Content-Based Image Retrieval System for Pulmonary Nodules Using Optimal Feature Sets and Class Membership-Based Retrieval

Shrikant A. Mehre¹ · Ashis Kumar Dhara^{1,2} · Mandeep Garg³ · Naveen Kalra³ · Niranjn Khandelwal³ · Sudipta Mukhopadhyay¹

Published online: 25 October 2018
© Society for Imaging Informatics in Medicine 2018

Abstract

Lung cancer manifests itself in the form of lung nodules, the diagnosis of which is essential to plan the treatment. Automated retrieval of nodule cases will assist the budding radiologists in self-learning and differential diagnosis. This paper presents a content-based image retrieval (CBIR) system for lung nodules using optimal feature sets and learning to enhance the performance of retrieval. The classifiers with more features suffer from the curse of dimensionality. Like classification schemes, we found that the optimal feature set selected using the minimal-redundancy-maximal-relevance (mRMR) feature selection technique improves the precision performance of simple distance-based retrieval (SDR). The performance of the classifier is always superior to SDR, which leans researchers towards conventional classifier-based retrieval (CCBR). While CCBR improves the average precision and provides 100% precision for correct classification, it fails for misclassification leading to zero retrieval precision. The class membership-based retrieval (CMR) is found to bridge this gap for texture-based retrieval. Here, CMR is proposed for nodule retrieval using shape-, margin-, and texture-based features. It is found again that optimal feature set is important for the classifier used in CMR as well as for the feature set used for retrieval, which may lead to different feature sets. The proposed system is evaluated using two independent databases from two continents: a public database LIDC/IDRI and a private database PGIMER-IITKGP, using three distance metrics, i.e., Canberra, City block, and Euclidean. The proposed CMR-based retrieval system with optimal feature sets performs better than CCBR and SDR with optimal features in terms of average precision. Apart from average precision and standard deviation of precision, the fraction of queries with zero precision retrieval is also measured.

Keywords Content-based image retrieval · Feature selection · CT images · Lung nodules · Lung cancer · Diagnosis of lung cancer · Self-learning tool of radiology

Introduction

Worldwide, lung cancer is responsible for the most cancer-related deaths in men, and the second most in women [31]. About 90% of lung cancers arise as a result of tobacco

use and are strongly correlated with cigarette smoking [21]. Only 15% of lung cancers are diagnosed at a localized stage, for which the 5-year survival rate is 54% [31]. Computed tomography (CT) is the best modality for detection of lung cancer. Lung cancer appears in the form of nodules, which are blob-like structures with a diameter of 3 to 30 mm. Screening of lung CT for nodules could increase the 5-year survival rate from 15 to 80% [11]. The diagnosis of lung nodules is essential for a further plan of treatment. The radiologists diagnose a lung nodule as a benign or malignant based on the visual information using CT images and the clinical information of the patient. Nowadays, hundreds of lung CT scans are generated in the hospitals, and these scans are stored using picture archiving and communication system (PACS). A content-based image retrieval (CBIR)

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s10278-018-0136-1>) contains supplementary material, which is available to authorized users.

✉ Sudipta Mukhopadhyay
smukho@gmail.com

Extended author information available on the last page of the article.

system for lung nodules will be helpful in retrieving the similar lung nodules with a diagnosis from experienced radiologists for a given query lung nodule. These retrieved lung nodule cases with the diagnosis will be helpful for radiologists to make a decision for the query lung nodule. Further, the budding radiologists could use the CBIR system for self-learning in the absence of experienced radiologists.

For making the CBIR system useful for the radiologists, it should need minimal user intervention. Initially, CBIR systems of lung nodules have used a representative slice of the nodule [16], [29] instead of the whole nodule for retrieval. These systems were dependent on the radiologists for the delineation of lung nodule boundary. A simple distance-based retrieval system [9] used an automated algorithm for nodule segmentation [7], which needs a seed point from the radiologist. After the nodule segmentation, similar nodule cases retrieved using several shape-, texture-, and margin-based features [9]. The use of the optimal feature set for retrieval is not properly explored. In this article, the efficacy of feature selection for retrieval is investigated using three different metrics.

Using only SDR, it is difficult to efficiently discriminate between types of lung nodules. The conventional classifier-based retrieval (CCBR) performs better than the SDR. However, retrieval performance is dependent on the accuracy of the classifier, since it retrieves the images from the winning class. In case of correct classification, CCBR achieves 100% precision, but for misclassification, retrieval precision is zero. This creates a huge deviation in retrieval performance. Moreover, use of many features for classification may lead to a reduction in performance. Hence, an optimal feature set selection will be crucial in improving the performance. A learning-based retrieval system, i.e., class membership-based retrieval (CMR) [23], [5], helps to overcome the limitation of CCBR by using a fuzzy class membership of the query nodule based on the confidence of the classifier. The CMR uses a weighted distance for high confidence of classifier and uses a simple distance in case of lack of high confidence. The search space is not limited to a single winning class. Thus, the CMR creates a chance to retrieve the cases in cases of misclassification also. The CMR scheme for texture images used the same set of features for classification and retrieval [23]. The proposed system aims to investigate the optimal feature set for classification and retrieval. Further, the analysis of retrieval techniques using different distance metrics is very important. The CMR approach has shown an improved performance for texture data [23]. In the proposed approach, we have analyzed the efficacy of the CMR approach for non-textural features, i.e., shape, margin, along with texture features.

The prior works of CBIR systems for medical images is discussed in “[Prior Works](#).” The proposed CBIR system

with feature selection approach is described in “[CBIR System for Pulmonary Nodules Using Optimal Feature Set and Class Membership-Based Retrieval](#).” The performance metrics and databases used to evaluate the proposed CBIR technique is explained in “[Performance Metrics](#)” and “[Database](#),” respectively. The performance of the proposed technique is compared with the various retrieval strategies, which are presented in “[Results and Discussions](#).”

Prior Works

Müller et al. [25] reported the use of CBIR system in clinical decision-making, medical education, and research. The benefits of the CBIR system motivated the researchers to develop a CBIR system for various kinds of medical images. Lehmann et al. [17] developed a mono-hierarchical multi-axial classification code for fast retrieval of medical images. Researchers have developed multiple retrieval systems for high-resolution CT (HRCT) images of lungs. The Comparison Algorithm for Navigating Digital Image Databases (CANDID) [14] obtains a global signature using various features like texture, shape, and color of each image stored in the database. The ASSERT [30] is a semi-automatic approach for retrieval of HRCT images, which uses a physician to mark the pathology bearing region (PBR) and anatomical landmarks. Features like texture, shape, edges, and intensity values were used for retrieval. The medGIFT [24] retrieval system is designed for multi-modal medical images. It uses open-source image finding tool with a picture archival and communication system. It uses textual labels and visual information for retrieval.

Lam et al. [16] developed a retrieval system named “BRISC” to retrieve slices of the lung nodule from a lung nodule database. The query lung nodule slice was delineated by the radiologists and the textural features using Haralick co-occurrence, Gabor filters, and Markov Random fields were extracted. For retrieval, two feature vectors were compared using the various distance metrics. Retrieved slices were considered to be relevant if it were the other slices of the query nodule. This technique was evaluated 2424 slices from 141 nodules and achieved a precision of 88% for top one retrieval. This system only retrieves the instances from the same nodule, hence could not be of much use in medical education.

Seitz et al. developed a CBIR system that uses a biggest representative slice from the nodule as a query [29]. The nodule boundary was delineated by the radiologists. The texture, size, shape, and intensity features were extracted to retrieve the similar nodules from the database. The genetic algorithm was used for feature selection and used Euclidean distance for comparing the similarity between two feature

vectors. The retrieval performance is evaluated using the LIDC database. They have considered nodules as benign with composite malignancy ratings of 1 or 2 and malignant nodule with ratings of 4 or 5. The retrieval performance was reported using 914 nodules from 399 subjects and achieved a precision of 85.95% at top three retrievals. This tool can be useful in medical education or differential diagnosis, but the delineation of nodule boundary may inhibit the usage by a radiologist.

In the CBIR system developed by Dhara et al. [9], a query nodule was segmented by semi-automated technique [7], which needs seed point from the user for segmentation. Various features like shape, margin, and texture features were extracted from the biggest representative slice and the whole nodule. Three distance metrics, i.e., Euclidean, City block, and Chebyshev, were used for retrieval. The CBIR system was evaluated using 891 nodules from 554 subjects of LIDC/IDRI database [1–3]. The retrieval performance was obtained for three configurations. In configuration 1, nodules with a composite rank of malignancy 1, 2 are considered as benign and 4, 5 as malignant. In configuration 2, nodules with a composite rank of malignancy 1, 2, 3 are considered as benign and 4, 5 as malignant, whereas in configuration 3, nodules with a composite rank of malignancy 1, 2 are considered as benign and 3, 4, 5 as malignant. The system achieved a precision of 82.14%, 75.91%, and 74.27% at top three retrievals for configuration 1, configuration 2, and configuration 3, respectively. In this system, retrieval is carried out using SDR. The system can be useful in self-learning and differential diagnosis of lung nodules.

The feature selection reduces the time required for feature extraction, training a model for classification, and similarity computation during retrieval. Moreover, it enhances the generalization of the classifier and reduces overfitting. The feature selection is extensively used in various machine learning applications. However, the efficacy of feature set selection for retrieval is not explored much. Seitz et al. used a genetic algorithm to find the optimal feature set for retrieval [29]. A set of relevant features can be selected by analyzing the area under the receiver operating characteristics (AUC) and the p -value of the feature under consideration. AUC and p -values can give the relevance of a feature with the target, but this technique analyzes single feature at a time and does not consider the redundancy between the two features. A minimal-redundancy-maximal-relevance (mRMR) feature selection technique considers both the relevance of the feature with target and redundancy between the selected features [26].

Mukhopadhyay et al. developed a learning-based retrieval framework using the fuzzy class membership for the query nodule [23]. In this approach, unlike the conventional classifier-based retrieval, a search space was

not limited to a single winning class. It assigned a weight to the distance between two feature vectors based on fuzzy class membership. The retrieval performance of the approach was evaluated on the three Brodatz texture databases, i.e., small size rotated (D1), medium size non-rotated (D2), and large size non-rotated (D3). This approach achieved a recall of 96.72%, 93.68%, and 86.01% in comparison with 63.93%, 74.71%, and 68.94% for D1, D2, and D3 databases with SDR. This retrieval approach had shown a better performance over the conventional classifier-based retrievals techniques for the texture database.

CBIR System for Pulmonary Nodules Using Optimal Feature Set and Class Membership-Based Retrieval

The proposed CBIR system is developed to retrieve similar lung nodules from a large lung CT image database using learning-based retrieval approach. For formation of a query nodule, the user provides a seed point for indication of query nodule. Considering a seed point as a centroid, a cubic volume of interest (VOI) of size (40 mm × 40 mm × 40 mm) is selected. The block diagram of the proposed CBIR system is given in Fig. 1. The steps in the proposed CBIR system are the query nodule segmentation (Fig. 2), nodule feature extraction, feature selection, and similar nodules retrieval.

Segmentation of Pulmonary Nodules

Accurate segmentation of lung nodules is important to improve the performance of the CBIR system. The lung nodules are segmented using the technique proposed by Dhara et al. [7]. The nodule segmentation technique can segment different types of lung nodules, such as solid, part-solid and non-solid, juxta-pleural, and juxta-vascular

Fig. 1 Block diagram of the proposed CBIR system for pulmonary nodules

Fig. 2 Block diagram of lung nodule segmentation proposed by Dhara et al. [7]

Fig. 3 Output of nodule segmentation algorithm for different types of nodules are **a** solid, **b** part-solid, **c** non-solid, **d** juxta-pleural, and **e** juxta-vascular. The ground truth of lung nodules is indicated by *green contours* (—) and the nodule segmentation output is indicated by *red contours* (—)

(Fig. 3 shows few examples of segmented nodules). Figure 2 shows the block diagram of the nodule segmentation algorithm. At first, the nodule segmentation framework classifies the input VOI as solid/part-solid or non-solid nodule by analyzing the texture of the core of the nodule. The two different segmentation algorithms are used to segment the solid/part-solid nodule and non-solid nodules, respectively.

For the solid/part-solid nodules, the VOI is thresholded (using -500 Hu), and the biggest connected component is selected for segmentation. The holes in the nodule are filled with the morphological closing operation, and an ellipsoid is fitted on the nodule. After the pre-processing of VOI, the overlap of boundary voxels of VOI and segmented object is examined for detection of attachment of pleura [15]. The pleural attachment is removed using the techniques of Kuhnigk et al. [15] and Moltz et al. [22] and the refinement of the pleural removal is achieved using morphological dilation [7]. For the juxta-vascular nodules, vessels are pruned using the geodesic distance map.

For the non-solid nodules, the VOI is thresholded (using -800 Hu) to remove the parenchyma, and the biggest connected component is selected for further segmentation. After filling the holes in the selected component, an anisotropic diffusion is applied to the gray-scale version of the thresholded image [27]. An ellipsoid is fitted on the segmented nodule for the further processing. The procedure for detection and removal of the pleural surface is the same for the solid/part-solid and non-solid nodules. For the removal of vessels attached to the nodules, multi-scale filtering using Hessian matrix [18] followed by adaptive thresholding and connected component analysis is used.

Nodule Feature Extraction

Lung nodules are represented in multidimensional feature space using feature vectors. Each feature vector corresponds to a lung nodule from a subject in the database. The features used to represent nodules can be grouped into several

categories like shape-based, margin-based and texture-based features [9]. Table 1 shows the list of features used in the proposed CBIR system.

3D Shape-based Features

The 3D shape features are computed using the binary mask of the segmented nodule. The 3D shape features like volume [10], sphericity [32], spiculation [10], lobulation [10], surface area of nodule [19], convex surface area [19], major axis length 3D [22], and minor axis length 3D [22] are used for retrieval.

2D Shape-based Features

These 2D shape-based features are extracted from the biggest representative slice of the segmented nodule [29]. The 2D shape-based features used in the CBIR system are area, perimeter, equivalent diameter 2D, convex area, convex perimeter, roughness, compactness, circularity, major axis length, and minor axis length.

Margin-based Features

The margin sharpness is an important diagnostic characteristic for estimation of malignancy of nodules [2]. It is represented using the features like acutance [28], [8] and histogram spread [33] of the averaged gradient of the nodule [8].

2D Texture-based Features

These texture features are extracted from the biggest axial slice of the nodule.

- **Haralick features:** Haralick features are computed using gray-level co-occurrence matrix (GLCM) [9], [13]. Six Haralick features are computed (viz. entropy,

Table 1 Feature set for representing a nodule with mRMR ranking

Type	Sl. no.	Feature name	mRMR rankings	
			LIDC/IDRI	PGIMER-IITKGP
3D shape	1	Sphericity	52	49
	2	Spiculation	5	5
	3	Lobulation	37	46
	4	Volume	51	12
	5	Equivalent diameter 3D	25	6
	6	Surface area of nodule	57	25
	7	Convex surface area	60	17
	8	Major axis length 3D	59	54
	9	Minor axis length 3D	21	28
Margin	10	HSAG	22	10
	11	Acutance 3D	58	47
Haralick 3D	12	Mean contrast	26	35
	13	Mean entropy	16	24
	14	Mean energy	24	30
	15	Mean inverse difference moment	29	43
	16	Mean sum variance	23	40
	17	Mean sum entropy	19	31
	18	Mean difference variance	7	44
	19	Mean information measure of correlation	4	34
	20	Range contrast	31	41
	21	Range entropy	48	2
	22	Range energy	28	23
	23	Range inverse difference moment	40	29
	24	Range sum average	9	22
	25	Range sum variance	36	18
	26	Range sum entropy	1	36
	27	Range sum squares	33	8
	28	Range difference entropy	44	39
	29	Range difference variance	45	14
2D shape	30	Area	47	15
	31	Convex Area	54	3
	32	Circularity	30	19
	33	Perimeter	55	38
	34	Convex perimeter	49	42
	35	Roughness	18	13
	36	Equivalent diameter 2D	13	20
	37	Major axis length	42	51
	38	Minor axis length	34	9
	39	Compactness	35	33
Haralick 2D	40	Contrast	41	37
	41	Entropy	8	11
	42	Energy	39	27
	43	Inverse difference moment	17	32
	44	Sum entropy	3	16
Gabor 2D	45	Difference entropy	14	7
	46	GaborSD_0.03	46	48

Table 1 (continued)

Type	Sl. no.	Feature name	mRMR rankings	
			LIDC/IDRI	PGIMER-IITKGP
HOG	47	GaborSD_0_04	2	26
	48	GaborSD_0_05	10	50
	49	GaborSD_45_03	50	52
	50	GaborSD_45_04	15	53
	51	GaborSD_45_05	20	55
	52	GaborSD_90_03	27	56
	53	GaborSD_90_04	32	57
	54	GaborSD_90_05	53	58
	55	GaborSD_135_03	56	59
	56	GaborSD_135_04	38	45
	57	GaborSD_135_05	43	60
	58	Mean of HOG	11	1
	59	Standard deviation of HOG	6	21
	60	Variance of HOG	12	4

energy, inverse difference moment, sum entropy, difference entropy, and contrast) using several separate GLCMs generated using different directions (0° , 45° , 90° , 135°) and distances as (1, 2, 3, 4) pixels. The mean of each computed Haralick feature from several GLCM is used as a feature in the proposed CBIR system.

- **Gabor features:** Gabor filters are widely used in texture analysis. In the proposed CBIR system, 12 Gabor filters obtained by considering 4 orientations (0° , 45° , 90° , 135°) and 3 frequencies (0.3, 0.4, 0.5) are used [9]. These Gabor filters are convolved with the biggest axial slice of the nodule, and standard deviations of the response images are considered as features for retrieval.
- **Histogram of oriented gradients (HOG):** This technique counts the occurrences of oriented gradients in a localized portion of an image [4]. The mean, variance, and standard deviation are used in the proposed CBIR system.

3D Haralick Features

Considering nine directions, nine GLCMs were computed for each nodule [12]. The mean and range of each Haralick feature over the nine directions is computed. The list of 3D Haralick features used in the proposed CBIR system is given in Table 1.

Feature Selection

Feature selection is a process of selecting a subset of relevant features from the set of extracted features. Many

extracted features can be redundant or irrelevant, and thus can be removed without incurring much loss in classification error. In the proposed CBIR system, we have selected mRMR as a feature selection technique, since the mRMR considers both the relevance of the feature with target and redundancy between the selected features [26].

Minimal-Redundancy-Maximal-Relevance

The mRMR feature selection technique was proposed by Peng et al. [26]. This technique uses maximal relevance criterion along with the minimal redundancy criterion for the selection of good features. The purpose of feature selection is to find a feature set S with m features $\{x_i\}$, which jointly have the largest dependency on the target class c . Max-relevance criterion is to find a set of features, which satisfies the condition:

$$\max D(S, c), D = \frac{1}{|S|} \sum_{x_i \in S} I(x_i; c) \quad (1)$$

where $I(x_i; c)$ is the mutual information between feature x_i and target class c .

It is highly possible that features selected using above criterion may have redundancy. When two features are highly dependent, the inclusion of these features will not bring much change in classification performance. Hence, the minimal redundancy condition is added to achieve better performance.

$$\min R(S), R = \frac{1}{|S|^2} \sum_{x_i, x_j \in S} I(x_i; x_j) \quad (2)$$

The combination of above two criteria is called as minimal-redundancy-maximal-relevance. For the optimiza-

tion of above two criteria, the operator $\Phi(D - R)$ is defined as follows:

$$\max \Phi(D, R), \Phi = D - R \quad (3)$$

To find the set of features, incremental feature selection methods are used. Suppose, we have already selected a feature set S_{m-1} with $(m - 1)$ features. Now, to select the m th feature from the set of $\{X - S_{m-1}\}$ features, optimization condition will be as follows:

$$\max_{x_j \in X - S_{m-1}} \left[I(x_j; c) - \frac{1}{m-1} \sum_{x_i \in S_{m-1}} I(x_i; x_j) \right] \quad (4)$$

Using mRMR, we get a prioritized list of features. The next step is to select an optimal set of features using a prioritized list of features. The procedure for selection of an optimal set of features for classification and retrieval is explained in “[Feature Set Selection for Classification](#)” and “[Feature Set Selection for Retrieval](#),” respectively.

Retrieval of Similar Nodules

For a given query, a set of similar nodules are retrieved by comparing the similarity between the feature vectors. In a simple distance-based retrieval system for lung nodules, the similarity of the query nodule feature vector is compared with all the stored nodule feature vectors in the database using different distance metrics. Computed distances are arranged in the ascending order for retrieval. Using these distances, most matching nodule cases are retrieved by the system. Here, the number of top retrievals is dependent on the user.

SDR-based CBIR framework for nodules is not very efficient because feature set used to represent a nodule may not be able to distinguish nodule types viz. benign and malignant with high accuracy. Classifier-based retrieval techniques classify the query nodule and retrieve the cases from the class label given by the classifier. This makes the search space for retrieval limited to the winning class. In this conventional classifier-based retrieval (CCBR) approach, retrieval performance is dependent on the classifier’s performance [20]. In this scenario, for correct classification of query nodule, retrieval performance will be 100%, whereas, for misclassification, CCBR provides completely erroneous retrieval. To take care of retrieval performance in both situations, class membership-based retrieval scheme was proposed by Mukhopadhyay et al. [23]. This retrieval approach does not limit the search space for retrieval to a single winning class. The CMR scheme is explained in “[Class Membership-based Retrieval](#),” and the approach for feature set selection for classification and retrieval is

explained in “[Feature Set Selection for Classification](#),” and “[Feature Set Selection for Retrieval](#),” respectively.

Class Membership-based Retrieval

Figure 4 shows the block diagram of class membership-based retrieval. This retrieval scheme consists of computation of fuzzy class membership for query nodule, selection of distance metric, and computation of similarity based on the class membership value. The procedure of CMR is explained as below:

- **Computation of fuzzy class membership for query nodule:** A multi-layered feedforward neural network with one hidden layer is trained using the nodule feature vectors and the corresponding targets. The output layer contains only two neurons, since there are only two nodule types viz. benign and malignant. Considering the feature vector dimension of n , the hidden layer contains $(2n + 1)$ neurons. The scaled conjugate gradient algorithm with cross-entropy as a cost function is used while training the network. A 4-fold cross-validation approach is used for evaluation of the classification as well as retrieval performance. For a query nodule X with feature vector $f(X)$, fuzzy class membership value for j^{th} class is given as follows:

$$\mu_j(f(X)) = \frac{o_j(f(X))}{\sum_{i=1}^c o_i(f(X))} \quad (5)$$

where $o_j(f(X))$ is the output of j^{th} neuron at the output layer for input feature vector $f(X)$ and c is the number of classes.

- **Weighted distance metric for retrieval:** In CCBR approach, search space is limited to a single winning class. In this approach for correct classification, retrieval performance is 100%, whereas for misclassification, it fails completely. To take care of these conditions, CMR uses weighted distance metric which

Fig. 4 Class membership-based retrieval

is dependent on the winning neuron output and fuzzy membership of query nodule. Use of this weighted metric is to assign a minimum penalty to a winning class and relatively higher penalty to the other classes. The CMR approach does not limit the search space for retrieval to a single class. Hence, there is a chance to retrieve the correct cases in case of misclassification, also. The weighted distance between query nodule X with feature vector $f(X)$ and the nodule Y_j with feature vector $f(Y_j)$ from the stored databases is given as follows:

$$d_w(X, Y_j) = \frac{1}{1 + (\xi \times \mu_j(f(X)))} \times d(X, Y_j) \quad (6)$$

where ξ is a non-negative integer, $\mu_j(f(X))$ is the fuzzy membership of query nodule X to output class j , $d(X, Y)$ is the distance between $f(X)$ and $f(Y)$ computed using any conventional distance metric. In this proposed CBIR, three conventional distance metrics, viz. Canberra, City block, and Euclidean, are used for analysis. Distance metrics between the feature vectors $A(a_1, a_2, a_3, \dots, a_n)$ and $B(b_1, b_2, b_3, \dots, b_n)$ are represented as follows:

$$\text{Canberra distance, } d(X, Y) = \sum_{i=1}^n \frac{|X_i - Y_i|}{|X_i| + |Y_i|} \quad (7)$$

$$\text{City block distance, } d(X, Y) = \sum_{i=1}^n |X_i - Y_i| \quad (8)$$

$$\text{Euclidean distance, } d(X, Y) = \sqrt{\sum_{i=1}^n (X_i - Y_i)^2} \quad (9)$$

where n is the dimension of each feature vector.

Feature Set Selection for Classification

In the proposed approach, we have used different feature sets for classification and retrieval. In the CCBIR and CMR, a neural network classifier is used. Training with a large set of features may lead to overfitting. Hence, to avoid overfitting and to reduce the time for training, an optimal feature set selection is important. To select an optimal set of features for classification, first, we have used mRMR feature selection technique for ranking the features, and then the cross-validation classification error is analyzed for sequential incremental feature selection using linear discriminant analysis (LDA). The number of features corresponding to a minimum classification error is considered as an optimal number of features for training the neural network in retrieval approaches like CCBIR and CMR. Figure 6a shows the performance of LDA classifier for the sequential incre-

mental feature selection, and the features corresponding to minimum cross-validation classification error is selected for classification using a neural network for learning-based retrieval approach. Here, it is visible that classification error using 22 features is less than the error using all 60 features.

Feature Set Selection for Retrieval

Generally, in a retrieval system, all the computed features are used for retrieval. A large number of features may lead to excessive feature calculations, and thus lead to time delay while feature extraction and retrieval. To reduce the time delay and to improve the performance of retrieval, a feature selection technique is used to find the optimal feature set for retrieval. For a selection of an optimal set of features for retrieval, the performance of CMR in terms of precision is analyzed for incremental feature selection. Radiologists will be interested in the top five to ten retrieved nodule cases. So, the error in precision is analyzed using the top seven retrieved nodule cases. In CMR, a classifier is used for retrieval, and the procedure of selecting an optimal feature set for the classification is already discussed in “[Feature Set Selection for Classification](#).” To find the optimal feature set for retrieval, we have used feature selection technique mRMR to find the rankings of the extracted features. After getting the ranking, the precision error is obtained by the 4-fold cross-validation using sequential incremental feature sets for CMR. The number of features corresponding to the minimum error in precision at top seven retrieved nodules is chosen as optimal feature set. Figure 7a shows the error in precision using Canberra distance for sequential incremental feature selection using feature rankings obtained through mRMR feature selection technique. It can be observed from the error response that error in precision using 60 features is more than the error using four features. To find the optimal feature set for SDR, the procedure is same except it does not require any classifier and cross-validation. A prioritized ranking of features is used followed by analysis of error in precision for sequential incremental feature selection for SDR.

Performance Metrics

The proposed retrieval system will be used by the radiologists for interpretation of different types of nodules. Radiologists will be interested to see the first few top retrievals for their education and interpretation. Hence, in this case, the high precision of the system is essential. The proposed CBIR system is evaluated using the following performance metrics.

Fig. 5 Complementary cumulative precision distribution

- **Average precision and standard deviation of precision:** The precision of the CBIR system is given as follows:

$$\text{Precision}(\%) = \frac{|\Phi_T \cap \Phi_R|}{|\Phi_T|} \times 100\% \quad (10)$$

where Φ_T is the set of top retrieved images from the database and Φ_R is the set of relevant images in the database. Mean and standard deviation of the precision are used to evaluate the performance of the proposed CBIR.

- **Complementary cumulative precision distribution (CCPD):** This is a graphical metric, proposed by Dash et al. [6], used to compare the performance of different retrieval techniques at a particular point considering all the images from the database as a query. Considering a constant number of top retrievals, CCPD graphically represents the fraction of query images at different values of precision. Figure 5 shows the CCPD at different precision values. As the CBIR system for medical images is a precision critical application, a fraction of queries with *zero* precision retrieval is considered as a figure of demerit for the evaluation of CBIR systems. The place in the graph showing fraction of queries having zero precision is shown in Fig. 5.

Database

The proposed CBIR system is evaluated on the two nodule databases: a publicly available Lung Image Database Consortium and Image Database Resource Initiative (LIDC/IDRI)

database [1–3] and a private database developed in collaboration with Postgraduate Institute of Medical Education & Research (PGIMER), Chandigarh.

LIDC/IDRI lung nodule database

The LIDC/IDRI database¹ is publicly available in The Cancer Imaging Archive (TCIA) [1]. Each CT slice in the database has a resolution of 512×512 pixels and pixel size ranging from 0.5 to 0.8 mm with 12-bit gray-scale resolution in Hounsfield units (HU). A group of four radiologists has evaluated the nodules and has provided the ratings for nine diagnostic characteristics of all nodules in an XML file. Malignancy rating of each nodule is given on a scale of 1:5, where rank-1 indicates the rare chance of malignancy and rank-5 indicates highly suspicious for malignancy. Since a nodule has malignancy ratings from the individual radiologist, the composite rank of malignancy

¹*Data Citation:* Armato III, Samuel G., McLennan, Geoffrey, Bidaut, Luc, McNitt-Gray, Michael F., Meyer, Charles R., Reeves, Anthony P., Clarke, Laurence P. (2015). Data From LIDC-IDRI. The Cancer Imaging Archive. <https://doi.org/10.7937/K9/TCIA.2015.LO9QL9SX>

Publication Citation: Armato SG III, McLennan G, Bidaut L, McNitt-Gray MF, Meyer CR, Reeves AP, Zhao B, Aberle DR, Henschke CI, Hoffman EA, Kazerooni EA, MacMahon H, van Beek EJR, Yankelevitz D, et al.: The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A completed reference database of lung nodules on CT scans. *Medical Physics*, 38: 915–931, 2011.

TCIA Citation: Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, Moore S, Phillips S, Maffitt D, Pringle M, Tarbox L, Prior F. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository, *Journal of Digital Imaging*, Volume 26, Number 6, December, 2013, pp 1045-1057.

is obtained by considering the mode of ratings, and in case of multiple modes, floor value of median of ratings is considered. The nodules with a composite rank of malignancy as “1,” “2” are considered to be benign, “4,” “5” as malignant, and “3” are not considered. The proposed CBIR system is evaluated on the 542 nodules from the 327 subjects. From the 542 nodules, benign and malignant nodules are 279 and 263, respectively.

PGIMER-IITKGP Lung Nodule Database

We have developed a private lung nodule database in collaboration with PGIMER, Chandigarh. The proposed CBIR system is evaluated on the 446 nodules from the 98 subjects. Each CT slice in the database has a resolution of 512×512 pixels and pixel size ranging from 0.47 to 0.95 mm with 16-bit gray-scale resolution in HU. A group of three radiologists, with consensus, has given ten diagnostic features for each lung nodule which are stored in a MATLAB data file. Malignancy rating of each nodule is given on a scale of 1:5, where rank-1 indicates the rare chance of malignancy and rank-5 indicates highly suspicious for malignancy. Here, also the nodules with the rank of malignancy as “1,” “2” are considered as benign, “4,” “5” as malignant, and “3” are not considered. Using this categorization, the PGIMER-IITKGP database contains 347 benign and 99 malignant nodules.

Results and Discussions

In the proposed CBIR using CMR, unlike the conventional method of using the same set of features for both classification and retrieval, two different feature sets are

used: one for classification and one for retrieval. The experiment is carried out on both databases. In this experiment, three different distance metrics are used viz. Canberra distance, City block distance, and Euclidean distance. The results of retrieval schemes like SDR, CCB, and CMR are compared on the basis of average precision, the standard deviation of precision, and the fraction of queries having zero precision retrieval using CCPD.

Feature Set Selection for Classification

To find the optimal feature set for classification out of total 60 features, at first, an mRMR feature selection is employed. Ranking of the features using mRMR feature selection technique for both the database is given in Table 1. After getting the rankings of all 60 features, cross-validation classification error for sequential incremental features is analyzed to obtain the optimal number of features. For LIDC/IDRI database, a minimum error is observed using a set of 22 features as shown in Fig. 6a. For the PGIMER-IITKGP database, as shown in Fig. 6b, a set of 31 features shows the minimum cross-validation classification error. For the LIDC/IDRI database, a neural network is trained with a set of 22 features, whereas for PGIMER-IITKGP database, it is trained with a set of 31 features.

Feature Set Selection for Retrieval

An optimal set of features for retrieval is obtained by analyzing the error in precision for incremental feature selection. Features rankings for sequential incremental feature selection is obtained using the mRMR feature selection technique. Figure 7 shows the CMR's performance in terms of average 4-fold cross-validation precision error

Fig. 6 Cross-validation classification error for sequential incremental feature selection using LDA for **a** LIDC/IDRI database, **b** PGIMER-IITKGP database

Fig. 7 Error in precision for sequential incremental feature selection for CMR for LIDC/IDRI database using **a** Canberra distance, **b** City block distance, and **c** Euclidean distance

for LIDC/IDRI database using different distance metrics. For Canberra distance, a feature set with four features achieves an average precision of 87.93%, whereas a feature set with 60 features achieves 87.53% average precision. For City block distance, average precision using eight features is 88.17% and using 60 features is 88.01%. Using Euclidean distance, a set of eight features achieves an average precision of 87.88% and 87.56% for 60 features. Figure 8 shows a similar set of results for PGIMER-IITKGP database. For Canberra distance, a feature set with 18 features achieves an average precision of 85.11%, whereas a feature set with 60 features achieves 84.05% average precision. For City block distance, average precision using 20 features is 84.79% and using 60 features is 83.89%. Using Euclidean distance, a set of 21 features achieves an average precision of 84.56 and 83.89% for 60 features. From these graphs, it is evident that use of an optimal feature set provides a gain in retrieval performance, and thus the time for feature extraction and retrieval will reduce.

The performance of the proposed CMR system is compared with the SDR with an optimal feature set. The procedure of an optimal feature set selection for

SDR is discussed in “[Feature Set Selection for Retrieval](#).”

Figure 9 shows the error in precision corresponding to the incremental feature set selection for the LIDC/IDRI database. For Canberra distance, feature set containing five features achieves a precision of 82.81% and a set of 60 features achieves 81.89% precision. Using City block distance, a set of 11 features achieves a precision of 82.95%, and a set of 60 features gives a precision of 82.45%. For Euclidean distance, a set of 22 features achieves 83% as compared to 82.18% with 60 features. Similarly, error in precision performance for incremental feature selection using PGIMER-IITKGP database is shown in Fig. 10. Using Canberra distance, a feature set containing six features achieves 79.02% precision as compared to a precision of 76.17% with 60 features. For City block distance, a set of 20 features achieves a precision of 78.19%, and a complete feature set gives a precision of 75.95%. For Euclidean distance, the precision of 77.74% and 76.36% is achieved using a set of three and 60 features, respectively. From Figs. 9 and 10, it is clear that the performance of SDR scheme using an optimal feature set is better than the performance with the complete feature set.

Fig. 8 Error in precision for sequential incremental feature selection for CMR for PGIMER-IITKGP database using **a** Canberra distance, **b** City block distance, and **c** Euclidean distance

Fig. 9 Error in precision for sequential incremental feature selection for SDR for LIDC/IDRI database using **a** Canberra distance, **b** City block distance, and **c** Euclidean distance

CMR Performance with Two Different Feature Sets

In the proposed CBIR framework, two different feature sets are used for classification and retrieval. From the analysis of classification error and precision error, we have selected a feature set for classification and retrieval using different distance metrics. A multi-layered feedforward neural network with one hidden layer is trained using 4-fold cross-validation approach. For the LIDC/IDRI database, a set of 22 features obtained using feature selection procedure is used for classification. The input layer of neural network contains 22 neurons, and hidden layer contains 45 neurons. A stochastic gradient back-propagation with cross-entropy as a cost function is used for training a neural network. For retrieval using CMR technique, a set of four features is used in case of Canberra distance, and a set of eight features is used in case of City block and Euclidean distance. For the PGIMER-IITKGP database, a set of 31 features is used for training of the classifier. The input layer of neural network contains 31 neurons, and hidden layer contains 63 neurons. For retrieval using CMR technique, a set of 18, 20, 21 features are used in case of Canberra distance, City block distance and Euclidean distance, respectively.

CMR Performance in Terms of Average Precision and Standard Deviation

Retrieval performance of the proposed CBIR system is compared with other retrieval approaches, i.e., CBIR system by Dhara et al. (SDR with all features) [9], SDR using an optimal feature set, CCBR, and CMR with the complete feature set. For LIDC/IDRI database, in case of SDR using an optimal feature set, retrieval is performed using a feature set of 5, 11, 22 features for Canberra, City block, and Euclidean distance, respectively. Tables 2, 3, 4, 5, 6, and 7 shows the performance of CMR approach with the proposed optimal feature sets for a different number of top retrievals (Ntop) using different distance metrics. The SDR with an optimal feature set achieves a marginal improvement over the CBIR system by Dhara et al. in retrieval performance. This shows that optimal feature selection improves the performance of the retrieval system. The same observation holds true for the CMR with all features and CMR with an optimal feature set. The CMR with an optimal feature set for classification and retrieval performs better than the CMR with all features. After using the optimal feature sets for classification and retrieval, proposed CMR retrieval with an

Fig. 10 Error in precision for sequential incremental feature selection for SDR for PGIMER-IITKGP database using **a** Canberra distance, **b** City block distance, and **c** Euclidean distance

Table 2 Comparison of retrieval approaches in terms of average precision for LIDC/IDRI database using Canberra distance

Method	No. of top images for retrieval							
	1	3	5	7	9	11	13	15
Dhara et al.	83.95	82.41	81.96	81.89	82	82.1	82.13	82.12
SDR with optimal features	82.10	83.15	82.95	82.81	82.68	82.52	82.34	82.19
CCBR	87.83	87.83	87.83	87.83	87.83	87.83	87.83	87.83
CMR with all features	88.93	88.07	87.60	87.48	87.60	87.67	87.64	87.71
CMR with optimal features	88.56	88.07	88.01	87.93	87.99	88.11	88.04	87.93

Italic entries were to emphasize the technique with higher average precision

Table 3 Comparison of retrieval approaches in terms of standard deviation of precision for LIDC/IDRI database using Canberra distance

Method	No. of top images for retrieval							
	1	3	5	7	9	11	13	15
Dhara et al.	36.74	29.48	28.02	27.47	27.11	26.73	26.52	26.08
SDR with optimal features	38.37	29.14	27.14	26.01	25.63	25.34	25.28	25.26
CCBR	32.73	32.73	32.73	32.73	32.73	32.73	32.73	32.73
CMR with all features	31.41	29.87	30.03	30.16	30.04	29.87	29.93	29.81
CMR with optimal features	31.86	29.53	29.11	29.17	28.99	28.83	28.91	29.04

Italic entries were to emphasize lower standard deviation

Table 4 Comparison of retrieval approaches in terms of average precision for LIDC/IDRI database using City block distance

Method	No. of top images for retrieval							
	1	3	5	7	9	11	13	15
Dhara et al.	83.95	82.53	82.44	82.45	82.33	82.09	81.86	81.88
SDR with optimal features	83.58	82.41	82.99	82.95	82.37	82.29	82.07	81.82
CCBR	87.83	87.83	87.83	87.83	87.83	87.83	87.83	87.83
CMR with all features	88.01	87.82	87.93	87.74	87.68	87.67	87.65	87.61
CMR with optimal features	88.38	88.25	88.15	88.17	88.01	87.99	88.01	88.01

Italic entries were to emphasize the technique with higher average precision

Table 5 Comparison of retrieval approaches in terms of standard deviation of precision for LIDC/IDRI database using City block distance

Method	No. of top images for retrieval							
	1	3	5	7	9	11	13	15
Dhara et al.	36.74	29.20	27.00	26.51	26.33	26.40	26.50	26.22
SDR with optimal features	37.08	29.82	27.55	27.06	26.71	26.43	26.16	25.94
CCBR	32.73	32.73	32.73	32.73	32.73	32.73	32.73	32.73
CMR with all features	32.52	31.58	31.05	31.26	31.21	31.20	31.13	31.13
CMR with optimal features	32.08	30.19	29.87	29.65	29.75	29.76	29.62	29.56

Italic entries were to emphasize lower standard deviation

Table 6 Comparison of retrieval approaches in terms of average precision for LIDC/IDRI database using Euclidean distance

Method	No. of top images for retrieval							
	1	3	5	7	9	11	13	15
Dhara et al.	83.39	82.72	82.18	82.18	82.02	82.09	82.09	81.99
SDR with optimal features	83.95	82.23	82.73	83.00	82.64	82.22	82.13	82.25
CCBR	87.83	87.83	87.83	87.83	87.83	87.83	87.83	87.83
CMR with all features	88.19	87.64	87.45	87.37	87.43	87.59	87.60	87.54
CMR with optimal features	88.56	88.07	87.71	87.88	87.90	87.82	87.78	87.71

Italic entries were to emphasize the technique with higher average precision

Table 7 Comparison of retrieval approaches in terms of standard deviation of precision for LIDC/IDRI database using Euclidean distance

Method	No. of top images for retrieval							
	1	3	5	7	9	11	13	15
Dhara et al.	37.25	28.92	27.06	26.71	26.32	26.12	25.71	25.57
SDR with optimal features	36.74	29.26	27.16	26.58	26.15	25.90	25.75	25.44
CCBR	32.73	32.73	32.73	32.73	32.73	32.73	32.73	32.73
CMR with all features	32.30	32.00	31.88	31.72	31.60	31.39	31.32	31.33
CMR with optimal features	31.86	30.01	29.74	29.32	29.15	29.12	29.12	29.11

Italic entries were to emphasize lower standard deviation

Table 8 Comparison of retrieval approaches in terms of average precision for PGIMER-IITKGP database using Canberra distance

Method	No. of top images for retrieval							
	1	3	5	7	9	11	13	15
Dhara et al.	78.70	76.76	76.59	76.17	75.39	75.46	74.99	75.05
SDR with optimal features	78.25	79.52	78.48	79.02	78.62	78.21	77.75	77.22
CCBR	83.17	83.17	83.17	83.17	83.17	83.17	83.17	83.17
CMR with all features	83.86	83.93	83.95	84.05	83.91	84.06	83.89	83.87
CMR with optimal features	85.65	85.50	85.02	85.11	84.83	84.65	84.56	84.39

Italic entries were to emphasize the technique with higher average precision

Table 9 Comparison of retrieval approaches in terms of standard deviation for PGIMER-IITKGP database using Canberra distance

Method	No. of top images for retrieval							
	1	3	5	7	9	11	13	15
Dhara et al.	40.99	31.45	29.21	28.09	28.03	27.55	26.97	26.57
SDR with optimal features	41.30	32.14	29.06	27.25	26.56	26.13	25.99	25.90
CCBR	37.44	37.44	37.44	37.44	37.44	37.44	37.44	37.44
CMR with all features	36.83	33.79	33.56	33.20	33.10	32.95	33.10	33.06
CMR with optimal features	35.10	33.21	33.17	32.70	32.63	32.62	32.57	32.65

Italic entries were to emphasize lower standard deviation

Table 10 Comparison of retrieval approaches in terms of average precision for PGIMER-IITKGP database using City block distance

Method	No. of top images for retrieval							
	1	3	5	7	9	11	13	15
Dhara et al.	78.03	75.71	76.01	75.98	76.56	76.29	76.27	76.04
SDR with optimal features	77.13	78.18	78.39	78.19	77.75	77.31	76.96	76.79
CCBR	83.17	83.17	83.17	83.17	83.17	83.17	83.17	83.17
CMR with all features	83.63	83.86	83.99	83.89	83.96	84.06	84.03	84.01
CMR with optimal features	<i>84.53</i>	<i>84.83</i>	<i>84.75</i>	<i>84.79</i>	<i>84.75</i>	<i>84.43</i>	<i>84.27</i>	<i>84.23</i>

Italic entries were to emphasize the technique with higher average precision

Table 11 Comparison of retrieval approaches in terms of standard deviation of precision for PGIMER-IITKGP database using City block distance

Method	No. of top images for retrieval							
	1	3	5	7	9	11	13	15
Dhara et al.	41.45	32.31	29.21	28.61	27.55	27.29	26.85	26.59
SDR with optimal features	42.05	<i>31.48</i>	<i>28.17</i>	<i>26.55</i>	<i>26.54</i>	<i>26.32</i>	<i>26.21</i>	<i>25.99</i>
CCBR	37.44	37.44	37.44	37.44	37.44	37.44	37.44	37.44
CMR with all features	37.04	34.16	33.70	33.63	33.40	33.07	33.10	33.16
CMR with optimal features	<i>36.20</i>	33.62	33.04	32.93	32.75	32.85	32.90	32.82

Italic entries were to emphasize lower standard deviation

Table 12 Comparison of retrieval approaches in terms of average precision for PGIMER-IITKGP database using Euclidean distance

Method	No. of top images for retrieval							
	1	3	5	7	9	11	13	15
Dhara et al.	78.25	76.83	76.41	76.36	76.73	76.46	76.16	76.25
SDR with optimal features	77.35	77.88	77.62	77.74	77.33	77.09	76.84	76.70
CCBR	83.17	83.17	83.17	83.17	83.17	83.17	83.17	83.17
CMR with all features	<i>84.53</i>	84.16	84.26	84.11	83.93	84.02	83.99	83.90
CMR with optimal features	<i>84.53</i>	<i>84.60</i>	<i>84.39</i>	<i>84.56</i>	<i>84.30</i>	<i>84.18</i>	<i>84.03</i>	83.93

Italic entries were to emphasize the technique with higher average precision

Table 13 Comparison of retrieval approaches in terms of standard deviation of precision for PGIMER-IITKGP database using Euclidean distance

Method	No. of top images for retrieval							
	1	3	5	7	9	11	13	15
Dhara et al.	41.30	32.48	29.55	28.27	<i>27.18</i>	26.86	26.87	26.38
SDR with optimal features	41.90	<i>31.59</i>	<i>29.33</i>	28.44	27.58	27.04	<i>26.64</i>	<i>26.13</i>
CCBR	37.44	37.44	37.44	37.44	37.44	37.44	37.44	37.44
CMR with all features	<i>36.20</i>	34.52	33.77	33.60	33.62	33.41	33.30	33.22
CMR with optimal features	<i>36.20</i>	34.29	33.90	33.25	33.18	33.19	33.10	33.15

Italic entries were to emphasize lower standard deviation

Table 14 Normalised standard deviation w.r.t. average precision of CBIR techniques for LIDC/IDRI database using Canberra distance at Ntop = 7

Method	Average (μ)	Standard Deviation (σ)	Norm. std. dev. ($\frac{\sigma}{\mu}$)
Dhara et al.	81.89	27.47	0.34
SDR with optimal features	82.81	26.01	0.31
CCBR	87.83	32.73	0.37
CMR with all features	87.48	30.16	0.34
CMR with optimal features	87.93	29.17	0.33

optimal feature set technique shows the better performance than both the SDR techniques, CCBR and CMR with all features. In some of the cases, CCBR performs better than the CMR with all features in terms of average precision. Hence, optimal feature set selection is important for better performance. Overall, from the Tables 2, 4, and 6, it is clear that the proposed CMR with an optimal feature set framework shows an improvement in average precision over competing retrieval techniques except for the few Ntop in case of Euclidean distance. Tables 3, 5, and 7 show that the proposed CMR with an optimal feature set framework has a better standard deviation of precision compared to CCBR. However, in most of the cases, the standard deviation of SDR with an optimal feature set is better than competing retrieval techniques, but SDR with an optimal feature set has lower average precision than CCBR and CMR.

For PGIMER-IITKGP database, Tables 8, 9, 10, 11, 12, and 13 show the performance of CMR approach for different Ntop using different distance metrics. The trend observed while analyzing the CMR's performance for LIDC/IDRI database is also consistent with the PGIMER-IITKGP database. In case of SDR using an optimal feature set, retrieval is performed using a feature set of 6, 20, 3 features for Canberra, City block, and Euclidean distance, respectively. Here, also, the SDR technique with an optimal feature set is performing better than its counterpart, i.e., CBIR system by Dhara et al. where all the features are used for retrieval. Though the CCBR achieves better performance than both the SDR techniques, proposed CMR with an optimal feature sets performs better than the CCBR performance as well as CMR with all features. In summary, from the Tables 8, 10, and 12, it is clear that the proposed CMR with an optimal feature set framework

shows an improvement in average precision over competing retrieval techniques. Tables 9, 11, and 13 show that the CMR with an optimal feature set has a better standard deviation of precision compared to CCBR, but not better than the standard deviation of SDR. However, SDR has lower average precision than CCBR and CMR.

The standard deviation of precision is high for all the retrieval techniques compared to the average precision, which will be clear from normalised standard deviation (w.r.t. average) of precision for Canberra distance and LIDC/IDRI database as shown in Table 14.

For some of the nodule queries, it is difficult to retrieve the similar samples from the same class, and for some queries, it is easy to retrieve the similar samples from the same class. This variation results in high standard deviation irrespective of techniques. The root cause of the variation is the efficacy (or lack of efficacy) of the features used for the retrieval task. Though the standard deviation of the CMR methods is high, it shows significant improvement over the SDR methods when compared with the help of p -values (using Student's t test), precision rank analysis and CCPD.

We have analyzed the statistical significance of the CBIR methods using a t test. Table 15 shows the p -values for the different CBIR techniques discussed in this paper for the LIDC/IDRI database.

From the analysis of p -values, it is evident that the precisions of the CCBR, CMR with all features and the CMR with optimal features are statistically significant when compared with the SDR with all features (Dhara et al.) and SDR with optimal features. Table 15 shows that the precisions of learning-based techniques, i.e., CCBR, CMR with all features and CMR with optimal features are not statistically different. Also, the difference between the

Table 15 p -value analysis of the precision of CBIR techniques for the LIDC/IDRI database using Canberra distance at the Ntop = 7 (the p -values with italic text indicates statistically significant values)

Method	Dhara et al.	SDR with optimal features	CCBR	CMR with all features
CMR with optimal features	<i>2.36E-04</i>	<i>1.19E-03</i>	4.78E-01	3.90E-01
CMR with all features	<i>8.04E-04</i>	<i>3.53E-03</i>	5.82E-01	
CCBR	<i>6.36E-04</i>	<i>2.70E-03</i>		
SDR with optimal features	2.85E-01			

Table 16 Precision rank of the SDR and CMR with optimal features and with all features for LIDC/IDRI database using Canberra distance at Ntop = 7 expressed in terms of percentage of total number of queries

Ranks	Higher (%)	Equal (%)	Lower (%)
SDR with optimal features Vs. SDR with all features	22.32	57.2	20.48
CMR with optimal features Vs. CMR with all features	7.38	85.61	7.01

precision of SDR with all features and SDR with optimal features is not statistically significant.

We have analyzed the precision rank of the CBIR techniques to show the efficacy of optimal feature selection with learning-based retrieval. Table 16 shows the precision rank of the SDR techniques and CMR techniques.

The precision rank of the SDR (CMR) techniques (Table 16) shows that for the 79.52% (93%) of the total number of queries, the SDR (CMR) with optimal features performs better than or as equal as the SDR (CMR) with all features. From the precision rank and mean precision, it is clear that the optimal feature selection helps the retrieval system to improve the performance.

The performance of the CMR with optimal features and CCBR technique is compared using the CCPD. The Fig. 11a shows the CCPD of CMR with optimal features and CCBR for LIDC/IDRI database using Canberra distance. The fraction of queries with zero precision retrieval using CMR with optimal features is 6.82%, whereas it is 12.17% for the CCBR technique. It shows that the CMR with optimal features helps to reduce the fraction of queries with zero precision retrieval.

From the average precision, *p*-value analysis and CCPD of precision of retrieval of LIDC/IDRI database using

Canberra distance, CMR with optimal features come out as the winner. The result remains same for other distances with LIDC/IDRI database and PGIMER-IITKGP database. The corresponding results are provided in the [Supplementary Document](#).

CMR performance evaluation using CCPD

The performance of the proposed CBIR framework with CCBR is compared using the graphical metric CCPD. Figures 11a, 12a, and 13a shows the performance comparison between two retrieval approaches, i.e., CCBR and CMR for LIDC/IDRI database using different distance metrics. This shows an improvement in a fraction of queries having zero precision retrieval for CMR over the CCBR. In CCBR approach, 12.17% of queries was having zero precision retrieval, whereas CMR using Canberra distance has reduced this fraction to 6.83% and using City block and Euclidean distance, it has reduced to 7.38% at Ntop = 7. This shows that the fraction of queries with zero precision retrieval using Canberra distance is less than other two distance metrics. Figures 11b, 12b, and 13b show the response of the fraction of queries with zero precision at different values of Ntop. This fraction of queries with zero

Fig. 11 Comparison of CMR and CCBR using CCPD for LIDC/IDRI database using Canberra distance **a** CCPD at fixed Ntop=7, and **b** fraction of queries with zero precision at different Ntop

Fig. 12 Comparison of CMR and CCBR using CCPD for LIDC/IDRI database using City block distance **a** CCPD at fixed $N_{top} = 7$, and **b** fraction of queries with zero precision at different N_{top}

precision decreases initially for increasing value of N_{top} but saturates after higher values of N_{top} .

For PGIMER-IITKGP database, the performance of the CMR with an optimal feature set is compared with CCBR using the CCPD. Figures 14a, 15a and 16a show the performance comparison in terms of the fraction of queries with zero precision retrieval between CCBR and CMR using different distance metrics at $N_{top} = 7$. In CCBR approach, 16.83% of queries were having zero precision retrieval, whereas CMR using Canberra distance has reduced it to

10.54% and using City block and Euclidean distance, it has reduced to 10.76 and 10.99%, respectively. Like LIDC/IDRI database, for PGIMER-IITKGP database, the performance of CMR using Canberra distance in terms of the fraction of queries with zero precision retrieval is better than the other two distance metrics. Figures 14b, 15b, and 16b show the response of fraction of queries with zero precision at different values of N_{top} . This fraction of queries with zero precision decreases initially but saturates after higher values of N_{top} .

Fig. 13 Comparison of CMR and CCBR using CCPD for LIDC/IDRI database using Euclidean distance **a** CCPD at fixed $N_{top} = 7$, and **b** fraction of queries with zero precision at different N_{top}

Fig. 14 Comparison of CMR and CCBR using CCPD for PGIMER-IITKGP database using Canberra distance **a** CCPD at fixed Ntop = 7, and **b** fraction of queries with zero precision at different Ntop

Comparison of the Distance Metrics Based on the Retrieval Performance

In the proposed CMR with an optimal feature set framework, we have used three distance metrics for analysis, i.e., Canberra, City block, and Euclidean. Figures 17 and 18 show the comparison of retrieval performance of SDR with an optimal feature set for three distance metrics for both the databases in terms of average precision and standard deviation of precision. For LIDC/IDRI database, in

most of the cases, both average precision and standard deviation of precision is better using Canberra distance. In case of PGIMER-IITKGP database, average precision using Canberra distance is better than the other two distance metrics, whereas standard deviation of precision for both City block and Canberra distance is marginally better than the Euclidean distance. Hence, from this analysis, we can conclude that retrieval performance of SDR with an optimal feature set using Canberra distance is marginally better than the other two distance metrics.

Fig. 15 Comparison of CMR and CCBR using CCPD for PGIMER-IITKGP database using City block distance **a** CCPD at fixed Ntop = 7, and **b** fraction of queries with zero precision at different Ntop

Fig. 16 Comparison of CMR and CCBR using CCPD for PGIMER-IITKGP database using Euclidean distance **a** CCPD at fixed $N_{top} = 7$, and **b** fraction of queries with zero precision at different N_{top}

Fig. 17 Comparison of SDR retrieval performance of three distance metrics, i.e., Canberra, City block, and Euclidean for LIDC/IDRI database **a** average precision and **b** standard deviation of precision

Fig. 18 Comparison of SDR retrieval performance of three distance metrics, i.e., Canberra, City block, and Euclidean for PGIMER-IITKGP database **a** average precision and **b** standard deviation of precision

Fig. 19 Comparison of CMR retrieval performance of three distance metrics, i.e., Canberra, City block, and Euclidean for LIDC/IDRI database **a** average precision and **b** standard deviation of precision

Figures 19 and 20 show the performance comparison of CMR with an optimal feature set using three distance metrics. For the LIDC/IDRI database, in most of the cases, average precision using City block and Canberra distance is better than the performance using Euclidean distance. However, the standard deviation of precision is better for Canberra distance than the other two distance metrics. For PGIMER-IITKGP database, both average precision and standard deviation of precision using Canberra distance are better than the performance using City block and Euclidean distance. Hence, we can conclude that CMR with optimal features using Canberra distance is performing better than the other two distance metrics.

Conclusion

We have investigated the combined effect of an optimal feature set selection and learning-based retrieval scheme, viz. CMR, on the performance of a retrieval system for lung nodules. We have observed that the SDR with an optimal feature set performs better than the SDR with all the features. We know that the learning-based retrieval techniques like CCBR and CMR perform better as compared to the SDR techniques for texture image retrieval [23]. The CCBR suffer from the high standard deviation of precision. We have used two different optimal feature sets for classification and retrieval in CMR. We have found that the CMR provides an improvement over

Fig. 20 Comparison of CMR retrieval performance of three distance metrics, i.e., Canberra, City block, and Euclidean for PGIMER-IITKGP database **a** average precision and **b** standard deviation of precision

SDR irrespective of the nature of features. The performance of CMR with optimal feature sets is better than the CMR without feature selection. In some of the cases, average precision of CMR without the optimal feature sets falls below than the CCBP. We have analyzed the retrieval performance with three different distance metrics, i.e., Canberra, City block and Euclidean distance. The proposed technique has shown an improvement in the retrieval performance (average precision) as compared to the competing techniques (viz. SDR and CCBP) irrespective of the distance metric. In addition, fraction of queries with zero precision retrieval using proposed method is less than the CCBP. We found that the retrieval performance of the proposed framework is better using Canberra distance than the City block or Euclidean. The proposed CBIR system will be helpful for budding radiologists in self-learning. Moreover, the performance of the CBIR system is dependent on the accuracy of nodule segmentation and the relevance of extracted features to discriminate nodule types. The future work is to improve the nodule segmentation and to search for the visual and clinical features to enhance the classifier's performance.

Acknowledgements The authors acknowledge the National Cancer Institute and the Foundation for the National Institutes of Health and their critical role in the creation of the free publicly available LIDC/IDRI Database used in this study.

Funding Information This study was funded by Ministry of Electronics and Information Technology, Government of India (grant no.: 1(2)/2013-ME&TMD/ESDA).

Compliance with Ethical Standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical Approval All procedures in this study involving human participants were performed in accordance with the ethical standards of the institution, and were approved by the research ethics boards at Indian Institute of Technology Kharagpur and Postgraduate Institute of Medical Education and Research, Chandigarh. This study does not contain any procedures involving animals.

Informed Consent Informed consent was obtained from all individual participants included in the study.

References

1. Armato III SG, McLennan G, Bidaut L, McNitt-Gray MF, Meyer CR, Reeves AP, Clarke LP: Data from LIDC-IDRI. The Cancer Imaging Archive. 2015. <https://doi.org/10.7937/K9/TCIA.2015.LO9QL9SX>
2. Armato III SG, McLennan G, Bidaut L, McNitt-Gray MF, Meyer CR, Reeves AP, Zhao B, Aberle DR, Henschke CI, Hoffman EA, Kazerooni EA, MacMahon H, Beek EJR, Yankelevitz D, Biancardi AM, Bland PH, Brown MS, Engelmann RM, Laderach GE, Max D, Pais RC, Qing DPY, Roberts RY, Smith AR, Starkey A, Batra P, Caligiuri P, Farooqi A, Gladish GW, Jude CM, Munden RF, Petkovska I, Quint LE, Schwartz LH, Sundaram B, Dodd LE, Fenimore C, Gur D, Petrick N, Freymann J, Kirby J, Hughes B, Castele AV, Gupta S, Sallam M, Heath MD, Kuhn MH, Dharaia E, Burns R, Fryd DS, Salganicoff M, Anand V, Shreter U, Vastagh S, Croft BY, Clarke LP: The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. *Med Phys* 38(2):915–931, 2011
3. Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, Moore S, Phillips S, Maffitt D, Pringle M, et al: The cancer imaging archive (TCIA): maintaining and operating a public information repository. *J Digit Imaging* 26(6):1045–1057, 2013
4. Dalal N, Triggs B, Schmid C: Human detection using oriented histograms of flow and appearance. In: *Computer vision—ECCV 2006*, pp. 428–441. Springer, 2006
5. Dash JK, Mukhopadhyay S, Gupta RD: Content-based image retrieval using fuzzy class membership and rules based on classifier confidence. *IET Image Process* 9(9):836–848, 2015
6. Dash JK, Mukhopadhyay S, Khandelwal N: Complementary cumulative precision distribution: a new graphical metric for medical image retrieval system. In: *SPIE Medical imaging*, pp 90,371s–90,371s. International society for optics and photonics, 2014
7. Dhara A, Mukhopadhyay S, Das Gupta R, Garg M, Khandelwal N: A segmentation framework of pulmonary nodules in lung CT images. *J Digit Imaging* 10:1007, 2015
8. Dhara AK, Mukhopadhyay S, Chakrabarty S, Garg M, Khandelwal N: Quantitative evaluation of margin sharpness of pulmonary nodules in lung CT images. *IET Image Process* 10(9):631–637, 2016
9. Dhara AK, Mukhopadhyay S, Dutta A, Garg M, Khandelwal N: Content-based image retrieval system for pulmonary nodules: Assisting radiologists in self-learning and diagnosis of lung cancer. *J Digit Imaging* 30(1):63–77, 2017
10. Dhara AK, Mukhopadhyay S, Saha P, Garg M, Khandelwal N: Differential geometry-based techniques for characterization of boundary roughness of pulmonary nodules in CT images. *Int J CARS* 11(3):337–349, 2016
11. Diederich S, Wormanns D, Semik M, Thomas M, Lenzen H, Roos N, Heindel W: Screening for early lung cancer with low-dose spiral CT: Prevalence in 817 asymptomatic smokers. *Radiology* 222(3):773–781, 2002
12. Han F, Wang H, Zhang G, Han H, Song B, Li L, Moore W, Lu H, Zhao H, Liang Z: Texture feature analysis for computer-aided diagnosis on pulmonary nodules. *J Digit Imaging* 28(1):99–115, 2014
13. Haralick RM, Shanmugam K, Dinstein IH: Textural features for image classification. *IEEE Trans Syst Man Cybern* 3(6):610–621, 1973
14. Kelly P, Cannon T, Hush D: Query by image example: the comparison algorithm for navigating image databases (CANDID) approach. In: *Proceedings of the SPIE*, 1995
15. Kuhnigk JM, Dicken V, Bornemann L, Bakai A, Wormanns D, Krass S, Peitgen HO: Morphological segmentation and partial volume analysis for volumetry of solid pulmonary lesions in thoracic ct scans. *IEEE Transactions on Medical Imaging* 25(4):417–434, 2006
16. Lam MO, Disney T, Raicu DS, Furst J, Channin DS: BRISC — an open source pulmonary nodule image retrieval framework. *J Digit Imaging* 20(1):63–71, 2007
17. Lehmann TM, Schubert H, Keyser D, Kohnen M, Wein BB: The IRMA code for unique classification of medical images. In: *Proceedings of SPIE Medical Imaging 2003*, pp 440–451, 2003

18. Li Z, Ma L, Jin X, Zheng Z: A new feature-preserving mesh-smoothing algorithm. *Vis Comput* 25(2):139–148, 2009
19. Lorensen WE, Cline HE: Marching cubes: a high resolution 3d surface construction algorithm. In: *ACM Siggraph computer graphics*, vol 21, pp 163–169. ACM, 1987
20. Ma WY, Manjunath BS: Texture features and learning similarity. In: *IEEE Computer society conference on computer vision and pattern recognition*, pp 425–430, 1996
21. Mishra S, Joseph RA, Gupta PC, Pezzack B, Ram F, Sinha DN, Dikshit R, Patra J, Jha P: Trends in bidi and cigarette smoking in India from 1998 to 2015, by age, gender and education. *BMJ Global Health* 1(1):e000,005, 2016
22. Moltz JH, Kuhnigk JM, Bornemann L, Peitgen H: Segmentation of juxtapleural lung nodules in CT scan based on ellipsoid approximation. In: *Proceedings of First International Workshop on Pulmonary Image Processing*. New York, pp 25–32, 2008
23. Mukhopadhyay S, Dash JK, Gupta RD: Content-based texture image retrieval using fuzzy class membership. *Pattern Recogn Lett* 34(6):646–654, 2013
24. Müller H, Lovis C, Geissbuhler A: The MedGIFT project on medical image retrieval. *Medical Imaging and Telemedicine*, Wujishan, China, 2005
25. Müller H., Michous N, Bandon D, Geissbuhler A: A review of content-based image retrieval systems in medical applications-clinical benefits and future directions. *Int J Med Inform* 73(1):1–23, 2004
26. Peng H, Long F, Ding C: Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans Pattern Anal Mach Intell* 27(8):1226–1238, 2005
27. Perona P, Malik J: Scale-space and edge detection using anisotropic diffusion. *IEEE Trans Pattern Anal Mach Intell* 12(7):629–639, 1990
28. Rangayyan RM, El-Faramawy NM, Desautels JL, Alim OA: Measures of acutance and shape for classification of breast tumors. *IEEE Trans Med Imaging* 16(6):799–810, 1997
29. Seitz KA Jr, Giuca AM, Furst J, Raicu D: Learning lung nodule similarity using a genetic algorithm. In: *Proceedings of SPIE Medical Imaging 2012*, pp 831537. San Deigo, USA, 2012
30. Shyu C, Brodley CE, Kak AC, Kosaka A, Aisen A: Broderick, I.: ASSERT: a physician-in-the-loop content-based retrieval system for HRCT image databases. *Comp Vision Image Underst* 75(2):111–132, 1999
31. Siegel R, Jemal A (2015) Cancer facts & figures 2015. American Cancer Society Cancer Facts & Figures
32. Sladoje N, Nyström I, Saha PK: Measurements of digitized objects with fuzzy borders in 2D and 3D. *Image Vis Comput* 23(2):123–132, 2005
33. Tripathi AK, Mukhopadhyay S, Dhara AK: Performance metrics for image contrast. In: *Proceedings of IEEE International Conference on Image Information Processing*, pp 1–4. Simla, India, 2011

Affiliations

Shrikant A. Mehre¹ · Ashis Kumar Dhara^{1,2} · Mandeep Garg³ · Naveen Kalra³ · Niranjana Khandelwal³ · Sudipta Mukhopadhyay¹

Shrikant A. Mehre
shrikant.mehre@gmail.com

Ashis Kumar Dhara
dear.ashis79@gmail.com

Mandeep Garg
gargmandeep@hotmail.com

Naveen Kalra
navkal2004@yahoo.com

Niranjana Khandelwal
khandelwaln@hotmail.com

- ¹ Department of Electronics and Electrical Communication Engineering, Indian Institute of Technology Kharagpur, Kharagpur, India
- ² Centre for Image Analysis, Uppsala University, Uppsala, Sweden
- ³ Department of Radiodiagnosis and Imaging, Post-graduate Institute of Medical Education and Research, Chandigarh, India