Ultrasound in Med. & Biol., Vol. ■, No. ■, pp. 1–9, 2015

0301-5629/$ - see front matter

ELSEVIER

http://dx.doi.org/10.1016/j.ultrasmedbio.2015.11.016

● *Original Contribution*

COMPUTER-AIDED DIAGNOSIS FOR BREAST ULTRASOUND USING COMPUTERIZED BI-RADS FEATURES AND MACHINE LEARNING METHODS

Juan Shan,* S. Kaisar Alam,[†‡§] Brian Garra,[¶‖] Yingtao Zhang,[#] and Tahira Ahmed[‖]

*Department of Computer Science, Seidenberg School of Computer Science and Information Systems, Pace University, New York, New York, USA; [†]Improlabs Pte Ltd, Valley Point, Singapore; [‡]Computational Biomedicine Imaging and Modeling Center (CBIM), Rutgers University, Piscataway, New Jersey, USA; [§]Department of Electrical & Electronic Engineering, Islamic University of Technology, Gazipur, Bangladesh; [¶]U.S. Food and Drug Administration, Silver Spring, Maryland, USA; [‖]Washington DC Veterans Affairs Medical Center, Washington, DC, USA; and [#]School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China

(*Received* 27 *May* 2015; *revised* 9 *November* 2015; *in final form* 13 *November* 2015)

Abstract—This work identifies effective computable features from the Breast Imaging Reporting and Data System (BI-RADS), to develop a computer-aided diagnosis (CAD) system for breast ultrasound. Computerized features corresponding to ultrasound BI-RADs categories were designed and tested using a database of 283 pathology-proven benign and malignant lesions. Features were selected based on classification performance using a "bottom-up" approach for different machine learning methods, including decision tree, artificial neural network, random forest and support vector machine. Using 10-fold cross-validation on the database of 283 cases, the highest area under the receiver operating characteristic (ROC) curve (AUC) was 0.84 from a support vector machine with 77.7% overall accuracy; the highest overall accuracy, 78.5%, was from a random forest with the AUC 0.83. Lesion margin and orientation were optimum features common to all of the different machine learning methods. These features can be used in CAD systems to help distinguish benign from worrisome lesions. (E-mail: jshan@pace.edu) © 2015 World Federation for Ultrasound in Medicine & Biology.

Key Words: **Breast cancer, Computer-aided diagnosis, Computerized features, Breast Imaging Reporting and Data System, BI-RADS, Machine learning, Receiver operating characteristic, Tissue characterization, Tumor classification, Ultrasonic imaging.**

INTRODUCTION

Breast cancer is the leading cause of cancer-related death in women (Cheng et al. 2010). In 2014, approximately 232,670 new cases of breast cancer were diagnosed and resulted in approximately 40,000 deaths in the United States (Siegel et al. 2014). Screening mammography is widely used and recommended for the early detection of breast cancer. Studies have indicated that the addition of ultrasound can increase the overall cancer detection rate and reduce the number of unnecessary biopsies (Costantini et al. 2006; Hwang et al. 2005). Screening ultrasound is becoming an important addition to routine breast cancer screening because of its superior ability in imaging dense breast tissue and its lack of ionizing radiation.

Despite its many advantages, however, the quality of ultrasound has been relatively low because of the intrinsic speckle noise and low contrast between different tissue types. Digital image processing techniques and machine learning methods have been applied to improve detection rate and increase specificity (Chen et al. 2003; Huang et al. 2006; Segyeong et al. 2004). Advances in the field of medical image processing has improved the ability of computer-aided diagnosis (CAD) to reduce background noise, improve image contrast, detect regions of interest, differentiate a tumor from background and therefore help differentiate benign from worrisome lesions (Drukker et al. 2006; Moon et al. 2013a; Shen et al. 2007). Among all these functionalities of CAD systems, classifying a tumor into benign or worrisome categories is the ultimate objective.

The performance of machine learning methods relies heavily on how well the characteristics of tumors are represented by digital features, which can be separated into two categories: knowledge based and statistic based.

Address correspondence to: Juan Shan, Department of Computer Science, Seidenberg School of CSIS, Pace University, 163 William Street, New York, NY 10038, USA. E-mail: jshan@pace.edu

Knowledge-based features are derived from the Breast Imaging Reporting and Data System (BI-RADS) lexicon (Mendelson et al. 2013), which is used to characterize lesions based on shape, margin, orientation, echo pattern and acoustic shadowing (Chen et al. 2003; Moon et al. 2013a; Song et al. 2005). The other category of features is obtained from statistical computation, such as auto-covariance coefficients and frequency domain features (Huang and Chen 2005; Mogatadakala et al. 2006). These features capture the correlation between pixels and do not necessarily correspond to any observable features in ultrasound images.

The BI-RADS lexicon aims to standardize mammography and ultrasound reports so that reports are clear, succinct and consistent among readers. Although all BI-RADS terms are descriptive, not quantitative, they need to be "translated" into computerized features so a CAD system can compute these features automatically. Several groups have proposed approaches to quantify BI-RADS features (André et al. 2007; Mainiero et al. 2005; Moon et al. 2013b), including a comprehensive study by Alam et al. (2011). For example, the most commonly used ultrasound BI-RADS feature is the "parallel" orientation, which corresponds to the "long axis of a lesion paralleling the skin line." To quantify this feature, an equivalent ellipse of the lesion was identified, and the ratio between the horizontal axis and the vertical axis of the ellipse was computed (Chen et al. 2004; Moon et al. 2013a; Sahiner 2007). If the ratio is larger than one, the tumor is more likely benign; if the ratio is less than one, it is more likely malignant.

For this study, we performed a complete translation of the entire ultrasound BI-RADS lexicon into digital features, which are used in machine learning methods for the purpose of developing an effective CAD system for breast ultrasound. We have proposed new and validated digital features to distinguish benign from worrisome lesions with the ultimate goal of improving the accuracy of breast cancer diagnosis.

METHODS

The database used in this study contains 283 breast ultrasound images. The images were collected subsequently without excluding any data by the Second Affiliated Hospital of Harbin Medical University (Harbin, China), using a VIVID 7 (GE, Horten, Norway) with a 5- to 14-MHz linear probe. The aperture of the transducer is 4 cm. To obtain the original ultrasound images, the techniques harmonics, spatial compounding and speckle reduction were not used. The average size of the images is 500 × 420 pixels. The tumors range from 0.5 to 6.5 cm, with a median size is 1.1 cm. Among them, 133 cases are benign and 150 cases are malignant. All lesions were validated by ultrasound-guided biopsy using a 14-gauge needle. Informed consent was obtained from all patients in this study. The study protocol was approved by the institutional ethics committee of the university. For each case, the boundary of the lesion was manually delineated by an experienced radiologist, and important findings were categorized into BI-RADS terms. The radiologist is a board-certified attending with more than 30 years of post-residency experience in breast ultrasound.

Computerization of BI-RADS features

In the fifth edition of BI-RADS Ultrasound (Mendelson et al. 2013), the dominant sonographic characteristics are grouped into five descriptive categories: shape, orientation, margin, echo pattern and posterior acoustic features. To quantify these BI-RADS features, multiple computerized features are proposed as discussed below and summarized in Table 1.

Shape. Irregular shape is a characteristic of malignancy. To capture this characteristic, an equivalent ellipse with the same second moments as the tumor region is used. As Figure 1 illustrates, the area difference between the tumor and its equivalent ellipse can describe how irregular the tumor is. This computable feature is called the area difference with equivalent ellipse (ADEE) and is defined as

$$\mathrm{ADEE} = \frac{A_{\mathrm{E}} + A^{\mathrm{T}} - A_{\mathrm{E}\cap\mathrm{T}}}{A^{\mathrm{T}}} \tag{1}$$

where A_{E} is the number of pixels in the equivalent ellipse, A_{T} is the number of pixels in the tumor region and $A_{\mathrm{E}\cap\mathrm{T}}$ is the number of pixels in the region where the tumor and the ellipse intersect.

Orientation. Orientation describes the direction of long axis of the tumor. If the long axis of the tumor parallels the skin line, the orientation is parallel, or "wider than tall"; otherwise, the orientation is anti-parallel, or "taller than wide." "Taller than wide" is a worrisome feature because malignant tumors have less compressibility and can grow across tissue planes. To quantify this feature, a minimum bounding box that covers the entire tumor is used. The edges of the bounding box should be parallel to the image boundaries. The ratio between the height and width of the box can characterize orientation as

$$R_{\text{Taller than wide}} = \frac{\text{height}}{\text{width}} \tag{2}$$

When the "taller than wide" ratio is larger than one, the tumor is likely to be malignant; otherwise, it is a sign of benignity.

Table 1. Summary of the proposed features that quantify each BI-RADS category

BI-RADS category	Feature formula	Description
Shape	$\text{ADEE} = \frac{A_{\text{E}} + A_{\text{T}} - A_{\text{E}\cap\text{T}}}{A_{\text{T}}}$	The more irregular the mass shape, the larger is the area difference.
Orientation	$R_{\text{Taller than wide}} = \frac{\text{height}}{\text{width}}$	When the "taller than wide" ratio is >1, the tumor is likely to be malignant.
Margin	$\text{AvgDiff} = \frac{\sum_{i\in\text{OUT}} \text{Diff(i)}}{\text{OUT}_N}$	Average intensity difference between inside and outside contours: A larger value indicates a less indistinct margin.
Margin	$\text{NumberPeaks} = \text{number of local maxima of } V_{\text{convex}}$	Malignancy feature: A larger number of peaks means the contour is bumpier.
Margin	$\text{AvgDistance} = \text{Average of } V_{\text{convex}}$	A larger value indicates a spiculated contour.
Margin	$ADCH = \frac{A_c - A_T}{A_T}$	Another perspective to capture and quantify the smoothness of the margin.
Echo pattern	$\text{Echogenicity} = \text{AvgIntensity}_{\text{surrounding}} - \text{AvgIntensity}_{\text{tumor}}$	Positive echogenicity indicates the tumor is hypo-echoic, whereas negative echogenicity indicates the tumor is hyper-echoic.
Echo pattern	$\text{Entropy} = -\sum_i P_i \log_2 P_i$	A large entropy indicates a heterogeneous tumor, whereas a small entropy indicates homogeneity.
Posterior feature	$\text{Shadow} = \overline{I_{\text{post}}} - \overline{I_{\text{tumor}}}$	Shadow is a feature that indicates malignancy.

Margin. Marginal characteristics are an important BI-RADS category in assessing the likelihood of malignancy. This BI-RADS category contains four subcategories focused on different characteristics of the tumor margin, namely: "indistinct," "angular," "microlobulated" and "spiculated," which are worrisome features. If none is present, the margin is described as circumscribed. Among the four subcategories, only indistinct margin is based on intensity level of pixels around the margin; the other three features are based on morphologic characteristics. We discuss the computerization of indistinct margin first.

Indistinct margin is defined as no clear demarcation between a mass and its surrounding tissue. To quantify this feature, the tumor contour shrinks into a smaller inside contour and enlarges to a larger outside contour, respectively. The inside, outside and the original contours are all delineated in the image. As Figure 2 illustrates, the original manual delineated contour is *white*, the inside contour is *red* and the outside contour is *blue*. The *yellow lines* represent three segments on which intensity difference vector Diff is computed. Each segment starts from a pixel on the outside contour and ends up at the closest pixel on the inside contour. If there is a sharp demarcation between the mass and surrounding tissue, the intensity difference between the average of the first half and the average of the second half of the segment should be relatively large, and vice versa. Figure 3 illustrates the two situations.

Mathematically, the Diff vector is defined as drawing the outside contour and inside contour along the tumor contour with a 20-pixel width on each side. For every pixel i on the outside contour,

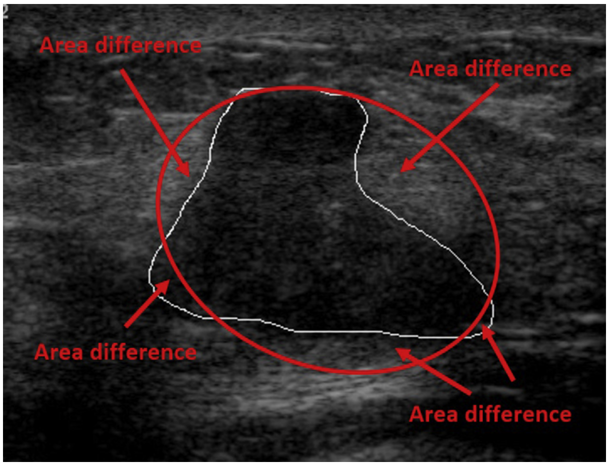

Fig. 1. Area difference between a breast tumor and its equivalent ellipse.

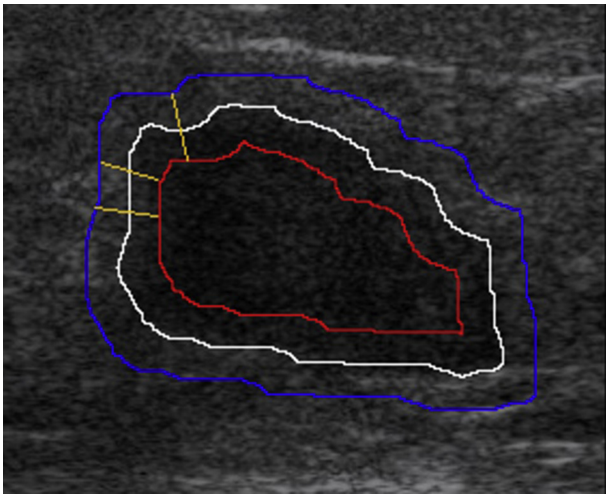

Fig. 2. Original contour, inside contour and outside contour of a lesion. Line segments connecting the outside contour and inside contour are used to compute the difference vector.

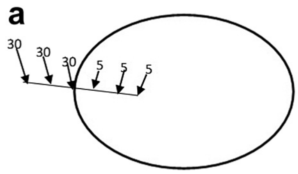

Fig. 3. Examples indicate how a value in the Diff vector is computed. (a) A distinct margin case: The difference in intensity between the average of the first half and average of the second half of the segment is 25. (b) Indistinct margin case: The difference in intensity between the average of the first half and the average of the second half of the segment is 15.

$$\mathrm{Diff}(i)=\overline{I_{\mathrm{out}}\ (i)}-\overline{I_{\mathrm{in}}(j)} \tag{3}$$

where i is the ith pixel on the outside blue contour, and j is the closest pixel to i on the inside red contour. $\overline{I_{\mathrm{out}}\ (i)}$ is the average intensity level of pixels on the outside half of the line segment ij, and $\overline{I_{\mathrm{in}}(j)}$ is the average intensity level of pixels on the other half of the line segment ij.

The computerized indistinct margin feature can be represented by the average of vector Diff, that is,

$$\mathrm{AvgDiff}=\frac{\sum_{i\in \mathrm{OUT}}\mathrm{Diff}(i)}{\mathrm{OUT}_N} \tag{4}$$

where i is a pixel on the outside contour, and N is the number of pixels on the outside contour.

The other margin features—“angular,” “microlobulated” and “spiculated”—focus on the smoothness of the contour. Their common characteristic is that some part of the margin extends away from the tumor body, with either a sharp angle, rounded microlobules or projected spicules. A digital feature is proposed to capture the common characteristic of these irregular shapes. A convex hull of the tumor is drawn, and a distance vector between the tumor contour and its convex hull is computed. For every pixel on the convex hull, its distance to the closest point on the tumor contour is saved in the distance vector V_{convex}:

For every pixel i on the tumor contour,

$$V_{\mathrm{convex}}(i)=\sqrt{\left(x_i-x_j\right)^2+\left(y_i-y_j\right)^2} \tag{5}$$

where i is the ith pixel on the convex hull, and j is the closest pixel to i on the tumor contour. x and y are the coordinates of the pixels.

Three computable features are extracted from the distance vector to describe an irregular margin: number of peaks on the distance vector (NumPeaks), average of the distance vector (AvgDistance) and area difference between the convex hull and tumor (ADCH). As Figure 4 illustrates, the numbers of peaks on the distance vector correspond to the number of valleys on the tumor contour, which are marked by *red stars*. The average of distance vector and area difference between the two contours can also depict the irregularity of the tumor margin, from different perspectives. The three digital features are defined in formulas as

$$\mathrm{NumPeaks}=\text{Number of local maxima of }V_{\mathrm{convex}} \tag{6}$$

$$\mathrm{AvgDistance}=\text{Average of }V_{\mathrm{convex}} \tag{7}$$

$$\mathrm{ADCH}=\frac{A_{\mathrm{c}}-A_{\mathrm{T}}}{A_{\mathrm{T}}} \tag{8}$$

where V_{convex} is the distance vector between the tumor boundary and the corresponding convex hull defined in Eq. (5), A_{C} is the number of pixels within the convex hull and A_{T} is the number of pixels within the tumor.

Echo pattern. Echo pattern is a feature defined in relation to fat. Tissue darker than fat is called “hypo-echoic,” and tissue lighter than the fat is called “hyper-echoic.” A complex echo pattern contains both hypo-echoic and hyper-echoic tissues. Hypo-echoic (darker than fat) tissue is a worrisome finding. Because it is defined relative to fat, and the intensity of fat is relative to the dynamic range of the entire image, it is hard to find a fixed intensity threshold to identify hypo-echoic or hyper-echoic. Instead, a dynamic computable feature is proposed to capture this feature. The average intensity of surrounding tissues should provide a good reference to describe the degree of hyper-echogenicity. So the intensity difference between the surrounding area and the tumor area is used to capture the degree of hyper-echogenicity:

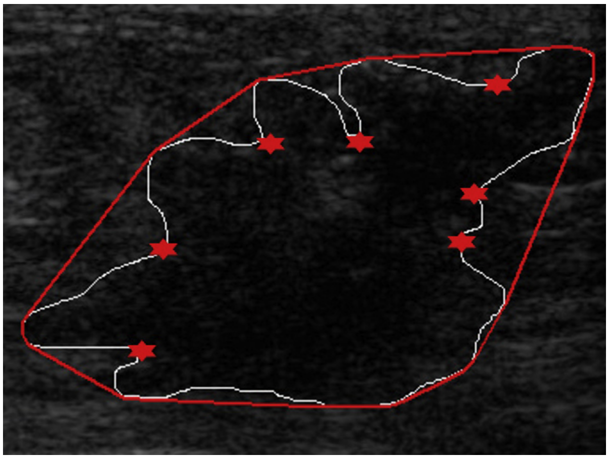

Fig. 4. Convex hull (*red contour*) of a malignant lesion (*white contour*), with peaks on the distance vector V_{convex} marked by *stars*.

$$\text{Echogenicity} = \text{AvgIntensity}_{\text{surrounding}} - \text{AvgIntensity}_{\text{tumor}} \tag{9}$$

The surrounding region should be a rectangular region that contains the tumor in its center and is about twice the size of the tumor. It should exclude shadow areas from the surrounding region to provide an accurate reference for average intensity. A large positive echogenicity indicates the tumor is hypo-echoic, whereas a negative echogenicity indicates the tumor is hyper-echoic.

In addition to hypo-echogenicity, the heterogeneous or complex echo pattern is also considered a worrisome finding. The heterogeneous echo pattern is a combination of darker and lighter components, so the entropy of pixel intensities within a heterogeneous tumor should be larger than the entropy of pixel intensities in a homogeneous tumor. The entropy feature is proposed to describe the degree of heterogeneity:

$$\text{Entropy} = -\sum_i P_i \log_2 P_i \tag{10}$$

Here, P_i is the probability that the intensity difference between two adjacent pixels is equal to i.

Posterior features. Acoustic shadowing is considered a hard finding that is worrisome for malignancy. Shadows are dark areas that appear immediately posterior to the tumors with decreasing or increasing shadow effect. Some tumors have complete posterior shadows, some have partial posterior shadows depending on the degree of desmoplasia of the tumor and some do not have shadows at all. To capture the shadowing feature, a rectangular region below the tumor is analyzed. The average intensity of this rectangular region is compared with the average intensity of the tumor. A negative or close-to-zero difference indicates the presentation of shadow, whereas a positive difference indicates no shadow.

$$\text{Shadow} = \overline{I_{\text{post}}} - \overline{I_{\text{tumor}}} \tag{11}$$

Here, $\overline{I_{\text{post}}}$ is the average intensity level of the rectangular region below the tumor and with similar size to the tumor.

Lesion size. Lesion size is not a BI-RADS feature. However, the fifth edition of BI-RADS Ultrasound (Mendelson et al. 2013) does mention that lesion size should be given to report important findings. For automatic tumor diagnosis, lesion size might be combined with other features to improve the performance of tumor classifiers. The number of pixels within the tumor contour could be used to represent lesion size.

Machine learning methods

Four machine learning methods were employed to distinguish benign from malignant tumors using the computerized features. For each machine learning method, feature selection was carried out to identify the optimum feature set using 10-fold cross-validation. The machine learning methods are briefly introduced as follows.

Decision tree. A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences. It is a rule-based decision tool. Decision trees are widely used in the field of pattern recognition, with an efficient training procedure and model construction. The decision tree algorithm implemented in the Weka Package was used (Hall et al. 2009) to test the classification performance of the proposed features.

Artificial neural network. Artificial neural network (ANN) is a self-learning method that imitates the properties of biological nervous systems and the functions of adaptive biological learning. An ANN is composed of an input layer, an output layer and one or more hidden layers. In this work, a single hidden layer with three neurons was employed as the network structure. The backpropagation algorithm is used to update the weights of neurons. The implementation of the backpropagation network in Weka Package (Hall et al. 2009) was used.

Support vector machine. The support vector machine (SVM) (Vapnik 1998) is a classification technique that seeks an optimal hyperplane to separate two classes of samples. The SVM has been reported to be a superior method in many classification problems. The software package SVMLIGHT (Joachims 1999) was adopted. 10-fold cross-validation was used to evaluate the performance of different feature combinations. The radial basis function (RBF) kernel with default parameter setting was chosen.

Random forest. Random forest is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the overall prediction of the individual trees. Random forests correct the overfitting problem of decision trees. The

Table 2. Means and SDs of the benign and malignant groups, with t-scores and *p* values obtained from Student's *t*-test

Feature	Benign	Malignant	*t*-Score	*p* Value*
f1. ADEE (shape)	0.191 ± 0.191†	0.324 ± 0.178	6.04	<0.00001
f2. Height/width (orientation)	0.417 ± 0.176	0.467 ± 0.169	2.64	0.009
f3. AvgDiff (margin)	0.508 ± 0.149	0.437 ± 0.141	4.10	0.00005
f4. NumPeaks (margin)	0.117 ± 0.128	0.281 ± 0.159	9.48	<0.00001
f5. AvgPeaks (margin)	0.198 ± 0.171	0.353 ± 0.165	7.74	<0.00001
f6. ADCH (margin)	0.133 ± 0.152	0.267 ± 0.164	7.13	<0.00001
f7: Echogenicity (echo pattern)	0.564 ± 0.169	0.519 ± 0.157	2.31	0.022
f8: Entropy (echo pattern)	0.416 ± 0.200	0.486 ± 0.168	3.17	0.002
f9: Shadow (posterior feature)	0.473 ± 0.216	0.426 ± 0.205	1.87	0.063
f10: Lesion size	0.242 ± 0.204	0.295 ± 0.174	2.34	0.020

* Two-tailed *t*-test.

† Mean ± standard deviation.

implementation of the random forest algorithm in the Weka Package (Hall et al. 2009) was used in this work.

Performance evaluation

Each computerized feature is evaluated separately using Student's *t*-test first. Then their combined prediction abilities on each of the machine learning methods are tested. A bottom-up feature selection approach is employed to find the feature combination that can give the best performance on each machine learning method. 10-fold cross-validation is carried out on the entire database to train and test the classifiers. Pathologic results are referred to as gold standards. Measures, including accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), Matthew's correlation coefficient (MCC) and ROC curve analysis, are used to evaluate the classification performance from different perspectives.

$$\mathrm{MCC} = (\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN})\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})} \tag{12}$$

Here, TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives and FN is the number of false negatives. MCC has been widely used as a performance measure for predicting models. Especially when the numbers of negative samples and positive samples are unequal, MCC gives a better evaluation than overall accuracy. Moreover, ROC curve analysis is also used to evaluate the CAD system. The MATLAB function PERFCURVE is used to carry out ROC curve analysis.

RESULTS

Student's t*-test for computerized BI-RADS features*

The mean value and standard deviation of each computerized BI-RADS feature for the benign and malignant groups are listed in Table 2. According to Student's *t*-test, seven features differed statistically between the benign and malignant groups, at significance level of 0.01 ($p < 0.01$): (f1) ADEE, (f2) height/width, (f3) AvgDiff, (f4) NumPeaks, (f5) AvgPeaks, (f6) ADCH and (f8) entropy. The other three features did not significantly differ at level 0.01, including (f7) echogenicity, (f9) shadow and (f10) lesion size. The analysis reveals that most of the proposed digital features have a strong ability to distinguish benign and malignant tumors, especially f4, f5 and f6 which are quantified from the margin category. The discriminating ability of combined features is tested in the next step.

Table 3. Performance of different feature combinations using decision tree

Iteration	Feature	Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	MCC	AUC (%)
1	{4}	67.8	76.0	58.7	67.5	68.4	0.353	68.7
2	{4, 3}	76.3	76.7	75.9	78.2	74.3	0.526	75.9
3	{4, 3, 10}	77.0	74.0	80.5	81.0	73.3	0.544	76.2
4	{4, 3, 10, 2}	76.7	73.3	80.5	80.9	72.8	0.537	79.6
5	{4, 3, 10, 2, 7}	77.7	74.0	82.0	82.2	73.7	0.559	80.3

PPV = positive predictive value; NPV = negative predictive value; MCC = Matthew's correlation coefficient; AUC = area under the ROC curve.

Table 4. Performance of different feature combinations using artificial neural network

Iteration	Feature	Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	MCC	AUC (%)
1	{4}	71.4	80.7	60.9	69.9	73.6	0.426	77.2
2	{4, 3}	72.4	77.3	66.9	72.5	72.4	0.446	77.3
3	{4, 3, 2}	74.9	74.7	75.2	77.2	72.5	0.498	78.6
4	{4, 3, 2, 6}	76.7	78.0	75.2	78.0	75.2	0.532	81.3
5	{4, 3, 2, 6, 5}	77.4	78.0	76.7	79.1	75.6	0.547	81.9
6	{4, 3, 2, 6, 5, 1}	78.1	78.0	78.2	80.1	75.9	0.561	82.3

PPV = positive predictive value; NPV = negative predictive value; MCC = Matthew's correlation coefficient; AUC = area under the ROC curve.

Feature selection: Bottom-up searching procedure

When the feature vector has multiple dimensions, less significant features might also contribute to the performance of the classifier when combined with other features. A bottom-up feature selection procedure is carried out to search the optimum feature set that performs best in classifying tumors. The feature pool is composed of the 10 digital features (as outlined in Table 2). The feature that gives the best performance is chosen first. Then additional features are added incrementally. At each step, the feature that can give the most improvement is chosen. This procedure is continued until addition of features decreases performance or all features have been included.

As mentioned earlier, four commonly used machine learning methods are employed. The initial assumption is that different machine learning methods will have different optimum feature sets. So bottom-up feature selection is carried out on each of these methods. The experimental results confirm this assumption. Tables 3–6 outline the feature selection procedure on each machine learning method. When a single feature is used, feature 4 (numPeaks) generates the best AUC performance on decision tree, ANN and SVM. The best single feature on random forest is feature 6 (ADCH), which is also derived from BI-RADS margin feature category as feature 4. This result indicates that margin features have a strong ability to distinguish benign from malignant tumors when used with machine learning methods. As the procedure continues, different features are added into each optimum feature set.

Two observations are important: First, different machine learning methods end up with different optimum feature sets. In other words, the optimum feature set is dependent on the classifier, and when the machine learning method of a CAD system is changed, the feature selection procedure should be redone to achieve the best performance on a particular machine learning method. Second, there are common features of the four optimum feature sets. Besides the best single feature 4 and feature 6, feature 2 (orientation) and feature 3 (indistinct margin) are included in every optimum feature set, which indicates that the distinguishing ability of these two features is independent of machine learning methods.

In the comparison of performance using multiple evaluation metrics, the question arises as to which metric should be considered as the overall metric, especially when different metrics indicate different trends. Multiple evaluation formulas are used in this study, and they provide evaluation from different perspectives. Here the authors chose the AUC (area under the ROC curve) since AUC provides an evaluation of the aggregated classification performance over the entire false positive rate range, and this metric is commonly used in evaluating CAD for breast cancer.

Finally, the best performance of each machine learning method using its own optimum feature set is compared in Table 7. Table 7 indicates that among the four machine learning methods, SVM achieved the best ROC performance, with an AUC of 84.2%, accuracy of 77.7%, sensitivity of 77.3% and specificity of 78.2%. Decision tree, ANN and random forest had AUCs of 80.3%, 82.3% and 83.7 respectively. However, when the alternate overall metric MCC was considered, the best MCC (0.572) was achieved by random forest with five features (f6, f10, f2, f9, f3). This feature set on random forest also gives the best accuracy, 78.5%, among

Table 5. Performance of different feature combinations using random forest

Iteration	Feature	Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	MCC	AUC (%)
1	{6}	70.3	74.0	66.2	71.2	69.3	0.403	74.7
2	{6, 10}	72.4	70.7	74.4	75.7	69.2	0.450	78.2
3	{6, 10, 2}	74.9	76.0	73.7	76.5	73.1	0.500	80.0
4	{6, 10, 2, 9}	74.9	74.7	75.2	77.2	72.5	0.498	82.7
5	{6, 10, 2, 9, 3}	78.5	75.3	82.0	82.5	74.7	0.572	82.8

AUC = area under the ROC curve.

Table 6. Performance of different feature combinations using support vector machine

Iteration	Feature	Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	MCC	AUC (%)
1	{4}	72.1	72.0	72.2	74.5	69.6	0.441	79.2
2	{4, 2}	73.9	73.3	74.4	76.4	71.2	0.477	80.4
3	{4, 2, 6}	75.6	78.0	72.9	76.5	74.6	0.510	83.2
4	{4, 2, 6, 3}	76.0	75.3	76.7	78.5	73.4	0.519	83.8
5	{4, 2, 6, 3, 10}	77.7	77.3	78.2	80.0	75.4	0.555	84.2

PPV = positive predictive value; NPV = negative predictive value; MCC = Matthew's correlation coefficient; AUC = area under the ROC curve.

all the other machine learning methods. Random forest and SVM both had superior classification ability in this tumor diagnosis task. Finally, ROC curves of different machine learning methods using their optimum feature sets are plotted in Figure 5.

According to performance achieved in this research, the proposed algorithms could be applied clinically to help radiologists detect cancers more accurately. The algorithms can be simply installed on radiologists' computers where digital ultrasound images are stored, or integrated into any existing ultrasound CAD system to complete the analysis. Intermediate results such as marking the peaks on the tumor contour could be displayed to emphasize the important features; classification results of the machine learning methods could provide guidance to help radiologists make the final decision.

DISCUSSION AND CONCLUSIONS

Computer-aided diagnosis for breast ultrasound is a field that has been extensively studied. A crucial task for a CAD system is discovering efficient computerized features to distinguish benign and malignant tumors. The fifth edition of the Ultrasound BI-RADS lexicon (Mendelson et al. 2013) was quantified into computerized features and evaluated using Student's *t*-test. Multiple features were combined to serve as input for machine learning methods, and the bottom-up feature selection procedure was used to find the optimum feature set. Four machine learning methods were employed as classifiers, and their performance was compared on the same database. The optimum feature set is separately reported.

Experimental results indicated that the digital features derived from margin categories (numPeaks and ADCH) have a strong ability to separate benign from worrisome lesions. Features derived from indistinct margin (AvgDiff) and orientation (Height/Weight) were contained in the optimum feature sets no matter which machine learning method was used. The distinguishing ability of these two features is independent of classifiers.

In addition to the effective features, the experimental results also indicated that different machine learning methods have different optimum feature sets, which means that feature selection should be conducted separately for each machine learning method to achieve the best classification performance using that method. Among the four machine learning methods, SVM achieved the highest AUC (84.2%) with five features. Ensemble classifier random forest performed better than single decision tree, which indicates better performance of clustered classifiers in a tumor classification task.

Future work could include, first, assembling other types of classifiers, such as SVM, into a cluster of classifiers to improve decision accuracy. Second, adding non-BI-RADS features, such as the statistical feature auto-covariance coefficients to the feature pool could provide more information for machine learning methods to improve classification performance. Third, the currently active research area deep learning provides a mechanism to extract features automatically through a self-learning network. Given that "good" features are the key to boosting classification accuracy, use of deep learning could be a promising new direction to obtain powerful features for automatic breast tumor classification. Finally, testing the robustness of the algorithm using images from

Table 7. Performance of different machine learning methods

Method	Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	MCC	AUC (%)
Decision tree	77.7	74.0	82.0	82.2	73.7	0.559	80.3
Artificial neural network	78.1	78.0	78.2	80.1	75.9	0.561	82.3
Random forest	78.5	75.3	82.0	82.5	74.7	0.572	82.8
Support vector machine	77.7	77.3	78.2	80.0	75.4	0.555	84.2

PPV = positive predictive value; NPV = negative predictive value; MCC = Matthew's correlation coefficient; AUC = area under the ROC curve.

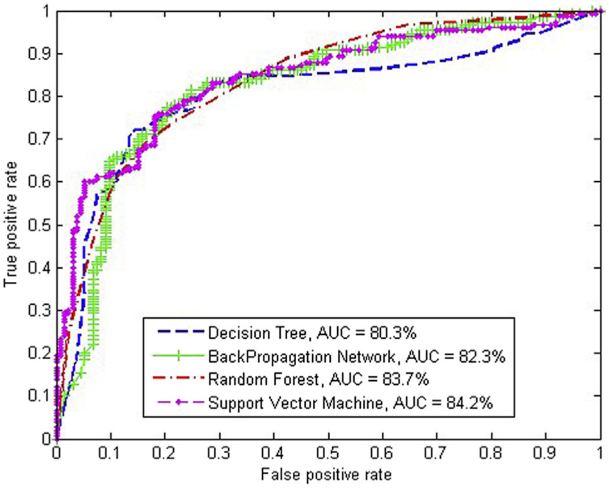

Fig. 5. Receiver operating characteristic curves of different machine learning methods. AUC = area under the ROC curve.

different systems with different acquisition parameters would also be a meaningful work.

Acknowledgments—Thanks are due to radiologists Dr. Jiawei Tian and Dr. Yanxin Su from the Second Affiliated Hospital of Harbin Medical University (China) for their efforts in collecting the images and labeling the database.

REFERENCES

Alam SK, Feleppa EJ, Rondeau M, Kalisz A, Garra BS. Ultrasonic multi-feature analysis procedure for computer-aided diagnosis of solid breast lesions. Ultrason Imaging 2011;33:17–38.

André M, Galperin M, Contro G, Omid N, Olson L, Comstock C, Richman K, O'Boyle M. Diagnostic performance of a computer-aided image analysis system for breast ultrasound. Acoust Imaging 2007;28:341–348.

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;226:504–514.

Chen S, Cheung Y, Su C, Chen M, Hwang T, Hsueh S. Analysis of sonographic features for the differentiation of benign and malignant breast tumors of different sizes. Ultrasound Med Biol 2004;23: 188–193.

Cheng HD, Shan J, Ju W, Guo YH, Zhang L. Automated breast cancer detection and classification using ultrasound images: A survey. Pattern Recog 2010;43:299–317.

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; 25:649–659.

Drukker K, Giger M, Metz C. Robustness of computerized lesion detection and classification scheme across different breast US platforms. Radiology 2006;238:834–840.

Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: An update. SIGKDD Explorations 2009;11.

Huang YL, Chen DR. Support vector machines in sonography: Application to decision making in the diagnosis of breast cancer. Clin Imaging 2005;29:179–184.

Huang Y, Wang K, Chen D. Diagnosis of breast tumors with ultrasonic texture analysis using support vector machines. Neural Comput Appl 2006;15:164–169.

Hwang KH, Lee JG, Kim JH, Lee HJ, Om KS, Yoon M, Choe W. Computer aided diagnosis (CAD) of breast mass on ultrasonography and scintimammography. In: Proceedings, Seventh International Workshop on Enterprise Networking and Computing in Healthcare Industry, HEALTHCOM 2005. New York: IEEE; 2005. p. 187–189.

Joachims T. Making large-scale SVM learning practical. In: Schölkopf B, Burges C, Smola A, (eds). Advances in kernel methods: Support vector learning. Cambridge: MIT Press; 1999.

Mainiero M, Goldkamp A, Lazarus E, Livingston L, Koelliker S, Schepps B, Mayo-Smith W. Characterization of breast masses with sonography: Can biopsy of some solid masses be deferred? J Ultrasound Med 2005;24:161–167.

Mendelson EB, Böhm-Vélez M, Berg WA, *et al*. ACR BI-RADS® Ultrasound. In: ACR BI-RADS® Atlas, Breast Imaging Reporting and Data System. Reston, VA: American College of Radiology; 2013.

Mogatadakala K, Donohue K, Piccoli C, Forsberg F. Detection of breast lesion regions in ultrasound images using wavelets and order statistics. Med Phys 2006;33:840–849.

Moon W, Lo CM, Cho N, Chang JM, Huang CS, Chen JH, Chang RF. Quantitative ultrasound analysis for classification of BI-RADS category 3 breast masses. J Digit Imaging 2013a;26:1091–1098.

Moon W, Lo CM, Cho N, Chang JM, Huang CS, Chen JH, Chang RF. Computer-aided diagnosis of breast masses using quantified BI-RADS findings. Comput Methods Programs Biomed 2013b;111: 84–92.

Sahiner B. Malignant and benign breast masses on 3-D US volumetric images: Effect of computer-aided diagnosis on radiologist accuracy. Radiology 2007;242:716–724.

Segyeong J, Yoon SY, Woo KM, Hee CK. Computer-aided diagnosis of solid breast nodules: Use of an artificial neural network based on multiple sonographic features. IEEE Trans Med Imaging 2004;23: 1292–1300.

Shen W, Chang R, Moon W, Chou Y, Huang C. Breast ultrasound computer-aided diagnosis using BI-RADS features. Acad Radiol 2007;14:928–939.

Siegel R, Ma J, Zou Z, Jemal A. Cancer statistics 2014. CA Cancer J Clin 2014;64:9–29.

Song JH, Venkatesh SS, Md EFC, Cary TW, Md PH. Artificial neural network to aid differentiation of malignant and benign breast masses by ultrasound imaging. Proc SPIE 2005;5750:148–152.

Vapnik V. Statistical learning theory. New York: Springer-Verlag; 1998.