Retrieval boosted computer-aided diagnosis of clustered microcalcifications for breast cancer

Hao Jing and Yongyi Yang^{a)}

Department of Electrical and Computer Engineering, Illinois Institute of Technology, 3301 South Dearborn Street, Chicago, Illinois 60616

Robert M. Nishikawa^{b)}

Department of Radiology, The University of Chicago, 5841 S. Maryland Avenue, Chicago, Illinois 60637-1463

(Received 13 September 2011; revised 14 December 2011; accepted for publication 16 December 2011; published 12 January 2012)

Purpose: The authors propose an image-retrieval based approach for case-adaptive classifier design in computer-aided diagnosis (CADx). The conventional approach in CADx is to first train a pattern-classifier based on a set of existing training samples and then apply this classifier to subsequent new cases. The purpose of this work is to improve the classification accuracy of a CADx classifier by making use of a set of known cases retrieved from a reference library that are similar to the case under consideration.

Methods: In the proposed approach, the authors will first apply image-retrieval to obtain a set of lesion images from a library of known cases that have similar image features to a case being diagnosed (i.e., query). These retrieved cases are then used to optimize a pattern-classifier toward boosting its classification accuracy on the query case. The basic idea is to put more emphasis on those cases that are similar to the query. The proposed approach is demonstrated first using a linear classifier and then extended to a nonlinear classifier induced by kernel principal component analysis.

Results: The proposed retrieval-driven approach was tested on a library of mammogram images from 1006 cases (646 benign and 360 malignant) obtained from multiple institutions and was demonstrated to yield significant improvement in classification performance. Measured by the area under the receiver operating characteristic curve (AUC), the case-adaptive approach could boost the classification performance of a linear classifier from $AUC = 0.7415$ to $AUC = 0.7807$; similar improvement was also obtained for a nonlinear classifier, with AUC boosted from 0.7527 to 0.7838.

Conclusions: Use of additional cases from a reference library that have similar image features can improve the classification accuracy of a CADx classifier on a query case. It can even outperform retraining the classifier with all the cases from the entire reference library. This implies that cases with similar image features are more relevant in defining the local decision boundary of the CADx classifier around the query. © 2012 American Association of Physicists in Medicine. [DOI: 10.1118/1.3675600]

Key words: clustered microcalcifications, computer-aided diagnosis, kernel principal component analysis, logistic regression

I. INTRODUCTION

Breast cancer is presently a common form of cancer diagnosed in women.¹ One of the important early signs of breast cancer in mammograms is the appearance of clustered microcalcifications (MC), which appear in 30%–50% of mammographically diagnosed cases. MCs are calcium deposits of very small dimension and appear as a group of granular bright spots in a mammogram (as illustrated in Fig. 1). Because of their subtlety in appearance in mammogram images, accurate diagnosis of microcalcification lesions as benign or malignant is a very challenging clinical problem for radiologists.² Studies show that a false-positive diagnostic imaging study leads to unnecessary biopsy of benign lesions, yielding a positive predictive value of only 20%–40%.²

Because of their importance in cancer diagnosis, there has been significant research in development of computer-

aided diagnosis (CADx) techniques for clustered MCs, the goal of which is to provide a second opinion to radiologists in their diagnosis to improve their performance and efficiency.^{2–4} Laboratory observer studies have shown that with CADx radiologists can improve their biopsy recommendation by sending more cancer cases and fewer benign cases to biopsy.^{3–6}

In CADx, a pattern classifier is used to predict the likelihood that a given lesion is malignant or benign (i.e., two-class classification) based on features extracted from the lesion. This classifier is typically trained on a set of existing, known cases (called training samples), which is subsequently applied to unknown cases (i.e., testing samples). For this purpose, many different machine learning methods have been studied. Here, we only briefly review some representative methods in terms of their underlying learning algorithms used and refer the interested reader to several recent review

FIG. 1. A mammogram image in CC view (left) and clustered microcalcifications in magnified view (right).

articles on this topic.^{7–10} Perhaps owing to its early success, a widely used technique by different researchers is artificial neural networks (ANN).^{11–17} In these methods, an ANN classifier is trained to discriminate between malignant and benign MC lesions using various features ranging from standard BI-RADS lexicon¹² to more quantitative image features. For example, Jiang *et al.* demonstrated an ANN algorithm to be even more accurate than radiologists in diagnosis of calcification lesions.¹⁴ Compared with ANN, linear discriminant analysis (LDA) is also widely used,^{18–22} which is more efficient to train. As in ANN, these methods also use a variety of lesion features, including shape features,¹⁸ morphological and texture features,¹⁹ gray-level image structure features,²⁰ radiologist-extracted features,²¹ and BI-RADS descriptors.²² Besides ANN and LDA, many other classifiers have also been investigated, such as the kernel Fisher discriminant (KFD), support vector machines (SVM), and relevance vector machines (RVM).²³ More recently, a classifier based on logistic regression (LR) was developed by Chhatwal *et al.*,²⁴ and a Bayesian network approach was investigated by Burnside *et al.*²⁵

In this work, we investigate a case-adaptive approach to boost the performance of a CADx classifier based on retrieval of cases with similar image features from a reference library. As illustrated in a diagram in Fig. 2, in traditional CADx, a classifier $f(\mathbf{x})$ is first optimized on a set of training samples, which is often limited in size; subsequently, this classifier is applied to classify a new case $\mathbf{x}$ under consideration (called query). In our approach, we will first obtain from a reference database (e.g., a PACS system) a set of known cases with similar features to the query case $\mathbf{x}$, and use these retrieved cases to adapt the classifier $f(\mathbf{x})$ so as to improve its classification accuracy on the query case. In recent years, content-based image retrieval (CBIR) has been studied as an alternative aid in diagnostic imaging,^{26–28} the goal of which is to provide radiologists with examples of lesions with known pathology that are similar to the lesion being evaluated (to boost their diagnostic accuracy). Our approach is similar in spirit, except that we will use retrieved cases to boost the performance of a CADx classifier instead of a human observer.

FIG. 2. Illustration of proposed adaptive approach for classification: for a case to be classified, a set of similar cases (retrieved from a reference library) is used to boost the classification accuracy of a baseline classifier (shown in dashed rectangle) on the case.

We note that this concept of using similar cases to boost the performance of a CADx classifier was first explored in our previous work,^{29,30} where a baseline SVM classifier was adjusted according to its classification accuracy on those images in the training set that are perceptually similar to the query image. For demonstration purpose, only a small dataset (consisting of 104 cases) was used in this preliminary study, and the MCs in the dataset were manually identified by experts. Also, for each query, the SVM classifier was optimized through a leave-one-out (LOO) procedure on the set of similar images, which becomes computationally expensive when the number of retrieved images is increased.

Based on this prior success, in this work, we further develop this case-adaptive CADx approach by making use of similar cases obtained from outside the training set. The idea is to take advantage of the existence of large medical databases such as modern PACS. To be realistic, we will make use of a large number of MC cases (1006) collected from multiple institutions and digitized with different scanners. This will allow us to demonstrate the proposed approach in the presence of heterogeneity across different databases. Such heterogeneity is known to lead to varied performance by a CAD system on cross-institutional databases.¹³ In addition, we will use image features derived from MCs detected by a computer detection algorithm (instead of manual identification). This is essential toward making the developed approach a useful clinical tool, because an MC cluster can have from a few to more than 100 MCs and it is impractical for radiologists to identify them all manually in a clinical environment. Of course, a computer detection technique will inevitably lead to both missed MCs and false positives. However, it has been demonstrated that existence of even moderate inaccuracy in the detection of MCs has little or no impact on the accuracy of a CADx classifier.^{31,32} Furthermore, to avoid the associated computational burden of the approach in our previous work,^{29,30} the case-adaptive approach in this work is first developed on a linear classifier based on logistic

regression, which is efficient to train and does not require a LOO optimization procedure for each query. This approach is then extended to a nonlinear classifier using the technique of kernel principal component analysis (KPCA) (Ref. 41) at little extra computational cost.

As noted above, CBIR has been studied recently by researchers as a useful tool for exploring known cases from a reference library that can assist radiologists in diagnosis. The idea is to provide evidence for case-based reasoning with information from the retrieved cases.³³ As an indication of the predictive value of retrieved cases, the correlation in disease condition between the query and the retrieved cases was examined in our previous work.^{34,35} The ratio of malignant cases among all retrieved cases was used as a useful predictor for the query;^{36,37} conceptually, this can be viewed as a K -nearest neighbor (KNN) classifier using only the retrieved cases. The similarity level between a retrieved case and the query was used by Zheng *et al.* as a weighting factor in the prediction.³⁸ A genetic algorithm was used by Mazurowski *et al.* to adjust the weighting factors of the retrieved cases.³⁹ An observer study was used by Nakayama *et al.* to investigate the potential diagnostic value of similar cases.⁴⁰ These studies provide evidence on the positive predictive value of similar cases in CADx, which we exploit to boost a CADx classifier in this study.

Finally, we mention in passing that there also exists a sizeable literature on studying features for classification of clustered microcalcifications.^{10,20,21,46} In this study, we will make use of a set of features reported in the literature, as our main focus is on how to improve the performance of an existing CADx classifier.

The rest of the paper is organized as follows: The development of the adaptive classification approach is given in Sec. II. Then a description of the dataset used in this study (including image features for characterization of MC lesions) is given in Sec. III. Details related to evaluation methods on CADx classification performance are described in Sec. IV. Experiment results and discussions are furnished in Sec. V. Finally, conclusions and future direction are discussed in Sec. VI.

II. ADAPTIVE CLASSIFICATION BOOSTED WITH RETRIEVAL OF SIMILAR CASES

In this work, the problem of classifying clustered MCs into malignant or benign is treated as a two-class classification problem. Let $f(\cdot)$ denote a classifier that has been trained on a set of training samples $\{(\mathbf{x}_i, y_i), i = 1, \dots, N\}$, where $\mathbf{x}_i \in R^n$ denotes the feature vector of each sample and $y_i \in \{-1, +1\}$ denotes its corresponding class label. Now, consider a new lesion with feature vector $\mathbf{x}$. In our approach, we will first obtain from a reference library a set of known cases that have similar image features to $\mathbf{x}$; our goal is to make use of these retrieved cases to improve the accuracy of the classifier $f(\cdot)$ on $\mathbf{x}$. Below we first describe our approach using a linear classifier, based on logistic regression, which is attractive in practice owing to its numerical simplicity. Afterward, we extend this approach to a nonlinear classifier using the technique of KPCA.⁴¹

II.A. Case-adaptive linear classifier

Mathematically, a linear classifier is of the form

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

where $\mathbf{x}$ is a vector denoting an input pattern (i.e., lesion), and $f(\mathbf{x})$ is the classifier output, which is typically compared against an operating threshold for decision on $\mathbf{x}$. In practice, the discriminant vector $\mathbf{w}$ and bias parameter b are both determined from training in a given problem.

We consider the approach of logistic regression for determining the unknown $\mathbf{w}$ and b from the training set $\{(\mathbf{x}_i, y_i), i = 1, \dots, N\}$,⁴² which is to maximize the following log-likelihood function:

$$L(\mathbf{w}, b) = \sum_{i=1}^N \log p(y_i | \mathbf{x}_i; \mathbf{w}, b), \quad (2)$$

where each probability term is given by

$$p(y_i = 1 | \mathbf{x}_i; \mathbf{w}, b) = \frac{1}{1 + \exp(-\mathbf{w}^T \mathbf{x}_i - b)}, \text{ and} \\ p(y_i = -1 | \mathbf{x}_i; \mathbf{w}, b) = 1 - p(y_i = 1 | \mathbf{x}_i; \mathbf{w}, b). \quad (3)$$

Now, consider a query lesion $\mathbf{x}$ to be classified. Let $\{(\mathbf{x}_i^{(r)}, y_i^{(r)}), i = 1, \dots, N_r\}$ be a set of N_r cases obtained from a reference library which are similar to $\mathbf{x}$. In our case-adaptive approach, we modify the objective function in Eq. (2) by these retrieved cases as

$$\tilde{L}(\mathbf{w}, b) = \sum_{i=1}^N \log p(y_i | \mathbf{x}_i; \mathbf{w}, b) \\ + \sum_{i=1}^{N_r} \beta_i \log p(y_i^{(r)} | \mathbf{x}_i^{(r)}; \mathbf{w}, b), \quad (4)$$

where the weighting coefficients $\beta_i > 1$ are used to put more emphasis on the retrieved samples, particularly those more similar to the query, the goal being to refine the decision boundary of the classifier in the neighborhood of the query $\mathbf{x}$. Indeed, the first term in Eq. (4) simply corresponds to the log-likelihood function in Eq. (2), while the second term can be viewed as a weighted likelihood of those retrieved similar samples. Intuitively, the retrieved samples are used to steer the pretrained classifier from Eq. (2) to achieve more emphasis in the neighborhood of the query $\mathbf{x}$.

Note that objective function in Eq. (4) has the same mathematical form as that in the original optimization problem in Eq. (2), which can be solved efficiently by the method of iteratively reweighted least square (IRLS).⁴²

II.B. Case-adaptive nonlinear classifier

While the concept of our adaptive approach is presented with a simple classifier above, it can also be applicable to a nonlinear classifier. Here, we extend it to a nonlinear classifier using the technique of KPCA,⁴¹ which is a nonlinear extension of the classical linear PCA widely used for feature extraction in pattern recognition problems. Indeed, it was demonstrated that a linear classifier with KPCA could

achieve comparable performance with that of a nonlinear SVM classifier.⁴¹

As in SVM classification, KPCA is also derived through an implicit nonlinear mapping defined via a kernel function. In concept, KPCA works as follows. Consider a set of given data points $\{\mathbf{x}_i, i = 1, \dots, N\}$. These points are first mapped into $\{\Phi(\mathbf{x}_i), i = 1, \dots, N\}$ in a high-dimensional space H through an implicit mapping $\Phi(\cdot)$ associated with a kernel function $K(\cdot, \cdot) = \Phi^T(\cdot)\Phi(\cdot)$; the linear PCA is then applied to these mapped data points in H for feature extraction. For a sample $\mathbf{x}$, its mapped vector $\Phi(\mathbf{x})$ can then be represented in terms of the PCA basis vectors in H , and classified according to this representation.

Interestingly, the KPCA decomposition of a vector $\mathbf{x}$ can be computed directly in terms of the kernel function $K(\cdot, \cdot)$ without the explicit need of the underlying mapping $\Phi(\cdot)$. Specifically, it can be computed as follows: first, form the kernel matrix $\mathbf{K} = [K(\mathbf{x}_i, \mathbf{x}_j)]_{i,j=1,\dots,N}$ from the set of data points $\{\mathbf{x}_i, i = 1, \dots, N\}$. Next, solve the eigenvectors/eigenvalues of $\mathbf{K}$ from $\mathbf{K}\boldsymbol{\alpha} = \lambda\boldsymbol{\alpha}$. Let the eigenvectors be denoted by $\boldsymbol{\alpha}^{(j)} = [\alpha_1^{(j)}, \dots, \alpha_N^{(j)}]^T, j = 1, \dots, N$, which are normalized to have unit length each. Then the projection of $\Phi(\mathbf{x})$ onto the j th PCA basis vector in H can be obtained in terms of $\boldsymbol{\alpha}^{(j)}$ as

$$\phi_j(\mathbf{x}) = \sum_{i=1}^N \alpha_i^{(j)} K(\mathbf{x}, \mathbf{x}_i), \quad (5)$$

for $j = 1, \dots, N$. Together these components $\phi_j(\mathbf{x})$ form the KPCA decomposition of $\mathbf{x}$.

For notational simplicity, it is assumed above that the mapped points $\{\Phi(\mathbf{x}_i), i = 1, \dots, N\}$ are centered in H , i.e., having zero mean. If not, the kernel matrix should be adjusted accordingly as $\tilde{\mathbf{K}} = \mathbf{K} - \mathbf{1}_N \mathbf{K} - \mathbf{K} \mathbf{1}_N + \mathbf{1}_N \mathbf{K} \mathbf{1}_N$, where $\mathbf{1}_N$ is an $N \times N$ matrix whose elements are all $\frac{1}{N}$.

In this work, the KPCA components associated with the largest eigenvalues are used as input features to the classifier. The adaptive linear classifier above can then be applied to these features. Specifically, let $\phi(\mathbf{x}_i) = [\phi_1(\mathbf{x}_i), \dots, \phi_M(\mathbf{x}_i)]^T$ denote the first M KPCA components of $\mathbf{x}_i, i = 1, \dots, N$.

The objective function in Eq. (4) can be written accordingly as

$$L(\mathbf{w}, b) = \sum_{i=1}^N \log p(y_i | \phi(\mathbf{x}_i); \mathbf{w}, b) + \sum_{i=1}^{N_r} \beta_i \log p(y_i^{(r)} | \phi(\mathbf{x}_i^{(r)}); \mathbf{w}, b). \quad (6)$$

Given the KPCA features, Eq. (6) has the same mathematical form as Eq. (4) and hence can be solved in the same way. Consequently, the resulting classifier will have a linear decision boundary in the mapped feature space. Because the underlying mapping function (associated with the kernel function) is typically nonlinear, this decision boundary corresponds to a nonlinear one in the original feature space.

In our experiments, we will use the Gaussian radial basis function (RBF) for the kernel function, which is widely used

in kernel methods. The RBF kernel function is of the following form:

$$K(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{y}\|^2}{\sigma^2}\right), \quad (7)$$

where the width parameter σ^2 is to be determined during training.

III. CONSTRUCTION OF DATASET

III.A. Database of MC cases

In this study, we use a dataset of digitized, standard-view, screen-film mammographic images collected from two sources: one is from the Department of Radiology, The University of Chicago (UC), and the other from the DDSM dataset maintained at The University of South Florida.⁴³ The DDSM dataset consists of images collected from three different institutions: The Massachusetts General Hospital (MGH), Wake Forest University School of Medicine (WFU), and Sacred Heart Hospital and ISMD, Incorporated (ISDM). It has 14 benign volumes, 15 cancer volumes, 12 normal volumes, and 2 benign without callback volumes. The normal and benign without callback volumes are not used in this work since they contain no lesions with verified pathology. To maximize the yield, we extracted all those cases with only MC lesions from all the benign and cancer volumes in DDSM.

For the UC images, all of the cases were pathologically proven by biopsy containing lesions with MCs. The malignant images were obtained in consecutive cases sent for biopsy at UC; likewise, the benign images were randomly selected from consecutive benign cases at UC. These malignant and benign cases were collected during an overlapping time period. They include a spectrum of representative, difficult-to-classify cases, because they were all sent for biopsy. Indeed, a set of 104 cases collected in a similar fashion at UC was used in a previous study;⁴ the average classification performance by a group of five attending radiologists on this set of cases yielded a value of only 0.61 in the area under the receiver operating characteristic (ROC) curve,⁴ which indicates that some of the cases were extremely subtle.

The mammogram images from the different institutions were digitized with five different scanners with different resolutions ranging from 100, 50, 43.5, to 42 μm . To reduce their differences, the mammogram images were first converted to a common resolution (100 μm) using cubic interpolation, and calibrated such that their gray levels correspond to the same optical density. Altogether, there were a total of 1006 cases (646 benign and 360 malignant) collected in the dataset. A breakdown of benign cases and malignant cases among the different institutions is given in Table I.

III.B. Detection of MCs

In the USF images, the ground truth of the MC lesions was already outlined using boundary chain codes. However, the individual MCs were not identified. In order to characterize the MC lesions, we applied the Difference of Gaussian (DoG) method to automatically detect and locate the

TABLE I. Distribution of MC cases in the dataset among different institutions.

Institution	UC	MGH	WFU	ISMD
Benign	277	194	155	20
Malignant	56	187	75	42
Total	333	381	230	62

individual MCs inside the lesion boundary.^{44,45} The DoG detector consists of two Gaussian kernels of different standard deviations, which are set to $\sigma_1 = 1.1$ and $\sigma_2 = 1.4$.⁴⁴ The MCs were detected from the DoG output with an adaptive local threshold $T_S = \mu_L + c \times \sigma_L$, where μ_L and σ_L are the local mean and standard deviation estimated from a 101×101 -pixel region centered at each peak location of DoG output, and c is a coefficient used to adjust the threshold for detection. The detected pixels were grouped into individual objects based on their connectivity, of which those with less than three pixels were discarded as false positives.

To optimize the detection performance, we first calibrated the threshold T_S by comparing the detection results against an expert human reader. For this purpose, we randomly selected 72 lesion ROIs from UC images and 72 lesion ROIs from USF images. By setting $c = 3.2$ in T_S , we could achieve an average true-positive detection rate of 84.6% and an false-positive detection rate of 24.1% when compared with that of the human marked results. This threshold was then applied for MC detection in the rest of the cases in the dataset.

We note that in the literature there have been more elaborate techniques developed for the purpose of detecting MC clusters from a mammogram image. In this study, the boundary of MC clusters is already given in each image, which reduces the complexity of the detection problem. Thus, we simply employed the DoG method, which was previously applied for detecting MCs in CADx applications.^{44,45}

III.C. Features for characterizing MC lesions

To characterize clustered MCs, in our previous work,²³ we used a set of eight features to discriminate between benign and malignant MC lesions:¹⁴ (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. These features were used to quantify both the properties of individual MCs and their geometric distribution in the lesion. In this work, we consider an expanded set of features by including a large number of features developed in the literature for characterizing clustered MCs (Refs. 10 and 46) and apply a feature selection procedure to automatically determine a set of input features for the CADx classifier. Below we describe these features in two major groups: the first is on the image features of individual MCs in a lesion, and the second is on their geometric properties as a cluster.

The following features are considered in this study for characterizing the individual MC objects. For a given MC object:

1. Area (or size) of the MC object, measured by the number of pixels in the MC;
2. Mean image intensity value of the MC pixels;
3. Standard deviation of the image intensity among the MC pixels;
4. Image contrast, computed as the difference between the mean intensity value of the MC and its surrounding background;
5. The effective volume of the MC (area times effective thickness);¹⁴
6. Compactness, measuring the roundness of the MC object;
7. Shape irregularity, measured by the variance of the distance from MC boundary to its geometric center;¹⁴
8. The mean intensity of a 11×11 image window centered at the MC;
9. The fourth order central moment of the MC object μ_{22} .^{47,48}

Note that the above feature quantities are defined for individual MCs. They were extracted from mammogram images after removal of background with a high-pass filter⁴⁹ in order to avoid the effect of inhomogeneous background. For a cluster of MCs, the maximum, minimum, mean, and standard deviation are computed for each of these features over all the MCs in the cluster, yielding a total of 36 features. The standard deviation values of these features are used to characterize the variations among the MCs in the cluster.

In addition, the following features are considered for characterizing the geometric distribution of MCs in a cluster:

1. Number of MCs in the cluster;
2. Mean of the distance from individual MCs to the geometric center of the cluster;
3. Standard deviation of the distance from individual MCs to the geometric center of the cluster;
4. Area of the cluster region;
5. Density of the MC cluster, measured by the number of MCs in a unit area;
6. Eccentricity of the cluster, computed as the ratio of the distance between the foci and the length of the major axis of the smallest enclosing ellipse of the cluster;
7. Compactness of the cluster, measuring the roundness of the cluster.

Altogether, the above yield a large number of features for quantifying an MC cluster. To determine the most relevant features for classification, we applied a sequential forward procedure⁵⁰ with a linear classifier using logistic regression on a set of training images (Sec. IV A). In the end, a total of nine features were selected for characterizing MC lesions, which are as listed in Table II.

The selected nine features were then used to form a vector $\mathbf{x}$ for each lesion in the dataset. Most of the cases in the dataset have two standard views (CC and MLO). Thus, a lesion can have as many feature vectors as the number of its available views. The issue of multiple views in CAD system has been explored in the literature.^{51–54} In our previous study,²³

TABLE II. Selected features for characterizing MCs in a cluster.

(a) Number of MCs in the cluster
(b) Density of the cluster, measured by the number of MCs in a unit area
(c) Mean of the MC size in the cluster
(d) Eccentricity of the cluster
(e) Standard deviation of the distance from individual MCs to the geometric center of the cluster
(f) Maximum of the mean intensity of MCs
(g) Mean of the average intensity in each MC window
(h) Standard deviation of the contrast of MCs
(i) Standard deviation of the fourth order central moment of MCs

it was determined that a simple average of features from multiple views could be more effective than several alternatives. This is because the features from the different views are highly correlated, and averaging can reduce potential noise in the feature values from individual views. In this study, the averaging approach is used for a lesion visible in multiple views.

IV. PERFORMANCE EVALUATION

IV.A. Experiment set up

To demonstrate the proposed approach for case-adaptive classification, we use the following setting in this study. We first partition the dataset of all 1006 cases randomly into two subsets, denoted by A and B, respectively, such that subset A consists of 175 cases (100 benign, 75 malignant), and subset B has the remaining 831 cases (546 benign, 285 malignant). The cases in A are used to train and optimize a baseline CADx classifier, while the cases in B are set aside as a library of known cases for retrieval, which is used to boost the baseline CADx classifier (trained on subset A). With such consideration, the number of cases in subset A is chosen to be similar to that typically used in CADx classification of MC cases in the literature, while a much larger number of cases is kept for subset B.

For efficient use of subset B, the cases in B are also used to test both the CADx classifier and its adaptive counterpart. To remove the effect of case distributions, we apply a bootstrapping methodology for testing the performance of the classifiers.^{59,60} A total of 10000 bootstrap sample sets are used, of which each is obtained by sampling with replacement from the cases in B; subset A is fixed during this process. The classifier performance is subsequently obtained over each bootstrap sample set.

To avoid any potential bias, we apply an LOO procedure when testing the adaptive classifier as follows: for each case used for testing, we exclude this case from B and retrieve known cases only from the rest of the data set to boost the CADx classifier. The resulting classifier is then applied to classify the test case. This is to ensure that the test case itself will not be used in any way for boosting the adaptive classifier.

Besides comparing the adaptive classifier with its baseline counterpart, we will also compare it with the CADx classifier that is trained with set A plus all but the test case in set B by

LOO. For comparison, this classifier is tested using the same bootstrap sample sets as the adaptive classifier.

To evaluate the classification performance, we conduct an ROC analysis, which is now routinely used for performance evaluation in classification tasks. An ROC curve is a plot of the classification sensitivity (i.e., true positive fraction) as the ordinate versus the specificity (i.e., false positive fraction) as the abscissa; for a given classifier, it is obtained by continuously varying the threshold associated with its decision function. As a summary measure of overall diagnostic performance, the area under an ROC curve (denoted by AUC) is often used. A larger AUC means better classification performance. In our experiments, the ROCKIT program⁵⁵ was used to calculate the AUC values for the different classifiers.

IV.B. Retrieval of similar cases

With all the cases in the library characterized by their features vectors, for a given query case $\mathbf{x}$, the similar cases to $\mathbf{x}$ can then be obtained by comparing its feature vector against that of library cases according to a similarity measure. These retrieved cases are then incorporated into the case-adaptive classifier in Eq. (4) via the weighting coefficients β_i . In this study, we implemented the following strategy for adjusting β_i according to the similarity level of a retrieved sample $\mathbf{x}_i^{(r)}$ to the query $\mathbf{x}$

$$\beta_i = 1 + k \frac{\alpha_i}{\max_{j=1, \dots, N_r} \{\alpha_j\}}, i = 1, \dots, N_r, \quad (8)$$

where α_i denotes the similarity measure between $\mathbf{x}_i^{(r)}$ and $\mathbf{x}$, and $k > 0$ is used to control the degree of emphasis on the retrieved samples relative to other training samples. The choice of the form in Eq. (8) is such that the weighting factor increases linearly with the similarity level of a retrieved case to $\mathbf{x}$, with the most similar case among the retrieved receiving maximum weight $1 + k$, which corresponds to k times more influence than the existing training samples in the objective function in Eq. (4).

As a similarity measure for retrieved cases, we used the Gaussian RBF kernel function

$$\alpha_i = \exp\left(-\|\mathbf{x}_i^{(r)} - \mathbf{x}\|^2 / \gamma^2\right), i = 1, \dots, N_r, \quad (9)$$

where γ is a scaling factor controlling the sensitivity of α_i with respect to the distance between the query and a retrieved case. In our experiments, the parameter γ was set to the 10th percentile of the distance between every possible image pairs in the training set. Such a choice is out of the consideration that most of the cases in a database are typically not similar to each other. Those cases with a large distance away from query $\mathbf{x}$ will receive a low similarity measure consequently.

We note that it is possible to extend the above adaptive scheme to other similarity measures other than the Euclidean distance above, such as the DANN measure,⁵⁶ perceptual

similarity,³⁵ and other measurements.^{57,58} This can be of interest for future research, as our main focus here is to demonstrate an approach to improve CADx performance with retrieved known cases. Hence, without unduly complicating the problem, the Euclidean distance measure was used in our experiments in favor of its simplicity.

V. RESULTS AND DISCUSSIONS

V.A. Case-adaptive linear classifier

We first tested the case-adaptive approach for the linear classifier described in Sec. II A. We summarize in Fig. 3 the performance results achieved by the case-adaptive linear classifier (ada-LR). To demonstrate the impact of retrieved cases, in the experiments we also varied the number of retrieved cases N_r for a query case. In Fig. 3, the area under the ROC curve, AUC, is shown for different values of N_r ; for comparison, the performance result by the baseline classifier (LR) (i.e., prior to boosting) is also shown. As described in Sec. IV A, these results were obtained from 10000 bootstrap sample sets of the cases in set B.

From Fig. 3, it can be seen that the AUC value of the adaptive classifier is notably higher than that of the baseline classifier. In particular, with $N_r=10$ the adaptive classifier achieved AUC=0.7663; with $N_r=50$, the AUC value of the adaptive classifier was further increased to 0.7807 compared with AUC=0.7415 for the baseline classifier (p -value=0.0001). No additional improvement was observed when N_r was further increased. We believe that this is because the number of similar cases for a given query case is limited in the reference library and the benefit of retrieving additional cases diminishes.

In addition, in Fig. 3, we also show the results obtained by the CADx classifier trained with all the cases in the dataset (LR-all), as described in Sec. IV A. As can be seen, while its performance is also higher than the baseline classifier (LR), owing to the use of more training cases, it is still some-

what lower than the retrieval-boosted adaptive classifier (ada-LR) (p -value=0.0164 when $N_r=50$).

Finally, we did an additional test to see whether the adaptive classifier could simply have been benefited from the use of additional cases from the library. In this test, an equal number of cases *randomly* selected from the library were used to boost the adaptive classifier. The results are also shown in Fig. 3. As can be seen, the performance of the adaptive classifier shows little or no improvement over the baseline classifier unless when N_r is considerably large. We believe that this is due to the fact the likelihood of having similar cases among the retrieved increases with N_r . For N_r small, the likelihood of retrieving similar cases is much reduced, and putting more emphasis on the retrieved cases actually adversely affects the adaptive classifier. This demonstrates that the performance improvement was indeed attributed to the retrieval of similar cases, not just the use of additional cases.

V.B. Case-adaptive nonlinear classifier

We next tested the case-adaptive nonlinear classifier described in Sec. II B. For the baseline KPCA classifier, the parameter σ^2 needs to be determined for the Gaussian RBF kernel; in addition, the number of KPCA components M in $\phi(\mathbf{x}_i) = [\phi_1(\mathbf{x}_i), \dots, \phi_M(\mathbf{x}_i)]^T$ also needs to be determined. In our experiments, we applied a leave-one-out procedure on the training set to determine these parameters. In the end, $\sigma^2=30$ and $M=10$ were used. To avoid any potential bias, the KPCA decomposition was also based on solely on the training set.

In Fig. 4, we show the results obtained by the adaptive KPCA classifier (KPCA-ada-LR), where the performance AUC is shown for different values of N_r . For comparison, the results are also given for both the baseline KPCA classifier (KPCA-LR) (i.e., prior to boosting) and the KPCA classifier trained with all the cases in the dataset (KPCA-LR-all). As can be seen, the performance of the adaptive classifier is notably higher than that of the baseline classifier. In

FIG. 3. Classification performance (AUC) achieved by the case-adaptive linear classifier (ada-LR). The number of retrieved cases N_r was varied from 6 to 200. For comparison, results are also shown for the baseline classifier (LR) and the classifier trained with all the available cases (LR-all). Results also shown for the case-adaptive classifier when boosted with randomly retrieved cases (ada-LR-random).

FIG. 4. Classification performance (AUC) achieved by the case-adaptive KPCA classifier (KPCA-ada-LR). The number of retrieved cases N_r was varied from 6 to 200. For comparison, the results are also shown for the baseline classifier (KPCA-LR) and the classifier trained with all the available cases (KPCA-LR-all).

particular, with $N_r=20$ the adaptive classifier achieved $AUC=0.7838$, compared with $AUC=0.7527$ for the baseline classifier (p -value = 0.0043). Similar to the case of the adaptive linear classifier above, the adaptive KPCA classifier also outperforms the classifier trained with all the cases.

By comparison with the results in Fig. 3, the adaptive KPCA classifier is noted to achieve only slightly higher performance than the adaptive linear classifier. This indicates that the decision boundary of the nonlinear classifier can be approximated well by that of the locally adaptive linear classifier. Interestingly, it is also noted that the adaptive KPCA classifier in Fig. 4 can adapt faster with the retrieved samples than the linear one, where the optimal performance was obtained with $N_r=20$, compared with $N_r=50$ in Fig. 3.

V.C. Effects of retrieval similarity

In the adaptive classifier (4), the weighting coefficient β_i was used to adjust the relative influence of a retrieved case on the classifier. In the above experiments, this coefficient was set according to Eq. (8), in which the parameter k was set to 10 and parameter γ was set as the 10th percentile of the intersample distances (as described in Sec. IV B). The performance of the adaptive classifiers was found to be rather robust to the choice of these parameters. To demonstrate this, in Fig. 5, we show the classification results on cases in set B obtained by the adaptive linear classifier in Fig. 3 when these two parameters were perturbed as follows: $k=5, 10$, and 20 , and $\gamma=1.81, 2.10$, and 2.46 (which correspond to the 5th, 10th, and 20th percentiles of the intersample distances, respectively). As can be seen, the performance of the adaptive classifier shows little variations with respect to these different parametric settings.

V.D. Execution time

In Fig. 6, we show the average execution times for classifying a test case by the adaptive classifiers; for comparison, results are also shown for their baseline counterparts. All these classifiers were implemented with MATLAB on a 2-GHz

FIG. 5. Classification performance of the adaptive linear classifier (LR-ada) with perturbed values for parameters k and γ in the similarity coefficient. The classifier performance shows little variation with respect to perturbations in these parameters.

FIG. 6. Average execution time for classifying an individual case by the different classifiers. The nonlinear classifiers are noted to be almost as fast as their linear counterparts, thanks to the use of the KPCA decomposition.

PC. As can be seen, while the adaptive classifiers were slower than their baseline counterparts, the time for classifying an individual case was around 7 ms on average; it is also noted that the nonlinear classifiers were almost as fast as their linear counterparts, thanks to the use of KPCA decomposition.

V.E. Further discussions

In this work, a fixed number N_r of similar cases were used to boost the classification accuracy of the query. Conceivably, a query would benefit more from having many similar cases in the reference library, since these retrieved cases would help better to refine the decision boundary in the local neighborhood of the query; on the contrary, such benefit would be minimal if the retrieved cases are not at all closely related to the query. This is taken into account in the weighting factor β_i in Eq. (8) for the adaptive classifier, which is used to adjust the relative influence of the retrieved cases to the classifier according to their distance from the query. It would be interesting to further investigate which type of cases has benefited more from such boosting.

Another interesting issue is the effect of the size of the training set on the adaptive classifier. In our experiments, we selected the size of the training set to be typical of that used in classification studies in the literature. It would be reasonable to expect that the performance improvement might diminish when the training set size becomes sufficiently large. To further examine this effect, we performed the following experiment: for the baseline linear classifier LR, different from the results above, we varied the size of the training set A to 100, 150, and 200, and tested the classification performance with a common set of cases (600) which were exclusive of set A. The baseline classifier LR achieved $AUC=0.7185, 0.7422$, and 0.7475 , respectively, for the above three training set sizes; correspondingly, the adaptive classifier ada-LR achieved $AUC=0.7643, 0.7661$, and 0.7697 , respectively, when the number of retrieved cases was $N_r=100$. As can be seen, the relative improvement of ada-LR over LR is greater when the size of training set A is 100; the classifier performance is similar when the training set size is further increased from 150 and 200. Furthermore, the results earlier in Figs. 3

and 4 also show that the adaptive classifier could still outperform the baseline classifier even when it was trained with the entire library. This indicates that the boosting can still be beneficial even when the training set is large.

Finally, in this study, the relevant cases were retrieved based on their feature similarity to the query. In the scenario that one class is greatly predominant over the other in the library, it would be necessary to apply a weighting scheme to adjust the relative boosting of the two classes, as would be done in traditional classifier training when samples are not well balanced between the classes. This can be a topic of future research.

VI. CONCLUSIONS

In this paper, we investigated the use of case-adaptive classifiers based on image-retrieval for computer-aided diagnosis. The idea is to use a set of retrieved cases with similar image features from a reference library to boost an existing baseline CADx classifier on a query case. The proposed approach was developed first using a linear classifier based on logistic regression, and then extended to a nonlinear classifier by using kernel PCA. We demonstrated this approach on a dataset of 1006 mammogram cases with microcalcification lesions collected from multiple institutions. The case-adaptive classifiers could achieve significant improvements in classification accuracy over the baseline classifiers; the performance of the adaptive classifiers was also demonstrated to be robust to variations in their parametric settings. Moreover, the adaptive classifiers were demonstrated to even outperform the scenario when the classifiers were trained with all the available cases. A potential drawback of the proposed approach is the added computational complexity associated with the adaptation of the classifiers. However, with the help of the kernel PCA and linear regression, the average run time of the adaptive classifiers was shown to be around 7 ms per case in our experiments. For simplicity, in this study, the Euclidean distance was used as the similarity measure in the adaptive classifiers; this would allow us to evaluate the benefit of the proposed approach without complications associated with the image retrieval step. In future work, we plan to further investigate the potential effect of using different similarity metrics on the performance of the proposed approach.

ACKNOWLEDGMENTS

This work was supported by the National Institutes of Health under grant NIBIB R01EB009905.

^{a)}Author to whom correspondence should be addressed. Electronic mail: yangyo@iit.edu

^{b)}Robert M. Nishikawa is a shareholder in Hologic, Inc. (Bedford, MA) and also receives research support and royalties.

¹*Cancer Facts and Figures 2011*, (American Cancer Society, Atlanta, GA, 2011).

²E. A. Sickles *et al.*, "Performance benchmarks for diagnostic mammography," *Radiology* **235**, 775–790 (2005).

³H. P. Chan *et al.*, "Improvement of radiologists' characterization of mammographic masses by using computer-aided diagnosis: An ROC study," *Radiology* **212**, 817–827 (1999).

⁴Y. Jiang, R. M. Nishikawa, R. A. Schmidt, C. E. Metz, M. L. Giger, and K. Doi, "Improving breast cancer diagnosis with computer-aided diagnosis," *Acad. Radiol.* **6**, 22–33 (1999).

⁵L. Hadjiiski *et al.*, "Improvement in radiologists' characterization of malignant and benign breast masses on serial mammograms with computer-aided diagnosis: An ROC study," *Radiology* **233**, 255–265 (2004).

⁶K. Horsch, M. L. Giger, C. J. Vyborny, L. Lan, E. B. Mendelson, and R. E. Hendrick, "Classification of breast lesions with multimodality computer-aided diagnosis: Observer study results on an independent clinical data set," *Radiology* **240**, 357–368 (2006).

⁷M. P. Sampat, M. K. Markey, and A. C. Bovik, "Computer-aided detection and diagnosis in mammography," in *Handbook of Image and Video Processing*, 2nd ed. (Elsevier/Academic, New York, 2005), Chap. 10.4.

⁸R. M. Nishikawa, "Current status and future directions of computer-aided diagnosis in mammography," *Comput. Med. Imaging Graph.* **31**, 224–235 (2007).

⁹R. M. Rangayyan, J. A. Fabio, and J. L. Desautels, "A review of computer-aided diagnosis of breast cancer: Toward the detection of subtle signs," *J. Franklin Inst.* **344**, 312–348 (2007).

¹⁰M. Elter and A. Horsch, "CADx of mammographic masses and clustered microcalcifications: A review," *Med. Phys.* **36**, 2052–2068 (2009).

¹¹Y. Wu, M. L. Giger, K. Doi, C. J. Vyborny, R. A. Schmidt, and C. E. Metz, "Artificial neural networks in mammography: Application to decision making in the diagnosis of breast cancer," *Radiology* **187**, 81–87 (1993).

¹²J. A. Baker, P. J. Kornguth, J. Y. Lo, M. E. Williford, and C. E. Floyd, "Breast cancer: Prediction with artificial neural network based on BI-RADS standardized lexicon," *Radiology* **196**, 817–822 (1995).

¹³J. Y. Lo, M. K. Markey, J. A. Baker, and C. E. Floyd, "Cross-institutional evaluation of BI-RADS predictive model for mammographic diagnosis of breast cancer," *Am. J. Roentgenol.* **178**, 457–463 (2002).

¹⁴Y. 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* **198**, 671–678 (1996).

¹⁵C. Markopoulos, E. Kouskos, K. Koufopoulos, V. Kyriakou, and J. Gogas, "Use of artificial neural networks (computer analysis) in the diagnosis of microcalcifications on mammography," *Eur. J. Radiol.* **39**, 60–65 (2001).

¹⁶M. Kallergi, "Computer-aided diagnosis of mammographic microcalcification clusters," *Med. Phys.* **31**, 314–326 (2004).

¹⁷L. Bocchi and J. Nori, "Shape analysis of microcalcifications using radon transform," *Med. Eng. Phys.* **29**, 691–698 (2007).

¹⁸H. P. Chan, D. Wei, K. L. Lam, S. B. Lo, B. Sahiner, M. A. Helvie, and D. D. Adler, "Computerized detection and classification of microcalcifications on mammograms," *Proc. SPIE* **2434**, 612–620 (1995).

¹⁹H. 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.* **25**, 2007–2019 (1998).

²⁰A. P. Dhawan, Y. Chitre, and C. Kaiser-Bonasso, "Analysis of mammographic microcalcifications using gray-level image structure features," *IEEE Trans. Med. Imaging* **15**, 246–259 (1996).

²¹J. Y. Lo, M. A. Gavrielides, M. K. Markey, and J. L. Jesneck, "Computer-aided classification of breast microcalcification clusters: Merging of features from image processing and radiologists," *Proc. SPIE* **5032**, 882–889 (2003).

²²S. Gupta, P. E. Chyn, and M. K. Markey, "Breast cancer CADx based on BI-RADS descriptors from two mammographic views," *Med. Phys.* **33**, 1810–1817 (2006).

²³L. Wei, Y. Yang, R. M. Nishikawa, and Y. Jiang, "A study on several machine-learning methods for classification of malignant and benign clustered microcalcifications," *IEEE Trans. Med. Imaging* **24**, 371–80 (2005).

²⁴J. Chhatwal, O. Alagoz, M. J. Lindstrom, C. E. Kahn, K. A. Shaffer, and E. S. Burnside, "A logistic regression model based on the national mammography database format to aid breast cancer diagnosis," *Am. J. Roentgenol.* **192**, 1117–1127 (2009).

²⁵E. S. Burnside *et al.*, "Probabilistic computer model developed from clinical data in national mammography database format to classify mammographic findings," *Radiology* **251**, 663–672 (2009).

²⁶H. Muller, N. Michoux, D. Bandon, and A. Geissbuhler, "A review of content-based image retrieval system in medical applications-clinical benefits and future directions," *Int. J. Med. Inf.* **73**, 1–23 (2004).

- ²⁷M. Rahman, T. Want, and B. Desai, "Medical image retrieval and registration: Towards computer assisted diagnostic approach," *Proceedings of IDEAS Workshop on Medical Information Systems: The Digital Hospital*, pp. 78–89 (2004).
- ²⁸B. Zheng, "Computer-aided diagnosis in mammography using CBIR approaches: Current status and future perspective," *Algorithms* **2**, 828–849 (2009).
- ²⁹L. Wei, Y. Yang, and R. M. Nishikawa, "Microcalcification classification assisted by content-based image retrieval for breast cancer diagnosis," *Pattern Recogn.* **42**, 1126–1132 (2009).
- ³⁰L. Wei, Y. Yang, and R. M. Nishikawa, "Retrieval-driven microcalcification classification for breast cancer diagnosis," *ISBI*, pp. 1260–1263 (2007).
- ³¹M. Kallergi, "Computer-aided diagnosis of mammographic microcalcification clusters," *Med. Phys.* **31**, 314–326 (2004).
- ³²Y. Jiang, R. M. Nishikawa, and J. Papaioannou, "Dependence of computer classification of clustered microcalcifications on the correct detection of microcalcifications," *Med. Phys.* **28**, 1949–1957 (2001).
- ³³A. Holt, I. Bichindaritz, R. Schmidt, and P. Perner, "Medical applications in case-based reasoning," *Knowl. Eng. Rev.* **20**, 289–292 (2005).
- ³⁴I. El-Naqa *et al.*, "A similarity learning approach to content based image retrieval: Application to digital mammography," *IEEE Trans. Med. Imaging* **23**, 1233–1244 (2004).
- ³⁵L. Wei, Y. Yang, R. M. Nishikawa, and M. N. Wernick, "Learning of perceptual similarity from expert readers for mammogram retrieval," *ISBI*, pp. 1356–1359 (2006).
- ³⁶C. E. Floyd, J. Y. Lo, and G. D. Tourassi, "Case-based reasoning computer algorithm that uses mammographic findings for breast biopsy decisions," *Am. J. Roentgenol.* **175**, 1347–1352 (2000).
- ³⁷A. O. Bilska-Wolak and C. E. Floyd, "Development and evaluation of a case-based reasoning classifier for prediction of breast biopsy outcome with BI-RADSTM lexicon," *Med. Phys.* **29**, 2090–2010 (2002).
- ³⁸B. Zheng *et al.*, "A method to improve visual similarity of breast masses for an interactive computer-aided diagnosis environment," *Med. Phys.* **33**, 111–117 (2006).
- ³⁹M. A. Mazurowski, P. A. Habas, J. M. Zurada, and G. D. Tourassi, "Decision optimization of case-based computer-aided decision systems using genetic algorithms with application to mammography," *Phys. Med. Biol.* **53**, 895–908 (2008).
- ⁴⁰R. Nakayama, H. Abe, J. Shiraishi, and K. Doi, "Potential usefulness of similar images in the differential diagnosis of clustered microcalcifications on mammograms," *Radiology*, **253**, 625–631 (2009).
- ⁴¹B. Schölkopf, A. Smola, and K. R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Comput.* **10**, 1299–1319 (1998).
- ⁴²C. M. Bishop, *Pattern Recognition and Machine Learning* (Springer, New York, 2006).
- ⁴³M. Heath, K. Bowyer, D. Kopans, R. Moore, and W. P. Kegelmeyer, "The digital database for screening mammography," *The Fifth International Workshop on Digital Mammography*, pp. 212–218 (2001).
- ⁴⁴M. F. Salfity, R. M. Nishikawa, Y. Jiang, and J. Papaioannou, "The use of *a priori* information in the detection of mammographic microcalcifications to improve their classification," *Med. Phys.* **30**, 823–831 (2003).
- ⁴⁵J. Dengler, S. Benrens, and H. F. Desaga, "Segmentation of microcalcifications in mammograms," *IEEE Trans. Med. Imaging* **12**, 664–669 (1993).
- ⁴⁶H. Soltanian-Zadeh, F. Rafiee-Rad, and S. Pourabdollah-Nejad, "Comparison of multiwavelet, wavelet, Haralick, and shape features for microcalcification classification in mammograms," *Pattern Recogn.* **37**, 1973–1986 (2004).
- ⁴⁷S. Liao and M. Pawlak, "On image analysis by moments," *IEEE Trans. Pattern Anal. Mach. Intell.* **18**, 254–266 (1996).
- ⁴⁸L. Shen, R. M. Rangayyan, and J. L. Desautels, "Application of shape analysis to mammographic calcifications," *IEEE Trans. Med. Imaging* **13**, 263–274 (1994).
- ⁴⁹I. 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. Imaging* **21**, 1552–1563 (2002).
- ⁵⁰I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.* **3**, 1157–1182 (2003).
- ⁵¹M. Kallergi and A. Vourtsi, "Two view CAD performance in digital mammography," *Med. Phys.* **37**, 3101 (2010).
- ⁵²B. Liu, C. E. Metz, and Y. Jiang, "An ROC comparison of four methods of combining information from multiple images of the same patient," *Med. Phys.* **31**, 2552 (2004).
- ⁵³M. Velikova, M. Samulski, P. Lucas, and N. Karssemeijer, "Improved mammographic CAD performance using multi-view information: A Bayesian network framework," *Phys. Med. Biol.* **54**, 1131 (2009).
- ⁵⁴M. Samulski and N. Karssemeijer, "Optimizing case-based detection performance in a multiview CAD system for mammography," *IEEE Trans. Med. Imaging* **30**, 1001 (2011).
- ⁵⁵C. E. Metz, B. A. Herman, and J. Shen, "Maximum-likelihood estimation of ROC curves from continuously-distributed data," *Stat. Med.* **17**, 1033–1053 (1998).
- ⁵⁶T. Hastie and R. Tibshirani, "Discriminant adaptive nearest neighbor classification," *IEEE Trans. Pattern Anal. Mach. Intell.* **18**, 601–616 (1996).
- ⁵⁷H. Cho, L. Hadjiiski, B. Sahiner, H. P. Chan, M. Helvie, C. Paramagul, and A. V. Nees, "Similarity evaluation in a content-based image retrieval (CBIR) CADx system for characterization of breast masses on ultrasound images," *Med. Phys.* **38**, 1820 (2011).
- ⁵⁸M. A. Mazurowski, J. M. Malofand, and G. D. Tourassi, "Comparative analysis of instance selection algorithms for instance-based classifiers in the context of medical decision support," *Phys. Med. Biol.* **56**, 473–89 (2011).
- ⁵⁹F. W. Samuelson and N. Petrick, "Comparing image detection algorithms using resampling," *Proceedings of 3rd IEEE International Symposium on Biomedical Imaging*, pp. 1312–1315 (2006).
- ⁶⁰A. I. Bandos, H. E. Rockette, and D. Gur, "A permutation test sensitive to differences in areas for comparing ROC curves from a paired design," *Stat. Med.* **24**, 2873–2893 (2005).