# Reliability analysis framework for computer-assisted medical decision systems

Piotr A. Habas[a)] and Jacek M. Zurada
*Computational Intelligence Laboratory, Department of Electrical and Computer Engineering, University of Louisville, Louisville, Kentucky 40292*

Adel S. Elmaghraby
*Department of Computer Engineering and Computer Science, University of Louisville, Louisville, Kentucky 40292*

Georgia D. Tourassi
*Digital Advanced Imaging Laboratories, Department of Radiology, Duke University Medical Center, Durham, North Carolina 27705*

(Received 31 May 2006; revised 20 December 2006; accepted for publication 20 December 2006; published 30 January 2007)

We present a technique that enhances computer-assisted decision (CAD) systems with the ability to assess the reliability of each individual decision they make. Reliability assessment is achieved by measuring the accuracy of a CAD system with known cases similar to the one in question. The proposed technique analyzes the feature space neighborhood of the query case to dynamically select an input-dependent set of known cases relevant to the query. This set is used to assess the local (query-specific) accuracy of the CAD system. The estimated local accuracy is utilized as a reliability measure of the CAD response to the query case. The underlying hypothesis of the study is that CAD decisions with higher reliability are more accurate. The above hypothesis was tested using a mammographic database of 1337 regions of interest (ROIs) with biopsy-proven ground truth (681 with masses, 656 with normal parenchyma). Three types of decision models, (*i*) a back-propagation neural network (BPNN), (*ii*) a generalized regression neural network (GRNN), and (*iii*) a support vector machine (SVM), were developed to detect masses based on eight morphological features automatically extracted from each ROI. The performance of all decision models was evaluated using the Receiver Operating Characteristic (ROC) analysis. The study showed that the proposed reliability measure is a strong predictor of the CAD system's case-specific accuracy. Specifically, the ROC area index for CAD predictions with high reliability was significantly better than for those with low reliability values. This result was consistent across all decision models investigated in the study. The proposed case-specific reliability analysis technique could be used to alert the CAD user when an opinion that is unlikely to be reliable is offered. The technique can be easily deployed in the clinical environment because it is applicable with a wide range of classifiers regardless of their structure and it requires neither additional training nor building multiple decision models to assess the case-specific CAD accuracy. 
[DOI: 10.1118/1.2432409]

Key words: mammography, computer-assisted detection (CAD), reliability analysis, artificial intelligence, receiver operating characteristic (ROC)

## I. INTRODUCTION

Although clinical integration of computer-assisted decision (CAD) systems is steadily increasing, they are still often perceived as "black boxes." Typically, CAD systems are developed to operate with fixed decision thresholds that are globally optimized to be overall accurate on the general population of prospective patients. Consequently, CAD systems provide users with only binary decisions that are basically thresholded versions of their continuous outputs.

The "black box" nature of CAD systems has been questioned before,[1] indicating limitations in clinical acceptance and ultimate effectiveness of such systems. As contemporary CAD tools are expected to be more interactive, several CAD investigators have been working towards these goals. For example, CAD systems are developed to provide visual justification of their opinions by retrieving and displaying reference cases that are similar to the query. However, the CAD user is still left unguided as to how to incorporate the additional visual information into the clinical decision making process. Enhancing CAD systems with the ability to provide a reliability measure for every decision they make would be an important step towards facilitating better human-computer interaction. We envision that a case-specific reliability measure will help physicians use CAD technology more effectively and robustly by identifying CAD opinions that are either highly reliable or questionable. The reliability measure could also be exploited as the basis of a mechanism to inform the physician as to whether the CAD system is expected to deliver a high level of diagnostic accuracy for a specific case.

Assessing the case-specific reliability of a CAD system, however, is a challenging task as it is affected by the uncertainty inherently present in the decision making process. Several studies have addressed the issue of uncertainty associated with the output of decision models. For the most part investigators have looked at two sources of uncertainty. Data-driven uncertainty is due to the variability present in the data and/or incomplete nature of the training set. Nix[2] and Qazaz[3] proposed extensions of traditional artificial neural network (ANN) architectures to compute the variance associated with noise inherently present in measurement data. A similar constant-variance approach presented by Edwards[4] was further extended by Papadopoulos[5] to model the noise variance as a function of the input vector. In contrast, model-driven uncertainty is most often due to the imperfection of the decision model (misspecification of its parameters) and/or the stochastic nature of the training/development process. Bishop[6] used a neural network example to demonstrate that a decision model trained on a given dataset develops a better representation of the data in regions of high data density. Furthermore, training algorithms used to develop neural networks do not guarantee that the final weight values generate the global minimum of the error function. Therefore, the model's output should be viewed as a random variable with variance equal to the model uncertainty variance that can be estimated using the linearization method[7,8] or the bootstrap technique.[9] Jiang[10] investigated the impact of the stochastic nature of the training process that gives ANNs the ability to avoid entrapment at local minima. The study demonstrated that even for the same set of training patterns and the same ANN architecture, one can develop various decision models achieving the same overall (global) performance but very different output values (and hence accuracy) for the same individual input pattern. In clinical applications, however, this type of variability is generally not recognized because a trained network is "frozen" and effectively applied as a deterministic decision model.

In this study, we propose a technique that provides CAD systems with the ability to assess the expected reliability of their individual outputs based on the local accuracy of the CAD system. In contrast to previously presented methods, the proposed technique can be easily employed in the clinical environment because it requires neither additional training to learn variability from the data nor building multiple decision models. Furthermore, the proposed technique is applicable with a wide range of classifiers including deterministic and nondeterministic neural networks, support vector machines, and potentially others.

This paper is organized as follows. Section II presents the theoretical framework of the proposed reliability analysis technique. It also provides a description of the experimental design and practical implementation of the technique with a variety of decision models developed for region-based mass detection on a mammographic database. Experimental results are presented in Sec. III along with a more detailed study of the impact of key implementation parameters. Summary and discussion of the study findings follow.

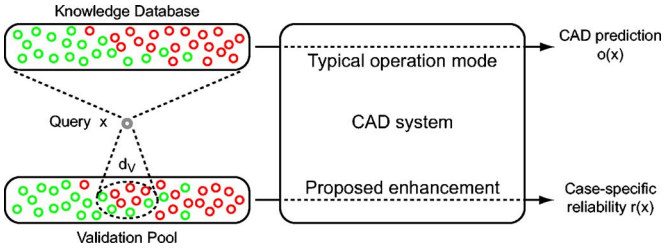



FIG. 1. Schematic representation of a typical CAD system enhanced with the proposed reliability analysis capability. Dark circles represent actually positive cases while light circles represent actually negative (normal) cases. The query case with unknown ground truth is shown in gray.

## II. METHODOLOGY

### A. Reliability assessment technique

The study is based on the premise that the query-specific reliability of a CAD system can be described in terms of the local accuracy of the system for cases highly similar to the query. The estimated local accuracy is utilized as a reliability measure of the CAD response to the query case. CAD predictions with higher associated reliability should be more accurate.

Traditional CAD systems are usually developed in the train-test mode. In the training mode, the CAD system extracts generalizable knowledge from a set of training cases. Then, in the testing mode, the CAD system is asked to make predictions $o(\mathbf{x}_i)$ for a number of previously unseen test cases $\{\mathbf{x}_i | i=1,2,\ldots\}$. Dependency between the training and testing sets is regulated by sophisticated data sampling plans such as leave-one-out, cross validation, or bootstrap.[11,12]

Following the train-test mode, we propose an additional step in the CAD development process. Specifically, for each prediction $o(\mathbf{x})$ made by the decision model, its case-specific reliability $r(\mathbf{x})$ is estimated as well. This additional step emphasizes individualized analysis of the case-dependent accuracy of the CAD system, and not only the assessment of its global accuracy as it is currently done.

Given a query case $\mathbf{x}$ and the corresponding CAD prediction $o(\mathbf{x})$, the proposed reliability estimate $r(\mathbf{x})$ can be calculated by assessing the accuracy of the CAD system for a set of cases $\{\mathbf{y}_i | i=1,2,\ldots,N\}$ highly similar to the query (i.e., the relevant set). Such a collection can be generated either by resampling from the training set or from an independent set reserved specifically for reliability analysis. Although the first approach maximizes knowledge extraction from the available data, it could potentially generate overoptimistic estimates due to the memorization effect present during the training mode. Reserving an independent set specifically for reliability analysis could reduce the impact of the potential memorization effect.

Taking into account usual data limitations, we developed a strategy that operates on two disjoint data subsets—a knowledge database (KD) and a validation pool (VP) as presented in Fig. 1. The knowledge database is used to develop and optimize the decision model. The above step is the typi-

cal implementation of a CAD system. When the trained CAD system is queried on a new unknown case $\mathbf{x}$, a CAD $o(\mathbf{x})$ is calculated.

Simultaneously, the validation set is used for assessing the local validation error for the CAD output value. Specifically, the validation pool is analyzed in search for cases similar to the query case with respect to a model-dependent distance measure $D$ defined in the feature space. If there are no cases in VP lying not farther than a predetermined threshold value $d_V$ from the query case, the query case is excluded from the reliability analysis and reported as a no-neighbor (NN) case. Otherwise, all similar cases $\{\mathbf{y}_i | i=1,2,\ldots,N\}$ are extracted to form a local case-specific validation set. The performance of the CAD system on the validation set is measured in terms of the mean square error (MSE) between target values $t(\mathbf{y}_i)$ and predicted values $o(\mathbf{y}_i)$ across all cases $\{\mathbf{y}_i | i=1,2,\ldots,N\}$ from the local validation set:

$$r(\mathbf{x}) = 1 - \frac{1}{N}\sum_{i=1}^{N}[t(\mathbf{y}_i) - o(\mathbf{y}_i)]^2 \quad \text{for all } \{\mathbf{y}_i | D(\mathbf{x},\mathbf{y}_i \le d_V)\}. \tag{1}$$

Following standard conventions in CAD development, the target values $t(\mathbf{y}_i)$ are assumed to be binary (0 if case $\mathbf{y}_i$ is negative, 1 if case $\mathbf{y}_i$ is positive). CAD predictions $o(\mathbf{y}_i)$ are assumed to be continuous in the range $[0,1]$ for all investigated decision models. As a result, reliability $r(\mathbf{x})$ can take continuous values ranging from 0 (in case of extremely high local validation error) to 1 (in case of perfect predictions for all cases $\mathbf{y}_i$ from the relevant set).

The reliability estimate $r(\mathbf{x})$ may be reported to the CAD user together with the CAD prediction $o(\mathbf{x})$ as a tuple $(o(\mathbf{x}), r(\mathbf{x}))$. Essentially, the two-level analysis described above results in a CAD system that not only provides its user with a prediction $o(\mathbf{x})$ for a given query $\mathbf{x}$, but also with additional information regarding the expected reliability $r(\mathbf{x})$ of this particular prediction. Theoretically, predictions with lower validation error should be more reliable as low validation error indicates that the CAD system is more accurate with cases similar to the query. Since the proposed technique does not make any *a priori* assumptions regarding the nature (e.g., deterministic, nondeterministic) or architecture (e.g., structure, kernel, etc.) of the underlying indecision model, it is universal and easily applicable with a variety of them.

## B. Dataset

The proposed reliability analysis framework was tested on the Directional Neighborhood Analysis (DNA) dataset—a private collection generated in our laboratory based on custom image analysis of screening mammograms selected from the Digital Database of Screening Mammography (DDSM).[13] The dataset is the result of our research attempts[14] to differentiate true masses from normal breast tissue by characterizing regional directional properties of the breast parenchyma.

The DNA dataset includes 1337 regions of interest ($512\times512$ pixels) extracted from DDSM mammograms digitized using the LUMISYS scanner. Out of them, 681 ROIs depict biopsy-proven masses (340 malignant, 341 benign) while the remaining 656 ROIs represent normal parenchyma. Although some ROIs were extracted from different mammographic views of the same patient, the 1337 ROIs used in the study are treated as independent cases. Based on the DNA technique, each ROI is described in terms of eight continuous-valued morphological features targeting the directional propagation of potential disturbances in breast tissue.[14]

## C. Decision models

Three different decision models were developed to investigate the feasibility of the proposed reliability analysis framework regardless of the nature and structure of the underlying decision model.

### 1. Back propagation neural network (BPNN)

Feed-forward back-propagation neural networks (BPNNs) are very popular with CAD systems as they have been shown to be universal classifiers and function approximators.[15] They have also demonstrated the ability to solve problems that are hard to describe in terms of mathematical formulas or analytical modeling.[16] In our experiments we used a BPNN network in a form of a fully connected three-layer perceptron. Three hidden neurons and the single output neuron had the unipolar sigmoidal activation function

$$f(net) = \frac{1}{1+\exp(-net)}. \tag{2}$$

Before the training of a new network, all layer weights and biases were initialized according to the Nguyen-Widrow algorithm[17] that chooses weight values in order to distribute the active region of each neuron in the layer approximately evenly across the layer's input space. Then, the network was trained for 100 cycles with the Levenberg-Marquardt back-propagation algorithm[18,19] at the learning rate of 0.1, regardless of the actually achieved performance.

### 2. Generalized regression neural network (GRNN)

A generalized regression neural network (GRNN)[20] is a three layer feed-forward network that can approximate a continuous function with an arbitrary accuracy.[21] The network stores training vectors $\mathbf{u}_i$ in a hidden layer of radial basis function (RBF) nodes[22] (one per pattern) computing

$$h_i = \exp\left(\frac{-\|\mathbf{x}-\mathbf{u}_i\|^2}{2\sigma^2}\right). \tag{3}$$

The output layer consists of a single neuron calculating a normalized weighted sum of outputs $h_i$ from all hidden neurons. The spread of RBF nodes $\sigma$ affects the smoothness of the function approximation—to fit data more closely, smaller values of $\sigma$ should be used.

### *3. Support vector machine (SVM)*

A support vector machine (SVM) is a learning procedure stemming from the statistical learning theory[23] and the principle of structural risk minimization. In contrast to BPNNs, it does not try to decrease the mean square error on the training set, but rather aims at minimizing the expected generalization error on new data. As a result, SVMs have been demonstrated to outperform other models in several real-world applications, including biomedical CAD systems.[24] For our experiments, an SVM with a Gaussian radial basis function kernel

$$K(\mathbf{x},\mathbf{y}) = \exp\left(\frac{-\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}\right) \quad (4)$$

was used. The kernel function $K(\mathbf{x},\mathbf{y})$ performs a nonlinear mapping between an input vector $\mathbf{x}$ and a support vector $\mathbf{y}$, before the final linear classification in the mapped space.

## D. Performance evaluation

The performance of the decision models and the impact of the proposed reliability assessment technique were evaluated using receiver operating characteristic (ROC) analysis,[25,26] typically used in CAD applications. Classification ability can be expressed with a set of indices calculated from ROC curves, from which the most commonly used is the area under the curve $A_z$.[27] The values of $A_z$ can range from 0.5 for chance to 1.0 for a perfectly operating classifier. ROC analysis was performed using ROCKIT software[28,29] version 1.0.1 provided by Charles Metz from Kurt Rossmann Laboratories for Radiologic Image Research at the University of Chicago.

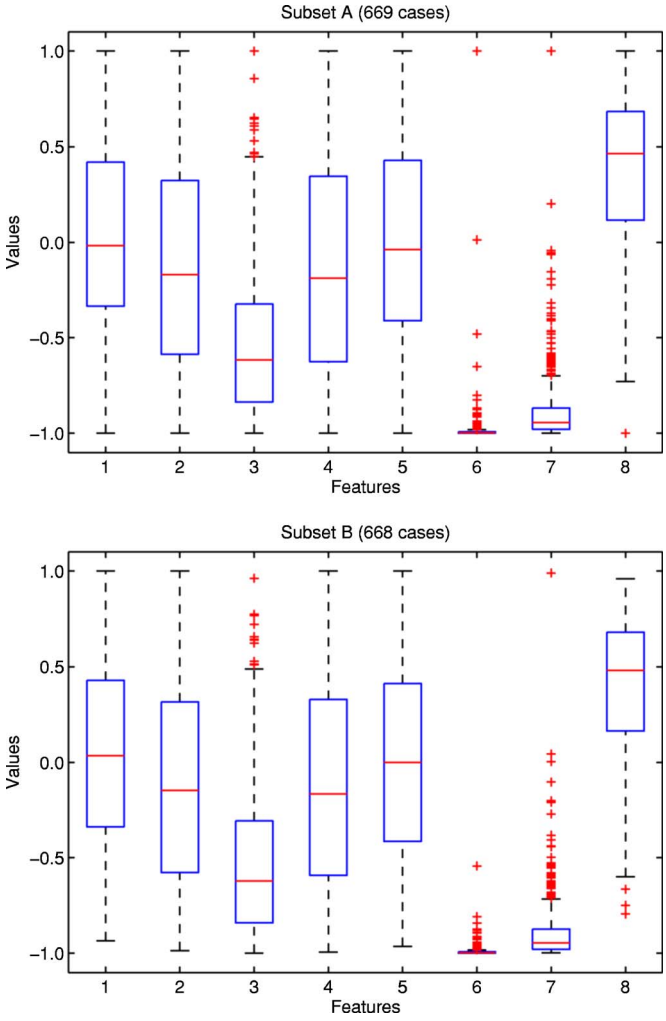



FIG. 2. Feature-based statistics of the cross-validation sets.

# III. RESULTS

## A. Data handling scheme

Data normalization is performed independently for each of the eight input features using standard line, transformation

$$x_n = 2\frac{x_o - x_{\min}}{x_{\max} - x_{\min}} - 1, \quad (5)$$

where $x_n$ denotes the normalized feature value, $x_o$ denotes the original feature value, and $x_{\min}$ and $x_{\max}$ denote the minimum and the maximum values of the feature among all data samples. As a result, the normalized values of each feature fall in the range $[-1,1]$.

After normalization, the available dataset is divided into two subsets A and B to serve as the knowledge database and the test set. To ensure a balanced representation of cases between the subsets, the dataset split is performed in a controlled manner based on feature values. The normalized data patterns are contained inside an eight-dimensional hypercube spreading between −1 and 1 along each axis. The origin $O=(0,0,\ldots,0)$ is the center of the hypercube and a point of reference for the normalized data. Having ranked all 1337 data samples according to their distance from the origin $O$, odd-numbered cases (total of 669) are gathered in a subset A, while all even-numbered cases (total of 668) are combined into a subset B. The resulting sets are balanced not only in terms of feature values (as shown in Fig. 2) but also in the prevalence of positive cases (353/669 for subset A, 328/668 for subset B).

In principle, the proposed reliability analysis technique requires three different datasets: (*i*) a knowledge database for training of the CAD system, (*ii*) a test set for testing of the system, and (*iii*) a validation pool for performing the reliability assessment of the test CAD outputs. To maximize the impact of the available data without compromising the statistical significance of the results, we applied the following data handling scheme.

Each full experiment consists of two cross-validation runs. First, subset A serves as the knowledge database and subset B serves as the test set. Then, the roles of each set are reversed, resulting in subset B serving as the knowledge database and subset A being the test set. In each cross-validation run, the subset serving as the knowledge database is used to train and optimize a decision model with a leave-one-out sampling scheme. When the optimal parameters of the decision model are determined (e.g., the number of hidden nodes and training iterations for the BPNN), the model is

trained on all training cases. At this stage it is considered ready for further testing and reliability analysis. To maximize the impact of the available data, the test set is used for reliability analysis following a pseudo leave-one-out sampling scheme. Specifically, each case $\mathbf{x}$ from the subset serving for testing is excluded once to serve as a test case and obtain a CAD prediction $o(\mathbf{x})$. Then, the remaining cases from the test set serve as the validation pool. Relevant cases $\mathbf{y}_i$ are drawn from the validation pool for estimation of the reliability measure $r(\mathbf{x})$. If no cases similar to the test case $\mathbf{x}$ are found, the test case $\mathbf{x}$ is assigned a no-neighbor (NN) label as the case-specific prediction reliability $r(\mathbf{x})$ cannot be assessed. Since the above experiments employ two disjoint sets for CAD training and reliability analysis, this data handling scheme is called "training-independent" reliability assessment.

Additionally, the above experiments are repeated using the training set (knowledge database) as the validation pool. This data handling scheme is called "training-dependent" reliability assessment. The motivation for performing separate experiments using an independent validation pool and using the training cases as the validation pool was formerly emphasized—to assess if and how the possible memorization of training cases affects the conclusions of the study.

For both data handling schemes, results from the twofold cross-validation runs are pooled together, resulting in 1337 prediction-reliability pairs $(o(\mathbf{x}),r(\mathbf{x}))$ used for performance evaluation.

### B. Reliability analysis

The following section presents the results of experiments performed to test the study hypothesis across three different decision models (BPNN, GRNN, and SVM) and two data handling schemes (i.e., training-independent and training-dependent reliability assessment). To verify the presence of a relationship between the proposed reliability metric and the $A_z$ performance of a CAD system, all predictions $o(\mathbf{x})$ are stratified according to their associated reliability values $r(\mathbf{x})$. Then, the values of $A_z$ performance are plotted as a function of the reliability range of each strata.

The width of the stratification bins is selected empirically to provide detailed analysis while ensuring that the bins contain enough cases for statistically meaningful estimation of the ROC area index. Consequently, for the reliability values between 1.00 and 0.75, the bins have the constant width of 0.05 regardless of the number of cases they contain. All cases with assigned reliability values below 0.75 are grouped together. The NN bin contains predictions for cases with no similar examples (neighbors) found in the validation pool. For consistency of the plots the same bin boundaries are used for all experiments.

The baseline $A_z$ value shown on the plots is the $A_z$ performance of the CAD system calculated for all cases before the reliability assessment technique is applied. The presence or absence of statistically significant differences in $A_z$ performance between different reliability-based groups is verified by the $t$ test at 95% confidence level.

#### 1. Effect of the neighborhood radius

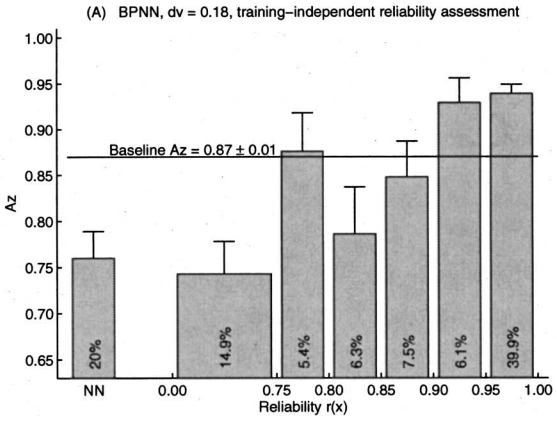
Before the proposed reliability assessment technique can be applied, the size of the neighborhood radius $d_V$ needs to be selected. We decided to use the same value of $d_V$ for all cases, though an input-dependent approach could also be applied. As the radius of the validation neighborhood shapes the number and the span of the input-dependent relevant set, it becomes a crucial parameter in the calculation of reliability values $r(\mathbf{x})$. Smaller values of $d_V$ result in a lower number of validation cases and may deteriorate the precision of reliability estimation. With large values of $d_V$, on the other hand, reliability $r(\mathbf{x})$ is calculated over cases not having much in common with the query case $\mathbf{x}$.

The following experiments were performed using the Euclidean distance as the preferred distance metric to assess case similarity. Although a number of metrics can be used to determine similarity between cases in the feature space, the Euclidean distance is a natural choice for two out of three applied decision models. Both the GRNN and the SVM use it as an integral part of their RBF kernels, either during the model's development (SVM) or testing on an unknown case (GRNN). On the other hand, due to its architecture, the BPNN does not perform direct computation of distances between cases, but for consistency the Euclidean distance was employed for all experiments.

Figure 3 demonstrates how the ROC area index varies as a function of the binned reliability index for three representative neighborhood distances and, for simplicity, focuses on the BPNN-based CAD system. Figure 3(a) presents results corresponding to a conservative neighborhood radius of $d_V=0.18$ while Figs. 3(b) and 3(c) show how the results change for progressively higher values of the neighborhood radius
$d_V=0.25$ and $d_V=0.35$, respectively. Each graph also highlights the baseline ROC area of the CAD system for easier comparison with the reliability-based binned performance. Note that the graphs also include percentages of cases that are stratified into each bin. In addition, pairwise statistical analysis of the ROC performance is performed for each pair of bins. The results are shown in the tables associated with each graph.

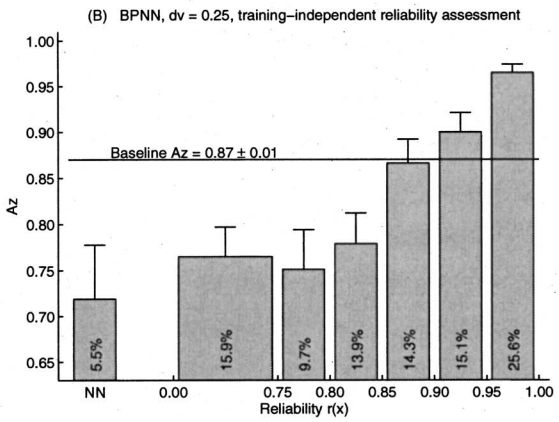
As demonstrated in Fig. 3, the neighborhood radius $d_V$ is an important factor in reliability assessment as it affects the range and span of local validation sets. Low values of $d_V$ result in imprecise reliability estimates that tend to be extreme [Fig. 3(a)]. No clear trend can be observed in relationship between the reliability and the $A_z$ performance. The extremity of reliability estimates is reflected by the fact that the majority of cases (54.8%) is assigned to either the highest (39.9% of cases) or the lowest (14.9% of cases) reliability bin. Each of the intermediate bins contains no more than 7.5% of cases. Low values of $d_V$ also result in an unacceptably high fraction of no-neighbor cases, which make them impractical in clinical applications.

High $d_V$ values [Fig. 3(c)], on the other hand, result in excessively large validation sets containing cases that are not really similar to the query case. A pessimistic trend in reli-



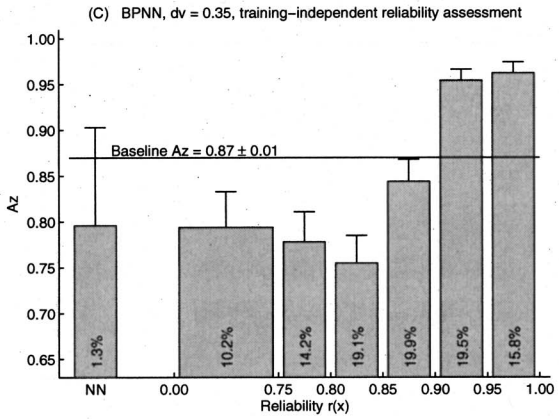

$p$-values for differences in $A_z$ indices between $r(\mathbf{x})$-based groups

| $r(\mathbf{x})$ | [0.00, 0.75) | [0.75, 0.80) | [0.80, 0.85) | [0.85, 0.90) | [0.90, 0.95) |
|---|---|---|---|---|---|
| [0.95, 1.00] | **≪0.01** | 0.17 | **<0.01** | **0.04** | 0.84 |
| [0.90, 0.95) | **≪0.01** | 0.29 | **0.01** | 0.10 | |
| [0.85, 0.90) | **0.03** | 0.62 | 0.28 | | |
| [0.80, 0.85) | 0.49 | 0.13 | | | |
| [0.75, 0.80) | **0.01** | | | | |




$p$-values for differences in $A_z$ indices between $r(\mathbf{x})$-based groups

| $r(\mathbf{x})$ | [0.00, 0.75) | [0.75, 0.80) | [0.80, 0.85) | [0.85, 0.90) | [0.90, 0.95) |
|---|---|---|---|---|---|
| [0.95, 1.00] | **≪0.01** | **≪0.01** | **≪0.01** | **≪0.01** | **<0.01** |
| [0.90, 0.95) | **≪0.01** | **<0.01** | **<0.01** | 0.31 | |
| [0.85, 0.90) | **0.01** | **0.02** | **0.04** | | |
| [0.80, 0.85) | 0.76 | 0.61 | | | |
| [0.75, 0.80) | 0.80 | | | | |




$p$-values for differences in $A_z$ indices between $r(\mathbf{x})$-based groups

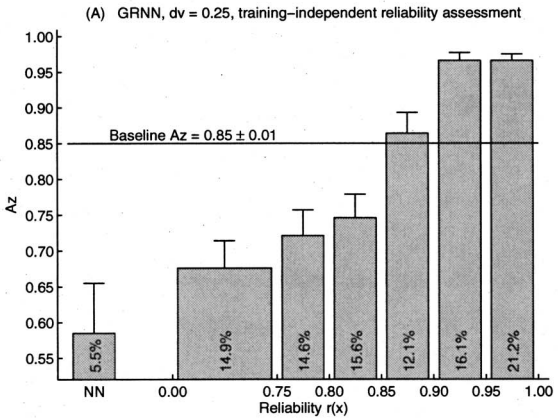
| $r(\mathbf{x})$ | [0.00, 0.75) | [0.75, 0.80) | [0.80, 0.85) | [0.85, 0.90) | [0.90, 0.95) |
|---|---|---|---|---|---|
| [0.95, 1.00] | **≪0.01** | **≪0.01** | **≪0.01** | **≪0.01** | 0.61 |
| [0.90, 0.95) | **≪0.01** | **≪0.01** | **≪0.01** | **≪0.01** | |
| [0.85, 0.90) | 0.27 | 0.11 | **0.02** | | |
| [0.80, 0.85) | 0.43 | 0.61 | | | |
| [0.75, 0.80) | 0.75 | | | | |

FIG. 3. Stratification of CAD predictions according to their reliability measure for the BPNN decision model and three representative values of the neighborhood radius: (a) $d_V$=0.18, (b) $d_V$=0.25, and (c) $d_V$=0.35. The corresponding tables list $p$ values of pairwise differences in $A_z$ performance among the various reliability strata; $p$ values in bold indicate statistically significant differences at 95% confidence level. **≪0.01** indicates a $p$ value smaller than 0.001.

ability estimation can be observed that results in assigning more cases to lower reliability bins. The highest reliability bin contains only 15.8% of cases (compared to 39.9% for $d_V$=0.18) and all other bins contain no less than 10% of cases. The fraction of no-neighbor cases is pleasingly low (about 1%).

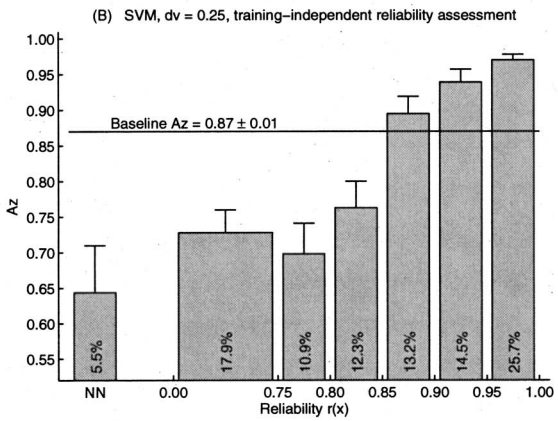
For the analyzed dataset, the neighborhood radius of $d_V$=0.25 proved to be the most clinically useful [Fig. 3(b)]. The fraction of no-neighbor cases without assigned reliability values is reasonably small (about 5%). The $A_z$ performance of the system for this group is lower than for cases with neighbors in the validation pool, which suggests that the no-neighbor cases are truly not similar to what the CAD system was trained with. For cases with neighbors, on the other hand, there exists a clear dependency between assigned reliability values and the $A_z$ performance of the CAD system. It is manifested by the presence of many statistically significant differences in $A_z$ values between groups of CAD predictions with different estimated reliability.

Similar dependency between reliability values and $A_z$ performance was observed for the remaining decision models—the generalized regression neural network (GRNN) and the




*p*-values for differences in $A_z$ indices between $r(\mathbf{x})$-based groups

| $r(\mathbf{x})$ | [0.00, 0.75) | [0.75, 0.80) | [0.80, 0.85) | [0.85, 0.90) | [0.90, 0.95) |
|---|---|---|---|---|---|
| [0.95, 1.00] | **≪0.01** | **≪0.01** | **≪0.01** | **<0.01** | 0.96 |
| [0.90, 0.95) | **≪0.01** | **≪0.01** | **≪0.01** | **<0.01** | |
| [0.85, 0.90) | **≪0.01** | **<0.01** | **<0.01** | | |
| [0.80, 0.85) | 0.16 | 0.61 | | | |
| [0.75, 0.80) | 0.38 | | | | |




*p*-values for differences in $A_z$ indices between $r(\mathbf{x})$-based groups

| $r(\mathbf{x})$ | [0.00, 0.75) | [0.75, 0.80) | [0.80, 0.85) | [0.85, 0.90) | [0.90, 0.95) |
|---|---|---|---|---|---|
| [0.95, 1.00] | **≪0.01** | **≪0.01** | **≪0.01** | **<0.01** | 0.11 |
| [0.90, 0.95) | **≪0.01** | **≪0.01** | **≪0.01** | 0.15 | |
| [0.85, 0.90) | **≪0.01** | **≪0.01** | **<0.01** | | |
| [0.80, 0.85) | 0.48 | 0.26 | | | |
| [0.75, 0.80) | 0.58 | | | | |

FIG. 4. Stratification of CAD predictions according to their reliability measure for (a) the GRNN and (b) the SVM decision models and the medium value of the neighborhood radius $d_V$=0.25. The corresponding tables list $p$ values of pairwise differences in $A_z$ performance among the various reliability strata; $p$ values in bold indicate statistically significant differences at 95% confidence level. **≪0.01** indicates a $p$ value smaller than 0.001.

support vector machine (SVM). Figure 4 clearly shows that as the reliability index becomes progressively lower, the performance of the CAD models decreases as well.

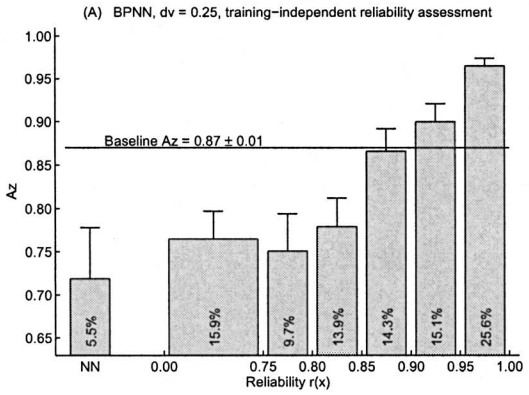
### 2. Effect of training-independent versus training-dependent reliability assessment

The purpose of this experiment is to investigate the impact of the data handling scheme and possible memorization of training cases on the quality of reliability estimation when cases relevant to the query are drawn from the training set (training-dependent reliability assessment) rather than from an independent validation pool (training-independent reliability assessment). Figure 5 presents the results of reliability-based stratification of predictions from the BPNN for the same medium value of the neighborhood radius ($d_V$=0.25) but different sources of relevant cases.

As shown in Fig. 5, the training-independent reliability assessment results in a stronger trend between estimated prediction reliability and the CAD accuracy. The differences in $A_z$ performance between consecutive reliability-based strata are larger and most of them turn out to be statistically significant. In case of training-dependent reliability assessment, there are no statistically significant differences in the $A_z$ index among four groups of cases with assigned reliability values between 0.75 and 0.95. Therefore, the resolution of the reliability-based stratification is effectively lowered to only three levels—low $[r(\mathbf{x})<0.75]$, medium $[0.75\leq r(\mathbf{x})<0.95]$, and high $[r(\mathbf{x})\geq 0.95]$.

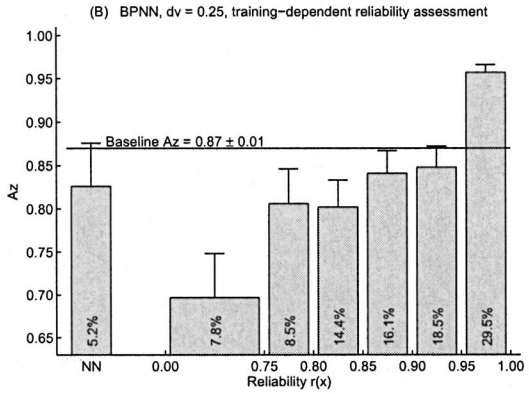
The training-independent reliability assessment produced a stronger correlation between reliability values and $A_z$ values; in other words, as reliability values increase, $A_z$ values also increase. The training-dependent reliability assessment, on the other hand, results in "overoptimistic" estimation of $r(\mathbf{x})$ values where more cases are assigned higher reliability scores. Since validation patterns are to some extent stored in the decision model (they were used for the model development/training), the validation error (calculated over a subset of them) tends to be underestimated. As a result, each of the top three reliability-based bins contains a larger fraction of cases than for the training-independent reliability assessment. However, the $A_z$ performance achieved by the CAD system on cases from these groups is lower, indicating that the real accuracy of the decision model was actually inferior to predicted by the reliability analysis.

Finally, the substantially better ROC performance of the CAD system on the NN cases for the training-dependent, compared to the training-independent, validation scenario comes as an unexpected observation. The discrepancy could be attributed to the fact that the NN groups contain 5% of the most unusual cases (without even a single similar case in the




$p$-values for differences in $A_z$ indices between $r(\mathbf{x})$-based groups

| $r(\mathbf{x})$ | [0.00, 0.75) | [0.75, 0.80) | [0.80, 0.85) | [0.85, 0.90) | [0.90, 0.95) |
|---|---|---|---|---|---|
| [0.95, 1.00] | **≪0.01** | **≪0.01** | **≪0.01** | **≪0.01** | **<0.01** |
| [0.90, 0.95) | **≪0.01** | **<0.01** | **<0.01** | 0.31 | |
| [0.85, 0.90) | **0.01** | **0.02** | **0.04** | | |
| [0.80, 0.85) | 0.76 | 0.61 | | | |
| [0.75, 0.80) | 0.80 | | | | |




$p$-values for differences in $A_z$ indices between $r(\mathbf{x})$-based groups

| $r(\mathbf{x})$ | [0.00, 0.75) | [0.75, 0.80) | [0.80, 0.85) | [0.85, 0.90) | [0.90, 0.95) |
|---|---|---|---|---|---|
| [0.95, 1.00] | **≪0.01** | **≪0.01** | **≪0.01** | **≪0.01** | **≪0.01** |
| [0.90, 0.95) | **0.01** | 0.38 | 0.25 | 0.85 | |
| [0.85, 0.90) | **0.01** | 0.47 | 0.34 | | |
| [0.80, 0.85) | 0.08 | 0.93 | | | |
| [0.75, 0.80) | 0.09 | | | | |

FIG. 5. Effect of the data handling scheme on reliability estimates: (a) training-independent reliability assessment and (b) training-dependent reliability assessment. The corresponding tables list $p$ values of pairwise differences in $A_z$ performance among the various reliability strata; $p$ values in bold indicate statistically significant differences at 95% confidence level. **≪0.01** indicates a $p$ value smaller than 0.001.

other data) and therefore the CAD predictions for these subsets may not follow the general trend. Nevertheless, it should be noted that for both scenarios the ROC area index achieved for the NN groups is significantly lower than the baseline performance of the system on the full set of cases.

Although the results shown in Fig. 5 correspond to the BPNN, the same trends regarding the training-independent and the training-dependent reliability assessment were observed for the other two decision models as well.

## IV. DISCUSSION

We have proposed a general theoretical framework for enhancing CAD systems with the ability to report expected reliability for each individual decision being made. The proposed framework capitalizes on the CAD system's validation on a separate set of cases that are reserved only for this purpose. These cases are chosen to reflect the patient population on which this system is employed or the population on which the system has been trained.

The proposed approach is based on the premise that the case-specific reliability of a prediction from a CAD system can be described in terms of the local accuracy the system. The technique analyzes the feature space neighborhood of the query case to dynamically select an input-dependent set of known cases relevant to the query. This set is used to assess the local (query-specific) accuracy of the CAD system. The estimated local accuracy is utilized as a reliability measure of the CAD response to the query case. It is expected that CAD predictions with higher associated reliability will be more accurate.

We tested the study hypothesis across three different decision models (BPNN, GRNN, and SVM) and two data handling schemes (training-independent and training-dependent reliability assessment) observing robustness and consistency of trends. The experimental results demonstrate that the proposed reliability measure is an accurate predictor of CAD performance regarding a particular query case. The presented reliability estimation technique does not require creating additional decision models or augmenting them with additional modules for accuracy assessment. Furthermore, it is easily applicable with virtually any decision model.

Special attention, however, needs to be given to a number of parameters such as (*i*) the similarity metric used to compare the query case to examples from the validation pool, (*ii*) the size of the neighborhood radius defining the span of the relevant set, and (*iii*) the error criterion describing the system's local performance on the relevant set. Although a variety of similarity measures can be considered, we believe that selection of the internal metric used by the underlying decision model is a natural and reasonable choice. In the case of models that do not make direct comparisons between patterns (e.g., BPNN), a similarity metric needs to be selected

empirically as the optimal choice may depend on the problem and the dataset at hand. In this study we followed a simple constant neighborhood radius approach, but in general $d_V$ may be a function of the query case and properties of its neighborhood in the validation pool, i.e., local class prevalence/balance and/or local data density. Under this scenario the choice of the optimal $d_V$ value(s) will be a result of numerical optimization rather that empirical selection. Finally, the mean squared error, although widely used as a performance measure for decision models developed in a train-test mode (e.g., BPNNs), may be substituted with other error criteria such as the modified perceptron error,[30] the classification figure of merit,[31] and potentially others.

There are two possible limitations in this study. First, since the number of similar relevant cases is variable, it is possible that it may affect the conclusions of the study. We have explored this possibility by monitoring the average number of relevant cases for CAD predictions assigned to every reliability-based bin shown in Figs. 3–5. We observed that the number of relevant cases does not seem to affect the numerical value of the reliability estimate. Although cases assigned higher reliability values tend to have larger number of neighbors in the validation pool, the differences are hot statistically significant. This finding was consistent across all decision models investigated in the study. Second, experimental results from two cross-validation runs were pooled for performance analysis. Although this is a common practice in CAD, we analyzed each fold separately as well. For all three decision models, we observed no statistically significant differences between $A_z$ estimates for groups of CAD predictions stratified into the same reliability-based bin but coming from different cross-validation folds.

Clinical integration of the proposed case-specific reliability measure is a challenging issue. The reliability measure could help with a risk stratification scheme to affirm the CAD user when a highly reliable opinion is offered or to alert him/her in case of a questionable one. There are different ways to integrate the reliability measure depending on the operating mode of the CAD system. One way is to use the reliability measure as a filtering mechanism for what is reported to the user. For example, we can apply a carefully chosen threshold to the reliability measure and if, for a given case, the reliability measure exceeds this threshold, then the CAD opinion will be presented to the user since the system is deemed reliable for the specific case. This type of integration is suitable for a computer-assisted detection system that presents visual cues to the radiologist. If an image location produces a CAD output above the system's operating threshold and its corresponding reliability is above the reliability threshold mentioned above, then the CAD opinion is deemed reliable and will be presented as a visual cue. A reliability-based risk stratification such as the one presented in the manuscript could be more appropriate for CAD systems that do not provide binary decisions (e.g., diagnostic systems that determine the likelihood of malignancy). The issue of clinical integration has been the focus of our research efforts leading to some promising preliminary results.[32]

Finally, the proposed reliability assessment framework could also form a foundation for quality monitoring practices that are currently lacking in the field of CAD. By continuous observation of the system's self-assessed experience and accuracy on new cases, we provide the CAD user with a quantitative measure of the relevance of the system to a particular clinical practice. If, for example, the percentage of no-neighbor cases (with no similar cases in the system's database to assess reliability) is increasing with time, this is an important indication that the practice's patient population has changed. Therefore, the CAD user should be warned that he/she cannot expect the same level of system's performance as before.

In conclusion, the proposed reliability analysis framework has the potential to serve as a mechanism for patient-specific customization of existing or future-developed CAD technology.

## ACKNOWLEDGMENTS

This work was partially supported by Grant No. R01 CA101911 from the National Cancer Institute and by the Systems Research Institute, Polish Academy of Sciences, Newelska 6, Warsaw, Poland.

[a]Electronic mail: habas@ci.louisville.edu

[1]E. A. Krupinski, "Computer-aided detection in clinical environment: benefits and challenges for radiologists," Radiology **231**, 7–9 (2004).
[2]D. A. Nix and A. S. Weigend, "Learning local error bars for nonlinear regression," in *Advances in Neural Information Processing Systems*, Vol. 7, edited by G. Tesauro, D. Touretzky, and T. Leen (MIT, Cambridge, 1995), pp. 489–496.
[3]C. S. Qszaz, "Bayesian error bars for regression," Ph.D. thesis, Aston University, Birmingham, 1996.
[4]P. J. Edwards, A. F. Murray, G. Papadopoulos, A. R. Wallace, J. Barnard, and G. Smith, "The application of neural networks to the papermaking industry," IEEE Trans. Neural Netw. **10**, 1456–1464 (1999).
[5]G. Papadopoulos, P. J. Edwards, and A. F. Murray, "Confidence estimation methods for neural networks: a practical comparison," IEEE Trans. Neural Netw. **12**, 1278–1287 (2001).
[6]C. M. Bishop, "Novelty detection and neural network validation," IEE Proc. Vision Image Signal Process., **141**, 217–222 (1994).
[7]J. R. Donaldson and R. B. Schnabel, "Computational experience with confidence regions and confidence intervals for nonlinear least squares," Technometrics **29**, 67–82 (1987).
[8]J. T. G. Hwang and A. A. Ding, "Prediction intervals for artificial neural networks," J. Am. Stat. Assoc. **92**, 748–757 (1997).
[9]B. Efron and R. J. Tibshirani, *An Introduction to Bootstrap* (Chapman & Hall, New York, 1993).
[10]Y. Jiang, "Uncertainty in the output of artificial neural networks," IEEE Trans. Med. Imaging **22**, 913–921 (2003).
[11]G. D. Tourassi and C. E. Floyd, "The effect of data sampling on the performance evaluation of artificial neural networks in medical diagnosis," Med. Decis Making **17**, 186–192 (1997).
[12]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).
[13]M. Heath, K. W. Bowyer, and D. Kopans, "Current status of the digital database for screening mammography," in *Digital Mammography* (Kluwer Academic, Dordrecht, 1998), pp. 457–460.
[14]N. H. Eltonsy, G. D. Tourassi, P. A. Habas, and A. S. Elmaghraby, "DNA: Directional neighborhood analysis for detection of breast masses in screening mammograms," Proc. SPIE **5747**, 38–47 (2005).
[15]C. M. Bishop, *Neural Networks for Pattern Recognition* (Oxford University Press, New York, 1995).
[16]J. M. Zurada, *Introduction to Artificial Neural Systems* (West, St. Paul,

1992).
[17]D. H. Nguyen and B. Widrow, "Neural networks for self-learning control systems," IEEE Control Syst. Mag. **10**, 18–23 (1990).
[18]K. Levenberg, "A method for the solution of certain problems in least squares," Q. Appl. Math. **2**, 164–168 (1944).
[19]D. Marquardt, "An algorithm for least-squares estimation of nonlinear parameters," SIAM J. Appl. Math. **11**, 431–441 (1963).
[20]D. F. Specht, "A general regression neural networks," IEEE Trans. Neural Netw. **2**, 568–576 (1991).
[21]P. D. Wasserman, *Advanced Methods in Neural Computing* (Wiley, New York, 1993).
[22]S. Chen, C. F. N. Cowan, and P. M. Grant, "Orthogonal least squares learning algorithm for radial basis function networks," IEEE Trans. Neural Netw. **2**, 302–309 (1991).
[23]V. N. Vapnik, *Statistical Learning Theory* (Wiley, New York, 1998).
[24]L. Wei, Y. Yang, R. M. Nishikawa, and Y. Jiang, "A study on several machine-learning methods for classification of malignant and benign clustered microcalcifications," IEEE Trans. Med. Imaging **24**, 371–380 (2005).
[25]C. E. Metz, "Basic principles of ROC analysis," Semin Nucl. Med. **8**, 283–298 (1978).
[26]C. E. Metz, "ROC methodology in radiologic imaging," Invest. Radiol. **21**, 720–733 (1986).
[27]J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," Radiology **143**, 29–36 (1982).
[28]C. E. Metz, B. A. Herman, and J. Shen, "Maximum-likelihood estimation of ROC curves from continuously-distributed data," Stat. Med. **17**, 1033–1053 (1998).
[29]C. E. Metz, B. A. Herman, and C. A. Roe, "Statistical comparison of two ROC curve estimates obtained from partially-paired datasets," Med. Decis Making **18**, 110–121 (1998).
[30]E. Barnard and D. Casasent, "A comparison between criterion functions for linear classifiers, with an application to neural nets," IEEE Trans. Syst. Man Cybern. **19**, 1030–1041 (1989).
[31]J. B. Hampshire and A. H. Waibel, "A novel objective function for improved phoneme recognition using time-delay neural networks," IEEE Trans. Neural Netw. **1**, 216–228 (1990).
[32]P. A. Habas, J. M. Zurada, A. S. Elmaghraby, and G. D. Tourassi, "Confidence-based stratification of CAD recommendations with application to breast cancer detection," Proc. SPIE **6144**, 1759–1766 (2006).