Max-AUC Feature Selection in Computer-Aided Detection of Polyps in CT Colonography

Jian-Wu Xu, *Member, IEEE*, and Kenji Suzuki, *Senior Member, IEEE*

***Abstract*—We propose a feature selection method based on a sequential forward floating selection (SFFS) procedure to improve the performance of a classifier in computerized detection of polyps in CT colonography (CTC). The feature selection method is coupled with a nonlinear support vector machine (SVM) classifier. Unlike the conventional linear method based on Wilks' lambda, the proposed method selected the most relevant features that would maximize the area under the receiver operating characteristic curve (AUC), which directly maximizes classification performance, evaluated based on AUC value, in the computer-aided detection (CADe) scheme. We presented two variants of the proposed method with different stopping criteria used in the SFFS procedure. The first variant searched all feature combinations allowed in the SFFS procedure and selected the subsets that maximize the AUC values. The second variant performed a statistical test at each step during the SFFS procedure, and it was terminated if the increase in the AUC value was not statistically significant. The advantage of the second variant is its lower computational cost. To test the performance of the proposed method, we compared it against the popular stepwise feature selection method based on Wilks' lambda for a colonic-polyp database (25 polyps and 2624 nonpolyps). We extracted 75 morphologic, gray-level-based, and texture features from the segmented lesion candidate regions. The two variants of the proposed feature selection method chose 29 and 7 features, respectively. Two SVM classifiers trained with these selected features yielded a 96% by-polyp sensitivity at false-positive (FP) rates of 4.1 and 6.5 per patient, respectively. Experiments showed a significant improvement in the performance of the classifier with the proposed feature selection method over that with the popular stepwise feature selection based on Wilks' lambda that yielded 18.0 FPs per patient at the same sensitivity level.**

***Index Terms*—Colonic polyps, computer-aided detection (CADe), feature selection, support vector machines (SVMs).**

I. Introduction

COLORECTAL cancer is one of the leading causes of mortality due to cancer in the United States [1]. Early detection is critical in reducing the risk of death due to colon cancer. However, early detection of polyps in CTC is difficult because of the similar appearance of various nonlesions. Therefore, there has been a great interest in the development of computer-aided detection (CADe) schemes for early detection of polyps in CTC to improve the detection sensitivity and specificity [2]–[4].

A CADe scheme generally consists of candidate detection followed by supervised classification [5]. The task of candidate detection is to achieve high detection sensitivity by including as many suspicious lesions as possible. A common approach to classification in a CADe scheme is to extract many texture, gray-level-based, geometric, and other features based on domain knowledge. However, not all of these extracted features might be helpful in discriminating lesions from nonlesions. Therefore, in the design of an effective classifier, it is critical to select the most discriminant features to differentiate lesions from nonlesions.

Feature selection has long been an active research topic in machine learning [6]–[8], because it is one of the main factors that determines the performance of a classifier. In the context of the CADe research field, one of the most popular feature selection methods is the stepwise feature selection based on Wilks' lambda coupled with linear discriminant analysis (LDA). The method has been applied in various CADe schemes because of its simplicity and effectiveness [9], [10]. Recently, feature ranking techniques have been applied for selection of relevant and informative features in CADe schemes [11], [12]. Campadelli *et al.* used the univariate Golub statistics to order individual features extracted from chest radiographs and chose a certain number of features with the highest positive and negative values [11]. Mutual information has been used to identify features that were highly correlated with the pathologic state of tissues from the trans-rectal ultrasound images in CADe of prostate cancer [12]. On the other hand, deterministic and stochastic feature selection methods were extensively employed for searching feature subset in the machine learning field. One of the most widely used deterministic feature searching approaches is the sequential forward floating selection (SFFS) and sequential backward floating selection (SBFS) [13]. SBFS has been used for selecting input features for artificial neural networks [14], [15]. SFFS has been used to search relevant features combined with various classifiers such as Naïve Bayes, a k-nearest-neighbor classifier, support vector machines (SVMs) [16], and AdaBoost [17], in different CADe schemes. Stochastic searching methodology consists of a genetic algorithm, particle swarm optimization, and others. A genetic algorithm has been used in lung nodule CADe [18] and in detecting pulmonary embolisms in CT images [19]. Mohamed and Salama applied particle swarm optimization in spectral multifeature analysis CADe of prostate cancer in trans-rectal ultrasound images [20].

Feature searching methods are classifier- and criterion-dependent. Different classifiers would select different sets of features given the same criterion. On the other hand, different selection criteria could result in distinctive feature sets even based on the same classifier. In the literature, classification

Manuscript received March 2, 2013; revised June 23, 2013; accepted August 5, 2013. Date of publication August 15, 2013; date of current version March 3, 2014. This work was supported in part by the National Cancer Institute/National Institutes of Health under Grant R01CA120549 and in part by the NIH under Grant S10 RR021039 and Grant P30 CA14599.

The authors are with the Department of Radiology, University of Chicago, Chicago, IL 60637 USA (e-mail: jwxu@uchicago.edu; suzuki@uchicago.edu).

Digital Object Identifier 10.1109/JBHI.2013.2278023

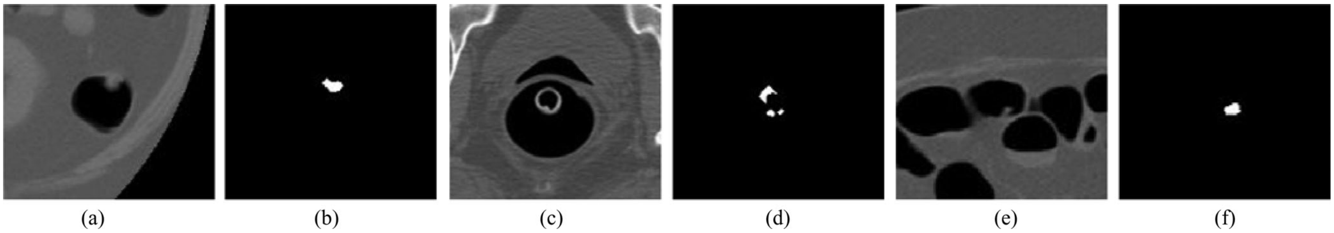


Fig. 1. Representative polyp and nonlesion detections and their corresponding segmented regions. (a) A true polyp; (c) a nonpolyp (rectal tube); (e) a nonpolyp (haustral fold); [(b), (d), and (f)] the corresponding segmented candidates.

accuracy [16], false-positive (FP) elimination rate [18], mean sensitivity of the free-response receiver operating characteristic (FROC) curve [21], pseudo-loss in the AdaBoost algorithm [17], and other general performance measures have been employed as the selection rules. However, a low FP rate at a high sensitivity region is necessary in order for a CADe scheme to be useful in clinical practice. The AUC value has been widely used in evaluation of CADe schemes in the literature [5]. The mean sensitivity criterion only measures the average sensitivity value in a predefined specificity range [21], which does not quantify how a CADe scheme performs in general as the AUC criterion does. In the machine learning community, the AUC value has also been used as a criterion for optimizing classifiers. Rakotomamonjy has proposed a novel form of an SVM by approximately maximizing the AUC value [22]. Marrocco *et al.* used a nonparametric linear classifier to maximize the AUC value [23]. These two methods did not involve feature selection. All features were used in the optimization of classifiers. Feature selections based on ranking [24] and perturbation [25] have been employed for the maximization of the AUC value in microarray and gene expression applications. However, feature ranking and perturbation considered only individual feature characteristics and did not take into account the collective discriminative power of feature combinations. In the classification, the collective discriminative power of combining multiple features matters most.

In this paper, we propose a feature-selection method that directly maximizes the AUC value for a CADe scheme coupled with a nonlinear SVM classifier. To test the performance of the proposed feature selection method, we compared it against the popular stepwise feature selection based on Wilks' lambda in CADe of polyps in CTC.

II. Materials

The CTC cases used in this study were acquired retrospectively at the University of Chicago Medical Center. The database consisted of 206 CTC datasets obtained from 103 patients. Each patient followed the standard CTC procedure with precolonoscopy cleansing and colon insufflation with room air or carbon dioxide. Fecal tagging was not employed in the CTC protocol. Both supine and prone positions were scanned with a multi-detector-row CT scanner (LightSpeed QX/i, GE Medical Systems, Milwaukee, WI) with collimations between 2.5 and 5.0 mm, reconstruction intervals of 1.25–5.0 mm, and tube currents of 60–120 mA with 120 kVp. Each reconstructed CT section had a matrix size of 512 $\times$ 512 pixels, with an in-plane pixel size of 0.5–0.7 mm. Optical colonoscopy was also performed for all patients. In this study, we used 5 mm as the lower limit on the clinically important size of polyps. The locations of polyps were confirmed by an expert radiologist based on CTC images, and pathology and colonoscopy reports. Fourteen patients had 25 colonoscopy-confirmed polyps, 11 of which were 5–9 mm and 14 were 10–25 mm in size. The dataset has been used in a previous study [10].

A lesion candidate detection algorithm was applied to the database. The initial detection algorithm was composed of 1) automatic knowledge-guided colon segmentation and 2) detection of polyp candidates based on the shape index and curvedness of the segmented colon [10]. The initial detection step missed one polyp and detected two polyps only in one view (supine or prone), yielding 24 detected lesions with 46 views, while detected 2624 nonlesions. The major sources of nonlesions included rectal tubes, stool, haustral folds, colonic walls, and the ileocecal valve. Therefore, the initial candidate detection algorithm achieved a 96% (24/25) by-polyp sensitivity with 25.5 (2624(103) FPs (i.e., nonlesions) per patient. Fig. 1 shows a representative polyp and two typical nonlesion detections and their corresponding segmented regions. Because the detection criterion was based on the shape index and curvedness, rectal tubes and haustral folds were typical FPs because of their similar shape appearances. A part of a rectal tube often exhibits a cap-like shape that is very similar to a part of a small polyp in appearance as shown in Fig. 1(c). Part of the rectal tube was falsely detected as a polyp with the segmented contour given in Fig. 1(d). A haustral fold produces large curvedness values, as does a polyp. Fig. 1(e) shows a typical haustral fold that was falsely detected as a polyp candidate because of its large curvedness. Fig. 1(f) presents the corresponding segmented region that has large curvedness values.

III. Methods

The structure of our proposed feature selection method coupled with a linear/nonlinear classifier is depicted in Fig. 2. The classification step consists of three major components: feature extraction from lesion candidates, SFFS feature selection based on the maximal AUC value criterion, and an SVM classifier operated on the optimal feature subset. The feature selection stage only occurred in the design stage. Once the optimal feature set was selected, the classification stage only consisted of feature extraction and the SVM classifier. The SVM classifier would classify each suspicious candidate into a lesion or a nonlesion, so that nonlesions from the previous detection step could be reduced while a high sensitivity would be maintained.

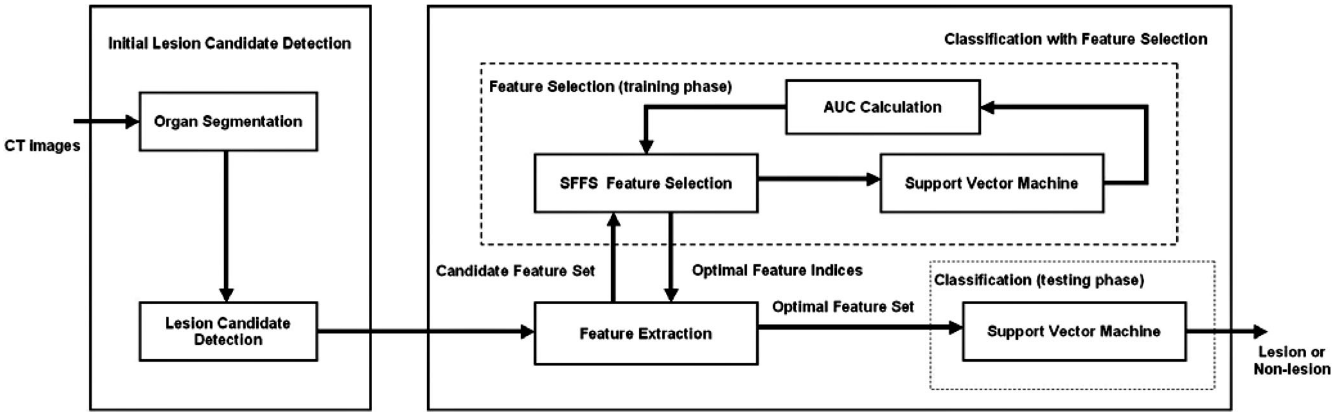


Fig. 2. Proposed feature selection in classification and the initial lesion candidate detection in a CADe scheme.

A. Feature Extraction

Feature extraction is one of the most important steps in a classification stage. We extracted 75 two-dimensional (2-D) and three-dimensional (3-D) morphologic, gray-level-based, and texture features from detected lesion candidates in CT images to form an initial feature set. 2-D features were calculated in the axial slice where the segmented candidate region had the largest area. 3-D features were computed in the overall segmented volume.

To compute features such as the contrast between a segmented candidate region and its outside, we created a ring structure for a 2-D case and a shell structure for a 3-D case surrounding a detected candidate, denoted as the band region. We performed a binary dilation operation on the detected candidates with a square-structuring element of 21 × 21 pixels and 11 × 11 pixels (a cube of 21 × 21 × 21 and 11 × 11 × 11 voxels for a 3-D case) [26]. The difference between the output dilated regions would be the final band regions. Therefore, the outside region was defined as a ring (a shell for a 3-D case) with a width of 5 pixels and 5 pixels away from the boundary of the detected candidate.

Gray-level information characterized lesion intensity information. Shape information such as radial and tangential gradient indices inside the lesions and in the band regions were computed. To make these features meaningful and discriminant, the delineation of the candidates is required to correspond closely to the real object boundaries. This, in turn, requires the accuracy of the hysteresis thresholding and clustering method employed in the detection of polyps. Histogram-based features were extracted to specify the range, distribution, and overlap of the voxel values in gray-level and edge-enhanced images inside and outside the delineated candidates.

B. Support Vector Machines

We used SVMs [27] as the classifier in our CADe scheme. SVMs are a machine-learning technique that maximizes the margin of separation between positive and negative classes. Given a set of N training data points $\{(x_i, y)\}_{i=1}^{N}$, where x_i is the feature vector with $x_i \in \Re^L$, and y_i is the class label with $y_i \in \{-1, 1\}$, the decision function for the SVM classifier can be written as

$$f(x) = \sum_{i=1}^{N} \alpha_i y_i K(x_i, x) + \alpha_0. \quad (1)$$

The parameters $\alpha_i \geq 0$ are called Lagrange multipliers that are optimized through quadratic programming. $K(x_i, x_j)$ is a symmetric nonnegative inner-product kernel. In the applications of SVMs, popular kernel functions include

dth degree polynomial function:

$$K(x_i, x_j) = \left(1 + x_i^{\mathbf{T}} x_j\right)^d \quad (2)$$

Gaussian kernel function:

$$K(x_i, x_j) = \exp(-\|x_i - x_j\|^2 / 2\sigma^2). \quad (3)$$

The optimal Lagrange multipliers $\alpha_i \geq 0$ in the optimal decision boundary (1) is computed through the maximization of the following objective function:

$$\max_{\alpha_i} \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \quad (4)$$

subject to the following constraints:

$$\sum_{i=1}^{N} \alpha_i y_i = 0$$

$$C \geq \alpha_i \geq 0, \qquad \text{for } i = 1, 2, \ldots, N$$

where C is a user-specified positive parameter. As the SVM can be reformulated through the regularized function estimation problem with a *hinge* loss criterion [27], it can be shown that the SVM classifier has property of large margin and is robust against outliers.

C. Maximal AUC SFFS Feature Selection

The proposed maximal AUC SFFS feature selection method adopted the wrapper approach where the searching procedure was coupled with an SVM classifier to yield the AUC value for evaluation in each step [28]. We used the AUC value from the ROC as the selection criterion, because it directly measures

TABLE I
MAXIMAL AUC SFFS FEATURE SELECTION METHOD I

```
Maximal AUC SFFS Feature Selection I

Initialization:
    Full feature set from CT images X, selected feature set at step
$F_0=\{\emptyset\}$, predefined feature number $l=75$, $k=0$.

while $k<=l$
    $x^{+} = \arg\max_{x \in X - F_k} J(F_k + \{x\})$ (5)
    $F_{k+1} = F_k + \{x^{+}\}$
    $k=k+1$
    if $k>2$
        $x^{-} = \arg\max_{x \in F_k} J(F_k - \{x\})$ (6)
        while $J(F_k-\{x^-\})>J(F_{k-1})$ and $k>2$
            $F_{k-1}= F_k - x^{-}$
            $k=k-1$
            if $k>2$
                $x^{-} = \arg\max_{x \in F_k} J(F_k - \{x\})$
            end
        end
    end
end

Output:
    Selected feature set $Y_l$
```

how a CADe scheme performs in general. It has been shown that the AUC value corresponds to the probability of correctly identifying if a case is normal or abnormal [29]. From a statistical perspective, the AUC value is also equivalent to the well-known nonparametric Wilcoxon statistic [29]. These connections provide alternative views of the AUC value and make it a suitable measure of the performance of a CADe scheme. The SFFS procedure selects features based on the collective discriminative power of a combination of features. This is different from the feature ranking or perturbation approach, where the selection is based on the individual discriminative power of features [24], [25].

In this study, we proposed two variants of the maximal AUC SFFS feature selection method. The first variant, denoted as *MaxAUCSVM*, selects features and stops the procedure until all combinations of features allowed in the SFFS procedure have been examined. On the other hand, the second variant, denoted as *MaxAUCSVMStat*, applies the statistical test between AUC values obtained from adding or deleting features to determine the stopping criterion. We provide detailed descriptions in the following.

Table I outlines the main procedure of the first variant of the proposed feature selection method. *MaxAUCSVM* starts with an empty selected feature set F_0. Then it begins to include one feature at a time that would maximize the AUC value, calculated via an SVM classifier, of the selected feature subset given a subset size. This is given in (5) where the criterion $J(F_k+\{x\})$ is the AUC value of the SVM classification with the selection feature set ($F_k+\{x\}$). Therefore, (5) guarantees that the selected feature would produce the maximal AUC value with the combination of the existing features in the subset. However, this step only includes features without removing any existing ones. It might be possible to increase the AUC value by removing some features from the selected subset. This is realized in (6) and onward. It starts with the selected feature subset, and removes one feature at a time if the remaining feature subset performs better than the one containing the feature to be removed. The procedure continues until the number of features in the selected subset reaches the total number of available features. The feature subset with the maximal AUC value would be selected as the final output of the procedure.

One characteristic of the first variant is that the inclusion or exclusion of a particular feature is judged by the difference of two AUC values, regardless of whether the difference is statistically significant or not. Therefore, the procedure does not stop until it finishes searching all necessary combinations that are allowed in the SFFS framework. However, this approach inevitably increases the computational time. Moreover, it tends to include more features even though the increase in the AUC value is not statistically significant. Hence, the selected feature set makes the classifier less reliable, given the relatively small dataset usually used in the development of CADe schemes. To mitigate these two issues associated with the first variant, we proposed a second variant of our feature selection method, denoted as *MaxAUCSVMStat*. The main difference is the criterion used to include a particular feature into the selected set. Only if the increase in the AUC value by including a particular feature is statistically significant, the method chooses that feature. On the other hand, we do not impose this condition for feature deletion, i.e., if the decrease or increase in the AUC value is not statistically significant, we will delete that feature from the selected subset, because the feature subset becomes more compact by doing so.

To perform statistical testing on the difference of two AUC values, we used a binormal model to estimate the AUC value from the outputs of the SVM classifier [30]. Given the null hypothesis that the two outputs from the SVM classifier with two different selected feature subsets arose from ROC curves with equal areas beneath them, we calculated the z-score statistic [29], defined as

$$z = \frac{A_{z1} - A_{z2}}{\sqrt{\mathrm{var}(A_{z1}) + \mathrm{var}(A_{z2}) - 2\mathrm{cov}(A_{z1}, A_{z2})}} \tag{7}$$

where A_{z1} and $\mathrm{var}(A_{z1})$ refer to the estimated AUC value and variance associated with the case of selected feature subset one, respectively, A_{z2} and $\mathrm{var}(A_{z2})$ are the corresponding quantities for feature subset two, and $\mathrm{cov}(A_{z1}, A_{z2})$ is the estimated covariance between two cases. These quantifies were estimated via maximum likelihood estimation method [25]. The z-score statistic is then referred to tables of the standard normal distribution. The value of z above a threshold, e.g., $z > 1.96$, is considered as evidence that the null hypothesis has to be rejected, and hence the difference between two AUC values is statistically significant (two-tailed p-value < 0.05). We estimated the AUC value based on the binormal model as [30]

$$A_z = \Phi\left(\frac{a}{\sqrt{1+b^2}}\right) \tag{8}$$

where Φ is the cumulative probability function of a standard normal distribution function, a and b are the intercept and slope parameters, respectively, that specify an ROC graph in the normal-deviate coordinate. The maximum likelihood method was employed to estimate the two parameters.

The two variants of the proposed maximal AUC SFFS feature selection methods have their own merits. The first variant, *MaxAUCSVM*, is able to explore all possible feature combinations allowed in the SFFS procedure and select the one that achieves the maximum AUC value. However, this is obtained at the expense of excessive computational time. On the contrary, the second variant, *MaxAUCSVMStat*, aims at reducing the computational time with the selected feature subset of possibly a smaller AUC value, because it can happen that the increase in the AUC value becomes statistically significant by including more features as the SFFS procedure continues. Therefore, it is a tradeoff between performance and computational time.

D. Study Design and Performance Evaluation Criteria

The proposed feature selection method consists of parameter optimization for an SVM classifier in the training phase. Given the small sample size of our database, it is critical to apply appropriate strategies for training and testing the proposed method in order to avoid over-fitting. Cross-validation is a popular method to reduce the bias and overcome over-fitting in machine learning when the sample size is small. Based on the number of available cases in the database, we used a fivefold cross-validation method to estimate the AUC value for the candidate feature subsets chosen by the SFFS procedure at each step, and also to optimize the parameters in the SVM classifier. All of the lesions and nonlesions obtained from one case appeared in either training data or testing data. There was no crossover of one case (patient) belonging to both training and testing samples. The purpose was to eliminate the bias that results from testing of a classifier trained with data samples from the same patient.

After we optimized the kernel parameter for the SVM classifier, we applied the leave-one-lesion-out cross-validation method to perform feature selection and reported the final results of the trained SVM classifier with the selected feature set. We compared the proposed method to the popular stepwise feature selection method based on Wilks' lambda coupled with an LDA classifier. The proposed feature selection framework is very generic. In fact, other classifiers, such as an LDA classifier or an artificial neural network, can be used instead of an SVM classifier. Therefore, we replaced the SVM classifier with an LDA classifier in Fig. 2 and denoted the method as *MaxAUCLDA*, based on the first variant of the SFFS procedure [31].

TABLE II
AUC Values for Different Kernel Functions With Different Parameters in the SVM Classifier for Colon Database

Model parameters			AUC Value
Polynomial	d	2	0.89
		4	0.90
		6	0.91
		8	0.87
Gaussian	σ	0.05	0.92
		0.1	0.93
		0.2	**0.94**
		0.7	0.92
		1	0.9
		5	0.91
		10	0.85

IV. Results

A. Optimization of the SVM Classifier

We used a fivefold cross-validation method to choose an optimal kernel function with suitable parameters in the SVM classifier. In this study, we only focused on the polynomial (3) and Gaussian (4) kernel functions. Feature vectors were normalized to a range between 0 and 1. Table II presents the AUC values indicating the performance of the SVM with different kernel functions and parameters for the colon database. For each parameter, we applied the feature selection method, *MaxAUCSVM*, to choose an optimal set of feature vectors. The AUC values were obtained for the SVM classifier with the optimal feature subset in the fivefold cross-validation procedure. We experimented with four different values for the d parameter in the polynomial kernel function and seven different kernel widths for σ in the Gaussian function. The SVM classifier with a Gaussian kernel function performed better than that with a polynomial kernel function, which suggested that the decision boundaries between lesions and nonlesions were highly nonlinear. The AUC values reached a maximum with a Gaussian kernel function with $\sigma = 0.2$. We used the Gaussian kernel function with the optimal parameters for the SVM classifier in the following experiments.

B. Comparison of Different Feature Selection Methods

After the parameter optimization of the SVM classifiers, we applied the proposed feature selection methods in a leave-one-lesion-out cross-validation procedure. To have a fair comparison, we used the same cross-validation procedure for the feature selection based on Wilks' lambda and *MaxAUCLDA*. The feature selection procedure shown in Fig. 2 produced one set of features. Then, we applied cross-validation to report the classification performance.

Fig. 3 plots the AUC values versus different selected feature subset sizes from the first variant of our proposed feature

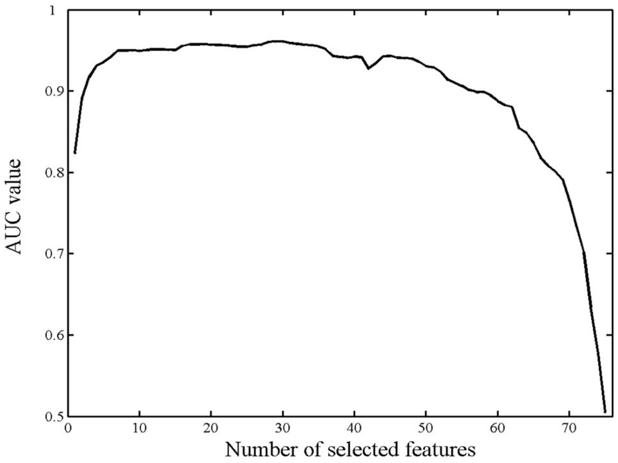


Fig. 3. AUC values versus different feature subsets selected by the proposed feature selection method, *MaxAUCSVM*.

selection method, *MaxAUCSVM*. As the number of selected features increases, the AUC value first increases, and it reaches its maximum when the feature subset size is 29. Then, the AUC value starts to decrease, which suggests that the added features cause the classifier performance to deteriorate. If we used all the extracted features without performing feature selection, the AUC value would be 0.51, which is slightly better than random guessing. This clearly illustrates the importance and necessity of feature selection in the classifier design to improve the overall performance of a CADe scheme. The first variant of the proposed feature selection method, *MaxAUCSVM,* explored all feature combinations in the SFFS procedure. The second variant, *MaxAUCSVMStat*, stopped the process at a feature subset size of 7, because the increase in the AUC value with a feature subset size of 8 was not statistically significant. The selected feature subset size and the corresponding AUC value achieved by the second variant of our proposed feature selection method were relatively smaller than those from the first variant. However, the advantage of the second variant over the first is a much lower computational cost.

To compare different selected feature subsets from different feature selection methods, we present the individual selected features, subset sizes, AUC values, and nonlesion reduction rate without removal of any lesion in Table III. The “X” mark denotes individual features selected by the method. The first variant of the proposed feature selection method, *MaxAUCSVM,* chose 29 features in total, out of which 25 were 3-D features and 4 were 2-D features. They include gray-level-based (such as feature numbers 3–7), shape-based (such as feature numbers 13 and 15), geometry-based (such as feature numbers 17 and 18), histogram-based (such as feature numbers 22 and 40), and other features. On the other hand, of the 7 features selected by the second variant, *MaxAUCSVMStat,* all were 3-D features, which suggests that 3-D features contain the most relevant and discriminatory information in distinguishing polyps from nonpolyps in CTC. These seven 3-D features include gray-level-based features on the contour of the candidate, sphere irregularity, and features derived from the edge-enhanced CT images. Note that 14 common features appear in the selected feature subsets by *MaxAUCSVM* and *MaxAUCLDA*. This accounts for around half of the selected features. However, their performance in terms of AUC values and nonlesion reduction rate without removal of any lesion is very different, which shows the substantial difference between a nonlinear SVM classifier and a LDA classifier. By comparing the feature subsets selected by *MaxAUCLDA* and the method based on Wilks’ lambda, we observe that there are 11 common features in total. The feature selection method, *MaxAUCLDA*, resulted in more than twice the features compared to that based on Wilks’ lambda. This observation suggests that different search procedures with different cost functions would have very different outcomes, even when the same classifiers were used.

C. *Performance Comparisons Among Different Feature Selection Methods*

Both variants of the proposed feature selection method yielded a much higher performance than did the ones based on Wilks’ lambda and *MaxAUCLDA*. The proposed *feature* selection methods achieved AUC values of 0.96 and 0.95, respectively, for the two variants, whereas the popular feature selection method based on Wilks’ lambda yielded an AUC value of 0.89. *MaxAUCLDA* produced an AUC value of 0.92. We performed statistical tests among different feature selection methods, as shown in Table IV. The results show that the differences in AUC values between the proposed feature selection methods and the other two (i.e., *MaxAUCLDA* and Wilks’ lambda) were statistically significant (with two-sided p-values < 0.05). However, the difference in AUC values between the two variants of the proposed feature selection methods was not statistically significant (with a two-sided p-value $= 0.06$). The FROC analysis provides more insights into the performance of different feature selection methods. Fig. 4 indicates that the first variant of the proposed feature selection method, *MaxAUCSVM*, was able to reduce 83.9% (2202/2624) of nonpolyps without removing any of the 24 polyps in a leave-one-lesion-out cross-validation test, i.e., a 96% (24/25) by-polyp sensitivity was achieved at an FP rate of 4.1 (422/103) per patient, whereas the second variant, *MaxAUCSVMStat*, eliminated 74.5% of nonpolyps without removal of any polyps and yielded a performance of 6.5 (669/103) FPs per patient at the same sensitivity. Although the difference in AUC values between the two variants was not statistically significant, the first variant was able to achieve a higher performance in terms of an FP rate per patient at the same sensitivity. The feature selection method based on Wilks’ lambda yielded a performance of 18.0 (1854/103) FPs per patient by eliminating 29.5% of nonpolyps. The feature selection method, *MaxAUCLDA*, yielded a performance in between, i.e., 10.0 (1030/103) FPs per patient by reducing 60.7% of nonpolyps. It is evident from these results that our proposed feature selection performed much better than did the popular one based on Wilks’ lambda and *MaxAUCLDA*.

We compared the computational costs of the proposed feature selection methods on a workstation (Intel, Xeon, 2.7 GHz, 1 GB RAM). The *MaxAUCSVM* took 23 h. The *MaxAUCSVMStat,*

TABLE III
COMPARISON OF SELECTED FEATURE SUBSETS BY DIFFERENT FEATURE SELECTION METHODS FOR THE COLON DATABASE

Feature #	Features	Feature subsets selected bv different methods			
		MaxAUC SVM	MaxAUC SVMStat	MaxAUC LDA	Wilks' lambda
1	Maximum gray levels inside the lesion			X	
3	Mean gray levels inside the lesion	X			
4	Median gray levels inside the lesion	X			
5	Standard deviation of gray levels inside the lesion	X		X	X
6	Maximum gray levels on the contour of the lesion	X	X		
7	Minimum gray levels on the contour of the lesion	X	X		
8	Mean gray levels on the contour of the lesion			X	X
9	Median gray levels on the contour of the lesion	X		X	
10	Standard deviation of gray levels on the contour of the lesion	X	X		
11	Summation over perimeter values of each 2D slice	X	X		
12	Sphericity	X		X	
13	Segmented lesion volume	X		X	
14	Surface area of the candidate	X		X	
15	Ratio of the overlapping volume between the candidate and a sphere (of the same volume) to the overall volume	X	X		
16	Radial gradient index (RGI) inside the lesion			X	X
17	Radial gradient index outside the lesion	X		X	
18	Tangential gradient index inside the lesion	X			X
20	Thresholds of top 10% histogram inside the lesion			X	X
22	Thresholds of bottom 10% histogram inside the lesion	X		X	X
25	Minimum range of the histogram inside the lesion	X			
26	Maximum range of the histogram outside the lesion			X	
27	Minimum range of the histogram outside the lesion			X	X
28/30	Maximum/minimum range of the histogram of pixel values in Sobel images inside the candidate			X	
31	Minimum range of the histogram of pixel values in Sobel images outside the candidate			X	
32	Full width at half of the histogram in gray scale image	X		X	X
39	Full width at 10% maximum of the histogram in Sobel image			X	
40	Histogram overlap in the gray scale images	X			
41	Histogram overlap in the Sobel images			X	
42	Voxel intensity difference	X			
44	Absolute distance between normalized histograms	X			
45	Shannon entropy of normalized histogram			X	
47	Matsutsita distance of normalized histograms			X	X
49	Voxel intensity difference in Sobel image	X		X	
50	Voxel separation in Sobel image			X	X
52	Shannon entropy of normalized histogram in Sobel image	X	X	X	
55	Mean voxel intensity in the Sobel image	X		X	X
57	Relative standard deviation in the Sobel image	X			
59	Average Sobel power value inside the 2D contour	X	X		X
60	Mean gray levels inside the 2D lesion			X	
61	Mean gray levels outside (band region) the 2D lesion	X		X	
62	Standard deviation of gray levels inside the 2D lesion				X
64/66	Area of the 2D contour and Circularity			X	
67	Ratio of overlapping area				X
69/74	RGI and entropy texture feature inside the 2D lesion	X		X	
Feature subset size		29	7	30	14
AUC value		0.96	0.95	0.92	0.89
Non-polyp reduction rate without removal of any polyp		83.9%	74.5%	60.7%	29.3%

on the other hand, only took 5 h. The results show that *MaxAUCSVMStat* is able to save computational cost during the training stage by performing a statistical test for an early stop. The difference in training time for the two variants is substantial. However, this is compensated for by a better performance of the *MaxAUCSVM*.

V. DISCUSSIONS

The novelty of our approach is to use a nonlinear classifier to select, train, and test relevant features directly and consistently. Previous studies such as the one in [21] applied an LDA classifier for selection of features during training, but used a nonlinear neural network classifier for testing (i.e., actual classification).

TABLE IV
STATISTICAL TESTS AMONG THE PERFORMANCE (AUC VALUES) OF FOUR DIFFERENT FEATURE SELECTION METHODS IN THE DISTINCTION BETWEEN POLYPS AND NONPOLYPS

	MaxAUC SVMStat (AUC= 0.95±0.01)	MaxAUC LDA (AUC= 0.92±0.01)	Wilks' lambda (AUC= 0.89±0.02)
MaxAUC SVM (AUC= 0.96±0.02)	0.06	0.03	0.02
MaxAUC SVMStat	–	0.02	0.04
MaxAUC LDA	–	–	0.01

The AUC values with standard errors and two-sided p values are shown.

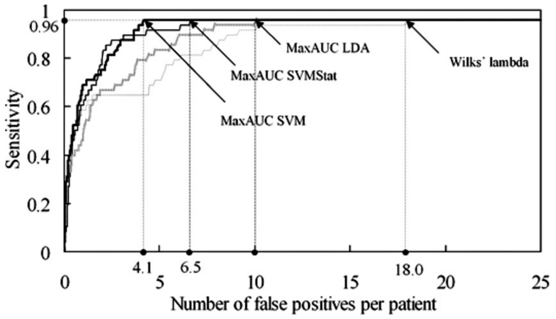


Fig. 4. FROC curves for the CADe schemes incorporating four different feature selection methods for the colon database. The performance of the initial candidate detection is shown on the far right with a 96.0% sensitivity at 25.5 FPs per patient.

This is not a principled approach towards feature selection because the features chosen by a linear classifier are not necessary optimal for a nonlinear classifier. This would be the reason why their method failed to achieve a higher performance against Wilks' lambda-based feature selection in their test [21]. Our proposed technique is based on a consistent, principled approach to feature selection and classification where both problems are handled at the same time. Other studies also used two different types of classifiers for feature selection and classifier testing (i.e., actual classification). For example, Bhooshan *et al.* [32] used the stepwise feature selection based on Wilks' lambda coupled with LDA for feature selection and Bayesian neural networks for classification. Lee *at al*. [33] applied a Gaussian kernel SVM for ranking individual features, but used a least-square SVM instead for actual classification. Their approaches were not optimal in terms of classification performance and the computational relevance between algorithms for feature selection and those for classification. Our technique presented a consistent, principled manner for feature selection and classification such that the selected features are indeed optimal for the final classifier used in the CADe scheme. Moreover, the AUC criterion we used in the selection of features reflects how a CADe scheme performs in general. It would be more suitable than the mean sensitivity of FROC used in [16] which only measured the performance of a CADe scheme in a certain specificity range. Li proposed FloatBoost to minimize classification error directly based on a backtrack mechanism [34]. The FloatBoost learning is different from our proposed feature selection method where SVM has been used for selection of an optimal feature subset by maximization of AUC value. Another novelty of our approach compared to other studies such as the ones in [21]–[25] and [33] is that the second variant of our method conducted statistical tests during the searching procedure that makes the feature selection step more reliable and efficient.

We used fivefold and leave-one-lesion out cross validation method to optimize the parameters in SVM and report performance. This procedure provided a robust and principled way to select a subset of optimal features while minimizing the risk of over-fitting given the relatively small number of true positive samples in the dataset.

VI. CONCLUSION

We have developed a maximal AUC SFFS feature selection method coupled with a nonlinear SVM classifier for CADe of polyps in CTC. The proposed method selected the most relevant features that would maximize the AUC value of the ROC curve. We presented two variants of the proposed method. Our feature selection method achieved a performance of 96% by-polyp sensitivity with 4.1 and 6.5 FPs per patient, whereas the conventional stepwise feature selection based on Wilks' lambda yielded the same sensitivity with 18.0 FPs per patient, and a maximal AUC SFFS one coupled with a LDA classifier achieved 10.0 FPs per patient at the same sensitivity level, in a leave-one-lesion-out cross-validation test in a CADe scheme for detection of polyps in CTC. One advantage of the second variant over the first one is its much lower computational cost by a factor of 4.6.

REFERENCES

[1] A. Jemal, R. Siegel, E. Ward, Y. Hao, J. Xu, and M. J. Thun, "Cancer statistics, 2010," *CA Cancer J. Clin.*, vol. 60, pp. 225–249, 2010.

[2] V. F. van Ravesteijn, C. van Wijk, F. M. Vos, R. Truyen, J. F. Peters, J. Stoker, and L. J. van Vliet, "Computer-aided detection of polyps in CT colonography using logistic regression," *IEEE Trans. Med. Imag.*, vol. 29, no. 1, pp. 120–131, Jan. 2010.

[3] H. Zhu, Z. Liang, P. J. Pickhardt, M. A. Barish, J. You, Y. Fan, H. Lu, E. J. Posniak, R. J. Richards, and H. L. Cohen, "Increasing computer-aided detection specificity by projection features for CT colonography," *Med. Phys.*, vol. 37, pp. 1468–1481, Apr. 2010.

[4] J. Yao, R. M. Summers, and A. K. Hara, "Optimizing the support vector machines (SVM) committee configuration in a colonic polyp CAD system," presented at the SPIE, Med. Imag., San Diego, CA, USA, vol. 5746, 2005.

[5] M. L. Giger, H.-P. Chan, and J. Boone, "Anniversary paper: History and status of CAD and quantitative image analysis: The role of medical physics and AAPM," *Med. Phys.*, vol. 35, pp. 5799–5820, 2008.

[6] I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh, *Feature Extraction: Foundations and Applications*. New York, NY, USA: Springer-Verlag, 2006.

[7] M. Tan, L. Wang, and I. W. Tsang, "Learning sparse SVM for feature selection on very high dimensional datasets," presented at the 27th Int. Conf. Mach. Learning, Haifa, Israel, 2010.

[8] Z. Xu, R. Jin, J. Ye, M. R. Lyu, and I. King, "Non-monotonic feature selection," presented at the Int. Conf. Mach. Learning, Montreal, Canada, 2009.

[9] B. Sahiner, N. Petrick, H.-P. Chan, L. M. Hadjiiski, C. Paramagul, M. A. Helvie, and M. N. Gurcan, "Computer-aided characterization of mammographic masses: Accuracy of mass segmentation and its effects on

characterization," *IEEE Trans. Med. Imag.*, vol. 20, no. 12, pp. 1275–1284, Dec. 2001.

[10] H. Yoshida and J. Nappi, "Three-dimensional computer-aided diagnosis scheme for detection of colonic polyps," *IEEE Trans. Med. Imag.*, vol. 20, no. 12, pp. 1261–1274, Dec. 2001.

[11] P. Campadelli, E. Casiraghi, and D. Artioli, "A fully automated method for lung nodule detection from postero-anterior chest radiographs," *IEEE Trans. Med. Imag.*, vol. 25, no. 12, pp. 1588–1603, Dec. 2006.

[12] S. Maggio, A. Palladini, L. De Marchi, M. Alessandrini, N. Speciale, and G. Masetti, "Predictive deconvolution and hybrid feature selection for computer-aided detection of prostate cancer," *IEEE Trans. Med. Imag.*, vol. 29, no. 2, pp. 455–464, Feb. 2010.

[13] P. Pudil, J. Novovicova, and J. Kittler, "Floating search methods in feature selection," *Pattern Recog. Lett.*, vol. 15, pp. 1119–1125, 1994.

[14] K. Suzuki, "Determining the receptive field of a neural filter," *J. Neural Eng.*, vol. 1, pp. 228–237, Dec. 2004.

[15] K. Suzuki, I. Horiba, and N. Sugie, "A simple neural network pruning algorithm with application to filter synthesis," *Neural Process. Lett.*, vol. 13, pp. 43–53, Feb. 2001.

[16] P.-W. Huang and C.-H. Lee, "Automatic classification for pathological prostate images based on fractal analysis," *IEEE Trans. Med. Imag.*, vol. 28, no. 7, pp. 1037–1050, Jul. 2009.

[17] A. Takemura, A. Shimizu, and K. Hamamoto, "Discrimination of breast tumors in ultrasonic images using an ensemble classifier based on the AdaBoost algorithm with feature selection," *IEEE Trans. Med. Imag.*, vol. 29, no. 3, pp. 598–609, Mar. 2010.

[18] L. Boroczky, L. Zhao, and K. P. Lee, "Feature subset selection for improving the performance of false positive reduction in lung nodule CAD," *IEEE Trans. Inf. Technol. Biomed.*, vol. 10, no. 3, pp. 504–511, Jul. 2006.

[19] S. C. Park, B. E. Chapman, and B. Zheng, "A multi-stage approach to improve performance of computer-aided detection of pulmonary embolisms depicted on CT images: Preliminary investigation," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 6, pp. 1519–1527, Jun. 2011.

[20] S. S. Mohamed and M. M. A. Salama, "Prostate cancer spectral multifeature analysis using TRUS images," *IEEE Trans. Med. Imag.*, vol. 27, no. 4, pp. 548–556, Apr. 2008.

[21] R. Hupse and N. Karssemeijer, "The effect of feature selection methods on computer-aided detection of masses in mammograms," *Phys. Med. Biol.*, vol. 55, pp. 2893–2904, 2010.

[22] A. Rakotomamonjy, "Optimizing area under ROC curve with SVMs," presented at the Proc. ROC Anal. Artif. Intell., Valencia, Spain, 2004.

[23] C. Marrocco, R. P. W. Duin, and F. Tortorella, "Maximizing the area under the ROC curve by pairwise feature combination," *Pattern Recog.*, vol. 41, pp. 1961–1974, 2008.

[24] W. Rui and T. Ke, "Feature selection for maximizing the area under the ROC curve," presented at the IEEE Int. Conf. Data Mining Workshops, Miami, FL,USA, 2009.

[25] J. Canul-Reich, L. O. Hall, D. Goldgof, and S. A. Eschrich, "Feature selection for microarray data by AUC analysis," presented at the IEEE Int. Conf. Syst., Man and Cybern., Singapore, 2008.

[26] L. Vincent, "Morphological transformations of binary images with arbitrary structuring elements," *Signal Process.*, vol. 22, pp. 3–23, 1991.

[27] V. N. Vapnik, *The Nature of Statistical Learning Theory*, 2nd ed. New York, NY, USA: Springer-Verlag, 1998.

[28] R. Kohavi and G. John, "Wrappers for feature subset selection," *Artif. Intell.*, vol. 97, pp. 273–324, 1997.

[29] J. Hanley and B. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, pp. 29–36, 1982.

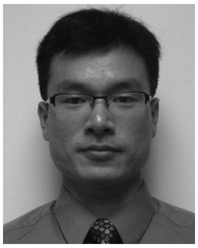
[30] C. E. Metz, B. A. Herman, and J. H. Shen, "Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data," *Statist. Med.*, vol. 17, pp. 1033–1053, May 15, 1998.

[31] J. Xu and K. Suzuki, "Computer-aided detection of hepatocellular carcinoma in hepatic CT: False positive reduction with feature selection," presented at the 8th IEEE Int. Symp. Biomed. Imaging (ISBI 2011), Chicago, IL, USA.

[32] N. Bhooshan, M. L. Giger, S. A. Jansen, H. Li, L. Lan, and G. M. Newstead, "Cancerous breast lesions on dynamic contrast-enhanced MR images: Computerized characterization for image-based prognostic markers," *Radiology*, vol. 254, pp. 680–690, Mar. 2010.

[33] S. H. Lee, J. H. Kim, N. Cho, J. S. Park, Z. Yang, Y. S. Jung, and W. K. Moon, "Multilevel analysis of spatiotemporal association features for differentiation of tumor enhancement patterns in breast DCE-MRI," *Med. Phys.*, vol. 37, pp. 3940–3956, Aug. 2010.

[34] S. Z. Li and Z. Zhenqiu, "FloatBoost learning and statistical face detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 9, pp. 1112–1123, Sept. 2004.



Jian-Wu Xu (M'07) received the B.S. degree in electrical engineering from Zhejiang University, Hangzhou, China, in 2002, and the Ph.D. degree in electrical and computer engineering from the University of Florida, Gainesville, USA, in 2007.

He was an Intern at Siemens Medical Solutions, Malvern, PA, USA, and RIKEN Brain Science Institute, Japan. He is currently a Postdoctoral Scholar with the Department of Radiology, University of Chicago, Chicago, IL, USA. His current research interests include computer-aided detection, medical image analysis, and adaptive signal processing.

Dr. Xu is a member the Tau Beta Pi and Eta Kappa Nu.

Kenji Suzuki (SM'04) received his Ph.D. degree in information engineering from Nagoya University in 2001.

From 1993 to 2001, he was with Hitachi Medical Corporation and then with Aichi Prefectural University as a Faculty Member. In 2001, he joined the Department of Radiology, University of Chicago, Chicago, IL, USA, where since 2006, he has been an Assistant Professor of Radiology, Medical Physics, and Cancer Research Center. His current research interests include computer-aided diagnosis and machine learning. He has authored or coauthored 230 papers (including 95 peer-reviewed journal papers). He has been the Editor-in-Chief and an Associate Editor of 27 leading international journals including *Medical Physics*, *International Journal of Biomedical Imaging*, and *Academic Radiology*.

Dr. Suzuki has received the Paul Hodges Award, three RSNA Certificate of Merit Awards and Research Trainee Prize, the Cancer Research Foundation Young Investigator Award, the SPIE Honorable Mention Poster Award, the IEEE Outstanding Member Award, and the Kurt Rossmann Excellence in Teaching Award.