A similarity study of content-based image retrieval system for breast cancer using decision tree

Hyun-chong Cho and Lubomir Hadjiiski[a)]
Department of Radiology, The University of Michigan, Ann Arbor, Michigan 48109-0904

Berkman Sahiner
U.S. Food and Drug Administration, 10903 New Hampshire Avenue, Silver Spring, Maryland 20993

Heang-Ping Chan, Mark Helvie, Chintana Paramagul, and Alexis V. Nees
Department of Radiology, The University of Michigan, Ann Arbor, Michigan 48109-0904

(Received 11 August 2012; revised 15 November 2012; accepted for publication 16 November 2012; published 19 December 2012)

Purpose: We are developing a decision tree content-based image retrieval (DTCBIR) CADx system to assist radiologists in characterization of breast masses on ultrasound images.
Methods: Three DTCBIR configurations, including decision tree with boosting (DTb), decision tree with full leaf features (DTL), and decision tree with selected leaf features (DTLs) were compared. For DTb, features of a query mass were combined first into a merged feature score and then masses with similar scores were retrieved. For DTL and DTLs, similar masses were retrieved based on the Euclidean distance between feature vectors of the query and those of selected references. For each DTCBIR configuration, we investigated the use of full feature set and subset of features selected by the stepwise linear discriminant analysis (LDA) and simplex optimization method, resulting in six retrieval methods and selected five, DTb-lda, DTL-lda, DTb-full, DTL-full, and DTLs-full, for the observer study. Three MQSA radiologists rated similarities between the query mass and computer-retrieved three most similar masses using nine-point similarity scale (9 = very similar).
Results: For DTb-lda, DTL-lda, DTb-full, DTL-full, and DTLs-full, average A_z values were 0.90 $\pm$ 0.03, 0.85 $\pm$ 0.04, 0.87 $\pm$ 0.04, 0.79 $\pm$ 0.05, and 0.71 $\pm$ 0.06, respectively, and average similarity ratings were 5.00, 5.41, 4.96, 5.33, and 5.13, respectively.
Conclusions: The DTL-lda is a promising DTCBIR CADx configuration which had simple tree structure, good classification performance, and highest similarity rating. © *2013 American Association of Physicists in Medicine*. [http://dx.doi.org/10.1118/1.4770277]

Key words: breast masses, computer-aided diagnosis, content-based image retrieval, decision tree, ultrasonography

I. INTRODUCTION

Breast cancer is one of the leading causes of death for women all over the world. In 2012, it was estimated that there would be approximately 226 870 newly diagnosed cases and 39 510 deaths in the United States.[1] Treatment at an early stage is important in order to reduce mortality from breast cancer. Screening mammography has been accepted as the major modality to reduce breast cancer mortality due to the increased detection of early cancers.[2] Despite the improvement in mammographic quality, however, the low sensitivity of screening mammography in dense breast remains a major limitation.[3] Ultrasonography (US) has been shown to be an useful modality for characterizing breast masses as malignant or benign.[4,5] Improved imaging techniques have enabled better characterization of sonographically visible breast lesions, allowing for less invasive management. In Ref. 5, a sensitivity of 98.4% and a specificity of 67.8% were achieved to distinguish 750 benign and malignant lesions using US. Taylor *et al.* showed that when sonography was combined with mammography in characterizing 761 breast masses, the specificity was improved from 51.4% to 63.8%, the positive predictive value was improved from 48% to 55.3%, and the sensitivity was improved from 97.1% to 97.9%.[6] Mammography is complemented by sonography for the diagnosis of breast masses in most breast imaging clinics of the United States. However, sonography is most effective if it is used for real-time evaluation by an experienced interpreter,[4–6] which in most clinical situations may not be practical. Moreover, there is considerable overlap in the sonographic characteristics between malignant and benign lesions due to the heterogeneous appearance of breast cancer. Many equivocal solid masses are recommended for biopsy, which increases health care costs and causes anxiety and morbidity to the patients. Currently, the positive predictive value for biopsy ranges only from 20% to 40%.[7–13] Therefore, it is very important to improve the accuracy of noninvasive methods of distinguishing malignant from benign masses and reduce the number of unnecessary biopsies.

Computer-aided diagnosis (CADx) is one of the research areas that have been explored to improve radiologists' accuracy in distinguishing between malignant and benign lesions. Earlier work on CADx for breast imaging focused on masses and microcalcifications on mammograms.[14–19] Although the

development of CADx systems for breast masses on US images started somewhat later than that on mammograms, numerous publications have appeared in the literature in the past 15 years. A survey of techniques for the computerized analysis of breast lesions on US images was provided by Cheng *et al.*[20] An incomplete sampling of work related to this study is presented below. Chen *et al.* used an artificial neural network (ANN) to classify 140 pathologically proven solid masses on US images[21] and obtained an area A_z under the receiver operating characteristic (ROC) curve of 0.96. In Ref. 22, Horsch *et al.* obtained an average A_z value of 0.87 from 11 independent experiments with their CADx system on a database of 400 cases. Sahiner *et al.*[23] studied computerized characterization of 102 breast masses on 3D US volumetric images and achieved an A_z value of 0.92. Sehgal *et al.*[24] used patient age, margin echogenicity, and angular variation of margin to distinguish malignant from benign masses using a logistic regression classifier. The leave-one-out testing A_z value was 0.87 on a dataset of 58 biopsy-proven masses. Joo *et al.*[25] trained an ANN classifier with five morphological US features to characterize masses using 584 histologically confirmed cases and to test it on an independent dataset of 266 cases. The test A_z value was 0.98. In Ref. 26, an automated method to segment breast masses on US images was designed by Cui *et al.*, achieving A_z values between 0.88 and 0.92.

To characterize a lesion, radiologists not only analyze individual features of that lesion, but they also depend on their recollection of clinically similar cases as references. Radiologists develop a pattern recognition memory of specific appearances of lesions and their characterization. This process has been labeled "Aunt Minnie" (Refs. 27 and 28) as an analogy to human recognition of facial features. The importance of image similarity for diagnostic decision making has prompted a wide research interest for the development of content-based image retrieval (CBIR) technology in medical imaging areas.[29,30] Several groups are developing methods to incorporate CBIR approaches into image database systems[31–35] and specifically for detection of masses on mammograms,[36–38] retrieval of liver lesions,[39] characterization of similar masses,[40,41] diagnosis of clustered microcalcifications on mammograms,[42] diagnosis of masses on US,[43,44] and diagnosis of masses on both mammograms and US.[43]

Previous work in CBIR for breast masses on US images included studies by Kuo *et al.*[45] and Chen *et al.*[46] Kuo *et al.* used three texture parameters, contrast, covariance, and dissimilarity at various pixel pair distances and a weighted Euclidean distance similarity measure (SM) for the retrieval of similar breast masses on US images. They retrieved the first k candidates with smaller distance values from the image database for differential diagnosis of malignant and benign lesions, and obtained a sensitivity of 94% with a specificity of 91% on a dataset of 129 malignant and 134 benign breast masses. Chen *et al.* used different texture feature spaces followed by principle component analysis and the Euclidean distance similarity measure. The performance of their image retrieval technique was evaluated by the separation between malignant and benign masses using tenfold cross validation on a dataset of 255 breast masses. The best combination of feature spaces yielded an A_z value of 0.925.

In our previous study,[44] we compared the effectiveness of seven SM: [i.e., Euclidean distance (ED), Manhattan distance, distance-weighted k-NN, correlation, cosine (Cos), linear discriminant analysis (LDA), and Bayesian neural network (BNN)] in a CBIR system. The performances of the CBIR CADx system were evaluated by radiologists' visual assessment of the similarity between the query and the retrieved masses. Although the BNN and LDA SMs had comparable classification performance that were higher than the other SMs in the CBIR CADx scheme, ED exhibited higher agreement (i.e., similarity ratings) from three radiologists' assessment than the Cos, LDA, and BNN measures.

To our knowledge, radiologists' visual similarity assessments have not been used by other research groups in the assessment of CBIR system retrieval of breast masses on US images and especially in CBIR systems based on decision trees.

In this study, we reported results for a CBIR system based on decision tree (i.e., DTCBIR) for CADx of US mass images that we are developing. The decision tree has advantages in some respects over other classifiers such as LDA and BNN, which were used in the CBIR CADx in our previous study.[44] A single decision tree classifier is a binary structure which is easy to understand and can be converted to rule sets that help improve the interpretation. At the same time a DTCBIR system is expected to provide more information than just likelihood of malignancy estimate to the radiologist by retrieving lesions similar to the query mass from the reference library and showing the known pathology of the retrieved masses as references to assist the radiologist in making diagnosis decision of the query mass. Moreover, the likelihood of malignancy of the query mass can be estimated by the DTCBIR system from the proportion of retrieved malignant and benign masses if the reference library is statistically representative of the population and the prevalence is properly taken into account.[47] Therefore, it is very important to investigate whether a given DTCBIR method can retrieve lesions that are considered to be similar by radiologists, what the best similarity measure is for image retrieval, and what the best method is for estimation of the likelihood of malignancy.

In this study, we investigated some of these issues for US image containing breast masses by comparing three DTCBIR configurations with or without additional feature selection steps and evaluating the performance of representative DTCBIR configurations in retrieving similar masses by radiologists' visual assessment. It is expected that this investigation will obtain relevant information for the design of a robust DTCBIR system for breast masses in US images.

II. MATERIALS AND METHODS

II.A. Dataset

A dataset was collected from the files of patients who had undergone breast US imaging in the Department of Radiology at the University of Michigan with approval by the

institutional review board (IRB). US images from 250 patients with breast masses (96 malignant and 154 benign) were acquired using a GE Logiq 700 scanner with an M12 linear array transducer for this study. All masses have biopsy proven pathology. A total of 488 US images were selected by two MQSA radiologists (R1 and R2) for these masses, as described below.

For each mass two optional orthogonal US views were selected by the radiologist R1 or R2, where the mass has been seen the best. However, if two orthogonal views were not available for a mass, only one view was selected. The patient cases were randomly partitioned into two subsets: S1 including 129 masses and S2 including 121 masses. The MQSA radiologist R1 selected US images corresponding to the biopsy-proven masses from the set S1 with the help of the pathology and radiology reports. The mass location was marked by the radiologist on every of the selected US images that correspond to the selected orthogonal or single US views. The second MQSA radiologist R2 selected and read images in the set S2 following the same procedure. As a result of the selection process, 258 images from 55 malignant and 74 benign masses were included in S1, and 230 images from 41 malignant and 80 benign masses were included in S2. In addition, the approximate center of the mass was provided by R1 and R2 for each image of the S1 and S2 subsets. The image selection and image annotation have been described in greater detail previously.[26]

II.B. Feature extraction and selection

An automated method proposed by Cui *et al.*[26] was used to segment breast masses on ultrasound images. A two-stage active contour model was used which automatically estimated an initial contour based on a manually identified point approximately at the mass center. By using the approximate mass centers from radiologists R1 and R2 two different computer segmentations were obtained for every image in the S1 and S2 datasets.

Based on the automated segmentation, we extracted features for the design of our DTCBIR system, which included width-to-height (WH), posterior shadowing (PS) and texture features. We defined WH ratio feature as a descriptor of the taller-than-wide shape of a sonographic mass, which is a good indication of malignancy.[5] The PS feature also has been reported to be useful for differentiation of malignant and benign masses. We described PS as the normalized average gray-level difference between the interior of the segmented mass and the darkest posterior strip.[23] The texture features were extracted from the spatial gray-level dependence (SGLD) matrices. Six texture features were extracted: information measures of correlations 1 and 2 (IC1 and IC2), difference entropy (DE), entropy (EN), energy (EY), and sum entropy (SM). The mathematical definitions of these features can be found in Ref. 48. The texture features were extracted from two disk-shaped regions containing the boundary of each mass, as well as the mass and normal tissue adjacent to the boundary of the mass. The areas for the upper and lower disk-shaped regions were selected to be equal, and their sum was the same as the area of the segmented mass. Six SGLD matrices (2 directions and 3 pixel-pair distances) were constructed for each disk-shaped region, and the total number of texture features extracted from an image containing the segmented mass was 72 (6 features $\times$ 6 SGLD matrices $\times$ the upper and the lower disk-shaped regions). The feature extraction methods have been described in greater detail previously.[23,44] Each feature was normalized to have values from 0 to 1, based on its own distribution in the training dataset.

The masses were classified as malignant or benign using a twofold cross-validation method and a LDA classifier[49] with stepwise feature selection. In the two cycles of twofold cross validation, the two data subsets S1 and S2 (see Sec. II.A) served alternately as the training and the test partition. The stepwise feature selection is based on the F statistics and uses three threshold values, F_{in} for feature entry, F_{out} for feature elimination, and tolerance of correlation. The appropriate values of these thresholds were not known *a priori*, and they were estimated using a leave-one-case-out resampling method and simplex optimization applied to the training set.[50]

The full set of features and the selected subset of features are used to characterize each mass. Table I shows the selected features sets for the four combinations of test set, training set, and centroid locations. The notation of each texture feature includes the information of direction, distance, and region. For example, IC2_0_2L denotes IC2 feature at direction $\theta = 0^\circ$, pixel-pair distance $d = 2$, and lower disk-shaped region.[44]

II.C. Retrieval methods by decision tree

A flowchart of our DTCBIR scheme is shown in Fig. 1. The mass on each US image from the reference database are segmented and the features characterizing the mass are stored in the reference library, composing a reference feature dataset. From every query sample submitted to the DTCBIR system to search for similar masses, the system first extracts the same features as that of the reference library. Using DTCBIR configurations, the similarity scores are then computed between the features of the query sample and those of the reference library. The system retrieves the masses in the reference library that are most similar to the query sample based on the similarity scores. In order to evaluate the effectiveness of the DTCBIR system in retrieval of similar masses, we conducted an observer study in which radiologists visually assessed the similarities between the query and the retrieved samples. Because our current reference library is still small, the DTCBIR system can only estimate a relative malignancy rating instead of the probability of malignancy for the query mass. ROC analysis of the relative malignancy rating estimated from the retrieved samples was used to evaluate the capability of the system in characterizing malignant and benign masses.

Using either the full set of features or the selected subset of features (described in Sec. II.B), we trained three decision trees. Each decision tree yielded a DTCBIR configuration, which could be used in our DTCBIR system. They included (1) decision tree with boosting (DTb), (2) decision tree with full leaf features (DTL), and (3) decision tree with selected leaf features (DTLs). One DTCBIR configuration (DTb) is

TABLE I. Selected feature sets by stepwise linear discriminant analysis and simplex optimization for the four combinations of test set, training set, and centroid location.

Selected features	Test S1 (train S2), centroids by R1	Test S2 (train S1), centroids by R1	Test S1 (train S2), centroids by R2	Test S2 (train S1), centroids by R2
WH	X	X	X	X
PS	X	X	X	X
IC1_0_2L		X		
IC1_0_4L	X		X	
IC1_0_4U			X	
IC1_90_2L	X	X	X	
IC1_90_6L				X
IC1_90_6U			X	
IC2_0_2L			X	
IC2_0_4L		X		
IC2_0_6L		X		
IC2_90_4U	X			
IC2_90_6L				X
IC2_90_6U	X			
DE_0_4U				X
DE_0_6L				X
DE_90_2L				X
DE_90_2U				X
EN_0_4U				X
EY_90_6L			X	

output-score-based and the other two DTCBIR configurations (DTL and DTLs) are input-feature-based. For the output-score-based DTCBIR configuration, the features of a query mass are combined first into a classifier score by decision tree with boosting, which is then compared to the classifier scores of the samples in the reference library to calculate the similarity scores. For the input-feature-based DTCBIR configuration, the individual features of a query mass are compared directly to the corresponding features of the samples in the reference library and the similarities of the individual features are combined into a similarity score for the pair. The six DTCBIR methods derived from the full feature set ("full") and the feature subset selected by the stepwise LDA and simplex optimization method ("lda") are denoted as DTb-lda, DTb-full, DTL-lda, DTL-full, DTLs-lda, and DTLs-full.

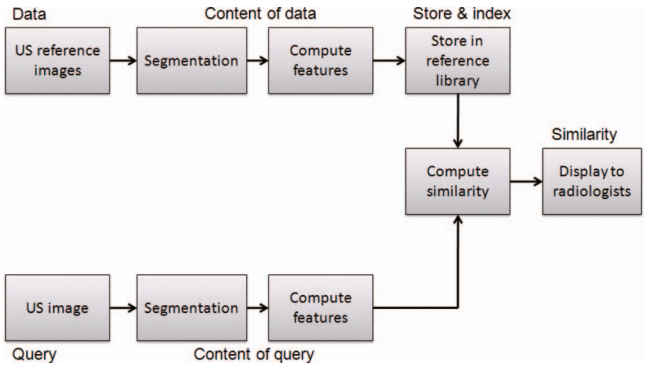

FIG. 1. The framework of our decision tree content-based image retrieval (DTCBIR) system.

The decision tree and DTCBIR configurations are described in detail below.

In our DTCBIR system, in order to retrieve k reference masses that have the highest similarity scores with the query mass, the k-nearest neighbor (k-NN) algorithm is used.

II.C.1. Decision trees

A decision tree is a simple tree structure where nonterminal nodes represent tests on one or more attributes and terminal nodes reflect decision outcomes. Each nonterminal node has a threshold associating with one or more features to divide the data into its descendants, and the process stops either when each terminal node only contains one class or all the selected attributes based on entropy have been exhausted.[51] Thus, decision tree can be used as a classification tool after the thresholds are set in the training process. The decision tree has advantages in some respects over other classifiers such as LDA, neural network, and support vector machine. For example, a single decision tree classifier is a binary structure, which is easy to understand, a decision tree can be converted to rule sets that help improve the interpretation, it can conveniently handle both continuous and discrete features, and it is independent of data distribution.

A well-known algorithm for constructing decision trees is C4.5.[52] This algorithm has been incorporated into the free classifier package WEKA (it is called J48 in WEKA) and is widely used in artificial intelligence. An updated version C5.0 that includes a boosting technique for improved performance is used in this study.

II.C.2. Decision tree with boosting retrieval

The basic idea behind boosting[53] is to design an ensemble of relatively simple base classifiers, and to combine the classifiers in this ensemble to improve the accuracy of the base classifier. When each classifier in the ensemble is being trained, the base classifier and the training method for the base classifier remain fixed, while the weighting of the training data changes. The first classifier of the ensemble, or the first iteration, uses unity weights for each case. At each subsequent iteration, the weights of the training cases are modified such that cases that have been misclassified in the previous iteration are weighted more heavily than those that are correctly classified. For a given test case, the C5.0 algorithm provides a confidence level that combines the results of the boosting iterations into a merged score.

In DTb, for every mass in the reference library a boosting merged score is calculated as described above. For the query mass, a boosting merged score is also calculated. The k-NN algorithm with the Euclidean distance is used for the retrieval scheme. All features were normalized before applying the Euclidean distance [Eq. (1)]

$$D(Q, R_i) = \sqrt{\sum_{s=1}^{n} (l_s(Q) - l_s(R_i))^2}, \quad (1)$$

where Q is the query mass, R_i is a reference mass i from the reference library, l_s is the sth feature, and n is the dimensionality of the feature space. A smaller distance D indicates a higher degree of similarity between the two compared masses. For the output-score-based DTb, the classifier score is the only feature so that $n = 1$ and l_1 is the boosting merged score and, from Eq. (1), it simply selects the k closest scores using the absolute difference between the query mass score and the scores of masses in the reference library.

From the k-NN algorithm, a characterization score which represents the relative malignancy rating of the query mass is computed as

$$P = \frac{1}{k} \sum_{i=1}^{k} B_i, \quad (2)$$

where k is the number of retrieved masses and B_i is a binary label indicating whether a retrieved mass is malignant (1) or benign (0) from the known pathology of each mass in the reference library. All DTCBIR configurations (DTb, DTL, and DTLs) used Eq. (2) for estimating the characterization scores in the retrieval scheme.

II.C.3. DTL and DTLs retrieval

Figure 2 shows an example of a single decision tree trained with the reference library masses. Within each leaf the final distribution of the malignant and benign images from the reference database are presented. The DT selected features (from full feature set or from LDA selected feature set) and the corresponding decision thresholds are also presented in Fig. 2. When a query mass is presented to the DT, it will be classified into one of the leaves.

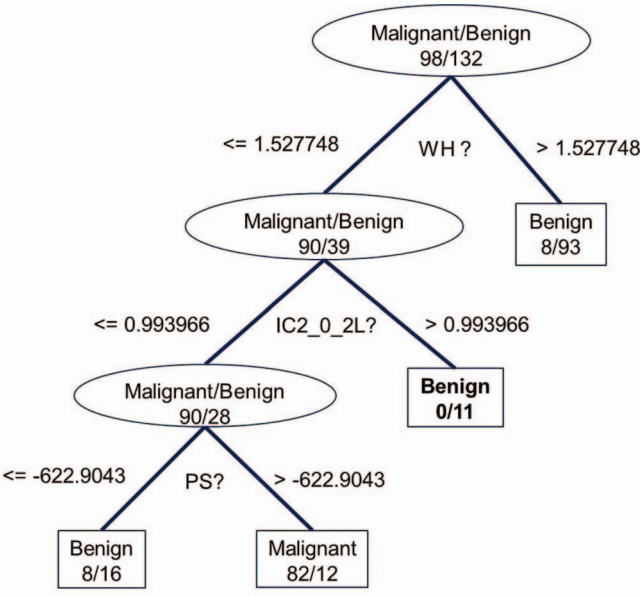

FIG. 2. A single decision tree (DTL-lda) using training set S1 with mass centroids marked by R2. Before training the decision tree, a subset of features was selected by the stepwise LDA and simplex optimization method. The numbers of malignant and benign masses reaching each leaf are indicated.

The similarity score of DTL is obtained by calculating the Euclidean distance [Eq. (1)] between the query mass features and the features of each reference mass within the same leaf into which the query mass was classified. Top k most similar masses are selected within the same leaf. However, if there are less than k masses in the leaf, we retrieved all masses. The feature space consists of the features selected by the single decision tree during training. For example, in Fig. 2, if a query mass reaches the benign leave [benign (0/11) marked in bold], the DTL will apply Euclidean distance measure using all three selected features (i.e., WH, IC2_0_2L, and PS) between the query and the 11 reference masses.

The similarity score of the DTLs is similar to the DTL, except that it uses only the features that are utilized to classify the query mass into the specific leaf. For example, in Fig. 2, if the query mass reaches the benign leaf, benign (0/11), the DTLs will use the Euclidean distance measure with only two corresponding features (IC2_0_2L and WH) between the query and the 11 reference masses.

II.D. Evaluation methods

II.D.1. Evaluation of classification performance of DTCBIR

The three DTCBIR configurations (DTb, DTL, and DTLs) used Eq. (2) to estimate the characterization scores. The characterization scores were then analyzed by the ROC methodology[54] and the area under ROC curve (A_z) was calculated using LABROC.[54] As described above, prior to image retrieval, the DTb method produces a one-dimensional classifier score. This score, termed DTb_{DI} score below, was also used to estimate the performance of the DTb method directly, without any involvement of the retrieval scheme. In most

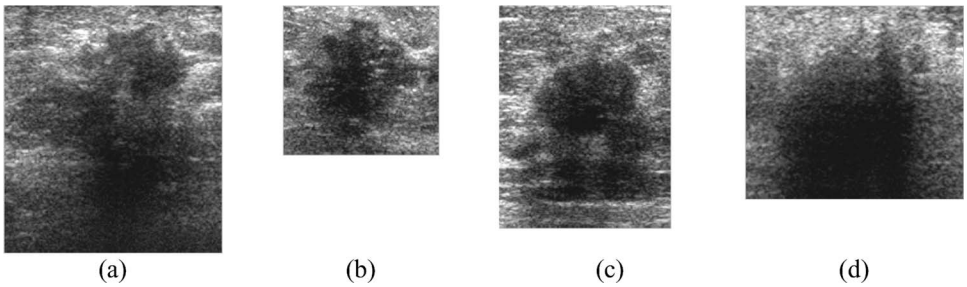

FIG. 3. A malignant query mass and three retrieved masses ($k = 3$) by our DTCBIR scheme using the DTL-lda method: (a) a malignant query mass, (b) first retrieved mass, (c) second retrieved mass, (d) third retrieved mass. The biopsy results of (a)–(d) are malignant. The similarity ratings from three radiologists (R1, R2, and R3) estimating the similarity between the query mass and the retrieved masses are: (b) 7, 8, and 8; (c) 8, 9, and 6; (d) 8, 6, and 7.

CADx studies, this direct performance evaluation is reported. However, in this study, since our main focus was image retrieval, the direct performance evaluations was reported only for comparison purposes, and the main performance evaluation of DTb was based on the calculation of characterization scores using Eq. (2).

II.D.2. Similarity evaluation by radiologists

The similarity between the query and the retrieved masses by the DTCBIR CADx system were evaluated by radiologists' visual similarity assessments. We used one of the four partitions for this experiment. The partition testing on set S1, training on set S2, segmentation initialized by R1, was selected because its A_z is close to the average A_z ($k = 3$) of the four partitions. The dataset for similarity study therefore included 121 reference library masses on 230 (79 malignant and 151 benign) images (S2 set) and 100 query masses from S1 on 100 (49 malignant and 51 benign) images. Finding a good balance among the number of observers, the proper number of similarity measures, the number of query masses, and the k value is difficult because of time constraint. We selected the number of query masses as 100. The 49 malignant and 51 benign masses were randomly selected from the malignant and benign subsets in S1, respectively. Among six retrieval methods (i.e., three DTCBIR configurations with and without feature selection), we selected five, DTb-lda, DTL-lda, DTb-full, DTL-full, and DTLs-full, for the observer study. For classification of masses as malignant and benign, our results [see Sec. III and Fig. 5(e)] indicated that DTb and DTL had better performance than DTLs for $k = 3$. DTLs-full was selected to represent the lowest A_z value. In this way, we attempted to cover the performance range for classification of masses as malignant or benign. For each query mass, 3 most similar masses ($k = 3$) were retrieved from the reference library with each method. Therefore, there were a total of 1500 (100 query masses × 3 most similar masses × 5 methods) pairs of query and retrieved masses. A graphical user interface was developed to present the image pairs of the query and retrieved masses to the radiologists. The observer assessed the two images of a pair that were displayed side-by-side in full resolution on a display workstation and was allowed to zoom and adjust the contrast and brightness of the images if needed. The mass pairs were mixed and presented to the radiologists in random order, one pair at a time. Three MQSA radiologists, with breast imaging experience of 9, 25, and 29 years, rated the similarity between the query mass and the computer-retrieved masses. They used a nine-point similarity scale (1 = very dissimilar, 3 = quite dissimilar, 5 = some degree of resemblance, 7 = quite similar, and 9 = very similar). The similarity ratings 2, 4, 6, and 8 were intermediate ratings. The radiologists were instructed to estimate the similarity as they would do in clinic by considering both the visual similarity and the similarity based on clinical malignant/benign descriptors. Examples of similarity evaluation by radiologists are shown on Figs. 3 and 4 with both malignant and benign query masses. Two of the three radiologists (R1, R2) were the same as the two that helped collect the dataset, marked the masses on the US images, and provided the centroid locations. However, none of the masses were viewed in pairs during dataset collection, and the collection of the dataset did not involve comparing the similarity of the masses. Furthermore,

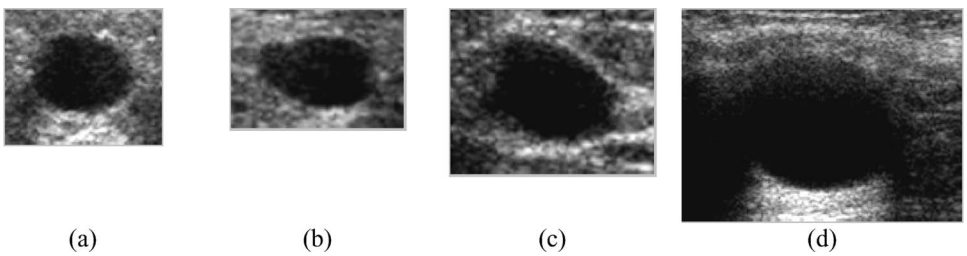

FIG. 4. A benign query mass and three retrieved masses ($k = 3$) by our DTCBIR scheme using the DTL-lda method: (a) a benign query mass, (b) first retrieved mass, (c) second retrieved mass, (d) third retrieved mass. The biopsy results of (a)–(d) are benign. The similarity ratings from three radiologists (R1, R2, and R3) estimating the similarity between the query mass and the retrieved masses are: (b) 8, 8, and 8; (c) 8, 7, and 7; (d) 7, 7, and 8.

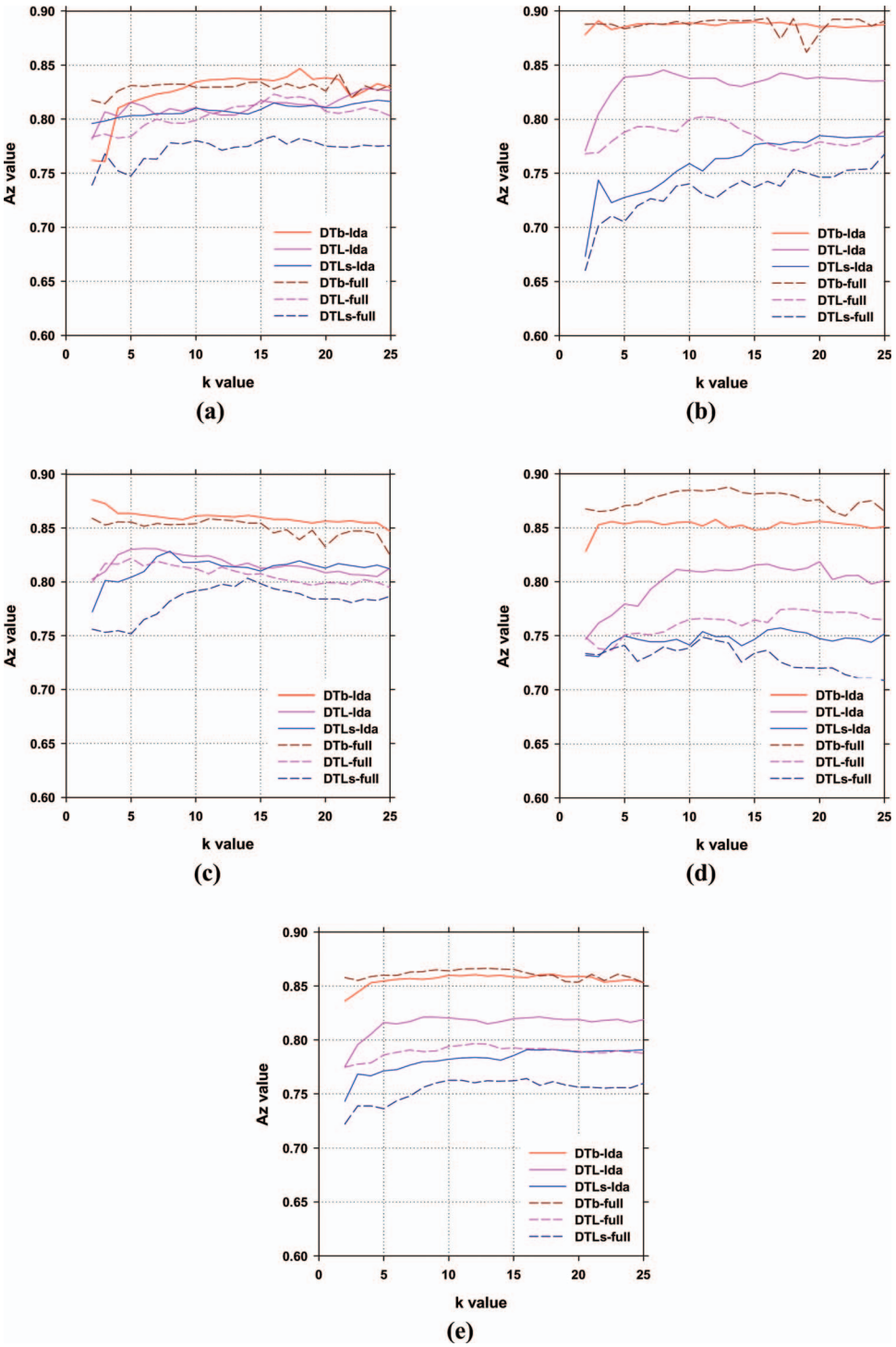

FIG. 5. The area under the ROC curves, A_z, for top k retrievals using segmentation initialized by mass centroids marked by R1 or R2 for the two cycles of cross validation: (a) Test set S1 (training set S2), mass centroids marked by R1, (b) Test set S2 (training set S1), mass centroids marked by R1, (c) Test set S1 (training set S2), mass centroids marked by R2, (d) Test set S2 (training set S1), mass centroids marked by R2, (e) Average of (a)–(d).

data collection was done two years before the similarity study and only ROI images were provided in the similarity study. Therefore, the participation in the similarity observer study of the two radiologists is not expected to introduce biases.

II.D.3. Normalized discounted cumulative gain (NDCG)

The performance of the DT retrieval methods were also evaluated by NDCG,[55] a standard technique used to measure the effectiveness of information retrieval algorithms when

graded truth is available, as represented by the nine-point radiologists' similarity scale in this study.

The discounted cumulative gain (DCG) is computed as

$$\mathrm{DCG}(k) = \sum_{i=1}^{k} \frac{2^{sim_i} - 1}{\log_2(1+i)}, \quad (3)$$

where i is the rank of the retrieved masses based on similarity measures, k is the total number of retrieved masses (here $k = 3$), and sim_i is the relevance value (i.e., similarity ratings from radiologists) of the result at rank i. NDCG is defined as

$$\mathrm{NDCG}(k) = \frac{\mathrm{DCG}(k)}{\mathrm{IDCG}(k)}, \quad (4)$$

where IDCG(k) denotes the DCG(k) value for an ideal ranked list for query mass. In a perfect algorithm, DCG(k) will be the same as the IDCG(k) producing a NDCG(k) of 1.0. The NDCG is used to measure the usefulness (gain) on a scale of 0 to 1 (or 0 to 100%) of k retrieved masses on the basis of their positions in the ranked list when the DT retrieval was used, and on the basis of their similarity to the query mass according to a separate similarity reference standard (the radiologist's rating on nine-point scale). The accumulated gain is evaluated with the weight of each retrieved lesion discounted at lower ranks. Thus, for a given k, higher NDCG(k) means more masses similar to the query mass are ranked ahead of dissimilar ones, with NDCG(k) equal to 1 implying perfect retrieval of k images.[39,56] The advantage of NDCG is that among the classifiers with the same accuracy, the classifier that can rank the similar masses higher will be rewarded more.

III. RESULTS

III.A. DTb_{DI} classification accuracy

Table II shows the A_z values for the DTb_{DI} with the two features sets (full features and the features selected by the stepwise LDA and simplex optimization) obtained directly from analysis of the classifier scores for the different dataset partitions. The average test A_z value for both DTb_{DI}-lda and DTb_{DI}-full were 0.86 ± 0.02. The maximum number of boosting iterations was 10.

III.B. Retrieval methods' characterization accuracy

The DT selected the WH ratio feature consistently in the four combinations of test set, training set, and centroid location. The PS feature was selected in three of four combinations in the case of DT selection from the full features and in all four combinations in the case of DT selection from the stepwise LDA preselected features. The malignant-versus-benign classification performance of the DTCBIR system for each dataset is presented in Fig. 5. The performance accuracy in terms of A_z of the DTCBIR depends on the number of the retrieved most similar masses, k. The dependence of A_z values on k averaged over the four datasets is shown in Fig. 5(e) for all methods. The average A_z values of the DTb-lda and DTb-full based systems remain relatively unchanged for all k, and those of other DTCBIR methods do not change substantially for $k \geq 10$. The C5.0 DT algorithm was executed with its default parameter values, except that we set the minimum of training instances within each leaf to be 10. This value was chosen because we observed that the resulting DTs have compact structure with relatively small number of selected features and, in DTCBIR system, it is good enough as a retrieved number. Overall, DT with boosting, DTb-lda and DTb-full, achieve a better performance compared to other single DT classification. The average A_z values of DTb-lda and DTb-full at $k = 3$ were 0.84 ± 0.04 and 0.86 ± 0.04, respectively. Table III shows the average A_z values of the six DTCBIR methods for several k values. Results for other k values can be found in Fig. 5(e). Table IV presents the A_z values for the 100 query images. The A_z value at $k = 3$ was significantly ($p \leq 0.01$) higher for the DTL-lda than the DTLs-full. However, the differences between DTL-lda and the rest of the retrieval methods did not reach statistical significance ($p \geq 0.1$).

TABLE II. The A_z values for DTb_{DI} classifiers with two feature set (full feature and LDA-selected feature set) for the four combinations of test set, training set, and centroid location.

Dataset	DTb_{DI}-lda	DTb_{DI}-full
Test S1 (train S2), centroid by R1	0.88 ± 0.02	0.88 ± 0.02
Test S2 (train S1), centroid by R1	0.85 ± 0.03	0.85 ± 0.03
Test S1 (train S2), centroid by R2	0.85 ± 0.02	0.86 ± 0.02
Test S2 (train S1), centroid by R2	0.87 ± 0.02	0.86 ± 0.02
Average	0.86 ± 0.02	0.86 ± 0.02

TABLE III. Average A_z values of the DTCBIR-CADx system using k-NN with six different retrieval methods for several k values. The average was performed over the test sets from the four combinations of test set, training set, and centroid location.

DTCBIR methods	$k = 3$	$k = 5$	$k = 10$	$k = 25$
DTb-lda	0.84 ± 0.04	0.86 ± 0.03	0.86 ± 0.03	0.86 ± 0.03
DTL-lda	0.80 ± 0.03	0.82 ± 0.03	0.82 ± 0.03	0.82 ± 0.03
DTLs-lda	0.78 ± 0.03	0.77 ± 0.03	0.78 ± 0.03	0.79 ± 0.03
DTb-full	0.86 ± 0.04	0.86 ± 0.03	0.86 ± 0.03	0.85 ± 0.03
DTL-full	0.78 ± 0.04	0.79 ± 0.03	0.80 ± 0.03	0.79 ± 0.03
DTLs-full	0.76 ± 0.04	0.73 ± 0.03	0.76 ± 0.03	0.75 ± 0.03

TABLE IV. The A_z values for the top k retrieval using the DTCBIR-CADx system trained by S2, centroid by R1, and 100 test query images from S1.

DTCBIR methods	$k = 3$	$k = 5$	$k = 10$	$k = 25$
DTb-lda	0.90 ± 0.03	0.89 ± 0.03	0.90 ± 0.03	0.89 ± 0.03
DTL-lda	0.85 ± 0.04	0.86 ± 0.04	0.85 ± 0.04	0.84 ± 0.04
DTb-full	0.87 ± 0.04	0.87 ± 0.04	0.87 ± 0.04	0.89 ± 0.03
DTL-full	0.79 ± 0.05	0.80 ± 0.05	0.81 ± 0.04	0.81 ± 0.04
DTLs-full	0.71 ± 0.06	0.72 ± 0.05	0.74 ± 0.05	0.76 ± 0.05

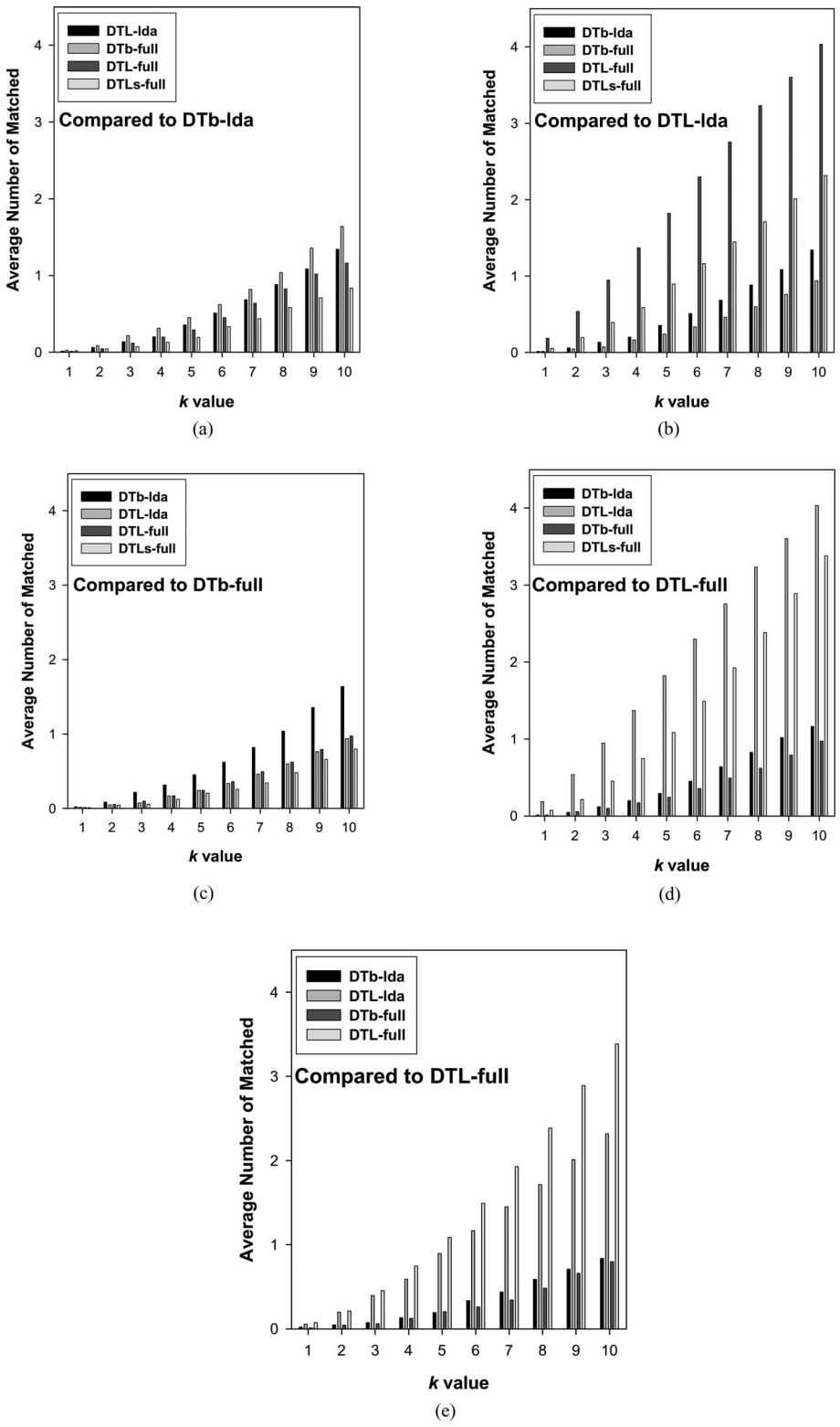

FIG. 6. The number of retrieved masses that were identical between two DTCBIR methods. The results for $k = 1$ to $k = 10$ are shown. The partitions: Test S1 (train S2), centroid by R1, was used. (a) DTb-lda, (b) DTL-lda, (c) DTb-full, (d) DTL-full, and (e) DTLs-full, was compared to the other four DTCBIR methods in each graph.

III.C. Number of similar masses retrieved by different methods

We studied the consistency of the different DTCBIR methods by comparing the number of identical masses retrieved by the different DTCBIR methods for a specified k. The DTL-lda, DTb-full, DTL-full and DTLs-full and DTb-lda are compared in Fig. 6 and Tables V and VI. For example, DTL-lda, DTb-full, DTL-full, and DTLs-full are compared with DTb-lda in Fig. 6(a), where the average number of identical

TABLE V. The average number of retrieved masses from the reference library that were identical (k = 3): test S1 (train S2), centroid by R1.

DTCBIR methods	DTCBIR configurations				
	DTb-lda	DTL-lda	DTb-full	DTL-full	DTLs-full
DTb-lda	3	0.14	0.22	0.12	0.07
DTL-lda	0.14	3	0.07	0.95	0.40
DTb-full	0.22	0.07	3	0.10	0.06
DTL-full	0.12	0.95	0.10	3	0.45
DTLs-full	0.07	0.40	0.06	0.45	3

masses retrieved by the different DTCBIR methods for a given k ($k = 1, \ldots, 10$) are shown. The same DTCBIR configuration with different features set (i.e., DTb-lda and DTb-full as well as DTL-lda and DTL-full) retrieved more identical masses than others. The DTL retrieved more masses, on average, that were identical to those retrieved by DTLs than DTb. The comparisons in Table VI show that, on average, 4.03 of the 10 masses retrieved by DTL-lda and DTL-full in 10-NN ($k = 10$) were the same. On the other hand, both DTb-lda and DTb-full retrieved less than two masses, on average, that were identical to those retrieved by DTL-lda, DTL-full, and DTLs-full.

III.D. Evaluation of retrieval methods by radiologists' visual assessment

Table VII shows the average similarity ratings of the three radiologists for all 100 query masses and the subset of malignant masses retrieved by the DTb-lda, DTL-lda, DTb-full, DTL-full, and DTLs-full methods and $k = 3$. The three radiologists' average similarity ratings were as follows: 5.00 for DTb-lda, 5.41 for DTL-lda, 4.96 for DTb-full, 5.33 for DTL-full, and 5.13 for DTLs-full. The average similarity ratings from the three radiologists were higher for the retrieval methods based on DTL compared to the ratings for the retrieval methods based on DTb. Statistical comparison was performed by finding the average similarity rating for each query mass (averaged over three readers and three retrieved masses) for each retrieval method, and then conducting a paired t-test of the average similarity ratings between pairs of retrieval methods. The radiologists' average similarity ratings were significantly ($p < 0.0001$) higher for the DTCBIR method based on DTL-lda than those based on a classifier score (DTb-lda or DTb-full). The average similarity ranking for DTL-lda was also significantly higher ($p < 0.0008$) than the ones for DTLs-full. However, the difference between DTL-lda and DTL-full did not reach statistical significance ($p = 0.27$). For malignant query masses, the average similarity ratings were 5.16 for DTb-lda, 5.52 for DTL-lda, 5.15 for DTb-full, 5.48 for DTL-full, and 5.09 for DTLs-full. We observed a tendency that one of the radiologists was giving lower similarity ratings than the other two. On average, the DTCBIR system retrieved masses that were moderately similar to the query masses based on radiologists' similarity assessments. Masses of higher similarities were retrieved for the malignant masses than for all query masses except for the DTLs-full retrieval method.

TABLE VI. The average number of retrieved masses from the reference library that were identical (k = 10): test S1 (train S2), centroid by R1.

DTCBIR methods	DTCBIR configurations				
	DTb-lda	DTL-lda	DTb-full	DTL-full	DTLs-full
DTb-lda	10	1.34	1.64	1.16	0.84
DTL-lda	1.34	10	0.94	4.03	2.32
DTb-full	1.64	0.94	10	0.97	0.80
DTL-full	1.16	4.03	0.97	10	3.38
DTLs-full	0.84	2.32	0.80	3.38	10

Table VIII shows the NDCG values of the DTCBIR-CADx system with five different DTCBIR methods for $k = 3$. The average NDCG values of DTL-lda and DTL-full have slightly higher values than those of DTb-lda, DTb-full or DTLs-full.

This study shows that among the six DTCBIR methods the DTL-lda is the best for searching similar masses and its A_z value is also acceptable (Tables IV and VII).

IV. DISCUSSION

We used two types of initial feature sets as input to DTCBIR (DTb, DTL, and DTLs) configurations. One set consisted of all extracted features and the other consisted of the features selected by the stepwise LDA and simplex optimization method. DT has its own feature selection, which is based on entropy,[57] and is performed within the DT training step. The DT feature selection proved to be efficient, because for $k = 3$, the average A_z values for DTb, DTL, and DTLs with the initial set of features selected by LDA are similar to the ones with the initial set of all features (Table III).

When compared to other DTCBIR methods, the DTL with LDA selected feature set and DTL with full feature set retrieved on average more masses that were identical (0.95 and 4.03 for $k = 3$ and $k = 10$, respectively, see Tables V and VI). DTL-full and DTLs-full were next (0.45 and 3.38 for $k = 3$ and $k = 10$, respectively). The masses retrieved by DTb-lda were more similar to those retrieved by DTb-full (0.22 and 1.64 for $k = 3$ and $k = 10$, respectively) than to those by the other three DTCBIR methods. However, masses retrieved by the input-feature-based configurations (e.g., DTL or DTLs) were very different from the masses that were retrieved by the output-score-based (DTb) configurations in the DTCBIR scheme because the input-feature-based methods used the individual features in the multidimensional feature space while the output-score-based methods used the merged classifier scores.[44]

The higher similarity of the query and retrieved masses for the input-feature-based configurations (DTL and DTLs) compared to the output-score-based configurations (DTb) was observed also based on the radiologists' visual assessments obtained in the observer study. On the other hand, the average A_z value of the output-score-based configurations (DTb) for the top retrievals (k) (Table IV) was higher than those of the

TABLE VII. The average similarity ratings of the three radiologists for all masses and the subset of malignant masses retrieved by the DTCBIR methods (k = 3).

DTCBIR methods	R1		R2		R3		Average	
	Total	Malignant	Total	Malignant	Total	Malignant	Total	Malignant
DTb-lda	4.72	4.65	5.27	5.65	5.01	5.19	5.00	5.16
DTL-lda	5.14	4.99	5.81	6.08	5.29	5.49	5.41	5.52
DTb-full	4.49	4.51	5.39	5.67	5.00	5.27	4.96	5.15
DTL-full	4.99	4.95	5.69	6.00	5.30	5.5	5.33	5.48
DTLs-full	4.88	4.64	5.41	5.45	5.09	5.17	5.13	5.09

input-feature-based configurations (DTL and DTLs). These results indicated that using the combined similarity of the individual features in the multidimensional feature space can match the similar masses more reliably than using a merged classifier score. The same classifier score can be obtained from many different combinations of the individual features and therefore masses retrieved based on similar classifier scores to the query mass may have large variations and very different individual features, which is also in agreement with our previous study.[44]

The performance of the CADx systems briefly reviewed in the Introduction varied considerably (A_z range of 0.87–0.98). However, it is difficult to directly compare the performances of different CADx systems because they generally depend on the subtlety of the abnormal cases and the size of the datasets used. The performance of our DT classifier without retrieval DTb_{DI} (A_z range of 0.85–0.88) was in the same range as the CADx systems of Horsch *et al.*[22] (A_z of 0.87) and Sehgal *et al.*[24] (A_z of 0.87), but was slightly lower than the conventional classifiers LDA and BNN (A_z range of 0.86–0.91) from our previous study[44] and some of the other studies.[23,25,26] The classification performance of our DTCBIR CADx systems (A_z range of 0.71–0.90) was comparable to Chen *et al.*[46] (A_z range of 0.893–0.925 for different feature combinations). However, our study is different from the studies of other investigators in that we investigated whether the DTCBIR CADx systems can retrieve lesions on US images that are considered to be similar by radiologists.

In this study, our focus was the design of a CBIR CADx system using decision trees. We also aimed at comparing the performance of the input-feature-based and output-score-based DTCBIR systems. The ultimate benchmark for a CADx system is the improvement in the performance of the radiologists when they are aided by the CADx system. The evaluation of DTCBIR CADx system performance is a relatively new area, and the tradeoffs between the performance of the standalone system for retrieval and classification as they are related to this ultimate benchmark are not yet known. Future observer studies will be needed to evaluate the relative importance of these two performance criteria.

There are limitations in our similarity study. Three radiologists participated in the observer study. They rated the similarity of a query mass to the top three ($k = 3$) retrieved masses. The total number of query masses was 100. Five DTCBIR methods were used for retrieval. These resulted in 1500 (3 × 5 × 100) readings performed by each radiologist. Even though the total number of readings was relatively large, the number of readings for each mass and the number of query masses were still small. One practical and important way to obtain more robust results is to increase the number of observers, which we plan to do in a future study. More reliable results may also be produced by increasing k; however, we have to choose carefully the value of k in order to avoid excessive reading times for the radiologists. Likewise, for the observer study five representative DT based SM methods were chosen from the six developed to reduce the number of readings needed. A search for a good balance among the number of query masses, the number of retrieved masses, the proper number of similarity measures, and the number of observers, will be carried out in the future. In addition, the available dataset for this study was relatively limited and future studies with larger datasets will be needed.

TABLE VIII. The NDCG values of the five different DTCBIR methods for k = 3.

DTCBIR methods	R1	R2	R3	Avg.
DTb-lda	0.81	0.83	0.86	0.83
DTL-lda	0.84	0.87	0.88	0.86
DTb-full	0.80	0.84	0.85	0.83
DTL-full	0.85	0.88	0.88	0.87
DTLs-full	0.82	0.84	0.89	0.85

V. CONCLUSIONS

In this study, we compared six different DTCBIR methods with full features and subset of features selected by the stepwise LDA and simplex optimization method (DTb-lda, DTb-full, DTL-lda, DTL-full, DTLs-full, and DTLs-full). Even though the DTb retrieval methods had the best classification performance (i.e., highest A_z) in the DTCBIR scheme while DTLs had the worst performance for $k = 3$, DTLs-full exhibited higher similarity ratings from the three radiologists' assessment than the DTb retrieval methods for the 100 query masses on average. In future investigations, we will study the relationship between the usefulness of the retrieved masses as

references for radiologists and the accuracy of estimating the likelihood of malignancy of the query mass. Future work includes applying the DTCBIR system to a larger and independent dataset, expanding the feature space, and combining the developed US characterization method with mammographic characterization method. The effects of the different DTCBIR CADx systems on the characterization of breast masses by radiologists will also be evaluated by observer study.

ACKNOWLEDGMENTS

This work is supported by USPHS Grant No. CA 118305. The authors are grateful to Charles E. Metz for the LABROC program.

[a]Author to whom correspondence should be addressed. Electronic mail: lhadjisk@umich.edu; Telephone: (734) 647-7428; Fax: (734) 615-5513.

[1]"American Cancer Society, "Cancer Facts & Figures 2012," " 2012, see www.cancer.org.
[2]L. L. Humphrey, M. Helfand, B. K. S. Chan, and S. H. Woolf, "Breast cancer screening: A summary of the evidence for the U.S. Preventive Services Task Force," Ann. Intern. Med. **137**, 347–360 (2002).
[3]K. Kerlikowske, D. Grady, J. Barclay, E. A. Sickles, and V. Ernster, "Effect of age, breast density, and family history on the sensitivity of first screening mammography," J. Am. Med. Assoc. **276**, 33–38 (1996).
[4]A. S. Hong, E. L. Rosen, M. S. Soo, and J. A. Baker, "BI-RADS for sonography: Positive and negative predictive values of sonographic features," Am. J. Roentgenol. **184**, 1260–1265 (2005).
[5]A. T. Stavros, D. Thickman, C. L. Rapp, M. A. Dennis, S. H. Parker, and G. A. Sisney, "Solid breast nodules: Use of sonography to distinguish between malignant and benign lesions," Radiology **196**, 123–134 (1995).
[6]K. J. W. Taylor *et al.*, "Ultrasound as a complement to mammography and breast examination to characterize breast masses," Ultrasound Med. Biol. **28**, 19–26 (2002).
[7]R. D. Rosenberg *et al.*, "Performance benchmarks for screening mammography," Radiology **241**, 55–66 (2006).
[8]M. Kriege *et al.*, "Efficacy of MRI and mammography for breast-cancer screening in women with a familial or genetic predisposition," New Engl. J. Med. **351**, 427–437 (2004).
[9]C. K. Kuhl, S. Schrading, C. C. Leutner, N. Morakkabati-Spitz, E. Wardelmann, R. Fimmers, W. Kuhn, and H. H. Schild, "Mammography, breast ultrasound, and magnetic resonance imaging for surveillance of women at high familial risk for breast cancer," J. Clin. Oncol. **23**, 8469–8476 (2005).
[10]M. O. Leach *et al.*, "Screening with magnetic resonance imaging and mammography of a UK population at high familial risk of breast cancer: A prospective multicentre cohort study (MARIBS)," Lancet **365**, 1769–1778 (2005).
[11]C. D. Lehman *et al.*, "Screening women at high risk for breast cancer with mammography and magnetic resonance imaging," Cancer **103**, 1898–1905 (2005).
[12]F. Sardanelli and F. Podo, "Breast MR imaging in women at high-risk of breast cancer: Is something changing in early breast cancer detection?," Eur. Radiol. **17**, 873–887 (2007).
[13]E. Warner *et al.*, "Surveillance of BRCA1 and BRCA2 mutation carriers with magnetic resonance imaging, ultrasound, mammography, and clinical breast examination," J. Am. Med. Assoc. **292**, 1317–1325 (2004).
[14]H. P. Chan, D. Wei, K. L. Lam, B. Sahiner, M. A. Helvie, D. D. Adler, and M. M. Goodsitt, "Classification of malignant and benign microcalcifications by texture analysis," Med. Phys. **22**, 938 (1995) (Abstract).
[15]Y. Jiang, R. M. Nishikawa, D. E. Wolverton, C. E. Metz, M. L. Giger, R. A. Schmidt, C. J. Vyborny, and K. Doi, "Malignant and benign clustered microcalcifications: Automated feature analysis and classification," Radiology **198**, 671–678 (1996).
[16]L. Shen, R. M. Rangayyan, and J. E. L. Desautels, "Application of shape analysis to mammographic calcifications," IEEE Trans. Med. Imaging **13**, 263–274 (1994).
[17]Z. Huo, M. L. Giger, C. J. Vyborny, U. Bick, P. Lu, D. E. Wolverton, and R. A. Schmidt, "Analysis of spiculation in the computerized classification of mammographic masses," Med. Phys. **22**, 1569–1579 (1995).
[18]B. Sahiner, H. P. Chan, N. Petrick, M. A. Helvie, M. M. Goodsitt, and D. D. Adler, "Classification of masses on mammograms using a rubber-band straightening transform and feature analysis," Proc. SPIE **2710**, 44–50 (1996).
[19]J. A. Baker, P. J. Kornguth, J. Y. Lo, and C. E. Floyd, "Artificial neural network: Improving the quality of breast biopsy recommendations," Radiology **198**, 131–135 (1996).
[20]H. D. Cheng, J. Shan, W. Ju, Y. H. Guo, and L. Zhang, "Automated breast cancer detection and classification using ultrasound images: A survey," Pattern Recogn. **43**, 299–317 (2010).
[21]D. R. Chen, R. F. Chang, and Y. L. Huang, "Computer-aided diagnosis applied to US of solid breast nodules by using neural networks," Radiology **213**, 407–412 (1999).
[22]K. Horsch, M. L. Giger, L. A. Venta, and C. J. Vyborny, "Computerized diagnosis of breast lesions on ultrasound," Med. Phys. **29**, 157–164 (2002).
[23]B. Sahiner, H. P. Chan, M. A. Roubidoux, M. A. Helvie, L. M. Hadjiiski, A. Ramachandran, G. L. LeCarpentier, A. Nees, C. Paramagul, and C. Blane, "Computerized characterization of breast masses on 3-D ultrasound volumes," Med. Phys. **31**, 744–754 (2004).
[24]C. M. Sehgal, T. W. Cary, S. A. Kangas, S. P. Weinstein, S. M. Schultz, P. H. Arger, and E. F. Conant, "Computer-based margin analysis of breast sonography for differentiating malignant and benign masses," J. Ultrasound Med. **23**, 1201–1209 (2004).
[25]S. Joo, Y. S. Yang, W. K. Moon, and H. C. Kim, "Computer-aided diagnosis of solid breast nodules: Use of an artificial neural network based on multiple sonographic features," IEEE Trans. Med. Imaging **23**, 1292–1300 (2004).
[26]J. Cui, B. Sahiner, H.-P. Chan, A. Nees, C. Paramagul, L. M. Hadjiiski, C. Zhou, and J. Z. Shi, "A new automated method for the segmentation and characterization of breast masses on ultrasound images," Med. Phys. **36**, 1553–1565 (2009).
[27]K. E. Applegate and D. V. B. Neuhauser, "Whose Aunt Minnie?," Radiology **211**, 292–292 (1999).
[28]L. Berlin, "Aunt Minnie's atlas and imaging-specific diagnosis," Radiology **204**, 278–278 (1997).
[29]H. Muller, N. Michoux, D. Bandon, and A. Geissbuhler, "A review of content-based image retrieval systems in medical applications: Clinical benefits and future directions," Int. J. Med. Inf. **73**, 1–23 (2004).
[30]H. Muller, A. Rosset, A. Garcia, J. P. Vallee, and A. Geissbuhler, "Informatics in radiology (infoRAD): Benefits of content-based visual data access in radiology," Radiographics **25**, 849–858 (2005).
[31]G. L. Gimelfarb and A. K. Jain, "On retrieving textured images from an image database," Pattern Recogn. **29**, 1461–1483 (1996).
[32]V. N. Gudivada and V. V. Raghavan, "Design and evaluation of algorithms for image retrieval by spatial similarity," ACM Trans. Inf. Sys. **13**, 115–144 (1995).
[33]J. Hafner, H. S. Sawhney, W. Equitz, M. Flickner, and W. Niblack, "Efficient color histogram indexing for quadratic form distance functions," IEEE Trans. Pattern Anal. Mach. Intell. **17**, 729–736 (1995).
[34]V. E. Ogle, "Chabot: Retrieval from a relational database of images," Computer **28**, 40–48 (1995).
[35]R. K. Srihari, "Automatic-indexing and content-based retrieval of captioned images," Computer **28**, 49–56 (1995).
[36]G. D. Tourassi, B. Harrawood, S. Singh, J. Y. Lo, and C. E. Floyd, "Evaluation of information-theoretic similarity measures for content-based retrieval and detection of masses in mammograms," Med. Phys. **34**, 140–150 (2007).
[37]S. C. Park, R. Sukthankar, L. Murnmert, M. Satyanarayanan, and B. Zheng, "Optimization of reference library used in content-based medical image retrieval scheme," Med. Phys. **34**, 4331–4339 (2007).
[38]X. H. Wang, S. C. Park, and B. Zheng, "Improving performance of content-based image retrieval schemes in searching for similar breast mass regions: An assessment," Phys. Med. Biol. **54**, 949–961 (2009).
[39]S. A. Napel, C. F. Beaulieu, C. Rodriguez, J. Y. Cui, J. J. Xu, A. Gupta, D. Korenblum, H. Greenspan, Y. J. Ma, and D. L. Rubin, "Automated retrieval of CT images of liver lesions on the basis of image similarity: Method and preliminary results," Radiology **256**, 243–252 (2010).

[40]C. Muramatsu, Q. Li, R. A. Schmidt, J. Shiraishi, and K. Doi, "Determination of similarity measures for pairs of mass lesions on mammograms by use of BI-RADS lesion descriptors and image features," Acad. Radiol. **16**, 443–449 (2009).
[41]H. Alto, R. M. Rangayyan, and J. E. L. Desautels, "Content-based retrieval and analysis of mammographic masses," J. Electron. Imaging **14**, 1–17 (2005).
[42]C. Muramatsu, Q. Li, R. Schmidt, J. Shiraishi, and K. Doi, "Investigation of psychophysical similarity measures for selection of similar images in the diagnosis of clustered microcalcifications on mammograms," Med. Phys. **35**, 5695–5702 (2008).
[43]K. Horsch, M. L. Giger, C. J. Vyborny, L. Lan, E. B. Mendelson, and R. E. Hendrick, "Classification of breast lesions with multimodality computer-aided diagnosis: Observer study results on an independent clinical data set," Radiology **240**, 357–368 (2006).
[44]H. C. Cho, L. Hadjiiski, B. Sahiner, H.-P. Chan, M. Helvie, C. Paramagul, and A. V. Nees, "Similarity evaluation in a content-based image retrieval (CBIR) CADx system for characterization of breast masses on ultrasound images," Med. Phys. **38**, 1820–1831 (2011).
[45]W. J. Kuo, R. F. Chang, C. C. Lee, W. K. Moon, and D. R. Chen, "Retrieval technique for the diagnosis of solid breast tumors on sonogram," Ultrasound Med. Biol. **28**, 903–909 (2002).
[46]D. R. Chen, Y. L. Huang, and S. H. Lin, "Computer-aided diagnosis with textural features for breast lesions in sonograms," Comput. Med. Imaging Graph. **35**, 220–226 (2011).
[47]J. Cui, B. Sahiner, H. P. Chan, J. Shi, A. V. Nees, C. Paramagul, and L. M. Hadjiiski, "A computer-aided diagnosis system for prediction of the probability of malignancy of breast masses on ultrasound images," Proc. SPIE **7260**, 72600L1–72600L7 (2009).
[48]R. M. Haralick, K. Shanmugam, and I. Dinstein, "Texture features for image classification," IEEE Trans. Syst. Man Cybern. **6**, 610–621 (1973).
[49]P. A. Lachenbruch, *Discriminant Analysis* (Hafner, New York, 1975).
[50]J. Wei, H.-P. Chan, B. Sahiner, L. M. Hadjiiski, M. A. Helvie, M. A. Roubidoux, C. Zhou, and J. Ge, "Dual system approach to computer-aided detection of breast masses on mammograms," Med. Phys. **33**, 4157–4168 (2006).
[51]H. D. Cheng, X. J. Shi, R. Min, L. M. Hu, X. P. Cai, and H. N. Du, "Approaches for automated detection and classification of masses in mammograms," Pattern Recogn. **39**, 646–668 (2006).
[52]J. R. Quinlan, *C4.5: Programs for Machine Learning*, (Morgan Kaufmann, San Mateo, CA, 1993).
[53]T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning* (Springer-Verlag, New York, 2001).
[54]C. E. Metz, B. A. Herman, and J. H. Shen, "Maximum-likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data," Stat. Med. **17**, 1033–1053 (1998).
[55]K. Jarvelin and J. Kekalainen, "Cumulated gain-based evaluation of IR techniques," ACM Trans. Inf. Sys. **20**, 422–446 (2002).
[56]L. L. Yin, G. X. Xu, M. Torii, Z. D. Niu, J. M. Maisog, C. Wu, Z. Z. Hu, and H. F. Liu, "Document classification for mining host pathogen protein-protein interactions," Artif. Intell. Med. **49**, 155–160 (2010).
[57]J. R. Quinlan, "Induction of decision trees," Mach. Learn. **1**, 81–106 (1986).