Artificial Intelligence in Medicine 50 (2010) 43–53

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

journal homepage: www.elsevier.com/locate/aiim

Computer-aided diagnosis of pulmonary nodules using a two-step approach for feature selection and classifier ensemble construction

Michael C. Lee [a,*], Lilla Boroczky [a], Kivilcim Sungur-Stasik [b], Aaron D. Cann [b], Alain C. Borczuk [b], Steven M. Kawut [b], Charles A. Powell [b]

[a] Philips Research North America, 345 Scarborough Road, Briarcliff Manor, NY 10510-2099, USA
[b] College of Physicians and Surgeons, Columbia University, 630 West 168th Street, P&S 8, Room 503, New York, NY 10032, USA

ARTICLE INFO

Article history:
Received 19 September 2008
Received in revised form 4 April 2010
Accepted 4 April 2010

Keywords:
Genetic algorithms
Linear discriminant analysis
Feature selection
Random subspace
Computer-aided diagnosis
Pulmonary nodules

ABSTRACT

Objective: Accurate classification methods are critical in computer-aided diagnosis (CADx) and other clinical decision support systems. Previous research has reported on methods for combining genetic algorithm (GA) feature selection with ensemble classifier systems in an effort to increase classification accuracy. In this study, we describe a CADx system for pulmonary nodules using a two-step supervised learning system combining a GA with the random subspace method (RSM), with the aim of exploring algorithm design parameters and demonstrating improved classification performance over either the GA or RSM-based ensembles alone.
Methods and materials: We used a retrospective database of 125 pulmonary nodules (63 benign; 62 malignant) with CT volumes and clinical history. A total of 216 features were derived from the segmented image data and clinical history. Ensemble classifiers using RSM or GA-based feature selection were constructed and tested via leave-one-out validation with feature selection and classifier training executed within each iteration. We further tested a two-step approach using a GA ensemble to first assess the relevance of the features, and then using this information to control feature selection during a subsequent RSM step. The base classification was performed using linear discriminant analysis (LDA).
Results: The RSM classifier alone achieved a maximum leave-one-out Az of 0.866 (95% confidence interval: 0.794–0.919) at a subset size of $s = 36$ features. The GA ensemble yielded an Az of 0.851 (0.775–0.907). The proposed two-step algorithm produced a maximum Az value of 0.889 (0.823–0.936) when the GA ensemble was used to completely remove less relevant features from the second RSM step, with similar results obtained when the GA-LDA results were used to reduce but not eliminate the occurrence of certain features. After accounting for correlations in the data, the leave-one-out Az in the two-step method was significantly higher than in the RSM and the GA-LDA.
Conclusions: We have developed a CADx system for evaluation of pulmonary nodule based on a two-step feature selection and ensemble classifier algorithm. We have shown that by combining classifier ensemble algorithms in this two-step manner, it is possible to predict the malignancy for solitary pulmonary nodules with a performance exceeding that of either of the individual steps.

1. Introduction

1.1. Background

Lung cancer is the most common cause of cancer death with over 215,000 new cases and 160,000 deaths estimated for the US alone in 2008 [1]. The central challenge in the diagnostic evaluation of patients with pulmonary nodules is to identify those nodules requiring intervention, while minimizing the number of invasive procedures performed on benign lesions. Conventionally, the evaluation of nodules is guided in part by a radiological assessment of the likelihood of malignancy of the nodules based on thoracic CT scans. To provide further objective guidance on the likelihood of malignancy, researchers have explored the development of computer-aided diagnosis (CADx) systems.

Machine learning methods have become one of the dominant approaches in CADx for medical imaging [2–5]. As a topic under the field of supervised learning, classifiers are developed and trained to label new cases according to a set of features derived from the data; these features may include the raw image data, higher level descriptive features derived from the images, or additional clinical data. However, using too many features in the classification algorithm can be problematic, particularly if there are irrelevant features. This can lead to overfitting, in which noise or irrelevant

* Corresponding author. Tel.: +1 914 945 6047; fax: +1 914 945 6580.
E-mail address: michael.c.lee@philips.com (M.C. Lee).

0933-3657/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.artmed.2010.04.011

features may exert undue influence on the classification decisions because of the modest size of the training data. Additionally, there may be redundancies in the extracted features. This study is focused on applying different methods for selecting a subset of features that can be used to create an accurate CADx system.

Genetic algorithms (GAs) are a well-known method for feature selection [6]. The design of these algorithms is motivated by concepts from evolutionary biology. In the problem of feature selection, different feature subsets are encoded on a series of 'chromosomes'. A population of different chromosomes evolves through a process of mating (swapping features), reproduction, and mutation. Selection pressure is exercised such that the 'fittest' chromosomes have the greatest chance to pass on their genetic information. In feature selection, fitness is often measured by the ability of a particular feature subset to perform correct classifications (i.e. the *wrapper* approach).

Another powerful technique in machine learning is the use of classifier ensembles, alternately known as ensembles of classifiers, committees of classifiers, or multiple classifier systems [7]. In a classifier ensemble, several classifiers independently classify an unknown data point. The outputs of all these classifiers are combined to create an ensemble output. The simplest combination method is voting, though many more powerful techniques have also been proposed, and in many instances may provide even higher classification performance [7–9]. We have previously described an ensemble classifier for computer-aided diagnosis of lung nodules where feature selection for the individual classifiers is performed using a GA and combination is performed using simple voting [10].

Diversity and quality are two key factors for the accuracy of an ensemble [11]. An extreme example of diversity through modifications of the feature space is the random subspace method (RSM) [12]. In this method, every classifier receives a randomly selected subset of the features chosen from the full feature space. This method has been shown to perform well in a wide variety of circumstances, even though no effort is made to choose optimal feature subsets [13].

Efforts have been made to increase the quality aspect of the base classifiers by using better-than-random feature subsets through GAs or deterministic algorithms. In some, but not all, test scenarios, these methods outperform random feature selection [11,14]. The use of GA-based ensembles in particular has been explored at length and has been shown in many cases to be competitive with or superior to other ensemble strategies [14–18]. The failure of GA-based ensembles to consistently outperform RSM ensembles may be due to overfitting of the training data or reduced diversity in the ensemble members. A number of authors have proposed methods for evolving the entire ensemble at once or otherwise explicitly including diversity or ensemble performance (as opposed to solely individual performance) into the fitness function [15–18].

To account for these competing concerns of diversity and quality, previous studies have employed feature reduction prior to a second feature selection step. Dunne et al propose repeated execution of a feature selection algorithm (forward or backward sequential selection algorithm or random hill climbing), either resampling the data or modifying the initial parameters in each trial [19]. They then calculated the frequency at which each feature was selected during the random trials. A single, final feature subset was created by selecting those features that appeared mostly frequently in these random trials. Similarly, Pranckeviciene et al. use multiple runs of a GA to reduce the feature space in a biological classification problem where the GA inputs vary only in their random seeds [20,21]. A combination of the resulting feature spaces was then used as a reduced feature space for follow-up classification via RSM ensembles or support vector machine based feature selection. Despite the instability of the GA, the union of the selected feature spaces provided a sufficiently meaningful reduction in the feature space that subsequent classification was improved. Bertoni et al. use a domain specific feature selection method to reduce the number of genes in a biological classification problem, then follow with RSM classification [22].

1.2. Introduction to the proposed system

This study presents the development of a lung CADx system that is based upon these two-step approaches to classifier ensemble creation. The current study is strongly motivated by the previously cited approaches for combining feature reduction with ensemble learning. Our first step uses multiple runs of a GA deliberately perturbed by data resampling in order to drive the feature reduction step. In the second step, we construct the final classifier ensemble using random subspace samples drawn from a feature space that is either reduced by the GA or using sampling probabilities derived from the GA. Diversity is thus driven by the RSM, whereas previous studies have explicitly included diversity within the GA itself [15–18].

The chief contributions of this work are (1) to propose three variations and extensions of the two-step approach using different methods to combine the GA results with the RSM; (2) to compare these three proposals against each other and the constituent ensembles; and (3) demonstrate these techniques on real-world lung imaging data, with a thorough investigation of the underlying parameters, including weighting functions and parameters, feature subset size, and ensemble classification rules.

2. Materials and methods

2.1. Overview

In this section, we will describe the methods used to develop our lung CADx system based on a two-step ensemble classifier. We first describe the lung nodule dataset, image processing, and feature extraction algorithms that are used to generate the feature sets. We then describe our implementation of two algorithms for creating feature selection ensembles: the RSM and GA approaches. Three proposals for combining these algorithms are then presented. The RSM, GA, and three new combined feature selection ensemble approaches are then tested using leave-one-out validation on the lung nodule data.

2.2. Lung nodule database

A retrospective dataset of 150 pulmonary nodules (76 benign; 74 malignant) from 140 patients was collected from patients scanned by computed tomography (CT) of the chest between 2000 and 2003 at the Columbia University Medical Center (New York, NY). Due to data transfer and segmentation difficulties in very diffuse or faint nodules, 25 nodules (13 benign; 12 malignant) from 21 patients were excluded from analysis.

Each nodule in the dataset was associated with a multi-slice CT volume with in-plane resolution range of 0.5–0.8 mm and slice thickness of 5 mm, clinical information, and "ground truth" diagnosis established from either biopsy results or from an observed two-year stability in nodule size. The clinical information for each patient included age, gender, lymph node status, presence of emphysema, prior chest surgery, presence of satellite nodules, as well as the morphological characteristics of the nodule as determined by a board certified radiologist with fellowship training in thoracic imaging. The morphology was reported in terms of density (solid, semi-solid, or ground-glass opacity), margin (smooth, lobulated, spiculated, or irregular), cavitation

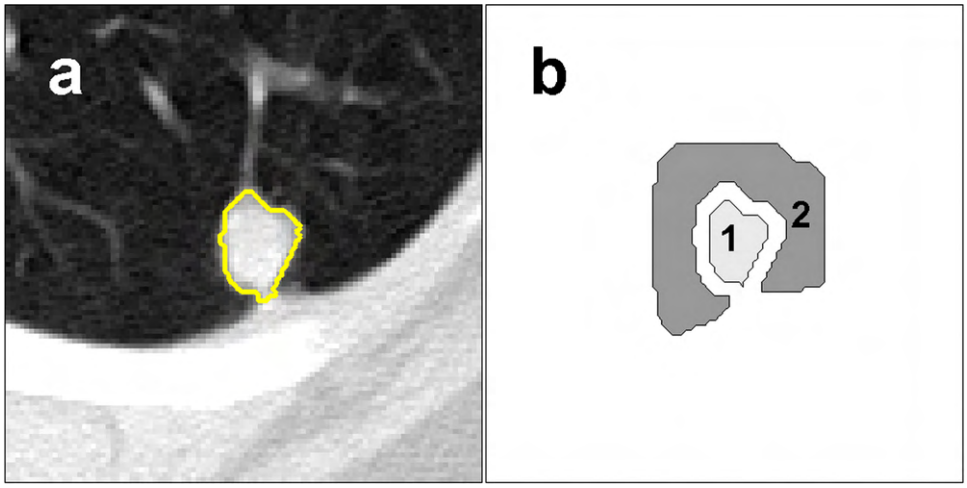

Fig. 1. Example of the sub-regions used to characterize the pulmonary nodule and its surroundings: (a) the original image with the nodule segmentation shown as a contour, (b) the two sub-regions (1 and 2) corresponding to the interior (nodule) and surroundings (parenchyma/background) of the nodule. Note that the lung wall has been excluded, so that subregion (2) does not extend fully into the lower part of the image as shown.

(present or absent), and calcification (present or absent). These clinical features were selected on the basis of medical experience and their clinical value has been reported in the research literature [2,23,24].

2.3. Image processing and analysis

For each case, 3D volumes of interest (VOIs) containing the pulmonary nodules and surrounding tissue were extracted from the CT scans. An example of this local volume of interest is shown in Fig. 1a. The location of the nodule as given by the radiologist was used as guidance in identifying the center of the VOI. In all cases, the VOI size in the axial plane was 101×101 voxels, while the number of slices in the VOI varied according to the extent of the nodule, with a minimum size of 21 slices. Linear interpolation was performed to insert interpolated slices between the native slices, resulting in 3D volumes with near-isotropic resolution. While this approach clearly does not add any new data, features calculated based on segmentations and grayscale data arising from this interpolation step have been shown to perform at as well or better than using the native slices alone [25].

A segmentation step was then used to identify three separate regions of each interpolated VOI, corresponding with the nodule, lung wall, and background (parenchyma). The segmentation algorithm has been previously reported and relies upon the techniques of distance transformations and region growing [26]. Briefly, a threshold was first used to binarize the CT volume. Region growing is then performed in which voxels are iteratively added to this core tumor region through an analysis of the relative distance of the voxel and its neighbors to the background. The stopping point for the growth is selected based on minimizing the surface integral of the distance transform values of the growth regions, leading to the exclusion of chest wall and vessels, as described previously [26,27].

The segmentation procedure was repeated on each nodule using initial Hounsfield unit (HU) thresholds varying from −800 to −300 HU, in 100 HU steps. The seed location was manually placed within the nodule, using the radiologist identified location as guidance. A 'best' threshold for each nodule was selected by visual inspection of the segmented nodules by experienced medical imaging researchers who were responsible for the development of the CADx algorithm used in this study. The segmentation results were also propagated to the original native resolution VOI by removing the interpolating slices. Four volumes were created through this process: (1) an interpolated VOI, (2) an interpolated VOI consisting only of segmentation labels, (3) a native resolution VOI, and (4) a native resolution VOI of the segmentation labels. Volumes (1) and (2) were used for computing 3D features, while volumes (3) and (4) were used for computing 2D and 2.5D features, as described in the following sections.

In addition to the nodule, lung wall, and background regions resulting from the segmentation, regions representing the interior and local surroundings of the nodule were also computed through a series of morphological operations. A margin of approximately six voxels separated these two regions. An example of this is shown in Fig. 1. This margin was used to reduce the potential impact of segmentation uncertainties on some of the features.

For each nodule, a total of 194 image-based features were computed from the processed VOIs. To improve robustness to any potential differences in slice thickness and to mimic the typical multi-slice axial analysis of human readers, the features were computed in 2D, 2.5D, and 3D. The 2D features were computed on the slice of the native resolution VOIs with the largest nodule cross-sectional area. The 2.5D features were computed as a weighted average of 2D features calculated on all the native slices, where the weights were defined by the cross-sectional area of the nodule on each slice. The 3D features were computed based upon the full interpolated VOIs. Grayscale features include statistical descriptions of the Hounsfield unit values of the interior and exterior regions of interest, as well as nodule contrast. Shape/size/margin features include area, volume, and perimeter, comparisons with circles, spheres, and ellipses (compactness, sphericity, circularity, ellipticity, and related statistics), statistical moments [28,29], Fourier descriptors [30], fractal dimension [31], and spiculation analysis [32]. Texture features include gray-level co-occurrence matrices (GLCM) [33], neighborhood gray-tone difference matrices (NGTDM) [34], and fractal dimension [35]. The GLCM matrices were characterized using entropy, energy, variance, inverse difference, and maximum probability, while NGTDM matrices were characterized using coarseness, busyness, complexity, strength, and contrast. A table of the feature categories and the number of features comprising these categories is given in Table 1; further discussion of the features is given in a related study of this CADx system [25].

Some of the clinical information described earlier is structured as categorical data: gender (male/female), lymph node status (pos/neg), presence of emphysema (yes/no), prior chest surgery (yes/no), presence of satellite nodules (yes/no), as well as the four

Table 1
Summary of feature types.

Feature type	# of features
Clinical	22
2D grayscale	7
2D shape/size/margin	22
2D texture	27
2.5D grayscale	7
2.5D shape/size/margin	22
2.5D texture	27
3D grayscale	11
3D shape/size/margin	21
3D texture	50

morphological characteristics of the nodule given in Section 2.2. The categorical clinical features were converted to binary values using 1-of-*C* encoding [36]. As an example, the density feature has three possible values: solid, semi-solid, ground-glass opacity. For this single clinical feature, three binary variables would be created: solid would be encoded as (1,0,0), semi-solid as (0,1,0), and ground-glass opacity as (0,0,1). In this way, 21 features were produced to represent the nine categorical variables; together with age, a total of 22 clinical features were thus used for each case. Thus, the final feature set contained 216 features characterizing each nodule.

2.4. Feature selection and classification

2.4.1. GA-based feature selection

The GA used in this study was the CHC (cross-generational elitist selection, heterogeneous recombination, and cataclysmic mutation) algorithm [37]. This algorithm combines a very rapid and aggressive search with highly disruptive crossover and mutation events to prevent premature convergence, and has previously been shown to be robust for feature subset selection [14,38,39]. Here, we adopt a previously described method for using GA-based feature selection for ensemble creation [14]. In this wrapper-based approach, the GA was coupled with a linear discriminant analysis (LDA) classifier, which was selected for its simplicity and promising performance over a range of ensemble and feature selection methods [13,40].

For the GA-based feature selection, potential feature subsets are encoded as 216-dimensional binary vectors in which a value of 1 indicates that the feature is used and a value of 0 indicates that the feature is not used. The GA begins with a randomly initialized population of 50 such binary feature selection vectors. Within each iteration, offspring are produced through random mating and recombination, without any selective bias in the choice of parents. Half uniform crossover (HUX) is used for recombination between the two parents, swapping exactly half of the differing bits. The only requirement for mating is that the two parents must be a Hamming distance of d apart; parents that do not meet this criteria are not returned to the pool of potential parents for that generation. The value of d is initialized to one fourth the total number of bits (216/4 = 54 in this study), and is decremented if no mating events occur in a generation or if there is no change in the population over two consecutive generations. No mutation occurs during the mating process. At each generation, a total of n offspring are generated, where $0 \leq n \leq 50$. When d is unity but no further offspring ($n = 0$) are produced during mating, the population has converged to a very low-diversity state. At that point, 'cataclysmic' mutation occurs, in which the best individual is retained intact, and all other members of the population are copies of this individual but mutated with a probability of mutation of c per bit. Previous studies have empirically found a c of 0.35 to yield high quality results, and so the cataclysmic mutation rate c was set to 0.35 in this study [37]. After a cataclysm, the value of the mating threshold d is reset to its initial value and offspring produced as before.

After the n offspring are generated, a fitness function is evaluated on each new chromosome, as defined here. For a predetermined training and testing data set that is held constant through the GA run, an LDA classifier is trained on the training data and applied to the testing data. The performance of the classifier is evaluated and the fitness is calculated as: $(A \times \text{accuracy}) + (B \times \text{Az}) + (C \times \text{sensitivity}) + (D \times \text{specificity})$, where Az represents the area under the ROC curve and A, B, C, and D were parameters selected to aid in the optimization. Accuracy, sensitivity, and specificity were always evaluated at a fixed point (a classification threshold of 0.5), while the Az was computed as usual over the whole ROC curve. Initial tests revealed $A = 1$, $B = 0.01$, $C = 0.002$, and $D = 0.001$ converged rapidly towards an optimum value, though varying these parameter values had little to no effect on the optimization. The primary objective of selecting values of A, B, C, and D was to aid in breaking ties in the chromosomes, thereby accelerating the convergence. The members of the previous generation and of all their offspring (total of $50 + n$ chromosomes) are sorted on the basis of this cost function and the 50 best chromosomes are retained. In cases of ties, the smallest feature subset is given priority. These chromosomes are then moved to the new generation, where the reproduction process begins anew. The process was stopped after 1000 generations and 1 cataclysm, or if the process reaches 100% accuracy on the test set. In practice, 100% accuracy on the relatively small test set was typically achieved after a few hundred iterations.

The GA ensemble creation process is presented schematically in Fig. 2. To create a single feature subset, the $m = 124$ training cases in each leave-one-out iteration are randomly split into a training set of b nodules and a test set of $124 - b$ nodules. Initial tests found

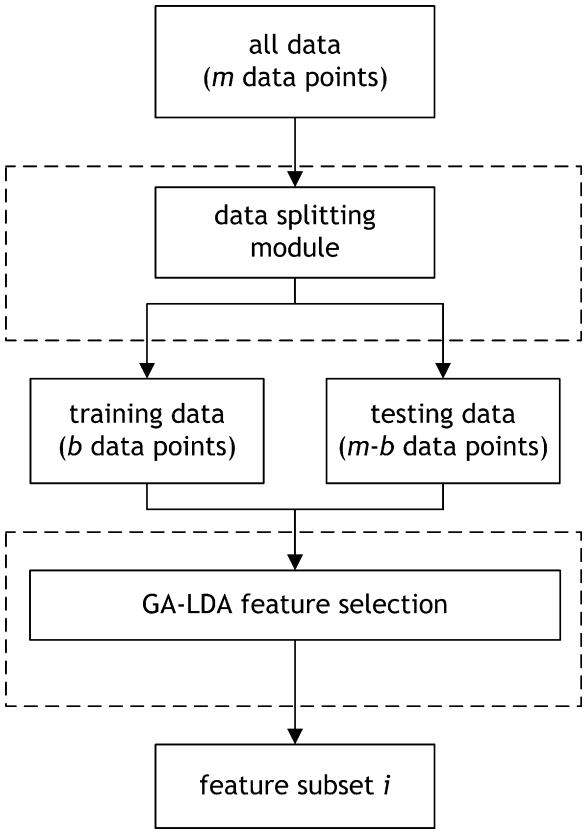

Fig. 2. Block diagram describing how a random split of the training data is used to create a single feature subset. This process is repeated k times to create the full GA-LDA classifier ensemble.

stable results when b was varied between 12 and 36, so a value of $b = 24$ was used in this study. The GA-LDA is then executed, assessing candidate feature subsets by training and testing on the split data. At the conclusion of the GA, the single feature subset with the highest fitness value is selected. This process is repeated k times to create an ensemble of k classifiers; throughout this study, k was set to 500. The likelihood estimations made by these classifiers are combined using majority voting; the estimated likelihood of malignancy is the fraction of the k classifiers reporting an estimated likelihood of malignancy greater than or equal to 0.5; exact ties with equal malignant and benign votes were assigned to the malignant group. Many other options for combination rules exist, including but not limited to weighted combinations, mixtures of experts [41], decision templates as formulated by Kuncheva et al. [9], and the Dempster–Shafer [42] combination rule. In selected instances, we supplement the voting results with the results for weighted voting (chosen because as a simple extension simple voting) and decision templates (chosen as an 'advanced' method) [43]. Briefly, weighted voting was performed using weights proportional to $\log[p/(1-p)]$, where p represents the training accuracy of each classifier [7]. In the unlikely event that perfect training performance was achieved ($p = 1$), the system would use an arbitrary value of 0.99 for p. For decision templates, decision profile matrices were constructed for each training input, with each row corresponding to a classifier and each column containing the likelihood of malignancy and benignity. The matrices were populated with the likelihood output each classifier produced for each class. A decision template for each class was produced by averaging these profiles across training samples. The likelihood of a new case belonging to a class is estimated by the squared Euclidean distance metric between that case and the decision templates. Decision templates are described in more detail in [9]. Because the differences between combination methods were small for this particular study, results are given for simple voting unless explicitly noted.

2.4.2. Random subspace method (RSM)

In an RSM ensemble, every classifier receives a randomly selected subset of the features [12]. Each of the k members of the ensemble is based on a feature subset of a predetermined mean length s. To diagnose an unknown case, the predictions made by all the classifiers are combined into a single output using a majority voting method. The final likelihood of malignancy is then the percentage of the k classifiers rendering a classification of malignant.

In addition to the ensemble size k, this method requires prior determination of a parameter s, related to the feature subset size. The size parameter s is used to randomly create the feature subsets as follows: a random 216-dimensional vector is initialized, sampling from a uniform random distribution on the unit interval. Features are active if the random value for that feature is less than $s/216$. Determination of an optimal subset size s is described in Section 3.

2.4.3. Two-step approach

The basic two-step method is illustrated schematically in Fig. 3, with the feature reduction process described in an illustrative example in Fig. 4. The approach is adapted from previously published methods for two-step ensemble classification [19,20]. In the first step, an ensemble of k feature subsets is produced through multiple runs of a GA-LDA algorithm with random data splits, as described in Section 2.4.1. The total number of occurrences of each feature through these k subsets is aggregated and sorted. The occurrence value for each feature ranges from 0 to k, indicating how many of the k independent GA runs used that particular feature in the final subset. The r features with the greatest number of occurrences are retained, while the other $216 - r$ features are discarded. The second step is the application of the RSM using the r remaining features, as described in Section 2.4.2. This two-step method therefore requires determination of two free parameters: the subset size s for the RSM and the number of retained features r to control the feature reduction. The results of a systematic search for optimal s and r parameters are described in Section 3.

Two additional variants on this two-step method were also studied in an effort to avoid using a hard threshold for the number of features retained. Both of these variants use the GA results to generate a probability density function. In the first variant, which we refer to as the "linear" approach, the features are ranked according to their frequencies of occurrence in the GA-ensemble subsets. The frequencies are then discarded, and a linear probability density function is created such that the probability for the highest ranked feature is β times the probability for the lowest ranked feature. The second step RSM then creates subsets of a defined length s by sampling without replacement from this probability density function. If $\beta = 1$, then all features are equally likely to be selected, and this reduces to the original RSM without any feature reduction. As β increases, higher ranked features appear increasingly more often. In the second variant, which we refer to as the "weighted" approach, the probability density function value for each feature is proportional to the frequency of occurrence of that feature raised to the power α. Again, the second step RSM creates subsets by sampling from this distribution. If $\alpha = 1$, then all features are equally likely to appear in the RSM ensemble, while higher values of α bias towards increased

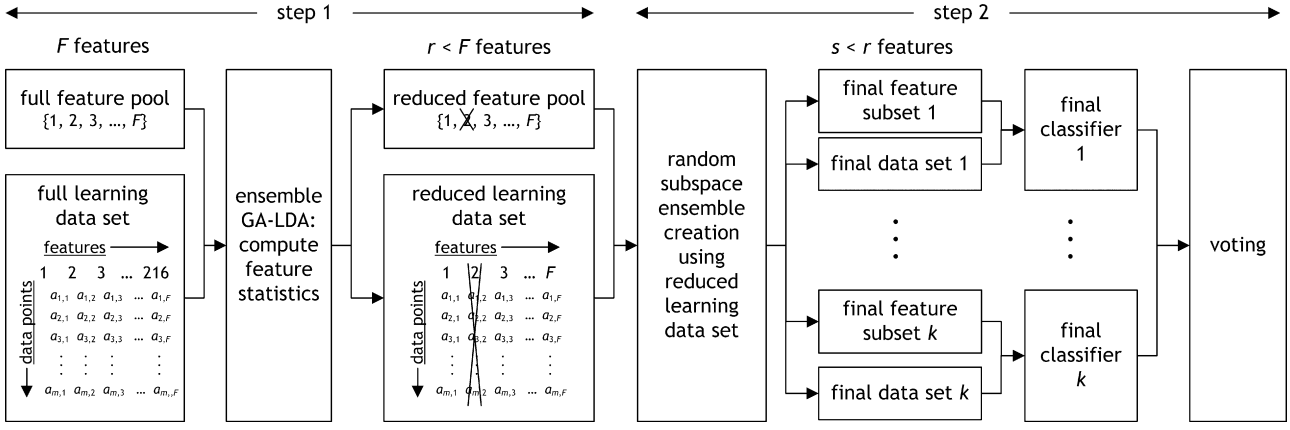

Fig. 3. Schematic of the two-step approach. In the present study, $F = 216$ (number of features), $m = 124$ (number of training cases in a leave-one-out iteration), and $k = 500$ (number of base classifiers in the ensemble). The values of s and r are varied during the study.

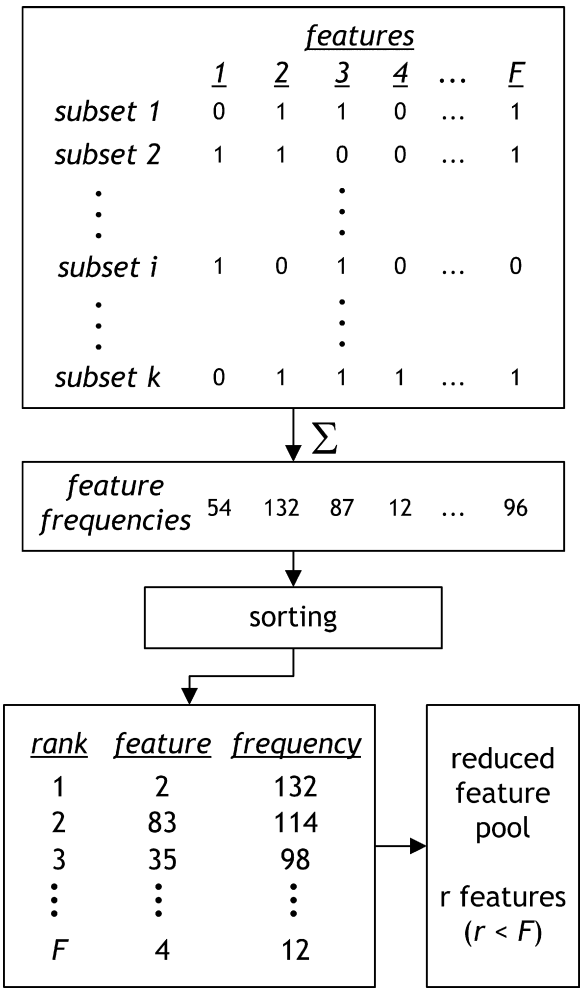

Fig. 4. Illustrative example of the feature reduction process used in the two-step algorithm. The feature subsets from the GA-LDA are represented as binary vectors in the upper panel. These are summed to determine the frequencies of occurrence for the various features. These frequencies are sorted and used to determine which features are retained in the two-step algorithm.

representation of highly ranked features. A search over the parameters space of s, α, and β is described in Section 3.

2.5. Validation

The diagnostic performance of the CADx system was tested through a leave-one-out procedure with ensembles consisting of $k = 500$ members. This large number of classifiers was chosen to reduce the influence of random selection of high or low performing subsets, particularly during the RSM. To reduce bias, both feature selection and classifier training were performed within each leave-one-out iteration [44].

For the RSM, 500 random feature subsets was created and applied to every iteration of the leave-one-out. The effect of s was determined by testing values in the set 4, 6, . . ., 100 during cross-validation on the training data. However, the same data were used for both selection and validation of the s parameter. Consequently, the results represent an optimistically biased estimate and are based on the best possible performance of the RSM on this dataset.

For the GA-based feature selection, each of the 125 leave-one-out iterations included running the GA-LDA 500 times on 124 nodules, thus creating 500 feature subsets. Each of the 500 classifiers was then trained on the 124 cases according to the selected feature subsets. The trained classifier was then applied to the one remaining case. In each of the 125 leave-one-out iterations, 500 separate GA runs were performed, resulting in a total of 62500 separate runs of the GA.

The two-step algorithms began with the aforementioned GA-LDA step. This was followed by feature reduction and random subspace ensemble creation, all within a single leave-one-out iteration. A range of values for r and s was tested to assess the algorithm under a variety of parameter choices. Both probabilistic variants were also tested by searching through s, α, and β values. Note that as a consequence of including both feature selection and classifier training in each leave-one-out iteration, a different ranking of features occurs in each iteration and a different set of features are removed through the application of the r parameter.

Statistical ROC analysis was performed using binormal fits of the observed data through a maximum likelihood approach [45]. Statistical testing was performed with a z-test to compare Az values, accounting for the correlation between the observed CADx output scores from the different ensembles. Empirical analysis of the ROC curves was also performed to estimate the Az value. The magnitude of the differences between CADx systems varied slightly from the binormal model, but there was no change in the rank-order of the methods or the conclusions of the study. Therefore, except where explicitly noted, Az results cited in the text are given for the binormal model.

3. Results

The range of Az values achieved by the RSM alone (i.e. without the GA-LDA feature reduction first step) is shown in Fig. 5. The maximum Az of 0.866 (95% confidence interval: 0.794–0.919) was realized with a subset size $s = 36$, though it is apparent that there is a plateau between a subset size of 20 and 40 where the RSM ensembles all yielded similar Az results. Using weighted voting, an Az of 0.863 was achieved, while decision templates yielded an Az of 0.835. The summary statistics for the RSM ensemble of mean subset size 36 are given in Table 2. Az values for the three ensemble combination rules used in this study are given in Table 3.

The GA-LDA method was used to create a set of 500 feature subsets for each leave-one-out iteration. As shown in Table 2, this resulted in an Az of 0.851 (0.775–0.907) for simple voting. This does not represent a statistically significant difference from the Az of the RSM ($p = 0.1104$; correlated $z = 1.60$). Other combination rules were also tested, with an Az for the GA-LDA of 0.849 and 0.827 for weighted voting and decision templates, respectively. Also shown in Table 2 is the mean Az taken across the individual classifiers the comprise each ensemble. Interestingly, the mean Az value of the individual GA-LDA ensemble members was nearly the same as the mean Az value of the individual random RSM members. This strongly suggests that because of the small test set used in the GA fitness function, the individual GA feature selection runs are creating feature subsets that fail to generalize well. However, in both the RSM and the GA-LDA, it is also evident that

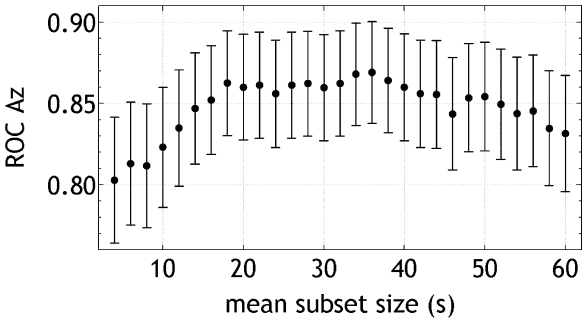

Fig. 5. Az values (±standard error) as a function of subset size parameter s, based on leave-one-out validation of RSM ensembles.

Table 2
Ensemble performance statistics.

	RSM[a]	GA-LDA	Two-step[b]	Weighted[c]	Linear[d]
Az (binormal)	0.866 ± 0.0318	0.851 ± 0.0335	0.889 ± 0.0286	0.876 ± 0.0304	0.884 ± 0.0293
Az (Wilcoxon)	0.869 ± 0.0328	0.853 ± 0.0346	0.891 ± 0.0300	0.871 ± 0.0326	0.879 ± 0.0316
Accuracy	0.79 (99/125)	0.77 (96/125)	0.84 (105/125)	0.80 (100/125)	0.81 (101/125)
Sensitivity	0.81 (50/62)	0.77 (48/62)	0.87 (54/62)	0.82 (51/62)	0.84 (52/62)
Specificity	0.78 (49/63)	0.76 (48/63)	0.81 (51/63)	0.78 (49/63)	0.78 (49/63)
Individual Az[e]	0.78	0.77	0.82	0.80	0.80

[a] Mean subset size of 36.
[b] $r = 95$, $s = 25$.
[c] $\alpha = 1.20$, $s = 30$.
[d] $\beta = 65$, $s = 25$.
[e] Mean Az of base classifiers.

Table 3
Effect of combination rules.

Combination	RSM[a]	GA-LDA	Two-step[b]	Weighted[c]	Linear[d]
Voting	0.866 ± 0.0318	0.851 ± 0.0335	0.889 ± 0.0286	0.876 ± 0.0304	0.884 ± 0.0293
Weighted vote	0.863 ± 0.0312	0.849 ± 0.0319	0.878 ± 0.0301	0.872 ± 0.0305	0.877 ± 0.0319
Decision template	0.835 ± 0.0336	0.827 ± 0.0357	0.871 ± 0.0334	0.850 ± 0.0349	0.852 ± 0.0347

[a] Mean subset size of 36.
[b] $r = 95$, $s = 25$.
[c] $\alpha = 1.20$, $s = 30$.
[d] $\beta = 65$, $s = 25$.

the ensemble approach can overcome the low quality of the individual LDA classifiers that was induced by randomness in the RSM or overfitting the GA-LDA to a small derivation data set.

The features subsets derived in the GA-LDA were used as the input into the proposed two-step algorithm. While each individual run of the GA-LDA only optimizes the feature set for a particular random data split, aggregated analysis of the runs over all the random splits indicates that there is clear pattern of dominant features. For illustrative purposes, the feature-by-feature frequencies of occurrence have been summed over all 125 leave-one-out iterations and plotted in Fig. 6. The peaks in the figure suggest that some features are dominant over others. The top 10 features were represented nearly three times as often as the 10 features that occurred least often. These top 10 features are listed in Table 4. The two features that occurred least often were the two gender features. This provides some assurance as to the relevance of the feature selection results, since gender has not been reported to have any significant influence on the likelihood of malignancy of lung nodules. However, all features are represented to some degree, suggesting that it is difficult to identify a single best feature subset in this data. Indeed, the most commonly occurring features appeared in less than 50% of the feature subsets.

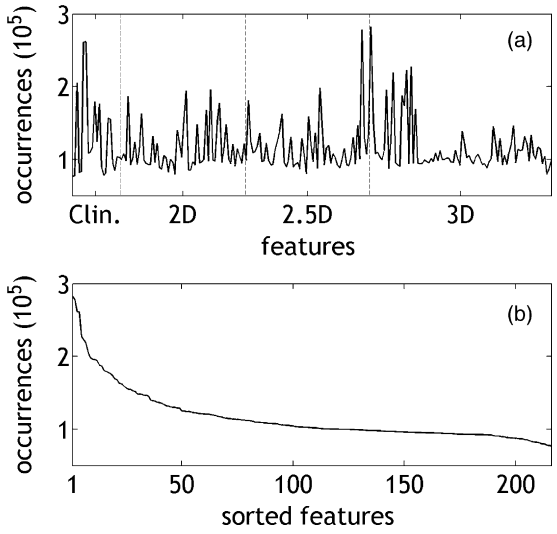

Fig. 6. (a) Feature frequency in the subsets GA-LDA subsets, summed over all leave-one-out iterations. (b) Sorted features, showing the optimal cutoff for retaining features in the two-step algorithm. Note that the summation over all leave-one-out iterations is done for illustrative purposes; in experiments, this summation was not performed.

This pool of features was then modified using the two-step approach, with r and s as parameters controlling respectively the size of the feature pool and the number of random features used. This is equivalent to choosing a y-axis cutoff position in Fig. 6(a) or an x-axis cutoff position in Fig. 6(b). The performance of the two-step ensembles in this parameter space is displayed in Fig. 7. The right-most vertical column of the figure represents random feature subsets drawn from the full 216-dimensional feature pool. Therefore, the contours at the right-most edge of the figure represent the same information as Fig. 5. Moving towards the left, the values of r are reduced, thus representing a reduction in the feature pool, with the features that occurred least often in the GA ensemble being removed first. The initial change is an increase in the width of the plateau of high performing subset sizes, followed by an increase in the maximum Az and accuracy. A peak eventually

Table 4
List of the 10 features that occurred most frequently in the GA-LDA ensemble, and the fraction of the GA selected subsets in which these features occurred.

Feature description	Type	% of subsets
Maximum HU	3D	45.2
NGTDM coarseness (outside)	2.5D	44.5
Prior chest surgery = no	Clinical	41.9
Prior chest surgery = yes	Clinical	41.8
HU standard deviation	3D	36.4
Contrast	3D	35.7
Elongated shape (moment-based)	3D	35.0
Age	Clinical	32.8
GLCM texture variance (in)	2.5D	31.7
GLCM texture variance (out)	2D	31.4

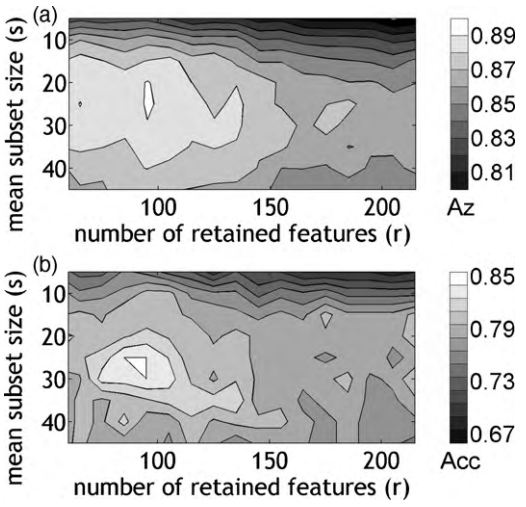

Fig. 7. Contour plot of (a) Az values and (b) accuracy (Acc) as a function of the r and s parameters for the two-step algorithm.

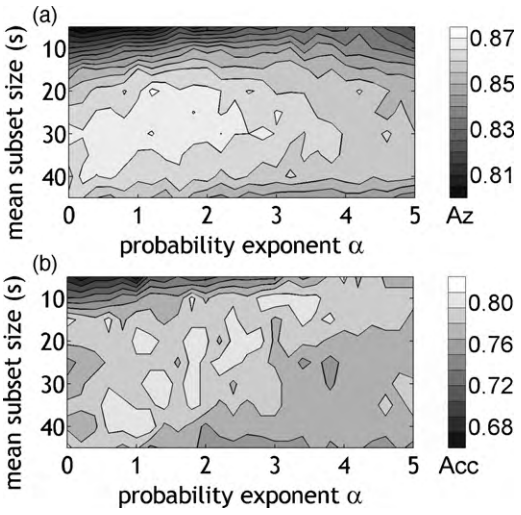

Fig. 9. Contour plot of (a) Az values and (b) accuracy as a function of the α and s parameters for the weighted probabilistic variant of the two-step algorithm.

appears, with a maximum Az reached at $s = 25$ and $r = 95$, yielding an Az of 0.889 (0.823–0.936). Using weighted voting or decision templates did not lead to an improvement of this Az (0.878, 0.871). While this particular choice of s and r represents a single high performing peak, it is clear from Fig. 7 that there is a wide range of parameter choices which provide similarly high Az and accuracy results. Note that this analysis is done in a leave-one-out fashion: because the training data changes at each leave-one-out iteration, the ranking of the features also changes at each iteration, as does the identity of the features that are removed and retained.

A statistically significant difference was found for the Az values between the RSM and two-step ensembles ($p = 0.033$) and between the GA-LDA and two-step ensembles ($p = 0.0018$). Parametric estimates of the ROC curves were obtained using binormal fitting and are displayed in Fig. 8. Using the natural cutoff point of 0.5, the two-step ensemble correctly classified 84% of the nodules (105/125), compared with 0.79 (99/125) and 0.77 (96/125) for the RSM and GA-LDA, respectively. Improvements and sensitivity and specificity were also observed, as reported in Table 2.

A grid search over parameter values in the probabilistic approaches to the two-step method resulted in a maximum Az of 0.876 (0.806–0.926) at a subset size $s = 30$ and $\alpha = 1.20$ for the weighted algorithm, and an Az of 0.887 (0.820–0.934) at a subset size $s = 25$ and $\beta = 65.0$ for the linear algorithm. No improvement in

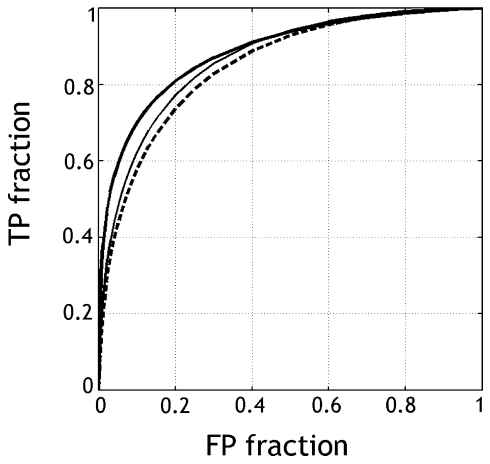

Fig. 8. ROC curves for the GA-LDA (dashed line), RSM (thin line), and two-step ensembles (thick line).

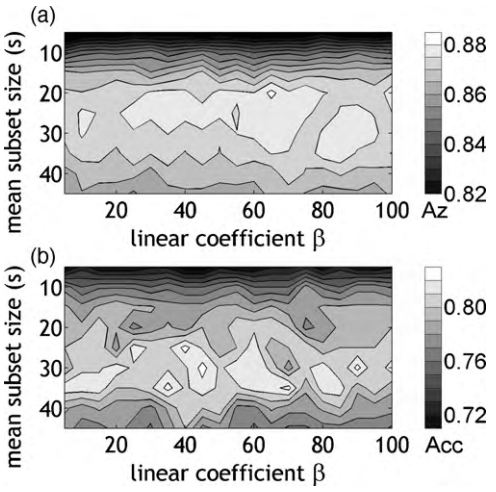

Fig. 10. Contour plot of (a) Az values and (b) accuracy as a function of the β and s parameters for the linear probabilistic variant of the two-step algorithm.

Az was noted when weighted voting or decision templates for either the weighted two-step (0.872, 0.850) or the linear two-step (0.877, 0.852). The linear algorithm significantly outperformed the RSM ($p = 0.012$) and GA-LDA ($p = 0.0048$), while the superiority of the weighted algorithm was less clear with a $p = 0.29$ when compared with the RSM, and $p = 0.045$ compared against the GA-LDA. However, as with the original two-step method, visualizing the parameter search results as in Figs. 9 and 10, it is clear that peaks in the Az and accuracy are achievable by changing the feature reduction parameters.

4. Discussion

The selection of high performing feature subsets is an important but difficult task in developing CADx systems. The use of ensemble classifiers that explicitly enable classification using multiple different subsets helps to mitigate this problem. This is clearly illustrated in the RSM that uses only random feature selection. Similarly, despite the risk of overfitting from a GA, the ensemble created by the GA-LDA can significantly outperform the individual LDA classifiers that constituted the ensemble. Indeed, in all cases studied here it was observed that the combined ensemble outperformed the mean individuals comprising the ensemble.

Using more advanced ensemble combination rules did not improve the classification in this study. However, the choice of an optimal combination rule is likely to be affected by the diversity and quality of the classifiers created by the feature selection strategy, and is a topic that requires further investigation. Nevertheless, despite the success of ensemble systems, it is likely that CADx system performance can be improved by controlling the feature pool through knowledge of the relative importance of features. This may include prior knowledge of the problem structure, i.e. domain knowledge, or it may be purely statistical in nature [19–22].

The first major contributions of the current work are to propose specific extensions to previously described methods for using combining GA and RSM-based classifier ensembles, and to compare these proposals against each other and against the GA and RSM ensembles individually. Such approaches may avoid the difficulty of crafting a GA scheme that can identify a single high quality feature subset while avoiding overfitting, and further take advantage of the simplicity of the RSM. In well-structured datasets with little redundancy in features, it is possible that a single well-designed GA may indeed identify ideal features for classification [38]. However, in many data sets, there may be large numbers of potentially redundant, correlated, or irrelevant features, and a limited data sample. Indeed, in the current study, tests on a real-world medical imaging data set demonstrate instability in the GA, whereby shifting a few data points from the GA fitness function training set to the testing set leads to a different final feature subset. This can be exploited to create classifier ensembles [14], but can it also be used to identify dominant and weak features (as distinct from identifying a single set of features that work well with each other). In this study, this feature ranking was used to constrain or bias the feature space in favor of those features that had the greatest contribution to the GA ensemble, and were thus judged to be likely to also contribute positively to the RSM ensembles.

Numerous alternative approaches exist to this problem of high-dimensionality and limited data. Projection of the data into new spaces through, for example, principal component analysis, may reduce the problem of correlated features, but there is a risk that some discriminatory power will also be lost during this transformation. Even if a single run of the GA is able to identify a high-performance feature subset, there is the risk that many equivalently high-performance subsets also exist undiscovered. Moreover, while such a feature subset might be appropriate for use in a classifier, it may not be appropriate to apply a second feature selection ensemble to this feature subset. There is no guarantee that dividing up this single high quality subset into smaller feature subsets through a second round of feature selection will lead to any improvement. We therefore propose that for many real-world data sets, it is advantageous to use multiple runs of a GA or other feature selection algorithm to build a statistical model for feature reduction.

In this study, a GA was used in the initial round of feature reduction. Alternative approaches to the feature reduction problem may include the use of other wrapper-based feature selection methods that have some level of instability; sequential selection methods were used by Dunne et al. for a similar purpose [19]. Pranckeviciene et al have previously demonstrated the use of multiple GA runs with varying random seeds, with each GA achieving a high confidence through the use of internal bootstrapping [21]. The union of these high confidence regions was shown to provide a meaningful reduction in the feature space that enhanced the performance subsequent feature selection steps. In contrast, we studied an approach where each GA is dedicated to optimizing a relatively simple objective function based on a small testing sample. The resulting feature sets are known to overfit the small testing sample, but the statistical combination of these feature sets for feature reduction was shown to overcome this limitation and improve the following ensemble step. The current approach and the previous studies it builds upon all represent variations on the theme of feature reduction through exploitation of algorithm variability.

While the RSM was used here as the second ensemble creation step because diversity was an explicit goal, it is possible that a GA-based ensemble may be equally useful in the second step. The principle of consecutive GAs has been previously reported as a means of handling large and complex feature spaces [46]. Significant prior research in GA ensembles has further described how diversity can be explicitly included in the GA search strategy [15–18]. Such techniques may be particularly valuable in the two-step framework if the feature space is first reduced using the statistical results of the GA. Instead of using an RSM to force diversity as proposed here, a diversity-driven GA-search strategy can then be employed as a second step. However, the outcome of using different feature selection ensembles for the first and second step has not been studied in this research. Identifying successful combinations of feature selection ensembles may prove to be an interesting future direction for this research.

It is likely that the advantages of this two-step approach are dependent on the specific dataset under study. We emphasize that the current analysis is for a CADx system operating on medical images acquired from patients. Imaging features extracted from this dataset are thus subject to the noise and uncertainty from the image acquisition and image segmentation process. Additionally, as multiple image features may describe similar concepts, there is likely some redundancy in the extracted features. It was clearly observed that there was a large variation in the frequency of occurrence for features in the GA runs. This strongly suggested a wide range in the relevance of the features to this particular classification problem. In situations in which there is less of a variation in feature relevance, the first step of GA-based feature reduction is less likely to produce a meaningful improvement in the feature space, as the differences in the frequency of feature occurrence will be much smaller.

These results indicate that for some real-world data sets, it may be difficult to isolate any single stable feature subset. Since feature selection is generally only performed prior to deployment, in CADx systems that aim to utilize only a single fixed feature subset, this instability is problematic. Indeed, the use of multiple feature subsets in an ensemble configuration can be considered a hedge against the uncertainty in using any one subset. Nevertheless, even in an ensemble classifier system, it is necessary to fix the collection of feature subsets prior to deployment of any CADx. This paper describes one potential method for mitigating the difficulty in choosing that final collection of feature subsets. In practice, however, any classifier system intended for use outside the laboratory must be thoroughly validated to provide statistical evidence of its robustness and generalizability; the difficulty in feature selection only further highlights this point.

The proposed method relies upon two parameters, s and r, to control how aggressively the feature space is reduced and the composition of the feature subsets used in the final ensemble. In the current study, the selection of specific values of s and r bias towards favorable cross-validation performance because the parameter values were chosen in a *post hoc* manner, with the goal of maximizing the Az statistic. However, the performance of the system was also tested using different parameter choices, as seen in Figs. 7, 9 and 10. Although the magnitude of the improvement achieved by the two-step approach varied, a performance improvement was evident through a wide range of parameter values. This can be visualized in the wide plateaus in which the two-step approach improved on both the RSM and GA-LDA. In particular, even arbitrarily chosen s and r parameters can

improve upon an RSM where the parameters are deliberately chosen as a best-case scenario. Therefore, while the magnitude of the effect is significantly biased by the *post hoc* parameter selection, we believe that the underlying trends are clear. While the results of this study are adequate for proving the concept of the two-step approach and exploring the parameter space and algorithmic design considerations, continuing work in this area will address specific performance comparisons through blind tests on independent data sets.

In this work, we have sought to reduce the impact of a hard threshold for r by using a "soft" interpretation of the GA-frequency distribution. Based on the observed frequency distributions, it is very likely that the features on either side of a hard r cutoff may have occurred at a similar frequency within the GA ensemble. Discarding the lower ranked feature in such an instance may not be optimal. Alternative approaches have therefore included probabilistic interpretations for the feature occurrence statistics. Features with low frequencies of occurrence in the GA ensemble are not completely rejected, but instead are used with low frequency in the second round of ensemble creation. In this study, these approaches also produced results competitive with the original two-step method. It should be noted, however, that such approaches do not fully address the problem of parameter selection, as control parameters α and β still exist for this probabilistic mapping. However, the influence of these parameters is reduced through the use of soft thresholds. Further research may include the use of different probability distributions, including, for example, functions that incorporate the notion of a fuzzy threshold.

The selection of the two-step parameter values also raises other practical considerations. When the computational complexity of searching a wide range of parameter values is added to the very large number of GA trials already executed, the overall computational burden of the approach may be significant. This computational burden is comparable to previously described approaches for GA-based ensembles [14,20,21]. However, the ensemble creation process as a whole and the evaluation of the fitness functions within a GA population are both highly parallelizable operations. Moreover, the current study is focused on the feature reduction and selection strategy and not on implementation optimization. The very large number of ensemble members and GA trials was selected in order to minimize the impact of statistical fluctuations in the frequency analysis used in feature reduction and remove that confounding factor. Since each ensemble member is created independently, the GA ensemble creation time is linear with ensemble size, considerable cost savings can be effected by reducing the ensemble size. The selection of an appropriate ensemble size may be a topic of interest for future studies on this approach.

It is also useful to reiterate that the ensemble creation process is only used during development of the computer-aided diagnosis system, and the actual computation cost of evaluating an unknown case is minimal. Moreover, to achieve full confidence in the control parameter values, it would be most helpful to have a second dataset for parameter selection. This is not unlike similar problems encountered with other machine learning techniques in which initial parameters must be selected. Alternative approaches may exist, in which, for example, r is selected on the basis of various heuristics, such as retaining a fixed percentile of features based on frequency of occurrence or excluding features whose frequency of occurrence are a number of standard deviations from the mean. Similarly, selection of the subset size s may be performed via various methods to estimate the latent dimensionality of the data, e.g. through principal components analysis. These alternative approaches will be explored in future studies as means of addressing the uncertainties around the parameter selection process.

5. Conclusion

We have developed and applied a two-step feature selection and ensemble classifier algorithm to the computer-aided diagnosis of pulmonary nodules. The current study is strongly motivated by previously reported results on approaches that combined use of feature reduction followed by ensemble learning. We have extended these prior works with a proposal for multiple different uses of the GA. A comparison was given of these different composite GA-RSM approaches against each other and the constituent GA and RSM ensembles. Finally, we have demonstrated these techniques on real-world lung imaging data, with a thorough investigation of the underlying parameters, including weighting functions and parameters, feature subset size, ensemble classification rules. We have shown that a classifier ensemble built in this fashion can achieve accurate and reliable prediction of malignancy for solitary pulmonary nodules superior to the two ensembles that comprise its constituent steps. We believe that continued technical work on improving the robustness and classification accuracy of machine learning systems is critical to the future creation and application of computer-aided diagnosis and clinical decision support systems.

Acknowledgments

The authors thank Ye Xu, Lalitha Agnihotri, Luyin Zhao, J. David Schaffer, and Eric Silfen for their helpful discussions on the algorithms and methods.

References

[1] Jemal A, Siegel R, Ward E, Hao YP, Xu JQ, Murray T, et al. Cancer statistics. Ca-a Cancer Journal for Clinicians 2008;58(March–April):71–96.
[2] Gurney JW. Determining the likelihood of malignancy in solitary pulmonary nodules with Bayesian analysis. Part I. Theory. Radiology 1993;186(February):405–13.
[3] Nakamura K, Yoshida H, Engelmann R, MacMahon H, Katsuragawa S, Ishida T, et al. Computerized analysis of the likelihood of malignancy in solitary pulmonary nodules with use of artificial neural networks. Radiology 2000;214(March):823–30.
[4] Way TW, Hadjiiski LM, Sahiner B, Chan HP, Cascade PN, Kazerooni EA, et al. Computer-aided diagnosis of pulmonary nodules on CT scans: Segmentation and classification using 3D active contours. Medical Physics 2006;33(July):2323–37.
[5] Suzuki K, Li F, Sone S, Doi K. Computer-aided diagnostic scheme for distinction between benign and malignant nodules in thoracic low-dose CT by use of massive training artificial neural network. IEEE Transactions on Medical Imaging 2005;24(September):1138–50.
[6] Siedlecki W, Sklansky J. A note on genetic algorithms for large-scale feature-selection. Pattern Recognition Letters 1989;10(November):335–47.
[7] Kuncheva LI. Combining pattern classifiers: methods and algorithms. Hoboken, NJ: John Wiley & Sons, Inc.; 2004.
[8] Kittler J, Hatef M, Duin RPW, Matas J. On combining classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence 1998;20(March):226–39.
[9] Kuncheva LI, Bezdek JC, Duin RPW. Decision templates for multiple classifier fusion: an experimental comparison. Pattern Recognition 2001;34(February):299–314.
[10] Boroczky L, Lee MC, Zhao L, Agnihotri L, Powell CA, Borczuk AC, et al. Computer-aided diagnosis for lung cancer using a classifier ensemble. International Journal of Computer Assisted Radiology and Surgery 2007;2(June):S362–4.
[11] Cunningham P, Carney J. Diversity versus quality in classification ensembles based on feature selection. In: Lopez de Mantaras R, Plaza E, editors. Lecture Notes in Artificial Intelligence 1810: Machine Learning: ECML 2000 11th European Conference on Machine Learning (May 31–June 2, 2000, Barcelona, Spain). Heidelberg, Germany: Springer; 2000. p. 109–16.
[12] Ho TK. The random subspace method for constructing decision forests. IEEE Transactions on Pattern Analysis and Machine Intelligence 1998;20(August):832–44.
[13] Skurichina M, Duin RPW. Bagging, boosting and the random subspace method for linear classifiers. Pattern Analysis and Applications 2002;5(June):121–35.
[14] Guerra-Salcedo C, Whitley D. Feature selection mechanisms for ensemble creation: a genetic search perspective. In: Freitas AA, editor. AAAI-99 and GECCO-99 Workshop on Data Mining with Evolutionary Algorithms: Research Directions. Menlo Park, CA: AAAI; 1999. p. 13–7.
[15] Tsymbal A, Pechenizkiy M, Cunningham P. Diversity in search strategies for ensemble feature selection. Information Fusion 2005;6:83–98.

[16] Tsymbal A, Pechenizkiy M, Cunningham P. Sequential genetic search for ensemble feature selection. In: Kaelbling LP, Saffiotti A, editors. Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence (July 30–August 5, 2005, Edinburgh, Scotland). Denver, CO: Professional Book Center; 2005. p. 877–82.
[17] Tsymbal A, Cunningham P, Pechenizkiy M, Puuronen S. Search strategies for ensemble feature selection in medical diagnostics. In: Krol M, Mitra S, Lee DJ, editors. Proceedings of the 16th IEEE Symposium on Computer-Based Medical Systems (June 26–27, 2003, New York, NY). Menlo Park, CA: IEEE; 2003. p. 124–9. doi: 10.1109/CBMS.2003.1212777.
[18] Opitz D. Feature selection for ensembles. In: Hendler J, Subramanian D, editors. Proceedings of the 16th National Conference on Artificial Intelligence (July 18–22, 1999, Orlando, Florida). Menlo Park, CA: AAAI Press; 1999 p. 379–84.
[19] Dunne K, Cunningham P, Azuaje F. Solutions to instability problems with sequential wrapper-based approaches to feature selection (technical note). Department of Computer Science, Trinity College, University of Dublin; 2002. Jan. Report No.: TCD-CS-2002-28.
[20] Pranckeviciene E, Baumgartner R, Somorjai R. Using domain knowledge in the random subspace method: application to the classification of biomedical spectra. Multiple Classifier Systems 2005;3541:336–45.
[21] Pranckeviciene E, Somorjai R, Baumgartner R, Jeon MG. Identification of signatures in biomedical spectra using domain knowledge. Artificial Intelligence in Medicine Nov 2005;35:215–26.
[22] Bertoni A, Folgieri R, Valentini G. Feature selection combined with random subspace ensemble for gene expression based diagnosis of malignancy. In: Apolloni B, Marinaro M, Tagliaferri R, editors. Biological and Artificial Intelligence Environments: Proceedings of the 15th Italian Workshop on Neural Nets (September 14–17, 2004, Perugia, Italy). Dordrecht, Netherlands: Springer; 2004. p. 29–35.
[23] Erasmus JJ, Connolly JE, McAdams HP, Roggli VL. Solitary pulmonary nodules. Part I. Morphologic evaluation for differentiation of benign and malignant lesions. Radiographics 2000;20(January–February):43–58.
[24] Li F, Sone S, Abe H, Macmahon H, Doi K. Malignant versus benign nodules at CT screening for lung cancer: comparison of thin-section CT findings. Radiology 2004;233(December):793–8.
[25] Xu Y, Lee MC, Boroczky L, Cann A, Borczuk AC, Kawut S, et al. Comparison of image features calculated in different dimensions for computer-aided diagnosis of lung nodules. In: Karssemeijer N, Giger ML, editors. SPIE Medical Imaging 2009: Computer-Aided Diagnosis (February 8–12, 2009, Lake Buena Vista, Florida). Bellingham, WA: SPIE; 2009. doi: 10.1117/12.807866.
[26] Wiemker R, Rogalla P, Blaffert T, Sifri D, Hay O, Shah E, et al. Aspects of computer-aided detection (CAD) and volumetry of pulmonary nodules using multislice CT. British Journal of Radiology 2005;78:S46–56.
[27] Wiemker R, Rogalla P, Hein P, Blaffert T, Rosch P. Computer-aided segmentation of pulmonary nodules: automated vasculature cutoff in thin- and thick-slice CT. In: Lemke HU, Vannier MW, Inamura K, Doi K, Reiber JHC, editors. International Congress Series. CARS 2003. Computer Assisted Radiology and Surgery. Proceedings of the 17th International Congress and Exhibition (June 25–28, 2003, London UK), 1256; 2003. p. 965–70. doi:10.1016/S0531-5131(03)00283-8.
[28] Hu M. Visual pattern recognition by moment invariants. IEEE Transactions on Information Theory 1962;8:179–87.
[29] Galvez JM, Canton M. Normalization and shape-recognition of 3-dimensional objects by 3D moments. Pattern Recognition 1993;26(May):667–81.
[30] Granlund GH. Fourier preprocessing for hand print character recognition. IEEE Transactions on Computers 1972;C 21:195–201.
[31] Kido S, Kuriyama K, Higashiyama M, Kasugai T, Kuroda C. Fractal analysis of small peripheral pulmonary nodules in thin-section CT—evaluation of the lung-nodule interfaces. Journal of Computer Assisted Tomography 2002;26(July–August): 573–8.
[32] Sahiner B, Chan HP, Petrick N, Helvie MA, Goodsitt MM. Computerized characterization of masses on mammograms: the rubber band straightening transform and texture analysis. Medical Physics 1998;25(April):516–26.
[33] Haralick RM, Shanmuga K, Dinstein I. Textural features for image classification. IEEE Transactions on Systems Man and Cybernetics 1973;SMC-3:610–21.
[34] Amadasun M, King R. Textural features corresponding to textural properties. IEEE Transactions on Systems Man and Cybernetics 1989;19(September–October):1264–74.
[35] Chen CC, Daponte JS, Fox MD. Fractal feature analysis and classification in medical imaging. IEEE Transactions on Medical Imaging 1989;8(June):133–42.
[36] Bishop CM. Neural networks for pattern recognition. New York: Oxford University Press; 1995.
[37] Eshelman L. The CHC adaptive search algorithm: how to have a safe search when engaging in nontraditional genetic recombination. In: Spitz, Bruce M, editor. Foundations of Genetic Algorithms (July 15–18, 1990, Indiana University, Bloomington, Indiana). San Mateo, CA: Morgan Kaufmann.
[38] Guerra-Salcedo C, Whitley DL. Genetic search for feature subset selection: a comparison between CHC and GENESIS. In: Koza, John R, Banzhaf, Wolfgang, Chellapilla, Kumar, Deb, Kalyanmoym Dorigo, Marco, Fogel, David B, Garzon, Max H, Goldberg, David E, Iba, Hitoshi, Riolo, Rick L, editors. Genetic Programming 1998: Proceedings of the Third Annual Conference (July 22–25, 1998, University of Wisconsin, Madison, Wisconsin). San Francisco, CA: Morgan Kaufmann.
[39] Lee MC, Nelson SJ. Supervised pattern recognition for the prediction of contrast-enhancement appearance in brain tumors from multivariate magnetic resonance imaging and spectroscopy. Artificial Intelligence in Medicine 2008;43(May):61–74.
[40] Duda RO, Hart PE, Stork DG. Pattern classification. New York: Wiley; 2001.
[41] Jacobs RA, Jordan MI, Nowlan SJ, Hinton GE. Adaptive mixtures of local experts. Neural Computation 1991;3:79–87.
[42] Rogova G. Combining the results of several neural network classifiers. Neural Networks 1994;7:777–81.
[43] Polikar R. Ensemble based systems in decision making. IEEE Circuits and Systems Magazine 2006;6:21–45.
[44] Li Q, Doi K. Reduction of bias and variance for evaluation of computer-aided diagnostic schemes. Medical Physics 2006;33(April):868–75.
[45] Metz CE, Herman BA, Shen JH. Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data. Statistics in Medicine 1998;17(May):1033–53.
[46] Mougiakakou SG, Valavanis IK, Nikita A, Nikita KS. Differential diagnosis of CT focal liver lesions using texture features, feature selection and ensemble driven classifiers. Artificial Intelligence in Medicine 2007;41(September):25–37.