Application of the mutual information criterion for feature selection in computer-aided diagnosis

Georgia D. Tourassi[a)]
Department of Radiology, Duke University Medical Center, Durham, North Carolina 27710

Erik D. Frederick
ChemCodes, Inc., Durham, North Carolina 27713

Mia K. Markey
Department of Biomedical Engineering, Duke University, Durham, North Carolina 27710

Carey E. Floyd, Jr.
Department of Radiology, Duke University Medical Center, Durham, North Carolina 27710 and Department of Biomedical Engineering, Duke University, Durham, North Carolina 27710

(Received 13 June 2001; accepted for publication 27 September 2001)

The purpose of this study was to investigate an information theoretic approach to feature selection for computer-aided diagnosis (CAD). The approach is based on the mutual information (MI) concept. MI measures the general dependence of random variables without making any assumptions about the nature of their underlying relationships. Consequently, MI can potentially offer some advantages over feature selection techniques that focus only on the linear relationships of variables. This study was based on a database of statistical texture features extracted from perfusion lung scans. The ultimate goal was to select the optimal subset of features for the computer-aided diagnosis of acute pulmonary embolism (PE). Initially, the study addressed issues regarding the approximation of MI in a limited dataset as it is often the case in CAD applications. The MI selected features were compared to those features selected using stepwise linear discriminant analysis and genetic algorithms for the same PE database. Linear and nonlinear decision models were implemented to merge the selected features into a final diagnosis. Results showed that the MI is an effective feature selection criterion for nonlinear CAD models overcoming some of the well-known limitations and computational complexities of other popular feature selection techniques in the field. © *2001 American Association of Physicists in Medicine.* [DOI: 10.1118/1.1418724]

Key words: mutual information, feature selection, computer-assisted diagnosis, acute pulmonary embolism

I. INTRODUCTION

Feature selection is the process of choosing a subset of features relevant to a particular application. During the selection process, a decision criterion is used to remove irrelevant or redundant features. Extensive research has led researchers to appreciate the importance of feature selection when developing computer-assisted diagnostic (CAD) tools. Optimized feature selection reduces data dimensionality and potentially removes noise, thus resulting in CAD tools that are not only more accurate but also more robust. Several CAD applications have demonstrated the positive impact of optimized feature selection.[1–8]

The most popular feature selection algorithm utilized in CAD is the stepwise linear discriminant analysis,[9] borrowed from linear statistics. It is designed to reduce the dimensionality of the feature vector by selecting in stepwise fashion the features that maximize the linear separability of the output classes. Intuitively though, a feature selection technique based on linear assumptions is not necessarily the optimal data preprocessing approach, particularly when the reduced feature set will be fed into a nonlinear decision algorithm such as the backpropagation neural network (BP–ANN). Consequently, weight pruning techniques[10] and genetic algorithms[11] have been proposed for feature selection with BP–ANNs as promising alternatives better tailored to the nonlinear nature of ANNs. Weight pruning techniques are based on the assumption that features connected with weights that are close to zero can be eliminated. With large data sets, the weight pruning approach can be computationally expensive because the network needs to be fully trained before any pruning occurs. A more popular and effective choice in CAD is the genetic algorithm (GA). GAs are a form of artificial intelligence inspired by the biological process of evolution. The basic approach is to create a population of randomly selected combinations of features. Each combination is considered a possible solution to the feature selection problem. Through the theory of natural selection and genetic recombination these solutions evolve into future populations where only diagnostically important combinations of features survive. The GA's main advantage is their ability to investigate many possible solutions simultaneously.[11] Although successful, GA algorithms are computationally very demanding, particularly as the number of available features increases.

The purpose of this study is to explore an information

theoretic approach to the feature selection problem for CAD applications. The approach utilizes the mutual information (MI) concept as the feature selection criterion. By definition, MI measures the information content of a given feature with regard to the decision task at hand.[12] Theoretically, the MI criterion offers three major advantages over other techniques.[12] First the MI measures general statistical dependence between variables, contrary to the linear correlation coefficient. Second, the MI is invariant to monotonic transformations performed on the variables, contrary to linear dimension reducers such as principal component analysis. Finally, the MI feature selection approach is independent of the decision algorithm, thus reducing computational complexity contrary to GAs.

The concept of mutual information has been applied in medical image processing, particularly for image registration tasks.[13–19] However, the application of MI as a feature selection criterion has been rather limited. In 1994, Battiti *et al.* presented the application of MI for pattern recognition using simulated data and benchmark databases.[20] In 1996, Zheng *et al.* applied MI for feature selection to perform pattern recognition using a radial-basis neural network.[21] Recently, Last *et al.* presented an information-fuzzy neural network for feature selection.[22] The study evaluated the impact of their feature selection technique on the diagnostic performance of a decision tree.

The focus of this paper is to demonstrate the application and impact of the MI criterion for feature selection when developing texture-based CAD tools for the automated diagnostic interpretation of medical images. Texture analysis methods produce a set of features with numerical values that have inherently different diagnostic significance. These values are assembled in a feature vector for texture characterization. Often, the feature vector is prohibitively large. Therefore, feature reduction is critical to optimize the diagnostic performance and robustness of the final CAD tool. Our study is based on a database of texture features extracted from perfusion lung scans for the diagnosis of acute pulmonary embolism. The ultimate goal is to develop a CAD tool that utilizes the smallest possible number of texture features without sacrificing diagnostic performance. Given this image database, we aim to apply and evaluate the effectiveness of the MI criterion in improving the feature selection process compared to other techniques extensively utilized in CAD.

II. MATERIALS AND METHODS

A. Definition of the mutual information concept and its properties

Mutual information (MI) is a basic concept in information theory. It is a measure of general interdependence between random variables.[12] Specifically, given two random variables (r.v.) X and Y, the mutual information $I(X;Y)$ is defined as follows:

$$I(X;Y)=H(X)+H(Y)-H(X,Y). \tag{1}$$

$H(\,)$ is the *entropy* of a random variable and measures the uncertainty associated with it. For a continuous random variable X, $H(X)$ is defined as

$$H(X)=-\int p(x)\log_2 p(x)dx. \tag{2}$$

If X is a discrete r.v., $H(X)$ is defined as follows:

$$H(X)=-\sum p(X)\log_2 p(X). \tag{3}$$

In both cases $p(X)$ represents the marginal probability distribution of r.v. X. Based on the above formulas, it is apparent why the entropy is often considered a measure of uncertainty. As an example, let the r.v. X represent the presence of a disease D. If there is no uncertainty about disease presence $[p(X=D)=1,\ p(X=\overline{D})=0]$ or disease absence $[p(X=D)=0,\ p(X=\overline{D})=1]$, then the entropy $H(X)$ equals zero. If however, there is high uncertainty about the presence or absence of disease $[p(X=D)=p(X=\overline{D})=0.5]$, then the entropy $H(X)$ equals 1. Generally, if any one of N diseases are equally possible $[p(X)=1/N]$, then $H(X)$ achieves its maximum value ($\log_2 N$). This value represents the highest possible uncertainty about the r.v. X.

Using the Bayes rule on conditional probabilities, Eq. (1) can be rewritten as

$$I(X;Y)=H(X)-H(X|Y)=H(Y)-H(Y|X). \tag{4}$$

Several observations can be made based on Eq. (4). $H(X)$ measures the *a priori* uncertainty of the r.v. X. $H(X|Y)$ measures the conditional *a posteriori* uncertainty of X after Y is observed. The mutual information $I(X;Y)$ measures how much the uncertainty of X is reduced if Y has been observed. It can be easily shown that if X and Y are generally independent, then $H(X,Y)=H(X)+H(Y)$. Consequently their mutual information is zero, i.e., observing Y does not reduce the uncertainty of X. If, however, $X=Y$ then $I(X;X)=H(X)$. Therefore, the entropy is also called self-information.

Since mutual information makes no assumptions about the nature of the relationship between variables, it is quite general and often regarded as a generalization of the linear correlation coefficient. If, however, X and Y are Gaussian random variables, it has been shown that their MI is a simple transformation of their linear correlation coefficient ρ:[23]

$$I(X;Y)=-\tfrac{1}{2}\log(1-\rho^2). \tag{5}$$

The concept of mutual information can be easily expanded to include more than two random variables. According to the chain rule,[12] the joint mutual information (JMI) between a set of features (X_1, X_2, X_3,…,X_N) and the outcome Y (i.e., diagnosis) is

$$I(X_1,X_2,X_3,\ldots,X_N;Y)=\sum_{i=1}^{N} I(X_i;Y|X_{i-1},X_{i-2},\ldots,X_1). \tag{6}$$

JMI was introduced to describe how much the information provided by the feature vector $(X_1,X_2,\ldots,X_N)$ decreases the uncertainty about the r.v. Y. Given a large set of features, it

is expected that some of them may be dependent on each other. Therefore, selecting features based on their individual MI with the output can produce subsets that contain informative yet redundant features. JMI is a more appropriate feature selection criterion since it can provide an optimal subset that contains not only the most relevant but also the least redundant features.

B. Estimation of the mutual information

This section describes the techniques employed in our study to estimate the mutual information between random variables. The MI between two random variables X and Y (i.e., a texture feature and the PE diagnosis, respectively) was estimated using the histogram approach. According to this method, the probability density function of each variable is approximated using a histogram. Then, the MI can be calculated according to the following equation:

$$I(X;Y)=\sum_{x}\sum_{y}P(X,Y)\log_2\frac{P(X,Y)}{P(X)P(Y)}, \tag{7}$$

where the summations are calculated over the appropriately discretized values of the random variables X and Y.[12] For each histogram bin, the joint probability distribution $P(X,Y)$ is estimated by counting the number of cases that fall into a particular bin and dividing that number with the total number of cases. The same technique is applied for the histogram approximation of the marginal distributions $P(X)$ and $P(Y)$.

The number of bins selected is a critical issue. In the past, several investigators have used a fixed number of bins chosen empirically. We chose to formalize the approach using the Gaussianity rule that was recently applied in speech processing.[24] With non-Gaussian data it has been proposed that $\log_2 N+1+\log_2(1+\hat{\kappa}\sqrt{N/6})$ is the proper number of bins for histogram estimation, where $\hat{\kappa}$ is the estimated kurtosis.[25] Since $\hat{\kappa}=0$ for Gaussian distributions, the proper number of bins is $\log_2 N+1$ for Gaussian data according to Sturge's rule.[25] Furthermore, since the distribution for the values of each feature are not known *a priori*, the range of values is divided into equal segments. For consistency, we calculated the mean μ and standard deviation σ of each feature. Then, the interval $[\mu-2\sigma,\mu+2\sigma]$ was divided into the predetermined number of equal segments. Any rare points falling outside the predetermined interval were assigned to the extreme left or right bins when calculating the histograms. The above rules were followed consistently for all features. Clearly, the random variable representing the diagnosis is discrete by nature. Each possible value (0 or 1) represents the absence or presence of disease, respectively.

The estimation of the JMI between a set of features and the PE diagnosis is more complex. Given a set of N features, there are 2^N possible subsets. An exhaustive search requires that the JMI is estimated for each one of the possible subsets. For example, with only 12 features, there are 4096 possible feature subsets to be considered. The feature subset that has the maximum JMI with the output variable is the optimal one. As the number of selected features increases, exponentially more data samples are required for reliable JMI estimation. Consequently, two heuristic approaches have been proposed for the task.[20,21] These techniques are based on selecting features in a stepwise mode so that each new feature selected has the highest individual MI with the output and the lowest possible JMI with the preselected features. Both methods start by selecting the single feature X_i that has the highest MI with the output variable Y. At each selection step, method A proceeds by selecting the candidate feature X_j that maximizes a weighted difference

$$I(X_j;Y)-\beta\sum_{k}I(X_k;X_j), \tag{8}$$

where k represents preselected features and j represents the candidate features.[20] Parameter β takes values between 0.5 and 1.0 and its optimal value is determined empirically. The alternative method B proceeds by selecting the candidate feature X_j with the smallest Euclidean distance from the point of maximum $I(X_j;Y)$ and minimum $\Sigma_k I(X_k;X_j)$, where k represents again preselected features.[21] It should be emphasized that both techniques are suboptimal and as such they are not guaranteed to provide consistently a better solution to the feature selection problem.

C. Application in CAD

The MI-based feature selection techniques were applied for the computer-aided diagnosis of acute pulmonary embolism (PE). The application was based on a database of 45 patients who underwent ventilation-perfusion scintigraphy due to suspicion of acute PE. The nuclear studies were performed and interpreted according to the PIOPED protocol at Duke University Medical Center.[26] Forty-one out of the 45 scans were identified as nondiagnostic for PE based on the original PIOPED criteria.[27] It should be noted that the differential diagnosis of PE in the selected patient sample was a particularly challenging task. The majority of the patients' lungs (84/90) had perfusion defects related to a variety of lung diseases often coexisting with pulmonary embolism (i.e., obstructive lung disease, pleural effusion, atelectasis, and parenchymal opacities). Thus, invasive pulmonary angiography was necessary to establish diagnosis. The presence of PE was angiographically confirmed in 13 patients and excluded in 32 patients. Furthermore, 18 out of the 90 lungs were identified as having perfusion defects related to PE.

The same image database has been utilized before for a study focusing on the fractal properties of the perfusion lung scans.[28] The present study aims to analyze the same images using statistical texture features. The study focused only on the posterior view of the perfusion lung scans analyzing each lung separately (for a total of 90 lungs). Initially, the natural edge of the normal lung was manually outlined by an experienced nuclear medicine physician. Then, texture analysis was performed on each outlined lung to extract several features. These texture features were based on the spatial gray level dependence (SGLD) matrix that describes local interactions of pairs of image pixels by taking into account their distance and angle direction.[29,30] By changing the distance and direction between the pixel pairs when defining the spa-

tial relationship, different SGLD matrices can be constructed. For our study, SGLD matrices were formed for three distances ($d=1,2,3$ pixels) and four directions ($\theta=0°$, 45°,90°,135°) for a total of 12 SGLD matrices. Twelve texture parameters were calculated from each SGLD matrix by weighted combination of the matrix elements. The texture parameters are known as Haralick's texture measures and include angular second moment, contrast, correlation, variance, inverse difference moment, sum average, sum variance, sum entropy, entropy, difference entropy, and two information correlation moments.[29,30] The total number of features was 144.

It has been shown before that as the dimensionality of the feature vector increases so does the bias introduced in the development and diagnostic evaluation of the CAD tool.[31,32] This bias can be a serious concern with limited databases such as in our case. In an effort to reduce potential bias, our study focused on only 12 texture features. Specifically, for each one of the 12 Haralick features (i.e., second angular moment, contrast, etc.) we formed the average across all distances and directions considered. Then, we focused on developing a CAD tool based on only those 12 ''average'' features. The CAD was designed to detect PE in a given lung of a perfusion lung scan. Therefore, each patient lung was described by 12 texture features and it was treated as a separate sample case. The overall objective was to maximize CAD performance while utilizing the smallest possible number of the available texture features.

D. Performance evaluation

The MI- and JMI-based algorithms described before were applied to the PE dataset to select diagnostically ''informative'' texture features. For comparison purposes, we also performed the popular stepwise linear discriminant feature selection. All selection techniques were performed on the entire dataset before any decision modeling was attempted.

Briefly, the stepwise linear discriminant analysis begins by selecting the feature with the largest difference in the mean values between the two classes. The difference is measured by the one way analysis of variance F statistic.[9] Subsequently, the next feature is added so that it improves the discrimination power of the model in combination with the existing finding. When a feature is entered into the statistical model, only its linear correlation with the pre-entered findings is considered. In addition, at any step, no finding can be removed. The selection process continues for as long as the discrimination power of the model improves. The final result of the stepwise selection process is a subset of features ordered in terms of importance. Conceptually, the stepwise feature selection is similar to JMI approximation technique described before. They are both iterative processes and at each step they select the next important feature based on the ones already selected. The difference lies in how they measure the statistical relationship between the candidate feature and the pre-entered features.

The selected features were then merged into a final diagnosis using a linear and a nonlinear decision model. The linear model was the linear discriminant analysis (LDA). The nonlinear model was an artificial neural network (ANN). The network had a three-layer, feed-forward architecture and it was trained using the Levenberg–Marquardt algorithm.[33] This algorithm has been shown to train feed-forward networks significantly faster than the usual gradient descent backpropagation algorithm. The number of input nodes varied depending on the number of texture features selected for the diagnostic task. The number of hidden nodes was determined empirically and it was found that four hidden nodes worked well. The output layer had a single node with target values 0 if PE was absent and 1 if PE was present. The results were based on the leave-one-out cross validation sampling scheme. The ANN experiments were performed with MATLAB software (The Math Works, Inc., Natick, MA). The stepwise feature selection and linear decision models were implemented with SAS/STAT software (SAS Institute, Inc., Cary, NC) using the processes STEPDISC and DISCRIM with the CROSSVALIDATE option.

The comparison of the MI-based vs the stepwise linear discriminant feature selection techniques was designed to test the hypothesis that a nonlinear feature selection technique is better suited to a nonlinear decision model. In addition, a GA-based feature selection was performed in combination with the ANN decision model. Comparing the MI-based feature selection technique to the GA will help demonstrate any potential benefits of using one technique over the other with ANNs.

The GA feature selection approach has been described before.[3,6,7] Generally, it starts with a large population of combinations of features, each one attempting to optimize CAD performance in a given database. The population is called a *generation*. Each feature combination is called a *chromosome* and represents a different solution to the feature selection problem. The chromosomes are copied to form a new population of chromosomes based on their *fitness function*. The fitness function measures the ''quality'' of the chromosome. For this application, an appropriate fitness function is the Receiver Operating Characteristic (ROC) area index achieved by the ANN that utilizes the particular feature combination represented by the chromosome. The higher the ROC area index of the ANN, the greater the probability that the chromosome will survive and contribute to the next generation. Consequently, the expected relative frequency of a chromosome in the new population is proportional to its fitness function (i.e., roulette wheel selection).[11] The replicated chromosomes undergo ''crossover'' so that components from the better performing chromosomes can be combined. Mutation can also take place by the occasional alteration of a randomly chosen gene. The above sequence of selection, crossover, and mutation continues until a best-performing feature combination emerges. The application of the GA feature selection technique is very time-consuming because it requires training and testing of several ANNs for many generations. Custom software written in the C programming language was used to implement the GA algorithm. The software ran on an UltraSPARC workstation (Sun Microsystems, Mountain View, CA). The GA search was implemented for

TABLE I. Normalized skewness and kurtosis for all features considered. Values that fail the Gaussianity test appear in *italic*.

Features	Skewness	Kurtosis
F_1: Angular second moment	*2.05*	0.48
F_2: Contrast	*−2.40*	*2.73*
F_3: Correlation	*−5.19*	*3.77*
F_4: Variance	−0.12	−0.51
F_5: Inverse difference moment	*2.71*	1.06
F_6: Sum average	−1.00	0.09
F_7: Sum variance	0.03	−0.55
F_8: Sum entropy	*−2.71*	0.98
F_9: Entropy	*−2.91*	0.98
F_{10}: Difference entropy	*−3.41*	1.45
F_{11}: Information measure of correlation-1	−0.06	1.52
F_{12}: Information measure of correlation-2	*−3.87*	*2.61*

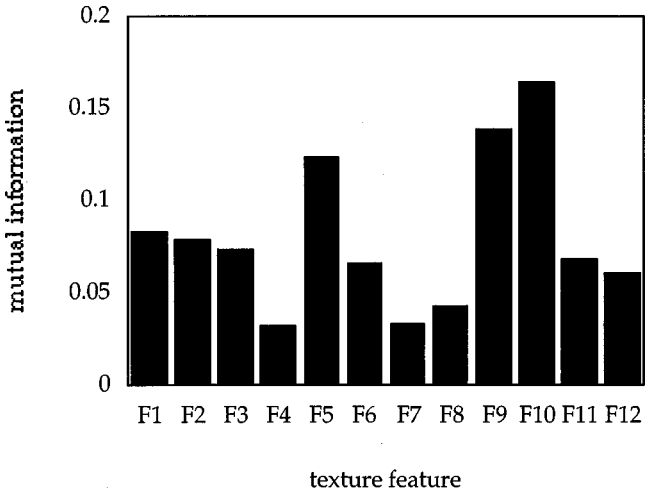

FIG. 1. The mutual information between each texture feature and diagnosis.

50 generations. Each generation contained 100 chromosomes. The crossover rate was set at 50% and the mutation rate was set at 0.1%. For each chromosome, an ANN was implemented and evaluated using the leave-one-out cross validation sampling scheme.

Diagnostic performance was measured based on ROC area indices. We used the ROCKIT software package developed by Metz *et al.* at the University of Chicago (http://www-radiology.uchicago.edu/krl/toppage11.htm) to fit ROC curves to the output responses of the decision models implemented in this study.

III. RESULTS

A. Non-Gaussianity of the texture features

To determine the number of bins required for the histogram estimation of the MI and JMI, we first tested whether the distribution of each feature is Gaussian. Several techniques have been proposed for testing the Gaussianity of a dataset. For this particular study we tested the normalized skewness and normalized kurtosis.[34] The null hypothesis for the above test is that a set of observations follows the Gaussian distribution if the normalized skewness and normalized kurtosis of the data follow the standard Gaussian distribution $N(0,1)$. The normalized skewness and normalized kurtosis are defined as follows:

$$S=\frac{1}{\sqrt{6N}\sigma^3}\sum_{i=1}^{N}(x_i-\bar{x})^3, \tag{9}$$

$$K=\frac{1}{\sqrt{24N}\sigma^4}\sum_{i=1}^{N}(x_i-\bar{x})^4-\sqrt{\frac{3N}{8}}, \tag{10}$$

where N, $\bar{x}$, σ are the sample size, sample mean, and sample standard deviation. At the significance level of $\alpha=0.05$, the critical values for the standard Gaussian distribution are ± 1.96. Table I reports the normalized skewness and kurtosis for each one of the 12 texture features of our dataset. As the table shows several features display non-Gaussian properties because the absolute value of either their normalized skewness, normalized kurtosis or both exceed the critical value 1.96. Depending on if a feature is Gaussian or not, we followed the formulas presented in Sec. II B to calculate the optimal number of bins. For our dataset, the number of bins ranged between 5 and 10. Furthermore, since our dataset consists of a mixture of Gaussian and non-Gaussian RV's, exploring the mutual information criterion as a feature selection technique is strongly justified. If there are diagnostically important relationships among the non-Gaussian features, then the MI can potentially capture them more efficiently than the linear correlation coefficient.

B. MI between a feature and diagnosis

Figure 1 shows the mutual information between each feature and the diagnosis. The estimation is based on all available cases of the dataset. Clearly, features F_{10}, F_9, and F_5 stand out as the most ''informative'' ones for the diagnosis of PE. This is quite interesting since individual ROC analysis shows that features F_2 and F_{11} are the most predictive of PE with independent ROC area indices of $A_Z(F_2)=0.73\pm0.06$ and $A_Z(F_{11})=0.70\pm0.06$, respectively. The remaining features showed individual ROC areas close to chance behavior ranging between 0.51 and 0.60 (Fig. 2).

C. Impact of feature selection technique on CAD performance

Table II summarizes our results. The table displays the subsets of features selected using the stepwise linear discriminant technique and the MI-based techniques. For each subset, the table shows the ROC area index achieved by the linear decision model (LDA) and the nonlinear ANN when they were used to merge the same subset of features into a final diagnosis. The table presents results for up to six selected features. The diagnostic performance of both decision models degraded as the number of selected features increased more than six.

When all 12 features are included, both the LDA and the ANN have clinically unacceptable diagnostic performance although the ANN ($A_Z=0.64\pm0.07$) appears to perform bet-

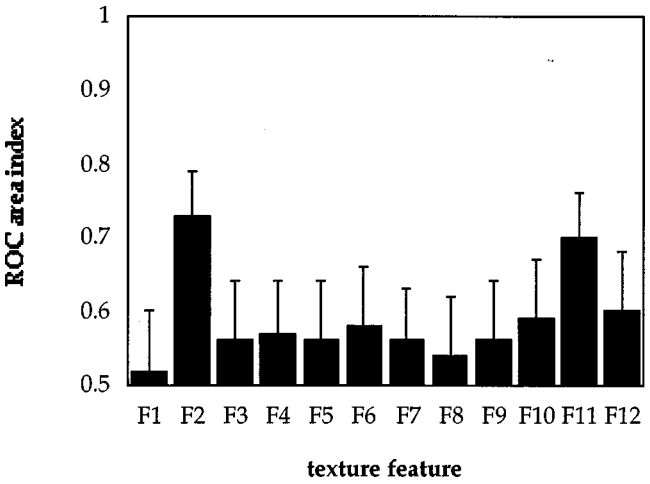

FIG. 2. The independent ROC area index with the corresponding error bar for each texture feature.

ter than the LDA ($A_Z = 0.53 \pm 0.07$). The difference however was not statistically significant, primarily due to the small number of positive cases.

It is apparent that the MI- and JMI-based techniques selected features very different from the ones selected with the stepwise linear technique. The different selections affected the diagnostic performance of both decision models. The best linear decision model achieved ROC area index of 0.70 ± 0.06. It utilized the first two features (F_2, F_{12}) selected by the stepwise linear technique. Contrary, the best ANN achieved a significantly better diagnostic performance ($A_Z = 0.82 \pm 0.06$). The ANN utilized five features (F_{10}, F_9, F_5, F_1, F_2) selected independently based on their MI content with respect to the diagnosis.

TABLE II. Comparison of feature selection methods in terms of ROC area.

Selected features	LDA	ANN
All	0.53 ± 0.07	0.64 ± 0.07
Stepdisc		
F_2, F_{12}	*0.70±0.06*	0.68 ± 0.07
F_2, F_{12}, F_4	0.69 ± 0.07	0.68 ± 0.06
F_2, F_{12}, F_4, F_{11}	0.70 ± 0.06	0.67 ± 0.07
$F_2, F_{12}, F_4, F_{11}, F_5$	0.69 ± 0.06	0.69 ± 0.07
$F_2, F_{12}, F_4, F_{11}, F_5, F_7$	0.67 ± 0.07	0.65 ± 0.09
MI		
F_{10}, F_9	0.58 ± 0.07	0.73 ± 0.07
F_{10}, F_9, F_5	0.61 ± 0.07	0.72 ± 0.05
F_{10}, F_9, F_5, F_1	0.59 ± 0.07	0.70 ± 0.08
$F_{10}, F_9, F_5, F_1, F_2$	0.60 ± 0.07	*0.82±0.06*
$F_{10}, F_9, F_5, F_1, F_2, F_3$	0.60 ± 0.07	0.75 ± 0.07
JMI-A ($\beta = 0.5$)		
F_{10}, F_9	0.58 ± 0.07	0.73 ± 0.07
F_{10}, F_9, F_3	0.55 ± 0.07	0.67 ± 0.07
F_{10}, F_9, F_3, F_2	0.64 ± 0.07	0.74 ± 0.08
$F_{10}, F_9, F_3, F_2, F_1$	0.62 ± 0.07	*0.78±0.07*
$F_{10}, F_9, F_3, F_2, F_1, F_4$	0.61 ± 0.07	0.73 ± 0.07
JMI-B (Euclidean)		
F_{10}, F_3	0.51 ± 0.08	0.64 ± 0.07
F_{10}, F_3, F_1	0.64 ± 0.07	0.66 ± 0.07
F_{10}, F_3, F_1, F_2	0.63 ± 0.07	*0.79±0.08*
$F_{10}, F_3, F_1, F_2, F_4$	0.61 ± 0.07	0.76 ± 0.06
$F_{10}, F_3, F_1, F_2, F_4, F_{11}$	0.67 ± 0.07	0.69 ± 0.07

It is interesting to observe that when the feature selection was optimized using a linear technique, the nonlinear decision model (ANN) was unable to improve upon the diagnostic performance of the linear decision model (LDA). However, the ANN was able to capitalize on the diagnostic potential of the MI-selected features while LDA showed a disappointing performance with any of the MI- or JMI-selected features.

The feature selection techniques based on the JMI concept produced subsets of features that worked better with the ANN than the LDA. However, neither one of the two JMI approximation techniques implemented in this study were able to achieve better diagnostic performance than the independently selected MI-based features. The differences though were not statistically significant.

Finally, the GA-optimized feature selection process was computationally very demanding. As mentioned before, the GA search was implemented for 50 generations. After approximately five generations, certain feature combinations started emerging as more diagnostically successful. In particular, features F_1, F_2, and F_5 remained present in the most successful chromosomes until the end. The GA recognized the feature subset $\{F_1, F_2, F_5, F_{10}\}$ as the optimal one by the end of the thirty-eighth generation. This feature subset is slightly different from the one selected using the MI-based approach $\{F_1, F_2, F_5, F_9, F_{10}\}$. Actually the feature subset $\{F_1, F_2, F_5, F_9, F_{10}\}$ was identified by the GA after only six generations. The GA-based search continued until it removed feature F_9. Although the overall ROC area index achieved by the ANN is the same using either subset, the GA eliminated a redundancy (F_9) that was maintained by the MI-based approach.

D. Effect of the MI approximation technique on feature selection

The number of bins selected for the histogram approximation of the probability distributions relevant in MI calculations is an important issue. More bins allow for more detailed representation of the probability distributions in Eq. (7). However, these details may be nothing more than noise caused by the small sample size in each bin. The potential estimation error can substantially alter the results of a study. Given the small number of positive cases, this is a serious concern with our dataset.

As mentioned before, it is a common practice to determine the number of bins empirically by trying several different values that remain fixed for all features. We explored this approach as well. Specifically, we performed the MI estimation by setting the number of bins to a fixed value (3, 5, 8, and 10) for all features. The lower and upper values were selected as follows. If we assume that all features are Gaussian, then according to Sturge's rule the appropriate number of bins is $(\log_2 N + 1)$. Since there are fewer number of positive ($N = 18$) than negative ($N = 72$) cases, a conservative estimate is five bins $(\log_2 18 + 1)$. In addition, we studied the

TABLE III. MI- and JMI-selected features based on histogram approximation with a fixed number of bins. For comparison, the table includes the features selected using a variable bin number determined according to the Gaussianity rule. The features are listed in the order they were selected.

	Three bins	Five bins	Variable bin number
MI	$F_2,F_{11},F_9,F_{10},F_5,F_3$	$F_{10},F_5,F_2,F_{11},F_1,F_9$	$F_{10},F_9,F_5,F_1,F_2,F_3$
JMI-A	$F_2,F_{10},F_5,F_9,F_8,F_1$	$F_{10},F_3,F_2,F_{12},F_{11},F_4$	$F_{10},F_9,F_3,F_2,F_1,F_4$
JMI-B	$F_2,F_{11},F_{12},F_1,F_3,F_4$	$F_{10},F_5,F_9,F_8,F_1,F_2$	$F_{10},F_3,F_1,F_2,F_4,F_{11}$

extreme case of using only three bins. Ten bins were selected as the upper end that is justified by the results in Sec. III A.

The following are the subsets S_N of features in the order of their ''information content'' where N is the fixed number of bins employed for histogram approximation:

$$S_3=\{F_2,F_{11},F_9,F_{10},F_5,F_3,F_{12},F_4,F_7,F_6,F_8,F_1\},$$

$$S_5=\{F_{10},F_5,F_2,F_{11},F_1,F_9,F_4,F_6,F_7,F_{12},F_8,F_3\},$$

$$S_8=\{F_{10},F_5,F_2,F_{11},F_6,F_9,F_{12},F_7,F_4,F_8,F_1,F_3\},$$

$$S_{10}=\{F_{10},F_5,F_{11},F_4,F_8,F_9,F_1,F_7,F_2,F_3,F_6,F_{12}\}.$$

The overall ordering of the features changed substantially depending on how many bins were used for the histogram approximation. However, certain texture features consistently prevailed as more important (F_2, F_5, F_9, F_{10}). It is interesting that when the very conservative number of three bins was employed, the two features with the highest MI (F_2 and F_{11}) are the ones that have independently the highest ROC area index. This result suggests that a very small number of bins may be more appropriate for this particular database. Implementation of the two JMI techniques using three and five bins produced the feature subsets presented in Table III. Although the JMI-A technique based on three bins and the JMI-B technique based on five bins worked the best, neither one selected feature subsets that achieved ANN ROC area indices higher than the best one reported in Table II. Both techniques, however, recognized that the feature combination $\{F_1,F_2,F_5,F_9,F_{10}\}$ is diagnostically important.

IV. DISCUSSION

By definition, a CAD tool can be considered as a system that reduces the uncertainty about a medical diagnosis by using information in the feature vector. Many CAD tools are based on elaborate image processing schemes that produce feature vectors with high dimensionality. In practice, adding new features to a CAD tool can often degrade its diagnostic performance. This effect is well known in pattern recognition as the ''curse of dimensionality.''[10] Consequently, optimizing feature selection in any CAD scheme is crucial.

The main focus of this paper was to investigate the mutual information concept as a feature selection criterion in CAD. This was primarily a comparative study based on a small database of clinically challenging perfusion lung scans. The study showed that the MI concept lends itself to some exciting possibilities not available with the feature selection techniques traditionally used in CAD. Some of them are discussed below.

Mutual information is a measure of information content. As such, it can be used to measure the information content of a feature or set of features and the desired output (i.e., correct diagnosis). Individual features or subsets of features with low information content can be excluded from a CAD scheme. Utilizing a criterion to eliminate irrelevant or redundant features is by no means a unique concept. Stepwise linear discriminant selection is based on the same idea. However, the MI selection criterion can offer advantages because it measures not only linear but general dependencies. If nonlinear dependencies exist among the available features and they are diagnostically important, then the MI-based approach is better suited to nonlinear decision models such as the popular ANNs. Our study certainly confirmed that. When feature selection was ''optimized'' using a linear technique, both the linear (LDA) and nonlinear (ANN) decision models achieved the same diagnostic performance. When the features were selected based on the more general MI concept, the ANN achieved a significantly better diagnostic performance than the LDA. It should be noted that the feature subset that proved to be the most diagnostically important for the ANN included non-Gaussian features.

Due to its computational simplicity, the stepwise linear selection is a common practice in CAD applications. However, researchers have recognized the limitations of feature selection techniques that capitalize on the linear relationships of features. Genetic algorithms were introduced in CAD to address precisely this issue. Although efficient, GAs are computationally very demanding because both feature selection and decision modeling are executed simultaneously by the GA. As the number of available features increases, the GA approach becomes impractical when implemented with a complex decision algorithm and an elaborate data sampling scheme. In our study, the GA approach was computationally very demanding primarily due to the leave-one-out sampling scheme. Our study showed that the MI-based selection technique is a competitive alternative. Given our database, the subsets of MI-selected features were similar to those identified by the GA as diagnostically important. Since the MI-based feature selection is executed before decision modeling is performed, it is a computationally affordable technique with databases that contain hundreds of available texture features.

There is an important issue related to the MI estimation from a finite set of samples. It has been shown before[35] that as the number of bins increases, the histogram approximation technique tends to overestimate the mutual information due to the finite data size effect. Authors have hypothesized that

this overestimation effect should not affect the relative ordering of the features.[20,21] However, this is true only if the data partition reflects the underlying probability distribution for each feature. That was certainly a critical issue with our dataset given the small number of positive cases. Feature ordering changed substantially based on the number of bins utilized for the estimation of MI and JMI. Although the overall ordering of the features changed depending on how many bins were used for the histogram approximation, certain texture features consistently prevailed as more important (F_1, F_2, F_5, F_9, F_{10}). How to determine the number and size of each bin is an ongoing research topic in information theory. Sophisticated adaptive discretization techniques have been proposed for the task.[36–38] In practice, we found that the variable bin number approach depending on the Gaussian or non-Gaussian nature of the data worked well for our study. However, a very conservative fixed selection for the number of histogram bins (three or five) produced very similar results.

At first glance, it is surprising that the JMI-based techniques were unable to improve the CAD performance achieved based on the MI-selected features (Table II). Such result implies that there were not necessarily strong redundancies among the diagnostically important features. Apparently, this is not true since the GA was able to recognize one (feature F_9). However, given the small number of truly positive cases, it is expected that the MI estimation between any two texture features was highly biased affecting the overall stepwise ordering of the selected features. The small dataset can cause substantial fluctuations on the estimated MI since many histogram bins remain ''unoccupied.'' These fluctuations increase the estimated MI because they are interpreted as important structure details of the joint probability distribution. Overestimating the MI between two texture features means overestimating their degree of redundancy which can affect substantially the results of the JMI-based feature selection techniques.

Another important issue with feature selection from limited datasets is the bias introduced when the same dataset is used for feature selection and classifier development.[31,39] Both the stepwise linear technique and the MI-based feature selection techniques were implemented using the whole dataset in our study. Therefore, the CAD performance may have been overestimated in both cases. However, due to the comparative nature of our study, this potential bias should have a limited effect on the conclusions. Theoretically, the GA approach is immune to that bias since the feature selection and decision modeling are performed simultaneously.

Finally, the CAD model that appears to work best for PE diagnosis is an ANN that utilizes four texture features: angular second moment (F_1), contrast (F_2), inverse difference moment (F_5), and difference entropy (F_{10}). The overall diagnostic performance of the ANN was 0.82 ± 0.06 based on a set of clinically challenging perfusion lung scans. The ANN performance is comparable to that achieved by an experienced nuclear medicine physician ($A_Z = 0.81 \pm 0.08$) and an ANN that utilizes multifractal textural information ($A_Z = 0.81 \pm 0.06$).[28] It should be noted however that the physician's diagnostic performance was based on the patients' complete set of imaging studies (perfusion scan, ventilation scan, and chest radiograph) and not only on the perfusion lung scan.

To summarize, our study showed that the mutual information is a promising feature selection criterion for nonlinear CAD models. It appears to bridge nicely the computational simplicity of the stepwise linear selection techniques with the effectiveness of the genetic algorithms. Special attention should be placed on MI estimation from limited datasets.

ACKNOWLEDGMENT

This work was supported by Grant No. RG 98-0324 from the Whitaker Foundation.

[a]Electronic mail: gt@deckard.mc.duke.edu

[1] J.Y. Lo, J.A. Baker, P.J. Kornguth, and C.E. Floyd, Jr., ''Computer-aided diagnosis of breast cancer: artificial neural network approach for optimized merging of mammographic features,'' Acad. Radiol. **2**, 841–850 (1995).

[2] M.F. McNitt-Gray, H-K. Huang, and J.W. Sayre, ''Feature-selection in the pattern classification problem of digital chest radiograph segmentation,'' IEEE Trans. Med. Imaging **14**, 537–547 (1995).

[3] B. Sahiner, H-P. Chan, D.T. Wei, N. Petrick, M.A. Helvie, and M.M. Goodsitt, ''Image feature selection by a genetic algorithm: Application to classification of mass and normal breast tissue,'' Med. Phys. **23**, 1671–1684 (1996).

[4] G.D. Tourassi, C.E. Floyd, Jr., and R.E. Coleman, ''Improved noninvasive diagnosis of acute pulmonary embolism using optimally selected clinical and chest radiographic findings'' Acad. Radiol. **3**, 1012–1918 (1996).

[5] H-P. Chan, B. Sahiner, K.L. Lam, N. Petrick, M.A. Helvie, M.M. Goodsitt, and D.D. Adler, ''Computerized analysis of mammographic microcalcifications in morphological and texture feature spaces,'' Med. Phys. **25**, 2007–2019 (1998).

[6] B. Sahiner, H-P. Chan, N. Petrick, M.A. Helvie, and M.M. Goodsitt, ''Design of a high-sensitivity classifier based on a genetic algorithm: application to computer-aided diagnosis,'' Phys. Med. Biol. **43**, 2853–2871 (1998).

[7] B. Zheng, Y-H. Chang, X-H. Wang, W.F. Good, and D. Gur, ''Feature selection for computerized mass detection in digitized mammograms by using a genetic algorithm,'' Acad. Radiol. **6**, 327–332 (1999).

[8] M.F. McNitt-Gray, E.M. Har, N. Wyckoff, J.W. Sayre, J.G. Goldin, and D.R. Aberle, ''A pattern classification approach to characterizing solitary pulmonary nodules imaged on high resolution CT: Preliminary results,'' Med. Phys. **26**, 880–888 (1999).

[9] R.J. Jennrich, ''Stepwise discriminant analysis,'' in *Statistical Methods for Digital Computers*, edited by K. Einslein, K. Ralston, and H.S. Wilf (Wiley, New York, 1977).

[10] C.M. Bishop, *Neural Networks for Pattern Recognition* (Clarendon, Oxford, 1995).

[11] D.E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning* (Addison-Wesley, Reading, MA, 1989).

[12] T.M. Cover and J.A. Thomas, *Elements of Information Theory* (Wiley, New York, 1991).

[13] C. Studholme, D.L.G. Hill, and D.J. Hawkes, ''Automated three-dimensional registration of magnetic resonance and positron emission tomography brain images by multiresolution optimization of voxel similarity measures,'' Med. Phys. **24**, 25–35 (1997).

[14] S. Sanjay-Gopal *et al.*, ''A regional registration technique for automated interval change analysis of breast lesions on mammograms,'' Med. Phys. **26**, 2669–2679 (1999).

[15] J.P.W. Pluim, J.B.A. Maintz, and M.A. Viergever, ''Image registration by maximization of combined mutual information and gradient information,'' IEEE Trans. Med. Imaging **19**, 809–814 (2000).

[16] L. Thurfjell *et al.*, ''Improved efficiency for MRI-SPET registration based on mutual information,'' Eur. J. Nucl. Med. **27**, 847–856 (2000).

[17] J.F. Krucker, C.R. Meyer, G.L. LeCarpentier, J.B. Fowlkes, and P.L. Carson, ''3D spatial compounding of ultrasound images using image-based nonrigid registration,'' Ultrasound Med. Biol. **26**, 1475–1488 (2000).
[18] G. Berks, A. Ghassemi, and D.G. von Keyserlingk, ''Spatial registration of digital brain atlases based on fuzzy set theory,'' Comput. Med. Imaging Graph. **25**, 1–10 (2001).
[19] M.A. Viergever, J.B.A. Maintz, W.J. Niessen, H.J. Noordmans, J.P.W. Pluim, R. Stokking, and K.L. Vincken, ''Registration, segmentation, and visualization of multimodal brain images,'' Comput. Med. Imaging Graph. **25**, 147–151 (2001).
[20] R. Battiti, ''Using mutual information for selecting features in supervised neural net learning,'' IEEE Trans. Neural Netw. **5**, 537–550 (1994).
[21] G.L. Zheng and S.A. Billings, ''Radial basis function network configuration using mutual information and the orthogonal least squares algorithm,'' Neural Networks **9**, 1619–1637 (1998).
[22] M. Last, A. Kandel, and O. Maimon, ''Information-theoretic algorithm for feature selection,'' Pattern Recogn. Lett. **22**, 799–811 (2001).
[23] W. Li, ''Mutual information functions versus correlation functions,'' J. Stat. Phys. **60**, 823–837 (1990).
[24] H.H. Yang, S. Van Vuuren, S. Sharma, and H. Hermansky, ''Relevance of time frequency features for phonetic and speaker-channel classification,'' Speech Commun. **31**, 35–50 (2000).
[25] W.N. Venables and B.D. Ripley, *Modern Applied Statistics with S-Plus* (Springer, New York, 1994).
[26] A. Gottschalk, J.E. Juni, H.D. Sostman, R.E. Coleman, J. Thrall, K.A. McKusick, J.W. Froelich, and A. Alavi, ''Ventilation-perfusion scintigraphy in the PIOPED study: Part I. Data collection and tabulation,'' J. Nucl. Med. **34**, 1109–1118 (1993).
[27] The PIOPED Investigators, ''Value of the ventilation/perfusion scan in acute pulmonary embolism: Results of the prospective investigation of pulmonary embolism diagnosis (PIOPED),'' J. Am. Med. Assoc. **263**, 753–2759 (1990).
[28] G.D. Tourassi, E.D. Frederick, C.E. Floyd, Jr., and R.E. Coleman, ''Multifractal texture analysis of perfusion lung scans as a computer aid for acute pulmonary embolism,'' Comput. Biol. Med. **31**, 15–25 (2001).
[29] R.M. Haralick, K. Shanmugan, and I. Dinstein, ''Textural features for image classification,'' IEEE Trans. Syst. Man Cybern. **3**, 610–621 (1973).
[30] R.M. Haralick, ''Statistical and structural approaches to texture,'' Proc. IEEE **67**, 786–804 (1979).
[31] H-P. Chan, B. Sahiner, R.F. Wagner, and N. Petrick, ''Classifier design for computer-aided diagnosis: Effects of finite sample size on the mean performance of classical and neural network classifiers,'' Med. Phys. **26**, 2654–2668 (1999).
[32] B. Sahiner, H-P. Chan, N. Petrick, R.F. Wagner, and L. Hadjiiski, ''Feature selection and classifier performance in computer-aided diagnosis: The effect of finite sample size,'' Med. Phys. **27**, 1509–1522 (2000).
[33] M.T. Hagan and M. Menhaj, ''Training feed-forward networks with the Marquardt algorithm,'' IEEE Trans. Neural Netw. **5**, 989–993 (1994).
[34] A. Stuart and J.K. Ord, *Kendall's Advanced Theory of Statistics* (Edward Arnold, Paris, 1994), Vol. 1.
[35] A. Treves and S. Panzeri, ''The upward bias in measures of information derived from limited data samples,'' Neural Comput. **7**, 399–407 (1995).
[36] A.M. Fraser and H.L. Swinney, ''Independent coordinates for strange attractors from mutual information,'' Phys. Rev. A **33**, 1134–1140 (1986).
[37] G.A. Darbellay and I. Vajda, ''Estimation of the information by an adaptive partitioning of the observation space,'' IEEE Trans. Inf. Theory **45**, 1315–1321 (1999).
[38] G.A. Darbellay, ''An estimator of the mutual information based on a criterion for independence,'' Comput. Stat. Data Anal. **32**, 1–17 (1999).
[39] M.A. Kupinski and M.L. Giger, ''Feature selection with limited datasets,'' Med. Phys. **26**, 2176–2182 (1999).