Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

Probabilistic Framework for Reliability Analysis of Information-Theoretic CAD Systems in Mammography

Piotr A. Habas, *Student Member, IEEE,* Jacek M. Zurada, *Fellow, IEEE,*
Adel S. Elmaghraby, *Senior Member, IEEE,* and Georgia D. Tourassi, *Member, IEEE*

***Abstract*— The purpose of this study is to develop and evaluate a probabilistic framework for reliability analysis of information-theoretic computer-assisted detection (IT-CAD) systems in mammography. The study builds upon our previous work on a feature-based reliability analysis technique tailored to traditional CAD systems developed with a supervised learning scheme. The present study proposes a probabilistic framework to facilitate application of the reliability analysis technique for knowledge-based CAD systems that are not feature-based. The study was based on an information-theoretic CAD system developed for detection of masses in screening mammograms from the Digital Database for Screening Mammography (DDSM). The experimental results reveal that the query-specific reliability estimate provided by the proposed probabilistic framework is an accurate predictor of CAD performance for the query case. It can also be successfully applied as a base for stratification of CAD predictions into clinically meaningful reliability groups (i.e., HIGH, MEDIUM, and LOW). Based on a leave-one-out sampling scheme and ROC analysis, the study demonstrated that the diagnostic performance of the IT-CAD is significantly higher for cases with HIGH reliability ($A_z = 0.92 \pm 0.03$) than for those stratified as MEDIUM ($A_z = 0.84 \pm 0.02$) or LOW reliability predictions ($A_z = 0.78 \pm 0.02$).**

I. Introduction

There exists a wide range of commercial and academic computer assisted detection (CAD) systems developed for the detection of malignancies in screening mammograms [1]. While these systems are based on various engineering principles, they are all designed to provide a second opinion to radiologists. Previous studies have shown that CAD technology has a positive impact on early breast cancer detection [2], [3], [4]. However, current CAD systems are still burdened by an excessive false positive rate. Consequently, radiologists (especially the less experienced) often find the task of recognizing and dismissing accordingly false positive CAD cues very challenging [5], [6], [7].

The high false positive rate is considered the main reason for radiologists' reluctance to trust CAD systems and this, in turn, often results in dismissal of correct CAD cues. Reducing the false positive rate would not have to be the only defense strategy to improve the clinical benefit of CAD if the second reader CAD systems were able to justify their opinions. Unfortunatelly, in the standard "black box" cueing capacity, CAD tools fall short of that role.

To address the need for more interpretive CAD, we proposed a technique that aims at providing the CAD user with an additional level of information beyond typical estimation of global accuracy on the general population of prospective cases [8]. The technique relied on the analysis of the feature space neighborhood of the query case and dynamically selected an input-dependent set of cases relevant to the query. Subsequently, this set was used to estimate the local (query-dependent) reliability of the CAD system. The study showed that the proposed reliability metric reported together with the CAD prediction is an accurate predictor of the system's query-specific performance.

The above reliability estimation scheme was originally proposed for feature-based CAD systems developed in the traditional train-test mode. However, knowledge-based CAD systems have been recently becoming increasingly popular in mammography. They aim at providing the user with evidence-based decision support by means of relating a new query case to other cases stored in a knowledge databank. A diagnosis is assigned to the new case by analogy or copying the answer if the match is close enough. Knowledge-based CAD systems in mammography are able to take full advantage of growing libraries of digital mammograms without a need of retraining. Thank to interactive nature, they allow physicians to formulate their own questions and get interpretable answers (e.g., the CAD response is often analogous to the odds-ratio). We have previously presented a knowledge-based CAD system for detection of masses in screening mammograms [9]. The main innovation of the system was that it did not rely on image-extracted features to assess case similarity, but rather applied information-theoretic principles to measure directly the global similarity between a new and an archived mammographic case. The purpose of the present study is to investigate if the previously proposed reliability analysis framework can be extended to featureless, information-theoretic CAD systems.

The article is organized as follows. Section II provides detailed information regarding the CAD system, the proposed probabilistic reliability analysis framework, the testing dataset, and the overall study design. Section III presents experimental results leading to the final conclusion in Section IV.

P. A. Habas is with the Computational Intelligence Laboratory, University of Louisville, Louisville, KY 40292, USA (phone: +1 502 852 3165; fax: +1 502 852 3940; email: p.habas@ieee.org).

J. M. Zurada is with the Department of Electrical and Computer Engineering, University of Louisville, Louisville, KY 40292, USA (email: j.zurada@ieee.org).

A. S. Elmaghraby is with the Department of Computer Engineering and Computer Science, University of Louisville, Louisville, KY 40292, USA (email: adel@louisville.edu).

G. D. Tourassi is with the Duke Advanced Imaging Laboratories, Department of Radiology, Duke University Medical Center, Durham, NC 27705, USA (email: georgia.tourassi@duke.edu). The work of G. D. Tourassi was sponsored in part by the National Cancer Institute grant R01 CA101911.

II. Methodology

A. Information-Theoretic CAD System

An information-theoretic CAD (IT-CAD) system in mammography is usually defined by three key elements: (i) a knowledge databank where mammographic cases with known ground truth (templates) are stored, (ii) a similarity metric used to assess a match between two cases/images, and (iii) a decision algorithm that calculates a CAD prediction regarding the query case.

The most critical component turns out to be the similarity metric which in our IT-CAD system is normalized mutual information. Mutual information (MI) is a measure of general interdependence between random variables [10] and a popular similarity measure for image registration. If X and Y represent two medical images, their mutual information $\text{MI}(X,Y)$ is expressed as

$$\text{MI}(X,Y) = \sum_x \sum_y P_{XY}(x,y) \log_2 \frac{P_{XY}(x,y)}{P_X(x)P_Y(y)} \qquad (1)$$

where $P_{XY}(x,y)$ is the joint probability density function (pdf) of the two images based on their corresponding pixel values. Equation (1) assumes that the image pixel values are samples of two random variables x and y and $P_X(x)$ and $P_Y(y)$ are the respective marginal pdfs. Mutual information is essentially an intensity-based measure of how much alike two images are. Theoretically, MI is a more effective and robust similarity index than traditional correlation metrics because it does not make any prior assumptions regarding linear relationship between the intensity values of the two images [10]. This assumption is often violated in mammograms, especially in the presence of tumors.

The present study utilizes normalized mutual information (NMI) to facilitate the probabilistic framework needed for reliability analysis.

$$\text{NMI}(X,Y) = \frac{2\text{MI}(X,Y)}{H(X)+H(Y)} \qquad (2)$$

The normalization incorporates the individual entropies $H(X)$ and $H(Y)$ of the two images X and Y and results in bounding the NMI values between 0 and 1. When $\text{NMI}(X,Y)=1$, the two compared images X and Y are identical. On the other hand, when $\text{NMI}(X,Y)=0$, images X and Y are completely unrelated.

The decision algorithm compares a new mammographic case Q presented to the system with all mass templates stored in the knowledge databank. The comparison is effectively a pairwise calculation of the normalized mutual information $\text{NMI}(Q,M_i)$ between the query case Q and each mass template M_i. Since the calculation of $\text{NMI}(Q,M_i)$ is based on the whole image, a correction component based on similarity $\text{NMI}(Q,N_j)$ between the query case Q and normal templates N_j is used to ensure that a high values of $\text{NMI}(Q,Mi)$ are not a result of matching backgrounds rather than potential abnormalities. The decision index $D(Q)$ for the query case is calculated as the difference between the average $\text{NMI}(Q,M_i)$ and the average $\text{NMI}(Q,N_j)$

$$D(Q) = \frac{1}{m}\sum_{i=1}^{m} \text{NMI}(Q,M_i) - \frac{1}{n}\sum_{j=1}^{n} \text{NMI}(Q,N_j) \qquad (3)$$

across all m mass templates M_i and n normal templates N_j stored in the knowledge databank.

B. Reliability Analysis

The previously presented reliability analysis framework was based on a hypothesis that uncertainty (error) and reliability are mutually exclusive. A query-specific uncertainty can be measured using an input-dependent set of cases similar to the query and evaluated in terms of the mean square error between the truth value and the CAD output. The experimental results confirmed the hypothesis and demonstrated that predictions with lower validation error are indeed more accurate and therefore more reliable.

Contrary to many previously investigated CAD systems for detection of mammographic masses, the proposed knowledge-based IT-CAD system does not undergo supervised learning procedure that is intended to minimize the difference between target values (binary coded ground truth) and the CAD predictions for a set of training cases. The decision algorithm is essentially a ranking scheme assigning each mammographic query case Q a continuous real-valued decision index $D(Q)$ with a higher score indicating a higher likelihood of containing a mass. Therefore, the reliability of CAD predictions need to be measured in terms of their relative rank (order) rather than absolute values.

For such CAD systems intended to provide discrimination between classes rather than reproduction of certain target values, we propose a probabilistic adaptation of the previous reliability analysis framework. It binds the reliability of a CAD output $D(Q)$ for a case Q with its rank among all predictions for a given class (mass or normal). CAD output is treated as a random variable z with conditional probability distribution functions $f(z|M)$ for mass cases and $f(z|N)$ for normal cases, respectively, approximated by the observed distribution of CAD outputs for cases stored in the knowledge databank (Fig. 1A).

Given Q is a mass case, the reliability of a prediction $D(Q)$ is expressed in terms of the (empirical) conditional cumulative probability $F(z|M)$ calculated at $z=D(Q)$

$$R_M(D(Q)) = F(D(Q)|M) = P(z \le D(Q)|M) \qquad (4)$$

as a higher value of $D(Q)$ means a higher likelihood of containing a mass and therefore more accurate/reliable prediction (Fig. 1B). Contrary, if Q is a normal case, the reliability of the CAD output $D(Q)$ is calculated as the 1's supplement of the conditional cumulative probability $F(z|N)$ for $z=D(Q)$

$$R_N(D(Q)) = 1 - F(D(Q)|N) = 1 - P(z \le D(Q)|N) \qquad (5)$$

because a lower value of $D(Q)$ means a more accurate/reliable prediction (Fig. 1C).

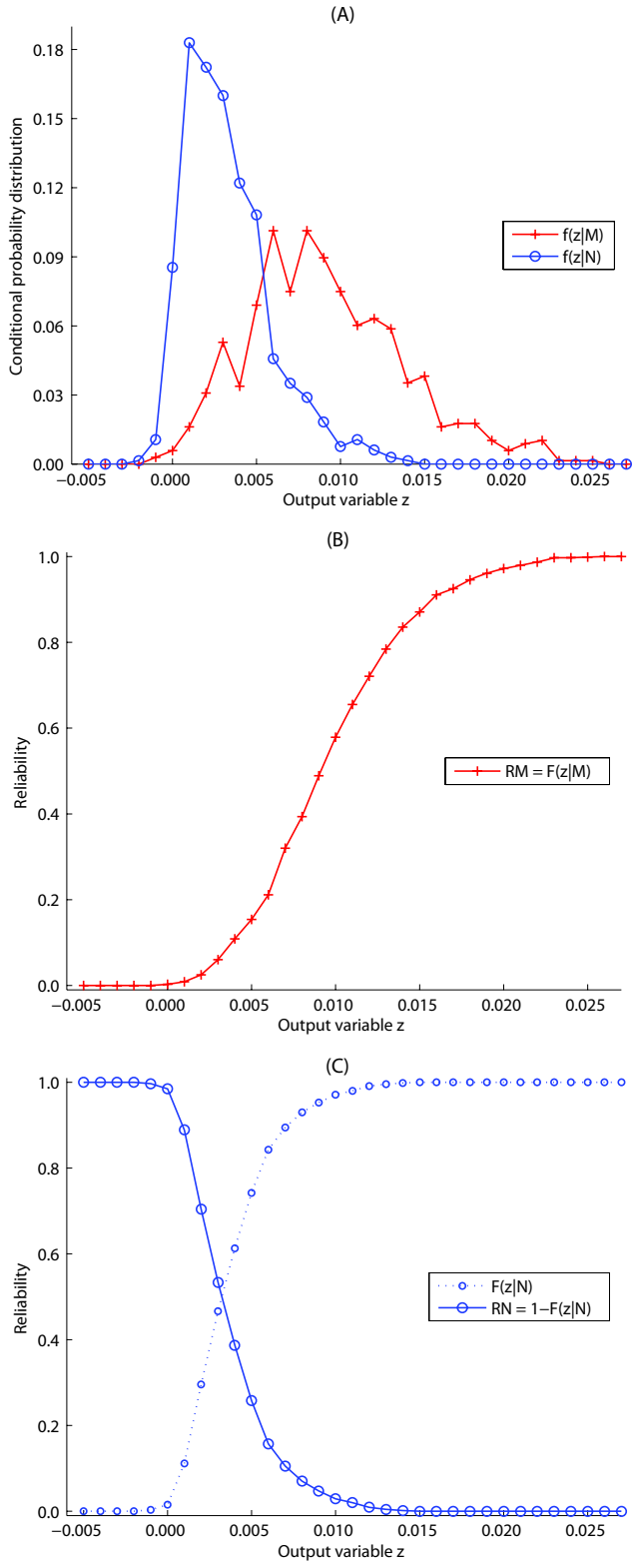

Fig. 1. Key components of the proposed probabilistic reliability analysis framework: (A) conditional probability distributions of CAD outputs for the two classes of mammographic ROIs, (B) reliability/accuracy of CAD predictions for mass templates, and (C) reliability/accuracy of CAD predictions for normal templates.

For a query case Q (with unknown ground truth), the reliability of a CAD prediction $D(Q)$ can be estimated based on reliability values assigned to CAD outputs for a set of cases relevant to the query. Although various metrics can be applied to compare similarity between cases, the internal measure used by the IT-CAD system (here NMI) seems to be a natural and reasonable choice. The size of the relevant set may be either query-dependent (if based on a minimum required similarity, may result in no relevant cases found) or constant (if always k most similar templates are retrieved). For the latter approach, followed in this study, the reliability function has a form

$$R(D(Q)) = \frac{1}{k}\left[\sum_i R_M(D(M_i)) + \sum_j R_N(D(N_j))\right] \quad (6)$$

where M_i $(i = 1, 2, \ldots)$ and N_j $(j = 1, 2, \ldots)$ denote the k templates from the knowledge databank (mass or normal) that are most similar to the query case Q in terms of NMI.

C. Dataset

The knowledge databank of the IT-CAD system is based on a collection of mammograms extracted from the Digital Database for Screening Mammography (DDSM) [11] that were digitized using the Lumisys scanner. First, $m = 681$ DDSM/Lumisys mammograms with annotated masses were selected to extract 512×512 pixels regions of interest (ROIs) centered on the known location of each abnormality. The mass cases comprised a wide range of mass shapes, mass margins, breast parenchymal density and were balanced in terms of malignancy status (340 malignant and 341 benign). Then, 512×512 pixels normal ROIs were extracted from 82 DDSM/Lumisys mammograms that were deemed normal during screening and remained normal after a 4-year follow up period. Two ROIs were randomly selected from each of four breast views per case (left CC, left MLO, right CC, right MLO) for a total of $n = 82 \times 2 \times 4 = 656$ normal ROIs. As a result, the ROI dataset used for this study contained a total of $m + n = 681 + 656 = 1,337$ ROIs.

D. Performance Evaluation

The performance of the CAD system and the impact of the proposed reliability assessment scheme were evaluated using Receiver Operating Characteristics (ROC) analysis [12], [13], typically used in CAD applications. An ROC curve can be generated by applying a number of thresholds to the output decision variable and plotting the true positive fraction (TPF) as a function of the false positive fraction (FPF) for each threshold. Among many summary indices calculated from ROC curves, the most commonly used is the area under the ROC curve (A_z) [14], ranging from 0.5 for chance guessing to 1.0 for a perfectly operating classifier/ranker. The presence of statistically significant differences in A_z performance between reliability-based strata was verified using Student's t test at 95% confidence level.

III. Results

Based on the leave-one-out sampling scheme, each of the 1,337 ROIs was excluded once to serve as the query case while the remaining 1,336 ROIs were used as the knowledge databank. Taking into account data limitation, the reliability assessment was implemented based on the same 1,336 ROIs, although keeping an independent set for this purpose may result in more accurate (unbiased) reliability estimation.

The baseline performance of the IT-CAD system in discriminating mass from normal ROIs using (3) as the decision variable on the full set of 1,337 ROIs was $A_z = 0.88 \pm 0.01$. Then, the proposed reliability analysis scheme with $k = 30$ was applied to calculate the reliability score (6) associated with each individual CAD prediction. Consequently, the CAD predictions were stratified into three clinically meaningful groups of different reliability (HIGH, MEDIUM, and LOW). The cut-off points between the bins were selected arbitrarily to maximize the differences in A_z performance between the reliability-based groups.

The HIGH group contains 22.7% of CAD predictions with the highest reliability values for which the IT-CAD system demonstrates discriminative ability of $A_z = 0.92 \pm 0.03$. The LOW strata encloses 38.9% of CAD recommendations with the lowest reliability scores for which the IT-CAD system achieves performance of $A_z = 0.78 \pm 0.02$. The MEDIUM group contains all remaining CAD predictions (38.4%) that can be classified by the IT-CAD system with performance of $A_z = 0.84 \pm 0.02$. All pairwise differences in A_z performance between the reliability-based strata are statistically significant at 95% confidence level.

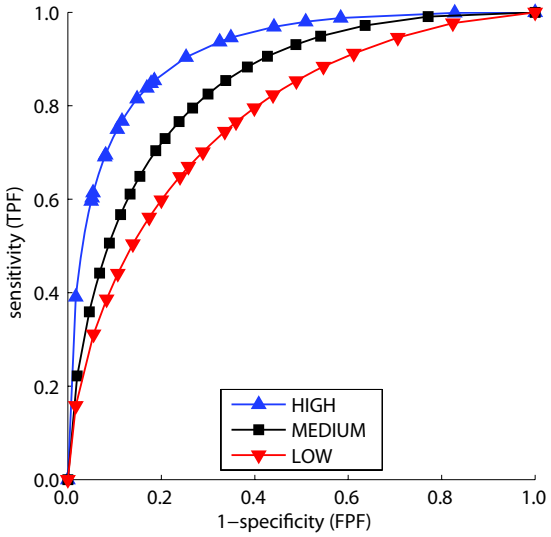

Fig. 2. Receiver Operating Characteristic (ROC) curves for 3 reliability-based groups (HIGH, MEDIUM, and LOW) of IT-CAD predictions.

The stratification results presented on Fig. 2 demonstrate that the proposed reliability score provides an additional level of diagnostic information that helps the CAD user recognize CAD recommendations that are more accurate than others.

IV. Conclusions

The latest research in CAD technology focuses on the interactive and interpretive aspect of CAD systems to facilitate their clinical acceptance and improve clinical effectiveness. Working towards this goal, we have developed a technique that allows a CAD system assess the case-specific reliability of its prediction for a query case, given prior experience with cases similar to the query. Our previous work focused on a reliability analysis technique for traditional feature-based CAD systems that follow the supervised learning scheme. The present study extended the technique to the more challenging information-theoretic CAD systems and demonstrated that with a proper probabilistic adaptation the reliability analysis framework is equally effective for evidence-based decision making as well.

References

[1] M. P. Sampat, M. K. Markey, and A. C. Bovik, "Computer-aided detection and diagnosis in mammography," in *Handbook of Image and Video Processing*. Elsevier Academic Press, 2005, pp. 1195–1217.

[2] L. J. Warren Burhenne, S. A. Wood, C. J. D'Orsi, S. A. Feig, D. B. Kopans, K. F. O'Shaughnessy, E. A. Sickles, L. Tabar, C. J. Vyborny, and R. A. Castellino, "Potential contribution of computer-aided detection to the sensitivity of screening mammography," *Radiology*, vol. 215, no. 2, pp. 554–562, 2000.

[3] R. L. Birdwell, D. M. Ikeda, K. F. O'Shaughnessy, and E. A. Sickles, "Mammographic characteristics of 115 missed cancers later detected with screening mammography and the potential utility of computer-aided detection," *Radiology*, vol. 219, no. 1, pp. 192–202, 2001.

[4] T. W. Freer and M. J. Ulissey, "Screening mammography with computer-aided detection: prospective study of 12,860 patients in a community breast center," *Radiology*, vol. 220, no. 3, pp. 781–786, 2001.

[5] B. Zheng, M. A. Ganott, C. A. Britton, C. M. Hakim, L. A. Hardesty, T. S. Chang, H. E. Rockette, and D. Gur, "Soft-copy mammographic readings with different computer-assisted detection cuing environments: preliminary findings," *Radiology*, vol. 221, no. 3, pp. 633–640, 2001.

[6] B. Zheng, R. G. Swensson, S. Golla, C. M. Hakim, R. Shah, L. Wallace, and D. Gur, "Detection and classification performance levels of mammographic masses under different computer-aided detection cueing environments," *Acad. Radiol.*, vol. 11, no. 4, pp. 398–406, 2004.

[7] E. A. Krupinski, "Computer-aided detection in clinical environment: benefits and challenges for radiologists," *Radiology*, vol. 231, no. 1, pp. 7–9, 2004.

[8] P. A. Habas, G. D. Tourassi, N. H. Eltonsy, A. S. Elmaghraby, and J. M. Zurada, "A novel technique for assessing the case-specific reliability of decisions made by cad tools," in *Medical Imaging 2005: Image Processing*, Proc. SPIE, vol. 5747, pp. 124–131, 2005.

[9] G. D. Tourassi, R. Vargas-Voracek, D. M. Catarious, and C. E. Floyd, "Computer-assisted detection of mammographic masses: a template matching scheme based on mutual information." *Med. Phys.*, vol. 30, no. 8, pp. 2123–2130, 2003.

[10] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley & Sons, Inc., 1991.

[11] M. Heath, K. W. Bowyer, D. Kopans, R. Moore, and P. Kegelmeyer, "The digital database for screening mammography," in *Proceedings of the 5th International Workshop on Digital Mammography*. Medical Physics Publishing, 2000.

[12] C. E. Metz, "Basic principles of ROC analysis," *Semin. Nucl. Med.*, vol. 8, no. 4, pp. 283–298, 1978.

[13] C. E. Metz, "ROC methodology in radiologic imaging," *Invest. Radiol.*, vol. 21, no. 9, pp. 720–733, 1986.

[14] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.