Journal of Electrical Engineering & Technology (2019) 14:435–443
https://doi.org/10.1007/s42835-018-00003-x

ORIGINAL ARTICLE

Evaluation of a Content-Based Image Retrieval Computer-Aided Diagnosis System for Breast Ultrasound Images Through Distance Similarity Measures

Min-jeong Kim[2] · Hyun-chong Cho[1]

Received: 8 March 2018 / Revised: 30 June 2018 / Accepted: 17 August 2018 / Published online: 4 January 2019
© The Korean Institute of Electrical Engineers 2019

Abstract

Purpose A content-based image retrieval (CBIR) computer-aided diagnosis (CADx) system using breast masses in ultrasound images has been developed and evaluated to assist radiologists with characterization processes. The purpose of this study is to improve the accuracy of breast cancer diagnoses by analyzing images and providing quantitative information to radiologists through the CADx system.

Methods Two morphological features and six texture features of breast masses were extracted to design how the CADx system retrieves a mass similar to a query mass in a reference library. Based on extracted features from breast masses, the CADx system retrieves masses which are similar to the query mass from the reference library using a k-nearest neighbor (*k*-NN) method. To evaluate the CBIR CADx system, 39 similarity measures (nine similarity families, F_0–F_8) based on the distance similarity were used. A receiver operating characteristic (ROC) analysis was conducted to evaluate the performance of the distance similarity measures.

Conclusions The F_0 family (Mahalanobis distance) measure used with the *k*-NN classifier provided slightly higher performance for the classification of malignant and benign masses as compared to those with the F_1–F_8 family measures.

Keywords Breast cancer · CADx system · Medical image processing · Ultrasound images

1 Introduction

For females, breast cancer is one of the most prevalent causes of death worldwide. Figure 1 shows the estimated number of new cases in 2017 in both Korea and the United States [1, 2], confirming that the incidence breast cancer is higher than average for both countries. Early detection and treatment through accurate screening and diagnosis is the most effective way to reduce the occurrence of breast cancer and associated mortality rate.

Types of medical images have increased due to the development of imaging technology, and diagnostic results can vary depending on the experience of the radiologist. Computer-aided diagnosis (CADx) systems have been studied to provide quantitative information useful for the diagnosis of this disease and to prevent misdiagnoses caused by incorrect interpretations or subjective judgments [3]. A CADx system specialized for breast ultrasound images can assist in the diagnosis of the radiologist by analyzing numerous forms of data and providing quantitative information.

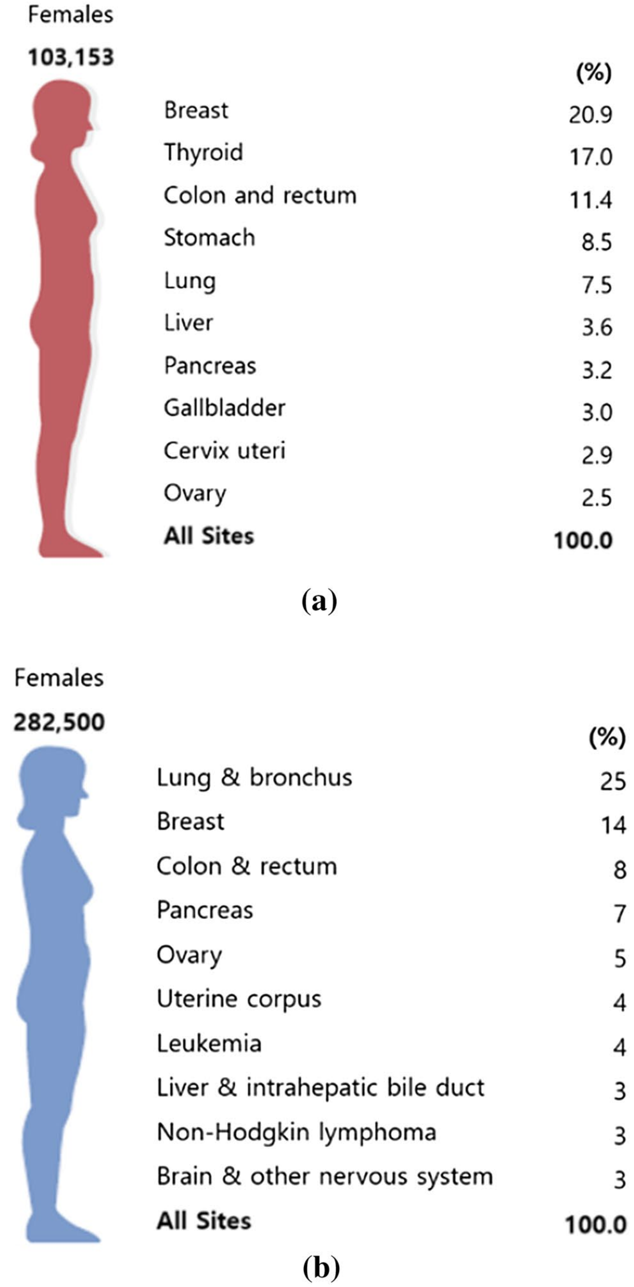
The importance of research using ultrasound imaging is increasing according to research findings which showed that x-rays used for breast cancer diagnoses affect the occurrence of breast cancer [4]. Ultrasound imaging is also a useful diagnostic method with which to distinguish between malignant and benign masses [5]. Generally, the margin of the mass is not clear in ultrasound images of cases of malignant breast cancer. On the other hand, the margin is smooth and clear in benign cases. Figures 2 and 3 correspondingly depict malignant and benign masses as imaged by a breast ultrasound system.

✉ Hyun-chong Cho
hyuncho@kangwon.ac.kr

Min-jeong Kim
mjeong9316@gmail.com

[1] Department of Electronic Engineering and Interdisciplinary Graduate Program for BIT Medical Convergence, Kangwon National University, Chuncheon, South Korea

[2] Interdisciplinary Graduate Program for BIT Medical Convergence, Kangwon National University, Chuncheon, South Korea



Females

103,153

	(%)
Breast	20.9
Thyroid	17.0
Colon and rectum	11.4
Stomach	8.5
Lung	7.5
Liver	3.6
Pancreas	3.2
Gallbladder	3.0
Cervix uteri	2.9
Ovary	2.5
All Sites	**100.0**

(a)

Females

282,500

	(%)
Lung & bronchus	25
Breast	14
Colon & rectum	8
Pancreas	7
Ovary	5
Uterine corpus	4
Leukemia	4
Liver & intrahepatic bile duct	3
Non-Hodgkin lymphoma	3
Brain & other nervous system	3
All Sites	**100.0**

(b)

Fig. 1 Most common types of estimated new cancer cases, 2017: **a** in Korea, **b** in the United States

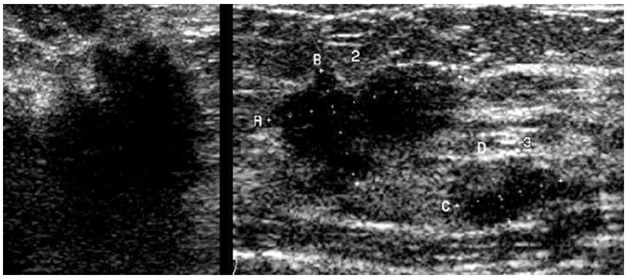


Fig. 2 Malignant masses in breast ultrasound images

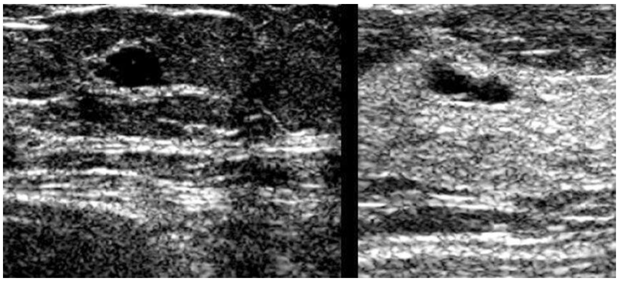


Fig. 3 Benign masses in breast ultrasound images

Research on a CADx system for breast ultrasound images is underway to assist radiologists by providing quantitative information about breast masses. After morphological and texture features are extracted from breast masses in ultrasound images, distance similarity measures are used to retrieve masses similar to a query mass in a reference library. Similarity distance measures are also very useful for solving many pattern-recognition issues, such as classification, clustering, and retrieval problems. Various similarity distance measures that are applicable to characterize malignant and benign masses are reviewed and categorized in both syntactic and semantic relationships. The purpose of this study is to design a CADx system for breast ultrasound images that improves the accuracy of breast cancer diagnoses using various similarity distance measures.

2 Methods

2.1 Data Set

In this study, we used records of patients who underwent breast imaging in the Department of Radiology at the University of Michigan to design the CADx system for breast ultrasound images. The use of the breast ultrasound data was approved by the Institutional Review Board (IRB), and all data were pathologically verified by biopsies [6].

In total, data from 250 patients were used, with 96 malignant and 154 benign masses included [6]. In this case, the radiologist selected two or more orthogonal ultrasound images that best represent each mass in the collected data. However, only a single ultrasound image was selected for certain masses not visible in orthogonal ultrasound images. The collected data were randomly classified into two sets. Set 1 (S_1) includes a total of 230 ultrasound images with 41 malignant and 80 benign masses and Set 2 (S_2) includes a total of 258 ultrasound images with 55 malignant and 74 benign masses.

2.2 Feature Extraction and Selection

To design the CADx system for breast ultrasound images, we segmented the breast masses in ultrasound images with an automated method using a previously designed active contour model [7]. The contour of the mass is automatically estimated from its center using this model. Morphological and texture features of the segmented masses were extracted, after which the features were used for characterization of a query mass for the CADx system.

We extracted two morphological features and six texture features based on automated segmentation in the design of the CADx system [6]. The morphological features represent the shape of the mass, such as the size or appearance of the mass, and the texture features represent the texture shown in the ultrasound image. In breast ultrasound images, a taller-than-wide shape is a good indicator of malignancy. Thus, the width-to-height ratio of the mass was extracted as a morphological feature. Another useful feature to differentiate between malignant and benign masses is the posterior shadowing feature, which is defined as the normalized average gray-level difference between the interior of the segmented mass and the darkest posterior strip [6]. The texture features extracted from spatial gray-level dependence matrices or co-occurrence matrices are information measures for correlations 1 and 2 and for the difference entropy, entropy, energy, and sum entropy [8]. Here, entropy is an indicator of uncertainty. A larger value indicates that the characteristics of the data are uncertain and a smaller value means that the characteristics of the data are biased toward one side. Conversely, energy is an indicator of uniformity.

To design the CADx system, the feature vectors of the selected masses were classified into training sets and test sets through a cross-validation method.

2.3 Distance Similarity Measures

To retrieve masses similar to the query mass, we extract feature vectors that are identical to the query mass in the reference library. After determining the measure similarity outcome between the feature vector of the query mass and the feature vector of the reference library, we retrieve masses similar to the query mass from the reference library. The performance of the CADx system was evaluated by a ROC (receiver operating characteristic) analysis.

In this study, 39 distance similarity measures were applied to retrieve masses similar to the query masses, and the performance of each similarity measure was evaluated. The 39 similarity measures were classified into nine categories according to the similarity of the notation. The nine families are Mahalanobis (F_0, 1 measure), the Minkowski family (F_1, three measures), the L_1 family (F_2, six measures), the Intersection family (F_3, seven measures), the Inner Product family (F_4, four measures), the Fidelity family or the Squared-chord family (F_5, three measures), the Squared L_2 family or the χ^2 family (F_6, eight measures), Shannon's entropy family (F_7, four measures), and Combinations family (F_8, three measures).

2.3.1 Mahalanobis Distance (F_0)

Mahalanobis distance refers to a method used to measure the degree of similarity considering the probability distribution of the data. It corresponds to the normalized Euclidean distance based on a covariance matrix [9]. Because it takes into account correlations between data instances, the performance of the Mahalanobis distance approach is usually better than that of the Euclidean distance [10]. The Mahalanobis distance can be calculated as follows,

$$D_{Mah} = \sqrt{\left(Q - P\left(r_j\right)\right) \sum^{-1} \left(Q - P\left(r_j\right)\right)^T}, \tag{1}$$

where Q is the feature vector of the query mass, $P(r_j)$ is the feature vector of the jth reference mass, and $\sum$ is the covariance matrix.

2.3.2 Minkowski Family (F_1)

The Minkowski family refers to a general metric distance [9], and it is a generalized measure of the Euclidean distance and the 'City block' distance. The Minkowski distance can be defined as shown below.

$$D_{Mk} = \sqrt[L]{\sum_{i=1}^{d} \left|q_i - p_i\left(r_j\right)\right|^L} \tag{2}$$

In this equation, q_i is the ith feature vector of the query mass, $p_i\left(r_j\right)$ is the ith feature vector of r_j, r_j is the j th reference mass, and d represents the dimensions of the feature space.

There are the Euclidean distance, City block distance, and Chebyshev distance in the Minkowski family. The Euclidean distance is the method most commonly used to find the distances between feature vectors in a multidimensional space [11]. It is a special case of the Minkowski distance where $L = 2$, and it can be expressed as Eq. (3). The City block distance represents the sum of the absolute differences between feature vectors [9]. It is $L = 1$ in Eq. (2) and is calculated as shown in Eq. (4). The Chebyshev distance measures distances while assuming only the most relevant dimensions [9]. In the case of $L = \infty$ in Eq. (2), the Chebyshev distance can be expressed as Eq. (5).

$$D_{Eu} = \sqrt{\sum_{i=1}^{d} \left(q_i - p_i\left(r_j\right)\right)^2} \tag{3}$$

$$D_{City} = \sum_{i=1}^{d} \left|q_i - p_i\left(r_j\right)\right| \tag{4}$$

$$D_{Cheb} = \max_i \left|q_i - p_i\left(r_j\right)\right| \tag{5}$$

2.3.3 L_1 Family (F_2)

The L_1 family uses absolute differences in an extended method of the City block distance. It is defined as the sum of the absolute differences between two feature vectors. We apply six distance similarity measures—Sorensen, Gower, Soergel, Kulczynski, Canberra, and Lorentzian—in this family. Each measure can be calculated from Eqs. (6)–(11).

$$D_{Sen} = \frac{\sum_{i=1}^{d} \left|q_i - p_i\left(r_j\right)\right|}{\sum_{i=1}^{d} \left(q_i + p_i\left(r_j\right)\right)} \tag{6}$$

$$D_{Gow} = \frac{1}{d} \sum_{i=1}^{d} \left|q_i - p_i\left(r_j\right)\right| \tag{7}$$

$$D_{Sgel} = \frac{\sum_{i=1}^{d} \left|q_i - p_i\left(r_j\right)\right|}{\sum_{i=1}^{d} \max_i \left(q_i, p_i\left(r_j\right)\right)} \tag{8}$$

$$D_{Kld} = \frac{\sum_{i=1}^{d} \left|q_i - p_i\left(r_j\right)\right|}{\sum_{i=1}^{d} \min_i \left(q_i, p_i\left(r_j\right)\right)} \tag{9}$$

$$D_{Can} = \sum_{i=1}^{d} \frac{\left|q_i - p_i\left(r_j\right)\right|}{q_i + p_i\left(r_j\right)} \tag{10}$$

$$D_{Lor} = \sum_{i=1}^{d} \ln\left(1 + \left|q_i - p_i\left(r_j\right)\right|\right) \tag{11}$$

The Gower distance in Eq. (7) scales the vector space into the normalized space and then uses the absolute difference [12]. The numerator signifies the difference and the denominator normalizes the difference in Canberra and Sorensen [9]. Canberra is obtained by dividing the absolute difference between the feature vectors by the sum of the feature vectors. The Lorentzian can be expressed as Eq. (11), which indicates the absolute difference between the feature vectors with a natural log function. At this time, 1 is added to ensure non-negative attributes and to avoid zero logs [12].

2.3.4 Intersection Family (F_3)

The intersection between feature vectors is a widely used form of similarity [12]. Intersection similarity measures can be transformed into L_1-based distance measures using this technique, i.e., $d(q,p) = 1 - s(q,p)$ or $d(q,p) = {}^{1}/_{s(q,p)}$. This family includes the Intersection, Wave Hedges, Czekanowski, Motyka Kulczynski, Ruzicka, and Tanimoto measures, which correspondingly are expressed as Eqs. (12)–(18).

$$\begin{aligned} D_{Its} &= 1 - s_{Its} = 1 - \sum_{i=1}^{d} \min_i \left(q_i, p_i\left(r_j\right)\right) \\ &= \frac{1}{2} \sum_{i=1}^{d} \left|q_i - p_i\left(r_j\right)\right| \end{aligned} \tag{12}$$

$$D_{WH} = \sum_{i=1}^{d} \left(1 - \frac{\min_i \left(q_i, p_i\left(r_j\right)\right)}{\max_i \left(q_i, p_i\left(r_j\right)\right)}\right) \tag{13}$$

$$s_{Cze} = \frac{2 \sum_{i=1}^{d} \min_i \left(q_i, p_i\left(r_j\right)\right)}{\sum_{i=1}^{d} \left(q_i + p_i\left(r_j\right)\right)} \tag{14}$$

$$D_{Mot} = 1 - s_{Mot} \frac{\sum_{i=1}^{d} \max_i \left(q_i, p_i\left(r_j\right)\right)}{\sum_{i=1}^{d} \left(q_i + p_i\left(r_j\right)\right)} \tag{15}$$

$$s_{Kls} = \frac{1}{D_{kld}} = \frac{\sum_{i=1}^{d} \min_i \left(q_i, p_i\left(r_j\right)\right)}{\sum_{i=1}^{d} \left|q_i - p_i\left(r_j\right)\right|} \tag{16}$$

$$s_{Ruz} = \frac{\sum_{i=1}^{d} \min_i \left(q_i, p_i\left(r_j\right)\right)}{\sum_{i=1}^{d} \max_i \left(q_i, p_i\left(r_j\right)\right)} \tag{17}$$

$$D_{Ta} = \frac{\sum_{i=1}^{d} q_i + \sum_{i=1}^{d} p_i\left(r_j\right) - 2 \sum_{i=1}^{d} \min_i \left(q_i, p_i\left(r_j\right)\right)}{\sum_{i=1}^{d} q_i + \sum_{i=1}^{d} p_i\left(r_j\right) - \sum_{i=1}^{d} \min_i \left(q_i, p_i\left(r_j\right)\right)} \tag{18}$$

2.3.5 Inner Product Family (F_4)

The Inner Product family is a method of using the *inner product* between feature vectors. There are similarity measures in this family that explicitly include the *inner product form* '$Q \cdot P$' in the definition [12]. We apply four measures—Cosine, Kumar-Hassebrook (PCE), Jaccard, and Dice—for the Inner Product family, as expressed by Eqs. (19)–(22) below.

$$s_{Cos} = \frac{\sum_{i=1}^{d} q_i p_i\left(r_j\right)}{\sqrt{\sum_{i=1}^{d} q_i^2}\sqrt{\sum_{i=1}^{d} p_i\left(r_j\right)^2}} \tag{19}$$

$$s_{PCE} = \frac{\sum_{i=1}^{d} q_i p_i\left(r_j\right)}{\sum_{i=1}^{d} q_i^2 + \sum_{i=1}^{d} p_i\left(r_j\right)^2 - \sum_{i=1}^{d} q_i p_i\left(r_j\right)} \tag{20}$$

$$D_{Jac} = \frac{\sum_{i=1}^{d} \left(q_i - p_i\left(r_j\right)\right)^2}{\sum_{i=1}^{d} q_i^2 + \sum_{i=1}^{d} p_i\left(r_j\right)^2 - \sum_{i=1}^{d} q_i p_i\left(r_j\right)} \tag{21}$$

$$s_{Dice} = \frac{2\sum_{i=1}^{d} q_i p_i\left(r_j\right)}{\sum_{i=1}^{d} q_i^2 + \sum_{i=1}^{d} p_i\left(r_j\right)^2} \tag{22}$$

2.3.6 Fidelity Family or Squared-Chord Family (F_5)

The Fidelity similarity is the sum of *geometric means*, and it is defined Eq. (23) [12]. It includes the Hellinger and Matusita. The squared-chord distance is referred to as Matusita without the square root. There are alternative representations using squared-chord distance for all Fidelity based measures. Equations (24), (25), and (26) denote Hellinger, Matusita, and Squared-chord, respectively.

$$s_{Fid} = \sum_{i=1}^{d} \sqrt{q_i p_i\left(r_j\right)} \tag{23}$$

$$D_{Hel} = 2\sqrt{1 - \sum_{i=1}^{d} \sqrt{q_i p_i\left(r_j\right)}} \tag{24}$$

$$D_{Mat} = \sqrt{\sum_{i=1}^{d} \left(\sqrt{q_i} - \sqrt{p_i\left(r_j\right)}\right)^2} \tag{25}$$

$$D_{Scho} = \sum_{i=1}^{d} \left(\sqrt{q_i} - \sqrt{p_i\left(r_j\right)}\right)^2 \tag{26}$$

2.3.7 Squared L_2 Family or χ^2 Family (F_6)

There are eight similarity measures using the squared Euclidean distance in this family, and the squared Euclidean distance is defined as Eq. (27) [12]. The eight similarity measures applied in this family are Squared Euclidean, Pearson χ^2, Neyman χ^2, Squared χ^2, Probabilistic Symmetric χ^2, Divergence, Clark, and Additive Symmetric χ^2. These can be calculated by Eqs. (27)–(34), respectively.

$$D_{SEu} = \sum_{i=1}^{d} \left(q_i - p_i\left(r_j\right)\right)^2 \tag{27}$$

$$D_{Pea} = \sum_{i=1}^{d} \frac{\left(q_i - p_i\left(r_j\right)\right)^2}{p_i\left(r_j\right)} \tag{28}$$

$$D_{Ney} = \sum_{i=1}^{d} \frac{\left(q_i - p_i\left(r_j\right)\right)^2}{q_i} \tag{29}$$

$$D_{Squ} = \sum_{i=1}^{d} \frac{\left(q_i - p_i\left(r_j\right)\right)^2}{q_i + p_i\left(r_j\right)} \tag{30}$$

$$D_{PSy} = 2\sum_{i=1}^{d} \frac{\left(q_i - p_i\left(r_j\right)\right)^2}{q_i + p_i\left(r_j\right)} \tag{31}$$

$$D_{Div} = 2\sum_{i=1}^{d} \frac{\left(q_i - p_i\left(r_j\right)\right)^2}{\left(q_i + p_i\left(r_j\right)\right)^2} \tag{32}$$

$$D_{Clk} = \sqrt{\sum_{i=1}^{d} \left(\frac{\left|q_i - p_i\left(r_j\right)\right|}{q_i + p_i\left(r_j\right)}\right)^2} \tag{33}$$

$$D_{Add} = \sum_{i=1}^{d} \frac{\left(q_i - p_i\left(r_j\right)\right)^2\left(q_i + p_i\left(r_j\right)\right)}{q_i p_i\left(r_j\right)} \tag{34}$$

2.3.8 Shannon's Entropy Family (F_7)

In this family, we apply four similarity measures that are probabilistic uncertainty or entropy concepts [12]. The four similarity measures are the Jeffreys, Topsoe, Jensen-Shannon, and Jensen difference measures, as correspondingly expressed in terms of the entropy as Eqs. (35)–(38).

$$D_{Jef} = \sum_{i=1}^{d} \left(q_i - p_i\left(r_j\right)\right) \ln \frac{q_i}{p_i\left(r_i\right)} \tag{35}$$

$$\begin{aligned} D_{Tsoe} = &\sum_{i=1}^{d} q_i \ln \frac{2q_i}{q_i + p_i\left(r_j\right)} \\ &+ \sum_{i=1}^{d} p_i\left(r_j\right) \ln \frac{2p_i\left(r_j\right)}{q_i + p_i\left(r_j\right)} \end{aligned} \tag{36}$$

$$\begin{aligned} D_{JSh} = \frac{1}{2} \Bigg[&\sum_{i=1}^{d} q_i \ln \frac{2q_i}{q_i + p_i\left(r_j\right)} \\ &+ \sum_{i=1}^{d} p_i\left(r_j\right) \ln \frac{2p_i\left(r_j\right)}{q_i + p_i\left(r_j\right)} \Bigg] \end{aligned} \tag{37}$$

$$\begin{aligned} D_{Jdiff} = &\sum_{i=1}^{d} \left(\frac{q_i \ln q_i + p_i\left(r_j\right) \ln p_i\left(r_j\right)}{2} \right) \\ &- \sum_{i=1}^{d} \left(\frac{q_i + p_i\left(r_j\right)}{2} \ln \frac{q_i + p_i\left(r_j\right)}{2} \right) \end{aligned} \tag{38}$$

2.3.9 Combinations (F_8)

There are three similarity measures in this family, referring to methods which utilize multiple measures [12]. First, Taneja utilizes *arithmetic* and *geometric means*, which can be expressed as Eq. (39). Second, Kumar-Johnson, which utilizes the arithmetic and geometric mean divergence, can be expressed as Eq. (40). Finally, we applied the average of City block distance and Chebyshev distance in the Minkowski family (F_1). This can be calculated by Eq. (41).

$$D_{Taj} = \sum_{i=1}^{d} \frac{q_i + p_i\left(r_j\right)}{2} \ln \left(\frac{q_i + p_i\left(r_j\right)}{2\sqrt{q_i p_i\left(r_j\right)}} \right) \tag{39}$$

$$D_{KuJ} = \sum_{i=1}^{d} \left(\frac{\left(q_i^2 - p_i\left(r_j\right)^2\right)^2}{2\left(q_i p_i\left(r_j\right)\right)^{3/2}} \right) \tag{40}$$

$$D_{Avg} = \frac{\sum_{i=1}^{d} \left| q_i - p_i\left(r_j\right) \right| + \max\limits_{i} \left| q_i - p_i\left(r_j\right) \right|}{2} \tag{41}$$

In this study, we applied 39 distance similarity measures to retrieve masses similar to query masses in the CADx system. According to the notation similarity of each measure, the 39 similarity measures were classified into nine categories. The F_0 family measures the similarity considering the probability distribution of the data, the F_1 family is a general metric distance, and the F_2 family uses absolute differences. The similarity measures in the F_3 family are a widely used form in which the intersection between feature vectors is used. The F_4 family uses the inner product between the feature vectors and the F_5 family is a method that uses the sum of the geometric means. Finally, there is the F_6 family, which uses the squared Euclidean distance, the F_7 family which uses the probabilistic uncertainty or entropy, and the F_8 family which uses multiple measures. Each has been described in depth [12].

3 Results and Discussion

To assist radiologists with breast cancer diagnoses, a CADx system was studied for breast ultrasound images. We applied 39 distance similarity measures based on a distance metric. The performance capabilities of the CADx system when applying each similarity measure were then analyzed and evaluated.

To design the CADx system for breast ultrasound images, breast masses in ultrasound images were classified into a training set and a test set through a cross-validation method. The 39 similarity measures were categorized into nine families based on the notation similarity in each case.

When S_1 is used as the training data and S_2 is used as the test data, the performance of the CADx system applying each similarity measure is presented in Table 1 and Fig. 4. Table 2 and Fig. 5 show the performance when S_2 is used as the training data and S_1 is used as the test data, where the value of k is the number of retrieved masses, referring to the number of breast ultrasound images shown in references to assist radiologists with their diagnoses. Tables 1 and 2 present numerical representations of the performance outcomes for the 39 distance similarity measures, and Figs. 4 and 5 are graphs of the performances for the nine families.

When using S_1 as the training data and S_2 as the test data, the performances of all similarity measures are similar. However, the performance of the F_0 family (the Mahalanobis distance) is slightly better than those of the others when the number of retrieved masses is increased (i.e., $k=25$–50). In addition, the average performance (i.e., $k=1$–50) of the F_0 family is better than those of the other similarity families (Table 1). The performance of each family is shown in Fig. 4, which indicates that the F_0 family is superior to the other families in large number of top retrieval masses

Table 1 The performances of the 39 similarity measures (training set S_1, Test set S_2)

		Avg (1–50)	Avg (25–50)
Family 0			
1	Mahalanobis	0.8767	0.8912
		0.8767	**0.8912**
Family 1			
2	Euclidean	0.8724	0.8882
3	City block	0.8733	0.8820
4	Chebyshev	0.8732	0.8862
		0.8730	**0.8855**
Family 2			
5	Sorensen	0.8748	0.8844
6	Gower	0.8733	0.8820
7	Soergel	0.8748	0.8844
8	Kulczynski d	0.8748	0.8844
9	Canberra	0.8727	0.8831
10	Lorentzian	0.8718	0.8789
		0.8737	**0.8829**
Family 3			
11	Intersection	0.8451	0.8670
12	Wave Hedges	0.8706	0.8790
13	Czekanowski	0.8748	0.8844
14	Motyka	0.8748	0.8844
15	Kulczynski s	0.8748	0.8844
16	Ruzicka	0.8748	0.8844
17	Tanimoto	0.8794	0.8901
		0.8706	**0.8819**
Family 4			
18	Cosine	0.8739	0.8872
19	Kumar–Hassebrook (PCE)	0.8737	0.8897
20	Jaccard	0.8737	0.8897
21	Dice	0.8737	0.8897
		0.8738	**0.8891**
Family 5			
22	Hellinger	0.8727	0.8882
23	Matusita	0.8727	0.8882
24	Squared-chord	0.8727	0.8882
		0.8727	**0.8882**
Family 6			
25	Squared Euclidean	0.8724	0.8882
26	Pearson	0.8736	0.8873
27	Neyman	0.8714	0.8877
28	Squa red	0.8729	0.8885
29	Probabil istic symmetric	0.8729	0.8885
30	Di vergence	0.8712	0.8835
31	Clark	0.8712	0.8835
32	Additive symmetric	0.8724	0.8883
		0.8723	**0.8869**
Family 7			
33	Jeffreys	0.8727	0.8882
34	Topsoe	0.8728	0.8884
35	Jensen–Shannon	0.8728	0.8884
36	Jensen d ifference	0.8728	0.8884
		0.8727	**0.8883**
Family 8			
37	Taneja	0.8726	0.8880
38	Kumar–Johnson	0.8726	0.8884
39	Avg (L1, L_inf)	0.8725	0.8868
		0.8726	**0.8877**

Average values are shown in bold

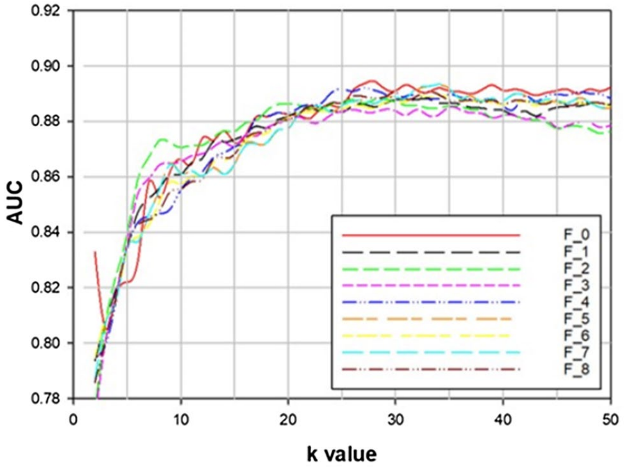


Fig. 4 The performances of the nine similarity families (Training set S_1, Test set S_2)

(k > 25). When using S_2 as the training data and S_1 as the test data, the results are similar to the previous results, as shown in Table 2 and Fig. 5.

In this study, we compared the performances of the CADx system when applying the 39 similarity measures. It was found that the performance of the F_0 family exceeded those of the other similarity families in both experiments for large number of top retrieval masses. The F_0 family computes the covariance distance taking into account the distribution of the given data. However, the other similarity families only consider the distance between two feature vectors. Therefore, the F_0 family (Mahalanobis distance) that calculates the covariance distance outperforms the other similarity measures on average.

Table 2 The performances of the 39 similarity measures (training set S_2, Test set S_1)

		Avg (1–50)	Avg (25–50)
Family 0			
1	Mahalanobis	0.8938	0.9092
		0.8939	**0.9092**
Family 1			
2	Euclidean	0.8893	0.9028
3	City block	0.8877	0.9047
4	Chebyshev	0.8835	0.8913
		0.8868	**0.8996**
Family 2			
5	Sorensen	0.8892	0.9034
6	Gower	0.8877	0.9047
7	Soergel	0.8892	0.9034
8	Kulczynski d	0.8892	0.9034
9	Canberra	0.8827	0.9016
10	Lorentzian	0.8834	0.9012
		0.8869	**0.9029**
Family 3			
11	Intersection	0.8637	0.8797
12	Wave Hedges	0.8851	0.9017
13	Czekanowski	0.8892	0.9034
14	Motyka	0.8892	0.9034
15	Kulczynski s	0.8892	0.9034
16	Ruzicka	0.8892	0.9034
17	Tanimoto	0.8701	0.8883
		0.8823	**0.8976**
Family 4			
18	Cosine	0.8970	0.9087
19	Kumar–Hassebrook (PCE)	0.8920	0.9035
20	Jaccard	0.8920	0.9035
21	Dice	0.8920	0.9035
		0.8932	**0.9048**
Family 5			
22	Hellinger	0.8895	0.9031
23	Matusita	0.8895	0.9031
24	Squared-chord	0.8895	0.9031
		0.8895	**0.9031**
Family 6			
25	Squared Euclidean	0.8893	0.9028
26	Pearson	0.8818	0.8977
27	Neyman	0.8788	0.8919
28	Squared	0.8845	0.9004
29	Probabilistic symmetric	0.8845	0.9004
30	Divergence	0.8880	0.9023
31	Clark	0.8880	0.9023
32	Additive symmetric	0.8793	0.8934
		0.884	**0.8989**
Family 7			
33	Jeffreys	0.8791	0.8932
34	Topsoe	0.8796	0.8942
35	Jensen–Shannon	0.8875	0.9015
36	Jensen difference	0.8875	0.9015
		0.8834	**0.8976**
Family 8			
37	Taneja	0.8874	0.9013
38	Kumar–Johnson	0.8792	0.8934
39	Avg (L1, L_inf)	0.8905	0.9054
		0.8857	**0.9000**

Average values are shown in bold

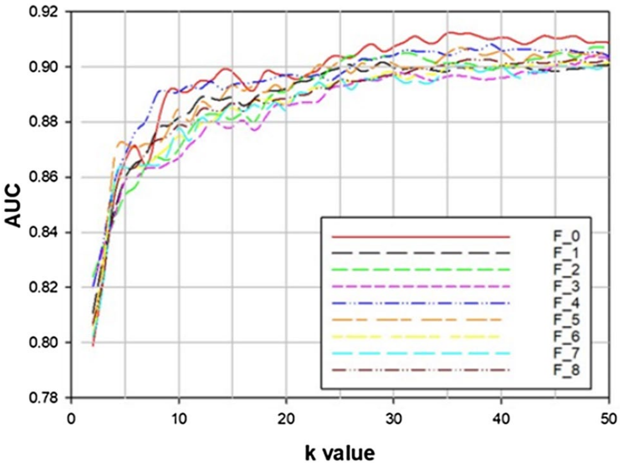


Fig. 5 The performances of the nine similarity families (Training set S_2, Test set S_1)

4 Conclusion

In this paper, a CADx system for breast ultrasound images was devised to assist radiologists in differentiating benign and malignant masses on ultrasound breast images. To design the CADx system, morphological and texture features were extracted from a database. The feature vectors of breast masses were then classified into a training set and a test set through a cross-validation method. Using a *k*-nearest neighbor (*k*-NN) method, we applied 39 distance similarity measures (nine similarity families, F_0–F_8) based on distance and compared the performance of the CADx system through an ROC analysis. The 39 distance similarity measures were classified in nine similarity families based on the notation similarity of each measure, as noted above. The 39 distance similarity measures were applied to retrieve masses similar to a query mass in a reference library.

It was found that the performances of each of the similarity measures did not show any significant differences. However, the performance of the F_0 family was greater than those of the other families when the number of retrieved masses is increased ($k \geq 25$). When the number of retrieved masses is low, the probability distribution cannot be confirmed because there are too few data instances to be considered. When the number of retrieved masses is increased, the probability distribution of the data can be fully taken into account. Therefore, the performance of the CADx system using the Mahalanobis distance (the F_0 family), which considers the probability distribution of the data, is superior to those of other families.

For larger number ($k > 10$) of top retrieval masses, the classification performance of all similarity measures continuously leveled off. 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 warrants further investigations.

Future work includes applying the CBIR CADx system to a larger and independent dataset, expanding the feature space, and combining the developed ultrasound image characterization method with mammographic characterization method. The effects of the different CBIR CADx systems on the characterization of breast masses by support vector machine (SVM) will also be evaluated.

Acknowledgements This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (no. 2017R1E1A1A03070297). This research was supported by the Ministry of Science and ICT (MSIT), Korea, under the Information Technology Research Center (ITRC) support program (IITP-2018-0-01433) supervised by the Institute for Information and Communications Technology Promotion (IITP).

References

1. Jung K-W, Won Y-J, Oh C-M, Kong H-J, Lee DH, Lee KH (2017) Prediction of cancer incidence and mortality in Korea, 2017. Cancer Res Treat 49:306–312
2. Siegel RL, Miller KD, Jemal A (2017) Cancer statistics, 2017. CA Cancer J Clin 67:7–30
3. Cho H, Hadjiiski L, Sahiner B, Chan HP, Paramagul C, Helvie M et al (2012) Interactive content-based image retrieval (CBIR) computer-aided diagnosis (CADx) system for ultrasound breast masses using relevance feedback. In: SPIE, medical imaging 2012, p 831509
4. Espín-López P, Martellosio A, Pasian M, Bozzi M, Perregrini L, Mazzanti A et al (2017) Breast cancer imaging at mm-waves: feasibility study on the safety exposure limits. In: Microwave conference (EuMC), 2016 46th European, pp 667–670
5. Hong AS, Rosen EL, Soo MS, Baker JA (2005) BI-RADS for sonography: positive and negative predictive values of sonographic features. Am J Roentgenol 184:1260–1265
6. Cho H, Hadjiiski L, Sahiner B, Chan HP, Helvie M, Paramagul C et al (2011) Similarity evaluation in a content-based image retrieval (CBIR) CADx system for characterization of breast masses on ultrasound images. Med Phys 38:1820–1831
7. Cui J, Sahiner B, Chan HP, Nees A, Paramagul C, Hadjiiski LM et al (2009) A new automated method for the segmentation and characterization of breast masses on ultrasound images. Med Phys 36:1553–1565
8. Haralick RM, Shanmugam K, Dinstein I (1973) Texture features for image classification. IEEE Trans Syst Man Cybern SMC-3:610–621
9. Belattar K, Mostefai S (2015) Similarity measures for content-based dermoscopic image retrieval: a comparative study. In: 2015 First international conference on new technologies of information and communication (NTIC), pp 1–6
10. Bo D, Zhangguan L, Cuixiao L (2015) An algorithm of image matching based on mahalanobis distance and weighted KNN graph. In: 2015 2nd international conference on information science and control engineering, Shanghai, 2015, pp 116–121
11. Bouhmala N (2016) How good is the euclidean distance metric for the clustering problem. In: 2016 5th IIAI international congress on advanced applied informatics (IIAI-AAI), pp 312–315
12. Cha SH (2007) Comprehensive survey on distance/similarity measures between probability density functions. Int J Math Model Meth Appl Sci 1(4):300–307

Min-jeong Kim She received the M.S. degree in Electrical and Electronic Engineering from Kangwon National University, South Korea in 2018. She is currently a researcher at National Cancer Center, South Korea.

Hyun-chong Cho He received the M.S. and Ph.D. degrees in Electrical and Computer Engineering from the University of Florida, USA in 2009. During 2010–2011, he was a Research Fellow at the University of Michigan at Ann Arbor, USA. From 2012 to 2013, he was a Chief Research Engineer in LG Electronics, South Korea. He is currently an Assistant Professor at Kangwon National University, South Korea.