Ultrasound in Med. & Biol., Vol. 33, No. 11, pp. 1688–1698, 2007
Copyright © 2007 World Federation for Ultrasound in Medicine & Biology
Printed in the USA. All rights reserved
0301-5629/07/$–see front matter

ELSEVIER

doi:10.1016/j.ultrasmedbio.2007.05.016

● *Original Contribution*

COMPUTER AIDED CLASSIFICATION SYSTEM FOR BREAST ULTRASOUND BASED ON BREAST IMAGING REPORTING AND DATA SYSTEM (BI-RADS)

WEI-CHIH SHEN,* RUEY-FENG CHANG,*† and WOO KYUNG MOON‡

*Department of Computer Science and Information Engineering, National Chung Cheng University, Chiayi, Taiwan, R.O.C.; †Department of Computer Science and Information Engineering, Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taipei, Taiwan, R.O.C.; and ‡Department of Radiology and Clinical Research Institute, Seoul National University Hospital, Seoul, Korea

(*Received* 16 *October* 2006; *revised* 7 *May* 2007; *in final form* 18 *May* 2007)

Abstract—Clinically, the ultrasound findings are evaluated by its sonographic characteristics and then assigned to assessment categories according to the definitions of Breast Imaging Reporting and Data System (BI-RADS) developed by the American College of Radiology. In this study, a computer-aided classification (CAC) system was proposed to classify the masses into assessment categories 3, 4 and 5, which simulated the clinical diagnosis of radiologists. Compared with current computer-aided diagnosis systems, the proposed CAC system classifies the indeterminate cases into BI-RADS category 4 for further diagnosis. Six hundred twenty-six cases were collected from three ultrasound systems and confirmed by pathology and retrospectively classified into categories 3, 4 and 5 by radiologists. The multinomial logistic regression model was trained as the CAC system for predicting the assessment category from the computerized BI-RADS features and from a set of machine-dependent factors. By using the machine-dependent factors to indicate the adopted ultrasound systems, the same regression model could be applied for the cases acquired from different ultrasound systems. A basic CAC system was trained by using the classification result of radiologists. A weighted CAC system, to improve the capacity of the basic CAC system in differentiating benign from malignant lesions, was trained by adding the pathologic result. Between the radiologists and the basic CAC system, a substantial agreement was indicated by Cohen's kappa statistic and the differences in either the performance indices or the A_Z of receiver operating characteristic (ROC) analysis were not statistically significant. For the weighted CAC system, the performance indices accuracy, sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) were 73.00% (457 of 626), 98.17% (215 of 219), 59.46% (242 of 407), 56.58% (215 of 380) and 98.37% (242 of 246), respectively; the A_Z was 0.94; and the correlation with the radiologists was also substantial agreement. The indices accuracy and specificity of weighted CAC system, compared with those of the radiologists, were improved by 5.91% and 8.85%, respectively and the indices of sensitivity and NPV, compared with those of a conventional CAD system, were improved by 10.5% and 5.21%, respectively; all improvements were statistically significant. To classify the mass into BI-RADS assessment categories by the CAC system is feasible. Moreover, the proposed CAC system is flexible because it can be used to diagnose the cases acquired from different ultrasound systems. (E-mail: rfchang@csie.ntu.edu.tw) **© 2007 World Federation for Ultrasound in Medicine & Biology.**

Key Words: **Breast cancer, Ultrasound image, BI-RADS, Computer-aided classification (CAC) system, Computer-aided diagnosis (CAD) system, Logistic regression.**

INTRODUCTION

Breast cancer is the leading cause of death among women worldwide (Parkin et al. 2005; Jemal et al. 2006). Due to early detection, intervention and postoperative treatment, mortality rates of breast cancer have been decreased. The screening by mammography has largely contributed to the early detection of breast cancer. Ultrasound, as an adjunct technique to mammography, can also increase the overall sensitivity of conventional breast imaging (Rizzatto 2001; Baker and Soo 2002). Stavros et al. (1995) proposed several useful sonographic features to describe a mass, such as ellipsoid shape, number of lobulations, ratio of width to anteroposterior dimension, marked hypoechoic region, punctuated calcification and posterior shadowing that help radiologists to

Address correspondence to: Professor Ruey-Feng Chang, Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan 10617, R.O.C. E-mail: rfchang@csie.ntu.edu.tw

accurately classify the masses as benign, indeterminate and malignant (sensitivity of 98.4% and negative predictive value [NPV] of 99.5%).

With the rapid development of computer applications, many researches have been carried out to develop a computer-aided diagnosis (CAD) system for differentiating malignant from benign masses (Chen et al. 1999, 2002, 2003; Horsch et al. 2002; Kuo et al. 2002; Joo et al. 2004; Drukker et al. 2004, 2005). To describe the characteristics of mass in a computer system, the sonographic features developed in previous clinical studies, such as the features developed by Stavros et al. (1995), are, therefore, quantified as "computerized features." Kim et al. (2005) evaluated the correlation between the computerized schemes and the grading of radiologists for six sonographic features. The computerized features are roughly classified into two types: morphologic features and acoustic features. The morphologic feature describes the formation of a mass, such as shape, orientation and marginal features (Chen et al. 2003; Joo et al. 2004) and the acoustic feature describes the correlation or the dissimilarity of pixels, such as lesion boundary, echo pattern and posterior shadowing features (Chen et al. 1999, 2002; Horsch et al. 2002). However, the pixel correction is always affected by the settings of the adopted ultrasound system. Hence, the acoustic features will be closely related to the ultrasound system from which images are acquired. Therefore, an unanticipated diagnosis may be made if the acoustic features are used to diagnose the cases acquired from different ultrasound systems in a CAD system. To apply the same CAD system for different ultrasound machines, Kuo et al. (2002) proposed a feature-adjustment scheme that includes the resolution, gray value distribution and quantization adjustments for different machines. A CAD system with feature-adjustment scheme was better than a system without adjustment scheme. The improvement in accuracy, specificity and positive predictive value (PPV) was statistically significant. Drukker et al. (2005) used four features selected by the computer from 46 features for two different data sets, indicating that the performance of their CAD system was robust for different ultrasound system used.

The Breast Imaging Reporting and Data System (BI-RADS) developed by the American College of Radiology (ACR 2003) standardizes the descriptions of sonographic characteristics and the assessment of the findings to facilitate communication and to improve the outcomes of monitoring (Obenauer et al. 2005; Eberl et al. 2006). To help radiologists grade the diagnosed masses, the sonographic descriptions are grouped into six classes: orientation, margin, lesion boundary, echo pattern and posterior acoustic characteristics. On the basis of these features, radiologists grade each mass as benign (category 2), probably benign (category 3), suspicious abnormality (category 4) and highly suggestive of malignancy (category 5). Because tumors in categories 3, 4 and 5 always require further evaluation, many studies focus on the reliability of radiologists' grading for these tumors (Arger et al. 2001; Zonderland et al. 2004; Hong et al. 2005; Costantini et al. 2006). Buchbinder et al. (2004) focused the capacity of a computer-aided classification (CAC) system on the lesions assigned to BI-RADS category 3 by at least two of four radiologists. Their CAC system scored 38 of 42 malignant lesions that were initially assigned to category 3 as category 4 or 5; that is, the CAC system accurately upgraded the category in 90% of the 42 lesions.

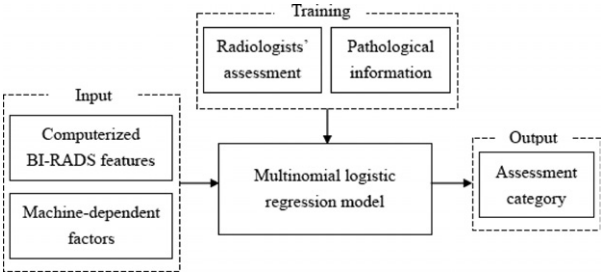

Fig. 1. The organization of proposed CAC system. The basic CAC system is trained by the radiologists' assessment, but the weighted CAC system is trained by associating the radiologists' assessment with the pathologic information.

In this study, a CAC system is proposed to classify the masses into the BI-RADS assessment categories 3, 4 and 5. The organization of which is shown in Fig. 1. The BI-RADS sonographic characteristics, shape, orientation, margin, lesion boundary, echo pattern and posterior acoustic features, are quantified by eight computerized features for characterizing the mass in the proposed CAC system. Besides, the adopted ultrasound systems represented by a set of machine-dependent factors, dummy variables (Hosmer and Lemeshow 2000), are also regarded as mass features. By these machine-dependent factors, the same CAC system could be applied for cases that were acquired from different ultrasound systems. Then, the classification results of radiologists are used to train a basic CAC system by the multinomial logistic regression (Hosmer and Lemeshow 2000). For increasing the diagnostic potential of the basic CAC system, a weighting strategy, using the pathologic information, is proposed to help the training process in constructing a weighted CAC system.

MATERIALS AND METHODS

This study was approved by the local ethics committee and informed consent was obtained from all patients included in the study. A total of 626 images were

obtained using ATL HDI 5000 (Advanced Technological Laboratory, Bothell, WA, USA) from June to July 2004, iU-22 (Philips Medical Systems, Bothell, WA, USA) from March to April 2004 and Voluson 730 expert (GE Medical systems, Kretz Ultrasound, Zipf, Austria) from April 2003 to February 2004. In all the scanning processes, the linear-array transducers were used, the images were captured at the largest diameters of the masses. Acoustic standoff pad was not used in these processes.

All diagnoses were proven by pathology either with fine-needle aspiration cytology (FNAC) or with core needle biopsy. The numbers of FNAC and cutting biopsies were 61 and 565, respectively. Two radiologists with 6 and 13 years experience were involved in the data collection. They concurrently retrospectively analyzed the static ultrasound images, assigned an ACR final assessment category and drew the margin of mass using a paint program (Microsoft Paint, version 5.2, Microsoft Inc, Seattle, WA, USA) by consensus. To avoid the influences on assigning the category and on deciding the mass margin, the radiologists were blind to the pathologic results. According to the classification result of the radiologists, the machine types and the pathologic result, the distribution of all adopted cases is summarized in Table 1.

Computerized BI-RADS features

In the masses category of ACR BI-RADS, the descriptive terms for evaluating the dominant features of mass are standardized as a lexicon and are grouped into six classes: shape, orientation, margin, lesion boundary, echo pattern and posterior acoustic features. For describing the shape and orientation of a mass, the corresponding best-fit ellipse is regarded as a baseline. Then, the semiminor axis and the angle of best-fit ellipse and the similarity between two shapes are defined as the computerized features. A distance map is transformed from the selected region-of-interest (ROI) and used to calculate the number of undulated and angular characteristics on the mass margin. Moreover, the outer and inner bands around the mass boundary are partitioned by the distance map and the difference between the two bands is used to evaluate the degree of abrupt interface across the lesion boundary. The echo pattern is evaluated by the average gray intensity and by the contrast within the mass. The difference between the two average gray intensities, the mass and the area behind the mass, is used to measure the posterior acoustic feature. Assume that a mass, circumscribed by a continuous closed boundary B with N_B pixels, is represented as region R containing N_R pixels; its BI-RADS sonographic features are quantified by the following eight computerized features.

Table 1. The distribution of all acquired cases is summarized by the classification results of radiologists, the machine types, and the pathological results

	Machine Type						
	HDI 5000		Voluson 730		Philips iU-22		
Radiologists	B	M	B	M	B	M	Total
Category 3	115	4	63	1	28	0	211
Category 4	68	39	103	38	27	17	292
Category 5	0	46	3	60	0	14	123
Total	272		268		86		626

Note: B is the number of benign cases. M is the number of malignant cases.

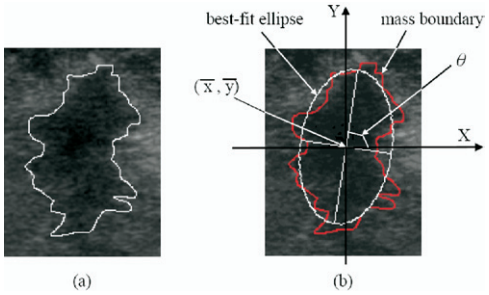

Fig. 2. The best-fit ellipse is found to roughly describe the mass shape. (a) A malignant mass. (b) The corresponding best-fit ellipse of (a).

A. Shape. A best-fit ellipse (Jain 1989) roughly describes the mass contour as shown in Fig. 2. Two shape features, S_b and S_{PR}, are then computed from the best-fit ellipse to quantify the mass shape. The feature S_b is defined by the length of the semiminor axis b of the best-fit ellipse, and the similarity, S_{PR}, between the mass contour and the best-fit ellipse can be measured by their perimeter ratio as

$$S_{PR} = \frac{N_B}{P_E} \tag{1}$$

where P_E is the perimeter of best-fit ellipse.

B. Orientation. The angle of major axis of the above best-fit ellipse is used to determine the mass orientation as shown in Fig. 2b. In general, the mass orientation is defined at the angles of 0 to $\frac{\pi}{2}$. But, the orientation of the mass shown in Fig. 3 is the smaller $\pi - \theta$, not θ. However, the angle of major axis of the best-fit ellipse is θ. Hence, the orientation feature, O_E, of mass is adjusted to be in the range 0 to $\frac{\pi}{2}$ as

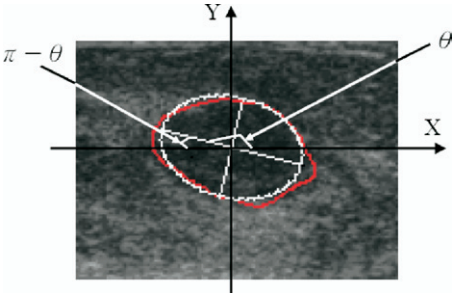

Fig. 3. The orientation of best-fit ellipse is close to the direction of mass. Occasionally, the supplementary angle is more appropriate to measure the parallel degree.

$$O_E = \begin{cases} \theta & if\ 0 \le \theta < \dfrac{\pi}{2} \\ \pi - \theta & if\ \dfrac{\pi}{2} \le \theta < \pi \\ \theta - \pi & if\ \pi \le \theta < \dfrac{3\pi}{2} \\ 2\pi - \theta & otherwise \end{cases} \quad (2)$$

where θ is the angle of the best best-fit ellipse.

C. Margin. The distance map (Jain 1989) is used to quantify the undulated and angular characteristics of the mass margin. For any pixel $P(x, y)$ of ROI, its eight neighbors are defined as

$$N_8(P) = \left\{ \begin{matrix} (x-1,\ y-1), (x,\ y-1),\ (x+1,\ y-1), (x-1,\ y), \\ (x+1,\ y), (x-1,\ y+1),\ (x,\ y+1), (x+1,\ y+1) \end{matrix} \right\}. \quad (3)$$

The distance between the pixel P in ROI and the boundary is recursively defined as

$$\text{distance}(P) = \min(\text{distance}(N_8(P))) + 1 \quad (4)$$

where $\min(\text{distance}(N_8(P)))$ is the minimum of the known distances of P's eight-neighbors. As the initial condition, the distances of mass boundary pixels are all set to 0. Different types of distance maps are shown in Fig. 4.

After computing the distance map, the maximum inscribed circle of the mass is found to segment the lobulate areas as shown in Fig. 5. The maximum distance within the lobulate area is used to estimate the degree of protuberance and to eliminate slighter undulation. When the interior maximum distance is less than three, the area is defined as slightly undulated and is ignored as shown in Fig. 5b. The undulation feature U is then defined by the number of lobulate areas observed. Furthermore, the

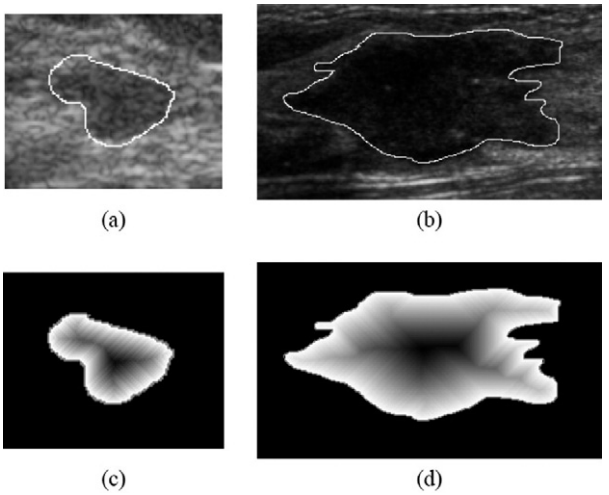

Fig. 4. The selected ROI is transformed into a distance map. (a) A benign mass. (b) A malignant mass. The formed maps of (a) and (b) are shown in (c) and (d), respectively. In both figures, the distance of each pixel in the mass is represented by the gray intensity and the black color denotes the largest distance.

local maxima within each lobulate area are found and grouped to compute the angular feature as shown in Fig. 6. The angular feature A is defined by the number of groups. Finally, the margin feature M_{UA} is defined as

$$M_{UA} = U + A. \quad (5)$$

D. Lesion boundary. The outer and inner bands around the mass boundary are used to quantify the degree of sharp demarcation. The distance map of a mass is also used to define the outer and the inner bands. As shown in Fig. 7, the outer and inner bands are defined by the pixels with a distance less than 10. The average gray intensities of these two bands with distance k can be then, respectively, defined as

$$avg\_outer = \frac{1}{N_{out}} \sum_{\text{distance}(P)=1}^{k} I(P) \text{ and}$$

$$avg\_inner = \frac{1}{N_{in}} \sum_{\text{distance}(P)=1}^{k} I(P) \quad (6)$$

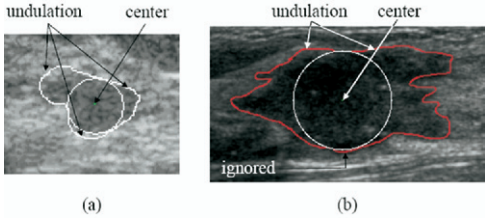

Fig. 5. The number of lobulate areas outside the maximum inscribed circle is defined as the undulation characteristic of margin class.

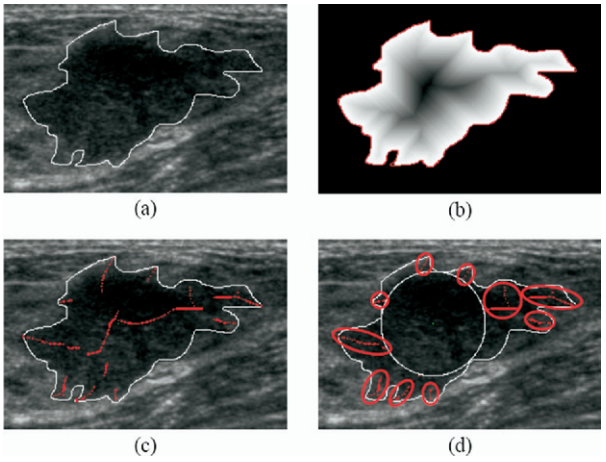

Fig. 6. The local maxima on the distance map are efficient for detecting the angular features. (a) A malignant mass. (b) The formed distance map of (a). (c) The local maxima on the map. (d) The local maxima in the lobulate areas are grouped and marked by ellipses. Ten angular features are observed here.

where $I(P)$ is the gray intensity of P, N_{out} is the number of pixels in the outer band and N_{in} is the number of pixels in the inner band. The lesion boundary feature LB_D is then defined as

$$LB_D = avg\_outer - avg\_inner. \quad (7)$$

In this study, the bandwidth is set to 3 by experiments.

E. Echo pattern. The average gray intensity EP_I is defined as

$$EP_I = \frac{1}{N_R}\sum_{P\in R} I(P) \quad (8)$$

where $I(P)$ is the gray intensity of mass pixel P. The histogram technique (Jain 1989) is useful for observing the distribution of gray intensities of a mass as shown in Fig. 8. For calculating the contrast within a mass, 25% brighter mass pixels are decided by a dynamic threshold to form the brighter group using the histogram. For example, the threshold is set to 51 and the brighter pixels contain 28.22% of mass pixels in Fig. 8a. The average gray intensity for the brighter group is then defined as:

$$avg\_bg = \frac{1}{N_{BP}}\sum_{P\in R \text{ and } I(P)\geq k} I(P) \quad (9)$$

where k is the dynamic threshold, $I(P)$ is the gray intensity of mass pixel P and N_{BP} is the number of brighter pixels. The contrast feature EP_C is then defined as:

$$EP_C = \frac{avg\_bg - EP_I}{EP_I}. \quad (10)$$

F. Posterior acoustic features. First, the area under the mass is identified for quantifying the posterior acoustic features. The width of the posterior area pw is two thirds of the mass width, mw and the height of posterior area, ph, is mass height, mh (but not exceeding 100 pixels). An example for defining the posterior area is shown in Fig. 9. Then, the average gray intensity of posterior acoustic area is defined as

$$avg\_pa = \frac{1}{N_{PA}}\sum_{P\in PA} I(P) \quad (11)$$

where $I(P)$ is the gray intensity of pixel P and N_{PA} is the number of pixels in the posterior acoustic area PA. The difference between the average gray intensity of the mass and that of the posterior acoustic area is used to evaluate the posterior acoustic characteristic PS_D as

$$PS_D = avg\_pa - EP_I \quad (12)$$

where EP_I is defined in eqn 8.

Machine-dependent factors

In the regression model, the dummy variable (Hosmer and Lemeshow 2000) is used to distinguish the subgroups of experimental cases and to correct the subgroup effect. A set of the dummy variables is useful to represent multiple subgroups in a single formula rather than to construct several formulas for each subgroup. In this study, a set of dummy variables, machine-dependent factors, is defined that represent the adopted ultrasound systems to apply the constructed CAC system to the cases acquired from various ultrasound systems. The machine-dependent factor M_i of case c is defined as

$$M_i(c) = \begin{cases} 1 \text{ if } c \text{ is acquired from machine } i \\ 0 \text{ otherwise} \end{cases}, \quad i = 1, 2, \ldots, n \quad (13)$$

where the experimental cases were collected from n

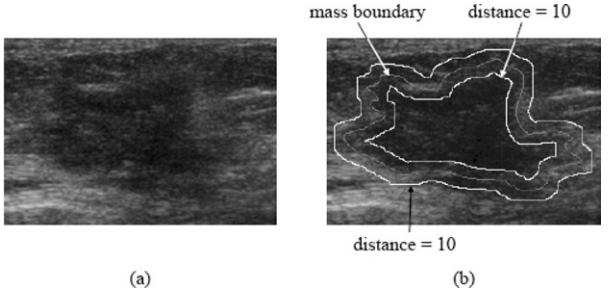

Fig. 7. The contour line can be used to find an outside and an inside band around the mass boundary. The degree of sharp demarcation on lesion boundary could be estimated by the difference between the average gray levels of these two bands.

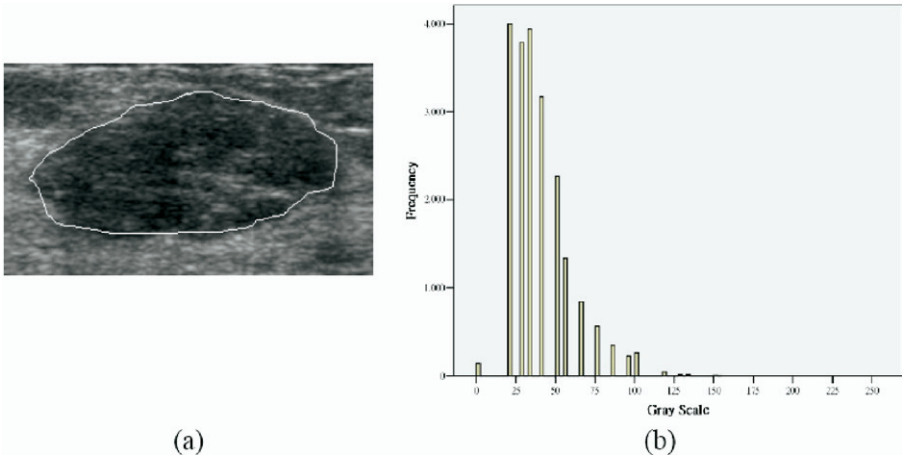

Fig. 8. The frequency of gray intensities within the mass is represented by the histogram. (a) A benign mass. (b) The histogram.

machines and the zero represents "not acquired from machine *i*". Using the reference cell coding (Hosmer and Lemeshow 2000), one of the adopted ultrasound systems was regarded as the reference group and its corresponding factor was omitted. The cases acquired from the reference group were then indicated by setting the remaining n-1 factors to zero. In this approach, the choice of the reference group is decided by the logistic regression model during the construction of the classification system. Thus, the occurrence probability P for a desired output could be predicted as:

$$\text{logit}(p) = c_0 + c_1 f_1 + c_2 f_2 + \ldots + c_m f_m + c_{m+1} M_1 + c_{m+2} M_2 + \ldots + c_{m+n-1} M_{n-1} \quad (14)$$

where c_i are the weighting coefficients for the corresponding m computerized BI-RADS features and n-1 machine-dependent factors, and machine n is regarded as the reference group.

The occurrence probability for any case acquired from machine n, *i.e.*, the reference group, can be predicted from a simplified function of eqn 14 as

$$\text{logit}(p) = c_0 + c_1 f_1 + c_2 f_2 + \ldots + c_m f_m \quad (15)$$

because $M_i = 0 \; for \; i = 1,2,\ldots,n-1$. Moreover, this function can be regarded as a baseline for estimating the occurrence probability of a case acquired from another machine. Compared with this simplified function, for any case acquired from machine k, $k \neq n$, the occurrence probability is predicted as

$$\text{logit}(p) = c_0 + c_1 f_1 + c_2 f_2 + \ldots + c_m f_m + c_{m+k}. \quad (16)$$

Evidently, this machine-specific occurrence probability for machine k is decided by the features in eqn 15 and by the extra weight c_{m+k}. This weight can be regarded as a treatment for adjusting the differences of computerized BI-RADS features of the machine k and those of the reference machine. The evaluations of features in different machines are then synchronized by these machine-dependent factors. Furthermore, the classification system is simply expended by addition of factors when new machines are adopted.

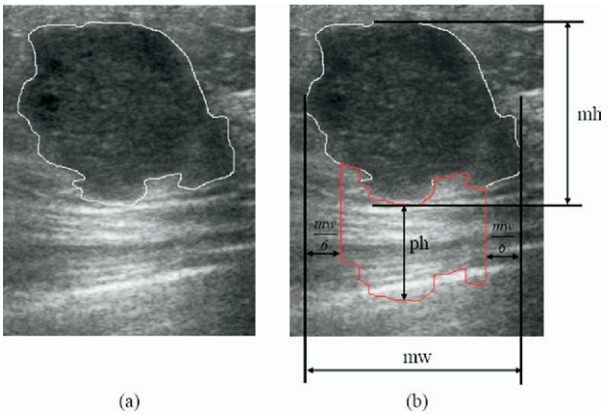

Fig. 9. The posterior acoustic area is defined relative to the shape of mass. (a) A malignant mass with width *mw* and height *mh*. (b) The defined posterior acoustic area with height *ph*.

Weighting strategy

Using the logistic regression model, a classification system is constructed mathematically by relating the probability of assessment categories on the computerized BI-RADS features and the machine-dependent factors. When the classification results of radiologists are used to train the regression model, a basic CAC system is constructed. For improving the potential of the basic CAC

Table 2. The mean values and standard deviations (mean ± SD) in the benign and malignant groups of pathological results and in the Categories 3, 4, and 5 of classification results of radiologists on each proposed feature

	Pathology		Radiologists' BI-RADS classification		
Feature	Benign	Malignant	Category 3	Category 4	Category 5
S_b	31.23 ± 15.71	64.98 ± 25.91	22.30 ± 9.38	45.67 ± 18.56	72.38 ± 27.37
S_{PR}	1.01 ± 0.08	1.16 ± 0.16	1.01 ± 0.09	1.04 ± 0.12	1.19 ± 0.16
O_E	11.57 ± 12.32	19.68 ± 20.74	10.53 ± 11.43	15.26 ± 16.98	19.05 ± 19.70
M_{UA}	4.05 ± 1.62	8.42 ± 2.98	3.43 ± 1.36	5.49 ± 2.17	9.47 ± 3.05
LB_D	23.94 ± 10.39	13.03 ± 6.61	28.39 ± 10.26	17.42 ± 8.49	12.36 ± 5.09
EP_I	47.84 ± 14.51	39.09 ± 13.52	49.51 ± 15.68	45.27 ± 14.03	35.50 ± 9.86
EP_C	0.46 ± 0.20	0.57 ± 0.26	0.45 ± 0.21	0.50 ± 0.21	0.59 ± 0.27
PS_D	43.19 ± 22.59	26.63 ± 27.11	45.69 ± 22.83	36.87 ± 25.11	24.44 ± 25.33

system for distinguishing the benign from malignant masses, the pathologic information is used to help the training process. Intuitively, the cases were classified to category 3 by radiologists because of the typical characteristics of benign masses, whereas the cases in category 5 have typical characteristics of malignant masses. Therefore, three possible strategies can be used to weight the cases in category 3, in category 5 or in both. The improvements in the diagnostic performances of the weighted CAC systems on specificity are insignificant when the strategy weighting category 5 or that weighting both categories 3 and 5 is used. However, the strategy weighting only category 3 increase the specificity without reducing the sensitivity. That is, this strategy classifies more benign cases into category 3 thus, reducing the number of unnecessary biopsies for benign cases. Hence, this strategy, weighting only category 3, is used. In the weight strategy, the benign cases are replicated once and the malignant cases are removed.

Experiment methods and performance evaluations

The k-fold cross-validation method (Stone 1974) is used to verify the classification performance of the CAC system. All the adopted cases are randomly partitioned into k subsets with similar size according to the machine type and the pathologic result. Note that the classification results of the radiologists are not used to partition the cases. Each subset is regarded as test set exactly once and classified by a CAC system trained by the other subsets. In, this study, k is set to 10. The agreement between the radiologists and CAC system was evaluated by Cohen's kappa statistic (Landis and Koch 1977). In kappa statistic, the agreement was considered slight if the k value was 0.20 or greater agreement; fair, if the value was in the range 0.21 to 0.40; moderate, if the value was in the range 0.41 to 0.60; substantial, if the value was in the range 0.61 to 0.80; and, almost perfect, if the value was in the range 0.81 to 1.00.

The classification results of CAC system were evaluated by the pathologic results, benign or malignant. At first, the likelihood of malignancy in each category is defined as

$$P(C_i) = \frac{N(C_{iM})}{N(C_i)} \tag{17}$$

where $N(C_i)$ is the number of cases in the category C_i and $N(C_{iM})$ is the number of malignant cases in the category C_i. A good CAC system should have $P(C_5) > P(C_4) > P(C_3)$. Moreover, five indices, accuracy, sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) are used to evaluate the diagnostic performance and are defined as

$$\begin{aligned} &\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN}),\\ &\text{Sensitivity} = \text{TP}/(\text{TP} + \text{FN}),\\ &\text{Specificity} = \text{TN}/(\text{TN} + \text{FP}),\\ &\text{PPV} = \text{TP}/(\text{TP} + \text{FP}), \text{ and}\\ &\text{NPV} = \text{TN}/(\text{TN} + \text{FN}) \end{aligned} \tag{18}$$

where TP is the number of malignant cases correctly

Table 3. Difference between the benign and malignant groups and the intercategory differences, which were both verified by Student's t-test. Before the calculations of Student's t-test, the Levene's test had been used for verifying the equality of variances

	p-value		
Feature	Benign and Malignant	Category 3 and 4	Category 4 and 5
S_b	<0.001	<0.001	<0.001
S_{PR}	<0.001	<0.001	<0.001
O_E	<0.001	<0.001	0.064
M_{UA}	<0.001	<0.001	<0.001
LB_D	<0.001	<0.001	<0.001
EP_I	<0.001	0.002	<0.001
EP_C	<0.001	0.005	<0.001
PS_D	<0.001	<0.001	<0.001

Table 4. The classification results of basic CAC system compared with the results obtained by radiologists

	Computerized classification result							
	Category 3		Category 4		Category 5			
Radiologists	B	M	B	M	B	M	Total	LM
Category 3	177	2	29	2	0	1	211	2.37%
Category 4	42	1	152	70	4	23	292	32.19%
Category 5	0	0	1	37	2	83	123	97.56%
Total	222		291		113		626	
LM	1.35%		37.46%		94.69%			

Note: B is the number of benign cases. M is the number of malignant cases. LM is the likelihood of malignancy.

classified as positive; TN, benign cases correctly classified as negative; FP, benign cases falsely classified as positive and FN, malignant cases falsely classified as negative. The overall diagnostic performance of classification results is measured by the area index of ROC curve (Park et al. 2004). When comparing two area indices, a larger index roughly indicates a better performance. Moreover, the significance of the difference between the two area indices could be evaluated by the z-test. The ROCKIT software (C. Metz, University of Chicago, Chicago, IL, USA) is used in the ROC analysis.

RESULTS AND DISCUSSION

Radiologists' classification result

The total of 626 experimental cases included 219 malignant and 407 benign lesions. Malignant masses included infiltrating ductal carcinoma ($n = 199$), infiltrating lobular carcinoma ($n = 7$) and ductal carcinoma *in situ* (DCIS) ($n = 13$). Benign lesions included fibroadenoma ($n = 253$) and fibrocystic changes ($n = 154$). The age of the patients ranged from 17 to 74 y (mean age, 48 y) and tumor diameter ranged from 0.5 to 4.2 cm (mean size, 1.5 cm). The *in situ* carcinomas were further classified as malignant. According to the data summarized in Table 1, the classification results of radiologists are evaluated as follows. The likelihood of malignancy according to the classification results of radiologists is 2.37% (5 of 211), 32.19% (94 of 292) and 97.56% (120 of 123) for categories 3, 4 and 5, respectively. The performance indices are as follows: accuracy, 67.09% (420 of 626); sensitivity, 97.72% (214 of 219); specificity, 50.61% (206 of 407); PPV, 51.57% (214 of 415); and NPV, 97.63% (206 of 211).

Data analysis of the proposed BI-RADS features

For each proposed computerized BI-RADS feature, the mean values and the standard deviations in the benign and malignant groups of pathologic result and in the categories 3, 4 and 5 of the classification result of radiologists are listed in the Table 2. The distribution of mean values on each proposed feature conforms to the clinical experiences in either increasing or decreasing order, regardless of the pathologic results or the classification results of radiologists. For example, clinically, the mass with a larger orientation angle corresponds to a higher likelihood of malignancy. The mean values of the orientation feature O_E in the benign and malignant groups and in the different classification categories all conform to such expectation. Moreover, the benign masses are statistically different from the malignant

Table 5. The classification results of weighted CAC system compared with the results obtained by radiologists

	Computerized classification result							
	Category 3		Category 4		Category 5			
Radiologists	B	M	B	M	B	M	Total	LM
Category 3	187	2	19	2	0	1	211	2.37%
Category 4	55	2	139	70	4	22	292	32.19%
Category 5	0	0	1	35	2	85	123	97.56%
Total	246		266		114		626	
LM	1.63%		40.23%		94.74%			

Note: B is the number of benign cases. M is the number of malignant cases. LM is the likelihood of malignancy.

Table 6. The performance indices of radiologists and of two proposed CAC systems. For each performance index, the significances of differences between the radiologists and the basic CAC system and between the radiologists and the weighted CAC system are verified using the *p*-value of χ-square statistic

		CAC system			
		Basic		Weighted	
Index	Radiologists	Score	*p*-value	Score	*p*-value
Accuracy	67.09%	69.49%	0.3623	73.00%	0.0224
Sensitivity	97.72%	98.63%	0.4754	98.17%	0.7363
Specificity	50.61%	53.81%	0.3617	59.46%	0.0112
PPV	51.57%	53.47%	0.5864	56.58%	0.1566
NPV	97.63%	98.65%	0.4316	98.37%	0.5684

masses on each computerized BI-RADS feature because their corresponding *P*-values are less than .001 at 95% confidence interval as listed in Table 3. For all the features, the mean differences between categories 3 and 4 are also statistically significant. The cases in the category 3 are also statistically different from those in category 4. The only insignificant feature is the difference in the orientation feature O_E between categories 4 and 5 concerning.

Basic CAC system

The classification results of basic CAC system are summarized and listed in Table 4. The agreement between the basic CAC system and the radiologists are substantially in agreement with the k value = 0.644. The likelihood of malignancy in categories 3, 4 and 5 on the basis of the CAC results are 1.35% (3 of 222), 37.46% (109 of 291) and 94.69% (107 of 113), respectively. The performance indices, accuracy, sensitivity, specificity, PPV and NPV, of the basic CAC system are 69.49% (435 of 626), 98.63% (216 of 219), 53.81% (219 of 407), 53.47% (216 of 404) and 98.65% (219 of 222), respectively. All performance indices showed slight improvement when compared with performance indices that were based on results of radiologists, but the differences were not statistically significant according to the chi-square statistic. The diagnostic capabilities of two classification results are further evaluated by the ROC curves as shown in Fig. 10. The area index A_Z for the basic CAC system is 0.9435 and that for radiologists is 0.9431. The difference between two A_Z values is not significant using the z-test.

Weighted CAC system

The classification results of weighted CAC system are also compared with the classification result of radiologists as summarized in Table 5. The agreement between the weighted CAC system and the radiologists are also in substantial agreement, with the k value = 0.644. The likelihood of malignancy in Categories 3, 4 and 5 based on the results of weighted CAC system is 1.63% (4 of 246), 40.23% (107 of 266) and 94.74% (108 of 114), respectively. The performance indices, accuracy, sensitivity, specificity, PPV and NPV, of the weighted CAC result are 73.00% (457 of 626), 98.17% (215 of 219), 59.46% (242 of 407), 56.58% (215 of 380) and 98.37% (242 of 246), respectively. Compared with the classification results of radiologists, all performance indices of this weighted CAC system clearly improved. Especially, the accuracy and specificity rates increased 5.91% and 8.85%, respectively, and all these improvements are statistically significant, with *P*-values = .0224 and .0112 using the chi-square statistic. Moreover, the high NPV rate of 98.37% indicates that only a few malignant cases are misclassified into category 3. That is, more benign cases are correctly classified into category 3 and the weighted strategy is very helpful. Furthermore, the A_Z of this system is 0.9448 as shown in Fig. 10. Its A_Z is also very close to that of radiologists and their difference is not significant using the z-test. The detailed performance results of the basic and weighted CAC systems are summarized in the Table 6.

Performance comparisons of conventional CAD and weighted CAC systems

A conventional CAD system, which classifies the cases into malignant and benign, is constructed by the

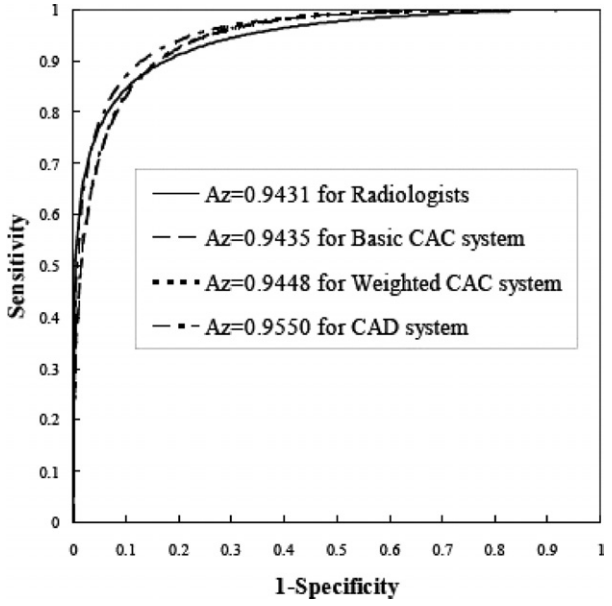

Fig. 10. The ROC curves for radiologists, basic CAC system, weighted CAC system and CAD system.

Table 7. The diagnostic results of the CAD system compared with the weighted CAC system. In the column "Adjustment," the results of category 4 are excluded from the performance evaluation, and in column "Original," the full results are summarized

			Weighted CAC system			
	CAD		Original (C3,C4,C5)		Adjustment (C3,C5)	
Pathology	Benign	Malignant	Benign	Malignant	Benign	Malignant
Benign	TN 368	FP 39	TN 242	FP 165	TN 242	FP 6
Malignant	FN 27	TP 192	FN 4	TP 215	FN 4	TP 108
Total	626		626		360	

logistic regression model and is compared with the proposed weighted CAC system. The A_Z of the CAD system is 0.9550 as shown in Fig. 10. When the diagnosis threshold is set to 0.4 for deciding whether the case is malignant or not, the diagnostic results and the diagnostic performances are listed in Tables 7 and 8.

Following the BI-RADS classifications, the malignant cases in categories 4 and 5 are considered true positive, and the benign cases in category 3 are considered true negative. The diagnostic results of the weighted CAC system are listed under the column "Original" in Tables 7 and 8. The indices sensitivity and NPV of the weighted CAC system, compared with the CAD system, are improved and the other three performance indices are diminished. The weighted CAC system is not better than the conventional CAD system because the indeterminate cases of category 4 are directly considered malignant. In the column "Adjustment" in Tables 7 and 8, the cases of category 4 are excluded from the performance evaluation. When the indeterminate cases were excluded, the diagnostic performances of weighted CAC system are obviously better than those of the CAD system and all improvements are statistically significant. This is the main advantage of the proposed CAC system because the current CAD system always classified the mass into benign and malignant. However, the indeterminate cases are classified into category 4 in the CAC system for further proof rather than giving a definite diagnostic result.

Table 8. The performance indices of CAD and weighted CAC systems are summarized from Table 7. The original and the adjustment results are compared with the CAD system, and the significances of differences are validated using the p-values of χ-square statistic

		Weighted CAC system			
		Original		Adjustment	
Index	CAD	Score	p-value	Score	p-value
Accuracy	89.46%	73.00%	<0.001	97.22%	<0.001
Sensitivity	87.67%	98.17%	<0.001	96.43%	0.0097
Specificity	90.42%	59.46%	<0.001	97.58%	0.0004
PPV	83.12%	56.58%	<0.001	94.74%	0.0026
NPV	93.16%	98.37%	0.0028	98.37%	0.0028

CONCLUSION AND FUTURE STUDIES

The proposed computerized BI-RADS features are proven to effectively quantify the sonographic characteristics of the mass and assist the construction of a CAC system. The differences in the performance indices and in the ROC analysis between the radiologists and the basic CAC system are not statistically significant. Classification of cases by the basic CAC system was similar to that done by radiologists. Because the classification of radiologists is not always correct, a weighted strategy is proposed to help the training of the multinomial logistic regression model, which improves the accuracy and sensitivity of the basic CAC system. Moreover, in clinical practice, the indeterminate cases of category 4 need further follow-up, and this follow-up potential is not provided by current CAD systems. However, the proposed CAC system has this potential to find the indeterminate cases. Furthermore, using the dummy technique, the proposed CAC system is found to be robust, even if the cases were acquired from different ultrasound systems and can be expanded when a new system is introduced.

In the proposed weighted CAC system, the radiologists circumscribe the mass boundary and it classifies only 59.46% (242 of 407) of the benign cases into category 3. In the future, a good automatic segmentation will make the CAC system more useful. Furthermore, increasing the capacities to distinguish the benign from malignant masses, more advanced features could be developed, and other classification methods have to be explored in the future study.

Acknowledgments—This work was supported by National Science Council, Taiwan (Grant NSC 95-2221-E-002-446).

REFERENCES

Arger PH, Sehgal CM, Conant EF, Zuckerman J, Rowling SE, Patton JA. Interreader variability and predictive value of US descriptions of solid breast masses: Pilot study. Acad Radiol 2001;8(4):335–342.

Baker JA, Soo MS. Breast US: Assessment of technical quality and image interpretation. Radiology 2002;223(1):229–238.

American College of Radiology (ACR). Breast imaging reporting and data system. Third Edition. Reston,VA: American College of Radiology, 2003

Buchbinder SS, Leichter IS, Lederman RB, Novak B, Bamberger PN, Sklair-Levy M, Yarmish G, Fields SI. Computer-aided classification of BI-RADS category 3 breast lesions. Radiology 2004;230(3): 820–823.

Chen CM, Chou YH, Han KC, Hung GS, Tiu CM, Chiou HJ, Chiou SY. Breast lesions on sonograms: Computer-aided diagnosis with nearly setting-independent features and artificial neural networks. Radiology 2003;226(2):504–514.

Chen DR, Chang RF, Huang YL. Computer-aided diagnosis applied to US of solid breast nodules by using neural networks. Radiology 1999;213(2):407–412.

Chen DR, Chang RF, Kuo WJ, Chen MC, Huang YL. Diagnosis of breast tumors with sonographic texture analysis using wavelet transform and neural networks. Ultrasound Med Biol 2002;28(10): 1301–1310.

Costantini M, Belli P, Lombardi R, Franceschini G, Mule A, Bonomo L. Characterization of solid breast masses: Use of the sonographic breast imaging reporting and data system lexicon. J Ultrasound Med 2006;25(5):649–659.

Drukker K, Giger ML, Metz CE. Robustness of computerized lesion detection and classification scheme across different breast US platforms. Radiology 2005;237(3):834–840.

Drukker K, Giger ML, Vyborny CJ, Mendelson EB. Computerized detection and classification of cancer on breast ultrasound. Acad Radiol 2004;11(5):526–535.

Eberl MM, Fox CH, Edge SB, Carter CA, Mahoney MC. BI-RADS classification for management of abnormal mammograms. J Am Board Fam Med 2006;19(2):161–164.

Hong AS, Rosen EL, Soo MS, Baker JA. BI-RADS for sonography: Positive and negative predictive values of sonographic features. AJR Am J Roentgenol 2005;184(4):1260–1265.

Horsch K, Giger ML, Venta LA, Vyborny CJ. Computerized diagnosis of breast lesions on ultrasound. Med Phys 2002;29(2):157–164.

Hosmer DW, Lemeshow S. Applied logistic regression. 2nd edition. New York: Wiley, 2000.

Jain AK. Fundamentals of digital image processing. Upper Saddle River, NJ: Prentice-Hall, 1989.

Jemal A, Siegel R, Ward E, Murray T, Xu J, Smigal C, Thun MJ. Cancer statistics, 2006. CA Cancer J Clin 2006;56(2):106–130.

Joo S, Yang YS, Moon WK, Kim HC. Computer-aided diagnosis of solid breast nodules: Use of an artificial neural network based on multiple sonographic features. IEEE Trans Med Imaging 2004; 23(10):1292–1300.

Kim KG, Cho SW, Min SJ, Kim JH, Min BG, Bae KT. Computerized scheme for assessing ultrasonographic features of breast masses. Acad Radiol 2005;12(1):58–66.

Kuo WJ, Chang RF, Moon WK, Lee CC, Chen DR. Computer-aided diagnosis of breast tumors with different US systems. Acad Radiol 2002;9(7):793–799.

Landis JR, Koch GG, The measurement of observer agreement fro categorical data. Biometrics 1977;33:159–174.

Obenauer S, Hermann KP, Grabbe E. Applications and literature review of the BI-RADS classification. Eur Radiol 2005;15(5):1027–1036.

Park SH, Goo JM, Jo CH. Receiver operating characteristic (ROC) curve: Practical review for radiologists. Korean J Radiol 2004;5(1): 11–18.

Parkin DM, Bray F, Ferlay J, Pisani P. Global cancer statistics, 2002. CA Cancer J Clin 2005;55(2):74–108.

Rizzatto GJ. Towards a more sophisticated use of breast ultrasound. Eur Radiol 2001;11(12):2425–2435.

Stavros AT, Thickman D, Rapp CL, Dennis MA, Parker SH, Sisney GA. Solid breast nodules: Use of sonography to distinguish between benign and malignant lesions. Radiology 1995;196(1):123–134.

Stone M. Cross-validatory choice and assessment of statistical predictors. J R Stat Soc, Series B 1974;36(1):111–147.

Zonderland HM, Pope TL, Jr., Nieborg AJ. The positive predictive value of the breast imaging reporting and data system (BI-RADS) as a method of quality assessment in breast imaging in a hospital population. Eur Radiol 2004;14(10):1743–1750.