Classification of Breast Masses Using Selected Shape, Edge-sharpness, and Texture Features with Linear and Kernel-based Classifiers

Tingting Mu,¹ Asoke K. Nandi,¹ and Rangaraj M. Rangayyan²

Breast masses due to benign disease and malignant tumors related to breast cancer differ in terms of shape, edge-sharpness, and texture characteristics. In this study, we evaluate a set of 22 features including 5 shape factors, 3 edge-sharpness measures, and 14 texture features computed from 111 regions in mammograms, with 46 regions related to malignant tumors and 65 to benign masses. Feature selection is performed by a genetic algorithm based on several criteria, such as alignment of the kernel with the target function, class separability, and normalized distance. Fisher's linear discriminant analysis, the support vector machine (SVM), and our strict two-surface proximal (S2SP) classifier, as well as their corresponding kernel-based nonlinear versions, are used in the classification task with the selected features. The nonlinear classification performance of kernel Fisher's discriminant analysis, SVM, and S2SP, with the Gaussian kernel, reached 0.95 in terms of the area under the receiver operating characteristics curve. The results indicate that improvement in classification accuracy may be gained by using selected combinations of shape, edge-sharpness, and texture features.

KEY WORDS: Breast masses, breast tumors, mammography, computer-aided diagnosis, feature selection, pattern classification, kernel-based classifiers, shape analysis, edge-sharpness analysis, texture analysis

INTRODUCTION

Worldwide, breast cancer is the most common form of cancer and the second most common cause of cancer deaths in females; the disease affects approximately 10% of all women at some stage of their life in the Western world.¹ Breast cancer may be detected via a careful study of clinical history, physical examination, and imaging with either mammography or ultrasound. However, definitive diagnosis of a breast mass may require, in some cases, fine-needle aspiration biopsy, core

needle biopsy, or excisional biopsy.² Mammography has been shown to be effective in screening asymptomatic women by detecting occult breast cancers and by reducing mortality by as much as 35% in women aged between 50 and 69 years.^{3,4} To improve the accuracy and efficiency of mammographic screening programs for the detection of early signs of breast cancer, a number of research projects are focusing on developing methods for computer-aided diagnosis to assist radiologists in diagnosing breast cancer, including works on image analysis⁵⁻²⁰ and computational intelligence²¹⁻²⁷ for efficient detection of breast cancer.

Breast tumors and masses usually appear in the form of dense regions in mammograms. Benign masses generally possess smooth, round, and well-circumscribed boundaries, as opposed to malignant tumors, which usually have spiculated, rough, and blurry boundaries.²⁸ Several shape features have been proposed for the classification of benign masses and malignant tumors.^{5,6,13-15} The need for measures to characterize the sharpness of a region

¹From the Department of Electrical Engineering and Electronics, The University of Liverpool, Brownlow Hill, L69 3GJ, Liverpool, UK.

²From the Department of Electrical and Computer Engineering, Schulich School of Engineering, University of Calgary, Calgary, Alberta, T2N 1N4, Canada.

Correspondence to: Asoke K. Nandi, Department of Electrical Engineering and Electronics, The University of Liverpool, Brownlow Hill, L69 3GJ, Liverpool, UK; tel: +44-151-7944525; fax: +44-151-7944540; e-mail: a.nandi@liverpool.ac.uk

Copyright © 2008 by Society for Imaging Informatics in Medicine

Online publication 28 February 2008

doi: 10.1007/s10278-007-9102-z

of interest (ROI) in an image has also been recognized, leading to different algorithms for the computation of measures of edge sharpness.^{6,7,14} In addition, subtle textural differences have been observed between benign masses and malignant tumors, with the former being mostly homogeneous and the latter showing heterogeneous texture.^{6,28} Methods of computing texture features have been proposed using the mass margin^{7,8,14} or ribbons of pixels around masses obtained using the “rubber band straightening transform.”²⁹

To study the incorporation of features representing multiple radiological characteristics in the analysis of breast masses, many combinations of shape, edge-sharpness, and texture features, formed on the basis of diagnostic significance and classification performance, have been evaluated using several pattern classification methods,^{6,14,23–25,27} including several linear classifiers, artificial neural networks (ANNs), and kernel-based classification methods. The classification accuracy of the relatively weak texture features reached a level comparable to that of shape features using ANNs with a selected topological structure²³ but was much lower than that of shape features using a linear classifier.²⁴ Nandi et al.²⁵ applied genetic programming (GP), associated with sequential forward (and backward) selection and statistical tests, to select combinations of shape, edge-sharpness, and texture features that are important for the purpose of classification using the GP classifier. Previous research works demonstrate that feature combinations with high classification performance using a specific classifier may not always be extended to other classifiers. Thus, in this paper, we propose to select combinations of shape, edge-sharpness, and texture features independent of any classifier, so that the selected combinations are suitable for use with several different classifiers. A genetic algorithm (GA)³⁰ is employed, instead of an exhaustive search of all possible subsets of features of the chosen cardinality, based on measures of data separability in the original feature space, such as alignment of the kernel with the target function,³¹ class separability,³² and normalized distance.³³ We also propose advanced kernel-based pattern classification algorithms that can yield higher classification accuracy when the features used are not well-separated.

Aizerman et al.³⁴ introduced the idea of using kernel functions in machine learning as inner

products in a corresponding feature space. Kernel methods in pattern analysis embed the data in a suitable feature space and then use algorithms based on linear algebra, geometry, and statistics to discover patterns in the embedded data. Several different kernel-based classifiers have been proposed: Boser et al.³⁵ combined kernel functions with large-margin hyperplanes, leading to kernel-based support vector machines (SVMs) that are highly successful in solving various nonlinear and nonseparable problems in machine learning. Fisher³⁶ proposed a method, which is well known as Fisher’s linear discriminant analysis (FLDA), to seek separating hyperplanes that best separate two or more classes of samples based on the ratio of the between-class scatter to the within-class scatter. Mika et al.³⁷ combined kernel functions with FLDA, leading to kernel Fisher’s discriminant analysis (KFDA). We recently proposed the strict two-surface proximal (S2SP) classifier^{38,39}: a classifier that seeks two cross proximal planes to fit the distribution of the samples in the empirical feature space by maximizing two strict optimization objectives with a “square of sum” optimization factor; kernel functions are employed to incorporate nonlinearity. In the present paper, we address the problem of identifying malignant breast tumors using FLDA and the linear versions of the SVM and the S2SP classifier, as well as KFDA and the nonlinear versions of the SVM and the S2SP classifier with the Gaussian kernel, using selected combinations of shape, edge-sharpness, and texture features.

IMAGE DATABASE AND FEATURE EXTRACTION

The digitized mammographic image set used in this study contains 111 ROIs extracted from mammograms, with 65 related to benign masses and 46 to malignant tumors, obtained by combining two image databases. One set of images was obtained from “Screen Test: Alberta Program for the Early Detection of Breast Cancer,”^{24,40} with 37 ROIs related to benign masses and 20 ROIs related to malignant tumors. The images were digitized using the Lumiscan 85 scanner at a resolution of 50 μm and 12 bits per pixel. The other set was obtained by using images containing masses from the Mammographic Image Analysis Society (MIAS, UK) database⁴¹ and the teaching library of the Foothills Hospital in Calgary,¹⁴ with a total

of 54 ROIs including 28 benign and 26 malignant types. The MIAS images were digitized at a resolution of 50 μm , whereas the Foothills Hospital images were digitized at a resolution of 62 μm . The diagnosis of each case was proven by biopsy. Mass or tumor ROIs were manually identified, and contours were drawn by a radiologist experienced in screening mammography. Twenty-two features were extracted from each ROI, including 5 shape features (C , F_{cc} , FF , SI , and FD), 3 edge-sharpness features (A , Co , and CV), and 14 texture features, which are explained in the following paragraphs.

- Shape features: Five shape features are considered in this study, including compactness (C), fractional concavity (F_{cc}), Fourier factor (FF), spiculation index (SI), and fractal dimension (FD). C is a simple measure of the efficiency of a contour to contain a given area and is defined in a normalized form as $1 - \frac{4\pi S}{D^2}$, where D and S are the contour perimeter and area, respectively.¹⁵ F_{cc} is the ratio of the cumulative length of the concave parts to the total length of the contour.¹⁵ Benign masses, due to their round or oval contours, result in low values of C and F_{cc} . On the other hand, contours of microlobulated or spiculated malignant tumors may be expected to have several significant concave portions, and hence, large values of F_{cc} , as well as C . FF is a measure related to the presence of roughness or high-frequency components in the contours.^{13,14} SI represents the degree of spicularity of a contour. Rangayyan et al.¹⁵ proposed an algorithm to compute SI based upon a polygonal model of the given contour and a combination of the segment lengths, base widths, and angles of possible spicules. Due to their effect on the surrounding tissues, most malignant tumors form narrow, stellate distortions around their boundaries and, hence, have higher values of SI than benign masses with smooth contours. FD can be used to characterize self-similarity, nested complexity, or space-filling properties, and was derived by using the two-dimensional ruler method.¹³
- Edge-sharpness features: Three edge-sharpness features are used in this study, including acutance (A), contrast (Co), and coefficient of variation (CV). A is a measure of the sharpness or change in density across a mass margin.¹⁴ Co

is a measure of contrast.⁷ CV is a feature based on the coefficient of variation of the edge strength computed at all points on the boundary of the ROI.⁷

- Texture features: Fourteen texture features were computed according to the definitions of Haralick et al.,⁴² using a ribbon of pixels around the margin of each mass,^{7,8,24} including angular second moment of energy (f_1), contrast (f_2), correlation (f_3), sum of squares (f_4), inverse difference moment (f_5), sum average (f_6), sum variance (f_7), sum entropy (f_8), entropy (f_9), difference variance (f_{10}), difference entropy (f_{11}), information measure of correlation (f_{12} and f_{13}), and maximal correlation coefficient (f_{14}). The texture features were computed using ribbons of a width of 8 mm obtained by dilating the mass boundaries after filtering and down-sampling the mammograms to an effective resolution of 200 μm per pixel.

Some of the shape features used are invariant to scaling (size) and rotation by design; others are normalized to remove the effect of spatial resolution and size. The texture and edge-sharpness features include normalization and should not be affected by the small differences in the pixel size used (50 and 62 μm).

FEATURE SELECTION

Given a set of l labeled training samples $z = \{(\mathbf{x}_i, y_i)\}_{i=1}^l \in (R^n \times Y)$, where R^n is the n -dimensional real feature space with a binary label space $Y = \{1, -1\}$ and $y_i \in Y$ is the label assigned to the sample $\mathbf{x}_i \in R^n$, the purpose of feature selection is to select a subset of relevant features to build robust classifiers based on measures of separability constructed from the training samples.

Measures for Feature Analysis

Alignment

The alignment measure was introduced by Cristianini et al.³¹ to measure the similarity between two kernel functions or between a kernel and a target function. To perform feature selection, we employ the alignment between the inner

product matrix ($\mathbf{E}$) of features in the original feature space and the target label matrix, given as

$$A_E = \frac{\langle \mathbf{E}, \mathbf{y}\mathbf{y}^T \rangle_F}{\sqrt{\langle \mathbf{E}, \mathbf{E} \rangle_F \langle \mathbf{y}\mathbf{y}^T, \mathbf{y}\mathbf{y}^T \rangle_F}}, \quad (1)$$

where $\mathbf{y}$ denotes the column vector of the labels of the training samples. The Frobenius product $\langle \cdot, \cdot \rangle_F$ between two Gram matrices $\mathbf{M}$ and $\mathbf{N}$ is defined as

$$\langle \mathbf{M}, \mathbf{N} \rangle_F = \text{tr}(\mathbf{M}\mathbf{N}). \quad (2)$$

where “tr” denotes the trace of a matrix.^{31,32} This quantity captures the degree of agreement between the input features and the given learning task. A larger value of A_E indicates a higher degree of agreement than a smaller value.

Class Separability

A quantity to measure the class separability of the training samples in the original feature space³² that we employ for feature selection is given by

$$J = \frac{\text{tr}(\mathbf{S}_B)}{\text{tr}(\mathbf{S}_W)}, \quad (3)$$

where $\mathbf{S}_B$ and $\mathbf{S}_W$ are the between-class and within-class scatter matrices, respectively, given by

$$\mathbf{S}_B = (\mathbf{m}^+ - \mathbf{m}^-)(\mathbf{m}^+ - \mathbf{m}^-)^T. \quad (4)$$

$$\mathbf{S}_W = \sum_{i=+, -} \sum_{\mathbf{x} \in C^i} (\mathbf{x} - \mathbf{m}^i)(\mathbf{x} - \mathbf{m}^i)^T, \quad (5)$$

$$\mathbf{m}^+ = \frac{1}{l^+} \sum_{i=1}^{l^+} \mathbf{x}_i^+, \quad (6)$$

$$\mathbf{m}^- = \frac{1}{l^-} \sum_{i=1}^{l^-} \mathbf{x}_i^-, \quad (7)$$

where $\mathbf{x}_i^+$ and $\mathbf{x}_i^-$ represent samples belonging to the positive (+) and negative (−) classes, respectively; l^+ and l^- denote the number of positive and negative training samples, respectively; and C^+ and C^- denote the sets of positive and negative training samples, respectively. A larger value of J indicates better class separability of the training set than a smaller value.

Normalized Distance

We use a measure of normalized distance³³ between the centers (mean vectors) of the two classes to evaluate their separability, given by

$$D = \frac{\|\mathbf{m}^+ - \mathbf{m}^-\|}{\sqrt{\frac{1}{l^+ - 1} \sum_{i=1}^{l^+} \|\mathbf{x}_i^+ - \mathbf{m}^+\|^2 + \frac{1}{l^- - 1} \sum_{i=1}^{l^-} \|\mathbf{x}_i^- - \mathbf{m}^-\|^2}}. \quad (8)$$

A larger value of D indicates better class separability of the training set than a smaller value.

Genetic Algorithm

GA operates on a number of potential solutions (a population), applying the principle of survival of the fittest to produce better and better approximations (individuals) to a solution (see “Tutorial” of Chipperfield et al.⁴³). Each individual is encoded as a string or chromosome composed over some alphabet, of which one commonly used representation is the binary alphabet $\{0,1\}$. The performance of each individual is assessed through an objective function and the fitness function. Highly fit individuals have a high probability of being selected for mating, whereas less fit individuals have a correspondingly low probability of being selected. The selected individuals are then recombined, using crossover, to produce the next generation with a probability P_c , which is to exchange genetic information between pairs or larger groups of individuals. A further genetic operator, called mutation, is then applied to the new individuals again with a low probability P_m , which is to ensure that the probability of searching a particular subspace of the problem space is never zero. The fractional difference between the size of the old population and the new population produced by selection, crossover, and mutation is termed as a generation gap. To maintain the size of the original

population, those new individuals that have larger fitness values than those of the old individuals are brought into the new population. Therefore, the new population includes the old individuals, which are fitter than the new individuals, and the remaining places are taken by the comparatively fitter of the new individuals. GA is terminated after a prespecified number of generations; a subsequent test may be applied to check the quality of the highly fit members of the population. If no acceptable solutions are found, GA may be restarted or a fresh search may be initiated.

One of the main issues in applying GA to the problem of feature selection is to map the search space into a representation suitable for genetic search. In this study, we consider each feature in the candidate feature set as a binary gene. Each possible feature combination is encoded into an n -bit binary string, where n is the total number of features available. An n -bit individual corresponds to an n -dimensional binary feature vector X , where $X_i=0$ represents elimination and $X_i=1$ indicates inclusion of the i th feature. The objective function of GA is set as one of the measures of alignment, class separability, and normalized distance, as described in “Measures for Feature Analysis.”

PATTERN CLASSIFICATION

The purpose of classification is to seek the best prediction of the label for an input sample $\mathbf{x}$ from the given set of labeled training samples $z = \{(\mathbf{x}_i, y_i)\}_{i=1}^l \in (R^n \times Y)$. A linear classifier achieves this by making a classification decision based on the value of a linear combination of the features. A kernel-based classifier performs the same task by applying the linear classification method in a kernel-transformed feature space κ , with a nonlinear mapping $\phi: R^n \rightarrow \kappa$, by using kernel functions as inner products in the feature space κ .

Linear Classifiers

Fisher's Linear Discriminant Analysis

FLDA seeks the separating function

$$f(\mathbf{x}) = \omega^T \mathbf{x} + b \quad (9)$$

by maximizing the following objective:

$$O(\omega) = \frac{\omega^T \mathbf{S}_B \omega}{\omega^T \mathbf{S}_W \omega}, \quad (10)$$

where ω and b are the weight vector and the bias of the separating hyperplane, respectively, and $\mathbf{S}_B$ and $\mathbf{S}_W$ are the between-class and within-class scatter matrices, respectively (see Eqs. 4 and 5). The optimal values of ω and b can be calculated by solving a generalized eigenvalue problem.⁴⁴ Letting $f^*(\mathbf{x})$ denote the derived optimal separating function, the label for an input sample is predicted by

$$Y_z(\mathbf{x}) = \text{sgn}(f^*(\mathbf{x})), \quad (11)$$

where $\text{sgn}(\mathbf{x})$ is 1 when $\mathbf{x} \geq 0$ and -1 otherwise, and Y_z is an estimate of the label for the input sample $\mathbf{x}$.

Support Vector Machines

SVMs seek the separating function based on the maximal margin rule⁴⁵ by solving the following constrained quadratic programming (QP) problem:

$$\max_{\beta} \sum_{i=1}^l \beta_i - \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l y_i y_j \beta_i \beta_j \mathbf{x}_i^T \mathbf{x}_j, \quad (12)$$

subject to

$$\sum_{i=1}^l y_i \beta_i = 0,$$

$$0 \leq \beta_i \leq \xi, \quad i = 1, 2, \dots, l,$$

where $\beta = [\beta_1, \beta_2, \dots, \beta_l]^T$ are Lagrange multipliers and ξ is the regularization parameter set by the user. Letting $\{\beta_i^*\}_{i=1}^l$ denote the optimal solution of the QP problem in Eq. 12, the optimal values of ω and b in the separating function given in Eq. 9 can be calculated as

$$\omega^* = \sum_{i=1}^l y_i \beta_i^* \mathbf{x}_i, \quad (13)$$

$$b^* = -\frac{1}{2S} \sum_{\mathbf{x} \in S_+ \cup S_-} \sum_{i=1}^l \beta_i^* y_i \mathbf{x}^T \mathbf{x}_i, \quad (14)$$

where S_+ and S_- are two sets of support vectors with the same size S but different labels of 1 and -1 . The label for an input sample $\mathbf{x}$ is then predicted by Eq. 11.

The Strict Two-Surface Proximal Classifier

Different from the discriminant classification methods, such as FLDA and SVMs, which predict the label based on one separating function, the S2SP classifier has been proposed to predict the label based on two proximal planes^{38,39}:

$$f_1(\mathbf{x}) = \omega_1^T \mathbf{x} + b_1 = 0, \quad (15)$$

$$f_2(\mathbf{x}) = \omega_2^T \mathbf{x} + b_2 = 0, \quad (16)$$

where ω and b are the weight vector (direction) and the bias of the proximal planes, respectively; the subscripts 1 and 2 denote the first and second planes, respectively; and the first plane is as close to the points of the positive class while being as far as possible from the points of the negative class, whereas the second plane is as close to the points of the negative class while being as far as possible from the points of the positive class. Compared with the proximal classification method of multisurface proximal SVMs,⁴⁶ the S2SP classifier eliminates the regularization term by employing a “square of sum” numerator, with consideration of the sign effect under the situation of misclassification with large projections onto the separating plane.

To obtain the first proximal hyperplane, the objective function to be maximized is

$$O_1(\omega_1, b_1) = \frac{[\mathbf{X}^- \omega_1 + \mathbf{e}b_1]^2}{\|\mathbf{X}^+ \omega_1 + \mathbf{e}b_1\|^2}; \quad (17)$$

the second proximal hyperplane is obtained by maximizing

$$O_2(\omega_2, b_2) = \frac{[\mathbf{X}^+ \omega_2 + \mathbf{e}b_2]^2}{\|\mathbf{X}^- \omega_2 + \mathbf{e}b_2\|^2}, \quad (18)$$

where the $l^- \times n$ matrix $\mathbf{X}^-$ represents samples from the negative class, the $l^+ \times n$ matrix $\mathbf{X}^+$ represents samples from the positive class, $\mathbf{e}$ is a column vector with all elements equal to 1, $[\text{vector}]$ is used to denote the sum of the elements of the vector, and $[\text{matrix}]$ is used to denote a column vector with the sum of each row. The optimal values of ω_1 , b_1 , ω_2 and b_2 can be calculated by solving two generalized eigenvalue problems.^{38,39} For linear classification in the original feature space, the two proximal planes serve as two ridge-like distribution models to fit the samples in the two classes. Letting $d_1(\mathbf{x})$ and $d_2(\mathbf{x})$ denote the Euclidean distance between a given sample $\mathbf{x}$ and the two proximal planes, respectively, one way to predict the label of $\mathbf{x}$ is to consider the values of $d_1(\mathbf{x})$, $d_2(\mathbf{x})$, and $\frac{d_1(\mathbf{x})}{d_2(\mathbf{x})}$ together using linear discriminant analysis.⁴⁴

Kernel-Based Classifiers

Kernel Fisher's Discriminant Analysis

In the kernel-transformed feature space κ , by expanding the weight vector ω of $f(\mathbf{x})$ in Eq. 9 into a linear summation of all training samples, the kernel-based separating function $f(\mathbf{x})$ becomes

$$f(\mathbf{x}) = \sum_{i=1}^l \alpha_i K(\mathbf{x}, \mathbf{x}_i) + b, \quad (19)$$

where $\{\alpha_i\}_{i=1}^l$ denote the summation weights and $K(\cdot, \cdot)$ is a kernel function used to compute the inner product matrix, the so-called kernel matrix, on pairs of samples in the kernel-transformed feature space κ . KFDA determines $f^*(\mathbf{x})$ by maximizing the Fisher criterion,⁴⁷ as

$$O(f) = \frac{(\mu^+ - \mu^-)^2}{(\sigma^+)^2 + (\sigma^-)^2}, \quad (20)$$

where

$$\begin{aligned}\mu^+ &= \frac{1}{l^+} \left(\sum_{i=1}^{l^+} f(\mathbf{x}_i^+) \right)^2, \\ \mu^- &= \frac{1}{l^-} \left(\sum_{i=1}^{l^-} f(\mathbf{x}_i^-) \right)^2, \\ (\sigma^+)^2 &= \frac{1}{l^+} \sum_{i=1}^{l^+} (f(\mathbf{x}_i^+) - \mu^+)^2, \\ (\sigma^-)^2 &= \frac{1}{l^-} \sum_{i=1}^{l^-} (f(\mathbf{x}_i^-) - \mu^-)^2,\end{aligned}$$

where μ^+ and μ^- denote the mean projections of the positive and negative samples, respectively, and σ^+ and σ^- are the corresponding standard deviations. By incorporating Eq. 19 into Eq. 20, the optimal values of $\{\alpha_i\}_{i=1}^l$ and b can be calculated by solving a generalized eigenvalue problem.⁴⁷ The label for an input sample $\mathbf{x}$ is then predicted by Eq. 11.

Kernel-Based Support Vector Machines

An SVM with embedded kernel functions determines the optimal separating function in the kernel-transformed feature space by solving the following constrained QP problem instead of Eq. 12:

$$\max_{\beta} \sum_{i=1}^l \beta_i - \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l y_i y_j \beta_i \beta_j K(\mathbf{x}_i, \mathbf{x}_j), \quad (21)$$

subject to

$$\sum_{i=1}^l y_i \beta_i = 0,$$

$$0 \leq \beta_i \leq \xi, i = 1, 2, \dots, l.$$

Letting $\{\beta_i^*\}_{i=1}^l$ denote the optimal solution of Eq. 21, the optimal value of the bias b can be calculated by

$$b^* = -\frac{1}{2S} \sum_{\mathbf{x} \in S_+ \cup S_-} \sum_{i=1}^l \beta_i^* y_i K(\mathbf{x}, \mathbf{x}_i). \quad (22)$$

The optimal separating function $f^*(x)$ is given by

$$f^*(\mathbf{x}) = \sum_{i=1}^l y_i \beta_i^* K(\mathbf{x}, \mathbf{x}_i) + b^*. \quad (23)$$

The label for an input sample $\mathbf{x}$ is then predicted by Eq. 11.

Kernel-Based Strict Two-Surface Proximal Classifier

In the kernel-transformed feature space κ , by expanding the direction vector of the hyperplane into a linear summation of all of the training samples, the two proximal hyperplanes are given as

$$f_1(\mathbf{x}) = \sum_{i=1}^l \alpha_{i1} K(\mathbf{x}, \mathbf{x}_i) + b_1 = 0, \quad (24)$$

$$f_2(\mathbf{x}) = \sum_{i=1}^l \alpha_{i2} K(\mathbf{x}, \mathbf{x}_i) + b_2 = 0, \quad (25)$$

where $\alpha_1 = [\alpha_{11}, \alpha_{21}, \dots, \alpha_{l1}]^T$ and $\alpha_2 = [\alpha_{12}, \alpha_{22}, \dots, \alpha_{l2}]^T$ are the summation weight vectors of the two proximal hyperplanes. To obtain the first proximal hyperplane, the objective function to be maximized is

$$O_1(\alpha_1, b_1) = \frac{[\mathbf{K}^- \alpha_1 + \mathbf{e} b_1]^2}{\|\mathbf{K}^+ \alpha_1 + \mathbf{e} b_1\|^2}; \quad (26)$$

the second proximal hyperplane is obtained by maximizing

$$O_2(\alpha_2, b_2) = \frac{[\mathbf{K}^+ \alpha_2 + \mathbf{e} b_2]^2}{\|\mathbf{K}^- \alpha_2 + \mathbf{e} b_2\|^2}, \quad (27)$$

where the $l^+ \times l$ matrix $\mathbf{K}^+$ represents the kernel matrix between the samples from the positive class and all the training samples and the $l^- \times l$ matrix

K^- represents the kernel matrix between the samples from the negative class and all the training samples. The optimal values of α_1 , b_1 , α_2 , and b_2 can be calculated by solving two generalized eigenvalue problems.^{38,39} For nonlinear classification based on kernel functions, the two proximal planes serve as two Bayesian models to fit the samples in the two classes. The label for an input sample $\mathbf{x}$ is then predicted using the same method as with the linear S2SP classifier by considering the values of $d_1(\mathbf{x})$, $d_2(\mathbf{x})$, and $\frac{d_1(\mathbf{x})}{d_2(\mathbf{x})}$ together (see “*The Strict 2-Surface Proximal Classifier*”).

EXPERIMENTS, RESULTS, AND COMPARATIVE ANALYSIS

Evaluation of Single Features

Distributions of the normalized values of the 22 features are shown in Figure 1, which

illustrates the separability and (or) overlap between the benign and malignant categories of the various features used. Five different measures were evaluated for each single feature, including the area A_z under the receiver operating characteristics (ROC) curve calculated by applying a sliding threshold with the LS-SVMLab1.5 toolbox,⁴⁸ the p value of the t test derived by employing the function “ttest2” in MATLAB, as well as alignment, class separability, and normalized distance as described in “*Image Database and Feature Extraction*.” The “ttest2” program performs a t test of the hypothesis that two independent samples come from distributions with equal means and returns the result of the test in H , as well as the p value. $H=0$ indicates that the null hypothesis (“means are equal”) cannot be rejected at the 5% significance level. $H=1$ indicates that the null hypothesis can be rejected at the 5% level. p value is the probability of observing the given result, or one more

Fig 1. Distribution of the normalized values of each feature for the 111 breast masses. The cross marks denote malignant tumors, and the circle marks denote benign masses.

Table 1. Values of the Various Measures of Separability Studied for Each Feature

Features	A_z	Alignment	Separability	Distance	p value
C	0.8980	0.4677	0.0345	5.2186	0.0000
SI	0.9134	0.5039	0.0401	4.8699	0.0000
F_{cc}	0.8866	0.4576	0.0331	5.5692	0.0000
FF	0.9100	0.4809	0.0364	5.2407	0.0000
FD	0.9137	0.4459	0.0315	3.9774	0.0000
Co	0.5843	0.0275	0.0011	0.1623	0.0776
A	0.6839	0.0913	0.0039	0.5414	0.0011
CV	0.7027	0.0550	0.0022	0.1903	0.0119
f_1	0.6863	0.0879	0.0037	0.4982	0.0013
f_2	0.7110	0.0922	0.0039	0.4584	0.0010
f_3	0.5575	0.0173	0.0007	0.1058	0.1624
f_4	0.6405	0.0402	0.0016	0.2393	0.0323
f_5	0.6829	0.0980	0.0042	0.8421	0.0007
f_6	0.5759	0.0242	0.0009	0.1779	0.0980
f_7	0.6398	0.0393	0.0016	0.2326	0.0341
f_8	0.6565	0.0712	0.0029	0.5038	0.0040
f_9	0.6853	0.0982	0.0042	0.7844	0.0007
f_{10}	0.7187	0.0971	0.0041	0.3577	0.0007
f_{11}	0.6896	0.0973	0.0041	0.8121	0.0007
f_{12}	0.5903	0.0272	0.0011	0.1882	0.0792
f_{13}	0.5082	0.0002	0.0000	0.0010	0.8794
f_{14}	0.5329	0.0064	0.0002	0.0347	0.3980

extreme, by chance if the null hypothesis is true. Larger values of A_z , alignment, class separability, and normalized distance and smaller values of the p value indicate stronger discriminating power. The calculated values of the five measures are recorded in Table 1; the features are ranked in Table 2 with each measure. A score was calculated by averaging the ranking numbers as given by the five measures for each feature; the resulting ranking of the features is recorded in the last column of Table 2. To ensure that only two edge-sharpness features would be included, features with scores less than 14 were selected to form the feature combination FS_{best1} with all the five shape features, two edge-sharpness features A and CV , and six texture features $f_1, 2, 5, 9, 10, 11$. To ensure that only one edge-sharpness feature would be included, features with scores less than 11 were selected to form the feature combination FS_{best2} with all the five shape features, one edge-sharpness feature A , and five texture features $f_{2,5,9,10,11}$.

Independent Feature Selection

Feature selection was performed using GA based on the three measures, as described in “[Measures for Feature Analysis](#),” calculated using all the 111

breast masses. Such a feature selection procedure is independent of any classifier. “Genetic Algorithm Toolbox for use with MATLAB (version 1.2)”⁴³ was employed to implement GA. For each GA individual, we set that at least one feature should be

Table 2. Ranking of the 22 Features Based on the Various Measures of Separability

Features	A_z	Alignment	Separability	Distance	p value	Average
C	4	3	3	3	3	2
SI	2	1	1	4	1	1
F_{cc}	5	4	4	1	4	3
FF	3	2	2	2	2	4
FD	1	5	5	5	5	5
Co	18	17	17	19	17	17
A	12	11	11	9	11	11
CV	8	14	14	16	14	14
f_1	10	12	12	11	12	12
f_2	7	10	10	12	10	10
f_3	20	20	20	20	20	20
f_4	15	15	15	14	15	15
f_5	13	7	7	6	7	8
f_6	19	19	19	18	19	19
f_7	16	16	16	15	16	16
f_8	14	13	13	10	13	13
f_9	11	6	6	8	6	6
f_{10}	6	9	9	13	9	9
f_{11}	9	8	8	7	8	7
f_{12}	17	18	18	20	18	18
f_{13}	22	22	22	22	22	22
f_{14}	21	21	21	21	21	21

selected from each feature category of shape, edge-sharpness, and texture. The objective function of GA was set as alignment, class separability, and normalized distance. The fitness function was calculated by using linear ranking. The population size was set as 50. The generation gap was set as 0.6. Two-point crossover was employed, with the corresponding crossover rate set as 0.8. Bit flip was used for mutation, with the corresponding mutation rate set as 0.05. The maximum number of generations was set as 500. In the last generation, 50 feature combinations were derived, corresponding to 50 final individuals, each represented by a 22-bit binary string.

The more frequently a feature appears in the GA individuals in the final population, the more important the feature is to provide better separability for classification. Thus, for each feature and each measure, a score of importance was calculated by counting the number of times the feature was included in the 50 individuals and divided by 50; the result is recorded in Table 3. Features with higher scores are more important for the purpose of classification. It can be seen from Table 3 that there is a clear gap between the important and unimportant features, with scores above 0.85 and

lower than 0.15, respectively. The five shape features and the texture feature f_9 possess high scores for all of the three measures (see Table 3). By running GA multiple times, the same three combinations of shape, edge-sharpness, and texture features, containing features with scores higher than 0.85, were derived based on the three objective functions: $FS_{\text{alignment}}$ with all the five shape features, one edge-sharpness feature CV , and one texture feature f_9 ; $FS_{\text{separability}}$ with all the five shape features, one edge-sharpness feature A , and one texture feature f_9 ; and FS_{distance} with five shape features, one edge-sharpness feature A , and two texture features $f_{6,9}$.

Pattern Classification with Selected Features

The following experiments were conducted with FS_{best1} , FS_{best2} , $FS_{\text{alignment}}$, $FS_{\text{separability}}$, and FS_{distance} ; the individual shape, edge-sharpness, and texture feature sets, as well as all of the 22 features combined without performing feature selection, for comparison. The input features were normalized to have zero mean and unit variance before being used by a classifier. Classification performance is shown in terms of the area A_z under the ROC curve and the corresponding standard error. Each ROC curve was generated by applying a sliding threshold to the output of the classifier with the LS-SVMLab1.5 toolbox.⁴⁸

Both the leave-one-out (LOO) and half-half-random (HHR) split procedures were used to evaluate the generalized performance of the classifiers with the features of the 111 breast masses. For the LOO procedure, each mass was used as the test sample once and the remaining 110 masses were used as the training samples; thus, 111 training-test trials were conducted. For the HHR split procedure, 100 training-test trials were conducted. In each trial, 25 benign masses and 25 malignant tumors were selected at random as the training samples (a total of 50 training samples), and the remaining 61 masses were used as the test samples.

Linear Classification

Nine different feature sets, including the shape, edge-sharpness, and texture feature sets; the whole set of 22 features; and the five selected combinations of shape, edge-sharpness, and texture features, were evaluated using FLDA, linear SVM

Table 3. Scores of Importance Derived by GA with Different Objective Functions for Feature Selection

Features	Alignment	Separability	Distance
C	0.95	0.92	0.88
SI	0.96	0.94	0.93
F_{cc}	0.91	0.89	0.92
FF	0.92	0.92	0.96
FD	0.89	0.89	0.94
Co	0.09	0.06	0.06
A	0.06	0.93	0.95
CV	0.92	0.03	0.06
f_1	0.06	0.07	0.08
f_2	0.08	0.07	0.10
f_3	0.07	0.10	0.05
f_4	0.07	0.05	0.11
f_5	0.05	0.07	0.05
f_6	0.08	0.12	0.91
f_7	0.08	0.09	0.06
f_8	0.05	0.06	0.08
f_9	0.91	0.89	0.91
f_{10}	0.08	0.08	0.08
f_{11}	0.03	0.05	0.09
f_{12}	0.06	0.10	0.10
f_{13}	0.06	0.11	0.05
f_{14}	0.07	0.09	0.08

The bold-faced numbers represent significantly better performance figures

Table 4. The Classification Performance of Different Feature Sets Using Linear Classifiers with the LOO Procedure

Feature Sets	LS2SP		FLDA		LSVM	
	A_z	SE	A_z	SE	A_z	SE
Shape	0.90	0.03	0.91	0.03	0.91	0.03
Gradient	0.67	0.05	0.68	0.05	0.65	0.06
Texture	0.64	0.05	0.65	0.05	0.65	0.05
All	0.88	0.03	0.88	0.03	0.89	0.03
$FS_{\text{alignment}}$	0.92	0.03	0.93	0.02	0.93	0.02
$FS_{\text{separability}}$	0.92	0.02	0.92	0.03	0.92	0.03
FS_{distance}	0.92	0.03	0.92	0.03	0.92	0.03
FS_{best1}	0.87	0.04	0.92	0.03	0.91	0.03
FS_{best2}	0.89	0.03	0.93	0.02	0.91	0.03

SE = standard error

(LSVM), and the linear S2SP (LS2SP) classifier with both the LOO and HHR split procedures. The corresponding classification accuracies in A_z values are recorded in Tables 4 and 5. The classification accuracies of FLDA, LSVM, and the LS2SP classifier are comparable; FLDA and LSVM performed slightly better than the LS2SP classifier. The best linear classification performance achieved is 0.93 in A_z value (see Tables 4 and 5).

From Tables 4 and 5, the following observations can be made: Shape features are the most significant features, with higher classification accuracy than the other two sets of edge-sharpness and texture features. The addition of edge-sharpness and texture features does not improve the classification performance of the shape features without feature selection. Feature selection using GA based on data-dependent measures improves the classification

performance over using the whole set of 22 features without feature selection. The feature combinations selected by GA ($FS_{\text{alignment}}$, $FS_{\text{separability}}$, and FS_{distance}) perform better than the combinations selected by ranking the features (FS_{best1} and FS_{best2}). However, shape features on their own provide classification performance similar to the performance provided by the feature combinations selected by GA.

Kernel-Based Classification

The same nine feature sets as listed in Section *Linear Classification* were evaluated using KFDA and the nonlinear versions of the SVM and the S2SP classifier, denoted by KSVM and KS2SP, respectively. The Gaussian kernel was employed to incorporate nonlinearity, with the form

$$K(\mathbf{x}_a, \mathbf{x}_b) = \exp\left(-\frac{\|\mathbf{x}_a - \mathbf{x}_b\|^2}{2\sigma^2}\right) \quad (28)$$

between two vectors $\mathbf{x}_a$ and $\mathbf{x}_b$, where σ is the kernel width set by the user. The value of σ for each classifier was determined by cross-validation in this experiment. Both the LOO and HHR split procedures were used to evaluate the three kernel-based classifiers; the classification accuracies in A_z values are recorded in Tables 6 and 7. The classification accuracies of KFDA, KSVM, and the KS2SP classifier are comparable (see Tables 6

Table 5. The Classification Performance of Different Feature Sets Using Linear Classifiers with the HHR Split Procedure

Feature Sets	LS2SP		FLDA		LSVM	
	A_z	SE	A_z	SE	A_z	SE
Shape	0.91	0.03	0.91	0.03	0.92	0.02
Gradient	0.67	0.06	0.66	0.07	0.66	0.07
Texture	0.62	0.06	0.61	0.07	0.65	0.06
All	0.83	0.09	0.84	0.05	0.89	0.04
$FS_{\text{alignment}}$	0.92	0.03	0.92	0.03	0.93	0.02
$FS_{\text{separability}}$	0.92	0.03	0.92	0.03	0.93	0.03
FS_{distance}	0.92	0.03	0.92	0.03	0.92	0.03
FS_{best1}	0.87	0.04	0.89	0.03	0.91	0.03
FS_{best2}	0.89	0.04	0.91	0.02	0.92	0.03

Each A_z value shown is the average of the A_z values over 100 HHR split classification experiments

SE = standard error

Table 6. The Classification Performance of Different Feature Sets Using the Kernel-based Classifiers with the LOO Procedure

Feature Sets	KS2SP		KFDA		KSVM	
	A_z	SE	A_z	SE	A_z	SE
Shape	0.92	0.03	0.93	0.02	0.91	0.03
Gradient	0.68	0.05	0.69	0.05	0.65	0.06
Texture	0.74	0.05	0.68	0.05	0.66	0.05
All	0.92	0.03	0.92	0.03	0.92	0.03
$FS_{\text{alignment}}$	0.95	0.02	0.93	0.02	0.94	0.02
$FS_{\text{separability}}$	0.94	0.02	0.93	0.02	0.94	0.02
FS_{distance}	0.94	0.02	0.93	0.02	0.93	0.02
FS_{best1}	0.94	0.02	0.93	0.02	0.93	0.02
FS_{best2}	0.94	0.02	0.93	0.02	0.94	0.02

SE = standard error

Table 7. The Classification Performance of Different Feature Sets Using the Kernel-based Classifiers with the HHR Split Procedure

Feature Sets	KS2SP		KFDA		KSVM	
	A_z	SE	A_z	SE	A_z	SE
Shape	0.92	0.03	0.92	0.03	0.93	0.02
Gradient	0.65	0.08	0.67	0.07	0.68	0.06
Texture	0.68	0.06	0.64	0.07	0.69	0.05
All	0.93	0.02	0.93	0.02	0.93	0.03
$FS_{\text{alignment}}$	0.93	0.02	0.93	0.02	0.94	0.03
$FS_{\text{separability}}$	0.94	0.03	0.93	0.03	0.94	0.02
FS_{distance}	0.94	0.03	0.93	0.03	0.94	0.03
FS_{best_1}	0.94	0.03	0.94	0.02	0.94	0.02
FS_{best_2}	0.94	0.02	0.94	0.02	0.94	0.02

Each A_z value shown is the average of the A_z values over 100 HHR split classification experiments

SE = standard error

and 7). Observations similar to those described in Section *Linear Classification* can also be made from Tables 6 and 7. The best nonlinear classification performance achieved is 0.95 in A_z value.

Comparative Analysis

We compared the classification performance of the three pairs of linear classifiers and their corresponding nonlinear versions with the Gaussian kernel (LS2SP/KS2SP, FLDA/KFDA, and LSVM/KSVM) and the performance of the selected combinations of shape, edge-sharpness, and texture features with that of the shape features on their own. Comparison of the ROC curves for pairs of the linear vs. kernel-based classifiers is shown in Figure 2 using the feature set FS_{best_2} , as well as the ROC curves of the feature set $FS_{\text{alignment}}$ and

Fig 2. ROC curves of different pairs of classifiers (or feature sets) with the LOO procedure.

Fig 3. ROC curves of KFDA, KSVM, and the KS2SP classifier using the features set FS_{best2} with the LOO procedure.

the five shape features on their own, all evaluated with the LOO procedure. Comparison of the ROC curves obtained by the three kernel-based classifiers is shown in Figure 3 using the feature set FS_{best2} , evaluated with the LOO procedure. The average A_z values over the nine feature sets with the LOO and HHR procedures, the training time for the feature set $FS_{\text{separability}}$ with the HHR split procedure, and the number of parameters required to be specified for each classifier are recorded in Table 8.

By using the ROCKIT software,⁴⁹ we also performed the area test⁵⁰ with the features set FS_{best2} to test the statistical significance of the differences between two ROC curves, for various kernel-based classifiers (KS2SP, KFDA, and KSVM), for the linear vs. kernel-based classifiers, and for the selected vs. the solely shape-based feature sets, based on the LOO procedure. The area test is a univariate z -score test of the difference between the areas under the two ROC curves (i.e., the difference in the overall diagnostic perfor-

Table 8. Comparison of the Linear and Kernel-based Classifiers Studied

Classifiers	A_z (LOO)	A_z (HHR)	Time (Second)	Parameters
LS2SP $\rightarrow$ KS2SP	0.84 $\rightarrow$ 0.89	0.84 $\rightarrow$ 0.87	0.80 $\rightarrow$ 0.93	0 $\rightarrow$ 1(σ)
FLDA $\rightarrow$ KFDA	0.86 $\rightarrow$ 0.87	0.84 $\rightarrow$ 0.87	0.65 $\rightarrow$ 0.68	0 $\rightarrow$ 1(σ)
LSVM $\rightarrow$ KSVM	0.85 $\rightarrow$ 0.87	0.86 $\rightarrow$ 0.88	2.21 $\rightarrow$ 2.56	1(ξ) $\rightarrow$ 2(ξ, σ)

The symbol " $\rightarrow$ " indicates the advantage to be gained or the additional cost as the linear version of a classifier is replaced by its kernel-based version

Table 9. Statistical Analysis of the Difference Between the ROC Curves of Pairs of Various Kernel-based Classifiers

	KS2SP (Group 1) vs. KFDA (Group 2)	KS2SP (Group 1) vs. KSVM (Group 2)	KFDA (Group 1) vs. KSVM (Group 2)
95% CI_1	[0.89, 0.97]	[0.89, 0.97]	[0.86, 0.97]
95% CI_2	[0.86, 0.97]	[0.88, 0.97]	[0.88, 0.97]
z-score	0.90	-0.57	-0.85
p value	0.37	0.57	0.39
95% CI_d	[-0.012, 0.033]	[-0.008, 0.004]	[-0.031, 0.012]

CI_1 denotes the asymmetric 95% CI for A_z of the classifiers in group 1. CI_2 denotes the asymmetric 95% CI for A_z of the classifiers in group 2. CI_d denotes the approximate 95% CI for the difference of A_z between each pair of classifiers

CI = confidence interval

mance of the two tests), of which a null hypothesis indicates the data sets arose from binormal ROC curves with equal areas beneath them.^{51,52} The computed value of the correlated test statistic (observed z score), the corresponding two-tailed p value, and the approximate 95% confidence interval for the difference are recorded in Tables 9, 10, and 11 for different pairs of classifiers and feature sets. The following observations can be made from the tables and figures mentioned above:

- The Gaussian kernel improves the performance of the linear classifiers, such as the LS2SP and LSVM, which indicates that advantages in classification accuracy may be gained by embedding kernel functions in the classifier (see Tables 8 and 10 and Fig. 2).
- The selected combinations of shape, edge-sharpness, and texture features improve the classification performance as compared to that of the shape features on their own (see Tables 6, 7, and 11 and Fig. 2).
- There is no statistically significant difference between the performance of the three kernel-based classifiers (see Tables 8 and 9 and Fig. 3).

- Kernel-based classifiers take longer training time than the corresponding linear classifiers (see Table 8).
- The SVM possesses lower training speed than the S2SP classifier and FLDA/KFDA (see Table 8).
- The S2SP classifier and FLDA/KFDA are more convenient for users than classifiers with a regularization parameter, such as SVMs, as there are fewer parameters required to be specified or predetermined (see Table 8).

Overall, the KS2SP classifier, based upon the S2SP classifier that we have recently proposed, provides the best performance with low computational cost and complexity of training.

DISCUSSION AND CONCLUSION

We have investigated and presented the results of classification of breast masses with a set of 111 regions in mammograms, with 46 related to malignant tumors and 65 to benign masses, each represented with 22 features including 5 shape factors, 3 edge-sharpness measures, and 14 texture

Table 10. Statistical Analysis of the Difference Between the ROC Curves of Pairs of Linear vs. Kernel-based Classifiers

	LS2SP (Group 1) vs. KS2SP (Group 2)	FLDA (Group 1) vs. KFDA (Group 2)	LSVM (Group 1) vs. KSVM (Group 2)
95% CI_1	[0.81, 0.94]	[0.87, 0.96]	[0.85, 0.96]
95% CI_2	[0.89, 0.97]	[0.86, 0.97]	[0.88, 0.97]
z-score	-2.23	-0.10	-2.23
p value	0.03	0.92	0.03
95% CI_d	[-0.102, -0.007]	[-0.021, 0.019]	[-0.047, -0.003]

CI_1 denotes the asymmetric 95% CI for A_z of the classifiers in group 1. CI_2 denotes the asymmetric 95% CI for A_z of the classifiers in group 2. CI_d denotes the approximate 95% CI for the difference of A_z between each pair of classifiers

CI = confidence interval

Table 11. Statistical Analysis of the Difference Between the ROC Curves of Pairs of Various Feature Sets

	Shape (Group 1) vs. All (Group 2)	$FS_{\text{alignment}}$ (Group 1) vs. Shape (Group 2)	FS_{best_1} (Group 1) vs. Shape (Group 2)
95% CI_1	[0.84, 0.96]	[0.90, 0.98]	[0.89, 0.97]
95% CI_2	[0.85, 0.96]	[0.84, 0.96]	[0.84, 0.96]
z-score	-1.37	1.36	0.96
p value	0.17	0.18	0.34
95% CI_d	[-0.017, 0.003]	[-0.012, 0.065]	[-0.018, 0.052]

CI_1 denotes the asymmetric 95% CI for A_z of the classifiers in Group 1. CI_2 denotes the asymmetric 95% CI for A_z of the classifiers in Group 2. CI_d denotes the approximate 95% CI for the difference of A_z between each pair of classifiers

CI = confidence interval

features. Before the classification stage, feature selection independent of any classifier was performed by GA based on three measures of data separability, including alignment of the kernel with the target function, class separability, and normalized distance. Three linear classifiers, including the classical pattern classification method of FLDA, the popular SVM based on the maximal margin rule, and the S2SP classifier that we have recently developed, as well as their corresponding nonlinear versions with the Gaussian kernel, were employed to perform the classification task.

Sahiner et al.⁶ obtained $A_z=0.87\pm0.02$ with a set of 249 films from 102 patients using a combination of morphological and texture features with stepwise feature selection and linear discriminant analysis; the result was improved to $A_z=0.91\pm0.02$ when the leave-one-case-out discriminant scores from different views of a mass were combined to obtain a summary score. One of our own previous studies²⁷ with a subset of 57 masses (from the set of 111 masses used in the present study), using a combination of shape, edge-sharpness, and texture features with the kernel partial least squares transformation, achieved classification accuracy in the range of $A_z=[0.99, 1.0]$ using FLDA, but with low robustness with respect to variations in the associated parameters. In the present study, with a more diverse data set of 111 masses, with the GA-selected feature combination $FS_{\text{alignment}}$, the linear classification performance reached 0.93 in A_z value using SVM, and the nonlinear classification performance reached 0.95 in A_z value using the S2SP classifier with good robustness around the associated parameters.

Principal component analysis (PCA) is a well-known feature dimensionality reduction tool,

which concentrates on significant projections of features; however, PCA does not always improve the classification performance of most classifiers. Furthermore, PCA does not identify specific features as the best-performing features. Compared with PCA, the feature selection methods used in the present study not only reduce the dimensionality of the features provided, but they also improve the classification performance of several classifiers and identify the strong features. Feature selection based on the performance of a specific classifier can determine a set of feature combinations with high classification performance; however, the results cannot always be extended to other classifiers, especially to nonlinear classifiers with different logic and structure. The methods used for feature selection in the present study are independent of the classifier; the selected combinations are suitable for use with several different classifiers. A limitation of the proposed approaches to feature selection is that the correlation existing between the given features is not accounted for, and consequently, the selected combinations of features may contain more than the minimal set of features that could provide similar classification performance. The increased classification accuracy values obtained indicate that the incorporation of features representing multiple radiological characteristics, such as edge sharpness, texture, and shape, instead of shape alone, could lead to improved representation and analysis of breast masses in mammograms by using advanced kernel-based classifiers associated with feature selection using GA based on measures of data separability. The proposed methods should find application in computer-aided detection and diagnosis of breast cancer.

ACKNOWLEDGMENT

T. Mu would like to acknowledge financial support from the Overseas Research Students Awards Scheme, UK; the Hsiang Su Coppin Memorial Scholarship Fund; and the University of Liverpool, UK. We thank the Catalyst Program of Research Services and the University Research Grants Committee of the University of Calgary, Canada, and the Medical Research Council (the Interdisciplinary Bridging Awards), UK, for financial support.

REFERENCES

- Duijm L, Groenewoud JH, Jansen FH, Fracheboud J, Beek M, de Koning HJ: Mammography screening in the Netherlands: delay in the diagnosis of breast cancer after breast cancer screening. *Br J Cancer* 91:1795–1799, 2004
- Breast Expert Workgroup, Cancer Detection Section, California Department of Health Service: Breast Cancer Diagnostic Algorithms for Primary Care Providers, 2005, 3rd edition. http://qap.sdsu.edu/screening/breastcancer/bda/pdf/CDS_Algorithms_2005a.pdf
- Cady B, Chung M: Mammographic screening: no longer controversial. *Am J Clin Oncol* 28(1):1–4, 2005
- Elmore JG, Armstrong K, Lehman CD, Fletcher SW: Screening for breast cancer. *J Am Med Assoc* 293(10):1245–1256, 2005
- Bruce LM, Adhami RR: Classifying mammographic mass shapes using the wavelet transform modulus-maxima method. *IEEE Trans Med Imaging* 18(12):1170–1177, 1999
- Sahiner BS, Chan HP, Petrick N, Helvie MA, Hadjiiski LM: Improvement of mammographic mass characterization using spiculation measures and morphological features. *Med Phys* 28(7):1455–1465, 2001
- Mudigonda NR, Rangayyan RM, Desautels JEL: Gradient and texture analysis for the classification of mammographic masses. *IEEE Trans Med Imaging* 19(10):1032–1043, 2000
- Mudigonda NR, Rangayyan RM, Desautels JEL: Detection of breast masses in mammograms by density slicing and texture flow field analysis. *IEEE Trans Med Imaging* 20(12):1215–1227, 2001
- Pisano E Ed.: Proceedings of the 7th International Workshop on Digital Mammography. Durham, NC, June 2004
- Guo Y, Sivaramakrishna R, Lu Ch, Suri JS, Laxminarayan S: Breast image registration techniques: a survey. *Med Biol Eng Comput* 44(1–2):15–26, 2006
- Morton MJ, Whaley DW, Brandt KR, Amrami KK: Screening mammograms: interpretation with computer-aided detection^a prospective evaluation. *Radiology* 239:375–383, 2006
- Oliver A, Freixenet J, Marti R, Pont J, Perez E, Denton ERE, Zwiggelaar R: A novel breast tissue density classification methodology. *IEEE Trans Inf Technol Biomed* DOI: 10.1109/TITB.2007.903514, 2007
- Rangayyan RM, Nguyen TM: Fractal analysis of contours of breast masses in mammograms. *J Digit Imaging* 20(3):223–237, 2007
- Rangayyan RM, El-Faramawy NM, Desautels JEL, Alim OA: Measures of acutance and shape for classification of breast tumors. *IEEE Trans Med Imaging* 16(6):799–810, 1997
- Rangayyan RM, Mudigonda NR, Desautels JEL: Boundary modelling and shape analysis methods for classification of mammographic masses. *Med Biol Eng Comput* 38(5):487–496, 2000
- Cascio D, Fauci F, Magro R, Raso G, Bellotti R, De Carlo F, Tangaro S, De Nunzio G, Quarta M, Forni G, Lauria A, Fantacci ME, Retico A, Masala GL, Oliva P, Bagnasco S, Cheran SC, Torres EL: Mammogram segmentation by contour searching and mass lesions classification with neural network. *IEEE Trans Nucl Sci* 53(5):2827–2833, 2006
- Doi K: Diagnostic imaging over the last 50 years: research and development in medical imaging science and technology. *Phys Med Biol* 51:R5–R27, 2006
- Doi K: Computer-aided diagnosis in medical imaging: historical review, current status and future potential. *Comput Med Imaging Graph* 31:198–211, 2007
- Rangayyan RM, Ayres FJ, Desautels JEL: A review of computer-aided diagnosis of breast cancer: toward the detection of subtle signs. *J Franklin Inst* 344(3–4):312–348, 2007
- Dominguez AR, Nandi AK: Improved dynamic-programming-based algorithms for segmentation of masses in mammograms. *Med Phys* 34(11):4256–4269, 2007
- El-Naqa I, Yang Y, Wernick MN, Galatsanos NP, Nishikawa RM: A support vector machine approach for detection of microcalcifications in mammograms. *IEEE Trans Med Imaging* 21(12):1552–1563, 2002
- Wei L, Yang Y, Nishikawa RM, Wernick MN, Edwards A: Relevance vector machine for automatic detection of clustered microcalcifications. *IEEE Trans Med Imaging* 24(10):1278–1285, 2005
- André TCSS, Rangayyan RM: Classification of breast masses in mammograms using neural networks with shape, edge sharpness, and texture features. *J Electron Imaging* 15(1):1–10, 2006 (article no. 013019)
- Alto H, Rangayyan RM, Desautels JEL: Content-based retrieval and analysis of mammographic masses. *J Electron Imaging* 14(2):1–17, 2005 (article no. 023026)
- Nandi RJ, Nandi AK, Rangayyan RM, Scutt D: Classification of breast masses in mammograms using genetic programming and feature selection. *Med Biol Eng Comput* 44(8):693–694, 2006
- Wei J, Chan H-P, Sahiner B, Hadjiiski LM, Helvie MA, Roubidoux MA, Zhou C, Ge J: Dual system approach to computer-aided detection of breast masses on mammograms. *Med Phys* 33(11):4157–4168, 2006
- Mu T, Nandi AK, Rangayyan RM: Classification of breast masses via nonlinear transformation of features based on a kernel matrix. *Med Biol Eng Comput* 45(8):769–780, 2007
- Homer MJ: Mammographic Interpretation: A Practical Approach, 2nd edition. Boston: McGraw-Hill, 1997
- Sahiner BS, Chan H-P, Petrick N, Helvie MA, Goodsitt MM: Computerized characterization of masses on mammograms: the rubber band straightening transform and texture analysis. *Med Phys* 25(4):516–526, 1998
- Buckles BP: Genetic Algorithms. Los Alamitos: IEEE Computer Society Press, 1992
- Cristianini N, Kandola J, Elisseeff A, Shawe-Taylor J: On Optimizing Kernel Alignment. Technical Report NC-TR-01-087. London: Royal Holloway University of London, 2001.
- Xiong H, Swamy MNS, Ahmad MO: Optimizing the kernel in the empirical feature space. *IEEE Trans Neural Netw* 16(2):460–474, 2005

33. Swain PH: Fundamentals of pattern recognition in remote sensing. In: Swain PH, Davis SM Eds. *Remote Sensing: The Quantitative Approach*. New York: McGraw-Hill, 1978, pp 136–187
34. Aizerman M, Braverman E, Rozonoer L: Theoretical foundations of the potential function method in pattern recognition learning. *Autom Remote Control* 25:821–837, 1964
35. Boser BE, Guyon IM, Vapnik VN: A training algorithm for optimal margin classifiers. In: *Proc. of the 5th Annual ACM Workshop on Computational Learning Theory*, 1992, pp 144–152
36. Fisher RA: The use of multiple measurements in taxonomic problems. *Ann Eugen* 7(2):179–188, 1936
37. Mika S, Rätsch G, Weston J, Schölkopf B, Muller K: Fisher discriminant analysis with kernels. In: *Proc. of IEEE Neural Networks for Signal Processing Workshop*, 1999, pp 41–48
38. Mu T, Nandi AK, Rangayyan RM: Strict 2-surface proximal classifier with application to breast cancer detection in mammograms. In: *Proc. of the 32nd Int'l Conf. on Acoustics, Speech, and Signal Processing, ICASSP*, volume 2, Honolulu, HI, April 2007, pp 477–480
39. Mu T, Nandi AK, Rangayyan RM: Strict 2-surface proximal classification of knee-joint vibroarthrographic signals. In: *Proc. of the 29th Annual Int'l Conf. of the IEEE Engineering in Medicine and Biology Society, EMBC*, Lyon, France, August 2007, pp 4911–4914
40. Alberta Cancer Board: Test: Alberta Program for the Early Detection of Breast Cancer. 2001/03 Biennial Report, Edmonton, Alberta, Canada, 2004. http://www.cancerboard.ab.ca/screenest/downloads/screenest_biennial_2001-03.pdf
41. The Mammographic Image Analysis Society digital mammogram database. Technical report. Imaging Science and Biomedical Engineering (ISBE): <http://www.isbe.man.ac.uk>
42. Haralick RM, Shanmugam K, Dinstein I: Textural features for image classification. *IEEE Trans Syst Man Cybern* SMC-3(6):610–622, 1973
43. Chipperfield AJ, Fleming PJ, Pohlheim H, Fonseca CM: *Genetic Algorithm Toolbox for use with MATLAB* (version 1.2). Sheffield: University of Sheffield, 1994
44. Duda RO, Hart PE, Stork DG: *Pattern Classification*, 2nd edition. New York: Wiley and Sons, 2001
45. Cortes C, Vapnik V: Support-vector networks. *Mach Learn* 20(3):273–297, 1995
46. Mangasarian OL, Wild EW: Multisurface proximal support vector machine classification via generalized eigenvalues. *IEEE Trans Pattern Anal Mach Intell* 28:69–74, 2006
47. Shawe-Taylor J, Cristianini N: *Kernel Methods for Pattern Analysis*. Cambridge: Cambridge University Press, 2004
48. Pelckmans K, Suykens JAK, Gestel TV, Brabanter JD, Lukas L, Hamers B, Moor BD, Vandewalle J: *LS-SVM lab1.5: Least squares support vector machines*. Katholieke University, Leuven, Belgium, 2003. Available at <http://www.esat.kuleuven.ac.be/sista/lssvmlab/toolbox.html>
49. Metz CE: ROCKIT 1.1B2 Beta version for WINDOWS operating system. Kurt Rossmann Laboratories, Department of Radiology, The University of Chicago, 2006. Available at http://www-radiology.uchicago.edu/krl/KRL_ROC/software_index6.htm
50. Metz CE, Wang P-L, Kronman HB: A new approach for testing the significance of differences between ROC curves measured from correlated data. In: Deconinck F Ed. *Information Processing in Medical Imaging*. The Hague: Martinus Nijhoff, 1984, pp 432–445
51. Metz CE, Herman BA, Shen J-H: Maximum-likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data. *Stat Med* 17:1033–1053, 1998
52. Dorfman DD, Alf E: Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals: χ^2 rating method data. *J Math Psychol* 6:487–496, 1969