANNs were trained in the same manner as the global BP-ANN (Section 3.4). Table 5-8 shows the features and Table 5-9 shows the network parameters that were used in training each of the local models. For some clusters, some of the seven available features (Section 1.4) were always zero; thus, only the features that had a maximum non-zero value for the cluster were used.

The ROC performance (Section 1.3.1) of the local LDA models is shown in Table 5-10. Most of the differences between the global LDA and local LDA were not significant in terms of either the AUC or the partial AUC: cluster A ($p = 0.30$, $p = 0.72$), B ($p = 0.11$, $p = 0.12$), C ($p = 0.06$, $p < 0.01$), D ($p = 0.86$, $p = 0.36$), E ($p = 0.65$, $p = 0.26$), F ($p = 0.02$, $p = 0.48$), and J ($p = 0.01$, $p = 0.01$). Notice that the global LDA was actually significantly better than the local LDA in terms of AUC and partial AUC for clusters C and J. The local LDA is significantly better than the global LDA in terms of the AUC, but not the partial AUC, for cluster F (the cluster profiles (Table 2-8) indicate that cluster F contains ill-defined, irregular or lobulated masses in older women). Thus, there was some performance benefit to building a modular LDA system using the clusters identified by agglomerative hierarchical clustering followed by K-Means. However, as shown in Figure 5-4, the improvement of the local LDA over the global LDA for cluster F was over the sensitivity range of 0.2 to 0.8, which is not clinically useful. Moreover, the local LDA was still significantly worse than the global BP-ANN (Section 3.4) in terms of the AUC ($p < 0.01$) and no different in terms of the partial AUC ($p = 0.06$) for cluster F.

The ROC performance (Section 1.3.1) of the local BP-ANN models is shown in Table 5-10. Most of the differences between the global BP-ANN (Table 4-11) and local BP-ANN were not significant in terms of either the AUC or the partial AUC: cluster A ($p = 0.01$, $p = 0.61$), B ($p = 0.36$, $p = 0.22$), C ($p = 0.87$, $p = 0.97$), D ($p = 0.64$, $p = 0.12$), E ($p = 0.08$, $p = 0.09$), F ($p = 0.95$, $p = 0.79$), and J ($p = 0.29$, $p = 0.68$). Notice that it is possible for the local model to not only fail to improve on the global model, but to actually be significantly worse than the global model. In particular, the local BP-ANN was actually significantly worse than the global BP-ANN for cluster A in terms of the AUC. Thus, there was no performance advantage in building a modular BP-ANN system using the clusters identified by agglomerative hierarchical clustering followed by K-Means.

Table 5-8. Indication of the features used to build the local LDA and local BP-ANN models for the clusters identified by agglomerative hierarchical clustering followed by K-Means (Section 2.4.1).

Cluster	Mass Margin	Mass Shape	Calc. Dist.	Calc. Morph.	Assocd. Findings	Special Findings	Age
A			X	X			X
B	X	X	X	X	X		X
C	X	X	X	X	X		X
D	X	X	X	X	X	X	X
E	X	X			X	X	X
F	X	X	X	X			X
J	X	X	X	X	X	X	X

Table 5-9. Network parameters used in training the local BP-ANNs for the clusters identified by agglomerative hierarchical clustering followed by K-Means (Section 2.4.1).

Cluster	1st layer learning rate	2nd layer learning rate	momentum	# hidden nodes	iterations
A	0.2	0.1	0.1	2	650
B	0.1	0.4	0.2	4	230
C	0.4	0.4	0.5	4	380
D	0.4	0.4	0.1	12	1070
E	0.1	0.2	0.1	8	490
F	0.5	0.3	0.1	10	590
J	0.2	0.5	0.4	6	50

Table 5-10. ROC performance of the local LDAs and local BP-ANNs round-robin trained on the clusters identified by agglomerative hierarchical clustering followed by K-Means (Section 2.4.1).

Cluster	N	Percent Malignant	LDA AUC	LDA partial AUC	BP-ANN AUC	BP-ANN partial AUC
A	101	51%	0.7202	0.2425	0.6907	0.3014
B	489	35%	0.6782	0.1839	0.6627	0.1883
C	360	50%	0.6738	0.0718	0.7172	0.1234
D	261	24%	0.6676	0.1908	0.6978	0.3472
E	426	19%	0.8320	0.2417	0.8372	0.2460
F	398	78%	0.7344	0.1323	0.7618	0.1921
J	96	69%	0.5787	0.0424	0.5889	0.0838

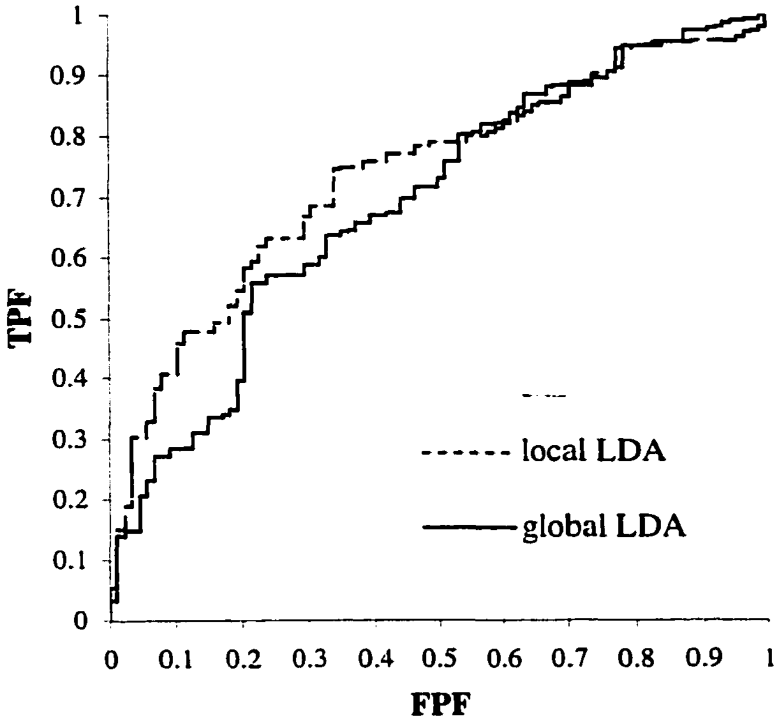


Figure 5-4. ROC curves for the global LDA (Section 3.2) and local LDA model for cluster F identified by agglomerative hierarchical clustering followed by K-Means (Section 2.4.1).

5.3.2 Self-Organizing Map

Recall that the Self-Organizing Map (SOM) was used to identify 16 clusters in the training data (Section 2.4.2). In the same manner as described for the global BP-ANN (Section 3.4), local BP-ANNs were round-robin trained on 7 clusters identified by the SOM (Section 2.4.2) that had approximately 200 or more cases (clusters #3, 4, 6, 8, 9, 13, and 14). Table 5-11 shows the features and Table 5-12 shows the network parameters that were used in training each of the local BP-ANNs. Of the seven available features (Section 1.4), the only features used were those that were non-zero for most of the cases in the cluster. Consequently, the models for clusters #3, 4, 6, and 8 used only mass findings and patient age while the models for clusters #9, 13, and 14 used only calcification findings and patient age.

The performance of the local BP-ANNS over the clusters identified by the SOM is shown in Table 5-13 (compare to the global BP-ANN in Table 4-13). The differences between the global BP-ANN and local BP-ANN were generally not significant in terms of either the AUC or the partial AUC: cluster #3 ($p = 0.15$, $p = 0.37$), #4 ($p < 0.01$, $p = 0.92$), #6 ($p < 0.01$, $p = 0.40$), #8 ($p = 0.85$, $p = 0.28$), #9 ($p = 0.64$, $p = 0.35$), #13 ($p = 0.21$, $p = 0.21$), and #14 ($p = 0.41$, $p = 0.30$). Notice that the AUC was actually significantly lower for the local BP-ANN than the global BP-ANN for clusters #4 and #6. Thus, there was no benefit to building a modular BP-ANN system for the clusters identified by the SOM in terms of the AUC or partial AUC.

Table 5-11. Indication of the features used in training the local BP-ANNs for the clusters identified by the SOM (Section 2.4.2).

Cluster	Mass Margin	Mass Shape	Calc. Dist.	Calc. Morph.	Assocd. Findings	Special Findings	Age
3	X	X					X
4	X	X					X
6	X	X					X
8	X	X					X
9			X	X			X
13			X	X			X
14			X	X			X

Table 5-12. Network parameters used in training the local BP-ANNs for the clusters identified by the SOM (Section 2.4.2).

Cluster	1st layer learning rate	2nd layer learning rate	momentum	# hidden nodes	iterations
3	0.1	0.1	0.1	15	130
4	0.4	0.1	0.1	14	680
6	0.8	0.1	0.1	15	50
8	0.3	0.1	0.1	10	710
9	0.4	0.1	0.1	3	360
13	0.6	0.1	0.1	2	410
14	0.2	0.1	0.1	32	580

Table 5-13. Performance of the local BP-ANNs round-robin trained on the clusters identified by the SOM (Section 2.4.2).

Cluster	N	Percent Malignant	AUC	partial AUC
3	190	45%	0.6449	0.1137
4	212	83%	0.6829	0.2556
6	301	6%	0.5646	0.0812
8	194	71%	0.7333	0.1651
9	313	52%	0.7220	0.1428
13	227	38%	0.6856	0.2721
14	378	39%	0.6864	0.1813

5.3.3 AutoClass

Recall that AutoClass was used to identify 5 clusters in the training data (Section 2.4.3). In the same manner as described for the global BP-ANN (Section 3.4), local BP-ANNs were round-robin trained on 4 clusters identified by AutoClass (Section 2.4.3) that had at least 100 or more cases (clusters α, β, γ, and δ). Table 5-14 shows the features and Table 5-15 shows the network parameters that were used in training each of the local BP-ANNs. For each of the seven available features (Section 1.4), a feature was not used if it was non-zero for only a few of the cases in the cluster.

The performance of the local BP-ANNS over the clusters identified by AutoClass is shown in Table 5-16 (compare to the global BP-ANN in Table 4-15). None of the differences between the global BP-ANN and local BP-ANN in the AUC or partial AUC were significant: α (p = 0.96, p = 0.45), β (p = 0.73, p = 0.10), γ (p = 0.13, p = 0.97), and δ (p = 0.86, p = 0.87). Thus, there was no benefit to building a modular BP-ANN system for the clusters identified by AutoClass in terms of the AUC or partial AUC.

Table 5-14. Indication of the features used in training the local BP-ANNs for the clusters identified by AutoClass (Section 2.4.3).

Cluster	**Mass Margin**	**Mass Shape**	**Calc. Dist.**	**Calc. Morph.**	**Assocd. Findings**	**Special Findings**	**Age**
α			X	X	X	X	X
β	X	X					X
γ	X	X			X		X
δ	X				X	X	X

Table 5-15. Network parameters used in training the local BP-ANNs for the clusters identified by AutoClass (Section 2.4.3).

Cluster	**1st layer learning rate**	**2nd layer learning rate**	**momentum**	**# hidden nodes**	**iterations**
α	0.1	0.1	0.1	17	370
β	0.1	0.1	0.1	3	90
γ	0.1	0.1	0.1	8	1240
δ	0.2	0.1	0.1	3	200

Table 5-16. Performance of the local BP-ANNs round-robin trained on the clusters identified by AutoClass (Section 2.4.3).

Cluster	**N**	**Percent Malignant**	**AUC**	**partial AUC**
α	961	43%	0.7201	0.1544
β	685	21%	0.7957	0.2872
γ	395	81%	0.7588	0.2681
δ	141	43%	0.7296	0.3107

5.4 Summary

In this section, we considered modular systems in which multiple classifiers were used to build a breast cancer CAD system by partitioning the input space into smaller domains, each of which was handled by a local model [41]. We investigated local models built specifically for the subsets defined by the three *a priori* partitions discussed previously: institution (Section 2.2.1), lesion type (Section 2.2.2), and patient age (Section 2.2.3). We also investigated the performance of local models built specifically for the clusters identified by the three unsupervised learning techniques previously described: agglomerative hierarchical clustering followed by K-Means (Section 2.4.1), SOM (Section 2.4.2), and AutoClass (Section 2.4.3). The local models used for each cluster were of the same variety as the global models described in Section 3. The performances of the local models were compared to the global models on the clusters (Section 4). Since we have used BP-ANN models extensively in our laboratory [14, 16, 21, 23, 71, 92] and the overall performance of the global BP-ANN was generally better than that of the other global models (Section 3.7), we used the global BP-ANN model as a "gold standard" to compare against.

Our study of local models for the institution subsets (Section 5.2.1) revealed several interesting trends (see also Lo, Markey, Baker, and Floyd [71]). We observed significant differences in the AUC and partial AUC index when comparing the performance on institution A of a BP-ANN trained on cases from institution B to one trained on cases from institution C (Table 5-2). Moreover, we observed that a threshold selected for a BP-ANN trained on cases from one institution did not generalize when that BP-ANN was applied to cases from another institution (Table 5-3). Thus, simply training a model on cases from one institution and applying it to cases at another institution is inadvisable. On the other hand, we observed that there was no benefit in terms of the AUC or partial AUC index in using a BP-ANN specifically trained for the cases at an institution rather than the global BP-ANN trained on cases mixed between the institutions. Also, recall that in Section 4.2.1, we showed that a threshold for the global BP-ANN that was selected using the cases mixed between institutions seemed to generalize to each institution separately. Thus, from the investigation of local models for the institution subsets, we concluded that mixing cases from multiple institutions might be helpful in overcoming the differences that exist between data sets collected at different institutions, but more work is needed on this important issue.

No benefit was seen for building a modular BP-ANN or modular CART system based on partitioning the data by lesion type (Section 5.2.2). The local CBR model was better than the global CBR model on calcifications, but it was still inferior to the global BP-ANN. The local CART model for masses (Figure 5-3) showed interesting connections to the global CART model (Figure 3-2) and to the profiles for clusters E, 6, and β (Table 2-10). The cluster profiles indicated younger women with well-circumscribed or obscured, oval shaped masses. Based on the cluster profiles and the CART models, a simple rule was devised which performed comparably on the entire training set to the global BP-ANN. Additional work is needed to study the relationship of this rule to current clinical practice and existing guidelines in the literature.

There was no advantage in building a modular BP-ANN for the data partitioned into subsets of older and younger women (Section 5.2.3). It should also be noted that the utility of partitioning based on age was called into question by other results (Section 4.2.3, Section 7.5.3, Section 7.7.2).

For partitions determined by agglomerative hierarchical clustering followed by K-Means (Section 5.3.1), SOM (Section 5.3.2), or AutoClass (Section 5.3.3), no benefit was seen for using a modular BP-ANN over the global BP-ANN. For cluster F determined by agglomerative hierarchical clustering followed by K-Means, a local LDA model was superior to the global LDA model. However, it was still inferior to the global BP-ANN on cluster F.

In conclusion, the modular systems considered here did not prove advantageous. Local models built for subsets determined by *a priori* knowledge or unsupervised learning did not result in significant improvements over the global BP-ANN. Other possible partitions or models could potentially provide significant performance gains. However, this seems unlikely and further work in this area is not expected to be fruitful.

6 Ensemble Systems: Combine Predictions of Multiple Models for Same Cases

6.1 Overview and Motivation

In Section 5, we investigated modular systems in which multiple classifiers were used by partitioning the input space into smaller domains, each of which was handled by a local model [41]. In this section, we describe ensemble systems, in which the same cases were used to train multiple models, whose predictions were then combined [41]. In other words, instead of building a separate model for some subset of cases, multiple models were built over a set of cases. That set can be the entire training set (Section 6.2) or just the cases in some cluster of interest (Section 6.3, Section 6.4). Simple ensembles of classifiers using voting or averaging to combine their predictions have shown promise in breast cancer CAD [42-44, 46].

We considered three approaches to combining classifiers and evaluating an ensemble system. (1) Compute a simple function of the continuous classifier outputs (*e.g.*, mean) and evaluate in terms of the AUC and partial AUC (Section 1.3.1). (2) Compute a simple function of the continuous classifier outputs (*e.g.*, mean) and evaluate in terms of the specificity at a fixed sensitivity (*e.g.*, 98%). (3) Apply a threshold to the continuous outputs of each classifier and then use a simple voting mechanism (*e.g.*, logical "and") to combine their binary predictions. Notice that simple combinations of the continuous model outputs involve assuming that the models produce outputs on comparable scales (*e.g.*, 0 to 1).

6.2 All Cases in Training Set

Of the global models considered, the global BP-ANN (Section 3.4) and global CBR (Section 3.5) were arguably the best since the global BP-ANN and global CBR showed the highest partial AUC values over all training cases (Table 3-1). In this section, ensembles of the global BP-ANN and global CBR models over all of the 2258 training cases were considered.

The round-robin outputs of the global BP-ANN and global CBR models were combined for each case. ROC analysis was performed on the vector resulting from applying a function (*e.g.*, min) to each pair of outputs (global BP-ANN, global CBR) for each case. While a variety of possible combination functions could be imagined, only a few were investigated here. Since all of the cases in this set were biopsied, a higher model output can be viewed as a more conservative prediction than a lower model output. Thus, we selected the "minimum" function as an example of a liberal combination and the "maximum" function as an example of a conservative combination. The "mean" function was used to give a middle-of-the-road combination that weighted the two input models equally. Table 6-1 shows the AUC and partial AUC performance results for combining the global BP-ANN and global CBR by taking the minimum, maximum, or mean of their outputs. By these measures, there were no advantages in combining these models in these ways since none of the ensemble systems outperformed the better of the two input models, the global BP-ANN. Table 6-2 shows the performance results in terms of the specificity for thresholds selected to give approximately 98% sensitivity. The mean or maximum combination methods appear to provide slightly better.specificity at 98% sensitivity, consistent with a general shift of the distribution of cases toward the higher end of the model outputs (*i.e.*, higher likelihood of cancer) with the very likely benign cases still being assigned low values (*e.g.*, the maximum of two small outputs is still small). However, the mean(global BP-ANN, global CBR) ensemble model failed to significantly (p = 0.65) improve the specificity over that of the global CBR, which had the highest specificity of the two input models.

Using the "or" function to combine binary outputs can be viewed as a conservative combination analogous to using the "maximum" function to combine the continuous outputs. Using the "or" function, a biopsy would be recommended if either input model recommended biopsy. The "and" function provides a more liberal combination scheme; a biopsy would be recommended only if both input models recommended biopsy. Table 6-3 shows the performance results in terms of sensitivity and specificity for "and" and "or" combinations of the binary votes of the classifier outputs. The binary votes were determined by applying thresholds to the global ANN (0.1842) and global CBR (0.1333) outputs that gave approximately 98% sensitivity. Keep in mind that the application of a threshold results in a single operating point and so ROC analysis is no longer possible. As expected, the "and" ensemble shows better specificity than the "or" ensemble or the input models, while the "or" ensemble shows better sensitivity than the "and" ensemble or the input models. Since the resulting ensemble models have both a different specificity and sensitivity than the input global models it is difficult to compare them. The "and" ensemble would spare 11% ((363-327)/327) more benign biopsies than the global CBR, but at the expense of delaying the diagnosis of 22% ((22-18)/18) more cancers. The tradeoffs in sensitivity and specificity are such that the "and" ensemble is unlikely to be significantly better than the global CBR model in either a statistical or clinical sense.

It is worth noting that "oracle" calculations can be used to determine an upperbound on ensembles of the global BP-ANN and global CBR. For example, an oracle for combining binary votes from BP-ANN and CBR (using the thresholds for 98% sensitivity described above) would output the correct answer if either of the models was correct. Such an oracle would perform with 98.7% sensitivity and 28.5% specificity (compare to Table 6-3).

Table 6-1. ROC performance metrics of ensembles of the global BP-ANN and global CBR models formed by taking minimum, maximum, or mean of their outputs. Notice that none of the ensemble systems outperformed the global BP-ANN (first row).

Model	AUC	partial AUC
global BP-ANN	0.8204	0.3456
global CBR	0.7875	0.3234
max(global BP-ANN, global CBR)	0.8072	0.3330
min(global BP-ANN, global CBR)	0.8109	0.3333
mean(global BP-ANN, global CBR)	0.8127	0.3366

Table 6-2. Performance in terms of the specificity at approximately 98% sensitivity of the ensembles of the global BP-ANN and global CBR models formed by taking minimum, maximum, or mean of their outputs. Notice that mean(global BP-ANN, global CBR) ensemble model failed to significantly (p = 0.65) improve the specificity over that of the global CBR (second row).

Model	Threshold	Sensitivity	Specificity
global BP-ANN	0.1842	965 / 982 = 98.3%	303 / 1276 = 23.7%
global CBR	0.1333	964 / 982 = 98.2%	327 / 1276 = 25.6%
max(global BP-ANN, global CBR)	0.2189	963 / 982 = 98.1%	341 / 1276 = 26.7%
min(global BP-ANN, global CBR)	0.1250	963 / 982 = 98.1%	315 / 1276 = 24.7%
mean(global BP-ANN, global CBR)	0.1772	964 / 982 = 98.2%	346 / 1276 = 27.1%

Table 6-3. Thresholds were applied to the global BP-ANN and global CBR models to give approximately 98% sensitivity. The resulting binary votes were combined by logical "or" and "and" functions.

Model	Sensitivity	Specificity
global BP-ANN	965 / 982 = 98.3%	303 / 1276 = 23.7%
global CBR	964 / 982 = 98.2%	327 / 1276 = 25.6%
OR(global BP-ANN, global CBR)	969 / 982 = 98.7%	267 / 1276 = 20.9%
AND(global BP-ANN, global CBR)	960 / 982 = 97.8%	363 / 1276 = 28.5%

6.3 *A priori* Subsets

In the previous Section 6.2, ensemble systems were built over the whole training set. In this section, we investigated ensemble methods for subsets defined by *a priori* partitions. While three *a priori* partitions were discussed previously (institution (Section 2.2.1), lesion type (Section 2.2.2), and patient age (Section 2.2.3)), we focused on the subsets defined by lesion type to investigate the potential of ensemble CAD systems. We compared the performance of the ensemble models to the performance of the global model on the lesion subsets (Section 4.2).

6.3.1 Lesion

As described in Section 2.2.2, the two major types of breast lesions are masses and calcifications. We investigated use of ensemble CAD systems for the subsets of mass and calcifications in the training data. In the following subsections, the round-robin continuous outputs of the global or local BP-ANN and CBR models were combined.

6.3.1.1 Mass

Ensembles of the global BP-ANN (Section 3.4) and global CBR (Section 3.5) models over the subset of mass lesions (Section 2.2.2) in the training data (Section 1.4) were investigated. Table 6-4 shows the AUC and partial AUC performance results for combining the global BP-ANN and global CBR by taking the minimum, maximum, or mean of their outputs on the mass cases. By these measures, there were no advantages in combining these models in these ways, since none of the ensembles outperformed the best input model, the global BP-ANN.

Table 6-4. ROC performance metrics of ensembles of the global BP-ANN and global CBR models formed by taking minimum, maximum, or mean of their outputs on the mass cases. Notice that none of the ensemble systems outperformed the global BP-ANN (first row).

Model	AUC	partial AUC
global BP-ANN	0.8848	0.4818
global CBR	0.8758	0.4617
max(global BP-ANN, global CBR)	0.8821	0.4811
min(global BP-ANN, global CBR)	0.8807	0.4617
mean(global BP-ANN, global CBR)	0.8828	0.4786

6.3.1.2 Calcifications

Ensembles of the global BP-ANN (Section 3.4) and global CBR (Section 3.5) models over the subset of calcification lesions (Section 2.2.2) in the training data (Section 1.4) were investigated. Table 6-5 shows the AUC and partial AUC performance results for combining the global BP-ANN and global CBR by taking the minimum, maximum, or mean of their outputs on the calcification cases. By these measures, there were no advantages in combining these models in these ways, since none of the ensembles outperformed the best input model, the global BP-ANN.

Recall that in Section 5 "local" models specific for different subsets or clusters of the data were built. Since the local CBR model was significantly better than the global CBR model on the calcification cases (Section 5.2.2), ensembles of the local CBR and local BP-ANN were also considered. Table 6-6 shows the AUC and partial AUC performance results for combining the local, calcification-specific BP-ANN and the local, calcification-specific CBR by taking the minimum, maximum, or mean of their outputs on the calcification cases. While the mean(local BP-ANN, local CBR) ensemble looks like an improvement in the partial AUC over the local BP-ANN and local CBR models, it was not significantly (p = 0.93) different from the global BP-ANN model on calcification lesions. Thus, there were no advantages in combining these models in these ways, since none of the ensembles outperformed the best input model, the local BP-ANN.

Table 6-5. ROC performance metrics of ensembles of the global BP-ANN and global CBR models formed by taking minimum, maximum, or mean of their outputs on the calcification cases. Notice that none of the ensemble systems outperformed the global BP-ANN (first row).

Model	**AUC**	**partial AUC**
global BP-ANN	0.7251	0.1811
global CBR	0.6413	0.1173
max(global BP-ANN, global CBR)	0.7010	0.1483
min(global BP-ANN, global CBR)	0.7102	0.1710
mean(global BP-ANN, global CBR)	0.7114	0.1688

Table 6-6. ROC performance metrics of ensembles of the local BP-ANN and local CBR models formed by taking minimum, maximum, or mean of their outputs on the calcification cases. The mean(local BP-ANN, local CBR) ensemble was not significantly (p = 0.93) different in partial AUC from the global BP-ANN model on calcification lesions.

Model	**AUC**	**partial AUC**
local BP-ANN	0.7305	0.1765
local CBR	0.6840	0.1747
max(local BP-ANN, local CBR)	0.7053	0.1729
min(local BP-ANN, local CBR)	0.7165	0.1752
mean(local BP-ANN, local CBR)	0.7096	0.1821

6.4 Unsupervised Learning Methods for Cluster Analysis

In the previous Section 6.3, we built ensemble models for subsets of the data defined by *a priori* knowledge. In this section we discuss ensemble models for subsets of the data identified by unsupervised learning. While we partitioned the data using three unsupervised learning methods (agglomerative hierarchical clustering followed by K-

Means (Section 2.4.1), Self-Organizing Map (Section 2.4.2), and AutoClass (Section 2.4.3), we focused on a single cluster identified by the SOM in order to investigate ensemble CAD systems. We compared the performance of the ensemble models to the global model on the cluster (Section 4.3).

Recall that the Self-Organizing Map (SOM) was used to identify 16 clusters in the training data (Section 2.4.2). Cluster #13 identified by the SOM (Section 2.4.2) was selected as the example cluster to on which to test ensemble approaches. Cluster #13 was chosen because the local BP-ANN was better than the global BP-ANN in terms of the partial AUC (Section 5.3.2), thought not significantly so (p = 0.21). Recall that there were 227 cases in the cluster and that the profiles (Section 2.3) of cluster #13 indicated that a typical case was a woman in her 50's with clustered, pleomorphic calcifications (Figure 2-12 and Figure 2-13). Since in general our models perform poorly on calcification lesions (Section 4.2.2 and Section 5.2.2), we were particularly interested in the possibility of getting any improvement on a cluster of calcification cases.

In addition to the global BP-ANN (Section 3.4) and local BP-ANN (Section 5.3.2) models, local LDA and local SVM models were built in the same manner described for the global LDA (Section 3.2) and global SVM (Section 3.3) models. Since the SVM outputs weren't on the same scale as the BP-ANN and LDA outputs, the SVM outputs were linearly rescaled to between zero and one. Table 6-7 shows the AUC and partial AUC performance results for the base models. For simplicity, we focused on combining the two most promising local models, the local BP-ANN and local SVM. We studied the effects of combining them by taking the minimum, maximum, or mean of their outputs. However, the most promising ensemble, min(local BP-ANN, local SVM), was not significantly better the global BP-ANN in terms of the AUC (p = 0.25) or the partial AUC (p = 0.19).

Table 6-7. ROC performance metrics of ensembles of the local BP-ANN and local SVM models formed by taking minimum, maximum, or mean of their outputs on the cases in cluster #13 identified by the SOM.

Model	AUC	partial AUC
global BP-ANN	0.7118	0.2266
local BP-ANN	0.6856	0.2721
local LDA	0.7000	0.2182
local SVM	0.7377	0.1707
min(local BP-ANN, local SVM)	0.7417	0.2726
max(local BP-ANN, local SVM)	0.6682	0.1633
mean(local BP-ANN, local SVM)	0.6747	0.2532

Given the disappointing performance of the simple combination functions (Table 6-7), we hypothesized that a more sophisticated combination function might be required. Thus, we also investigated using a perceptron and a BP-ANN with a hidden layer to combine the local BP-ANN and local SVM outputs. Notice that this approach entailed doing a "round-robin of the round-robin" (Section 1.3.3). That is, a BP-ANN or perceptron model was trained in a round-robin fashion to combine the round-robin outputs of other models. The purpose of employing round-robin sampling was to avoid training and testing on the same cases since that makes it difficult to gauge the potential for generalization of the model. Four ensemble models were considered. (1) A perceptron to combine the local BP-ANN and local SVM round-robin outputs. (2) A perceptron to combine the local BP-ANN and local SVM round-robin outputs and the three input features (Calcification Morphology, Calcification Distribution, and Age). (3) A BP-ANN with a hidden layer to combine the local BP-ANN and local SVM round-robin outputs. (4) A BP-ANN with a hidden layer to combine the local BP-ANN and local SVM round-robin outputs and the three input features (Calcification Morphology, Calcification Distribution, and Age). The input features were included in some ensembles because it was hypothesized that patterns in the input features could be valuable in determining how to combine the model outputs. Table 6-8 summarizes the network parameters used for these four ensemble models.

Notice that ensembles (2) and (4) required unusually high numbers of iterations to train, by an order of magnitude. A minimum in the MSE on the held-out cases ("testing MSE") was not observed for the perceptron for ensemble (2); the training was arbitrarily cut off. When the learning rate was increased, a minimum was observed, but that perceptron did not achieve as low of testing MSE as the one selected. If the training was arbitrarily cutoff at 100 iterations (a value more consistent with the other ensembles), the performance was more similar to that of ensemble (1). A minimum in the testing MSE was observed for ensemble (4), but many more iterations than usual were required to reach that point. If the training was arbitrarily cutoff at 100 iterations, the performance was more similar to that of ensemble (3).

Table 6-8. The network parameters for the four neural network-based ensemble models. (1) perceptron(local BP-ANN, local SVM), (2) perceptron(local BP-ANN, local SVM, input features), (3) BP-ANN(local BP-ANN, local SVM), (4) BP-ANN(local BP-ANN, local SVM, input features).

Ensemble	**1st layer learning rate**	**2nd layer learning rate**	**momentum**	**# hidden nodes**	**iterations**
(1) perceptron	0.1	0.1	0.1	-	50
(2) perceptron	0.1	0.1	0.1	-	3080
(3) BP-ANN	0.1	0.1	0.1	6	140
(4) BP-ANN	0.9	0.1	0.1	8	3440

Table 6-9 shows the AUC and partial AUC performance results for the four neural network-based ensemble methods. The round-robin performance results for forming an ensemble by training a BP-ANN on the feature inputs and the round-robin outputs of the local BP-ANN and local SVM were extremely good. In particular, ensemble (4) performed with AUC = 0.9192 and partial AUC = 0.4579 as compared to the local BP-ANN model performance of AUC = 0.6856 and partial AUC = 0.2721. These performances on a cluster of calcification cases were unprecedented. Since we suspected that these results were optimistic due to the "round-robin of the round-robin" sampling performed, we tested the generalization of ensemble (4) to the evaluation set (Section 1.4) that had been set aside to allow for independent testing to resolve exactly this sort of generalization issue. Our concern was the "round-robin of the round-robin" resulted in a lack of independence between the training and testing cases. In order to test the generalization, a single model was required and round-robin sampling actually produces N separate models. Thus, train-on-all local BP-ANN and local SVM models using the parameters determined from the round-robin training were built on the training data. In other words, a single BP-ANN (and likewise a single SVM model) was built by training without round-robin sampling, but using the same model parameters selected from round-robin training (*e.g.*, the number of hidden nodes). Those train-on-all outputs and the feature inputs were used to build a train-on-all BP-ANN using the parameters determined from the round-robin training. In other words, a single BP-ANN ensemble model was built by training on all the training case inputs and outputs, but using the network parameters determined from round-robin training. The cases in the evaluation set (Section 1.4) that mapped to cluster #13 were then passed through the ensemble (4) system. The resulting performance was AUC = 0.7476 and partial AUC = 0.1891, which was much lower than the optimistic results from the 'round-robin of the round-robin' on the training cases (Table 6-9). Moreover, the performance of the ensemble (4) system on the evaluation set was not significantly different from that of the global BP-ANN on the evaluation set in terms of either the AUC (0.7153, p = 0.26) or the partial AUC (0.2687, p = 0.34). Thus, round-robin training of ensemble models based on both feature inputs and round-robin local model outputs seems inadvisable. While the number of weights in the combination BP-ANN of the ensemble (4) system was not excessive (57 weights) and the number of inputs was small (3 input features + 2 model outputs), the small number of cases (227) may also have contributed to the overtraining.

Multi-stage CAD systems are common in the literature (*e.g.*, [42, 108, 109]). For example, a CAD system might extract features, select among the extracted features, and merge the selected features with a machine learning technique. Each stage often involves parameter optimization, which requires some form of sampling, such as round-robin sampling. Particular care must be taken with regard to sampling and evaluation in multi-stage CAD systems. The optimistic results that were presented in this section should be taken as a general cautionary tale and not as strange behavior exhibited only by this particular combination of methods for this particular task.

Table 6-9. The ROC performance of the four neural network-based ensemble models. (1) perceptron(local BP-ANN, local SVM), (2) perceptron(local BP-ANN, local SVM, input features), (3) BP-ANN(local BP-ANN, local SVM), (4) BP-ANN(local BP-ANN, local SVM, input features).

Ensemble	AUC	partial AUC
global BP-ANN	0.7118	0.2266
(1) perceptron	0.6617	0.2345
(2) perceptron	0.7134	0.1606
(3) BP-ANN	0.6604	0.2362
(4) BP-ANN	0.9192	0.4579

6.5 Summary

In this section, we considered ensemble systems, in which the same cases were used to train multiple models, whose predictions were then combined in a two-stage classifier [41]. The models used were the same types as the models described in Sections 3 and 5.

For the global BP-ANN (Section 3.4) and global CBR (Section 3.5), simple combinations (min, max, mean) of the continuous round-robin model outputs were considered and evaluated in terms of the AUC, partial AUC, and the specificity at 98% sensitivity. Thresholds were also applied to the continuous model outputs to give binary predictions which were combined by logical "and" and "or" functions. No improvement in performance was seen over all the training cases or over the mass and calcification subsets (Sections 6.2 and 6.3).

Combining the outputs of a local BP-ANN and local CBR on calcifications by taking the mean showed some promise, but was not significantly better than using the global BP-ANN (Section 6.3.1.2). Likewise, combining the outputs of the local BP-ANN and local SVM on SOM cluster #13 by taking the minimum showed some promise, but was not significantly better than using the global BP-ANN (Section 6.4).

Our efforts to combine the outputs of the local BP-ANN and local SVM on SOM cluster #13 using a BP-ANN model tell an important cautionary tale (Section 6.4). Round-robin training of a BP-ANN to combine the input features and the round-robin outputs of the local BP-ANN and local SVM appeared to be extremely successful, but failed to generalize when tested on the evaluation set. This training approach appears to be fundamentally flawed and should be avoided.

In conclusion, the ensemble methods considered here did not prove advantageous. The simplistic combination schemes did not result in significant improvements and more complicated combination schemes were found to be unduly optimistic. However, it should be noted that there isn't any way to know *a priori* what models should be combined or how they should be combined. Thus, additional work in this area could be beneficial and may be warranted. For example, resampling techniques such as boosting could be investigated [110-112].

7 Evaluation

As discussed in Section 1.3, evaluation of the breast cancer CAD systems was critical. A major concern is the ability of CAD systems to generalize. That is, to perform on a similar but previously unseen data set in approximately the same way as it performed on the data set used to construct it. In order to address this issue, the data were randomly partitioned into two halves (Sections 1.3.3 and 1.4). In Sections 2, 3, 4, 5, and 6 we described the results of unsupervised and supervised learning techniques applied to the training half of the data set. The cluster analysis (Section 2) was performed on the entire training half while the predictive models (Sections 3, 4, 5, and 6) were trained using round-robin sampling. In this section, we describe the generalization tests performed with the evaluation half of the data set. By using round-robin sampling on the training half and performing a final verification using the evaluation set, we are assessing the system using three portions of the data in a training-testing-validation strategy.

Testing for generalization on a held-out evaluation set is still important even if round-robin sampling was used in training. The results from round-robin sampling may be biased because, for example, the results from round-robin training were considered in choosing the model parameters (*e.g.*, the number of hidden nodes in a BP-ANN). Although round-robin sampling is very efficient for using all cases for training and testing without direct overlap, each case contributes indirectly to the choice of model parameters, and thus there is not a truly independent test set. The problems that we encountered with the "round-robin of the round-robin" in Section 6.4 illustrate the importance of testing for generalization on a held-out evaluation set.

A superior form of the training-testing-validation strategy would entail resampling the cases into the training, testing, and validation portions. That is, we would need to repeat the random partitioning of data into "training" and "evaluation" sets and optimize the models each time using round-robin sampling on the "training" set and then validate the round-robin results by testing on the held-out "evaluation" set. While this approach would be ideal, we did not pursue it due to the computational time that would be required.

Notice that in the previous section we were looking for an improvement over our "gold standard" model and so we hoped to see significant differences in the comparisons we performed. On the other hand, in this section we are testing for generalization and so we hope to see failures to demonstrate a significant difference, which would indicate that the results from the training set would apply to a similar but new data set.

7.1 Evaluation Data Set

As described in Section 1.4, the data were randomly partitioned into training and evaluation sets. Approximately the same percentage of the cases was malignant in the evaluation set (42%) as was in the training set (43%, Table 1-2). The same distribution of cases from the three institutions was observed in the training and evaluation sets: Duke (33%), UPenn (22%), and DDSM (45%, Table 1-1). The same distribution of lesion types was seen with the evaluation set as was observed with the training set (rightmost column of Table 2-3): masses (1079 / 2177 = 50%), calcifications (896 / 2177 = 41%), and other lesions (202 / 2177 = 9%). See Section 2.2.2 for definitions of the lesion types. While the percent of women in the evaluation set who were older than 55 years (1053 /

2177 = 48%) was lower than that observed in the training set (50%, Table 2-6), the difference was not significant (p = 0.33, Chi-square test for independence).

7.2 Map Evaluation Cases to SOM Clusters in Training Set

After performing cluster analysis (Section 2), one may wish to identify to which of the clusters a new case would belong. One simple method was used to map the evaluation cases to the SOM clusters (Section 2.4.2). The evaluation data were normalized in the same manner as the training data. The centroids (means) of the clusters in the training data were computed. Each evaluation case was mapped to the cluster for which the Euclidean distance from the case to the cluster centroid was smallest. The distribution of evaluation cases across the SOM clusters was similar to what was seen with the training cases (Table 7-1). The percent of the cases that were malignant in each cluster was similar for the training and evaluation sets. This provided some reassurance that the partition into training and evaluation sets was fair and that the clustering technique was robust, given that the clustering results were not considered in the original random split of the data into training and evaluation sets.

Table 7-1. Distribution of the training and evaluation cases across the clusters identified by the SOM (Section 2.4.2).

	Training		Evaluation	
Cluster	N	Percent Malignant	N	Percent Malignant
1	68	26%	69	30%
2	91	14%	130	8%
3	190	45%	129	43%
4	212	83%	201	86%
5	8	-	9	-
6	301	6%	237	8%
7	89	24%	127	22%
8	194	71%	207	65%
9	313	52%	238	45%
10	29	31%	69	28%
11	95	69%	99	83%
12	1	-	0	-
13	227	38%	276	38%
14	378	39%	324	32%
15	3	-	5	-
16	59	68%	57	70%

7.3 Generalization of Global Models to Evaluation Set

In Section 3, five global models were considered for the training set. The generalization of the two most promising (Section 3.7) models, BP-ANN (Section 3.4) and CBR (Section 3.5), was tested using the evaluation set (Sections 1.4 and 7.1).

A global BP-ANN (Section 3.4) was trained on all the training cases, using the network parameters determined from the round-robin training, and applied to the evaluation set. The performance on the training set refers to the analysis of the round-robin model outputs. There was no significant difference (unpaired z-test) in the performance of the global BP-ANN on the training and evaluation sets (Table 7-2) in terms of the AUC ($p = 0.35$) or the partial AUC ($p = 0.29$). The generalization of a threshold selected to give 98% sensitivity on the training set is shown in Table 7-3. The

Table 7-2. Generalization of the global BP-ANN and global CBR models to the evaluation set in terms of the AUC and partial AUC. Standard deviations were estimated by bootstrap sampling (Section 1.3.1). While both models showed good generalization in terms of the AUC, the difference in the partial AUC between the training and evaluation sets was borderline significant (p = 0.05) for CBR.

	Training		**Evaluation**	
Model	**AUC**	**partial AUC**	**AUC**	**partial AUC**
BP-ANN	0.820 ± 0.009	0.347 ± 0.022	0.832 ± 0.009	0.312 ± 0.025
CBR	0.788 ± 0.009	0.324 ± 0.019	0.798 ± 0.010	0.263 ± 0.025

Table 7-3. Generalization on the evaluation set of threshold selected to give 98% sensitivity on the training set for the global BP-ANN and global CBR. The thresholds selected on the training set generalized slightly better for the BP-ANN than the CBR.

		Training		**Evaluation**	
Model	**Threshold**	**Sensitivity**	**Specificity**	**Sensitivity**	**Specificity**
BP-ANN	0.1842	965 / 982 = 98.3%	303 / 1276 = 23.4%	884 / 904 = 97.8%	296 / 1273 = 23.3%
CBR	0.1333	964 / 982 = 98.2%	327 / 1276 = 25.6%	873 / 904 = 96.6%	316 / 1273 = 24.8%

7.4 Generalization of Rule-Based Method to the Evaluation Set

In Section 5.2.2, a simple classification rule was proposed based on the cluster profiles and the CART models: if the Mass Margin was well-circumscribed or obscured and the age was less than 59 years and there were no calcifications, associated findings, or special findings, then don't biopsy, otherwise do biopsy. On the training set, this rule gave 961 / 982 = 98% sensitivity and 336 / 1276 = 26% specificity. On the evaluation set, this rule gave 886 / 904 = 98% sensitivity and 339 / 1273 = 27% specificity. Thus, the rule generalized well to the evaluation set and performed comparably to the global BP-ANN (Table 7-3).

7.5 Global BP-ANN Performance on the *A Priori* Subsets in Evaluation Set

In Section 4.2, we investigated the performance of the round-robin trained, global BP-ANN on the *a priori* subsets (Section 2.2) in the training data. In this section, we examined the performance of the global BP-ANN trained on the training data and tested on the evaluation data over the *a priori* subsets (institution, lesion type, patient age). We checked to see if the same trends were observed on the evaluation cases as had been seen on the training cases (*e.g.*, better performance on masses than calcifications). The global BP-ANN performance on the training set (round-robin outputs) and evaluation set was compared in terms of the AUC and partial AUC.

7.5.1 Institution

The difference in the global BP-ANN performance on the institution subsets in the training (Table 4-1) and evaluation (Table 7-4) sets was not significant for Duke (AUC $p = 0.36$, partial AUC $p = 0.82$) or UPenn (AUC $p = 0.36$, partial AUC $p = 0.51$). The difference in AUC was not significant ($p = 0.64$) for DDSM. However, the partial AUC for the DDSM subset was significantly ($p = 0.04$) lower on the evaluation set (0.261 ± 0.036) than on the training set (0.355 ± 0.029). Thus, there are some potentially important differences between the randomly sampled training and evaluation sets even for this very large database.

We compared the performance of the global BP-ANN across the institution subsets in the evaluation set. As with the training set, the global BP-ANN performance is relatively constant across the institution subsets in the evaluation set (Table 7-4). However, there is one exception. The difference in the partial AUC for the Duke and

DDSM subsets was borderline significant (p = 0.049, Table 7-5). In other words, for the global BP-ANN trained on a the mixed institution training set, the difference in performance on the Duke and DDSM subsets in the evaluation was borderline significant in the high sensitivity region. Thus, the issue of cross-institutional analysis continues to warrant further study.

Table 7-4. ROC performance on the institution subsets for the global BP-ANN, trained on the training set and tested on the evaluation set. The standard deviations were estimated by bootstrap sampling on the network outputs (Section 1.3.1).

	AUC	**partial AUC**
Duke	0.841 ± 0.015	0.367 ± 0.040
UPenn	0.843 ± 0.018	0.339 ± 0.052
DDSM	0.817 ± 0.014	0.261 ± 0.036

Table 7-5. Statistical comparisons of the performance of the global BP-ANN on the institution subsets in the evaluation set.

	AUC	**partial AUC**
Duke vs. UPenn	p = 0.93	p = 0.67
Duke vs. DDSM	p = 0.24	*p = 0.05*
UPenn vs. DDSM	p = 0.25	p = 0.22

7.5.2 Lesion

The differences in the global BP-ANN performance on the masses in the training and evaluation sets were not significant (AUC p = 0.57, partial AUC p = 0.56). Likewise, the differences in the global BP-ANN performance on the calcifications in the training and evaluation sets were not significant (AUC p = 0.57, partial AUC p = 0.36). As on the training set, the global BP-ANN performance on the masses was significantly (p < 0.01) better than that on calcifications (Table 7-6) in the evaluation set.

Table 7-6. ROC performance on the mass and calcification subsets for the global BP-ANN trained on the training set and tested on the evaluation set (both were mixes of cases that included mass, calcifications, and other lesions). The standard deviations were estimated by bootstrap sampling on the network outputs (Section 1.3.1).

	AUC	partial AUC
Mass	0.893 ± 0.010	0.448 ± 0.015
Calcification	0.711 ± 0.018	0.147 ± 0.027

7.5.3 Patient Age

The difference in the global BP-ANN performance on the subset of younger women in the training and evaluation set was not significant (AUC p = 0.09, partial AUC p = 0.12). However, the difference in the global BP-ANN performance on the subset of older women in the training and evaluation set was significant in AUC (p < 0.01) though the difference in partial AUC was not significant (p = 0.11).

The comparison of the global BP-ANN on the younger and older women was very different for the training (Section 4.2.3) and evaluation sets (Table 7-7). On the training set the model performed better on the younger women than on the older women (AUC p < 0.01, partial AUC p < 0.01) while on the evaluation set the model performed somewhat better on the older women than on the younger women (AUC p = 0.04, partial AUC = 0.97). These confounding results on the age subsets are apparently due to sampling effects (Section 7.7.2). Thus, there are some potentially important differences introduced by sampling even for this very large database.

Table 7-7. ROC performance on the subsets of younger and older women for the global BP-ANN trained on the training set and tested on the evaluation set. The standard deviations were estimated by bootstrap sampling on the network outputs (Section 1.3.1).

	AUC	partial AUC
Age ≤ 55	0.792 ± 0.015	0.270 ± 0.040
Age > 55	0.832 ± 0.012	0.268 ± 0.033

7.6 Global BP-ANN Performance on the SOM Clusters in Evaluation Set

As described in Section 7.3, the global BP-ANN was trained on the training set and tested on the evaluation set and a threshold (0.1842) was applied that had given approximately 98% sensitivity on the training set. On the evaluation set, the overall sensitivity was also approximately 98% (884 / 904) and the specificity was approximately 23% (296 / 1273). In other words, 316 evaluation cases (296 actual negatives and 20 actual positives) fell below the threshold. As described in Section 7.2, each of the evaluation cases was mapped to a cluster in the training set identified by the SOM. Table 7-8 shows the distribution of BP-ANN's true negative and false negative classifications across the clusters in the evaluation set. As was seen on the training set (Table 4-14, Section 4.3.2), the majority of the cancers in the evaluation set that the BP-ANN would have incorrectly referred to follow up (14 / 20 = 70%) and the majority of the benign lesions in the evaluation set that the BP-ANN would have correctly spared biopsy (198 / 296 = 67%) were in SOM cluster #6.

Table 7-8. The global BP-ANN (Sections 3.4 and 7.3) was applied to the evaluation set (Sections 1.4 and 7.1). A threshold of 0.1842, which gave approximately 98% sensitivity on the training set, was applied. The distributions of the true negatives and false negatives across the clusters identified by the SOM (Sections 2.4.2 and 7.2) are shown. There were 1273 actual negatives and 904 actual positives.

Cluster	N	Percent Malignant	True Negatives	False Negatives
1	69	30%	19	4
2	130	8%	64	2
3	129	43%	0	0
4	201	86%	0	0
5	9	-	1	0
6	237	8%	198	14
7	127	22%	4	0
8	207	65%	0	0
9	238	45%	0	0
10	69	28%	2	0
11	99	83%	0	0
12	0	-	0	0
13	276	38%	0	0
14	324	32%	8	0
15	5	-	0	0
16	57	70%	0	0

7.7 Resampling Effects on Global BP-ANN Model

Previously we described how even the creation of an independent evaluation set for validation may still be biased by the specific cases sampled into the training vs. evaluation sets. For example, if more cases of type "A" and less of type "B" happened to be selected for the training set, then the resulting model may fail to generalize to the evaluation set containing more "B" and less "A" cases. Some form of cross-validation or bootstrap sampling would be necessary to resolve this concern, but doing so would be prohibitively expensive.

7.7.1 Resampling Effects: Overall BP-ANN Performance

In this section, we present a preliminary study of the effect of using different random splits of the data into training and evaluation sets (Section 1.4). For each training set, the BP-ANN was trained in a round-robin manner using the network parameters determined from the primary split used throughout the dissertation (Section 3.4). A BP-ANN was trained on all the training cases using the same network parameters and tested on the evaluation cases for each split of the data. Table 7-9 shows the variation in the ROC performance (Section 1.3.1) of the BP-ANN across the splits of the data into training and evaluation sets. The row labeled "primary" shows the results for the split of the data used throughout the dissertation. The variability in performance due to sampling is more pronounced for the partial AUC (standard deviation ~0.020, ~5% error) than the AUC (standard deviation ~0.005, ~0.5% percent error). This is not unexpected since the partial AUC is based on a smaller fraction of the cases, but is unfortunate since this measure is more clinically relevant than the AUC. Notice that the AUC values for the primary split are around 60-80% of a standard deviation from the estimated mean while the partial AUC values for the primary split are around 110-150% of a standard deviation from the estimated mean. Over all, the primary split does not appear to be an outlier in terms of the over all performance relative to that seen on the other random splits. However, the observed variability in performance due to sampling despite the large size of the data set implies that caution should be taken in interpreting any small differences reported in this dissertation, particularly in the partial AUC.

Table 7-9. Performance of the global BP-ANN for several random splits of the data into training and evaluation sets.

	Percent Malignant		**AUC**		**partial AUC**	
	Train.	**Eval.**	**Train.**	**Eval.**	**Train.**	**Eval.**
Primary	43%	42%	0.8204	0.8315	0.3456	0.3098
S1	43%	42%	0.8130	0.8343	0.2824	0.3594
S2	42%	43%	0.8238	0.8196	0.3212	0.3298
S3	41%	44%	0.8170	0.8353	0.3231	0.3263
S4	44%	41%	0.8156	0.8211	0.2825	0.3407
S5	42%	43%	0.8233	0.8269	0.3037	0.3423
S6	43%	42%	0.8158	0.8283	0.2881	0.3449
S7	42%	43%	0.8165	0.8258	0.3046	0.3051
S8	42%	43%	0.8146	0.8294	0.3123	0.3454
S9	44%	41%	0.8152	0.8270	0.3327	0.3091
S10	42%	43%	0.8242	0.8270	0.3394	0.3114
Mean	43%	42%	0.8181	0.8278	0.3123	0.3295
Stdev.	1%	1%	0.0040	0.0048	0.0223	0.0185

7.7.2 Resampling Effects: BP-ANN Performance on Age Subsets

As described in Section 7.5.3, the comparison of the global BP-ANN on the younger and older women was very different for the training (Section 4.2.3) and evaluation sets (Table 7-7). In this section, we investigated the role that sampling into training and evaluation sets played in this phenomenon. The same resampling into training and evaluation sets and BP-ANN models were used as were discussed in previous Section 7.7.1.

We found that the there was no statistically significant difference (unpaired z-test, p = 0.76) between the mean AUC across the resamples for the younger age subset (0.8121 ± 0.0135) and the older age subset (0.8092 ± 0.0131). Likewise, there was no statistically significant difference (unpaired z-test, p = 0.16) between the mean partial AUC across the resamples for the younger age subset (0.2988 ± 0.0480) and the older age subset (0.2252 ± 0.0217). Thus, the apparent differences we had observed between the older and younger women were apparently artifacts of which cases were sampled into the training and evaluation sets.

7.8 Summary

In this section, we described the generalization tests performed with the evaluation half of the data set. By using round-robin sampling on the training half and performing a final verification using the evaluation set, we assessed the system using three independent portions of the data in a training-testing-validation strategy. This approach can help overcome the bias inherent in using round-robin sampling alone.

Over all, the global BP-ANN and global CBR models generalized well to the evaluation set (Section 7.3). Our conclusions regarding generalization were mixed for the comparison of the performance of the global BP-ANN across clusters in the training and evaluation sets. Certain trends were were clearly apparent with both the training and evaluation sets: (a) better performance on masses than calcifications (Section 7.5.2), (b) similar size and malignancy fraction of the clusters identified by the SOM (Section 7.2), and (c) similar behavior of the global BP-ANN in focusing on a particular cluster (#6) in terms of the correct and incorrect recommendations for follow up (Section 7.6). Other conclusions from working with the training data were somewhat weakened by the analysis of the evaluation cases. The cross-institutional studies (Section 7.5.1) suggest that sampling may still be affecting our ability to discern institutional effects, even with a large data set. In particular, one borderline significant difference in the partial AUC was observed between two institution subsets in the evaluation set that was not seen in the training set. The apparent trend with the age subsets observed on the training cases was reversed on the evaluation cases (Section 7.5.3). Resampling experiments with the BP-

8 Summary, Conclusions, and Future Work

The purpose of this study was to investigate modular and ensemble systems of machine learning methods for computer-aided diagnosis (CAD) of breast cancer. The CAD methods were developed to reduce the number of benign biopsies. In this section, we summarize the major results of the dissertation.

In Section 1, we provided an overview of breast cancer and mammography (Section 1.1), the role that computer-aided diagnosis can play (Section 1.2), and modular and ensemble breast cancer CAD systems performed (Section 1.2.1). While mammography is valuable for early detection of breast cancer, it has a high false-positive rate. A CAD system for referring benign lesions to short-term follow-up instead of biopsy could spare women discomfort, anxiety, and expense and potentially improve the cost-effectiveness of mammographic screening programs. Evaluation of CAD systems is critical since the consequences of a delayed diagnosis of cancer can be dire. We described the importance of ROC analysis (Section 1.3.1) and sampling (Section 1.3.3) in evaluating breast cancer CAD systems. In this study, considerable attention was paid to the issue of model generalization; thus, the data were partitioned into training and evaluation halves and furthermore round-robin sampling was used in building models on the training half. We presented a case study of the relationship between performance metrics from ROC analysis and one commonly used in developing breast cancer CAD systems (Section 1.3.2). From this case study we concluded that predictive models that minimize the mean squared error may also maximize the area (AUC) under the ROC curve, but may not maximize the partial area (partial AUC) under the high sensitivity region of the ROC curve. The partial AUC is more clinically relevant than the AUC since a breast cancer

CAD system must maintain high sensitivity (*i.e.*, delaying the diagnosis of breast cancer is generally worse than a benign biopsy). Finally, we supplied a detailed description of the data set used for the remainder of this dissertation (Section 1.4). It is worth noting that the database used in this study was very large and was comprised of cases mixed from multiple institutions (Duke, UPenn, DDSM). The database was randomly split into two halves: 2258 cases for training and 2177 cases for evaluation. Each breast lesion underwent biopsy and was described by six mammographic findings (BI-RADS™) and patient age.

In Section 2, we described the clusters that could be identified in the training set using *a priori* knowledge (Section 2.2) or unsupervised learning techniques (Section 2.4). Subsets defined by institution (Section 2.2.1), lesion type (Section 2.2.2), and patient age (Section 2.2.3) were considered. Three unsupervised learning techniques were used: agglomerative hierarchical clustering followed by K-Means (Section 2.4.1), Self-Organizing Map (Section 2.4.2), and AutoClass (Section 2.4.3). Some agreement was seen between the clusters identified by the three unsupervised learning methods. Of particular interest, all three identified a cluster of mostly benign masses. Using profiling techniques (Section 2.3), we described a typical case in these clusters as a younger woman with a well-circumscribed or obscured, oval-shaped mass (clusters E, 6, and β).

In Section 3, we investigated five supervised machine learning models for predicting biopsy outcome from mammographic findings and patient age using the training set: linear discriminant analysis (LDA, Section 3.2), support vector machines (SVM, Section 3.3), back-propagation artificial neural networks (BP-ANN, Section 3.4), case-based reasoning (CBR, Section 3.5), and classification and regression trees (CART, Section

3.6). Each of these "global" models was built in a round-robin fashion on the training set. We found that the non-linear models (BP-ANN, CBR, CART) were generally superior to the linear models (LDA, SVM) for this task. In our previous work with smaller data sets we had not been able to demonstrate the superiority of non-linear models over linear models for this task. The global BP-ANN and global CBR models were considered the most promising since they showed the highest partial AUC values over all training cases (Table 3-1). For the BP-ANN, the AUC = 0.820 ± 0.009 and the partial AUC = 0.347 ± 0.022. For the CBR, the AUC = 0.788 ± 0.009 and the partial AUC = 0.324 ± 0.019.

In Section 4, we examined the performance of the global models (Section 3) over the clusters identified using *a priori* knowledge (Section 2.2) and unsupervised learning (Section 2.4). The global models were trained in a round-robin fashion over the entire training set and then evaluated in terms of their performance on the clusters of cases. One of the most striking results of this dissertation was that CAD systems trained on a mixture of lesion types performed much better on masses than on calcifications (Section 4.2.2). The study of the institutional effects suggests that models built on cases mixed between institutions may overcome some of the weaknesses of models built on cases from a single institution (Section 4.2.1, Section 5.2.1). There were no significant differences in the AUC or partial AUC of the global BP-ANN in pair-wise comparisons on the institution subsets (Table 4-1, Table 4-2) and the same threshold on the BP-ANN gave approximately 98% sensitivity and 23% specificity on all three of the institution subsets (Table 4-5). While we observed that there was no benefit in terms of the AUC or partial AUC index in using a BP-ANN specifically trained for the cases at an institution rather than the global BP-ANN trained on cases mixed between the institutions, we also found that a BP-ANN trained on cases from one institution did not always perform the same way (AUC, partial AUC, sensitivity and specificity for a fixed threshold) when tested on cases from another institution (Section 5.2.1). Thus, further cross-institutional studies of breast cancer CAD systems are still needed. Another very interesting result is that each of the unsupervised methods identified a cluster that accounted for the majority of the BP-ANN's recommendations for follow up (clusters E, 6, and β, Section 4.4). In other words, each clustering technique identified a cluster that contained the majority of the benign cases that would have been correctly referred to follow-up and the majority of the malignant cases that would have been incorrectly referred to follow-up.

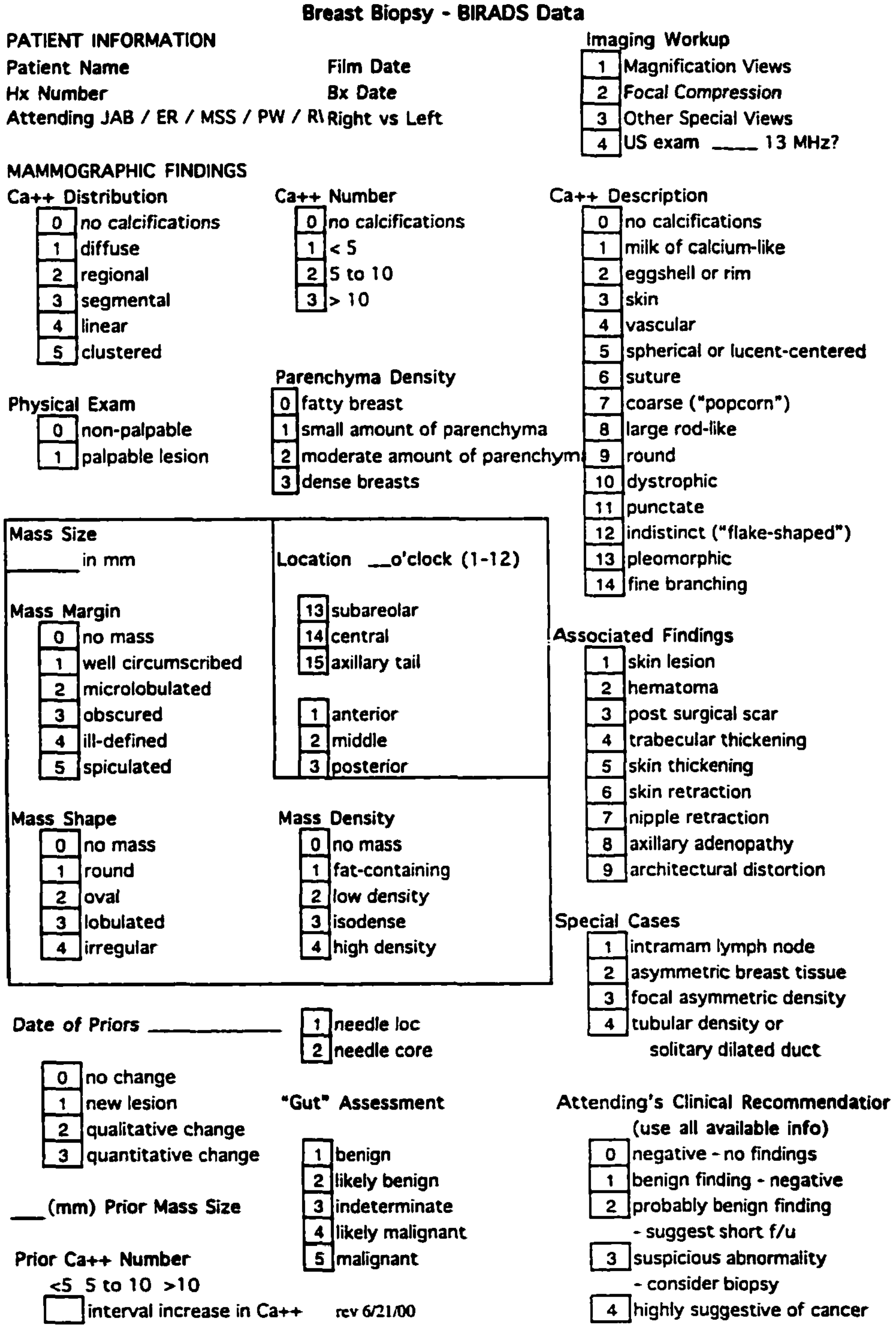
In Section 5, we developed modular CAD systems by building local models specifically for each of the clusters identified using *a priori* knowledge (Section 2.2) and unsupervised learning (Section 2.4). Each "local" model was trained in a round-robin fashion only on the cases in a cluster identified in the training data. While some local models were superior to some global models, we were unable to build a modular CAD system that was better than the global BP-ANN model, which was considered to be a "gold standard" since we have used BP-ANN models extensively in our laboratory and the overall performance of the global BP-ANN was generally better than that of the other global models (Section 3.7). We consider it unlikely that additional work with similar modular systems would prove fruitful. However, the cluster analysis and local models also lead us to an unexpected, interesting result. We developed a simple diagnostic rule from the local CART model for masses and the profiles of the mass clusters identified by the unsupervised learning methods (clusters E, 6, and β): if the Mass Margin was well-circumscribed or obscured and the age was less than 59 years and there were no calcifications, associated findings, or special findings, then don't biopsy, otherwise do biopsy. On the 2258 training cases, this rule gave 961 / 982 = 98% sensitivity and 336 / 1276 = 26% specificity. In other words, this rule performed comparably to the global BP-ANN with a threshold of 0.1842 (965 / 982 = 98% sensitivity, 303 / 1276 = 24% specificity). There are several potential advantages of a simple rule over more complicated models. First, such a rule would be trivial to implement. Second, its simplicity makes it more understandable and thus clinicians may more readily accept it. Third, the transparency of the rule allows for more direct comparisons to clinically accepted criteria and guidelines. Comparison with current clinical criteria is an important area for future work.

In Section 6, we investigated ensemble CAD systems in which the same cases were used to train multiple models, whose predictions were then combined. Simple combinations (min, max, mean) of the continuous round-robin model outputs were considered and evaluated in terms of the AUC, partial AUC, and the specificity at 98% sensitivity. Thresholds were also applied to the continuous model outputs to give binary predictions which were combined by logical "and" and "or" functions. However, these simplistic combination schemes did not result in significant improvements. We also investigated using a perceptron and a BP-ANN with a hidden layer to combine continuous model outputs and feature inputs. These more complicated combination schemes using "round-robin of round-robin" sampling were found to be unduly optimistic. However, it should be noted that there isn't any way to know *a priori* what models should be combined or how they should be combined. Thus, additional work in this area could be beneficial and may be warranted.

In Section 7, we tested the generalization of the results observed on the training half of the data to the evaluation half of the data. A major concern is the ability of CAD systems to perform on a new data set in approximately the same way as it performed on the data set used to construct it (*i.e.*, to generalize). In order to address this issue, the data were randomly partitioned into two halves for training and evaluation (Sections 1.3.3 and 1.4). In Sections 2, 3, 4, 5, and 6 we described the results of unsupervised and supervised learning techniques applied to the training set and in Section 7 we tested the generalization of the results to the evaluation set. Over all, the global BP-ANN and global CBR models generalized well to the evaluation set (Section 7.3). However, resampling experiments (Section 7.7) with the BP-ANN suggested that the error on the AUC due to random fluctuations between the training and evaluation sets may be approximately 0.005 and the error on the partial AUC may be approximately 0.02. In particular, studies with the age (Section 7.5.3, Section 7.7.2) and institution (Section 7.5.1) subsets suggested that sampling might still have affected our ability to discern some effects, even with such a large data set, since differences were seen between the training and evaluation set results. For example, the partial AUC of the global BP-ANN on the DDSM subset was significantly lower for the evaluation set than for the training set. The better performance on masses than calcifications seen with the training set was readily apparent on the evaluation set (Section 7.5.2). In particular, the global BP-ANN performed significantly better on the masses than on the calcifications in the evaluation set in terms of both the AUC and the partial AUC. The correlation of a particular cluster with the BP-ANN recommendations for follow up was also confirmed with the evaluation set (Section 7.6; clusters E, 6, and β). The simple classification rule based on the cluster profiles and the CART models (Section 5.2.2) generalized well to the evaluation set and performed comparably to the global BP-ANN (Section 7.4).

In conclusion, a comprehensive study was undertaken to use machine learning techniques for the computer-aided diagnosis of breast cancer. In particular, the goal was to increase the specificity of mammography-induced breast biopsy. This is a timely and significant problem in biomedical engineering. One the largest data sets of its type was assembled from three independent institutions. A wide variety of modeling techniques were evaluated, individually and in tandem with each other. The data were likewise analyzed as a global whole and in terms of subsets. The overall intent was to engineer modular and ensemble systems using this large data set and the rich variety of tools available. Somewhat to our surprise, these systems tended to match but not exceed the performance of a classic feed-forward, back-propagation artificial neural network. As a result of this endeavor, however, we clearly identified both the potential promises and problems inherent in the use of a large, heterogeneous data set, *e.g.*, issues such as generalization across institutions and important difference between subtypes of cases such as masses vs. calcifications. We hope that these discoveries will move computer-aided diagnosis of breast cancer closer to eventual clinical implementation.

Appendix 1: Duke Data Collection Form



Breast Biopsy - BIRADS Data

PATIENT INFORMATION

Patient Name
Hx Number
Attending JAB / ER / MSS / PW / R\

Film Date
Bx Date
Right vs Left

Imaging Workup

1 Magnification Views
2 *Focal Compression*
3 Other Special Views
4 US exam ____ 13 MHz?

MAMMOGRAPHIC FINDINGS

Ca++ Distribution

0 *no calcifications*
1 diffuse
2 regional
3 segmental
4 linear
5 clustered

Physical Exam

0 non-palpable
1 palpable lesion

Ca++ Number

0 no calcifications
1 < 5
2 5 to 10
3 > 10

Parenchyma Density

0 fatty breast
1 small amount of parenchyma
2 moderate amount of parenchym
3 dense breasts

Ca++ Description

0 no calcifications
1 milk of calcium-like
2 eggshell or rim
3 skin
4 vascular
5 spherical or lucent-centered
6 suture
7 coarse ("popcorn")
8 large rod-like
9 round
10 dystrophic
11 punctate
12 indistinct ("flake-shaped")
13 pleomorphic
14 fine branching

Mass Size
______ in mm

Mass Margin

0 no mass
1 well circumscribed
2 microlobulated
3 obscured
4 ill-defined
5 spiculated

Location __o'clock (1-12)

13 subareolar
14 central
15 axillary tail

1 anterior
2 middle
3 posterior

Mass Shape

0 no mass
1 round
2 oval
3 lobulated
4 irregular

Mass Density

0 no mass
1 fat-containing
2 low density
3 isodense
4 high density

Associated Findings

1 skin lesion
2 hematoma
3 post surgical scar
4 trabecular thickening
5 skin thickening
6 skin retraction
7 nipple retraction
8 axillary adenopathy
9 architectural distortion

Special Cases

1 intramam lymph node
2 asymmetric breast tissue
3 focal asymmetric density
4 tubular density or solitary dilated duct

Date of Priors __________

0 no change
1 new lesion
2 qualitative change
3 quantitative change

___(mm) Prior Mass Size

Prior Ca++ Number
<5 5 to 10 >10
□ interval increase in Ca++

1 needle loc
2 needle core

"Gut" Assessment

1 benign
2 likely benign
3 indeterminate
4 likely malignant
5 malignant

rev 6/21/00

Attending's Clinical Recommendatior
(use all available info)

0 negative - no findings
1 benign finding - negative
2 probably benign finding - suggest short f/u
3 suspicious abnormality - consider biopsy
4 highly suggestive of cancer

Appendix 2: DDSM Sample Files

Example 1

DDSM Case 1828 OVERLAY file

```
TOTAL_ABNORMALITIES 1
ABNORMALITY 1
LESION_TYPE MASS SHAPE ASYMMETRIC_BREAST_TISSUE MARGINS
ILL_DEFINED
ASSESSMENT 4
SUBTLETY 4
PATHOLOGY MALIGNANT
TOTAL_OUTLINES 1
ABNORMALITY 2
LESION_TYPE CALCIFICATION TYPE FINE_LINEAR_BRANCHING
DISTRIBUTION LINEAR
ASSESSMENT 4
SUBTLETY 4
PATHOLOGY MALIGNANT
TOTAL_OUTLINES 1
```

DDSM Case 1828 ICS file

```
ics_version 1.0
filename A-1828-1
DATE_OF_STUDY 17 12 1996
PATIENT_AGE 64
FILM
FILM_TYPE REGULAR
DENSITY 2
DATE_DIGITIZED 4 2 1999
DIGITIZER HOWTEK 43.5
SEQUENCE
LEFT_CC LINES 6871 PIXELS_PER_LINE 3886 BITS_PER_PIXEL 12
RESOLUTION 43.5 OVERLAY
LEFT_MLO LINES 6601 PIXELS_PER_LINE 3676 BITS_PER_PIXEL 12
RESOLUTION 43.5 OVERLAY
RIGHT_CC LINES 6196 PIXELS_PER_LINE 3556 BITS_PER_PIXEL 12
RESOLUTION 43.5 NON_OVERLAY
RIGHT_MLO LINES 6586 PIXELS_PER_LINE 3466 BITS_PER_PIXEL 12
RESOLUTION 43.5 NON_OVERLAY
```

The OVERLAY and ICS files are shown above for the mediolateral oblique (MLO) view of the left breast of case 1828 in cancer volume 11 in the Digital Database for Screening Mammography (DDSM). Since there were two "PATHOLOGY" values in the OVERLAY file, it was parsed as two lesions or cases for this study. Notice that the Mass Shape value of "asymmetric breast tissue" was translated into "Mass Shape = No Mass = 0" and "Special Findings = asymmetric breast tissue = 2". The patient age was parsed from the ICS file. The encodings of these cases as specified by Table 1-3 are shown below.

ID	4412	4082
Biopsy Outcome	1	1
Calcification Distribution	0	4
Calcification Morphology	0	14
Mass Margin	4	0
Mass Shape	0	0
Associated Findings	0	0
Special Findings	2	0
Age	64	64

Example 2

DDSM Case 3125 OVERLAY file

```
TOTAL_ABNORMALITIES 1
ABNORMALITY 1
LESION_TYPE CALCIFICATION TYPE PUNCTATE-PLEOMORPHIC
DISTRIBUTION CLUSTERED
ASSESSMENT 4
SUBTLETY 2
PATHOLOGY BENIGN
TOTAL_OUTLINES 1
```

DDSM Case 3125 ICS file

```
ics_version 1.0
```

```
filename B-3125-1
DATE_OF_STUDY 15 1 1997
PATIENT_AGE 61
FILM
FILM_TYPE REGULAR
DENSITY 3
DATE_DIGITIZED 4 3 1998
DIGITIZER LUMISYS LASER
SEQUENCE
LEFT_CC LINES 4688 PIXELS_PER_LINE 2712 BITS_PER_PIXEL 12
RESOLUTION 50 OVERLAY
LEFT_MLO LINES 4704 PIXELS_PER_LINE 2640 BITS_PER_PIXEL 12
RESOLUTION 50 OVERLAY
RIGHT_CC LINES 4768 PIXELS_PER_LINE 2640 BITS_PER_PIXEL 12
RESOLUTION 50 NON_OVERLAY
RIGHT_MLO LINES 4720 PIXELS_PER_LINE 2672 BITS_PER_PIXEL 12
RESOLUTION 50 NON_OVERLAY
```

The OVERLAY and ICS files are shown above for the mediolateral oblique (MLO) view of the left breast of case 3125 in benign volume 1 in the Digital Database for Screening Mammography (DDSM). Since there was one "PATHOLOGY" value in the OVERLAY file, it was parsed as one lesion or case for this study. Notice that there was more than one value for Calcification Morphology, punctate-pleomorphic, so the one more suspicious for malignancy, pleomorphic, was used. The patient age was parsed from the ICS file. The encoding of this case as specified by Table 1-3 is shown below.

ID	3093
Biopsy Outcome	0
Calcification Distribution	5
Calcification Morphology	13
Mass Margin	0
Mass Shape	0
Associated Findings	0
Special Findings	0
Age	61

Appendix 3: Feature Histograms

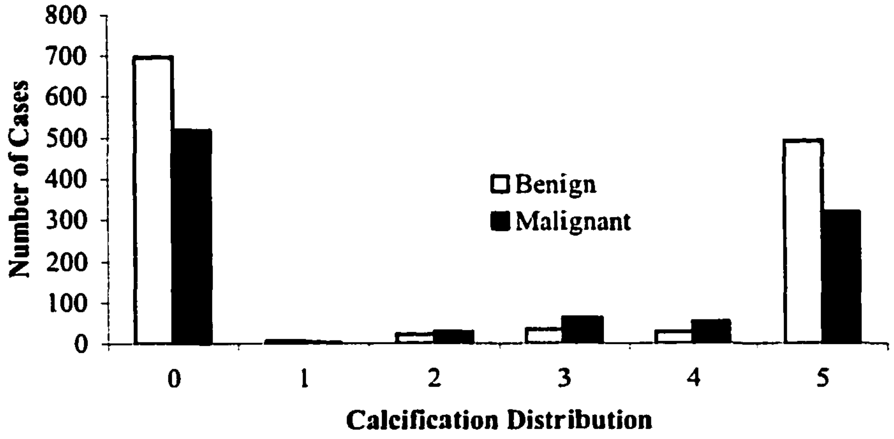


Figure A3 - 1. Distribution of the Calcification Distribution in the training set.

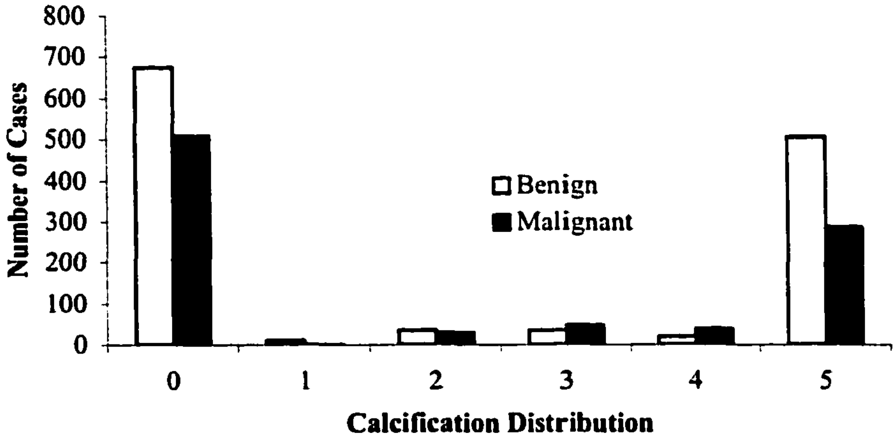


Figure A3 - 2. Distribution of the Calcification Distribution in the evaluation set.

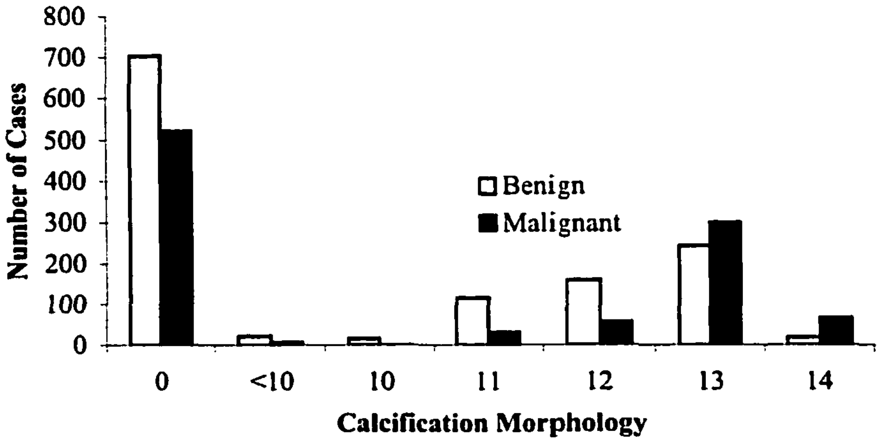


Figure A3 - 3. Distribution of the Calcification Morphology in the training set.

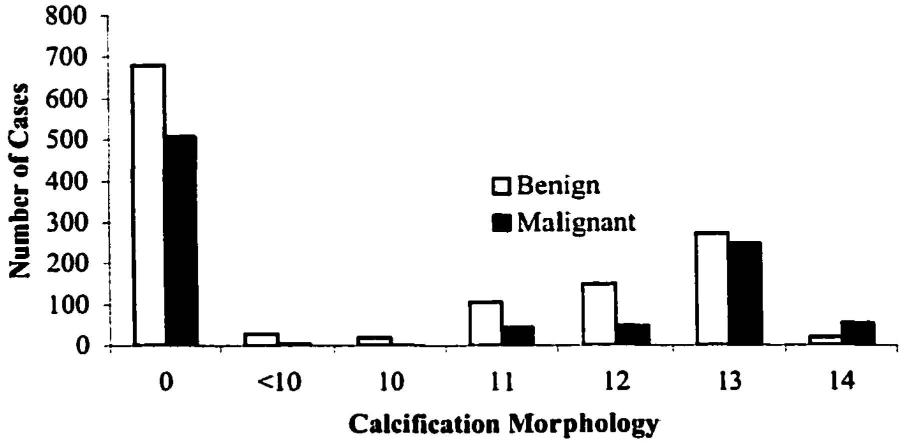


Figure A3 - 4. Distribution of the Calcification Morphology in the evaluation set.

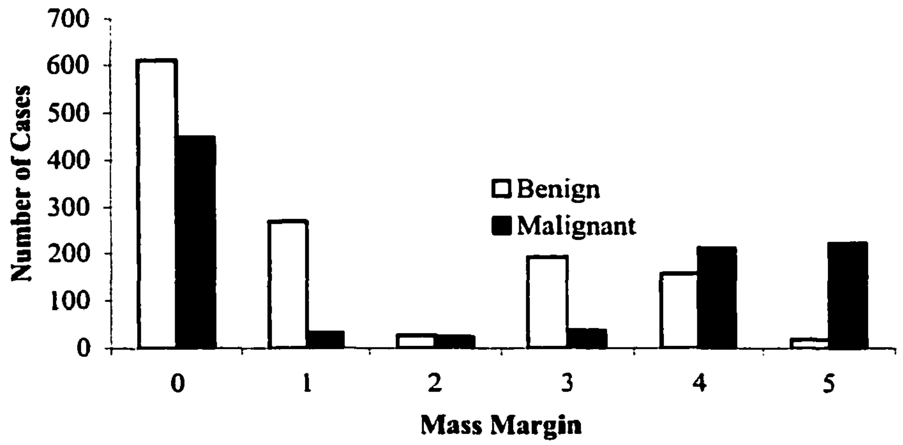


Figure A3 - 5. Distribution of the Mass Margin in the training set.

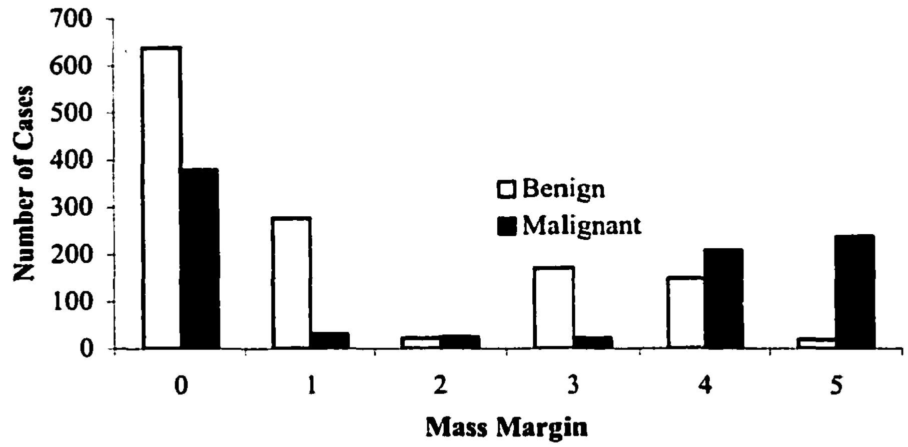


Figure A3 - 6. Distribution of the Mass Margin in the evaluation set.

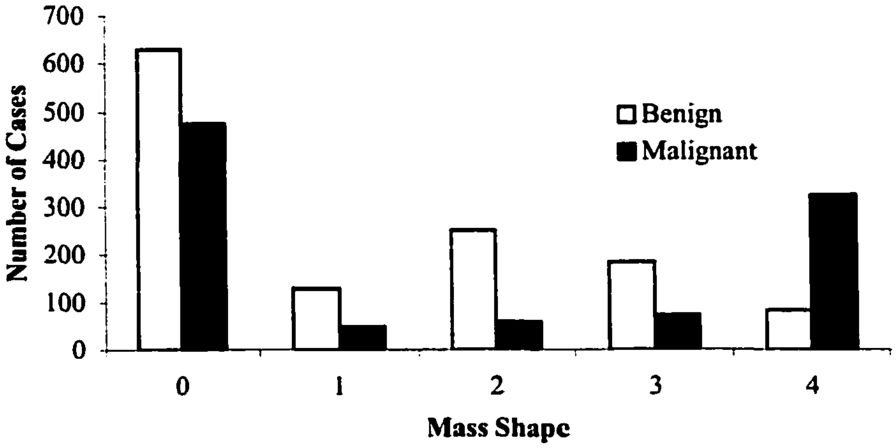


Figure A3 - 7. Distribution of the Mass Shape in the training set.

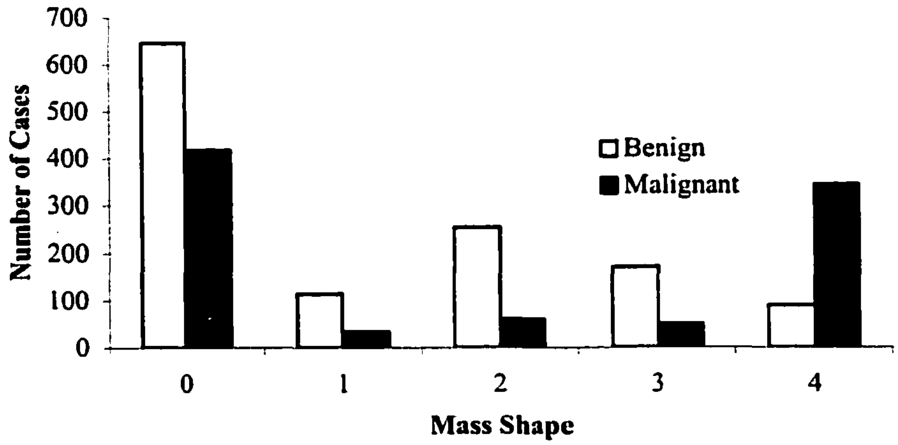


Figure A3 - 8. Distribution of the Mass Shape in the evaluation set.

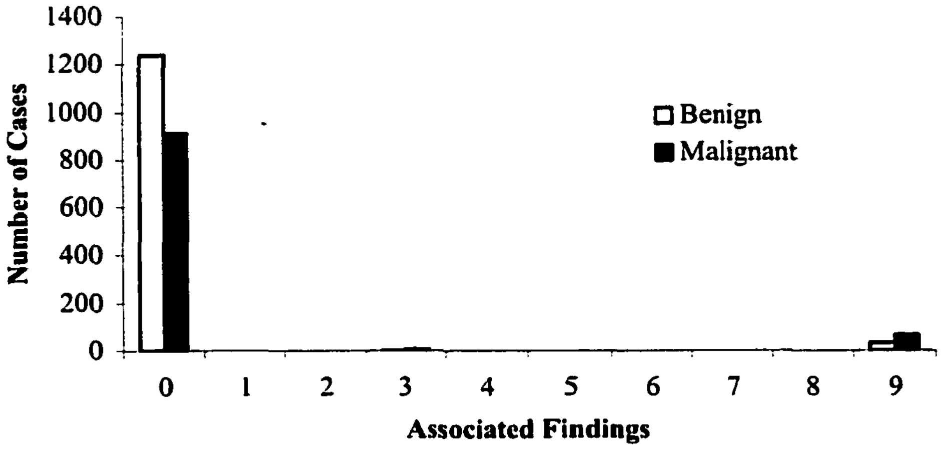


Figure A3 - 9. Distribution of the Associated Findings in the training set.

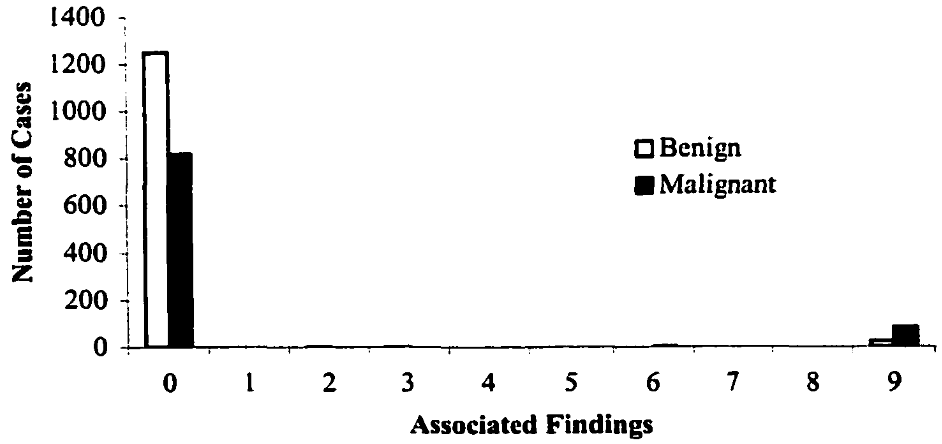


Figure A3 - 10. Distribution of the Associated Findings in the evaluation set.

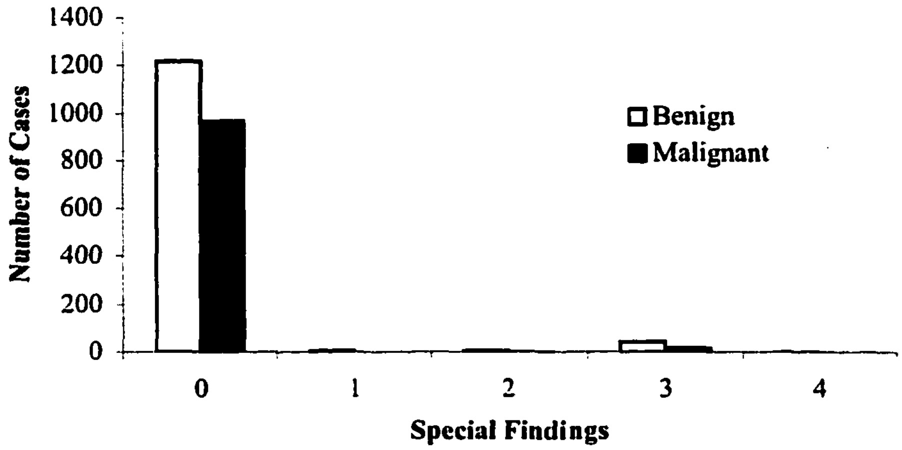


Figure A3 - 11. Distribution of the Special Findings in the training set.

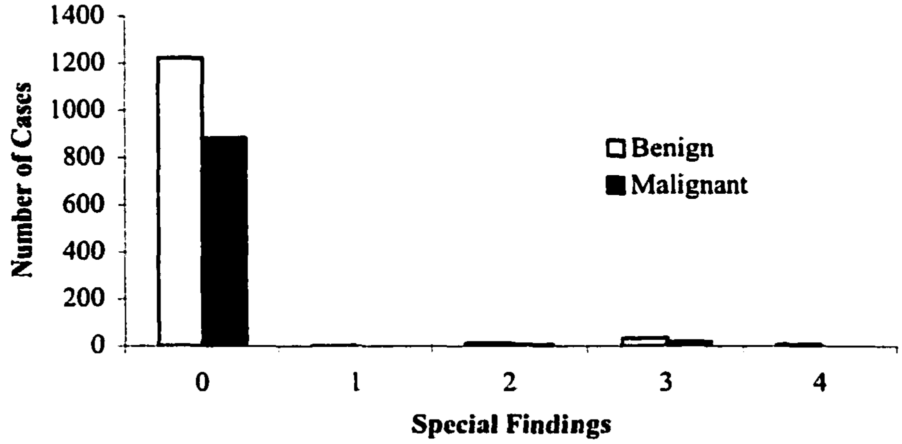


Figure A3 - 12. Distribution of the Special Findings in the evaluation set.

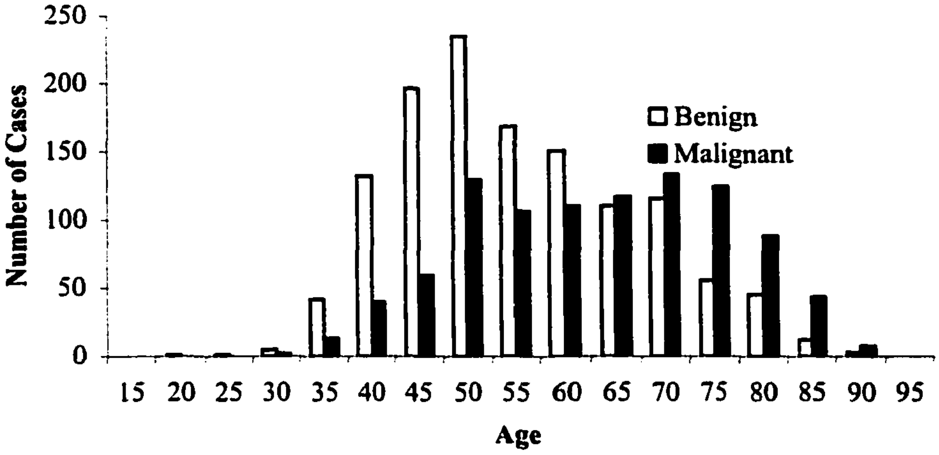


Figure A3 - 13. Distribution of Patient Age in the training set.

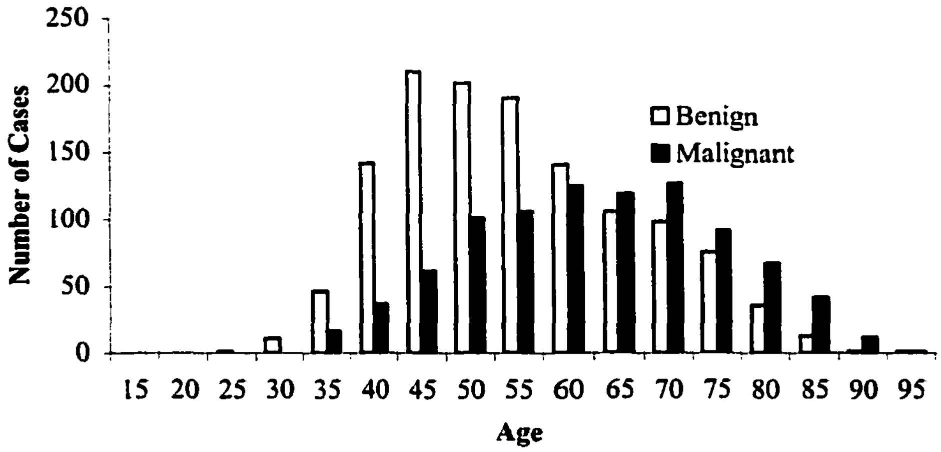


Figure A3 - 14. Distribution of Patient Age in the evaluation set.

12. Orel, S.G., *et al.*, BI-RADS categorization as a predictor of malignancy. Radiology, 1999. **211**(3): p. 845-50.

13. Lacquement, M.A., D. Mitchell, and A.B. Hollingsworth, Positive predictive value of the Breast Imaging Reporting and Data System. Journal of the American College of Surgeons, 1999. **189**(1): p. 34-40.

14. Baker, J.A., *et al.*, Breast cancer: prediction with artificial neural network based on BI-RADS standardized lexicon. Radiology, 1995. **196**(3): p. 817-822.

15. Baker, J.A., P.J. Kornguth, and C.E. Floyd, Jr, Breast imaging reporting and data system standardized mammography lexicon: observer variability in lesion description. AJR. American Journal of Roentgenology, 1996. **166**(4): p. 773-8.

16. Lo, J.Y., *et al.*, Predicting breast cancer invasion with artificial neural networks on the basis of mammographic features. Radiology, 1997. **203**(1): p. 159-163.

17. Tourassi, G.D., *et al.*, A Neural Network Approach to Breast Cancer Diagnosis as a Constraint Satisfaction Problem. Medical Physics, 2001. **28**(5): p. 804-811.

18. Floyd, C.E., Jr, J.Y. Lo, and G.D. Tourassi, Cased-based reasoning computer algorithm that uses mammographic findings for breast biopsy decisions. AJR. American Journal of Roentgenology, 2000. **175**: p. 1347-1352.

19. Berg, W.A., *et al.*, Breast Imaging Reporting and Data System: inter- and intraobserver variability in feature analysis and final assessment. AJR. American Journal of Roentgenology, 2000. **174**(6): p. 1769-77.

20. Kerlikowske, K., *et al.*, Variability and accuracy in mammographic interpretation using the American College of Radiology Breast Imaging Reporting and Data System. Journal of the National Cancer Institute, 1998. **90**(23): p. 1801-9.

21. Baker, J.A., *et al.*, Artificial neural network: improving the quality of breast biopsy recommendations. Radiology, 1996. **198**: p. 131-135.

22. Lo, J.Y., *et al.*, Computer-aided diagnosis of breast cancer: artificial neural network approach for optimized merging of mammographic features. Academic Radiology, 1995. **2**(10): p. 841-850.

23. Lo, J.Y., *et al.*, Effect of patient history data on the prediction of breast cancer from mammographic findings with artificial neural networks. Academic Radiology, 1999. **6**(1): p. 10-15.

24. Doi, K., *et al.*, Computer-aided diagnosis in radiology: potential and pitfalls. European Journal of Radiology, 1999. **31**(2): p. 97-109.

25. Vybomy, C.J., M.L. Giger, and R.M. Nishikawa, Computer-aided detection and diagnosis of breast cancer. Radiologic Clinics of North America, 2000. **38**(4): p. 725-740.

26. Giger, M.L., Computer-aided diagnosis of breast lesions in medical images. Computing in Science and Engineering, 2000. **2**(5): p. 39-45.

27. Giger, M.L., N. Karssemeijer, and S.G. Aramato, III, Guest editorial computer-aided diagnosis in medical imaging. IEEE Transaction on Medical Imaging, 2001. **20**(12): p. 1205-1208.

28. Giger, M.L., Computer-aided diagnosis in radiology. Academic Radiology, 2002. **9**: p. 1-3.

29. Physician Insurers Association of America, Breast Cancer Study, 2002.

30. Tourassi, G.D., Journey toward computer-aided diagnosis: role of image texture analysis [editorial; comment]. Radiology, 1999. **213**(2): p. 317-20.

31. Chan, H.P., *et al.*, Computerized analysis of mammographic microcalcifications in morphological and texture feature spaces. Medical Physics, 1998. **25**(10): p. 2007-19.

32. Sahiner, B., *et al.*, Improvement of mammographic mass characterization using spiculation measures and morphological features. Medical Physics, 2001. **28**(7): p. 1455-1465.

33. Petrick, N., *et al.*, Automated detection of breast masses on mammograms using adaptive contrast enhancement and texture classification. Medical Physics, 1996. **23**(10): p. 1685-96.

34. Sahiner, B., *et al.*, Design of a high-sensitivity classifier based on a genetic algorithm: application to computer-aided diagnosis. Phys Med Biol, 1998. **43**(10): p. 2853-71.

35. Kupinski, M.A., M.A. Anastasio, and M.L. Giger. Multiobjective genetic optimization of diagnostic classifiers used in computerized detection of mass lesions in mammography. in SPIE Medical Imaging 2000: Image Processing. 2000.

36. Gavrielides, M.A., *et al.*, Segmentation of suspicious clustered microcalcifications in mammograms. Medical Physics, 2000. **27**(1): p. 13-22.

37. Wu, Y., *et al.*, Artificial neural networks in mammography: Application to decision making in the diagnosis of breast cancer. Radiology, 1993. **187**: p. 81-87.

38. Floyd, C.E., Jr, *et al.*, Prediction of breast cancer malignancy using an artificial neural network. Cancer, 1994. **74**(11): p. 2944-2948.

39. Jiang, Y., *et al.*, Malignant and benign clustered microcalcifications: automated feature analysis and classification. Radiology, 1996. **198**(3): p. 671-8.

40. Chan, H.P., *et al.*, Computerized classification of malignant and benign microcalcifications on mammograms: texture analysis using an artificial neural network. Physics in Medicine and Biology, 1997. **42**(3): p. 549-67.

41. Sharkey, A.J.C., ed. Combining Artificial Neural Nets: Ensemble and Modular Multi-Net Systems. Perspectives in Neural Computing, ed. J.G. Taylor. 1999, Springer-Verlag.

42. Li, L., *et al.*, False-positive reduction in CAD mass detection using a comptetive classification strategy. Medical Physics, 2001. **28**(2): p. 250-258.

43. Rymon, R., *et al.*, Incorporation of a set enumeration trees-based classifier into a hybrid computer-assisted diagnosis scheme for mass detection. Acad Radiol, 1998. **5**(3): p. 181-7.

44. Zheng, B., Y.H. Chang, and D. Gur, Mass detection in digitized mammograms using two independent computer-assisted diagnosis schemes. AJR. American Journal of Roentgenology, 1996. **167**(6): p. 1421-4.

45. Zheng, B., Y.H. Chang, and D. Gur, Adaptive computer-aided diagnosis scheme of digitized mammograms. Academic Radiology, 1996. **3**(10): p. 806-14.

46. Zheng, B., *et al.*, Performance gain computer-assisted detection schemes by averaging scores generated from artificial neural networks with adaptive filtering. Medical Physics, 2001. **28**(11): p. 2302-2308.

47. Huo, Z., *et al.*, Automated computerized classification of malignant and benign masses on digitized mammograms. Academic Radiology, 1998. **5**(3): p. 155-68.

48. Huo, Z., M.L. Giger, and C.E. Metz, Effect of dominant features on neural network performance in the classification of mammographic lesions. Physics in Medicine & Biology, 1999. **44**(10): p. 2579-95.

49. Paquerault, S., *et al.*, Improvement of computerized mass detection on mammograms: Fusion of two-view information. Medical Physics, 2002. **29**(2): p. 238-247.

50. Jiang, Y., R.M. Nishikawa, and J. Papaioannou, Dependence of computer classification of clustered microcalcifications on the correct detection of microcalcifications. Medical Physics, 2001. **28**(9): p. 1949-1957.

51. Horsch, K., *et al.*, Computerized diagnosis of breast lesions on ultrasound. Medical Physics, 2002. **29**(2): p. 157-164.

52. Metz, C.E., Basic principles of ROC analysis. Sem Nuc Med, 1978. **8**: p. 283-298.

53. Metz, C.E., ROC methodology in radiologic imaging. Investigative Radiology, 1986. **21**: p. 720-733.

54. McClish, D.K., Analyzing a portion of the ROC curve. Medical Decision Making, 1989. **9**: p. 190-195.

55. Dwyer, A.J., In pursuit of a piece of the ROC. Radiology, 1996. **201**: p. 621-625.

56. Jiang, Y., C.E. Metz, and R.M. Nishikawa, A receiver operating characteristic partial area index for highly sensitive diagnostic tests. Radiology, 1996. **201**: p. 745-750.

57. Efron, B. and R.J. Tibshirani, An Introduction to the Bootstrap. Monographs on Statistics and Applied Probability, ed. D.R. Cox, *et al.* 1993, New York, NY: Chapman & Hall.

58. Sickles, E.A., Probably benign breast lesions: when should follow-up be recommended and what is the optimal follow-up protocol. Radiology, 1999. **213**: p. 11-14.

59. Rubin, E., Six-month follow-up: an alternative view. Radiology, 1999. **213**: p. 15-18.

60. Sickles, E.A., Commentary on Dr Rubin's Viewpoint. radiology, 1999. **213**: p. 19-20.

61. Rubin, E., Commentary on Dr Sickles's viewpoint. Radiology, 1999. **213**: p. 21.

62. Kupinski, M.A. and M.A. Anastasio, Multiobjective genetic optimization of diagnostic classifiers with implications for generating receiver operating characteristic curves. IEEE Transactions on Medical Imaging, 1999. **18**(8): p. 675-685.

63. Schwarzer, G., W. Vach, and M. Schumacher, On the misuses of artificial neural networks for prognostic and diagnostic classification in oncology. Statistics in Medicine, 2000. **19**: p. 541-561.

64. Sargent, D.J., Comparison of artificial neural networks with other statistical approaches - Results from medical data sets. Cancer, 2001. **91**(8): p. 1636-1642.

65. Giger, M.L., *et al.*, Computerized analysis of lesions in US images of the breast. Academic Radiology, 1999. **6**(11): p. 665-74.

66. Sahiner, B., *et al.*, Computerized characterization of masses on mammograms: the rubber band straightening transform and texture analysis. Medical Physics, 1998. **25**(4): p. 516-26.

67. McNitt-Gray, M.F., *et al.*, A pattern classification approach to characterizing solitary pulmonary nodules imaged on high resolution CT: preliminary results. Med Phys, 1999. **26**(6): p. 880-8.

68. Knutzen, A.M. and J.J. Gisvold, Likelihood of malignant disease for various categories of mammographically detected, nonpalpable breast lesions. Mayo Clinic Proceedings, 1993. **68**: p. 454-460.

69. Lo, J.Y., *et al.* 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.

70. Tourassi, G.D. and C.E. Floyd, The effect of data sampling on the performance evaluation of artificial neural networks in medical diagnosis [see comments]. Medical Decision Making, 1997. **17**(2): p. 186-92.

71. Lo, J.Y., *et al.*, Cross-institutional evaluation of BI-RADS predictive model for mammographic diagnosis of breast cancer. AJR. American Journal of Roentgenology, 2002. **178**: p. 457-463.

72. Heath, M., K.W. Bowyer, and D. Kopans, Current status of the Digital Database for Screening Mammography, in Digital Mammography, N. Karssemeijer, M. Thijssen, and J. Hendriks, Editors. 1998, Kluwer Academic Publishers. p. 457-460.

73. Mitchell, T.M., Machine Learning. 1997: WCB/McGraw-Hill.

74. Duda, R.O., P.E. Hart, and D.G. Stark, Pattern Classification. Second ed. 2001, New York: John Wiley & Sons.

75. Russell, S. and P. Norvig, Artificial Intelligence: A Modern Approach. Prentice Hall Series in Artificial Intelligence, ed. S. Russell and P. Norvig. 1995, Upper Saddle River: Prentice-Hall, Inc.

76. Karssemeijer, N. and J.H. Hendriks, Computer-assisted reading of mammograms. European Radiology, 1997. **7**(5): p. 743-8.

77. Castellino, R.A., J. Roehrig, and W. Zhang, Improved Computer-aided Detection (CAD) Algorithms for Screening Mammography. Radiology, 2000. **217(P)**: p. 400.

78. Chan, H.P., *et al.*, Improvement of radiologists' characterization of mammographic masses by using computer-aided diagnosis: an ROC study. Radiology, 1999. **212**(3): p. 817-27.

79. Jiang, Y., *et al.*, Improving breast cancer diagnosis with computer-aided diagnosis. Academic Radiology, 1999. **6**(1): p. 22-33.

80. Kahn, C.E., Jr., *et al.*, Construction of a Bayesian network for mammographic diagnosis of breast cancer. Computers in Biology & Medicine, 1997. **27**(1): p. 19-29.

81. Schaller, H.N., Constraint Satisfaction Problems, in Optimization Techniques, C.T. Leondes, Editor. 1998, Academic Press: San Diego, CA. p. 209-248.

82. Everitt, B.S., Cluster Analysis. 3rd ed. 1993: Arnold.

83. Gordon, A.D., Classification. 2nd ed. Monographs on Statistics and Applied Probability, ed. D.R. Cox, *et al.* 1999: Chapman & Hall/CRC.

84. Sharma, S., Applied Multivariate Techniques. 1996: John Wiley & Sons, Inc.

85. Hartigan, J.A., Clustering Algorithms. Wiley Series in Probability and Mathematical Statistics, ed. R.A. Bradley, *et al.* 1975, New York: John Wiley & Sons.

86. Hartigan, J.A. and M.A. Wong, A k-means clustering algorithm. Applied Statistics, 1979. **28**(1): p. 100-108.

87. Markey, M.K., *et al.* Cluster analysis of BI-RADS descriptions of biopsy-proven breast lesions. Medical Imaging 2002: Image Processing, Proceedings of the SPIE 4684:363-370 (2002). San Diego, CA.

88. Kohonen, T., Self-Organizing Maps. Springer Series in Information Sciences, ed. T.S. Huang, T. Kohonen, and M.R. Schroeder. 1995: Springer-Verlag.

89. Chen, D., R.F. Chang, and Y.L. Huang, Breast cancer diagnosis using self-organizing map for sonography. Ultrasound in Medicine & Biology, 2000. **26**(3): p. 405-11.

90. Cheeseman, P., *et al.* Bayesian Classification. in National Conference on Artificial Intelligence (AAAI-88). 1988. St. Paul, MN: Morgan Kaufmann Publishers.

91. Cheeseman, P. and J. Stutz, Bayesian Classification (AutoClass): Theory and Results, in Advances in Knowledge Discovery and Data Mining, U.M. Fayad, *et al.*, Editors. 1996, AAAI Press/MIT Press.

92. Markey, M.K., J.Y. Lo, and C.E. Floyd, Jr, Differences between computer-aided diagnosis of breast masses and that of calcifications. Radiology, 2002. **223**(2): p. 489-493.

93. Bilska-Wolak, A.O. and C.E. Floyd, Jr. Breast biopsy prediction using a case-based reasoning classifier for masses versus calcifications. Medical Imaging 2002: Image Processing, Proceedings of the SPIE 4684:661-665 (2002) San Diego, CA.

94. Cristianini, N. and J. Shawe-Taylor, An introduction to support vector machines: and other kernel-based learning methods. 2000, Cambridge, United Kingdom: Cambridge University Press.

95. Akanda, A., W.H. Land, and J.Y. Lo, Application of support vector machines to breast cancer classification using mammogram and history data, in Smart Engineering System Design: Neural Networks, Fuzzy Logic, Evolutionary Programming, Data Mining, and Complex Systems, C.H. Dagli, *et al.*, Editors. 2001, ASME Press: New York, NY. p. 839-846.

96. Land, W.H., Jr, *et al.* Application of support vector machine machines to breast a cancer screening using mammogram and history data. Medical Imaging 2002: Image Processing, Proceedings of the SPIE 4684:636-642 (2002). San Diego, CA.

97. Rumelhart, D.E. and J.L. McClelland, eds. Parallel Distributed Processing: Explorations in the Microstructures of Cognition. 1986, The MIT Press: Cambridge, Massachusetts.

98. Bishop, C.M., Neural Networks for Pattern Recognition. 1995: Oxford University Press.

99. Hertz, J., K. Anders, and R.G. Palmer, Introduction to the Theory of Computation. Santa Fe Institute Studies in the Science of Complexity. 1991: Addison-Wesley.

100. Huo, Z., *et al.*, Computerized classification of benign and malignant masses on digitized mammograms: a study of robustness. Academic Radiology, 2000. **7**(12): p. 1077-1084.

101. Chan, H.P., *et al.*, Classifier design for computer-aided diagnosis: effects of finite sample size on the mean performance of classical and neural network classifiers. Med Phys, 1999. **26**(12): p. 2654-68.

102. Kolodner, J., Case-Based Reasoning. 1993, San Mateo: Morgan Kaufmann Publishers, Inc.

103. Breiman, L., *et al.*, Classification and Regression Trees. The Wadsworth Statistics/Probability Series, ed. P. Bickel, W. Cleveland, and R. Dudley. 1984, Belmont: Wadsworth International Group.

104. Chambers, J.M. and T.J. Hastie, eds. Statistical Models in S. 1992, Wadsworth & Books/Cole Advanced Books & Software: Pacific Grove, California.

105. Kegelmeyer, W.P., *et al.*, Computer-aided mammographic screening for spiculated lesions. Radiology, 1994. **191**(2): p. 331-337.

106. Kuo, W.J., *et al.*, Datamining with decision trees for diagnosis of breast tumor in medical ultrasound images. Breast Cancer Research and Treatment, 2001. **66**: p. 51-57.

107. Markey, M.K., J.Y. Lo, and C.E. Floyd, Jr. Differences in computer aided diagnosis of breast cancer: masses vs. calcifications. in Chicago 2000: World Congress on Medical Physics and Biomedical Engineering. 2000. Chicago, IL.

108. Nagel, R.H., *et al.*, Analysis of methods for reducing false positives in the automated detection of clustered microcalcifications in mammograms. Med Phys, 1998. **25**(8): p. 1502-6.

109. Qian, W., *et al.*, Digital mammography: comparison of adaptive and nonadaptive CAD methods for mass detection. Academic Radiology, 1999. **6**(8): p. 471-80.

110. Land, W.H., T. Masters, and J.Y. Lo. Application of a new evolutionary programming / adaptive boosting hybrid to breast cancer diagnosis. in IEEE Congress on Evolutionary Computation Proceedings. 2000.

111. Land, W.H., *et al.* New Results in Breast Cancer Classification Obtained from an Evolutionary Computation/Adaptive Boosting Hybrid Using Mammogram and History Data. in IEEE Mountain Workshop on Soft Computing in Industrial Applications. 2001. Virginia Tech, Blacksburg, Virginia.

112. Land, W.H., *et al.* Application of adaptive boosting to EP-derived multi-layer feedforward neural networks (MLFN) to improve benign/malignant breast cancer classification. in SPIE Medical Imaging 2001: Image Processing. 2001.

Publications

Refereed Journals

[1] M. V. Boland, **M. K. Markey**, R. F. Murphy, "Automated recognition of patterns characteristic of subcellular structures in fluorescence microscopy images", Cytometry 33:366-375 (1998).

[2] **M. K. Markey**, M. V. Boland, R. F. Murphy, "Toward objective selection of representative microscopy images", Biophysical Journal 76:2230-2237 (1999).

[3] G. D. Tourassi, **M. K. Markey**, J. Y. Lo, C. E. Floyd, Jr., "A neural network approach to breast cancer diagnosis as a constraint satisfaction problem", Medical Physics 28:804-811 (2001).

[4] G. D. Tourassi, E. D. Frederick, **M. K. Markey**, C. E. Floyd, Jr., "Application of the mutual information criterion for feature selection in computer-aided diagnosis", Medical Physics 28:2394-2402 (2001).

[5] J. Y. Lo, **M. K. Markey**, J. A. Baker, C. E. Floyd, Jr., "Cross-institutional evaluation of BI-RADS predictive model for mammographic diagnosis of breast cancer", American Journal of Roentgenology 178:457-463 (2002).

[6] **M. K. Markey**, J. Y. Lo, C. E. Floyd, Jr., "Differences between the computer-aided diagnosis of breast masses and that of calcifications", Radiology 223:489-493 (2002).

[7] **M. K. Markey**, J. Y. Lo, R. Vargas-Voracek, G. D. Tourassi, C. E. Floyd, Jr., "Perceptron error surface analysis: a case study in breast cancer diagnosis", Computers in Biology and Medicine 32:99-109 (2002).

Non-refereed Publications

[1] **M. K. Markey**, V. T. Tang, M. LaBarbera, "Swimming kinematics of *Argopecten irradians*", National Conference on Undergraduate Research, Western Michigan University, (1994). (abstract)

[2] M. V. Boland, **M. K. Markey**, R. F. Murphy, "Automated classification of protein localization patterns", 36th American Society for Cell Biology Annual Meeting, Molecular Biology of the Cell 7: 908-908 Suppl. (1996). (abstract)

[3] M. V. Boland, **M. K. Markey**, R. F. Murphy, "Classification of protein localization patterns obtained via fluorescence light microscopy", Proceedings of the 19th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (1997), pg. 594-597.

[4] **M. K. Markey**, M. V. Boland, R. F. Murphy, "Towards objective selection of representative microscopy images", 37th American Society for Cell Biology Annual Meeting, Molecular Biology of the Cell 8: 2012-2012 Suppl. (1997). (abstract)

[5] **M. K. Markey**, J. Y. Lo, C. E. Floyd, Jr., "Differences in computer aided diagnosis of breast cancer: masses vs. calcifications", World Congress on Medical Physics and Biomedical Engineering (2000). (abstract)

[6] G. D. Tourassi, E. D. Frederick, **M. K. Markey**, C. E. Floyd, Jr., "Application of an information theoretic approach for feature selection in the computer-aided diagnosis of acute pulmonary embolism", Radiological Society of North America Annual Meeting, Radiology 221:547-547 Suppl. (2001). (abstract)

[7] **M. K. Markey**, J. Y. Lo, G. D. Tourassi, C. E. Floyd, Jr., "Cluster analysis of BI-RADS descriptions of biopsy-proven breast lesions", Medical Imaging 2002: Image Processing, Proceedings of the SPIE 4684:363-370 (2002).