Analysis of perceived similarity between pairs of microcalcification clusters in mammograms

Juan Wang, Hao Jing, and Miles N. Wernick

Department of Electrical and Computer Engineering, Medical Imaging Research Center, Illinois Institute of Technology, 3440 South Dearborn Street, Chicago, Illinois 60616

Robert M. Nishikawa

Department of Radiology, University of Pittsburgh, 3362 Fifth Avenue, Pittsburgh, Pennsylvania 15213

Yongyi Yang^{a)}

Department of Electrical and Computer Engineering, Medical Imaging Research Center, Illinois Institute of Technology, 3440 South Dearborn Street, Chicago, Illinois 60616

(Received 13 September 2013; revised 7 February 2014; accepted for publication 26 March 2014; published 16 April 2014)

Purpose: Content-based image retrieval aims to assist radiologists by presenting example images with known pathology that are visually similar to the case being evaluated. In this work, the authors investigate several fundamental issues underlying the similarity ratings between pairs of microcalcification (MC) lesions on mammograms as judged by radiologists: the degree of variability in the similarity ratings, the impact of this variability on agreement between readers in retrieval of similar lesions, and the factors contributing to the readers' similarity ratings.

Methods: The authors conduct a reader study on a set of 1000 image pairs of MC lesions, in which a group of experienced breast radiologists rated the degree of similarity between each image pair. The image pairs are selected, from among possible pairings of 222 cases (110 malignant, 112 benign), based on quantitative image attributes (features) and the results of a preliminary reader study. Next, the authors apply analysis of variance (ANOVA) to quantify the level of variability in the readers' similarity ratings, and study how the variability in individual reader ratings affects consistency between readers. The authors also measure the extent to which readers agree on images which are most similar to a given query, for which the Dice coefficient is used. To investigate how the similarity ratings potentially relate to the attributes underlying the cases, the authors study the fraction of perceptually similar images that also share the same benign or malignant pathology as the query image; moreover, the authors apply multidimensional scaling (MDS) to embed the cases according to their mutual perceptual similarity in a two-dimensional plot, which allows the authors to examine the manner in which similar lesions relate to one another in terms of benign or malignant pathology and clustered MCs.

Results: The ANOVA results show that the coefficient of determination in the reader similarity ratings is 0.59. The variability level in the similarity ratings is proved to be a limiting factor, leading to only moderate correlation between the readers in their readings. The Dice coefficient, measuring agreement between readers in retrieval of similar images, can vary from 0.45 to 0.64 with different levels of similarity for individual readers, but is higher for average ratings from a group of readers (from 0.59 to 0.78). More importantly, the fraction of retrieved cases that match the benign or malignant pathology of the query image was found to increase with the degree of similarity among the retrieved images, reaching average value as high as 0.69 for the radiologists (p -value $< 10^{-4}$ compared to random guessing). Moreover, MDS embedding of all the cases shows that cases having the same pathology tend to cluster together, and that neighboring cases in the plot tend to be similar in their clustered MCs.

Conclusions: While individual readers exhibit substantial variability in their similarity ratings, similarity ratings averaged from a group of readers can achieve a high level of intergroup consistency and agreement in retrieval of similar images. More importantly, perceptually similar cases are also likely to be similar in their underlying benign or malignant pathology and image features of clustered MCs, which could be of diagnostic value in computer-aided diagnosis for lesions with clustered MCs.

© 2014 American Association of Physicists in Medicine. [<http://dx.doi.org/10.1118/1.4870959>]

Key words: clustered microcalcifications, content based image retrieval, similarity rating, computer aided diagnosis, multidimensional scaling (MDS)

1. INTRODUCTION

Breast cancer is the most frequently diagnosed nonskin cancer in women in the United States. It is estimated that there are about 224 680 new breast cancer cases and 39 510 breast cancer deaths among women in the United States in 2012.¹ Numerous studies have shown that early detection of breast cancer can both reduce the mortality and increase treatment options.¹ Mammography is an effective screening tool for diagnosis of breast cancer, which can detect about 80%–90% of breast cancers in women without symptoms.¹ However, it can also yield a large number of false positives. It has been reported that benign lesions are often sent for biopsy, yielding a positive predictive value of only 20%–40%.²

Clustered microcalcifications (MCs) can be an important early sign of breast cancer in mammograms. MCs are tiny calcium deposits which appear as a group of granular bright spots in a mammogram image. They can be found in both malignant and benign cases in screening mammography.³ However, because of their subtlety in appearance in mammogram images, accurate diagnosis of MC lesions as benign or malignant is a very challenging clinical problem for radiologists.

There has been a great deal of effort devoted to development of computer-aided diagnosis (CADx) methods to assist in the interpretation of MC lesions in mammograms. The purpose of CADx is to provide a computerized “second-opinion” to radiologists in their diagnosis. Research results have demonstrated that CADx has the potential to improve radiologists’ performance in breast cancer diagnosis. For example, Jiang *et al.*^{4,5} showed that the use of a CADx system could both improve the diagnostic accuracy and reduce the variability among radiologists in their interpretation of mammograms with MC lesions. Similar improvement in diagnostic accuracy with CADx was also reported in mammograms with mass lesions.^{6–8}

The traditional CADx approach for MC lesions has been to use a classifier to estimate the likelihood of malignancy based on quantitative image features of a lesion.^{9–12} For example, Jiang *et al.*¹³ showed that by presenting this estimated likelihood as a second opinion, radiologists can improve their diagnostic accuracy and biopsy recommendations. However, one often-cited challenge in such a CADx approach is that it may be difficult for a radiologist to optimally incorporate a numerical value such as likelihood of malignancy into his or her decision without additional supporting evidence.

To alleviate this concern, we have been developing an alternative CADx approach for MC lesions, the goal of which is to assist radiologists by presenting them with a set of images similar to the one being evaluated, along with the known pathology of these reference images.^{14,15} The known reference images are obtained from a library of cases based on the concept of content-based image retrieval (CBIR). Such an approach has also been studied by other researchers and applied for different lesion types and imaging modalities.^{16–20} While these studies have collectively demonstrated the promise of using similar cases to assist in diagnosis, a major challenge

facing the CBIR approach is how to retrieve images that are truly perceived to be similar to the lesion under consideration. This problem is particularly challenging for mammogram lesions with clustered MCs, because their appearance is often subtle and difficult to interpret.

We have investigated the feasibility of using observer studies to measure the subjective impression of similarity by the radiologists in their interpretation of mammogram lesions with clustered MCs. With such an approach, readers are used to rate the degree of similarity between MC lesions presented to them; the numerical ratings thus obtained are then analyzed for consistency among readers. In a pilot study,²¹ we considered two methods for obtaining the radiologists’ perception of similarity between pairs of MC lesions (30 image pairs were used): quantitative scoring using a numerical scale, and paired comparison. The results showed that the readers were able to judge similarity of clustered MCs in a consistent way and there was good agreement between the similarity ratings by the two methods. Subsequently, we further studied the feasibility of modeling the notion of perceptual similarity by a machine learning algorithm.^{14,15} In an independent study, Muramatsu *et al.*²² used a set of 114 pairs of images with MCs (formed with 19 query images, six pairs for each) and demonstrated good consistency even among three different types of readers (breast radiologists, general radiologists, and nonradiologists) in their similarity ratings.

While these prior studies show that the notion of perceptual similarity can be quantified in a consistent way across readers, the similarity scores from individual readers are subject to high variability. This is due to both the subjective nature of the task and, more importantly, the subtlety of the image features of clustered MCs. The purpose of this study is twofold. First, we seek to quantify the degree of variability in the similarity ratings from individual readers, and investigate the effect of this variability on agreement in retrieval of similar images between readers. Our focus is particularly on image pairs that are relatively similar, as the purpose of using CBIR for CADx is to obtain similar cases. Second, we explore how the radiologists’ similarity rating might be related to the characteristics underlying the MC images. For example, are perceptually similar cases also similar in pathology? How are their similarity ratings related to the image features of the clustered MCs? To the best of our knowledge, these important issues have not been yet studied systematically.

In the first part of the study, we conduct a reader study to collect similarity ratings on a set of 1000 image pairs from a group of experienced breast radiologists. These 1000 image pairs are selected from a set of 222 cases, of which all have mammograms with clustered MCs. To ensure that these image pairs are sufficiently similar, we apply a two-step procedure to select images for pairing with each of the 222 cases based on the image features of the clustered MCs and the scoring from a preliminary reader study. We then apply analysis of variance (ANOVA) to quantify the effects of image pairs and reader variability on the similarity ratings. We also examine how the variability level in the reader ratings is related to the degree of consistency across readers. Furthermore, we apply the reader ratings in the context of retrieving

similar images to a given query, and quantify the agreement among readers by measuring the degree of overlap in their retrieved results (using the Dice measure from information retrieval).

In the second part of the study, to explore the potential of using similar cases in diagnosis of breast cancer, we apply the similarity scores from the readers to retrieve the most similar cases for each of the 222 cases, and measure the pathology matching ratio (fraction of retrieved images sharing the same benign or malignant pathology as the query image). We investigate how this matching ratio may change with the level of similarity in the retrieved images, the purpose being to determine whether MC lesions that are more perceptually similar are also more similar in pathology.

Furthermore, to study how the similarity scores from the readers might be correlated with the underlying image features among different cases, we apply multidimensional scaling (MDS) to embed the different cases in a two-dimensional (2D) plot, in which lesions are represented as points and the distances among the points reflect their degree of similarity as judged by the readers. Thus, cases that are similar are placed in close vicinity of each other while those not similar are placed further apart. Such a plot allows one to study how neighboring cases (i.e., cases similar to each other) may relate to one another. In particular, we will examine the relationships among the cases in several aspects, including: (1) case pathology (benign or malignant), (2) spatial distribution patterns of their clustered MCs, and (3) image pairs of clustered MCs that are highly similar.

2. MATERIALS AND METHODS

2.A. Reader study

2.A.1. Image dataset

The dataset used in this study was collected by the Department of Radiology at the University of Chicago (U of C). It consisted of 365 screen-film mammogram images from 222 cases, all of which had been pathologically proven by biopsy containing lesions with MCs. These images were obtained with a Lumiscan 85 film digitizer (Lumisys; Sunnyvale, CA). Among them, 110 cases are malignant (23 invasive, 87 DCIS), and 112 are benign (39 fibrocystic change, five ADH, one atypia, 67 unknown). The malignant images were obtained in consecutive cases sent for biopsy at the U of C. Likewise, the benign images represent consecutive benign cases at the U of C. These malignant and benign cases were collected during an overlapping time period. Among the 222 cases, 143 have images in both cranio-caudal (CC) and mediolateral-oblique (MLO) views. The MCs in each mammogram were manually identified by a group of experienced radiologists. Overall, there were 12 702 MCs identified in these mammograms. In our study, these MCs were used as ground truth for extraction of the image features for quantifying the lesions. Regions of interest (ROIs) were cropped in size of 512×512 or 1024×1024 pixels with a spatial resolution of 0.1 mm/pixel based on the sizes of the lesion areas.

2.A.2. Selection of image pairs for scoring

With 222 cases in the dataset, there can be potentially as many as 24 531 pairs of different cases formed for comparing their similarity (the number of possible mammogram pairs is even higher given that many cases have multiple views). However, it would be too time-consuming and also unnecessary to score all of them in a reader study, because most of the cases are not similar to each other and we are mainly interested in obtaining similar cases in a CBIR-based CADx system. Based on our past experience, we expect that for each case there will be on average fewer than ten other cases in the dataset that are found to be reasonably similar.¹⁵ Therefore, in order to economize on the reader resources, we apply the following two-step procedure to select only those potentially similar image pairs for reader scoring: (1) preselect an initial set of 2000 image pairs formed by cases with similar image features; and (2) conduct a preliminary reader study on these 2000 pairs, based on which a subset of 1000 image pairs is selected for the final study. This procedure is described in detail below.

2.A.2.a. Step 1: Initial set of 2000 image pairs. First, the mammogram images in the dataset were quantified by a set of descriptive features of their clustered MCs. The mammogram images were then selected based on this set of image features. Specifically, for each case in the dataset, a total of 15 other cases were selected for pairing based on their image features (as described below). For 222 cases, this yielded 3330 image pairs, among which only 2270 were unique image pairs (duplicate pairs were produced when two images were paired with each other when each as query). Afterward, a subset of 2000 image pairs was selected randomly from these unique pairs, resulting in an average of about 18 image pairs for each of the 222 case. When pairing two cases, only one pair of mammogram images from the two cases (one image from each case) was selected. When multiple views existed for the two cases, the image pair with the shortest distance between their image features was selected among all the possible pairs that could be formed from the two cases, and no pair was formed from the same case.

To characterize the clustered MCs, we used a set of nine descriptive features commonly used in computer-aided diagnosis of MC lesions in mammograms in the literature.^{9,12,23–29} These nine features can be categorized into three groups: (1) image features describing individual MCs (Feature Group 1), including the standard deviation of the image contrast values of the MCs, and the maximum and the standard deviation of the sizes of the MCs,^{9,12,23,24} (2) spatial clustering features of the MCs (Feature Group 2), including the number of MCs in a cluster, the area of the cluster, and the compactness of cluster,^{12,24,25} and (3) texture-based features (Feature Group 3), including the energy, contrast, and correlation derived from the gray-level cooccurrence matrices (GLCM).^{26–29} For completeness, the detailed definitions of these features are given in Appendix.

For each case as query, 15 pairing cases were selected using the following procedure from the rest of the cases in the dataset: First, three cases were selected as having the smallest Euclidean distance to the query separately based on *each* of

the following groups of features: (1) all the nine features, (2) Feature Group 1, and (3) Feature Group 2. Next, two cases were selected as having the smallest Euclidean distance to the query based on Feature Group 3. Finally, four cases were selected randomly from the dataset without using their features. The rationale for these randomly selected cases is to include cases that may potentially be perceptually similar but may differ in their quantitative image features.

2.A.2.b. Step 2: Final set of 1000 image pairs. To further refine the set of potentially similar pairs for reader scoring, we conducted a preliminary reader study on the initial set of 2000 image pairs selected above. As described below, five readers scored the similarity level for each of these 2000 image pairs based on their clustered MCs. Afterward, the scores from the five readers were averaged for each image pair. The averaged scores were then used to select a subset of 1000 image pairs from the initial set of 2000 image pairs. Specifically, for each case in the dataset, three pairs were selected with the highest similarity scores, and two pairs were selected with the shortest Euclidean distance (based on all nine features) among the remaining pairs. This process was repeated for each of the 222 cases in turn, resulting in a total of 1155 unique image pairs. A subset of 1000 pairs were then selected randomly from these unique pairs. Among these image pairs, 126 were from the random selected pairs above. There were about 9.8 image pairs on average for each case in the dataset.

2.A.3. Reader scoring sessions

2.A.3.a. Reader study on final set of 1000 image pairs. The reader study was carried out by a group of five radiologists for the 1000 image pairs, based on their perceptual similarity, using a discrete scale from 0 (most dissimilar) to 10 (most similar). These five radiologists are MQSA-qualified breast imagers with between 2 and 20 years of experience. To reduce the effects of reader fatigue, these 1000 pairs were randomly allocated into four separate reading sessions, with 250 pairs in each session, and a minimum of 24 h between sessions. These reading sessions were preceded by a “precalibration” session, during which the readers were provided with a set of five “anchor” image pairs spanning the rating scale (their scores were 1, 3, 5, 7, 9, respectively), which was determined upon consensus from a panel of six radiologists in our previous study.¹⁵ These anchor pairs were also made available for recall by the readers in subsequent scoring sessions. The purpose of this precalibration session was to establish a uniform measure of perceptual similarity among the readers. During the scoring sessions, the image pairs were presented to the reader one pair at a time in a random order; the images in each pair were displayed side-by-side and the reader scored their level of similarity. The pathology information of the cases was not revealed to the readers during their scoring.

The reader study was conducted in a dedicated observer-study reading room with the lights turned off. The image pairs were displayed on a Totoku 5 MP monitor (Totoku Electric Co Ltd., Japan). During the study the readers were allowed to adjust the brightness and contrast, and magnify or minify the images.

2.A.3.b. Preliminary study on initial set of 2000 image pairs. As explained above, a preliminary study was performed on the initial set of 2000 image pairs for the purpose of refining the choice of potentially similar image pairs for the final scoring. This reader study was carried out by a group of five nonradiologists with a format similar to the reader study above. These readers were researchers in breast imaging with a minimum of 5 years of experience. However, due to the larger number of image pairs in the dataset, a total of ten separate sessions were used, with 200 pairs randomly assigned to each session.

2.B. Analyses of reader scores

2.B.1. Quantify reader variability in similarity reading

Because of the subjective nature in interpretation of clustered MCs in mammogram images, readers can vary in their similarity scores. While the use of a precalibration session may help reduce this variation, some individual differences may still remain. To suppress such apparent differences, we first transform the similarity scores from individual readers into *z*-scores; i.e., the set of scores from each reader is normalized to have zero mean and unit standard deviation. Because of its linearity, this normalization step preserves the relative similarity ranking of a set of image pairs; thus, two readers who differ in their similarity scores only by a constant scaling factor and a constant offset would rank the relative ordering of retrieved images in an identical way, and the normalization step will result in identical scores for the two readers. Similarly, when comparing the scores averaged from a group of readers, the averaged group scores are also normalized to have zero mean and unit standard deviation. Henceforth, we will describe the individual readers by their normalized scores, and the average of a group of readers by the average of the individual normalized scores upon further normalization.

To quantitatively analyze the reader scores, we model the similarity score given by a reader to an image pair as

$$SC = s + n, \quad (1)$$

where *s* denotes the ensemble mean score of the image pair (among the reader population) and *n* denotes the deviation (i.e., modeling error) from the mean. The error component *n* characterizes the variability in the reader score *SC* resulting from the subjectivity of judging perceptual similarity. Our goal is to determine the level of the error term *n* from the reader data collected on 1000 image pairs. For this purpose, we will apply the ANOVA technique,^{30,31} which is a statistical tool for analysis of linear observation models. By ANOVA, the variations in the readers' similarity scores are decomposed into two sources, one due to the differences in the image pairs and the other due to the error term. A statistical test is performed to determine the significance of the contribution by the image pairs in the reader scores. The mean of squares for the error term provides an unbiased estimate of the population variance.³¹

2.B.2. Inter-reader consistency in similarity reading

To quantitatively measure the agreement between two readers, we use the Pearson correlation coefficient ρ . By definition, the correlation coefficient measures the linear association between a pair of variables (the similarity scores from two readers in our case). Its value is 1 when the two variables are in perfect agreement with each other, nearly 0 when they have a little agreement. The correlation coefficient also allows us to analyze the relationship between reader consistency and the ANOVA results, as will be discussed later in Sec. 3.

To visualize the relationships among the readers, we apply MDS,³² a data visualization method by which data points having known relationships to one another can be represented in a two- or three-dimensional scatter plot. In a MDS plot, the data points are arranged in such a way that the distance between any two points reflects the relative degree of “dissimilarity” between those points. Specifically, consider a set of N data points. The MDS seeks to embed these data points in a lower-dimensional space by minimizing the following objective function:

$$\sigma_1^2 = \frac{\sum w_{ij}[d(\mathbf{x}_i, \mathbf{x}_j) - \delta_{ij}]^2}{\sum w_{ij}\delta_{ij}^2}, \quad (2)$$

where $\mathbf{x}_i$ are the embedded data points, δ_{ij} are their pairwise proximity measure, w_{ij} are their corresponding weight factors (specified by users), and $d(\mathbf{x}_i, \mathbf{x}_j)$ denotes the Euclidean distance between points $\mathbf{x}_i$ and $\mathbf{x}_j$, i.e., $d(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{x}_i - \mathbf{x}_j\|$. Consequently, the original data points are represented by the resulting points $\mathbf{x}_i$, $i = 1, \dots, N$, which are in a two- or three-dimensional MDS space (for easy visualization). The MDS plot of the data points is obtained by the embedded points $\mathbf{x}_i$, $i = 1, \dots, N$. The quantity σ_1 is known as Stress-1,³² which measures the goodness of fit of the MDS model. In our application, we represent each reader as a point in a 2D plane, wherein the dissimilarity between a pair of readers is defined according to their correlation coefficient. Specifically, consider two readers i and j having correlation coefficient ρ_{ij} . Their corresponding points in the 2D MDS plot are to be separated approximately by distance $d_{ij} \approx 1 - \rho_{ij}$. Thus, readers close in the MDS plot are those exhibiting higher inter-reader consistency in their readings. In our implementation, the iterative majorization method³² was used for optimizing the objective function in Eq. (2). Also, unless specified otherwise, the weight factor w_{ij} was set to 1 in the subsequent results.

2.B.3. Evaluate inter-reader agreement in retrieval of similar cases

The inter-reader consistency analysis described in Sec. 2.B.2 pertains to overall agreement between readers on the entire set of similarity scores. However, in CBIR we are more interested in readers agreement about image pairs with a high level of similarity, as these are the ones that may actually be retrieved. Conceptually, this could be accomplished by the following procedure: for each query, have readers to select cases from a reference dataset that are most similar to the query, then analyze the overlap among

the selected images. However, such a procedure would be impractical because it would require every reader to compare every case in the dataset with the query.

To avoid this problem, we utilize the similarity scores already collected in the reader study as follows. For each query case, we use the reader scores to retrieve the most similar cases among those paired with the query for scoring (in the set of 1000 pairs); those cases with scores higher than a pre-defined similarity threshold are retrieved for the query. This process is repeated for each case in the dataset being used in turn as the query.

To measure the agreement between the results retrieved by different readers, we use the Dice measure from information retrieval for assessing the similarity between two sets of retrieval results.^{33,34} Suppose that for a given query the retrieved sets of cases by two readers are denoted by A and B , respectively. Then the agreement between the two sets is measured by the Dice coefficient, which is defined as

$$D(A, B) = \frac{2|A \cap B|}{|A| + |B|}, \quad (3)$$

where $|\cdot|$ denotes the cardinality of a set. A higher Dice coefficient indicates better agreement between the two sets, with 0 indicating no overlap and 1 indicating perfect agreement. In particular, when the two sets are equal in size, the Dice coefficient corresponds to the fraction of cases in common between the two sets. For example, suppose A and B each have two retrieved cases, of which one is common between them; then the Dice coefficient is 0.5. The Dice coefficient is averaged over all the query cases when comparing two different readers.

2.B.4. Explore the relationship between perceptual similarity and case pathology

Given that the ultimate goal of obtaining similar cases is to assist the diagnosis of the query case being under consideration, it is interesting to explore how the retrieved similar cases might relate to the latter in terms of the underlying pathology. For example, are similar looking cases also correlated in their pathology? To investigate the potential interconnection between the notion of perceptual similarity and pathology, we explore two methods by making use of the similarity scores collected in the reader study. The first method is to examine the pathology matching ratio in the context of CBIR, where the fraction of the cases sharing the same benign or malignant pathology as the query among all the retrieved is computed. The second method is to apply the MDS technique to embed the different cases according to their perceptual similarity so that one can explore how similar cases are related to each other in terms of both their benign or malignant pathology and underlying image features.

2.B.4.a. Disease matching ratio of similar cases. As in the analysis of agreement in retrieval of similar cases above, we will use each case in turn as the query and retrieve according to the reader scores the most similar cases among those paired with the query (in the set of 1000 pairs). The pathology of the query cases is then compared with that of the top retrieved case, and the matching fraction between the two

(averaged over all query cases) is computed. In the scenario that there are ties among the top retrieved images for a given query, the fraction among the ties with matched pathology to the query is used. In addition, because the number of cases in the dataset is somewhat limited, not all the queries may have similar cases to them. Thus, a query case without sufficiently similar retrieval is not included in computing the matching ratio, since only the cases that are similar to the query are of interest. A retrieved case is deemed to be sufficiently similar to the query if its similarity score is above a preset threshold. By varying this threshold, we can also examine how the pathology matching ratio may vary with the level of similarity among the images.

To remove the effect of the case distribution in the dataset, we will adopt a bootstrapping resampling procedure³⁵ on the query cases to obtain the statistics of the resulting pathology matching ratio. The significance of the reader similarity score is obtained by comparison with retrieval by random guessing.

2.B.4.b. MDS embedding of similar cases. To explore the potential underlying relationship among the different cases in the dataset, we use the MDS technique to embed all the cases in the dataset on a 2D plot, where cases that are perceptually similar are placed to be close to each other, and vice versa. For this purpose, the proximity measure between a pair of cases is defined according to their similarity score. Specifically, consider two cases i and j with similarity score SC_{ij} ; their corresponding points in the MDS space will be separated approximately by distance $\delta_{ij} \approx \frac{1}{3.75+SC_{ij}}$, where a constant offset 3.75 (over three standard-deviations) is added to insure that d_{ij} is positive. Thus, similar cases will be accordingly placed closely to each other in the plot. It is noted that, unlike the MDS plot of the readers above, here the pairwise proximity measure is available for only a small set of the image pairs; thus, we will use the weighted MDS technique³² where the unscored image pairs are assigned a weight of 0 and consequently not included in the objective function. With the resulting MDS plot, we will explore the distribution of the different cases in the dataset by also examining the image features of their underlying clustered MCs. In particular, we will examine how the similar cases may relate to each other in their image patterns.

3. RESULTS AND DISCUSSIONS

3.A. Reader variability and consistency in similarity reading

3.A.1. Level of variability in similarity scoring

To quantitatively evaluate the level of variability of the readers in scoring the image pairs, we applied one-way ANOVA analysis to the scores from the five radiologists on the set of 1000 image pairs. The results are summarized in Table I, where the following entities are given: source of variability (SoV), sum of squares (SS), degrees of freedom (DoF), and mean squares (MS). It is noted that the p -value for the effect of image pairs is nearly 0 ($\ll 10^{-4}$), indicating that the difference among the image pairs is a significant contributor to the total variance in the reader scores. In other words, the

TABLE I. One-way ANOVA table for five radiologists.

SoV	SS	DoF	MS	F statistics	p -value
Image pairs	2952.8	999	2.9558	5.7894	≈ 0
Residuals	2042.2	4000	0.5105		
Total	4995.0	4999			

reader scores do vary according to the level of similarity in the image pairs. Indeed, from Table I, the coefficient of determination is $R^2 = 1 - \frac{SS_{\text{of residuals}}}{SS_{\text{Total}}} = 0.5912$, which shows that about 59% of the total variation in the reader scores is attributed to the source of image pairs. Moreover, from Table I it is noted that the variance of the residual (i.e., modeling error) is estimated to be 0.5105, which is rather high given that the variance of the (normalized) scores from the individual readers is 1. This is likely associated with the fundamental difficulty by the readers in their interpretation of mammogram images with clustered MCs. As to be shown below, this error level is also responsible for the consistency level among the individual readers.

For comparison, one-way ANOVA analysis was also applied to the scores from the five nonradiologists on the same set of 1000 image pairs. The results are summarized in Table II. From these results, the coefficient of determination is $R^2 = 1 - \frac{2422.8}{4995.0} = 0.5150$, which is lower than that from the radiologists above. This indicates that the level of variability is higher in the scores from the nonradiologists. Indeed, from Table II the variance of the residual is 0.6057, which is higher than that of the radiologists above. This is likely attributed to the experience level of the readers in reading mammogram images with clustered MCs.

3.A.2. Reader consistency and MDS visualization

To evaluate the consistency among the different readers in their reading of the image pairs with clustered MCs, we computed the Pearson correlation coefficients in a pairwise fashion for all the five radiologists. The results are shown in Table III, where each reader is denoted by R_i , $i = 1, \dots, 5$. Owing to the symmetry of the correlation coefficient, the entries in the lower triangular portion are omitted in Table III. As can be seen, among the different readers the correlation coefficient values range from 0.4250 (between R_3 and R_5) to 0.5507 (between R_1 and R_2). Moreover, the average correlation coefficient among the different readers is 0.4889 (standard deviation is 0.0450). From Table III, we can see that among the individual readers R_1 is most correlated with the others (average correlation coefficient = 0.5161), while R_5 is the least correlated (average correlation coefficient = 0.4547).

TABLE II. One-way ANOVA table for five nonradiologists.

SoV	SS	DoF	MS	F statistics	p -value
Image pairs	2572.2	999	2.5748	4.2510	≈ 0
Residuals	2422.8	4000	0.6057		
Total	4995.0	4999			

TABLE III. Pearson correlation coefficients for five radiologists.

	R_1	R_2	R_3	R_4	R_5
R_1	1	0.5507	0.5095	0.5424	0.4618
R_2	...	1	0.4595	0.5486	0.4682
R_3	...	...	1	0.4600	0.4250
R_4	...	...	...	1	0.4638
R_5	...	...	...	...	1

These results indicate the correlation level among the individual readers is only moderate. As to be explained below, this is largely attributed to the uncertain level in their similarity scores.

For comparison, in Table IV we show the correlation coefficients for the scores from the five nonradiologists on the same set of 1000 image pairs. In the table, each reader is denoted by nR_i , $i = 1, \dots, 5$. As can be seen, the correlation coefficient values range from 0.2751 (between nR_1 and nR_3) to 0.5225 (nR_2 and nR_4); the average correlation coefficient among the different readers is 0.3941 (standard deviation = 0.0738). Compared with the radiologist readers in Table III, these results show that the average inter-reader correlation is much lower and the variation is larger among the nonradiologist readers. As to be demonstrated below, this is also consistent with the larger variability level in their similarity scores.

In addition, we also computed the inter-reader correlation coefficients between the radiologists and nonradiologists on the same set of 1000 image pairs. The results are shown in Table V. The average correlation coefficient between these two groups is 0.3873 (standard deviation = 0.0499). Interestingly, the average value is similar to that among the nonradiologists above.

Finally, based on the inter-reader correlation results in Tables III–V, we show in Fig. 1 a MDS embedding plot of all the different readers (Stress-1 = 0.18; $w_{ij} = 1$ was used). In this plot, each reader is represented by a point (indicated by a dot), and the pairwise distances between the different points indicate how correlated their corresponding readers are. Thus, the plot provides an intuitive visualization of how the different readers are related to each other. Besides the individual readers, in Fig. 1 we also include the following two group readers: “ $\bar{R}$ ” is the average of the similarity scores of the radiologists, and “ $\bar{nR}$ ” is the average of the nonradiologists. From the MDS plot, it can be seen that the radiologists are all placed within a small cluster on the left half of the plot, while the nonradiologists are all placed on the right half and

TABLE IV. Pearson correlation coefficients for five nonradiologists.

	nR_1	nR_2	nR_3	nR_4	nR_5
nR_1	1	0.3520	0.2751	0.4290	0.3406
nR_2	...	1	0.3996	0.5225	0.4081
nR_3	...	...	1	0.3976	0.3310
nR_4	...	...	...	1	0.4851
nR_5	...	...	...	...	1

TABLE V. Pearson correlation coefficient between radiologists and nonradiologists.

	R_1	R_2	R_3	R_4	R_5
nR_1	0.4107	0.4756	0.3494	0.4530	0.3930
nR_2	0.3768	0.3780	0.3822	0.3947	0.3994
nR_3	0.4088	0.3590	0.3462	0.4058	0.3356
nR_4	0.4479	0.4537	0.4282	0.4230	0.4384
nR_5	0.2742	0.3601	0.2928	0.3435	0.3536

more scattered. This is consistent with the higher correlation coefficients among the radiologists in Tables III and IV. It is also observed that nR_1 , nR_3 , and nR_5 are individually placed further away from other readers, reflecting their lower inter-reader agreement with other readers (Table V). More importantly, it is noted that two average readers “ $\bar{R}$ ” and “ $\bar{nR}$ ” are closely placed next to each other in the plot, indicating that the averaged scores from the radiologists are highly similar to that from the nonradiologists. Indeed, the correlation coefficient between them is 0.6977, much higher than that between any individual readers. This indicates that despite the higher variability level in the individual scores of the nonradiologists, their averaged scores can be highly similar to that of radiologists. As explained below, this can be attributed to the reduced variability level in the averaged scores from the different readers.

3.A.3. Relationship between scoring variability and reader consistency

In the above analyses, the ANOVA results were used to quantify the level of variability in the similarity scores from a group of readers, while the Pearson correlation coefficients were used to characterize the level of consistency between individual readers in their readings. Interestingly, both

FIG. 1. MDS plot of the five radiologists, five nonradiologists, and their group averages.

lower variability level and higher inter-reader correlation were observed in the scores of the radiologists when compared to that of the nonradiologists. Below we demonstrate that the inter-reader correlation is mainly attributed to the level of variability in the reader scores.

As in ANOVA, consider the following additive model for the similarity score on an image pair from reader i :

$$SC_i = s_i + n_i, \quad (4)$$

where SC_i denotes the reader score, s_i denotes its expectation, and n_i denotes the variability error. For convenience, assume that the reader scores are prenormalized to have zero mean and unit standard deviation. Then, the correlation coefficient between any two readers i and j can be computed as

$$\rho_{ij} = E[SC_i SC_j] = E[(s_i + n_i)(s_j + n_j)]. \quad (5)$$

Under the assumption that the error terms are independent among the different readers and that $E[n_i] = 0$, we have

$$\rho_{ij} = E[s_i s_j]. \quad (6)$$

Consider the scenario that the individual mean scores s_i, s_j of an image pair are not too different from the ensemble mean s in Eq. (4), i.e., $s_i \approx s, s_j \approx s$. The correlation coefficient between two readers can be approximated as

$$\rho_{ij} \approx E[s^2] = 1 - \sigma_n^2, \quad (7)$$

where σ_n^2 is the variance of the residual term in the linear model in Eq. (4). Thus, the correlation coefficient is directly related to the variance of the residual term, which is determined by the variability level of the readers.

From the ANOVA results of the radiologists in Table I, the variance of the residuals is 0.5105. According to Eq. (7), the inter-reader correlation coefficient is expected to be $\rho_{ij} = 1 - 0.5105 = 0.4895$. Interestingly, this closely matches the average value (0.4889) of the correlation coefficients in Table III.

Similarly, for the nonradiologists, the variance of the residuals is 0.6057 from Table II. Thus, the inter-reader correlation coefficient is expected to be $\rho_{ij} = 1 - 0.6057 = 0.3943$. This also closely matches the average value (0.3941) of the correlation coefficients in Table IV.

Finally, in the MDS plot above the average readers “ $\bar{R}$ ” and “ $\bar{nR}$ ” were observed to be highly correlated. This can also be explained by Eq. (7): the noise level σ_n^2 is reduced by averaging compared with individual scores; consequently, the value of the correlation coefficient gets higher than that of the individual scores.

3.B. Reader agreement in retrieval of similar images

We first evaluated the reader agreement in retrieval of similar images among the individual radiologists. With five readers, each reader was compared with another, resulting in ten different pairings. For each pairing of readers, their scores were applied separately for retrieval of similar images and their agreement was measured by the Dice coefficient (as described in Sec. 2.B.3). The results are summarized in Fig. 2, where both the average and standard deviation of the Dice coefficient (from the ten different reader pairings) are shown.

FIG. 2. The Dice coefficients for retrieval of similarity images based on scores of individual readers (denoted by “Individual”) and group averages (denoted by “Group”).

These results were obtained for each case in the dataset used in turn as query; in the scenario that a query failed to yield similar cases for retrieval according to both of the paired readers, it was excluded in computing the Dice measure. In these results, the following different values were used for defining the similarity threshold for retrieval: 0, 0.25, and 0.5. Note that 0 corresponds to the mean of all the reader scores (i.e., retrieving images with above-average similarity); 0.25 corresponds to $\frac{1}{4}$ standard deviation above the mean, and 0.5 corresponds to $\frac{1}{2}$ standard deviation above the mean (top 15.9% of the similarity scores).

We next evaluated the reader agreement in retrieval of similar images when their average scores were used. The MDS results earlier showed that the average scores of the radiologists were highly correlated with that of the nonradiologists. Thus, to maximize the number of readers for averaging, the scores from both radiologists and nonradiologists were pooled together in this comparison. From this pool of ten readers, we randomly divided them into two separate groups (five readers each). Afterward, the similarity scores were averaged for each group, and used for retrieval of similar images. The agreement between the two groups was then measured by the Dice coefficient. This process was repeated 20 times, during each of which the ten readers were regrouped. The obtained results are also shown in Fig. 2 for comparison.

From Fig. 2, it can be seen that the Dice coefficient is considerably higher for the group average readers than that of the individual readers. Consistent with our previous analysis results of the reader consistency and variability, the group average scores yield better agreement in retrieval of similar images. It is also noted that for both individual readers and group averages, the Dice coefficient values get decreased with the similarity threshold increased (from 0.64 to 0.45 for individual readers, and from 0.78 to 0.59 for group averages). This indicates that the reader agreement is more hampered on images that are more similar. That is, the readers are more likely to agree in judging a pair of images that are less similar than those that are more similar. We note that this behavior is also consistent with the observation model in Eq. (1). Consider the ratings from two observers on a given image pair. From

Eq. (1), if one observer rates it higher than the similarity threshold (indicating likely high mean value), then the other observer is also likely to rate it higher than the same threshold; in addition, as the threshold decreases, the probability increases for a rating to exceed the threshold. Thus, one would expect the Dice coefficient to increase with decreasing threshold, and vice versa.

3.C. Relationship between similarity scores and case disease conditions

3.C.1. Disease matching ratio in retrieved similar cases

To explore whether perceptually similar cases are also similar in their disease conditions, we computed the disease matching ratio between a query case and the top retrieved case based on their similarity. The results are summarized in Fig. 3, where the average scores from the five radiologists (i.e., $\bar{R}$) were used for retrieval of similar images for each query. To demonstrate the effect of the similarity level of the retrieved cases, we varied the similarity threshold for retrieval from 0.5 to 1. These results were obtained from a total of 10 000 bootstrap samples on the query cases in the dataset, based on which the mean and standard deviation of the matching ratio were obtained. For comparison, similar results are also given in Fig. 3 for retrieval using the average scores of the five nonradiologists ($\bar{nR}$). As reference, the expected matching ratio by retrieval with random guessing is also shown.

As can be seen in Fig. 3, the disease matching ratio for $\bar{R}$ is increased from 0.6205 to 0.6911 as the similarity threshold varied from 0.5 to 1. A similar increasing trend is also observed for $\bar{nR}$. These results indicate that more similar images among the retrieval are also more likely to be similar in disease to the query. Compared to retrieval by random guessing, both $\bar{R}$ and $\bar{nR}$ have achieved a higher disease matching ratio [at similarity threshold 1, their p -values are nearly 0 ($<10^{-4}$) and 0.0068, respectively]. The disease matching ratio is also observed to be higher for $\bar{R}$ than $\bar{nR}$ (at similarity threshold 1, p -value is 0.0477). This could be attributed to two possible factors: one is the lower level of variability in the scores of the

radiologists, and the other is that the radiologists likely used image features more related to the disease when comparing the clustered MCs.

We note that it might be interesting to examine further how the similarity ratings might be correlated with the pathology in terms of the disease types, e.g., disease matching ratio of DCIS vs that of invasive cancers. However, it would require that the cases had been balanced among the different pathology types to warrant statistical and experimental rigor. Nevertheless, this can be of interest in future research.

3.C.2. MDS distribution of perceptually similar cases

Finally, to explore the potential relationship among the different cases as revealed by the similarity scores from the readers, we applied the MDS technique to embed all the cases in the dataset in a 2D plot for visual assessment. Due to the fact that the pairwise similarity scores are available for only those image pairs scored by the human readers, we used the scores from both the five radiologists (on the set of 1000 image pairs) and the five nonradiologists (on the set of 2000 image pairs). The scores from the ten readers were averaged on their common set of 1000 image pairs (denoted by S_1), and the scores from the five nonradiologists were averaged for the remaining 1000 image pairs (denoted by S_2). The weight factor in the MDS proximity measure was adjusted for the different pairs according to the level of similarity and the level of variability as follows: for image pair i with score SC_i ,

$$w_i = \begin{cases} 1 & \text{if } i \in S_2 \text{ and } SC_i \leq 0, \\ 1.5 & \text{if } i \in S_1 \text{ and } SC_i \leq 0, \\ 2 & \text{if } i \in S_2 \text{ and } SC_i > 0, \\ 3 & \text{if } i \in S_1 \text{ and } SC_i > 0. \end{cases} \quad (8)$$

The rationale for such a choice is to assign a higher weight value to pairs that are more similar and scored by more readers. The resulting MDS plot of all the cases is shown in Fig. 4 (Stress-1 = 0.19), where cancer cases and benign cases are represented by different symbols (dot for cancer, square for benign). For those cases with multiple views, the different views are shown separately in the plot.

As can be seen from Fig. 4, while there is no clear separation between cancer and benign cases, it can be seen that there are more cancer cases (and fewer benign cases) in the right half of the plot than in the left half. More importantly, cases of same disease tend to be clustered together locally. For example, about 70.0% of the cases are malignant in the rectangular region (indicated by dashed lines) in the middle-right section of Fig. 4, whereas only 34.1% of the cases are malignant in the rectangular region in the middle-left section of the plot. Interestingly, this is consistent with the higher disease matching ratio achieved by top retrieved images earlier in Fig. 3.

Furthermore, to explore how the readers' notion of similarity might be related to the image features of the clustered MCs, we also show in Fig. 4 for a number of example cases the spatial distribution patterns of their underlying clustered MCs; these examples are selected throughout the different

FIG. 3. Disease matching ratio achieved by retrieval based on similarity scores of radiologists and nonradiologists.

FIG. 4. MDS embedding of perceptually similar cases in the dataset, wherein cancer cases are represented by “red dots” and benign cases are represented by “blue squares.” The spatial distribution patterns of clustered MCs are shown for some sample cases, where the spatial locations of the MCs are indicated by “green plus signs.”

regions in the plot. For each example, the spatial locations of the individual MCs in a cluster are indicated by “+” signs. From Fig. 4, it can be seen that the neighboring cases tend to have MC clusters similar in size (i.e., number of MCs) and shape (i.e., spatial distribution pattern of MCs). Moreover, it is noted that the MC clusters tend to be bigger in size (more MCs) and more irregular in shape in the right half of the plot than in the left half. For example, the average number of MCs in a cluster is 38.97 for the cases in the marked rectangular region in the middle-right section of Fig. 4, compared to 25.24 for those in the marked rectangular region in the middle-left section of the plot. Similarly, the average compactness of the cluster is 0.65 for the cases in the middle-right rectangular region, compared to 0.69 for those in the middle-left rectangular region. Interestingly, this also coincides with the observation above that there are also more cancer cases in the right half. This indicates that the readers likely considered the clustering features of MCs (both the number of MCs and their spatial distribution) when scoring image pairs.

Finally, to closely examine the image features of similar cases, we consider a few selected sample pairs for which the cases in each pair were closely placed next to each other in the MDS plot. In Fig. 5, we show two sample image pairs selected from the upper-right and lower-right regions of the MDS plot; the cases in both pairs are all cancer. Similarly, in Fig. 6 we show two sample image pairs selected from the middle-left and lower-left regions of the MDS plot; the cases in both pairs are all benign. For better visualization, we have adjusted the images in these figures by suppressing the tissue background so the individual MCs become more visible. In Fig. 5, the

MCs are noted to be more irregular and have higher contrast; in Fig. 6, the MCs are noted to be more round and have lower contrast. Consistent with the observation above, these results indicate that the readers likely also considered the features

FIG. 5. Examples of neighboring image pairs selected from the upper-right (a) and lower-right (b) regions in the MDS plot. All cases are malignant.

FIG. 6. Examples of neighboring image pairs selected from the middle-left (a) and lower-left (b) regions in the MDS plot. All cases are benign.

of individual MCs (such as their contrast and shape) in their scoring.

4. CONCLUSION

In this study, we investigated the level of variability in subjective reader ratings of similarity between MC lesions, and examined its effects in the context of retrieval of similar images toward CADx. We also explored how the similarity ratings from readers might relate to the characteristics underlying the clustered MCs in the mammogram images [from disease condition (benign or cancer) to image features]. For these purposes, we conducted an observer study to collect similarity scores on a set of image pairs selected from 222 mammogram cases with a group of human readers. The ANOVA results showed that only up to approximately 59% of the total variation in the similarity scores could be attributed to the effect of the presented images. The variability level in the similarity scores is also a main contributor to the somewhat moderate agreement between individual readers in their ratings. In spite of variations among individual readers, the average similarity ratings from a group of readers could achieve a high level of reader consistency and reader agreement in retrieval of similar images. Moreover, it was found that perceptually similar cases are also likely similar in their underlying pathology and their clustered MCs. These results indicate that perceptually similar lesions could be of diagnostic value in CADx for clustered MCs. In future work, we plan to further investigate whether retrieving similar cases would aid radiologists in their diagnosis of mammogram lesions with clustered MCs.

ACKNOWLEDGMENT

This work was supported by NIH/NIBIB under Grant No. R01EB009905. R. M. Nishikawa receives royalties from Hologic, Inc.

APPENDIX: IMAGE FEATURES OF CLUSTERED MCS

The following nine features are used for quantifying the clustered MCs in a lesion:

- (1) The standard deviation of the size of individual MCs in a cluster, where the size of the individual MC is measured by the number of pixels in the MC;
- (2) The maximum of the size of the individual MCs in a cluster;
- (3) The standard deviation of the contrast of the individual MCs in a cluster, where the contrast of the MC is computed as mean pixel value of the MC above that of the background;
- (4) The number of MCs in a cluster;
- (5) The area of the cluster A , computed as the number of pixels in the cluster region which is constructed by the morphology dilation of the MCs;
- (6) The compactness of cluster C , a measure of the roundness of the cluster region, it is calculated by

$$C = \frac{4\pi A}{P^2},$$

where A and P are the area and perimeter of the cluster.

(7)–(9) The energy, contrast, and correlation of the textures computed from the GLCM. The GLCM is a well-established robust tool for extracting second texture information in an image.²⁹ It represents the distribution of joint probability of occurrence of a pair of gray-level values separated by a given displacement d . In other words, it calculates how often a pixel with gray-level value i occurs adjacent to a pixel with the gray-level value j . In this study, the GLCM is obtained by setting $d = [10, 0]$, that is, the distance of 10 pixels in the horizontal direction is used. Then the contrast, correlation, and energy of the GLCM are computed as

$$\text{Contrast} = \sum_{i,j} |i - j|^2 p(i, j),$$

$$\text{Correlation} = \sum_{i,j} \frac{(i - \mu_i)(j - \mu_j)p(i, j)}{\sigma_i \sigma_j},$$

$$\text{Energy} = \sum_{i,j} p(i, j)^2,$$

where p is the GLCM, μ_i and μ_j denote the mean of the GLCM in the row and column directions, respectively, and σ_i and σ_j are the standard deviation of the GLCM in the row and column direction separately. Among the nine features, the first three are the image features of the MCs, the middle three are spatial clustering features of the MCs, and the last three are texture based features.

^{a)} Author to whom correspondence should be addressed. Electronic mail: yangyo@iit.edu

- ¹Cancer Facts and Figures (American Chemical Society, Atlanta, GA, 2012).
- ²E. A. Sickles, D. L. Miglioretti, R. Ballard-Barbash, B. M. Geller, J. W. T. Leung, R. D. Rosenberg, R. Smith-Bingman, and B. C. Yankaskas, "Performance benchmarks for diagnostic mammography," *Radiology* **235**, 775–790 (2005).
- ³D. B. Kopans, *Breast Imaging*, 3rd ed. (Williams and Wilkins, Baltimore, MD, 2007).
- ⁴Y. Jiang, R. M. Nishikawa, R. A. Schmidt, C. E. Metz, and K. Doi, "Relative gains in diagnostic accuracy between computer-aided diagnosis and independent double reading," *Medical Imaging 2000: Image Perception and Performance* (SPIE Press, Bellingham, WA, 2000), pp. 10–15.
- ⁵Y. Jiang, R. M. Nishikawa, R. A. Schmidt, A. Y. Toledano, and K. Koi, "Potential of computer-aided diagnosis to reduce variability in radiologists' interpretations of mammograms depicting microcalcifications," *Radiology* **220**, 787–794 (2001).
- ⁶Z. Hou, M. L. Giger, C. J. Vyborny, and C. E. Metz, "Breast cancer: Effectiveness of computer-aided diagnosis – Observer study with independent database of mammograms," *Radiology* **224**, 560–568 (2002).
- ⁷K. 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).
- ⁸H. P. Chan, B. Sahiner, M. A. Helvie, N. Petrick, M. A. Roubidoux, T. E. Wilson, D. D. Adler, C. Paramagul, J. S. Newman, and S. Sanjay-Gopal, "Improvement of radiologists' characterization of mammographic masses by using computer-aided diagnosis: An ROC study," *Radiology* **212**, 817–827 (1999).
- ⁹Y. 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).
- ¹⁰W. J. H. Veldkamp and N. Karssemeijer, "Automated classification of clustered microcalcifications into malignant and benign types," *Med. Phys.* **27**, 2600–2608 (2000).
- ¹¹L. Wei, Y. Yang, and R. M. Nishikawa, "A study on several machine-learning methods for classification of malignant and benign clustered microcalcifications," *IEEE Trans. Med. Imaging* **24**, 371–380 (2005).
- ¹²I. Andreadis, G. M. Spyrou, and K. S. Nikita, "A comparative study of image features for classification of breast microcalcifications," *Meas. Sci. Technol.* **22**, 114005 (2011).
- ¹³Y. Jiang, R. M. Nishikawa, R. A. Schmidt, C. E. Metz, M. L. Giger, and K. Doi, "Improving breast cancer diagnosis with computer-aided diagnosis," *Acad. Radiol.* **6**, 22–33 (1999).
- ¹⁴I. Ei-Naqa, Y. Yang, N. P. Galatsanos, R. M. Nishikawa, and M. N. Wernick, "A similarity learning approach to content-based image retrieval: Application to digital mammography," *IEEE Trans. Med. Imaging* **23**, 1233–1244 (2004).
- ¹⁵L. Wei, Y. Yang, M. N. Wernick, and R. M. Nishikawa, "Learning of perceptual similarity from expert readers for mammogram retrieval," *IEEE J. Sel. Top. Signal Process.* **3**, 53–61 (2009).
- ¹⁶L. Yang, R. Jin, R. Sukthankar, B. Zheng, L. Mummert, M. Satyanarayanan, M. Chen, and D. Jukic, "Learning distance metrics for interactive search-assisted diagnosis of mammograms," *Medical Imaging 2007: Computer-Aided Diagnosis* (SPIE Press, Bellingham, WA, 2007), p. 6514H.
- ¹⁷G. 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).
- ¹⁸C. 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).
- ¹⁹B. Zheng, A. Lu, L. A. Hardesty, J. H. Sumkin, and C. M. Hakim, "A method to improve visual similarity of breast masses for an interactive computer-aided diagnosis environment," *Med. Phys.* **33**, 111–117 (2006).
- ²⁰A. M. Aisen, L. S. Broderick, H. Winer-Muram, C. E. Brodley, A. C. Kak, C. Pavlopoulou, J. Dy, C. Shyu, and A. Marchiori, "Automated storage and retrieval of thin-section CT images to assist diagnosis: System description and preliminary assessment," *Radiology* **228**, 265–270 (2003).
- ²¹R. M. Nishikawa, Y. Yang, D. Huo, M. Wernick, C. A. Sennett, J. Papaioannou, and L. Wei, "Observer's ability to judge the similarity of clustered calcifications on mammograms," *Medical Imaging 2004: Image Perception, Observer Performance, and Technology Assessment* (SPIE Press, Bellingham, WA, 2004).
- ²²C. Muramatsu, Q. Li, K. S. R. Schmidt, J. Shiraishi, G. M. Newstead, and K. Doi, "Experimental determination of subjective similarity for pairs of clustered microcalcifications on mammograms: Observer study results," *Med. Phys.* **33**, 3460–3468 (2006).
- ²³H. Soltanian-Zadeh, F. Rafiee-Rad, and S. P.-N. D., "Comparison of multi-wavelet, wavelet, haralick, and shape features for microcalcification classification in mammograms," *Pattern Recog.* **37**, 1973–1986 (2004).
- ²⁴H. Jing, Y. Yang, and R. M. Nishikawa, "Retrieval boosted computer-aided diagnosis of clustered microcalcifications for breast cancer," *Med. Phys.* **39**, 676–685 (2012).
- ²⁵J. Wang and Y. Yang, "Spatial density modeling for discriminating between benign and malignant microcalcification lesions," in *Proceedings of the International Symposium on Biomedical Imaging: From Nano to Macro, San Francisco, 2013* (IEEE, Piscataway, NJ, 2013), pp. 133–136.
- ²⁶A. P. Dhawan, Y. Chitre, C. Kaiser-Bonasso, and M. Moskwitz, "Analysis of mammographic microcalcifications using gray-level image structure features," *IEEE Trans. Med. Imaging* **15**, 246–259 (1996).
- ²⁷J. K. Kim and H. W. Park, "Statistical textural features for detection of microcalcifications in digitized mammograms," *IEEE Trans. Med. Imaging* **18**, 231–238 (1999).
- ²⁸S. Nowshad and U. Farooq, "Analysis of mammogram for detection breast cancer using wavelet statistical features," *Int. J. Image Process. Vis. Sci.* **1**, 1–6 (2012).
- ²⁹A. N. Karahahiou, I. S. Boniatis, S. G. Skiadopoulos, F. N. Sakellariopoulos, N. S. Arikidis, E. A. Likaki, G. S. Panayiotakis, and L. I. Costari-dou, "Breast cancer diagnosis: Analyzing texture of tissue surrounding microcalcifications," *IEEE Trans. Inf. Technol. Biomed.* **12**, 731–738 (2008).
- ³⁰P. Lea, T. Næs, and M. Rødbotten, *Analysis of Variance for Sensory Data* (Wiley, New York, NY, 1997).
- ³¹D. C. Montgomery, *Design and Analysis of Experiments* (Wiley, New York, NY, 2012).
- ³²I. Borg and P. J. F. Groenen, *Modern Multidimensional Scaling: Theory and Application* (Springer, New York, NY, 2005).
- ³³L. Egghe and R. Rousseau, "Classical retrieval and overlap measures satisfy the requirements for rankings based on a Lorenz curve," *Inf. Process. Management* **42**, 106–120 (2006).
- ³⁴G. Kowalski, *Information Retrieval Architecture and Algorithms* (Springer, New York, NY, 2011).
- ³⁵F. W. Samuelson and N. Petrick, "Comparing image detection algorithms using resampling," in *Proceedings of the International Symposium on Biomedical Imaging: From Nano to Macro, Arlington, 2006* (IEEE, Piscataway, NJ, 2006), pp. 1312–1315.