A Support Vector Machine Approach for Detection of Microcalcifications

Issam El-Naqa, *Student Member, IEEE*, Yongyi Yang*, *Member, IEEE*, Miles N. Wernick, *Senior Member, IEEE*, Nikolas P. Galatsanos, *Senior Member, IEEE*, and Robert M. Nishikawa

***Abstract*—In this paper, we investigate an approach based on support vector machines (SVMs) for detection of microcalcification (MC) clusters in digital mammograms, and propose a successive enhancement learning scheme for improved performance. SVM is a machine-learning method, based on the principle of structural risk minimization, which performs well when applied to data outside the training set. We formulate MC detection as a supervised-learning problem and apply SVM to develop the detection algorithm. We use the SVM to detect at each location in the image whether an MC is present or not. We tested the proposed method using a database of 76 clinical mammograms containing 1120 MCs. We use free-response receiver operating characteristic curves to evaluate detection performance, and compare the proposed algorithm with several existing methods. In our experiments, the proposed SVM framework outperformed all the other methods tested. In particular, a sensitivity as high as 94% was achieved by the SVM method at an error rate of one false-positive cluster per image. The ability of SVM to outperform several well-known methods developed for the widely studied problem of MC detection suggests that SVM is a promising technique for object detection in a medical imaging application.**

***Index Terms*—Computer-aided diagnosis, kernel methods, microcalcifications, support vector machines.**

I. Introduction

In THIS paper we propose the use of support vector machine (SVM) learning to detect microcalcification (MC) clusters in digital mammograms. SVM is a learning tool originated in modern statistical learning theory [1]. In recent years, SVM learning has found a wide range of real-world applications, including handwritten digit recognition [2], object recognition [3], speaker identification [4], face detection in images [5], and text categorization [6]. The formulation of SVM learning is based on the principle of structural risk minimization. Instead of minimizing an objective function based on the training samples [such as mean square error (MSE)], the SVM attempts to minimize a bound on the generalization error (i.e., the error made by the learning machine on test data not used during training). As a result, an SVM tends to perform well when applied to data outside the training set. Indeed, it has been reported that SVM-based approaches are able to significantly outperform competing methods in many applications [7]–[9]. SVM achieves this advantage by focusing on the training examples that are most difficult to classify. These "borderline" training examples are called *support vectors*.

In this paper, we investigate the potential benefit of using an SVM-based approach for object detection from medical images. In particular, we consider the detection of MC clusters in mammograms. There are two main reasons for addressing this particular application using SVM. First, accurate detection of MC clusters is itself an important problem. MC clusters can be an early indicator of breast cancer in women. They appear in 30–50% of mammographically diagnosed cases. In the United States, women have a baseline risk of 5%–6% of developing cancer; 50% of these may die from the disease [10]. Second, because of the importance of accurate breast-cancer diagnosis and the difficulty of the problem, there has been a great deal of research to develop methods for automatic detection of MC clusters. Therefore, the problem of MC cluster detection is one that is well understood, and provides a good testing ground for comparing SVM with other more-established methods. The strong performance of SVM in our studies indicates that SVM indeed can be a useful technique for object detection in medical imaging.

In the proposed approach, MC cluster detection is accomplished through detection of individual MCs using an SVM classifier. MCs are small calcium deposits that appear as bright spots in a mammogram (see Fig. 1). Individual MCs are sometimes difficult to detect due to their variation in shape, orientation, brightness and size (typically, 0.05–1 mm), and because of the surrounding breast tissue [11]. In this paper, an SVM is trained through supervised learning to classify each location in the image as "MC present" or "MC absent."

A difficult problem that arises in training a classifier for MC detection is that there are a very large number of image locations where no MC is present, so that the training set for the "MC absent" class can be impractically large. Thus, there arises an issue of how to select the training examples so that they well represent the class of "MC absent" locations. To solve this

Manuscript received April 9, 2002; revised September 4, 2002. This work was supported by the National Institutes of Health (NIH)/National Cancer Institute (NCI) under Grant CA89668. The work of R. M. Nishikawa was supported in part by NIH/NCI under Grant CA60187. R. M. Nishikawa is a shareholder in R2 Technology, Inc. (Los Altos, CA). It is the University of Chicago Conflict of Interest Policy that investigators disclose publicly actual or potential significant financial interests that may appear to be affected by the research activities. The Associate Editor responsible for coordinating the review of this paper and recommending its publication was N. Karssemeijer. *Asterisk indicates corresponding author.*

I. El-Naqa, M. N. Wernick, and N. P. Galatsanos are with the Department of Electrical and Computer 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.

R. M. Nishikawa is with the Department of Radiology, The University of Chicago, Chicago, IL 60637 USA.

Digital Object Identifier 10.1109/TMI.2002.806569

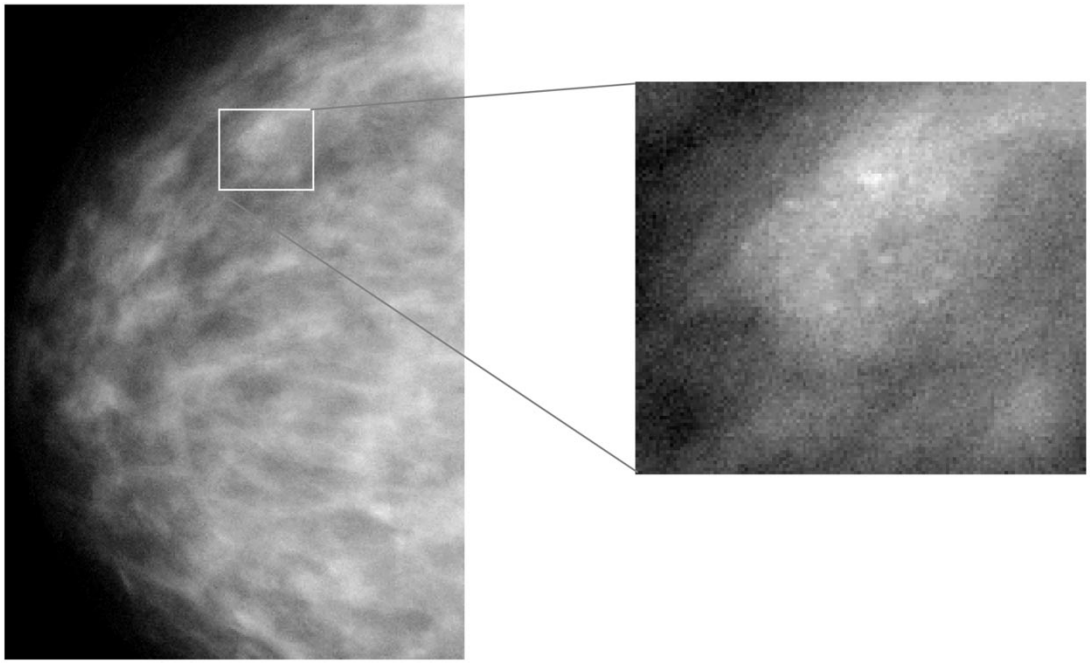

Fig. 1. (left) Mammogram in craniocaudal view. (right) Expanded view showing MCs.

problem we propose a solution that we call *successive enhancement-learning* (SEL) to select the training examples. SEL selects iteratively the "most representative" MC-absent examples from all the available training images while keeping the total number of training examples small. Numerical results demonstrate that this approach can improve the generalization ability of the SVM classifier.

We developed the proposed SVM approach using a database of 76 clinical mammograms containing 1120 MCs. These mammograms were divided equally into two subsets, one used exclusively for training and the other exclusively for testing. Compared to several other existing methods, the proposed approach yielded superior performance when evaluated using free-response receiver operating characteristic (FROC) curves. It achieved sensitivity as high as 94% with only about one false-positive MC cluster per mammogram. This figure of merit is difficult to compare with previous reports in the literature because, as we will show, the sensitivity measure depends strongly on the way MC clusters are defined. However, within each of our studies we maintained a uniform definition for clusters to allow for meaningful comparisons.

The rest of the paper is organized as follows. A brief review of the literature on MC detection is provided in the remainder of this section. A background on SVM learning is furnished in Section II. The use of an SVM for MC detection is formulated in Section III. An evaluation study of the proposed SVM approach is described in Section IV, and the experimental results are presented in Section V. Finally, conclusions are drawn in Section VI. A proof of convergence of the proposed SEL scheme is given in the Appendix.

There exist many methods for MC detection (a thorough review of various methods can be found in Nishikawa [12]). There is also a commercial computer-aided diagnosis system developed (e.g., high detection sensitivity is claimed in [13]). The following is a brief review of some representative methods for detection of MCs. Karssenmeijer [14] developed a statistical Bayesian image analysis model for detection of MCs. Nishikawa *et al.* [15] investigated a method based on a difference image technique followed by morphological post-processing. Wavelet-based approaches have been proposed in [16]–[18]. In [16], a decimated wavelet transform and supervised learning are combined for the detection of MCs, while in [17] and [18] an undecimated wavelet transform and optimal subband weighting are used. A detection scheme is proposed in [19] for the automatic detection of clustered MCs using multiscale analysis based on the Laplacian-of-Gaussian filter and a mathematical model describing an MC as a bright spot of a certain size and contrast. Dengler *et al.* [20] used methods based on a weighted difference-of-Gaussian (DoG) filter for spot detection and morphological operators to extract shape features. Gurcan *et al.* [21] developed a method based on higher order statistics. Cheng *et al.* [22] applied fuzzy logic for MC detection. Pfrench *et al.* [23] presented a two–dimensional adaptive lattice algorithm to predict correlated clutters (i.e., the tissue structure) in the mammogram. Li *et al.* [24] proposed using fractal background modeling, taking the difference between the original and the modeled image, which results in enhanced MC detection. Bankman *et al.* [25] developed a method based on region-growing in conjunction with active contours, wherein the seed points are selected as the local maxima found by an edge-detection operator. Mixed wavelet components, gray-level statistics, and shape features were used to train a two-stage multilayer neural network (TMNN) for detection of individual MC objects [26]. Recently, Bazzani *et al.* [27] proposed a method for MC detection based on multiresolution filtering analysis and statistical testing, in which an SVM classifier was used to reduce the false detection rate. This approach is quite different from ours in that it used

extracted image features (including area, average pixel value, edge gradient, degree of linearity, and average gradient) as the basis for detection, while our approach does not attempt to extract any explicit image features. Instead, we directly use finite image windows as input to the SVM classifier, and rely on the capability of the SVM to automatically learn the relevant features for optimal detection.

II. Review of SVM Learning for Classification

In this paper, we treat MC detection as a two-class pattern classification problem. At each location in a mammogram, we apply a classifier to determine whether an MC is present or not. We refer to these two classes throughout as "MC present" and "MC absent." Let vector $\mathbf{x} \in R^n$ denote a pattern to be classified, and let scalar y denote its class label (i.e., $y \in \{\pm 1\}$). In addition, let $\{(\mathbf{x}_i, y_i), i = 1, 2, \ldots, l\}$ denote a given set of l training examples. The problem is how to construct a classifier [i.e., a decision function $f(\mathbf{x})$] that can correctly classify an input pattern $\mathbf{x}$ that is not necessarily from the training set.

A. Linear SVM Classifiers

Let us begin with the simplest case, in which the training patterns are linearly separable. That is, there exists a linear function of the form

$$f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b \tag{1}$$

such that for each training example $\mathbf{x}_i$, the function yields $f(\mathbf{x}_i) \geq 0$ for $y_i = +1$, and $f(\mathbf{x}_i) < 0$ for $y_i = -1$. In other words, training examples from the two different classes are separated by the hyperplane $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b = 0$.

For a given training set, while there may exist many hyperplanes that separate the two classes, the SVM classifier is based on the hyperplane that maximizes the separating margin between the two classes (Fig. 2) [7], [9]. In other words, SVM finds the hyperplane that causes the largest separation between the decision function values for the "borderline" examples from the two classes. Mathematically, this hyperplane can be found by minimizing the following cost function:

$$J(\mathbf{w}) = \tfrac{1}{2}\,\mathbf{w}^T\mathbf{w} = \tfrac{1}{2}\,||\mathbf{w}||^2 \tag{2}$$

subject to the separability constraints

$$\mathbf{w}^T\mathbf{x}_i + b \geq +1, \qquad \text{for } y_i = +1$$

or

$$\mathbf{w}^T\mathbf{x}_i + b \leq -1, \qquad \text{for } y_i = -1;\ i = 1, 2, \ldots, l. \tag{3}$$

Equivalently, these constraints can be written more compactly as

$$y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1; \qquad i = 1, 2, \ldots, l. \tag{4}$$

This specific problem formulation may not be useful in practice because the training data may not be completely separable by a hyperplane. In this case, slack variables, denoted by ξ_i, can be introduced to relax the separability constraints in (4) as follows:

$$y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i, \qquad \xi_i \geq 0;\ i = 1, 2, \ldots, l. \tag{5}$$

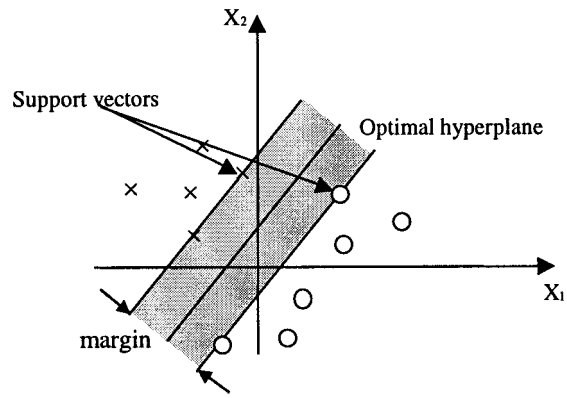

Fig. 2. SVM classification with a hyperplane that maximizes the separating margin between the two classes (indicated by data points marked by "×"s and "○"s). Support vectors are elements of the training set that lie on the boundary hyperplanes of the two classes.

Accordingly, the cost function in (2) can be modified as follows:

$$J(\mathbf{w}, \xi) = \tfrac{1}{2}\,||\mathbf{w}||^2 + C\sum_{i=1}^{l}\xi_i \tag{6}$$

where C is a user-specified, positive, regularization parameter. In (6), the variable ξ is a vector containing all the slack variables $\xi_i, i = 1, 2, \ldots, l$.

The modified cost function in (6) constitutes the so-called *structural risk*, which balances the *empirical risk* (i.e., the training errors reflected by the second term) with model complexity (the first term) [28]. The regularization 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.

B. Nonlinear SVM Classifiers

The linear SVM can be readily extended to a nonlinear classifier by first using a nonlinear operator $\Phi(\cdot)$ to map the input pattern $\mathbf{x}$ into a higher dimensional space $\boldsymbol{\mathcal{H}}$. The nonlinear SVM classifier so obtained is defined as

$$f(\mathbf{x}) = \mathbf{w}^T\Phi(\mathbf{x}) + b \tag{7}$$

which is linear in terms of the transformed data $\Phi(\mathbf{x})$, but nonlinear in terms of the original data $\mathbf{x} \in R^n$.

Following nonlinear transformation, the parameters of the decision function $f(\mathbf{x})$ are determined by the following minimization:

$$\min\ J(\mathbf{w}, \xi) = \tfrac{1}{2}\,||\mathbf{w}||^2 + C\sum_{i=1}^{l}\xi_i \tag{8}$$

subject to

$$y_i(\mathbf{w}^T\Phi(\mathbf{x}_i) + b) \geq 1 - \xi_i, \qquad \xi_i \geq 0;\ i = 1, 2, \ldots, l. \tag{9}$$

C. Solution of SVM Formulation

Using the technique of Lagrange multipliers, one can show that a necessary condition for minimizing $J(\mathbf{w}, \xi)$ in (8) is that

the vector $\mathbf{w}$ is formed by a linear combination of the mapped vectors $\Phi(\mathbf{x}_i)$, i.e.,

$$\mathbf{w} = \sum_{i=1}^{l} \alpha_i y_i \Phi(\mathbf{x}_i) \tag{10}$$

where $\alpha_i \geq 0$, $i = 1, 2, \ldots, l$, are the Lagrange multipliers associated with the constraints in (9).

Substituting (10) into (7) yields

$$f(\mathbf{x}) = \sum_{i=1}^{l} \alpha_i y_i \Phi^T(\mathbf{x}_i)\Phi(\mathbf{x}) + b = \sum_{i=1}^{l} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b \tag{11}$$

where the function $K(\cdot, \cdot)$ is defined as

$$K(\mathbf{x}, \mathbf{z}) \equiv \Phi^T(\mathbf{x})\Phi(\mathbf{z}). \tag{12}$$

The Lagrange multipliers $\alpha_i \geq 0, i = 1, 2, \ldots, l$, are solved from the dual form of (8), which is expressed as

$$\max\ W(\alpha_1, \alpha_2, \ldots, \alpha_l) = \sum_{i=1}^{l} \alpha_i - \tfrac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \tag{13}$$

subject to

$$0 \leq \alpha_i \leq C, \qquad i = 1, 2, \ldots, l \tag{14}$$

$$\sum_{i=1}^{l} \alpha_i y_i = 0. \tag{15}$$

Notice that the cost function $W(\alpha_1, \alpha_2, \ldots, \alpha_l)$ is convex and quadratic in terms of the unknown parameters α_i. In practice, this problem is solved numerically through quadratic programming.

Analytic solutions of (13) are not readily available, but it is still informative to examine the conditions under which an optimal solution is achieved. The Karush–Kuhn–Tucker optimality conditions for (13) lead to the following three cases for each α_i:

1) $\alpha_i = 0$. This corresponds to $y_i f(\mathbf{x}_i) > 1$. In this case, the data element $\mathbf{x}_i$ is outside the decision margin of the function $f(\mathbf{x})$ and is correctly classified.
2) $0 < \alpha_i < C$. In this case, $y_i f(\mathbf{x}_i) = 1$. The data element $\mathbf{x}_i$ is strictly located on the decision margin of $f(\mathbf{x})$. Hence, $\mathbf{x}_i$ is called a *margin support vector* of $f(\mathbf{x})$.
3) $\alpha_i = C$. In this case, $y_i f(\mathbf{x}_i) < 1$. The data element $\mathbf{x}_i$ is inside the decision margin (though it may still be correctly classified). Accordingly, $\mathbf{x}_i$ is called an *error support vector* of $f(\mathbf{x})$.

Note that most of the training examples in a typical problem are correctly classified by the trained classifier (case 1), i.e., only a few training examples will be support vectors. For simplicity, let $\mathbf{s}_j$, α_j^*, $j = 1, 2, \ldots, l_s$, denote these support vectors and their corresponding nonzero Lagrange multipliers, respectively, and let y_j denote their class labels. The decision function in (11) can, thus, be simplified as

$$f(\mathbf{x}) = \sum_{j=1}^{l_s} \alpha_j^* y_j \Phi^T(\mathbf{s}_j)\Phi(\mathbf{x}) + b = \sum_{j=1}^{l_s} \alpha_j^* y_j K(\mathbf{s}_j, \mathbf{x}) + b. \tag{16}$$

Note that the decision function is now determined directly by the support vectors $\mathbf{s}_j$, $j = 1, 2, \ldots, l_s$, which are determined by solving the optimization problem in (13) during the training phase.

D. SVM Kernel Functions

Notice that the nonlinear mapping $\Phi(\cdot)$ from R^n to $\boldsymbol{\mathcal{H}}$ never appears explicitly in either the dual form of SVM training in (13) or the resulting decision function in (16). The mapping $\Phi(\cdot)$ enters the problem only implicitly through the kernel function $K(\cdot, \cdot)$, thus, it is only necessary to define $K(\cdot, \cdot)$, which implicitly defines $\Phi(\cdot)$. However, when choosing a kernel function $K(\cdot, \cdot)$, it is necessary to check that it is associated with the inner product of some nonlinear mapping. Mercer's theorem states that such a mapping indeed underlies a kernel $K(\cdot, \cdot)$ provided that $K(\cdot, \cdot)$ is a positive integral operator [28], [29], that is, for every square-integratable function $g(\cdot)$ defined on R^n the kernel $K(\cdot, \cdot)$ satisfies the following condition:

$$\iint K(\mathbf{x}, \mathbf{y}) g(\mathbf{x}) g(\mathbf{y})\, d\mathbf{x}\, d\mathbf{y} \geq 0. \tag{17}$$

Examples of kernels satisfying Mercer's condition include polynomials and radial basis functions (RBFs), which will be discussed in Section III.

III. SVM Formulation for Microcalcification Detection

In this section, we present a supervised SVM learning framework for detection of MCs in which an SVM is first trained using existing mammograms. The ground truth of MCs in these mammograms is assumed to be known *a priori*. A detailed formulation of the SVM learning framework is presented in the following discussion. A performance evaluation of the method is presented in Section IV.

A. Input Feature Vector

Individual MCs are well localized in a mammogram; therefore, to detect whether an MC is present at a given location, it is sufficient to examine the image content within a small neighborhood around that location. Thus, we define the input pattern to the SVM classifier to be a small $M \times M$ pixel window centered at the location of interest.

The window should be chosen large enough to contain an MC, but small enough to avoid potential interference from neighboring MCs. A small window size is also favorable for computational reasons. In our study, the mammograms were digitized at a resolution of 0.1 mm/pixel, and we chose $M = 9$. Our experiments indicated that the results were not very sensitive to the choice of M (e.g., similar performance was achieved when $M = 7$ was used).

To suppress the image background and, thus, restrict intraclass variation among the training patterns, we begin by applying a sharp high-pass filter to each mammogram. This filter was designed as a linear-phase finite impulse response filter with 3-dB cutoff frequency $\omega_c = 0.125$ and length 41. As an example, we show in Fig. 3 the result after filtering the mammogram in Fig. 1 with this filter. The filter appears to be effective in reducing the inhomogeneity of the background.

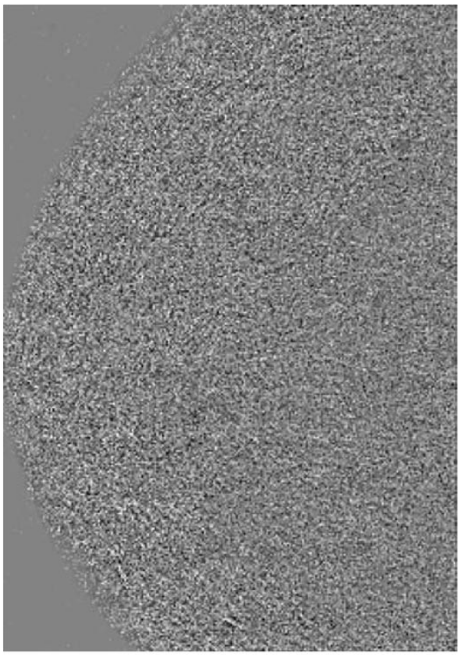

Fig. 3. The mammogram in Fig. 1 after background removal by a high-pass filter designed for the purpose.

To summarize, if we let $\mathbf{f}$ denote the entire mammogram, and W be a windowing operator that extracts the $M \times M$ window centered at a particular location, then the input feature vector $\mathbf{x}$ is extracted as follows:

$$\mathbf{x} = W[H\mathbf{f}] \tag{18}$$

where H denotes the high-pass filter for background removal. Note that the vector $\mathbf{x}$ is of dimension M^2 (81 in this study), and is formed at every image location where an MC is to be detected [the fact that $\mathbf{x}$ varies with location is not explicitly indicated in (18) for notational simplicity].

The task of the SVM classifier is to decide whether the input vector $\mathbf{x}$ at each location is an MC pattern $(y = +1)$ or not $(y = -1)$.

B. SVM Kernel Functions

The kernel function in an SVM plays the central role of implicitly mapping the input vector (through an inner product) into a high-dimensional feature space. In this paper, we consider two kernel types: polynomial kernels and Gaussian RBFs. These are among the most commonly used kernels in SVM research, and are known to satisfy Mercer's condition [28]. They are defined as follows.

1) Polynomial kernel:

$$K(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^T\mathbf{y} + 1)^p \tag{19}$$

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

2) Gaussian RBF kernel:

$$K(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{y}\|^2}{2\sigma^2}\right) \tag{20}$$

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

Notice that in both cases the kernel function serves essentially as a similarity measure between $\mathbf{x}$ and $\mathbf{y}$. In particular, the polynomial kernel function in (19) assumes its maximum when $\mathbf{x}$ and $\mathbf{y}$ are aligned in the same direction (with their respective lengths fixed); while the Gaussian RBF kernel function in (20) assumes its maximum when $\mathbf{x}$ and $\mathbf{y}$ are identical. The associated parameters, order p in (19) and width σ in (20), are determined during the training phase.

C. Preparation of Training Data Set

The procedure for extracting training data from the training mammogram set is as follows. For each MC location in a training-set mammogram, a window of $M \times M$ image pixels centered at its center of mass is extracted; the vector formed by this window of pixels, denoted by $\mathbf{x}_i$, is then treated as an input pattern for the "MC present" class $(y_i = +1)$. "MC absent" samples are collected $(y_i = -1)$ similarly, except that their locations are selected randomly from the set of all "MC absent" locations in the training mammograms. In this procedure, no window in the training set is allowed to overlap with any other training window. The reason for using only a random subset of "MC absent" examples is that there are too many "MC absent" examples to be used at once practically.

D. Model Selection and SVM Training

Once the training examples are gathered, the next step is to determine the SVM decision function in (16). In this process, we must decide the following variables: the type of kernel function, its associated parameter, and the regularization parameter C in the structural risk function. To optimize these parameters, we applied m-fold cross validation [8] to the training-mammogram set. This procedure consists of the following steps. First, divide randomly all the available training examples into m equal-sized subsets. Second, for each model-parameter setting, train the SVM classifier m times; during each time one of the m subsets is held out in turn while all the rest of the subsets are used to train the SVM. The trained SVM classifier is then tested using the held-out subset, and its classification error is recorded. Third, the classification errors are averaged to obtain an estimate of the generalization error of the SVM classifier. In the end, the model with the smallest generalization error will be adopted. Its performance will be evaluated using FROC analysis (Section IV).

As explained in Section II, the training of the SVM classifier is accomplished by solving the quadratic optimization problem in (13). While in principle this can be done using any existing general-purpose quadratic programming software, it should be noted that the number of training examples (hence, the number of unknowns) used in this study is large (on the order of several thousand). Fortunately, numerically efficient algorithms have been developed for solving the SVM optimization problem [8]. These algorithms typically take advantage of the fact that most of the Lagrange multipliers in (13) are zero. In this paper, we adopted a technique called *successive minimal optimization* (SMO) [30]–[32]. The basic idea of this technique is to optimize the objective function in (13) iteratively over a pair of variables (i.e., two training samples) at a time. The solution can be found analytically for each pair, thus, faster convergence can be achieved. We found in this study that the SMO algorithm is typically five to ten times faster than a general-purpose quadratic optimization algorithm.

E. Insight on the SVM Classifier

Consider the SVM decision function in (16), which is expressed in terms of the support vectors $\mathbf{s}_j$, $j = 1, 2, \ldots, l_s$. Let l_s^1 denote the number of support vectors that belong to the "MC present" class and, for notational simplicity, let them be denoted in an ordered fashion as $\mathbf{s}_j, j = 1, 2, \ldots, l_s^1$. Then, we can rewrite $f(\mathbf{x})$ as

$$\begin{aligned} f(\mathbf{x}) &= \sum_{j=1}^{l_s^1} \alpha_j y_j K(\mathbf{s}_j, \mathbf{x}) + \sum_{j=l_s^1+1}^{l_s} \alpha_j y_j K(\mathbf{s}_j, \mathbf{x}) + b \\ &= \sum_{j=1}^{l_s^1} \alpha_j K(\mathbf{s}_j, \mathbf{x}) - \sum_{j=l_s^1+1}^{l_s} \alpha_j K(\mathbf{s}_j, \mathbf{x}) + b. \end{aligned} \quad (21)$$

Replacing $K(\cdot, \cdot)$ by the inner product of the mapping $\Phi(\cdot)$ in (12) and making use of the symmetry of the inner product, we obtain

$$f(\mathbf{x}) = \Phi^T(\mathbf{x}) \left[\sum_{j=1}^{l_s^1} \alpha_j \Phi(\mathbf{s}_j) - \sum_{j=l_s^1+1}^{l_s} \alpha_j \Phi(\mathbf{s}_j) \right] + b. \quad (22)$$

Defining

$$\Phi^* \equiv \sum_{j=1}^{l_s^1} \alpha_j \Phi(\mathbf{s}_j) - \sum_{j=l_s^1+1}^{ls} \alpha_j \Phi(\mathbf{s}_j) \quad (23)$$

we have

$$f(\mathbf{x}) = \Phi^T(\mathbf{x}) \Phi^* + b. \quad (24)$$

Note that, when expressed as in (24), the SVM decision function assumes the form of a template-matching detector in the nonlinear-transform space $\mathcal{H}$: the vector Φ^* can be viewed as a known template, against which the input pattern $\mathbf{x}$ is compared in the $\mathcal{H}$ space. A careful examination of the form of the template Φ^* provides further insight to the SVM classifier. The first sum in (23) is composed of support vectors from the "MC present" class, while the second sum consists of those from the "MC absent" class. Naturally, a large positive matching score is expected when an input pattern $\mathbf{x}$ is from the "MC present" class; similarly, a large but negative matching score is expected when $\mathbf{x}$ is from the "MC absent" class.

Furthermore, by definition, support vectors are those training examples found to be either on or near the decision boundaries of the decision function. In a sense, they consist of the "borderline," difficult-to-classify examples from each class. The SVM classifier then defines the decision boundary between the two classes by "memorizing" these support vectors. This in philosophy is quite different from a neural network, for example, that is based on minimization of MSE.

In an interesting study in [33], where a neural network was trained for MC detection, it was reported that better performance was achieved when the neural network was trained with a set of "difficult cases" (identified by human observers) than with the whole available data set. In our method, the "difficult cases" are automatically identified by the SVM during training.

F. Successive Enhancement Learning

The support vectors define the decision boundaries of the SVM classifier; therefore, it is essential that they well represent their respective classes. As mentioned earlier, in a mammogram there are vastly more examples available from the "MC absent" class than from the "MC present" class. Yet, in training only a small fraction of them can practically be used. As such, a potential concern is whether this fraction of randomly selected training samples can represent the "MC absent" class well.

To address this issue we propose an SEL scheme to make use of all the available "MC absent" examples. The basic idea is to select iteratively the "most representative" "MC absent" examples from all the available training images while keeping the total number of training examples small. Such a scheme improves the generalization ability of the trained SVM classifier (as shown experimentally in Section IV). The proposed algorithm is summarized below. A proof of convergence of the proposed algorithm is given in the Appendix.

SUCCESSIVE ENHANCEMENT–LEARNING ALGORITHM:

1. Extract an initial set of training examples from the available training images (e.g., through random selection). Let $Z = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_l, y_l)\}$ denote this resulting set of training examples.
2. Train the SVM classifier $f(\mathbf{x}) = \sum_{j=1}^{ls} \alpha_j y_j K(\mathbf{s}_j, \mathbf{x}) + b$ with Z.
3. Apply the resulting classifier $f(\mathbf{x})$ to all the mammogram regions (except those in Z) in the available training images and record the "MC absent" locations that have been misclassified as "MC present."
4. Gather N new input examples from the misclassified "MC absent" locations; update the set Z by replacing N "MC absent" examples that have been classified correctly by $f(\mathbf{x})$ with the newly collected "MC absent" examples.
5. Re-train the SVM classifier with the updated set Z.
6. Repeat steps 3–5 until convergence is achieved.

In Step 1, the training set size l is typically kept small for numerical efficiency. Consequently, the training examples represent only a small fraction of all the possible mammogram regions. The purpose of steps 3 and 4 is to identify those difficult "MC absent" examples in the training mammograms that were not included in the initial training set Z. In Step 4, there may be several ways for gathering the new "MC absent" examples. One is simply to select the N most-misclassified "MC absent" locations [i.e., those with the most positive values of $f(\mathbf{x})$]. This is referred to as the *greedy* approach. An alternative would be to select randomly among all those misclassified "MC absent" locations. In our studies, we experimented with both approaches. In Step 6, the numerical convergence of the algorithm is determined by monitoring the change in support vectors during each iteration.

IV. Performance Evaluation Study

A. Mammogram Data Set

We developed and tested the proposed algorithm using a data set collected by the Department of Radiology at The University of Chicago. This data set consists of 76 clinical mammograms, all containing multiple MCs. These mammograms are of dimension 1000×700 pixels, with a spatial resolution of 0.1 mm/pixel and 10-bit grayscale. Collectively, there are a total of 1120 MCs in these mammograms, which were identified by a group of experienced mammographers. These mammograms were obtained at The University of Chicago which are representative of cases that contain clustered MCs that are difficult to detect.

In this study, we divided the data set in a random fashion into two separate subsets, each of which consisted of 38 images. One of these subsets was used exclusively during the training phase of the proposed algorithm, and is hereafter designated as the *training-mammogram set*; the other subset was used exclusively during the testing phase, and is designated as the *test-mammogram set*. At no time was a test-set image used in any way in the training procedure, and *vice versa*.

B. Performance Evaluation Method

To summarize quantitatively the performance of the trained SVM classifier, we used FROC curves [34]. An FROC curve is a plot of the correct detection rate (i.e., true-positive fraction) achieved by a classifier versus the average number of false positives (FPs) per image varied over the continuum of the decision threshold. An FROC curve provides a comprehensive summary of the trade-off between detection sensitivity and specificity.

We constructed the FROC curves by the following procedure. First, the trained SVM classifier was applied with varying thresholds to classify each pixel in each test mammogram as "MC present" or "MC absent." Because several neighboring pixels may be part of an MC, it is necessary next to group the pixels classified as "MC present" to form MC objects. This was accomplished by a morphological processing procedure described in [15], where isolated spurious pixels were removed. Finally, MC clusters were identified by grouping the objects that have been determined by the algorithm to be MCs.

In our implementation, we adopted a criterion recommended by Kallergi *et al.* [35] for identifying MC clusters. Specifically, a group of objects classified as MCs is considered to be a true positive (TP) cluster only if: 1) the objects are connected with nearest-neighbor distances less than 0.2 cm; and 2) at least three true MCs should be detected by the algorithm within an area of 1 cm^2. Likewise, a group of objects classified as MCs is labeled as an FP cluster provided that the objects satisfy the cluster requirement but contain no true MCs. It was reported [35] that such a criterion yields more-realistic performance than several other alternatives.

It bears repeating here that, to ensure a realistic evaluation, the FROC curves in this study were all computed using only the test-mammogram set. As mentioned before, this set of 38 mammograms, chosen randomly, was held aside at the beginning of the study, and was never used by any of the training algorithms.

C. Other Methods for Comparison

For comparison purposes, the following four existing methods for MC detection were also considered in this study: 1) image difference technique (IDT) [15]; 2) DoG method [20]; 3) wavelet-decomposition (WD)-based method [17], [18]; and 4) a TMNN method [26]. We selected these because they are well-known methods that are representative of two main approaches that are widely used: template-matching techniques and learning-based methods.

The following is a summary of the parameter values we used when implementing the four methods for comparison. For the DoG method, the values of the kernel width (σ) used for the positive and negative Gaussian kernels were 0.75 and 4, respectively. The weight associated with the positive kernel was 0.8. For the WD method, four-octave decomposition was used where an additional voice was inserted between octaves 2 and 3, and one between octaves 3 and 4. For the TMNN method, a three-layer feed-forward neural network with six neurons in the hidden layer was used in the first stage; and another three-layer feed-forward neural network with eight neurons in the hidden layer was used for the second stage. The 15-component feature vector described in [26] was used.

While it was nearly impossible to obtain the globally optimal parametric setting for each algorithm, care was taken in our implementation so that it is as faithful to its original description in the literature as possible. Whenever feasible, these methods were typically run under multiple parameter settings and the one yielding the best results was chosen for the final test.

A final note is that both the WD and TMNN methods are learning-based, thus training was required. The same training-mammogram set was used for these methods as for the proposed SVM method. All the methods were evaluated using the same test-mammogram set.

V. Experimental Results

A. SVM Training and Model Selection

The training-mammogram set contained 547 MCs. Consequently, 547 examples were gathered for the "MC present" class from this set of mammograms. In addition, twice as many "MC absent" examples were selected by random sampling from these mammograms. Thus, there were 1641 training examples in total. A tenfold cross-validation procedure was used for training and testing the SVM classifier under various model and parametric settings.

We also experimented with using an increased number of "MC absent" examples in training (e.g., up to five times more than the number of MC examples), but no significant improvement was observed in the generalization error of the resulting SVM classifier. We believe this is largely due to the redundancy among the vast collection of "MC absent" examples. This partly motivated our proposed SEL training scheme for the SVM classifier. In this regard, the SEL is an informed scheme for selecting the "MC absent" samples for training, making use of both the current state of the SVM classifier in training and all the available "MC absent" samples.

In our evaluations, we used generalization error as a figure of merit. Generalization error was defined as the total number of incorrectly classified examples divided by the total number of examples classified. Generalization error was computed using only those examples held-out during training.

In Fig. 4(a), we summarize the results for the trained SVM classifier when a polynomial kernel was used. The estimated generalization error is plotted versus the regularization parameter C for kernel order $p = 2$ and $p = 3$. Similarly, in Fig. 4(b) we summarize the results when the Gaussian RBF kernel was used; here, the estimated generalization error is plotted for different values of the width σ(2.5, 5, and 10).

For the polynomial kernel, we found that the best error level is achieved when $p = 3$ and C is between 1 and 10; interestingly, a similar error level was also achieved by the Gaussian RBF kernel over a wide range of parameter settings (e.g., when $\sigma = 5$ and C is in the range of 100–1000). These results indicate that the performance of the SVM classifier is not very sensitive to the values of the model parameters. Indeed, essentially similar performance was achieved when σ was varied from 2.5 to 5.

Having determined that the SVM results do not vary significantly over a wide range of parameter settings, we will focus for the remainder of the paper on a particular, representative configuration of the SVM classifier, having a Gaussian RBF kernel with $\sigma = 5$ and $C = 1000$.

Some insight about the SVM classifier can be gained by looking at the support vectors produced by the training procedure. The number of support vectors in the representative case that we studied was approximately 12% of the total number of training examples and the training time is around 7s (implemented in MATLAB on a Pentium III 933-MHz PC). Fig. 5 shows some examples of the support vectors obtained for both "MC present" and "MC absent" image windows. For comparison, some randomly selected examples from the training set are also shown. Note that, as expected, some of the support vectors indeed appear to be the difficult-to-classify, "borderline" cases; i.e., the "MC present" support vectors are MCs that could be mistaken for background regions, and the "MC absent" support vectors are background regions from the training set that look like MCs.

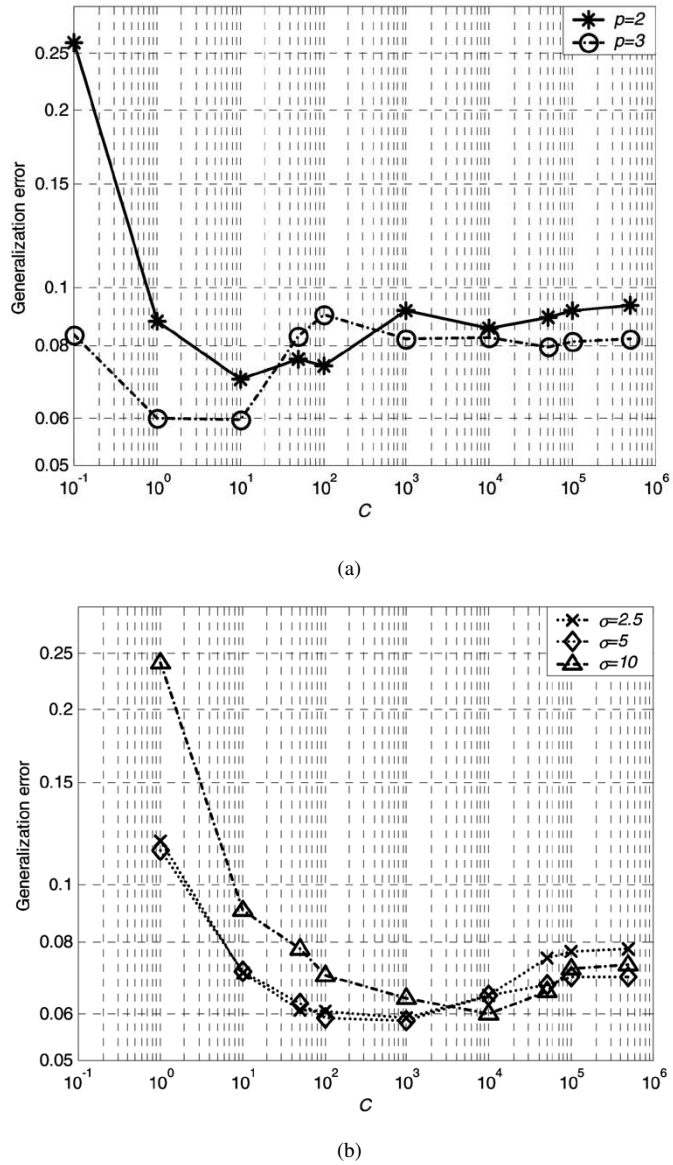

Fig. 4. Plot of generalization error rate versus regularization parameter C achieved by trained SVM classifiers using (a) a polynomial kernel with orders two and three and (b) a Gaussian RBF kernel with width $\sigma = 2.5, 5$, and 10.

B. Effect of Successive Enhancement Learning

The SVM classifier (with the representative parameters described previously) was then further trained using the proposed SEL scheme on the training mammogram set. For this purpose, a total of additional 50 000 nonoverlapping, "MC absent" sample windows were randomly selected from the training-mammogram set. Collectively these samples together with the previous 1641 training samples cover as much as 15% of the total training-mammogram areas. The proposed SEL scheme was then applied with this set of 50 000 samples. Note that this slightly deviates from the original description of the SEL scheme in that only a subset of the mammogram background areas (rather than all the mammogram regions) were used. We find this is sufficient to demonstrate the effect of the SEL scheme. For testing the resulting trained SVM, 5000 additional nonoverlapping, "MC absent" samples were randomly selected from the remaining mammogram areas of the training-mammogram set. These 5000 samples were then used to compute the generalization error rate of the trained SVM classifier with SEL. Both the greedy approach and random selection were tested. Up to $N = 50$ misclassified "MC absent" samples were selected during each iteration.

In Fig. 6, we show a plot of the generalization error rate achieved by the trained SVM classifier for the first nine iterations. Note that in both cases there is a significant drop in the generalization error rate after the first two iterations, and diminishing gain from subsequent iterations. We believe this indicates that most of the "difficult" "MC absent" examples were effectively selected by the proposed SEL scheme during the first two iterations. Also, note that the random SEL approach outperformed the greedy method in Fig. 6. This is possibly due to the fact that the latter always selects the most misclassified samples during each iteration, which may not necessarily be most representative of the "MC absent" class; on the other hand,

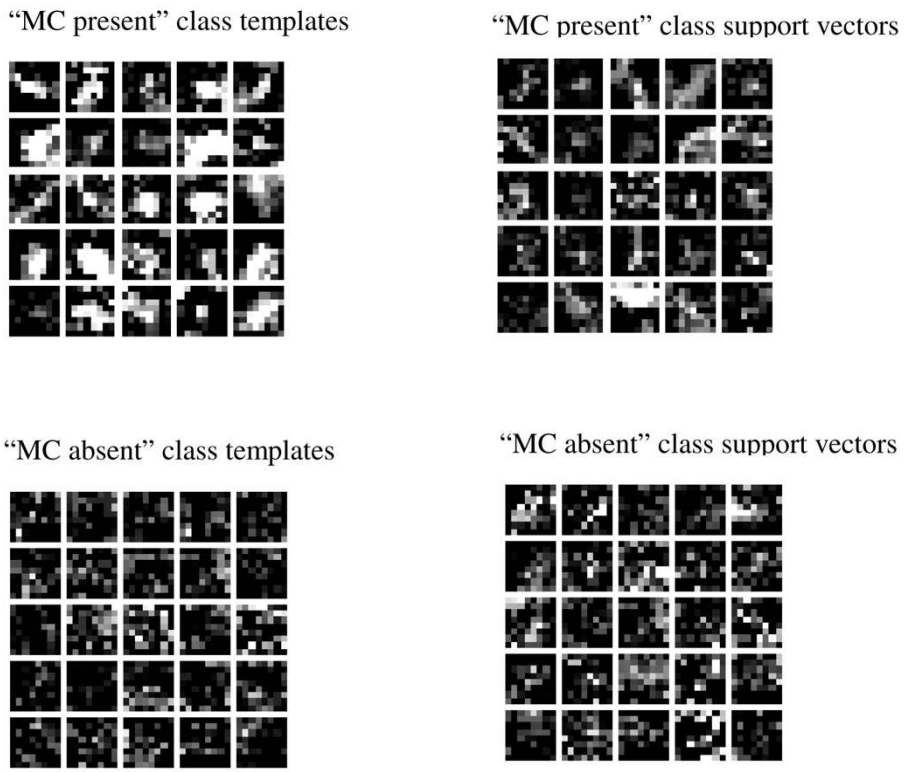

Fig. 5. Examples of 9×9 image windows and support vectors. Image windows with and without MCs are shown at top-left and bottom-left, respectively. Support vectors representing the "MC present" and "MC absent" classes of image windows are shown at top-right and bottom-right, respectively. Note that the SVs represent the borderline examples from each class that are difficult to categorize ("MC present" SVs could be mistaken for "MC absent" image regions; "MC absent" SVs might be mistaken for MCs. The support vectors shown are for the case of a SVM with Gaussian kernel ($\sigma = 5$, and $C = 1000$).

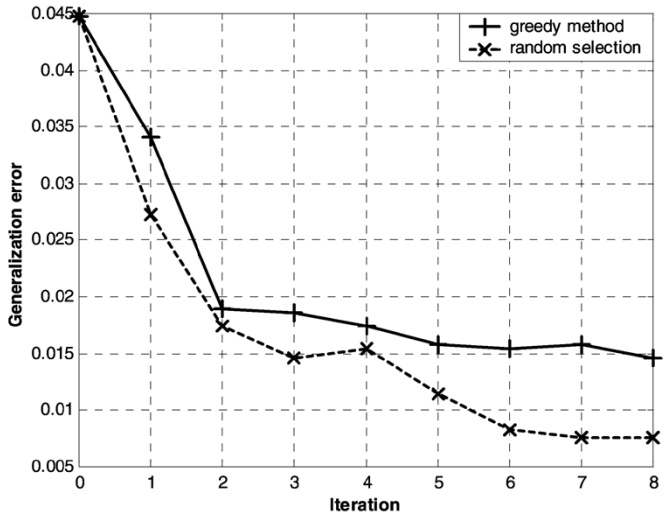

Fig. 6. Plot of generalization error rate of the trained SVM classifier using SEL versus the number of iterations.

the random approach selects samples from all the misclassified samples, leading to the possibility of selecting more-representative samples as the iterations progress. This random SEL trained SVM was used in the rest of the evaluation study.

C. Performance Evaluation

The performance of the proposed SVM approach, along with the other methods, is summarized by the FROC curves in Fig. 7. As can be seen, the SVM classifier offers the best detection result, and is improved by the proposed SEL scheme. The SVM achieves a sensitivity of approximately 85% when the false-positive (FP) rate is at an average of one FP cluster per image.

The FROC results obtained here for WD and IDT filteringare very similar to those described in the original reports of these methods [15], [17], [18]. For the DoG method (for which no FROC information is given in its original report), the detection rate is close to that of the IDTF when the FP rate is around two FP clusters per image. This is not surprising because both methods operate under a similar principle (the detection kernels in both cases behave like a bandpass filter). In addition, the FROC results indicate that the TMNN method outperforms the other three methods we compared (WD, IDTF, and DoG) when the FP rate is above one FP cluster per image. The numerical FROC results we obtained for the TMNN are somewhat different from those in its original report. There are several possible explanations: 1) the mammogram set used was different; 2) the detection criterion for MC clusters used in performance evaluation was different; and 3) in the original work [26] the MC clusters used for training were also included in testing.

In Fig. 8, we demonstrate that the method of defining MC clusters has an influence on the FROC curves, making it difficult to compare reported results in the literature that were derived using various criteria. The results in Fig. 8, which differ from those in Fig. 7, were obtained when the nearest-neighbor-distance threshold for MC cluster detection was increased from 0.2 cm to 0.3 cm. In particular, the sensitivity of the SVM approach increased to nearly 90% at

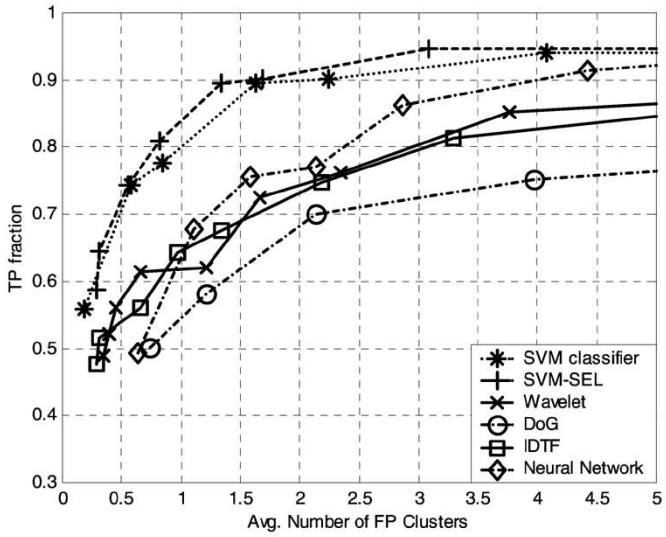

Fig. 7. FROC comparison of the methods tested. A higher FROC curve indicates better performance. The best performance was obtained by a successive learning SVM classifier, which achieves around 85% detection rate at a cost of one FP cluster per image. The nearest neighbor distance threshold used for cluster detection is 0.2 cm.

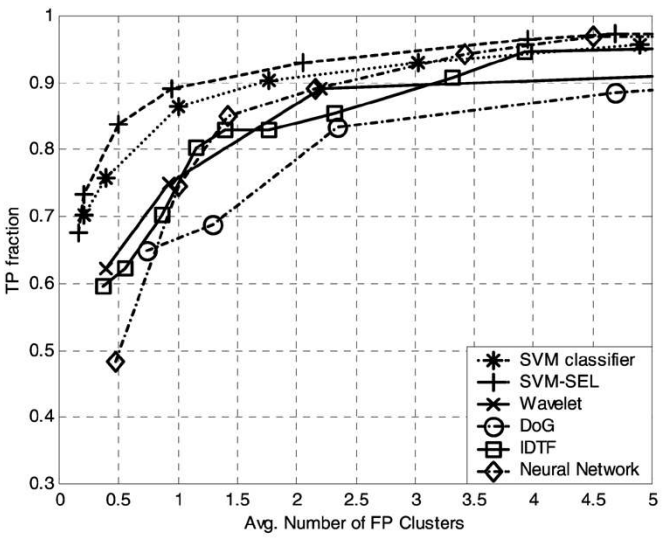

Fig. 8. FROC curves of the methods tested. The best performance was obtained by a successive learning SVM classifier, which achieves around 90% detection rate at a cost of one FP cluster per image. The nearest neighbor distance threshold used for cluster detection is 0.3 cm.

an FP rate of one FP cluster per image. Similarly, when the nearest-neighbor-distance threshold is increased further to 0.5 cm, the sensitivity of the SVM approach increased to as high as 94% while the FP rate remains at one FP cluster per image. The FROC curves in this case are shown in Fig. 9. Note that, while different criteria may affect the numerical FROC results, the relative ordering of performance of the methods is preserved.

VI. Conclusion

In this paper, we proposed the use of an SVM for detection of MCs in digital mammograms. In the proposed method, an SVM classifier was trained through supervised learning to test at every location in a mammogram whether an MC is present or not. The formulation of SVM learning is based on the principle of structural risk minimization. The decision function of the trained SVM classifier is determined in terms of support vectors that were identified from the examples during training. The result is that the SVM classifier achieves low generalization error when applied to classify samples that were not included in training. In addition, the proposed SEL scheme can further lead to improvement in the performance of the trained SVM classifier. Experimental results using a set of 76 clinical mammograms demonstrate that the proposed framework is very insensitive to the choice of several model parameters. In our experiments, FROC curves indicated that the SVM approach yielded the best performance when compared to several existing methods, owing to the better generalization performance by the SVM classifier.

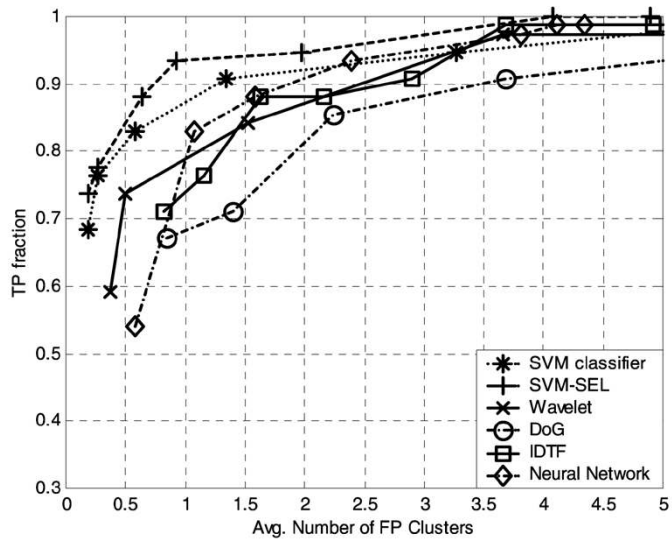

Fig. 9. FROC curves of the methods tested. The best performance was obtained by a successive learning SVM classifier, which achieves around 94% detection rate at a cost of one FP cluster per image. The nearest neighbor distance threshold used for cluster detection is 0.5 cm.

Appendix
Proof of the Successive Enhancement Learning Algorithm

In this section, we provide a proof for the convergence of the proposed successive enhancement learning (SEL) algorithm. This proof follows a similar approach to one given by Osuna *et al.* [5] for a decomposition strategy for SVM training with a large data set. Here, we apply it to prove convergence of the proposed SEL algorithm.

Let $Z = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_l, y_l)\}$ denote a subset of the training examples, and let $B = \{(\mathbf{x}_{l+1}, y_{l+1}), (\mathbf{x}_{l+2}, y_{l+2}), \ldots, (\mathbf{x}_L, y_L)\}$ denote the remainder of the training set so that the entire training set is represented by $Z_T = Z \cup B$.

Thus, the original dual problem in (13) can be extended as follows:

$$\max W_T(\alpha_1, \alpha_2, \ldots, \alpha_l, \alpha_{l+1}, \ldots, \alpha_L) = \sum_{i=1}^{L} \alpha_i - \tfrac{1}{2} \sum_{i=1}^{L} \sum_{j=1}^{L} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \quad \text{(A-1)}$$

subject to

$$0 \le \alpha_i \le C \qquad \text{for } i = 1, 2, \ldots, L; \quad \text{(A-2)}$$

and

$$\sum_{i=1}^{L} \alpha_i y_i = 0. \quad \text{(A-3)}$$

Observe that the original problem in (13) now becomes only a subproblem of (A-1). Indeed, let $(\alpha_1^*, \alpha_2^*, \ldots, \alpha_l^*)$ denote an optimal solution to (13), i.e., solution of training the SVM with subset Z. Let α^* denote the vector formed by $(\alpha_1^*, \alpha_2^*, \ldots, \alpha_l^*, \alpha_{l+1}^*, \ldots, \alpha_L^*)$, with $\alpha_j^* = 0$ for $j = l+1, \ldots, L$. Then α^* automatically satisfies both the constraints in (A-2) and (A-3) and, thus, is a feasible solution to (A-1).

Let $\mathbf{x}_s$ denote a margin support vector from the "MC absent" class obtained when the SVM is trained with Z, that is, $0 < \alpha_s^* < C$ and $f(\mathbf{x}_s) = -1$. In addition, let E denote the index set of those N examples in B that have been selected to update the training set Z. Note that these N examples have been misclassified by the trained $f(\mathbf{x})$.

Let δ_s be a positive constant such that $\alpha_s^* - \delta_s > 0$. Now consider a vector $\hat{\alpha} = (\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_l, \hat{\alpha}_{l+1}, \ldots, \hat{\alpha}_L)$ having components

$$\hat{\alpha}_j = \begin{cases} \alpha_s^* - \delta_s, & j = s \\ \dfrac{\delta_s}{N}, & j \in E \\ \alpha_j^*, & \text{otherwise.} \end{cases} \quad \text{(A-4)}$$

Then

$$W_T(\hat{\alpha}) - W_T(\alpha^*) = \sum_{i=1}^{L} \hat{\alpha}_i - \sum_{i=1}^{L} \alpha_i^* - \tfrac{1}{2} \sum_{i=1}^{L} \sum_{j=1}^{L} \hat{\alpha}_i \hat{\alpha}_j y_i y_j \times K(\mathbf{x}_i, \mathbf{x}_j) + \tfrac{1}{2} \sum_{i=1}^{L} \sum_{j=1}^{L} \alpha_i^* \alpha_j^* y_i y_j K(\mathbf{x}_i, \mathbf{x}_j).$$

From (A-4), we have $\sum_{i=1}^{L} \hat{\alpha}_i = \sum_{i=1}^{L} \alpha_i^*$ and, thus

$$W_T(\hat{\alpha}) - W_T(\alpha^*) = -\tfrac{1}{2} \sum_{i=1}^{L} \sum_{j=1}^{L} (\hat{\alpha}_i \hat{\alpha}_j - \alpha_i^* \alpha_j^*) y_i y_j K(\mathbf{x}_i, \mathbf{x}_j). \quad \text{(A-5)}$$

Let

$$S_1 = \sum_{i=1}^{l} \sum_{j=1}^{l} (\hat{\alpha}_i \hat{\alpha}_j - \alpha_i^* \alpha_j^*) y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \quad \text{(A-6)}$$

$$S_2 = \sum_{i=l+1}^{L} \sum_{j=1}^{l} (\hat{\alpha}_i \hat{\alpha}_j - \alpha_i^* \alpha_j^*) y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \quad \text{(A-7)}$$

and

$$S_3 = \sum_{i=l+1}^{L} \sum_{j=l+1}^{L} (\hat{\alpha}_i \hat{\alpha}_j - \alpha_i^* \alpha_j^*) y_i y_j K(\mathbf{x}_i, \mathbf{x}_j). \quad \text{(A-8)}$$

Noting that $K(\mathbf{x}_i, \mathbf{x}_j)$ is symmetric, we have

$$W_T(\hat{\alpha}) - W_T(\alpha^*) = -\tfrac{1}{2}(S_1 + 2S_2 + S_3). \quad \text{(A-9)}$$

Furthermore, since $y_s = f(\mathbf{x}_s) = -1$, we have

$$S_1 = (\delta_s)^2 K(\mathbf{x}_s, \mathbf{x}_s) - 2\delta_s \sum_{i=1}^{l} \alpha_i^* y_i K(\mathbf{x}_i, \mathbf{x}_s) = (\delta_s)^2 K(\mathbf{x}_s, \mathbf{x}_s) - 2\delta_s[-1 - b]. \quad \text{(A-10)}$$

Noting that $\hat{\alpha}_i = \delta_s / N$ and $y_i = -1$ for $i \in E$, we obtain

$$\begin{aligned} S_2 &= \sum_{i \in E} \sum_{j=1}^{l} (\hat{\alpha}_i \hat{\alpha}_j) y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \\ &= -\frac{\delta_s}{N} \sum_{i \in E} \sum_{j=1}^{l} \hat{\alpha}_j y_j K(\mathbf{x}_i, \mathbf{x}_j) \\ &= -\frac{\delta_s}{N} \sum_{i \in E} \left[\sum_{j=1}^{l} \alpha_j^* y_j K(\mathbf{x}_i, \mathbf{x}_j) + \delta_s K(\mathbf{x}_i, \mathbf{x}_s) \right] \\ &= -\frac{\delta_s}{N} \sum_{i \in E} [f(\mathbf{x}_i) - b + \delta_s K(\mathbf{x}_i, \mathbf{x}_s)] \\ &= -\frac{(\delta_s)^2}{N} \sum_{i \in E} K(\mathbf{x}_i, \mathbf{x}_s) + \frac{\delta_s}{N} \sum_{i \in E} f(\mathbf{x}_i) + b\delta_s \end{aligned} \quad \text{(A-11)}$$

and

$$S_3 = \left(\frac{\delta_s}{N} \right)^2 \sum_{i \in E} \sum_{j \in E} K(\mathbf{x}_i, \mathbf{x}_j). \quad \text{(A-12)}$$

Therefore

$$\begin{aligned} W_T(\hat{\alpha}) - W_T(\alpha^*) &= \delta_s \left[1 + \frac{1}{N} \sum_{i \in E} f(\mathbf{x}_i) \right] - \frac{(\delta_s)^2}{2} \\ &\quad \times \left[K(\mathbf{x}_s, \mathbf{x}_s) - \frac{2}{N} \sum_{i \in E} K(\mathbf{x}_i, \mathbf{x}_s) + \frac{1}{N^2} \sum_{i \in E} \sum_{j \in E} K(\mathbf{x}_i, \mathbf{x}_j) \right]. \end{aligned} \quad \text{(A-13)}$$

When δ_s is chosen sufficiently small, the second-order term in (A-13) is negligible and, thus

$$W_T(\hat{\alpha}) - W_T(\alpha^*) \approx \delta_s \left[1 + \frac{1}{N} \sum_{i \in E} f(\mathbf{x}_i) \right]. \quad \text{(A-14)}$$

By selection, we have $f(\mathbf{x}_i) > 0$ for $i \in E$. Thus, $W_T(\hat{\alpha}) - W_T(\alpha^*) > 0$. Therefore, the extended objective function in (A-1) can be further improved by training the SVM with the newly updated set Z. A successive application of this procedure will eventually lead to an optimal solution of (A-1), which implies that the generalization error of the trained SVM will also be improved.

This proof also shows that, when retrained with the updated set Z, a reasonable choice of the starting point for the optimization algorithm is α^*.

ACKNOWLEDGMENT

N. P. Galatsanos acknowledges the fruitful discussions on SVM with Prof. S. Theodoridis and N. Kaloupsidis at the Department of Informatics, the University of Athens, Athens, Greece.

REFERENCES

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

[2] B. Scholkopf, S. Kah-Kay, C. J. Burges, F. Girosi, P. Niyogi, T. Poggio, and V. Vapnik, "Comparing support vector machines with Gaussian kernels to radial basis function classifiers," *IEEE Trans. Signal Processing*, vol. 45, pp. 2758–2765, Nov. 1997.

[3] M. Pontil and A. Verri, "Support vector machines for 3-D object recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, pp. 637–646, June 1998.

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

[5] 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.

[6] T. Joachims, "Transductive inference for text classification using support vector machines," presented at the *Int. Conf. Machine Learning*, Slovenia, June 1999.

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

[8] K. R. Muller, S. Mika, G. Ratsch, K. Tsuda, and B. Scholkopf, "An introduction to kernel-based learning algorithms," *IEEE Trans. Neural Networks*, vol. 12, pp. 181–201, Mar. 2001.

[9] M. N. Wernick, "Pattern classification by convex analysis," *J. Opt. Soc. Amer. A*, vol. 8, pp. 1874–1880, 1991.

[10] *Cancer Facts and Figures 1998*. Atlanta, GA: American Cancer Society, 1998.

[11] M. Lanyi, *Diagnosis and Differential Diagnosis of Breast Calcifications*. Berlin, Germany: Springer-Verlag, 1988.

[12] R. M. Nishikawa, "Detection of microcalcifications," in *Image-Processing Techniques for Tumor Detection*, R. N. Strickland, Ed. New York: Marcel Dekker, 2002.

[13] J. Roehrig, T. Doi, A. Hasegawa, B. Hunt, J. Marshall, H. Romsdahl, A. Schneider, R. Sharbaugh, and W. Zhang, "Clinical results with R2 Imagechecker system," in *Digital Mammography*, N. Karssemeijer, M. Thijssen, J. Hendriks, and L. van Erning, Eds. Boston, MA: Kluwer Academic, 1998, pp. 395–400.

[14] N. Karssemeijer, "A stochastic model for automated detection calcifications in digital mammograms," in *Proc. 12th Int. Conf. Information Medical Imaging*, Wye, U.K., July 1991, pp. 227–238.

[15] R. M. Nishikawa, M. L. Giger, K. Doi, C. J. Vyborny, and R. A. Schmidt, "Computer aided detection of clustered microcalcifications in digital mammograms," *Med. Biol. Eng. Compu.*, vol. 33, pp. 174–178, 1995.

[16] H. Yoshida, K. Doi, and R. M. Nishikawa, "Automated detection of clustered microcalcifications," in *Digital Mammograms Using Wavelet Transform Techniques, Medical Imaging*. Bellingham, WA: SPIE (Int. Soc. Opt. Eng.), 1994, pp. 868–886.

[17] R. N. Strickland and H. L. Hahn, "Wavelet transforms for detecting microcalcifications in mammograms," *IEEE Trans. Med. Imag.*, vol. 15, pp. 218–229, Apr. 1996.

[18] ——, "Wavelet transforms methods for object detection and recovery," *IEEE Trans. Image Processing*, vol. 6, pp. 724–735, May 1997.

[19] T. Netsch, "A scale-space approach for the detection of clustered microcalcifications in digital mammograms," in *Digital Mammography, Proc. 3rd Int. Workshop Digital Mammography*, Chicago, IL, 1996, pp. 301–306.

[20] J. Dengler, S. Behrens, and J. F. Desaga, "Segmentation of microcalcifications in mammograms," *IEEE Trans. Med. Imag.*, vol. 12, pp. 634–642, Dec. 1993.

[21] M. N. Gurcan, Y. Yardimci, A. E. Cetin, and R. Ansari, "Detection of microcalcifications in mammograms using higher order statistics," *IEEE Signal Processing Letters*, vol. 4, pp. 213–216, Aug. 1997.

[22] H. Cheng, Y. M. Liu, and R. I. Freimanis, "A novel approach to microcalcifications detection using fuzzy logic techniques," *IEEE Trans. Med. Imag.*, vol. 17, pp. 442–450, June 1998.

[23] P. A. Pfrench, J. R. Zeidler, and W. H. Ku, "Enhanced detectability of small objects in correlated clutter using an improved 2-D adaptive lattice algorithm," *IEEE Trans. Image Processing*, vol. 6, pp. 383–397, Mar. 1997.

[24] H. Li, K. J. Liu, and S. Lo, "Fractal modeling and segmentation for the enhancement of microcalcifications in digital mammograms," *IEEE Trans. Med. Imag.*, vol. 16, pp. 785–798, Dec. 1997.

[25] N. Bankman, T. Nizialek, I. Simon, O. Gatewood, I. N. Weinberg, and W. R. Brody, "Segmentation algorithms for detecting microcalcifications in mammograms," *IEEE Trans. Inform. Technology in Biomedicine*, vol. 1, pp. 141–149, June 1997.

[26] S. Yu and L. Guan, "A CAD system for the automatic detection of clustered microcalcifications in digitized mammogram films," *IEEE Trans. Med. Imag.*, vol. 19, pp. 115–126, Feb. 2000.

[27] A. Bazzani, A. Bevilacqua, D. Bollini, R. Brancaccio, R. Campanini, N. Lanconelli, A. Riccardi, and D. Romani, "An SVM classifier to separate false signals from microcalcifications in digital mammograms," *Phys. Med. Biol.*, vol. 46, pp. 1651–1663, 2001.

[28] B. Scholkopf, C. Burges, and A. Smola, *Advances in Kernel Methods: Support Vector Learning*. Cambridge, MA: MIT Press, 1999.

[29] S. Haykin, *Neural Networks: A Comprehensive Foundation*, 2nd ed. Englewood Cliffs, NJ: Prentice-Hall, 1999.

[30] C. Chang, C. Hsu, and C. Lin, "The analysis of decomposition methods for support vector machines," *IEEE Trans. Neural Networks*, vol. 11, pp. 1003–1008, July 2000.

[31] J. Platt, "Fast training of support vector machines using sequential minimal optimization," in *Advances in Kernel Methods: Support Vector Learning*, B. Schölkopf, C. Burges, and A. J. Smola, Eds. Cambridge, MA: MIT Press, 1999, pp. 185–208.

[32] S. Keerthi, S. Shevade, C. Bhattacharyya, and K. Murthy, "Improvements to Platt's SMO algorithm for SVM classifier design," *Neural Computation*, vol. 13, pp. 637–649, Mar. 2001.

[33] R. H. Nagel, R. M. Nishikawa, and K. Doi, "Analysis of methods for reducing false positives in the automated detection of clustered microcalcifications in mammograms," *Med. Phys.*, vol. 25, no. 8, pp. 1502–1506, 1998.

[34] P. C. Bunch, J. F. Hamilton, G. K. Sanderson, and A. H. Simons, "A free-response approach to the measurement and characterization of radiographic-observer performance," *J. Appl. Eng.*, vol. 4, 1978.

[35] M. Kallergi, G. M. Carney, and J. Garviria, "Evaluating the performance of detection algorithms in digital mammography," *Med. Phys.*, vol. 26, no. 2, pp. 267–275, 1999.