A Study on Several Machine-Learning Methods for Classification of Malignant and Benign Clustered Microcalcifications

Article *in* IEEE Transactions on Medical Imaging · April 2005

DOI: 10.1109/TMI.2004.842457 · Source: PubMed

Robert M. Nishikawa
University of Pittsburgh

A Study on Several Machine-Learning Methods for Classification of Malignant and Benign Clustered Microcalcifications

Liyang Wei, *Student Member, IEEE*, Yongyi Yang*, *Senior Member, IEEE*, Robert M. Nishikawa, and Yulei Jiang

***Abstract*—In this paper, we investigate several state-of-the-art machine-learning methods for automated classification of clustered microcalcifications (MCs). The classifier is part of a computer-aided diagnosis (CADx) scheme that is aimed to assisting radiologists in making more accurate diagnoses of breast cancer on mammograms. The methods we considered were: support vector machine (SVM), kernel Fisher discriminant (KFD), relevance vector machine (RVM), and committee machines (ensemble averaging and AdaBoost), of which most have been developed recently in statistical learning theory. We formulated differentiation of malignant from benign MCs as a supervised learning problem, and applied these learning methods to develop the classification algorithm. As input, these methods used image features automatically extracted from clustered MCs. We tested these methods using a database of 697 clinical mammograms from 386 cases, which included a wide spectrum of difficult-to-classify cases. We analyzed the distribution of the cases in this database using the multidimensional scaling technique, which reveals that in the feature space the malignant cases are not trivially separable from the benign ones. We used receiver operating characteristic (ROC) analysis to evaluate and to compare classification performance by the different methods. In addition, we also investigated how to combine information from multiple-view mammograms of the same case so that the best decision can be made by a classifier. In our experiments, the kernel-based methods (i.e., SVM, KFD, and RVM) yielded the best performance ($A_z = 0.85$, SVM), significantly outperforming a well-established, clinically-proven CADx approach that is based on neural network ($A_z = 0.80$).**

***Index Terms*—Clustered microcalcifications, computer-aided diagnosis, kernel methods, mammography, relevance vector machine, support vector machine.**

I. Introduction

IN this paper, we investigate the use of several state-of-the-art machine-learning methods for differentiating malignant from benign clustered microcalcifications (MCs) in digital mammograms. Clustered MCs can be an important early indicator of breast cancer in women. They appear in 30%–50% of mammographically diagnosed cases. For example, Fig. 1 shows a mammogram with a cluster of MCs. Though commonly seen on mammograms, MCs are often difficult to diagnose accuately. This greatly compromises the quality of radiologists' biopsy recommendations, which is an important issue in breast cancer diagnosis. It is reported that among those with radiographically suspicious, nonpalpable lesions who are sent for biopsy, only 15%–34% are found to actually have malignancies [1], [2].

There has been a great deal of research in recent years to develop computerized methods that potentially could assist radiologists in differentiating benign from malignant MCs. Using a computer-aided diagnosis (CADx) scheme, radiologists could incorporate the output from the computer into their decision. Indeed, several studies have demonstrated that CADx can help improve radiologists' ability in differentiating malignant breast lesions from benign ones [3]–[10]. In particular, Jiang *et al.* [4] developed an automated computer scheme that was demonstrated to classify clustered MCs more accurately than radiologists. This scheme made use of an artificial neural network, which was trained to predict the likelihood of malignancy based on quantitative image features automatically extracted from the clustered MCs. It was subsequently demonstrated in [5] that when used as a diagnostic aid, this scheme could also lead to significant improvement in radiologists' performance in distinguishing between malignant and benign clustered MCs.

The following is a brief review of other representative methods for classification of clustered MCs. Wu *et al.* [6] and Lo *et al.* [11] have independently develop CADx schemes that employ artificial neural networks where the image features were used were extracted by the radiologist. Linear discriminant analysis was used by Chan *et al.* in [12] for classification of MCs. In [13], Dhawan developed a radial basis function neural network for classifying hard-to-diagnose cases of MCs. Neural networks were also used in [14] for classifying MCs. Bayes classification (linear and quadratic) and K-nearest-neighbor method were used in [15] and [16]. In [17], Zaiane *et al.* proposed an approach based on association rule mining.

Built on the success of the work by Jiang *et al.* [4], [5], we have implemented several promising machine-learning tools developed most recently in statistical learning theory, and have investigated whether these state-of-the-art learning methods could further lead to more accurate classification of malignant from benign clustered MCs. The methods we consider were: support vector machine (SVM) [18], kernel Fisher discriminant (KFD) [19], relevance vector machine (RVM) [20], and committee ma-

Manuscript received July 1, 2004; revised November 22, 2004. This work was supported in part by the National Institutes of Health (NIH)/National Cancer Institute (NCI) under Grant CA89668. The Associate Editor responsible for coordinating the review of this paper and recommending its publication was N. Karssemeijer. *Asterisk indicates corresponding author.*

L. Wei is with the Department of Biomedical Engineering, Illinois Institute of Technology, Chicago, IL 60616 USA.

*Y. Yang is with the Department of Electrical and Computer Engineering, Illinois Institute of Technology, 3301 South Dearborn Street, Chicago, IL 60616 USA (e-mail: yy@ece.iit.edu).

R. M. Nishikawa is with the Department of Radiology, The University of Chicago, Chicago, IL 60637 USA. He is also a shareholder in R2 Technology, Inc. (Sunnyvale, CA).

Y. Jiang is with the Department of Radiology, The University of Chicago, 5841 South Maryland Avenue, Chicago, IL 60637 USA.

Digital Object Identifier 10.1109/TMI.2004.842457

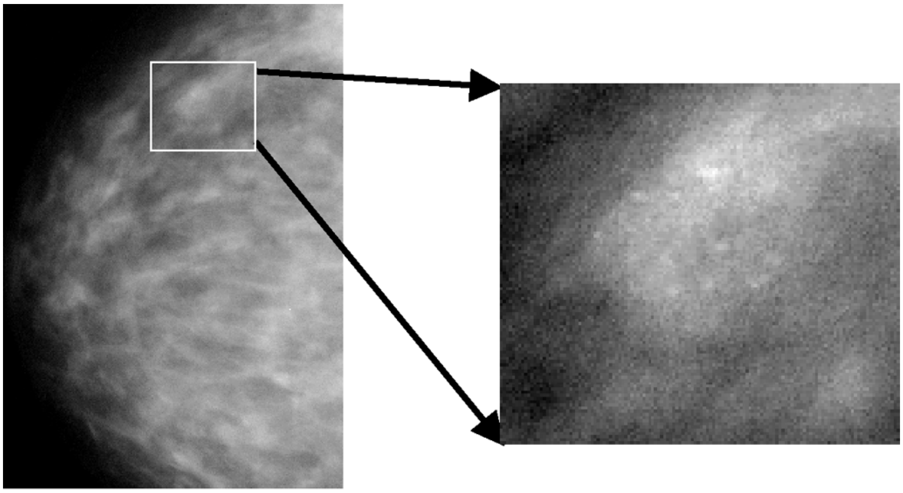


Fig. 1. *Left*: a CC view mammogram; *right*: expanded view showing clustered MCs. MCs are small granule-like deposits of calcium, and appear as bright spots in a mammogram.

chines (including ensemble averaging [21] and AdaBoost [22]). To our best knowledge, most of these methods are still relatively new to applications in medical imaging, even though they have been actively studied in the machine learning community in recent years.

In addition to evaluating classifiers, we have investigated different methods for combining the results from multiple views of the same cases, since MCs often appear in both of the two standard-view mammograms: mediolateral oblique (ML) and craniocaudal (CC) views.

The rest of the paper is organized as follows: An overview of the different machine-learning methods considered in this study is provided in Section II. A description of the mammogram data set is furnished in Section III. An evaluation study of the learning methods using the mammogram data set is described in Section IV, and the experiment results are presented and discussed in Section V. Finally, conclusions are drawn in Section VI.

II. Description of the Machine-Learning Methods Considered

In this paper, classification of malignant from benign clustered MCs is treated as a two-class pattern classification problem, i.e., a microcalcification cluster (MCC) under consideration is either malignant or benign. To begin, let vector $\mathbf{x} \in R^n$ denote a pattern to be classified, and let scalar d denote its class label (i.e.,$d \in \{\pm 1\}$). In addition, let $\{(\mathbf{x}_i, d_i),\ i = 1, 2, \cdots, N\}$ denote a given set of N training examples, where each sample $\mathbf{x}_i$ has a known class label d_i. The problem is how to determine a classifier $f(\mathbf{x})$ (i.e., a decision function) that can correctly classify an input pattern (not necessarily from the training set).

Below we provide a brief introduction to the different classifier models that are considered in this paper.

A. Support Vector Machines (SVM)

SVM is a constructive learning procedure rooted in statistical learning theory [18]. It is based on the principle of structural risk minimization, which aims at minimizing the bound on the generalization error (i.e., error made by the learning machine on data unseen during training) rather than minimizing the mean square error over the data set [18]. As a result, an SVM tends to perform well when applied to data outside the training set.

Indeed, SVM has been found to outperform competing methods in several real-world applications (see, for example, [23]–[27]). In our own work [28], we developed an SVM based approach for detection of clustered MCs in mammograms, and demonstrated using clinical mammogram data that such an approach could outperform several well-known methods in the literature.

For classification, a (nonlinear) SVM classifier in concept first maps the input data vector $\mathbf{x}$ into a higher dimensional space $\mathbf{H}$ through an underlying nonlinear mapping $\Phi(\mathbf{x})$, then applies linear classification in this mapped space. That is, an SVM classification function can be written in the following form:

$$f_{\mathrm{SVM}}(\mathbf{x}) = \mathbf{w}^T \mathbf{\Phi}(\mathbf{x}) + b \tag{1}$$

where parameters $\mathbf{w}$, b are determined from the training data samples. This is accomplished through minimization of the following so-called *structural risk* function:

$$J(\mathbf{w}, \boldsymbol{\xi}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N}\xi_i,$$
$$\text{subject to } d_i f_{\mathrm{SVM}}(\mathbf{x}_i) \geq 1 - \xi_i,\ \xi_i \geq 0;\ i = 1, 2, \ldots, N \tag{2}$$

where C is a user-specified, positive parameter, ξ_i are slack variables. In particular, when the two classes are separable, minimizing the structural risk in (2) amounts to maximizing the separating margin between the two classes.

The cost function in (2) constitutes a balance between the *empirical risk* (i.e., the training errors reflected by the second term) and model complexity (the first term) [18]. The parameter C controls this trade-off. The purpose of using model complexity to constrain the optimization of empirical risk is to avoid *overfitting*, a situation in which the decision boundary too precisely corresponds to the training data, and thereby fails to perform well on data outside the training set.

A training sample $(\mathbf{x}_i, d_i)$ is called a *support vector* when $d_i f_{\mathrm{SVM}}(\mathbf{x}_i) \leq 1$. Introducing a so-called *kernel function* $K(\mathbf{x},\mathbf{y}) \equiv \Phi(\mathbf{x})^T\Phi(\mathbf{y})$, we can rewrite the SVM function $f_{\mathrm{SVM}}(\mathbf{x})$ in (1) in a kernel form as follows:

$$f_{\mathrm{SVM}}(\mathbf{x}) = \sum_{i=1}^{N_s} \alpha_i K(\mathbf{x}, \mathbf{s}_i) + b \tag{3}$$

where $\mathbf{s}_i$, $i = 1, 2, \cdots, N_s$, denote the *support vectors*. In general, support vectors constitute only a small fraction of the training samples $\{\mathbf{x}_i, i = 1, 2, \cdots, N\}$.

From (3), we can directly evaluate the decision function through the kernel function $K(\cdot,\cdot)$ without the need to specifically addressing the underlying mapping $\Phi(\cdot)$. In this paper, we consider two kernel types: polynomial kernels and Gaussian radial basis functions (RBFs). These are among the most commonly used kernels in SVM research, and are known to satisfy Mercer's condition [30]. They are defined as follows.

1) Polynomial kernel:

$$K(\mathbf{x},\mathbf{y}) = (\mathbf{x}^T\mathbf{y} + 1)^p \tag{4}$$

where $p > 0$ is a constant that defines as the kernel order.

2) RBF kernel:

$$K(\mathbf{x},\mathbf{y}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}\right) \tag{5}$$

where $\sigma > 0$ is a constant that defines the kernel width.

B. Kernel Fisher Discriminant (KFD)

Fisher's (linear) discriminant is a powerful technique for pattern classification in statistical data analysis. It is based on the principle of projecting the data onto a one-dimensional space so that the two classes are well separated (by maximizing the so-called Rayleigh coefficient) [31], [32]. KFD is an extension of Fisher's linear discriminant to the kernel space. Like SVM, the KFD first maps the input data vector $\mathbf{x}$ into a higher dimensional space [through an underlying nonlinear mapping $\Phi(\mathbf{x})$], but applies Fisher's linear classification in this mapped space [19], [33]. This effectively yields a nonlinear discriminant with respect to the original vector $\mathbf{x}$.

As in SVM, let $K(\cdot,\cdot)$ denote the kernel function corresponding to the underlying nonlinear mapping $\Phi(\mathbf{x})$. We can write the KFD function $f_{\mathrm{KFD}}(\mathbf{x})$ as follows:

$$f_{\mathrm{KFD}}(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i K(\mathbf{x}, \mathbf{x}_i) + b. \tag{6}$$

The coefficients α_i, $i = 1, 2, \cdots, N$, are determined by maximizing the Rayleigh coefficient corresponding to $f_{\mathrm{KFD}}(\mathbf{x})$ [33].

The kernel function $K(\cdot,\cdot)$ in (6) must satisfy the Mercer conditions. As for SVM, we consider polynomial kernels and RBF kernels in this paper.

While the KFD decision function $f_{\mathrm{KFD}}(\mathbf{x})$ in (6) has a similar form to that of the SVM $f_{\mathrm{SVM}}(\mathbf{x})$ in (3), there is also a distinctive difference between the two. In $f_{\mathrm{KFD}}(\mathbf{x})$, all the training samples are used for the decision, whereas in $f_{\mathrm{SVM}}(\mathbf{x})$ only a subset of the training samples (namely the support vectors) are used. It is reported that the KFD can perform just as well as the SVM [33]. This is also observed in this paper (Section V).

C. Relevance Vector Machines (RVM)

RVM is a statistical learning technique developed recently by Tipping [20] based on Bayesian estimation for regression (and classification) problems. Its key feature is that it can yield a solution function that depends on only a very small number of training samples (called *relevance vectors*). It is reported that in several benchmark studies RVM can yield nearly identical performance to, if not better than, that of SVM while using far fewer relevance vectors than the number of support vectors for SVM [20]. Compared to SVM, RVM does not need the tuning of a regularization parameter (C) during the training phase.

For an input vector $\mathbf{x}$, an RVM classifier models the distribution of its class label $d \in \{+1, -1\}$ using logistic regression as

$$p(d = 1|\mathbf{x}) = \frac{1}{1 + \exp\left(-f_{\mathrm{RVM}}(\mathbf{x})\right)} \tag{7}$$

where $f_{\mathrm{RVM}}(\mathbf{x})$ is given by

$$f_{\mathrm{RVM}}(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i K(\mathbf{x}, \mathbf{x}_i) \tag{8}$$

where $K(\cdot,\cdot)$ is a kernel function, and $\mathbf{x}_i$, $i = 1, 2, \cdots, N$, are the training samples.

According to Tipping [20], the parameters α_i, $i = 1, 2, \cdots, N$, in $f_{\mathrm{RVM}}(\mathbf{x})$ are determined using Bayesian estimation. Toward this end, a sparse prior is introduced on α_i. Specifically, these parameters are assumed to be statistically independent and each obeys a zero-mean, Gaussian distribution with variance λ_i^{-1}; furthermore, a so-called *hyper-prior* (based on the Gamma distribution) is assumed on the variance λ_i^{-1}, which is used to force the parameters α_i to be highly concentrated around 0, thereby leading to very few nonzero terms in $f_{\mathrm{RVM}}(\mathbf{x})$.

The parameters α_i in (8) are then obtained by maximizing the posterior distribution of the class labels given the input vectors. This is equivalent to maximizing the following objective function:

$$J(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \log p(d_i|\mathbf{x}_i) + \sum_{i=1}^{N} \log p\left(\alpha_i|\lambda_i^*\right) \tag{9}$$

where the first summation term corresponds to the likelihood of the class labels, and the second term corresponds to the prior on the parameters α_i, in which λ_i^* denotes the maximum a posteriori estimate of the hyper-parameter λ_i. In the resulting solution, only those samples associated with nonzero coefficients α_i, called relevance vectors, will contribute to the decision function $f_{\mathrm{RVM}}(\mathbf{x})$.

In (8), the kernel function $K(\cdot,\cdot)$ is used to form expansion basis functions for $f_{\mathrm{RVM}}(\mathbf{x})$, and in theory, is not limited by the Mercer conditions as in the case of SVM or KFD. For simplicity, the polynomial kernels and the RBF kernels are used in this paper.

D. Artificial Neural Networks (ANN)

ANN is a powerful learning technique that has been applied successfully to a variety of fields. As pointed out in the introduction section, Jiang *et al.* [4], [5] designed a feedforward neural network (FFNN) for automated classification of MCCs. For comparison purposes, we also consider this network in this paper.

The FFNN used in [4], [5] is a three-layer network, trained by the well-known backpropagation algorithm. It has eight input neurons (corresponding to eight features, which are given in Section III), six neurons in the hidden layer, and one output neuron. In this paper, the number of neurons in the hidden layer is varied during the training phase for optimal performance.

E. Committee Machines

Committee machines (also called *committee methods*) operate on the principle that by combining the output from a group of learning machines (i.e., classifiers in our case), one can achieve a decision function that is superior in performance to any of the individual ones. In this paper, we consider committee machines formed by the following two popular strategies: 1) *ensemble averaging* [21]; and 2) *boosting* [21], [22].

1) Ensemble Averaging: The ensemble averaging approach simply combines the output of a collection of classifiers to produce an averaged output. In this paper, we use the FFNN described above as the basis classifiers, and apply ensemble averaging to combine them. The basis classifiers are obtained by training the FFNN with many different random starting points. It is known in neural network training that the final solution is dependent on the choice of initial starting points, owing to the fact that the underlying optimization problem is nonconvex and has multiple (local) minima. By ensemble averaging, we can effectively reduce the variance in the final solution by removing its dependency on the choice of initial starting points [30]. Consequently, in our experiments (Section V) for FFNN we will report results obtained with a committee of 100 trained networks rather than that from a single network.

2) AdaBoost: Boosting is a method that strategically combines a collection of "weak" classifiers (whose individual performance is only slightly better than random guessing) to form a stronger classifier. In particular, we consider the AdaBoost [22] method, which is perhaps the most popular and has been applied with great success in machine learning problems [34], [35]. Its basic idea is as follows: Given a training data set, we repeatedly apply a chosen learning algorithm to obtain a sequence of (weak) classifiers, denoted by $f_m(\mathbf{x})$, $m = 0, 1, \cdots, M-1$, by successively modifying this training set at each step. This is accomplished by associating a weight factor with each training sample (all samples are equally weighted at $m = 0$) and systematically adjusting this weight factor such that a misclassified sample by $f_m(\mathbf{x})$ will receive an increased weight value in the training of subsequent $f_{m+1}(\mathbf{x})$. This will allow those difficult-to-classify training samples to have ever-increasing influence, and consequently, a subsequent classifier will focus more on the samples that have been misclassified by its predecessors. In the end, we combine all the obtained classifiers using a weighted average, i.e.,

$$f_{\mathrm{AdaBoost}}(\mathbf{x}) = \sum_{m=0}^{M-1} w_m f_m(\mathbf{x}) \tag{10}$$

where the weighting factors w_m are determined such that those more accurate classifiers in the sequence will have more influence on the final decision function.

In this paper, we use a single-neuron FFNN as the base learning algorithm, which amounts to a simple linear classifier when used separately. We apply AdaBoost to construct a more powerful classifier from a collection of such simple classifiers.

III. Description of Mammogram Data Set

A. Mammogram Data Set

The different classifier models were developed and tested using a data set collected by the Department of Radiology at the University of Chicago. This data set consisted of 697 mammograms from 386 clinical cases, of which all had lesions containing clustered MCs which were histologically proven. Among them 75 were malignant, and the rest (311) were benign. Furthermore, most of these cases have two standard-view mammograms: ML and CC views. The clustered MCs were identified by a group of experienced researchers. For computer analysis, all the mammograms in the data set were digitized with a spatial resolution of 0.1 mm/pixel and 10-bit grayscale.

It is noted that this data set includes a subset consisting of 53 cases (19 malignant and 34 benign, all collected between 1985 and 1988) which was used in the work of Jiang *et al.* on development of CADx classifiers using FFNN [4], and another subset consisting of 104 cases (46 malignant and 58 benign, 90% of them collected between 1990 and 1996) which was used in their subsequent work [5]. The rest of the cases were collected from 1994 to 1995 and included cases that contained MCCs but were judged clinically benign. All these cases were considered benign based on 2-year follow up.

The data set includes a wide spectrum of cases that are judged to be difficult to classify by radiologists. For example, among the subset of 53 cases used in [4], 51% of them were difficult to classify as retrospectively reviewed by a radiologist [4]. Also, among the subset of 104 cases used in [5], many of them are extremely difficult to classify [5]; the average classification performance by a group of five attending radiologists on these cases yielded a value of only 0.62 in the area under the receiver operating characteristic (ROC) curve [5].

B. Feature Extraction and Processing

For automated classification, Jiang *et al.* [4], [5] used the following eight features, all computed from the mammogram image, to characterize an MCC: 1) the number of MCs in the cluster; 2) the mean effective volume (area times effective thickness) of individual MCs; 3) the area of the cluster; 4) the circularity of the cluster; 5) the relative standard deviation of the effective thickness; 6) the relative standard deviation of the effective volume; 7) the mean area of MCs; and 8) the second highest MC-shape-irregularity measure. The numerical values of all these features were normalized to be within the range between 0 and 1.

A detailed description of these features can be found in [4]. These features were selected to have intuitive meanings that correlate qualitatively to features used by radiologists [4]. This provides an important common ground for the computer scheme to achieve high classification performance and for radiologists to interpret the computer results. Specifically, "benign clusters tend to be smaller and more round, whereas malignant clusters tend to be larger and irregular. These are the typical characteristics of the spatial distributions of benign MCs associated with adenosis and malignant ductal MCs" [4], which can be described by features 3 and 4; "Benign clusters tend to have fewer and smaller MCs compared with malignant clusters. These characteristics of benign clusters correspond to punctate and lobular calcifications which are often interpreted as benign" [4], which can be described by features 1 and 2; "Malignant MCs tend to have wider variation in size within a cluster, indicating the pleomorphic nature of some malignant MC clusters" [4], which can be described by features 5 and 6; and finally, "for a given size, malignant MCs tend to be more irregular than those that are benign. This characteristic corresponds to one of the most important clinical indications of malignancy: linear or branching MCs" [4], which can be described by features 7 and 8.

It is noted that most of these features are also closely related to the categorical descriptions for MCCs defined in BI-RADS [36]. The results from [4] and [5] demonstrate that these features indeed have good predictive value.

Consequently, we used these same eight features in this study, in order to facilitate the comparison of our new classifier models with the ANN scheme in [4] and [5].

For preparation of training and testing samples for the classifier models, the eight features are extracted for each MCC in the mammogram data set; the vector formed by these eight feature values, denoted by $\mathbf{x}_i$, is then treated as an input pattern, and is labeled as $y_i = +1$ for a malignant case, and $y_i = -1$ otherwise. Together, $(\mathbf{x}_i, y_i)$ forms an input-output pair. There are in total 697 such pairs obtained from the whole mammogram data set. These pairs are subsequently for training and testing of the classifier models.

C. Learning With Multiple Views

As pointed out above, most of the cases (292, to be exact) in the data set have multiple views. Consequently, each lesion yields as many input-output pairs $(\mathbf{x}_i, y_i)$ as the number of views available. These input-output pairs are inherently correlated, because they result from the same case. Therefore, these correlated input-output pairs should be grouped by case such that those from the same case will be used either for training or for testing, but not for both.

We investigate the following three different strategies for making use of information from multiple views: 1) direct method; 2) averaging method; and 3) joint method. In the direct method, the multiple input-output pairs are directly treated as separate samples, but grouped together either for training or for testing. In testing a classifier, a case is classified as malignant when any of its multiple views are classified as malignant. This method was used in the previous work of Jiang *et al.* [4], [5].

In the averaging method, the multiple input-output pairs of each case are averaged to produce a single sample. That is, the feature vectors from the multiple views are simply averaged before applied to a classifier.

Finally, in the joint method, we form an augmented 16-dimensional vector by stacking together the two feature vectors from the ML and CC views of each lesion. The resulting vector is then used as input to the classifier models. In the data set, there are 94 cases with only single views available; for each of these cases we simply replicated the available view to substitute for the missing one.

D. Distribution of the Cases

With all the cases represented as vectors in the input space, it is desirable to study how they are distributed in this space. For example, it would be interesting to inspect whether the malignant (or benign) cases are distributed in any particular structure. Unfortunately, this is not an easy task in the input space since it is eight-dimensional. Instead, we use a technique called *multidimensional scaling* (MDS) [29] which allows us to explore the data set in a lower dimensional space [two-dimensional (2-D) in our case].

MDS is a versatile technique for representation and analysis of a set of objects based on their mutual similarity (or proximity) measurements. The basic idea of MDS is to embed the objects of interest as points in a low-dimensional (typically 2–D or three-dimensional) space such that the geometric distances between the points in this space are in accordance with the similarity measurements between the corresponding objects. The resulting representation in this lower dimensional space enables one to visualize the relationship among the objects (which could be quite complex) in a rather intuitive manner. Because of this property, MDS has gained popularity in a variety of data analysis applications.

In Fig. 2(a), we show a 2-D MDS plot of the cases in the dataset, wherein the MCCs are represented as points (indicated by "+" for malignant cases, and "o" for benign ones) in the 2-D plane; the mutual distances between the points reflect the Euclidean distances between the feature vectors of the corresponding MCCs. For clarity, only 75 of the 311 benign cases (chosen randomly) are shown in the plot.

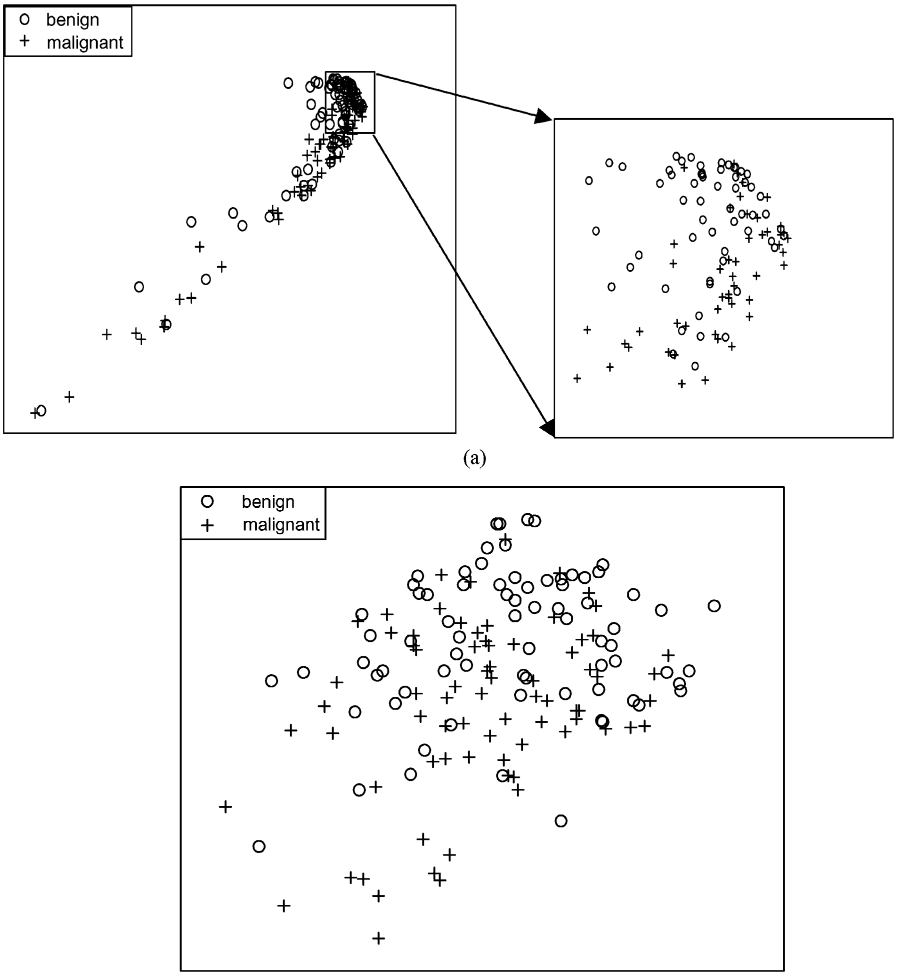
For comparison, we show in Fig. 2(b) a scatter-plot of the same cases as in Fig. 2(a) based on the first two principal components of the data through principal component analysis (PCA).

The MDS plot reveals that the MCCs are closely distributed within a comet-shaped region, with cases most densely distributed near the head region (enclosed by a box in the plot). This indicates that most of the cases are also closely distributed in the original input space. Most importantly, from the plot there is no clear pattern in how the malignant cases are distributed differently from the benign ones; rather, the malignant cases are closely neighbored by benign ones, and *vice versa*. This implies that MCCs with similar feature vectors can be quite different in pathology. This, of course, poses a significant challenge to the classifiers. We feel that this is also a manifestation of the fact that some of the cases in the data set are very difficult to classify.

IV. Performance Evaluation Study

A. Machine Training and Parameter Selection

Note that each classifier model described in Section II is typically associated with a few model parameters that need to be fine-tuned during training for best performance. Specifically, for the kernel methods, namely SVM, KFD, and RVM, we need to decide the type of kernel function to be used (i.e., polynomial versus RBF) and its associated parameter (i.e., the order p for the polynomial, or the kernel width σ for RBF); in addition, for SVM we also need to determine the regularization parameter C in (2). For FFNN, we need to decide the number of layers and the number of neurons for each hidden layer. For AdaBoost, we



(a)

(b)

Fig. 2. (a) MDS embedding of the dataset in the 2-D plane, and (b) PCA embedding of the dataset in the 2-D plane.

need to determine the best number of rounds for boosting [i.e., M in (10)].

To determine the fine-tuning parameters for each classifier model, we apply a *leave-one-out* cross validation procedure [37]. For M cases used, this procedure is as follows: for each parameter setting, train the classifier model M times; during each time one of the M cases is held out in turn while the remaining $M-1$ cases are used to train the classifier; the trained classifier is then used to classify the held-out case, and the classification result is recorded. In the end, the classification results for the M cases are used to obtain an estimate of the generalization error of the classifier model. The parameter setting with the smallest generalization error is chosen.

B. Training With Unbalanced Samples

In the mammogram data set we have an unbalanced distribution of cases between the malignant class and the benign class, with the benign cases outnumbering the malignant ones by a margin of more than 4:1. Aimed to improving the classification performance, several strategies exist for dealing with unbalanced data sets for classifier training, including resizing the training data sets, adjusting misclassification costs, and recognition-based learning (learning from the minority class) [38]. The approach of resizing training sets is through over-sampling the minority class and/or under-sampling the majority class. Here, we experiment with over-sampling the minority class by simply adding duplicates of the minority-class samples to the training set. While over-sampling does not introduce any new information in the training set, it implicitly increases the cost associated with misclassification of the minority samples. We will explore whether such an approach for adjusting the balance of the samples between the two classes in training can be beneficial for training of classifiers using the mammogram data set.

C. ROC Analysis

To evaluate the performance of a classifier, we use the so-called receiver operating characteristic (ROC) analysis, which is now used routinely for many classification tasks. An ROC curve is a plot of the classification sensitivity [i.e., true positive fraction (TPF)] as the ordinate versus the specificity [i.e., false positive fraction (FPF)] as the abscissa; for a given classifier, it is obtained by continuously varying the threshold associated with its decision function. Thus, at any given FPF, an ROC curve with a higher TPF corresponds to better classification performance. As a summary measure of overall diagnostic performance, the area under an ROC curve (denoted by A_z) is often used.

TABLE I
RESULTS OBTAINED USING DIFFERENT STRATEGIES FOR LEARNING WITH MULTIPLE VIEWS FOR SVM AND FFNN

	Direct method	Averaging method	Joint method
FFNN	0.7943 (0.0267)	0.8007 (0.0266)	0.7516 (0.0323)
SVM	0.8324 (0.0265)	0.8545 (0.0259)	0.7927 (0.0307)

We will perform evaluation experiments using the following two procedures: a *single-set* procedure, and a *multi-set* procedure. In the single-set procedure, we treat all the available cases in a single set and apply the leave-one-out procedure described above to train and test a classifier model. In the end, the test results for all cases are analyzed using the ROCKIT program of Dr. Metz [39].

In the multi-set procedure, we partition all the available cases in a random fashion into three equal-sized subsets, denoted by A, B, and C, respectively, with each containing 25 malignant cases. For each classifier model, we train it twice: one with A, and the other with B, each time with the following two-step procedure: 1) apply the leave-one-out cross validation procedure to determine the tuning parameters; and 2) retrain the classifier model with the tuned parameters using all the cases in the training set. This effectively yields two "readers" for the same classifier model, one from each training set. Both readers are then tested using C. This amounts to a multiple-reader, multiple-case design in medical imaging. In the end, the test results are analyzed using the LABMRMC program of Dr. Metz [40]. This program uses jackknifing and ANOVA techniques to test the statistical significance of the differences between the classifier models and between the training sets.

The single-set procedure has the advantage that all the cases will be used for both training and testing of the classifiers; its downside is that the testing results are also used in the fine-tuning of the classifier models, which could lead to an overly optimistic bias in the evaluation results. On the other hand, the multi-set procedure not only avoids such a bias by completely isolating the testing samples from model training and fine-tuning, but also allows us to generalize the comparisons to the training sets and test samples. Its downside, of course, is that each time only one third of the samples are used either for training or for testing. The smaller training dataset will negatively bias our measured performance, while the smaller testing dataset will increase the variance in measured performance [41], [42].

V. EXPERIMENTAL RESULTS AND DISCUSSIONS

A. Learning With Multiple Views

Using the single-set procedure, we first performed experiments to compare the three different strategies given in Section III-C for combining the information from the multiple views of each MCC, namely: 1) direct method; 2) averaging method; and 3) joint method. In Table I, we present the resulting classification performance, summarized by A_z (the area under the ROC curve) and its standard deviation (in parentheses), from each of the three different strategies; for brevity, here the results are shown only for SVM and FFNN, as they also exemplify those by the other learning methods (as will be clear from later results). Also, for FFNN the results in Table I were obtained by ensemble averaging over a committee of 100 trained networks. This is also true for the rest of the section when results of FFNN are given.

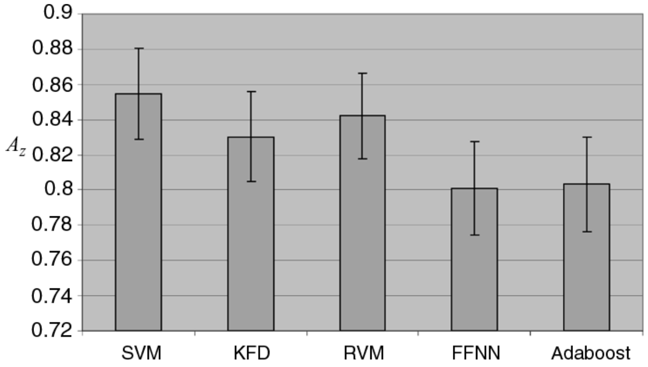


Fig. 3. Classification results obtained with different classifier models.

The results in Table I indicate that for both SVM and FFNN the averaging method consistently leads to the best classification performance (though only slightly better than the direct method). This may seem rather counter-intuitive, because of the following two reasons: 1) the direct method leads to twice as many training samples as in the other two methods, which in principle could lead to better training of the classifiers; and 2) the joint method makes simultaneously available the information from multiple views to a classifier, which could lead to more informative decision by the classifier. Yet, both methods seem to under-perform the averaging method.

Our explanation is as follows: the samples from the multiple views of an MCC are inherently correlated; indeed an ensuing calculation yields that the cross-correlation of the feature vectors between ML and CC views is 0.86 for the cases in the data set. The results in Table I demonstrate that having the multiple views as separate samples in training added little value when compared to using their averages as samples. By averaging the multiple views, one can eliminate some of the potential noise originated in feature extraction, which could be the cause for the improved performance. On the other hand, the joint method causes the dimension of the input vector to be doubled, which significantly increases the complexity of learning. Given the limited number of training samples available, the benefit of such an approach did not materialize in our experiment.

Consequently, the averaging method is used in all subsequent results.

B. Performance by Different Learning Models

We summarize in Fig. 3 the classification results, obtained using the single-set procedure, for all the classifier models described in Section II; the area under the ROC curve A_z is shown for each method. We list in Table II the estimate of A_z and its standard deviation, obtained using the ROCKIT program, for

TABLE II
CLASSIFICATION RESULTS OBTAINED WITH DIFFERENT CLASSIFIER MODELS

	SVM	KFD	RVM	FFNN	AdaBoost
A_z	**0. 8545**	0.8303	0.8421	0.8007	0.8034
Std. Dev.	0.0259	0.0254	0.0243	0.0266	0.0268

TABLE III
PARAMETRIC SETTINGS OF DIFFERENT CLASSIFIER MODELS

	SVM	KFD	RVM	FFNN	AdaBoost
Parameters	Order-2 polynomial kernel, C=700	Order-2 polynomial kernel	Order-2 polynomial kernel	3 layers, 6 hidden neurons, 100 seeds	30 rounds

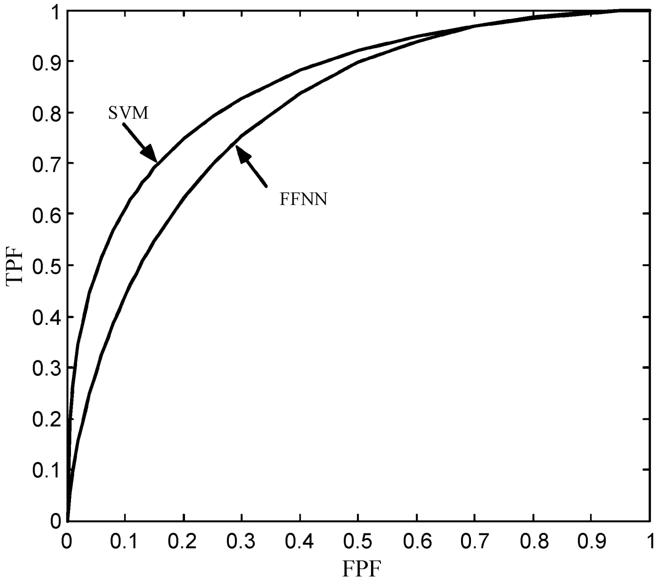


Fig. 4. The ROC curves for SVM and FFNN obtained using the one-set procedure.

each classifier model. In addition, we list in Table III the parametric settings resulted from the training procedure for the different classifier models.

These results demonstrate that the kernel methods (SVM, KFD, and RVM) are similar in performance (in terms of A_z), whereas FFNN and AdaBoost are similar; the former group outperforms the latter.

In particular, a statistical comparison between SVM and FFNN using the ROCKIT program yields a two-tailed p-value of 0.0199 (one-tailed p-value is 0.0099) for rejecting the null hypothesis that their corresponding ROC curves have the same area under them; moreover, an approximate 95% confidence interval on the difference in A_z between the two methods is computed to be (0.0091, 0.1056). The corresponding ROC curves for SVM and FFNN are shown in Fig. 4. As can be seen, SVM yields a notably higher ROC curve than FFNN.

We point out that the strong performance by SVM over FFNN in this particular task is not surprising. As explained earlier in Section III, the dataset used in our study includes cases that are difficult to classify, and the MDS plot reveals that the malignant cases and benign cases are closely distributed in the feature space and, moreover, no clear separation between the two classes seems to exist. In such a case, a learning method solely based on minimization of training errors (such as FFNN) can potentially suffer from over-fitting, leading to poor generalization beyond the training samples.

TABLE IV
EVALUATION RESULTS OBTAINED USING MULTI-SET PROCEDURE. A 95% CONFIDENCE INTERVAL FOR THE DIFFERENCE IN A_z BETWEEN SVM AND FFNN IS (.0009, .0822)

	SVM	FFNN
A_z	0.8138	0.7722
Std. Dev.	0.0492	0.0462

Among the kernel methods, SVM yields the highest A_z value. In our experiment, approximately 30% of the training samples were found to be support vectors in the SVM decision function, while for RVM only about 3% of the training samples were relevance vectors.

Furthermore, to investigate the robustness of the SVM classifier, we also evaluated its classification performance by varying both the parameter of the kernel function and the regularization parameter C. The results are shown in Fig. 5. As can be seen, the classification performance degrades gracefully as the SVM is perturbed from its optimal parametric setting (order-2 polynomial kernel with $C = 700$). A similar set of results was also obtained when the RBF kernel was used, but omitted here for brevity.

The results in Fig. 5 also demonstrate that the performance of the SVM classifier depends on the proper choice of the parameters. As the regularization parameter C is increased toward ∞, the SVM performance starts to degrade. This is not unexpected, since the empirical risk term (training errors) in the cost function in (2) starts to dominate as C is increased toward ∞, eventually leading to over-fitting (hence poor generalization).

C. Multi-Set Comparison

We next conducted further comparison between SVM and FFNN by using the multiple-set procedure. The results are summarized in Table IV, in which the estimated A_z value and its standard deviation, obtained using the LABMRMC program, are given for each method. A statistical comparison between the

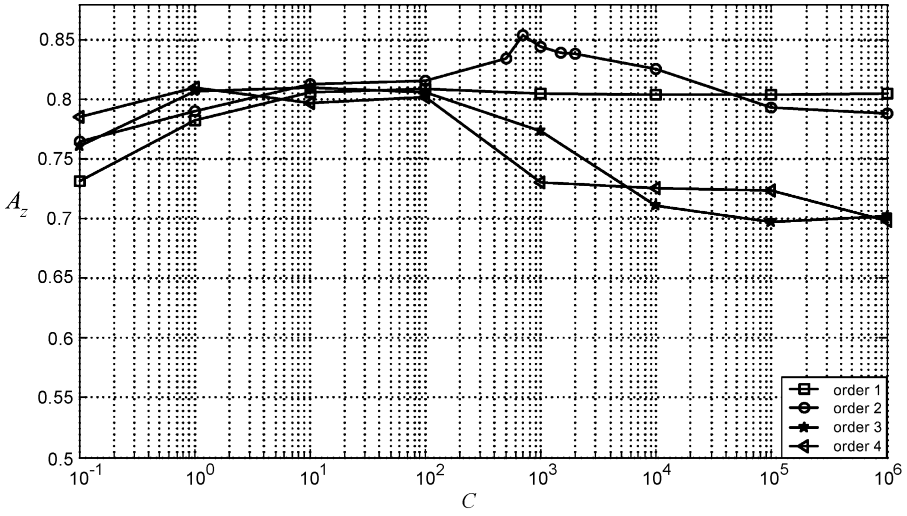


Fig. 5. The classification performance by SVM under different parametric settings.

TABLE V
EXECUTION TIME (IN SEC.) FOR DIFFERENT CLASSIFIER MODELS. FOR FFNN, THE RESULTS GIVEN ARE FOR A SINGLE NETWORK OF THE 100-MEMBER COMMITTEE

	SVM	KFD	RVM	FFNN	AdaBoost
Training	1.0	3.7	12.6	5.3	40.2
Testing	< 0.001	<0.001	< 0.001	0.01	0.31

two methods yields a p-value of 0.0453 for rejecting the null hypothesis that the two methods have the same area under their ROC curves; a 95% confidence interval on the difference in A_z between SVM and FFNN is computed to be (.0009, .0822). Furthermore, the difference comparison between the two readers (i.e., two training sets) corresponds to a p-value of 0.1043 (not deemed significant at a level of 0.05).

Compared to their counterparts in Table II, the estimated values of A_z in Table IV are notably decreased, and their confidence intervals are increased. This is expected, because the training sets and test set are much reduced in size in the multi-set procedure, leading to increased generalization errors and increased uncertainty and a negative bias in A_z estimates; however, the multi-set procedure removes any potential overly-optimistic bias that is inherent with the single-set procedure. Nevertheless, it is rather interesting that the two procedures yield nearly the same confidence intervals on the *difference* in A_z between SVM and FFNN. Thus, the results from the two different procedures corroborate with each other conclusively that SVM yields improved classification performance over FFNN.

D. Training With Unbalanced Samples

Using the single-set procedure, we experimented with over-sampling the malignant cases in the training set. For SVM, when the malignant cases are over-sampled by a factor of two, three, and four, the resulting A_z is 0.8222, 0.8131, and 0.8126, respectively; similarly, for FFNN, the resulting A_z is 0.7934, 0.7988, and 0.7967, respectively. Compared with no over-sampling (Table II), the over-sampling approach did not seem to improve the classification performance with the given database.

E. Execution Time

In our experiments, all the algorithms were implemented with MATLAB on a Pentium III 933-MHz PC. We summarize in Table V the execution times for the different classifier models. For each classifier model, the training time was computed as the average time taken by each training step in the single-set procedure (386 steps in total); the testing time was computed as the average time taken by the trained model to classify each of the 386 cases in the database. For FFNN, the results given in Table V are the average times taken by only a single network in the 100-member committee.

VI. CONCLUSION

In this paper, we investigated the use of SVM, KFD, RVM, and committee machines for classification of clustered MCs in digital mammograms. These different classifier models were trained through supervised learning to classify whether a cluster of MCs is malignant or benign, based on quantitative image features extracted from the MCs. These classifiers were tested using a database of 697 clinical mammograms, which included a wide spectrum of difficult-to-classify cases. Results obtained from two different sets of experiments demonstrated that the kernel based methods (i.e., SVM, KFD, and RVM) yielded the best performance, outperforming that of FFNN and AdaBoost; furthermore, these methods were also computationally advantageous both in training and in testing. The results also demonstrated that the averaging method offered the best performance for combining information from multiple-view mammograms.

REFERENCES

[1] A. M. Knutzen and J. J. Gisvold, "Likelihood of malignant disease for various categories of mammographically detected, nonpalpable breast lesions," *Mayo Clin. Proc.*, vol. 68, pp. 454–460, 1993.

[2] D. B. Kopans, "The positive predictive value of mammography," *AJR*, vol. 158, pp. 521–526, 1992.

[3] M. L. Giger, "Overview of computer-aided diagnosis in breast imaging," in *Comput.-Aided Diagnosis in Medical Imaging*. Amsterdam, The Netherlands: Elsevier, 1998, pp. 167–176.

[4] Y. Jiang, R. M. Nishikawa, E. E. Wolverton, C. E. Metz, M. L. Giger, R. A. Schmidt, and C. J. Vyborny, "Malignant and benign clustered microcalcifications: automated feature analysis and classification," *Radiology*, vol. 198, pp. 671–678, 1996.

[5] Y. Jiang, R. M. Nishikawa, R. A. Schmidt, C. E. Metz, M. L. Giger, and K. Doi, "Improving breast cancer diagnosis with computer-aided diagnosis," *Academic Radiol.*, vol. 6, pp. 22–33, 1999.

[6] Y. Wu, M. L. Giger, K. Doi, C. J. Vyborny, R. A. Schmidt, and C. E. Metz, "Application of neural networks in mammography: applications in decision making in the diagnosis of breast cancer," *Radiology*, vol. 187, pp. 81–87, 1993.

[7] Z. Huo, M. L. Giger, C. J. Vyborny, and C. E. Metz, "Effectiveness of CAD in the diagnosis of breast cancer: an observer study on an independent database of mammograms," *Radiology*, vol. 7, pp. 1077–1084, 2000.

[8] C. J. D'Orsi, D. J. Getty, J. A. Swets, R. M. Pickett, S. E. Seltzer, and B. J. McNeil, "Reading and decision aids for improved accuracy and standardization of mammographic diagnosis," *Radiology*, vol. 184, pp. 619–622, 1992.

[9] J. A. Baker, P. J. Kornguth, J. Y. Lo, and C. E. Floyd Jr., "Artificial neural network: improving the quality of breast biopsy recommendations," *Radiology*, vol. 198, pp. 131–136, 1996.

[10] H. P. Chan, B. Sahiner, M. A. Helvie, N. Petrick, M. A. Roubidoux, T. E. Wilson, D. D. Adler, C. Paramagul, J. S. Newman, and S. Sanjay-Gopal, "Improvement of radiologists' characterization of mammographic masses by using computer-aided diagnosis: an ROC study," *Radiology*, vol. 212, pp. 817–827, 1999.

[11] J. Y. Lo, J. A. Baker, P. J. Kornguth, J. D. Iglehart, and C. E. Floyd, "Predicating breast cancer invasion with artificial neural networks on the basis of mammographic features," *Radiology*, vol. 203, pp. 159–163, 1997.

[12] H. P. Chan, B. Sahiner, K. L. Lam, N. Petrick, M. A. Helvie, M. M. Goodsitt, and D. D. Adler, "Computerized analysis of mammographic microcalcifications in morphological and texture feature spaces," *Med. Phys.*, vol. 25, pp. 2007–2019, 1998.

[13] A. Dhawan, Y. Chitre, C. Bonasso, and K. Wheeler, "Radial-basis-function-based classification of mammographic microcalcifications using texture features," in *Proc. 17th Annu. Int. Conf. IEEE Engineering in Medicine and Biology Society*, vol. 1, 1995, pp. 535–536.

[14] E. Kouskos, C. Markopoulos, K. Revenas, K. Koufopoulos, V. Kyriakou, and J. Gogas, "Computer-aided preoperative diagnosis of microcalcifications on mammograms," *Acta Radiologica*, vol. 44, pp. 43–46, 2003.

[15] R. J. Ferrari, P. M. Azevedo-Marques, A. F. Frere, S. K. Kinoshita, and L. A. R. Spina, "Characterization of breast cancer using statistical approaches," in *Computer-Aided Diagnosis in Medical Imaging*. Amsterdam, The Netherlands: Elsevier, 1998, pp. 281–286.

[16] W. J. H. Veldkamp, N. Karssemeijer, J. D. M. Otten, and J. H. C. L. Hendriks, "Automated classification of clustered microcalcification into malignant and benign types," *Med. Phys.*, vol. 27, pp. 2600–2608, 2000.

[17] O. R. Zaiane, M.-L. Antonie, and A. Coman, "Mammography classification by an association rule-based classifier," in *Proc. MDK/KDD 2002: Int. Workshop Multimedia Data Mining*, 2002, pp. 62–69.

[18] V. Vapnik, *Statistical Learning Theory*. New York: Wiley, 1998.

[19] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K.-R. Muiller, "Fisher discriminant analysis with kernels," in *Neural Networks for Signal Processing IX*. Piscataway, NJ: IEEE, 1999, pp. 41–48.

[20] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Machine Learn. Res.*, no. 1, pp. 211–244, 2001.

[21] S. Haykin, *Neural Network—A Comprehensive Foundation*, 2nd ed. Upper Saddle River: Prentice-Hall, 1999.

[22] Y. Freund and R. Schapire, "A decision-theoretic generalization of online learning and an application to boosting," *J. Comput. Syst. Sci.*, vol. 55, pp. 119–139, 1997.

[23] M. Pontil and A. Verri, "Support vector machines for 3-D object recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, no. 6, pp. 637–646, Jun. 1998.

[24] V. Wan and W. M. Campbell, "Support vector machines for speaker verification and identification," in *Proc. IEEE Workshop on Neural Networks for Signal Processing*, Sydney, Australia, Dec. 2000, pp. 775–784.

[25] E. Osuna, R. Freund, and F. Girosi, "Training support vector machines: application to face detection," in *Proc. Computer Vision and Pattern Recognition*, Puerto Rico, 1997, pp. 130–136.

[26] T. Joachims, "Transductive inference for text classification using support vector machines," in *Proc. 16th Int. Conf. Machine Learning*, Bled, Slovenia, Jun. 1999, pp. 200–209.

[27] C. J. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining Knowledge Discovery*, vol. 2, pp. 121–167, 1998.

[28] I. El-Naqa, Y. Yang, M. N. Wernick, N. P. Galatsanos, and R. M. Nishikawa, "A support vector machine approach for detection of microcalcifications," *IEEE Trans. Med. Imag.*, vol. 21, no. 12, pp. 1552–1563, Dec. 2002.

[29] I. Borg and P. Groenen, *Modern Multidimensional Scaling: Theory and Applications*. Berlin, Germany: Springer-Verlag, 1997.

[30] B. Ripley, *Pattern Recognition and Neural Networks*. Cambridge, U.K.: Cambridge Univ. Press, 1996.

[31] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Eugenics*, vol. 7, pp. 179–188, 1936.

[32] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. San Diego, : Academic, 1990.

[33] B. Scholkopf and A. Smola, *Learning With Kernels—Support Vector Machines, Regularization, Optimization and Beyond*. Cambridge, MA: MIT Press, 2002.

[34] H. Schwenk and Y. Bengio, "Adaboosting neural networks: application to on-line character recognition," in *Proc. Int. Conf. Artificial Neural Networks (ICANN'97)*, 1997, pp. 967–972.

[35] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Int. Conf. Computer Vision and Pattern Recognition*, Hawaii, 2001, pp. 511–518.

[36] American college of radiology, Reston, VA, *American College of Radiology Breast Imaging-Reporting and Data Systems (BI-RADS)*, 3rd ed., 1998.

[37] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Berlin, Germany: Springer-Verlag, 2001.

[38] N. Kapkowicz, ""Learning from imbalanced data sets: a comparison of various strategies," *Proceedings of Learning From Imbalanced Data Sets, AAAI Workshop*," AAAI Press, Menlo Park, CA, Tech. Rep. WS-00-05, 2000, pp. 10–15.

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

[40] D. D. Dorfman, K. S. Berbaum, and C. E. Metz, "Receiver operating characteristic rating analysis: generalization to the population of readers and patients with the Jackknife method," *Investigat. Radiol.*, vol. 27, pp. 723–731, 1992.

[41] H. P. Chan, B. Sahiner, R. F. Wagner, and N. Petrick, "Classifier design for computer-aided diagnosis: effects of finite sample size on the mean performance of classical and neural network classifiers," *Med. Phys.*, vol. 26, pp. 2654–2668, 1999.

[42] B. Sahiner, H. P. Chan, N. Petrick, R. F. Wagner, and L. Hadjiiski, "Feature selection and classifier performance in computer-aided diagnosis: the effect of finite sample size," *Med. Phys.*, vol. 27, pp. 1509–1522, 2000.