Computer-aided Detection: The Impact of Machine Learning Classifier and Image Feature Selection on Scheme Performance

Xingwei Wang, Dror Lederman, Jun Tan, Bin Zheng

Department of Radiology
University of Pittsburgh
Pittsburgh, PA 15213, USA

wangx6@upmc.edu, ledermand@upmc.edu, tanj@upmc.edu, zhengb@upmc.edu
doi: 10.4156/ijiip.vol1.issue1.4

Abstract

Computer-aided detection and diagnosis (CAD) schemes have been developed and applied to detect suspicious lesions depicted on biomedical images. After identifying initial candidates for the targeted suspicious lesions, most CAD schemes use a pre-trained multi-image-feature based machine learning classifier to classify these candidates into two groups of positive and negative detections. Although a large number of image features and machine learning classifiers have been developed and tested using different image databases, selecting the optimal image features and a machine learning classifier remains a challenged issue in CAD development. In this study, we assembled two independent image datasets for training and testing. We optimized four machine learning classifiers (namely, artificial neural network, support vector machine, Bayesian belief network, and k-nearest neighbor algorithm), which were trained and tested using the same dataset with two sets of image features. The results showed that using the first feature set, the case-based classification performance of four classifiers measured with the normalized areas under FROC-type performance curves (AUCs) ranged from 0.925 to 0.943 without statistically significant difference (p > 0.05). When using the second image feature set, AUC values of four classifiers significantly reduced to the range from 0.886 to 0.903 (p < 0.01). This study suggested that although these four classifiers were built based on different machine learning concepts, their actual performance levels were likely to converge to the similar level when using the same image features and an independent testing dataset. Thus, selecting image features rather than a machine learning classifier plays a more important role in determining CAD performance.

Keywords: Computer-aided detection (CAD), Machine learning, Feature selection

1. Introduction

Developing computer-aided detection and/or diagnosis (CAD) schemes has been attracting rapidly growing interest in biomedical imaging research field during the last two decades. Aiming to assist clinicians (e.g., radiologists and pathologists) more accurately, consistently, and efficiently reading and interpreting biomedical images in the busy clinical environment, a large number of CAD schemes have been developed and applied to detect a variety of diseases or image biomarkers, which include but are not limited to the detection of suspicious breast lesions (e.g., masses and micro-calcification clusters) [1], lung nodules [2], pulmonary embolism (PE) [3], interstitial lung diseases (ILD) [4], colon polyps [5], abnormal metaphase chromosomes [6], and fluorescent in situ hybridization (FISH) labeled genetic biomarkers [7]. A number of studies have also investigated whether and/or how using CAD as “a second reader” might affect the performance level of radiologists in reading and interpreting medical images. For example, commercially available CAD schemes have been routinely used in the clinical practice to assist radiologists reading and interpreting film and digital based screening mammograms in USA and some other countries around the world. Several large scale prospective clinical studies showed that using CAD helped radiologists detect more breast cancers associated with micro-calcification clusters [8, 9], while the others showed that using CAD had little impact on radiologists’ overall performance in both cancer detection and recall rates [10] or even reduced radiologists’ performance level measured by the area under the receiver operating characteristic (ROC) curves [11]. Although there is no universal agreement on whether using CAD schemes at current performance level
can actually improve radiologists’ performance in interpreting medical images [12], our previous study
demonstrated that only using “highly performing” CAD improved radiologists’ performance in
detecting suspicious breast masses and micro-calcification clusters depicted on mammograms, while
using “poorly performing” CAD schemes with the higher false-positive cueing rates reduced
radiologists’ performance [13]. Therefore, both effective improvement and objective assessment of
CAD scheme performance are important and have scientific merit.

CAD schemes typically involve multiple stages in detecting and/or classifying suspicious lesions
depicted on biomedical images. Some schemes automatically search for and detect the suspicious
candidate regions (e.g., computer-aided detection schemes used as the second reader in the screening
environment) and some are only applied to process the manually selected suspicious lesions (e.g.,
computer-aided diagnosis schemes used to classify between malignant and benign lesions). No matter
which approach is used, after identifying and segmenting the regions of the identified suspicious
lesions or abnormalities, CAD schemes compute a set of image features from the segmented regions
and use a pre-trained or optimized machine learning classifier to classify these suspicious regions into
two groups of positive and negative detections. Although a large number of machine learning methods
based on different learning models or concepts [14] have been investigated and tested in developing
CAD schemes of biomedical images, which machine learning method is the “best” choice in
classifying suspicious abnormalities or lesions depicted on a variety of medical images has not been
fully investigated. One previous study compared several machine learning methods that used eight
image features and a leave-one-out cross validation method to develop classifiers that were applied to
classify between the malignant and benign breast micro-calcification clusters. The study reported that
the kernel-based statistical learning methods (e.g., a support vector machine) significantly
outperformed the back-propagation based iterative learning method (e.g., an artificial neural network)
[15]. To further test or verify whether one type of machine learning method or classifier is always
superior to the other types of classifiers in developing CAD schemes, we used a different approach in
this study to test and compare four popular machine learning methods and classifiers that have been
widely used in previous CAD studies. Two sets of image features were extracted and computed from
each suspicious breast mass region initially detected by a CAD scheme. Each machine learning
classifier was trained and tested using two independent image datasets to reduce the potential bias in
using the leave-one-out cross validation method. Based on the free-response ROC (FROC) analytical
methodology, we then compared and analyzed performance level differences among these four
classifiers and their relationship to the feature selection.

2. Materials and Methods

2.1. Four machine learning methods and classifiers

In CAD schemes of detecting cancer or other types of diseases, the pre-optimized machine learning
classifiers are typically used to classify the targeted candidate regions into two groups (classes) of
positive and negative detections. For example, Figure 1 demonstrates applying a CAD scheme to detect
suspicious breast mass regions depicted on two cranio-caudal (CC) view digital mammograms of the
left and right breasts. In this example, CAD detected and marked two suspicious mass regions (one
true-positive and one false-positive). The purpose of training machine learning classifier is to optimally
classify the suspicious regions initially detected and/or segmented by CAD scheme. Specifically, by
letting a suspicious region be represented by an $n$-dimensional feature vector $x \in \mathbb{R}^n$ and scale $d$
denote the group label (e.g., 1 – positive group and 0 – negative group), a machine learning method is
used to train and determine a classifier or a decision function $f(x)$ that is able to optimally classify
each feature vector $x$ into one of two groups. Although there are only two groups with targeted label
of 1 and 0 in CAD schemes, most of the classifiers trained by the machine learning methods typically
generates the continuous detection (or classification) scores ranging from 0 to 1. The larger the score,
the higher the likelihood of the test region being positive is. Thus, an operating threshold (e.g., 0.55)
needs to be determined and applied to actually divide the classified regions into the positive and
negative groups. From a large number of different machine learning methods that have been applied to
determine the classifiers (or decision functions) in developing CAD schemes, we only focused in this
study on comparing four popular machine learning classifiers that have also been independently optimized and applied using different image datasets and feature vectors in our own previous studies for different applications. These classifiers were built based on an artificial neural network (ANN) [16], a Bayesian belief network (BBN) [17], a support vector machine (SVM) [18], and a k-nearest neighbor (KNN) instance-based learning method [19]. A brief description on the optimization and application of these four machine learning methods in CAD schemes of medical images is provided following.

![Figure 1](image)

**Figure 1.** An example of applying CAD to detect suspicious breast mass regions. Two original CC view digital mammograms of the right and left breast are shown in (a) and (b), respectively. A true-positive mass region is detected and marked on right breast (c) and a CAD-cued false-positive region is marked on left breast (d).

A) An artificial neural network

The development of an artificial neural network (ANN) was originally inspired in part by the observations that (1) biological learning systems are built of very complex webs of interconnected neurons and (2) the information processing abilities of the biological neural systems follow from highly parallel processes operating on representations that are distributed over many neurons [14]. Thus, an ANN is to mimic the highly parallel computation based on distributed representation. One important advantage of ANN is that it provides a relatively robust approach to approximate general and explicit target functions with potentially noisy and/or incomplete training data. Although many different types of ANNs have been developed and tested in different application tasks, a three-layer feed-forward ANN is typically sufficient for the most applications including CAD schemes of biomedical images. In a feed-forward ANN used in our study, the relationship between the input (feature) neurons ($x_i$) and the output (decision) neuron ($Y$) is determined by

$$Y = g\left[ \sum_{j=1}^{m} w_{j} g\left( \sum_{i=1}^{n} w_{ji} x_{i} + \theta_{in} \right) + \theta_{hid} \right]$$

where $g(z) = 1/(1 + e^{-z})$, $w_{ji}$ is the weight from the $j$th hidden neuron to the output neuron, $w_{ji}$ is the weight from $i$th input neuron to the $j$th hidden neuron, $\theta_{in}$ and $\theta_{hid}$ are two bias neurons in the input and hidden layer of the ANN, respectively. A nonlinear sigmoid function,
\[ O_{pj} = \frac{1}{1 + e^{-\sum_{i} w_{pi} O_{pi} + \theta_i}} \]
is used as the activation function for each process neuron, where \( O_{pj} \) is the \( j \)th element of the output pattern produced by the input pattern \( O_{pi} \). Then, using a back-propagation method, the weights that link between the neurons are iteratively adjusted and computed as follows:

\[ \Delta w_{pj}(k + 1) = \eta \delta_{pj} O_{pi} + \alpha \Delta w_{pj}(k) \]

where \( \eta \) is the learning rate, \( \alpha \) is momentum that determines the effect of past weight changes on the current changes, \( k \) is the number of iterations, and \( \delta_{pj} \) is the error between the desired and actual ANN output value. Using a set of training data (feature vectors), the ANN is iteratively trained to reduce the error (i.e., minimize the difference between the desired and actual ANN output values). The ANN then builds a single “global” optimization target function to cover the entire case domain. Based on our previous experience in training ANNs to reduce over-training and maintain higher level of robustness for the CAD scheme [16], we selected and implemented a fixed training iteration number (1000) and a relatively large ratio between the momentum and learning rate (0.9 and 0.01) into the ANN training protocol in this study.

B) A support vector machine

A support vector machine (SVM) uses a constructive machine learning process based on the statistical learning theory. Unlike ANN that minimizes the mean square error over the training dataset, SVM aims to minimize the bound on the generalization error (e.g., error made by the learning machine on data unseen during training) [20]. To classify test cases (regions) into two groups of positive and negative detections in CAD schemes of medical images, the SVM first maps the input feature vector \( x \) into a higher dimensional space using a specific nonlinear mapping function, \( \Phi(x) \). The SVM classification function can be expressed as:

\[ S(x) = w^T \Phi(x) + \beta \]

where \( w, \beta \) are parameters that are determined based on the selected training samples (or feature vectors). Basically, the SVM builds a non-linear hyper-plane of classification. The cases (or representing feature vectors) distributed on two sides of the hyper-plane are classified into two groups. In theory, SVM should achieve higher generalization than ANN in testing new “unseen” testing datasets [15].

In our studies, we selected and modified a publicly available SVM software package (SVM-Light) [21] to build the SVM for classifying between positive and negative regions. To build a practical SVM classifier using a computerized scheme, SVM can often be simplified into the following function:

\[ S(x) = \sum_{i=1}^{n} \alpha_i K(x_i, s_i) + b \]

where \( S_i, i = 1, 2, \ldots, n \) denote \( n \) selected support vectors, where \( n \) is typically much smaller than the total samples (\( N \)) in the training dataset. Based on this function, many different types of SVM classifiers can be built by selecting different classification models or kernels \( K(x_i, s_i) \). In our study we selected a polynomial function based statistic model \( K = (a \ast b + c)^d \) to build the hyper-plane of SVM. Since the large \( d \) value makes the hyper-plane more complex and also reduces the robustness of the SVM classifier, based on our previous experience in SVM optimization [18] we selected \( d = 3 \) in this study.

C) A Bayesian belief network

Bayesian belief network (BBN) is another popular machine learning method based on the well-established statistic learning approach. The primary advantage of a BBN is that it can provide the flexibility for specifying dependence and independence of variables in a reasonably natural way through the network topology, which makes it possible for the structure of a BBN to be examined by human experts to identify the specific relationships between the selected features [22]. The BBN topography represents the joint probability distribution of a problem domain by exploiting the
dependencies between variables and by capturing the uncertain knowledge of a given problem in a natural and efficient way. Unlike the ANN and SVM in which the iterative training or learning processes are required to minimize the errors to the specific targets, the BBN utilizes all training data (samples) to compute the joint probabilities once the node topology of the network is determined. Thus, to compute these joint probabilities each node (feature) must be represented by a relatively small number of discrete states. In our study, the feature values computed from all training samples were divided into five discrete states with equal sample distribution [17]. For example, if the decision node (for classifying between the positive and negative cases) has \( n \) “parent” nodes, a set of \( m \) joint probabilities can be computed to generate a joint-conditional probability table:

\[
P_1(\text{Positive} = \text{yes} | F_1 = \text{state} 1, F_2 = \text{state} 1, \ldots, F_n = \text{state} 1),
\]

\[
P_2(\text{Positive} = \text{yes} | F_1 = \text{state} 2, F_2 = \text{state} 1, \ldots, F_n = \text{state} 1),
\]

\[
\ldots \ldots
\]

\[
P_m(\text{Positive} = \text{yes} | F_1 = \text{state} 5, F_2 = \text{state} 5, \ldots, F_n = \text{state} 5).
\]

Hence, based on the feature states of a test case, the probability value of this region being positive for cancer or the other diseases can be computed from the pre-established probability table.

However, due to the substantial difference between human and computer vision, manually or subjectively defining the dependency and/or independency between the CAD-extracted and computed features is very difficult. To solve this problem and build a BBN with an optimal topology, we used a publicly available BBN optimization tool (BN Power Constructor and Predictor [23]) in this study to build the BBN by automatically searching for an optimal topology of the network and the corresponding features that better reflect the logic structure inherent in this specific classification task using the training datasets (feature vectors).

D) A \( k \)-nearest neighbor searching algorithm

The \( k \)-nearest neighbor (KNN) algorithm is one of the most widely used classifiers when using the local data (instance) based machine learning methods [14]. It has been commonly used in many of the recently developed CAD schemes based on content-based image retrieval (CBIR) approaches [24]. Similar to an adaptive learning approach, KNN algorithm is able to provide an option of selecting a different hypothesis or local approximation to the target function for each unknown test (query) sample. When applying a KNN based classifier, a feature vector based searching algorithm is first applied to search for and sort a set of the “computationally similar” regions of interest (ROIs) in the reference (training) dataset. The similarity is measured by the distance between a queried ROI \( (y_q) \) and a reference ROI \( (x_i) \) in a multi-dimensional \( (n) \) feature space.

\[
d(y_q,x_i) = \sqrt{\sum_{r=1}^{n} (f_r(y_q) - f_r(x_i))^2}
\]

In KNN, the smaller the distance, the higher the degree of computed “similarity” is between any two compared regions. The computed distances between the test region and each of all stored reference regions in the training (reference) dataset are sorted (rank ordered) from the smallest to the largest. The first \( K \)-regions in the list are selected as the \( K \)-most similar (the best matched) reference regions. A distance weight is defined as:

\[
w_i = \frac{1}{d(y_q,x_i)^2}
\]

Thus, a detection score representing the probability of test region being actually positive is computed as:

\[
P_{TP} = \frac{\sum_{i=1}^{N} W_i^{TP}}{\sum_{i=1}^{N} W_i^{TP} + \sum_{j=1}^{M} W_j^{TP}}, \quad N + M = K.
\]
where $N$ is the number of true-positive regions and $M$ is the number of CAD-generated false-positive regions that are selected in the set of the $K$ “most similar” reference regions. Based on our previous optimization experiment [19], we selected $K = 15$ in this study.

### 2.2. Training and testing image datasets

From a full-field digital mammography (FFDM) image database pre-established in our group [25], we randomly selected cases and assembled two independent datasets for training and testing each machine learning classifier. The training dataset includes 250 positive and 250 negative cases. The testing dataset includes 125 positive and 125 negative cases. Each positive case has one verified malignant mass that is visually detectable in both cranio-caudal (CC) and medio-lateral oblique (MLO) views of examination. Each case also has four images (namely CC and MLO view mammmograms of the left and right breast).

A CAD scheme previously developed for the digitized screen-film mammograms (SFM) in our group [26] was converted and re-optimized for the FFDM images [25]. This CAD scheme was applied “as is” to the all images involved in both training and testing datasets. In brief, the CAD scheme used three image processing and feature analysis stages to detect and classify suspicious mass regions depicted on each FFDM image (Figure 2).

![Figure 2. A flow diagram of our CAD scheme to detect suspicious breast mass regions depicted on each single FFDM image.](image)

The first stage of CAD scheme uses a difference-of-Gaussian filtering method to identify all possible suspicious regions. This stage typically detects between 10 and 30 initially regions per image depending on breast tissue density and pattern distribution. The second stage of CAD scheme applies a multilayer topographic region growth algorithm to segment the identified suspicious mass regions. For each growth layer, a growth threshold is adaptively selected based on measurements of the region’s contrast and a set of simple intra- and inter-layer classification rules on size, growth ratio, contrast, and shape factor of the region in question is applied to eliminate a large fraction of the initially identified regions. This stage typically reduces the number of suspicious mass regions to less than 5 per image. For each segmented suspicious mass regions, CAD scheme also computes a set of 36 initial image features. By selecting a set of image features, the third stage of CAD scheme uses a pre-trained machine learning classifier to further classify the segmented regions into the positive and negative groups. In this study, we only used the first two stages of CAD scheme to detect and segment suspicious mass regions depicted on each processed FFDM image. We then retrained and tested four classifiers used in the third stage of the scheme.

From all CAD-segmented suspicious regions depicted on all images in the training dataset, we randomly selected 250 true-positive mass regions and 250 CAD-segmented false-positive (negative) regions. The feature vectors extracted from all of these 250 positive and 250 negative regions were selected to build a new region or feature-based training dataset that was actually used to train each of the four machine learning classifiers. Although the majority of true-positive masses are visible and initially detectable in both CC and MLO views of the same cases, only one region was randomly selected from one case to increase the diversity of training dataset. The 250 CAD-segmented false-positive regions were also randomly selected from 250 negative cases in the training dataset (one for each case). However, for the testing dataset, all CAD initially detected and segmented suspicious regions (including both true-positive and false-positive) in the second stage of the scheme (Figure 2)
were kept and used to test the performance of each of the four classifiers. Hence, after training, each classifier was applied to process and classify all of these segmented suspicious mass regions depicted on the entire testing dataset.

2.3. Classifier testing and performance assessment

For each initially detected and segmented suspicious mass region, our computer scheme extracted and computed 36 morphological and pixel value (intensity) distributed image features as we have previously reported [16]. Using a genetic algorithm, an optimal feature set (vector) including 14 features was selected to build the final classifier used in our previous CAD scheme [25, 26]. In this study, we selected and used two feature vectors \( (x_1, x_2) \). One was the previously selected optimal vector with 14 features and one also included 14 features that were randomly selected from the initial feature pool of 36 features. These two feature vectors were separately used to train and test four machine learning classifiers (ANN, SVM, BBN, and KNN).

After training each classifier using the training dataset with one feature vector \( (x_1, x_2) \) and the classifier training protocol as we have optimized in our previous studies [16-19], the classifier was applied to all detected and segmented suspicious regions in the testing dataset. The classifier-generated detection or classification scores (the likelihood of the test region being positive) of all test regions were first processed and analyzed by a ROC curve fitting program that uses an expanded bi-normal model and the maximum likelihood estimation method (ROCKIT [27]). Since this is a CAD scheme that is allowed to detect multiple suspicious mass regions depicted on each case or image, we used a previously developed method to convert ROC curve into FROC curve [28]. From the FROC curve, we detected and reported the overall detection sensitivity of the CAD scheme at a specific operating threshold that generates a false-positive rate of 0.3 per image. This false-positive rate is similar to that generated by the commercially available CAD schemes used in the clinical practice to date [26]. Since in the clinical application only the suspicious (segmented) regions that have detection scores greater than the operating threshold (Figure 2) are marked on the images and the other regions with detection scores that are smaller than the operating threshold are discarded, we compared four detection sensitivity levels generated by our CAD scheme implemented with four different machine learning classifiers at the same false-positive rate. In addition, since the total number of true-positive mass regions and false-positive regions detected by the second stage of our CAD scheme were fixed for testing all four classifiers, we also compared the areas under the normalized FROC curves (AUC) using the method we have previously reported [29] and their statistically significant difference (p values). Finally, both the case-based and region-based performance assessments (comparisons) were conducted and reported in this study. In case-based assessment, a true-positive mass is counted as detected if either one or two regions depicted on CC and MLO view is detected, while in region-based assessment, two regions associated with the same mass depicted on two views are independently counted and analyzed.

3. Results

From 250 malignant mass regions (representing 125 masses that are visible on two views), the first two stages of our CAD scheme initially detected and segmented 221 true-positive mass regions that are associated with 125 masses. Thus, CAD scheme initially achieved 100% (125/125) case-based sensitivity and 88.4% (221/250) region-based sensitivity. CAD scheme also detected 3182 false-positive regions depicted on 250 cases (including 125 positive and 125 negative cases) or 1000 FFDM images (four images per case). This resulted in an initial CAD-generated false-positive rate of 3.18 per image (3182/1000). After applying each of the four classifiers (ANN, SVM, BBN, and KNN) to classify these 221 true-positive mass regions and 3182 false-positive regions, the classification results were compared and summarized in the following tables. Tables 1 shows the case-based classification performance levels of the four classifiers measured by the areas under the normalized FROC curves (AUC) and their standard deviations using two sets of training and testing feature vectors, respectively. When using the “optimal” feature vector \( (x_1) \), the computed AUC values ranged from 0.924 (BBN) to 0.943 (SVM). Although SVM yielded the highest AUC value, it is not statistically significantly
different from the AUCs yielded by the other three classifiers (i.e., $p = 0.914$ for ANN, $p = 0.091$ for BBN, and $p = 0.101$ for KNN). When using the randomly selected (un-optimized) feature vector ($x_2$), ANN yielded the highest AUC value and it is also not statistically significantly different from the AUCs yielded when using the other classifiers (i.e., $p = 0.069$ for SVM, $p = 0.107$ for BBN, and $p = 0.283$ for KNN). Table 2 shows the corresponding region-based performance levels of the four classifiers. In region-based classification, SVM yielded the highest performance level (AUC value) when using both feature vectors. The statistically significant differences in AUC values between SVM and two classifiers were detected when using feature vector $x_1$ (e.g., $p = 0.038$ between SVM and BBN and $p = 0.001$ between SVM and KNN), while the $p$ value between the AUCs generated by SVM and ANN was not significantly different ($p = 0.078$). There were also no statistically significant differences between SVM and the other classifiers when using feature vector $x_2$ (i.e., $p = 0.296$ for ANN, $p = 0.073$ for BBN, and $p = 0.080$ for KNN). However, when comparing the performance level difference of the four classifiers between using feature vector $x_1$ and $x_2$, we found that both case-based and region-based AUC values (shown in Tables 1 and 2) were significantly different from each other ($p < 0.01$).

**Table 1.** Comparison of case-based classification performance levels measured by the areas under normalized FROC curves (AUC) and standard deviations of the four classifiers trained and tested using two different image feature vectors

<table>
<thead>
<tr>
<th>Classifier</th>
<th>AUC using feature vector $x_1$</th>
<th>AUC using feature vector $x_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANN</td>
<td>0.942 ± 0.010</td>
<td>0.903 ± 0.012</td>
</tr>
<tr>
<td>SVM</td>
<td>0.943 ± 0.010</td>
<td>0.886 ± 0.011</td>
</tr>
<tr>
<td>BBN</td>
<td>0.924 ± 0.011</td>
<td>0.880 ± 0.017</td>
</tr>
<tr>
<td>KNN</td>
<td>0.925 ± 0.013</td>
<td>0.890 ± 0.013</td>
</tr>
</tbody>
</table>

**Table 2.** Comparison of region-based classification performance levels measured by the areas under normalized FROC curves (AUC) and standard deviations of the four classifiers trained and tested using two different image feature vectors

<table>
<thead>
<tr>
<th>Classifier</th>
<th>AUC using feature vector $x_1$</th>
<th>AUC using feature vector $x_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANN</td>
<td>0.882 ± 0.013</td>
<td>0.809 ± 0.015</td>
</tr>
<tr>
<td>SVM</td>
<td>0.891 ± 0.012</td>
<td>0.821 ± 0.013</td>
</tr>
<tr>
<td>BBN</td>
<td>0.871 ± 0.012</td>
<td>0.803 ± 0.016</td>
</tr>
<tr>
<td>KNN</td>
<td>0.866 ± 0.014</td>
<td>0.806 ± 0.015</td>
</tr>
</tbody>
</table>

Figure 3 shows an example of both case-based and region-based FROC-type performance curves of the CAD scheme implemented with a SVM based classifier trained and tested using two feature vectors. From the computed FROC curves, Tables 3 and 4 summarize and compare the case-based and region-based differences of classification performances measured by detection sensitivity at the same operating threshold (with false-positive rate of 0.3 per image) among four classifiers. For each reported sensitivity level, the lower and upper bounds of the asymmetric point-wise 95% confidence interval (CI) are also presented in each table. Although the detected sensitivity levels are related to the AUC values reported in Tables 1 and 2, these two performance indices are not 100% correlated. For example, SVM always has the higher performance level than BBN when using AUC as the performance assessment index. However, by using sensitivity level at the operating threshold as the performance index, BBN yielded the higher sensitivity than SVM when using feature vector $x_2$. In summary, no matter which performance index was used, the results indicated when using the same feature vector extracted from the same training and testing datasets, these four machine learning classifiers tended to yield very comparable performance level (in particular for the case-based performance assessment). However, when using different feature vectors, the performance level of each of these four classifiers was significantly changed ($p < 0.01$).
Computer-aided Detection: The Impact of Machine Learning Classifier and Image Feature Selection on Scheme Performance
Xingwei Wang, Dror Lederman, Jun Tan, Bin Zheng

Figure 3. Comparison of case-based (a) and region-based (b) FROC-type performance curves when training and testing a SVM classifier using two feature sets (vectors).

Table 3. Comparison of case-based classification performance levels measured by the detection sensitivity at the false-positive rate of 0.3 per image and 95% confidence interval (CI) of the four classifiers trained and tested using two different image feature vectors

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Sensitivity using feature vector $x_1$</th>
<th>Sensitivity using feature vector $x_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANN</td>
<td>82.6% [76.4%, 87.7%]</td>
<td>69.4% [62.3%, 75.9%]</td>
</tr>
<tr>
<td>SVM</td>
<td>82.8% [76.7%, 87.9%]</td>
<td>60.2% [52.6%, 67.9%]</td>
</tr>
<tr>
<td>BBN</td>
<td>76.3% [69.7%, 82.2%]</td>
<td>67.7% [60.6%, 74.3%]</td>
</tr>
<tr>
<td>KNN</td>
<td>78.9% [72.3%, 84.3%]</td>
<td>64.8% [57.4%, 71.6%]</td>
</tr>
</tbody>
</table>

Table 4. Comparison of region-based classification performance levels measured by the detection sensitivity at the false-positive rate of 0.3 per image and 95% confidence interval (CI) of the four classifiers trained and tested using two different image feature vectors

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Sensitivity using feature vector $x_1$</th>
<th>Sensitivity using feature vector $x_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANN</td>
<td>60.9% [56.1%, 65.3%]</td>
<td>44.7% [39.6%, 49.8%]</td>
</tr>
<tr>
<td>SVM</td>
<td>62.7% [58.1%, 67.1%]</td>
<td>41.5% [36.4%, 46.7%]</td>
</tr>
<tr>
<td>BBN</td>
<td>55.9% [50.9%, 60.7%]</td>
<td>43.7% [38.7%, 48.7%]</td>
</tr>
<tr>
<td>KNN</td>
<td>58.4% [53.6%, 62.9%]</td>
<td>44.5% [39.4%, 49.5%]</td>
</tr>
</tbody>
</table>

4. Discussion

Many different machine learning methods have been investigated and applied in developing multi-feature based classifiers that are implemented in CAD schemes of biomedical images. In this study we investigated and compared four popular machine learning methods or classifiers (ANN, SVM, BBN, and KNN) used in CAD development. Among them, three are global data based machine learning classifiers and one is a local instance based classifier. Each machine learning method or classifier has its unique characteristics including both advantages and limitations [14]. The ANN, SVM, and BBN are trained to build a single “global” optimization target function in attempt to cover the entire data domain for the future “never seen” testing cases using different learning approaches. These classifiers are capable of learning and estimating the global based target function using the relatively noisy or incomplete training data. However, over-fitting the training data is an important issue or risk in some of these classifiers, which produces poorer classification results when applying to the new testing dataset. Previous studies have reported that the global data based classifiers using the statistical learning models or concepts (i.e., SVM and BBN) should achieve the higher robustness level than the purely data-driven learning methods (e.g., ANN) [15, 17]. As a local instance-based learning method, KNN adaptively builds different local approximations to the target function depending on the similar
references (neighbors) of the new testing (queried) case. Although this has great advantage when the target function is very complex (in particular for many applications in biomedical imaging field) because KNN can still be described by a collection of less complex local approximations, KNN is likely to be more sensitive to the data noise and diversity (distribution) of the reference database. In addition, the local instance-based learning methods require more computational power (spend more computational time) in testing each new case than the classifiers using a global targeted decision function. Hence, selecting which classifier in developing CAD schemes of biomedical images to achieve the “best” detection or classification performance is considered an interesting and difficult issue faced by the CAD developers.

Based on a common training dataset and a common “never seen” testing dataset, we in this study investigated and compared the performance levels of a CAD scheme implemented with four popular machine learning classifiers. We found that once a feature vector was determined and the appropriate training was conducted, all four classifiers in general tended to generate very comparable classification performance (without statistically significant difference). The results indicated that although the classifiers based on statistical learning models were in general able to achieve higher robust level, taking the necessary measures to minimize the potential over-fitting bias in training ANN (i.e., limiting the training iteration and the size of feature vector, as well as maintaining the larger ratio between the training momentum and learning rate), the optimized ANN can also produce the comparable and robust classification performance when applied to the “never seen” testing datasets. Our results are different from the previous study in which SVM achieved significantly higher performance than ANN using a leave-one-out cross validation method [15]. Such difference needs to be further investigated.

In this study, we also selected two sets of image features (or two feature vectors) extracted from each suspicious mass region and compared performance of four classifiers when using these two feature vectors. The experimental results showed when using these two feature vectors extracted from the same training and testing dataset, the performance levels were significantly changed for each of all four classifiers. This consistent trend suggests that it is the selection of feature vector rather than the specific machine learning classifier that determines the overall CAD scheme performance level. Therefore, in CAD development one should select and use a machine learning method or classifier that one has the “best” experience (knowledge) to optimally train and use. Meanwhile, to improve CAD performance, more research efforts should be focused on building the optimal training database, increasing accuracy in lesion segmentation (for computing reliable features), and selecting effective or optimal feature vectors.

5. Acknowledgement

This work is supported in part by Grant CA77850 to the University of Pittsburgh from National Cancer Institute, National Institutes of Health, USA.

6. References

Computer-aided Detection: The Impact of Machine Learning Classifier and Image Feature Selection on Scheme Performance
Xingwei Wang, Dror Lederman, Jun Tan, Bin Zheng