Diagnostic Accuracy Study

Medicine®

OPEN

# A multi-feature image retrieval scheme for pulmonary nodule diagnosis

Guohui Wei, PhD[a,c,*], Min Qiu, MD[b], Kuixing Zhang, MD[a], Ming Li, MD[a], Dejian Wei, MD[a], Yanjun Li, MD[a], Peiyu Liu, PhD[c], Hui Cao, MD[a,*], Mengmeng Xing, MD[a], Feng Yang, MD[a]

**Abstract**

Deep analysis of radiographic images can quantify the extent of intra-tumoral heterogeneity for personalized medicine.

In this paper, we propose a novel content-based multi-feature image retrieval (CBMFIR) scheme to discriminate pulmonary nodules benign or malignant. Two types of features are applied to represent the pulmonary nodules. With each type of features, a single-feature distance metric model is proposed to measure the similarity of pulmonary nodules. And then, multiple single-feature distance metric models learned from different types of features are combined to a multi-feature distance metric model. Finally, the learned multi-feature distance metric is used to construct a content-based image retrieval (CBIR) scheme to assist the doctors in diagnosis of pulmonary nodules. The classification accuracy and retrieval accuracy are used to evaluate the performance of the scheme.

The classification accuracy is $0.955 \pm 0.010$, and the retrieval accuracies outperform the comparison methods.

The proposed CBMFIR scheme is effective in diagnosis of pulmonary nodules. Our method can better integrate multiple types of features from pulmonary nodules.

**Abbreviations:** AUC = the area under the curve, CAD = Computer-Assisted Diagnosis, CBIR = content-based image retrieval, CBMFIR = content-based multi-feature image retrieval, CBMFIR = content-based multi-feature image retrieval, CNN = convolutional neural network, CT = computed tomography, DSDC = differential scatter discriminant criterion, ELM = extreme learning machine, RF = random forest, SSM-DML = semisupervised multiview distance metric learning, SVM = support vector machine.

**Keywords:** content-based image retrieval, distance metric learning, multi-feature, pulmonary nodule, similarity metric

## 1. Introduction

Lung cancer has become one of the most fatal malignant cancers in the world.[1] Early diagnosis could improve the chances of recovery dramatically. Currently, it has been proven that deep analysis of radiographic images can inform and quantify the microenvironment and the extent of intra-tumoral heterogeneity for personalized medicine.[2] Computed tomography (CT) is the best means of screening for lung cancer. Therefore, CT-based image analysis of lung cancer plays a crucial role in computer-assisted diagnosis (CAD).

In general, the challenges of CAD mainly include feature extraction and diagnostic discrimination. In feature extraction, current researches mainly focus on designing new features or feature selection to improve the description and differentiation of images,[3,4] such as morphological and texture features,[5–7] shape features,[8] feature selection,[9] radiomics features,[10,11] deep learning features.[3,12] However, most of them (excepting deep features) suffer from the intra-class variation and inter-class ambiguity problem. Deep features will encounter feature fusion problems with other features. For diagnostic discrimination, a number of classical classifiers are selected for diagnosis, such as support vector machine (SVM),[6,13] Random Forest (RF),[14] convolutional neural network (CNN).[15,16] However, each classifier has a suitable object.

As one of the CAD methods, content-based image retrieval (CBIR) can not only help doctors diagnose tumors benign or malignant but give a selection of similar annotated cases for doctors' reference. The advantage is to help doctors make a diagnosis with reference to the existing similar case diagnosis. It is

*Editor: Michael Masoomi.*

*Compliance with Ethical Standards*

*This study was funded by the National Natural Science Foundation of China (No. 81973981), the Science Foundation of Shandong University of Traditional Chinese Medicine (No. 2018zk02), Shandong Medical Health Technology Development Plan (No. 2018WS206), and Medical Education Research Project of Chinese Medical Association (No. 2018B-N02156).*

*The authors declare no conflict of interests.*

*This article does not contain any studies with human participants performed by any of the authors.*

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

*[a] School of Science and Engineering, Shandong University of Traditional Chinese Medicine, [b] Affiliated Hospital of Jining Medical University, [c] Shandong Provincial Key Laboratory for Distributed Computer Software Novel Technology, Jinan, China.*

*[*] Correspondence: Guohui Wei, Hui Cao, School of Science and Engineering, Shandong University of Traditional Chinese medicine, Jinan 250355, China (e-mails: bmie530@163.com, caohui63@163.com).*



*How to cite this article: Wei G, Qiu M, Zhang K, Li M, Wei D, Li Y, Liu P, Cao H, Xing M, Yang F. A multi-feature image retrieval scheme for pulmonary nodule diagnosis. Medicine 2020;99:4(e18724).*

*Received: 23 September 2019 / Received in final form: 5 December 2019 / Accepted: 12 December 2019*

*http://dx.doi.org/10.1097/MD.0000000000018724*

useful for medical research, CAD, radiotherapy and evaluations of surgery outcome as well. In CBIR field of breast lesions, researchers have done a lot of exploration.[17–20] For lung lesions, Ma et al[21] proposed a CBIR method to retrieve CT imaging signs. However, few of researches concentrated on pulmonary nodule classification. Our group is engaged in the research of medical image retrieval for pulmonary nodule diagnosis.[22–24]

In CBIR, all medical images can be represented as vector collection. As mentioned above, this is similar to the feature extraction in CAD. Therefore, it is important to extract appropriate features to represent medical images. Recent research[3] indicated multiple types of features can better represent pulmonary nodules and achieve higher classification accuracy. However, multi-feature fusion is a problem that needs to be solved because unifying multiple features into one vector is not optimal. Besides the multi-feature fusion problem, similarity measurement of tumor images is another critical issue. During retrieval process, the query image's features are then compared with the features of indexed images using a defined similarity measurement algorithm. The measurements can rank the images in order of the similarity. The similarity measurement usually requires learning a distance metric. Recently, distance metric learning has attracted the attention of researchers. However, the traditional distance metric learning is based on the hypotheses that data is represented by a single feature vector. It is incapable of multiple features. Due to multiple features usually have different physical properties, straightforwardly unifying multiple features to a long feature vector is not optimal. Since this would lead to curse-of- dimensionality and over-fitting problems. Semisupervised multiview distance metric learning (SSM-DML) algorithm proposed by Ref. 25 learns a muiltiview distance metric from multiple features sets to measure the similarity between cartoon data, which is under the umbrella of graph-based semisupervised learning. However, SSM-DML is graph-based and it simply calculates the distance metric between image features without considering the semantic relevance, which are learned from the labeled data. Furthermore, this algorithm is proposed for cartoon data, it is not necessarily suitable for medical tumor images.

In this paper, we propose a novel content-based multi-feature image retrieval (CBMFIR) scheme for computer-aided diagnosis of pulmonary nodules. This scheme considers developing a distance metric learning method named Multi-feature Distance Metric Learning to measure the similarity of pulmonary nodules. This new method explores multi-feature fusion problem with an integrating optimal algorithm. The learned distance metric measures the similarity of pulmonary nodules based on the semantic relevance.[26] Based on learned distance metric, we develop a novel CBIR scheme to help doctors search for similar cases and differentiate benign from malignant pulmonary nodules.

## 2. Materials and methods

### 2.1. Image dataset

For developing and testing a new CBMFIR scheme, a reference pulmonary nodule image dataset was assembled from the public available LIDC-IDRI lung CT scan images, which contained 1018 independent examination cases. In the assembled pulmonary nodule dataset, 746 nodule ROIs were extracted, in which 375 nodules were experts-identified malignant and 371 nodules were experts-identified benign. After obtaining $N = 746$ nodules $X = [x_1, \ldots, x_N]$, we extracted Haralick textures (Denote as Data1) and density related features (Denote as Data2) to represent pulmonary nodules. The Haralick texture features are connected to a 26-dimensional vector, while density related features are unified into a 2-dimensional vector. Detailed research process can be referred to our published papers.[22–23]

### 2.2. Content-based multi-feature image retrieval scheme

**2.2.1. Overview of distance metric learning.** Research in distance metric learning[27] is driven by the need to find meaningful low-dimensional manifolds that capture the intrinsic structure of high-dimensional data. In this section, we present a novel distance metric learning algorithm.

Denote the sample dataset as $X = [x_1, \ldots, x_n] \in \Re^{d*n}$, with $x_i \in \Re^d$ being the $i$th sample in the input space and $n$ being the total number of samples. For better presentation, we also denote a distance metric $d_M(x_i, x_j)$ as a Mahalanobis distance between $x_i$ and $x_j$, which is defined as:

$$d_M(x_i, x_j) = \sqrt{(x_i - x_j)^T M (x_i - x_j)} \tag{1}$$

In Eq. (1), $^T$ denotes the transpose of a vector or a matrix, $M$ is a positive semi-definite matrix. If $M = I$, $d_M(x_i, x_j)$ corresponds to Euclidean distance. If $M$ is restricted to be a diagonal matrix, $d_M(x_i, x_j)$ represents a distance metric in which the different axes are given different weights. More generally, $M$ represents a set of Mahalanobis distance. Because $M$ is a positive semi-definite matrix, it can be decomposed into $M = AA^T$. Hence, Eq. (1) can be rewritten as:

$$d_M(x_i, x_j) = \sqrt{(x_i - x_j)^T A A^T (x_i - x_j)} = ||A^T (x_i - x_j)|| \tag{2}$$

Therefore, learning such a distance metric is actually equivalent to finding a transformation of Euclidean distance between samples in the original high-dimensional space. During recent years, a variety of techniques[27] have been proposed to learn such an optimal Mahalanobis distance metric $d_M(x_i, x_j)$ from training data that are given in the form of side information. We want to obtain $A$ from the semantic relevance.

**2.2.2. Similarity metric.** We define similarity measures as semantic relevance.[26] Semantic relevance can be presented by side information, which means that if 2 nodules have same labels, they are semantic relevance. Therefore, we study transformation matrix $A$ according to semantic relevance.

For semantic relevance, it describes the class separability, which requires the separability measure increase when the size of the between-class scatter matrix increases or the size of the within-class scatter matrix is smaller. This can be described by the Differential Scatter Discriminant Criterion (DSDC) model,[28] it is defined as:

$$A = arg \max_{A^T A = I} \left( tr\left(A^T S_B A\right) - \rho tr\left(A^T S_W A\right)\right) \tag{3}$$

The variation is defined as:

$$\begin{aligned} A &= arg \min_{A^T A = I} \left( tr\left(A^T S_W A\right) - \rho tr\left(A^T S_B A\right)\right) \\ &= arg \min_{A^T A = I} \left( tr\left(A^T (S_W - \rho S_B) A\right)\right) \end{aligned} \tag{4}$$

In (4), $S_W$ is the within-class scatter matrix, $S_B$ is the between-class scatter matrix. $\rho$ is a nonnegative tuning parameter, which balances the relative merits of minimizing the within-class scatter

to the maximization of the between-class scatter. The learned matrix $A$ is the transformation matrix. With matrix $A$, we can calculate Mahalanobis distance between nodule images.

Define $L = S_W - \rho S_B$, Eq. (4) can be rewritten as:

$$A = arg \min_{A^T A = I} tr(A^T L A) \tag{5}$$

***2.2.3. Multi-feature distance metric learning.*** Multiple types of features usually have different physical properties. Therefore, it is not optimal for straightforwardly concatenating multiple features into a long feature vector. This would cause over-fitting and curse-of-dimensionality problems. Especially, if the number of samples is not large enough, it is difficult to learn a robust distance metric in a high-dimensional feature space.

In this section, we extend single feature similarity metric to multi-feature spaces. We apply multiple types of features to learn multiple transformation matrices to construct multi-feature distance metric. We linearly combine the similarity metrics constructed from multiple feature sets through the weights $\alpha_i$ and add a regularizer to the weights. Thus, the objective function is as follows:

$$\begin{aligned} &\phi(\alpha, A^{(1)}, A^{(2)}, \ldots, A(k)) = \sum_{k=1}^{K} \alpha_k tr(A^{(k)^T} L A^{(k)}) + \lambda ||\alpha||^2 \\ &s.t. \sum_{k=1}^{K} \alpha_k = 1 \end{aligned} \tag{6}$$

where $A^{(k)}$ is the $k$ transformation matrix learning from the $k$ feature set, $\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_k]^T$

Therefore, the objective function (5) is proposed to learn a distance metric for each feature set, while the objective function (6) is constructed to integrate the information of the feature sets with the combination weights. This strategy reduces the model complexity and alleviates the over-fitting problem.

To solve the objective function (6), firstly, transformation matrices $A^{(k)}|_{k=1}^{K}$ should be given according to Eq. (5). Afterwards, a lagrangian multiplier method is employed to obtain the optimum solution. With the Lagrange multiplier $\eta$, the objective function turns to:

$$L(\alpha, \eta) = \sum_{k=1}^{K} \alpha_k tr(A^{(k)T} L A^{(k)}) + \lambda ||\alpha||^2 - \eta(\sum_{k=1}^{K} \alpha_k - 1) \tag{7}$$

By setting the partial derivatives of $L(\alpha, \eta)$ with respect to $\alpha$ and $\eta$ to be zeros, we get:

$$\begin{cases} \dfrac{\partial L}{\partial \alpha_1} = tr(A^{(1)T} L A^{(1)}) + 2\lambda\alpha_1 - \eta = 0 \\ \vdots \\ \dfrac{\partial L}{\partial \alpha_K} = tr(A^{(K)T} L A^{(K)}) + 2\lambda\alpha_K - \eta = 0 \\ \dfrac{\partial L}{\partial \eta} = \sum_{k=1}^{K} \alpha_k - 1 = 0 \end{cases} \tag{8}$$

Combining the above equations, we get:

$$\sum_{k=1}^{K} tr(A^{(k)T} L A^{(k)}) + 2\lambda \sum_{k=1}^{K} \alpha_k - \eta K = 0 \tag{9}$$

Since $\sum_{k=1}^{K} \alpha_k = 1$, we can obtain:

$$\eta = \frac{\sum_{k=1}^{K} tr(A^{(k)T} L A^{(k)}) + 2\lambda}{K} \tag{10}$$

Putting this equation into (9), we can obtain:

$$\begin{aligned} \alpha_k &= \frac{\eta - tr(A^{(K)T} L A^{(K)})}{2\lambda} \\ &= \frac{2\lambda + \sum_{k=1}^{K} tr(A^{(k)T} L A^{(k)}) - K tr(A^{(k)T} L A^{(k)})}{2\lambda K} \end{aligned} \tag{11}$$

By multi-feature distance metric learning, we can obtain the Mahalanobis distance $d^{(k)}(x_i, x_j)$ and weight value$\alpha_k$ corresponding to the $k$ feature set. Therefore, the multi-feature Mahalanobis distance between sample $x_i$ and $x_j$ can be calculated as:

$$d(x_i, x_j) = \sum_{k=1}^{K} \alpha_k d^{(k)}(x_i, x_j). \tag{12}$$

***2.2.4. Multi-feature image retrieval scheme.***

**Retrieval Scheme**

..Given a sample set $X = [x_1, x_2, \ldots, x_n] \in \Re^{d*n}$, and the number of classes c.

1. Solve (5) with eigenvalue decomposition and compute the transformation matrix $A^{(k)}|_{k=1}^{K}$ corresponding to the $k$ image feature set.
2. Calculate weight value $\alpha_k$ according to (11).
3. Compute the multi-feature Mahalanobis distance $d(x_i, x_j)$ between sample $x_i$ and $x_j$ according to (12).
4. With the multi-feature Mahalanobis distance $d(x_i, x_j)$, sort the distances we obtained, the retrieval inclusion is the smallest ones.
5. For diagnosis, a classification likelihood value of the queried nodule is computed to measure the malignancy of a nodule.

***2.2.5. CBMFIR scheme for pulmonary nodule diagnosis.*** With the obtained multi-feature Mahalanobis distance, we propose a CBMFIR scheme to assist doctors in diagnosing pulmonary nodules. CBMFIR-based pulmonary nodule diagnosis mainly includes 2 parts: (1) Retrieval example reference; (2) Computer-aided diagnosis.

1. Retrieval example reference. Image retrieval can retrieve many images similar to the query image. The doctor can refer to the diagnostic experience of the retrieved similar tumor images before diagnosing pulmonary nodule benign or malignant or determining whether a biopsy is necessary.
2. Computer-aided diagnosis. According to the retrieval examples, a malignant likelihood value of the query nodule can be calculated to measure the malignancy of this nodule. The formula is as follows ($K$ is the number of retrieval examples, $M$ is the number of malignant nodule):

$$P_q = \frac{M}{K} \tag{13}$$

Giving a threshold of $P_q$ (such as $P_T = 0.5$), if $P_q \geq P_T$, we conclude that the query nodule is malignant, otherwise, it is benign.

### 2.3. Performance assessment

In our real experiments, we randomly selected 400 nodule images from the pulmonary nodule dataset to serve as the training set, in which approximately 200 nodules were benign and about 200 nodules were malignant. The remaining 346 pulmonary nodule images were used as the testing dataset. All the experiment evaluations were run in Windows 7, MATLAB R2014a, Intel Core(TM) i5–5200U CPU and 4GB RAM.

To demonstrate the feasibility of the proposed scheme (CBMFIR) for diagnosis of pulmonary nodule lesions, extensive experiments are performed to analyze the diagnostic performance of CBMFIR algorithm in 2 settings, classification accuracy and retrieval accuracy. Therefore, we compare our proposed diagnostic scheme to several existing metric methods, including Information-Theoretic Metric Learning (ITML),[29] Large Margin Nearest Neighbor (LMNN),[28] SSM-DML,[25] Kernel based Differential Scatter and Patch Alignment Distance Metric (KDPDM)[22] and other diagnosis algorithms SVM, ELM. The experiments of performance assessment are performed in the context of pulmonary nodule dataset. We firstly introduce the parameter effects of our proposed scheme, and then compare the performance of our scheme with that of existing algorithm, including the pulmonary nodule classification accuracy and retrieval accuracy. Finally, a retrieval example is given to illustrate the feasibility of the proposed retrieval scheme.

**2.3.1. Parameter configurations.** In this subsection, the effects of several parameters are analyzed. The investigation of parameter configurations is performed based on an image retrieval task and the assembled pulmonary nodule dataset. In the experiments, some factors are configured, the tradeoff parameter $\rho$ in Eq. (4), $\lambda$ in Eq. (6). In these experiments, every experiment was repeated for 10 times with different randomly training nodules. We calculated the average performance over 10 rounds of experiments.

In our experiments, ROC curve can be drawn with varying the threshold of the malignant probability in (13). Thus the area under the curve (AUC) is used to analyze the effects of the parameters. A larger AUC value indicates a better classification performance. The AUC value calculated in the figures is a mean value of 10 experimental results.

**2.3.2. Diagnosis performance assessment.** We compare our scheme to 4 state-of-the-art algorithms for learning distance functions and distance metrics: ITML, LMNN, SSM-DML, and KDPDM. We selected KDPDM with the orthogonal case. Euclidean distance is included as comparative references.

The diagnosis performance of CBMFIR is evaluated with 2 metrics: classification accuracy and retrieval accuracy.[22] Classification accuracy means the extent to which malignant nodules can be detected on the basis of the nodule image that are retrieved. We firstly select $K$ nearest neighbor nodules with the learned Mahalanobis distance metric, and then calculate the probability of the query sample belonging to malignant nodule. With the obtained probabilities for query nodules, ROC curve can be drawn with varying the threshold of the malignant probability. Thus the AUC value from the ROC curve is used to evaluate the classification accuracy.

The second metric, retrieval accuracy, reflects the proportion of retrieval nodules that are semantic relevant (ie, in the same semantic class) to the query nodule. Retrieval accuracy is calculated by the leave-one-nodule-out method in the test dataset. The result of retrieval accuracy can be depicted by a performance curve, each value is a function of the number of retrieved nodules. According to leave-one -nodule-out manner, in the test dataset, one nodule is used as the query image, the rest of the nodules are the retrieval dataset. We calculate the distance metrics between the query nodule and the rest retrieval ones, then rank the Mahalanobis distances in ascending order. The formula of retrieval accuracy is constructed as follows:

$$r(q_i^k) = \frac{\sum_{j=1}^{k} \delta(y_i = y_j)}{k} \tag{14}$$

$r(q_i^k)$ is the proportion of nodules identical to the query nodule label in the first $k$ ranked nodules. $y_i$ is the $i$th query nodule.

## 3. Results

### 3.1. Parameter configurations

In Eq. (4), the effects of the tradeoff parameter $\rho$ is investigated. We vary $\rho$ with [10–8,10–6,10–4,10–2,10–1,1,101,102,104,106,108]. Figure 1 shows the mean classification accuracies and the corresponding standard deviations when parameter $\rho$ varies from $10^{-8}$ to $10^{-8}$. From this figure, we can conclude that the proposed CBMFIR scheme is sensitive to $\rho$. The performance curve has a fluctuation when $\rho = 1$. When $\rho > 1$, the performance of this scheme will drop to 0.8. This illustrates that our scheme is not suitable for a large parameter $\rho$. In this experiment, we fixed the tradeoff parameter $\lambda = 1$, the number of nodules retrieved each time is fixed at 15.

We then analyze the effect of parameter $\lambda$ in Eq. (7). For fair comparison, we set parameter $\lambda$ within the range [10–3,10–2,10–1,1,10,102,103]. Figure 2 reports the accuracy curve with respect to $\lambda$. It can be seen that the performance of our scheme is relatively stable when $\lambda \geq 1$. This demonstrates that our scheme prefers larger parameter $\lambda$, the number of nodules retrieved each time is fixed at 15.

### 3.2. Diagnosis performance assessment

Classification accuracy and retrieval accuracy are used to evaluate the diagnosis performance of the proposed CBMFIR scheme. For classification accuracy, we firstly compare the classification performance with different features:

1. the combined features with our scheme (CBMFIR);
2. the straightforwardly concatenating multiple features D1 and D2 into a long feature vector (D1D2);
3. the concatenating D2 and D1 into a vector (D2D1), as reported in Table 1.

According to the results of the comparison, our scheme has a better classification accuracy than that of the other features, which demonstrates that straightforwardly unifying multiple features to a long feature vector is not optimal. We then compare the classification accuracy of our scheme with that of the state-of-the-art distance metric learning algorithms. Table 2 shows AUC values for CBMFIR and the baseline methods. Euclidean distance metric has the worst classification accuracy. The proposed CBMFIR has the better classification accuracy than that of other comparison algorithms. Finally, we analyze the diagnostic

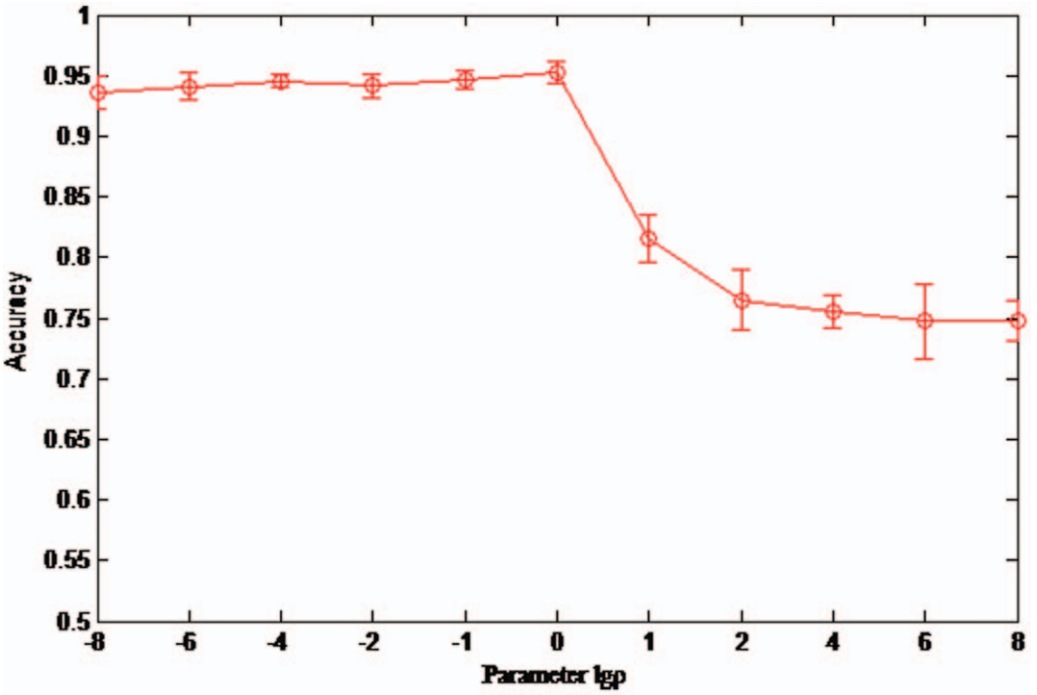


**Figure 1.** The accuracy with different $\rho$.

performance of CBMFIR with some classical diagnostic algorithms. The feature set of the classical diagnostic algorithms is a feature vector through concatenating multiple features D1 and D2. Table 3 reports the comparison results. It can be concluded that CBMFIR performs best in pulmonary nodule diagnosis.

For retrieval accuracy, we compare the retrieval performance between CBMFIR with other state-of-the-art distance metric

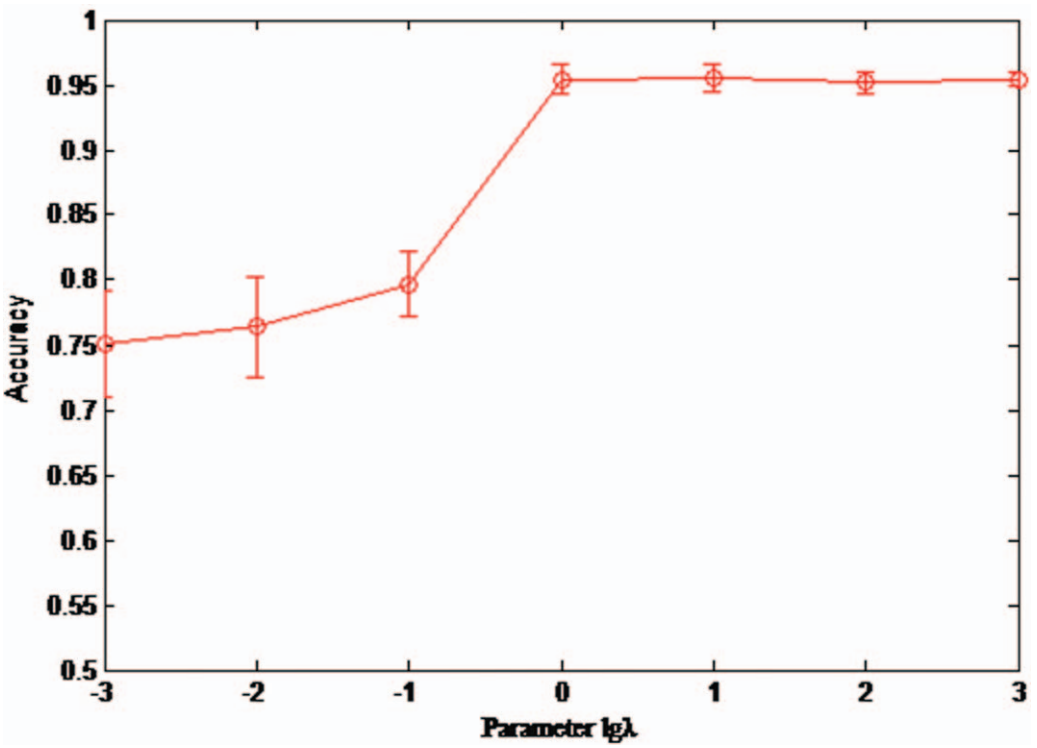


**Figure 2.** The accuracy with different $\lambda$.

**Table 1**

**Comparison of the classification accuracy with different features.**

| Features | AUC (mean ± std) |
|---|---|
| D1D2 | 0.942 ± 0.010 |
| D2D1 | 0.945 ± 0.009 |
| CBMFIR | **0.955 ± 0.010** |

**Table 2**

**Comparison of the classification accuracy of distance metric learning algorithms.**

| Algorithms | AUC (mean ± std) |
|---|---|
| Euclidean | 0.933 ± 0.013 |
| ITML | 0.942 ± 0.013 |
| LMNN | 0.952 ± 0.008 |
| SSM-DML | 0.940 ± 0.006 |
| KDPDM | 0.936 ± 0.008 |
| CBMFIR | **0.955 ± 0.010** |

**Table 3**

**Comparison of the classification accuracy of classical diagnostic algorithms.**

| Algorithms | AUC (mean ± std) |
|---|---|
| SVM | 0.910 ± 0.008 |
| ELM | 0.896 ± 0.013 |
| CBMFIR | **0.955 ± 0.010** |

learning algorithms: ITML, LMNN, KDPDM, and SSM-DML. European distance metric is included as a comparative reference. Figure 3 reports the retrieval accuracy of the comparative algorithms. Among them, CBMFIR performs best in pulmonary nodule retrieval. This indicates that multi-feature used in CBMFIR can effectively improve the retrieval accuracy. The retrieval performance of other algorithms is slightly weaker, especially when rank ≥15, the performance of ITML is significantly reduced. The retrieval performance of Euclidean distance is not bad. One possible reason is that the dimension of feature extraction is not large.

### *3.3. Retrieval examples*

Figure 4 reports four retrieval examples returned by CBMFIR. In Figure 4A and C, the data set being retrieved is the training dataset; in Figure 4A and D, the data set being queried is the testing dataset. According to Figure 4, perfect search results will arrange nodules in order of increasing Mahalanobis distance metrics. Based on the diagnostic information of the retrieval results, doctors can make an evaluation of the query nodule and decide if a pathological examination is needed.

## 4. Discussion

In this work, we propose and demonstrate the feasibility of developing a multi-feature image retrieval scheme for pulmonary nodule diagnosis. A multi-feature distance metric learning algorithm is proposed to measure the similarity of pulmonary nodules. This study has many unique features and experimental observations. First, a CBIR scheme is used to help doctors evaluate pulmonary nodules benign or malignant before pathological experiments. A retrieval set with diagnostic reports is provided to doctors for reference.

Second, multiple types of features (texture features and density related features) are used to represent pulmonary nodules. Texture features are the computer's point of view to identify pulmonary nodules. Density related features are the doctors'

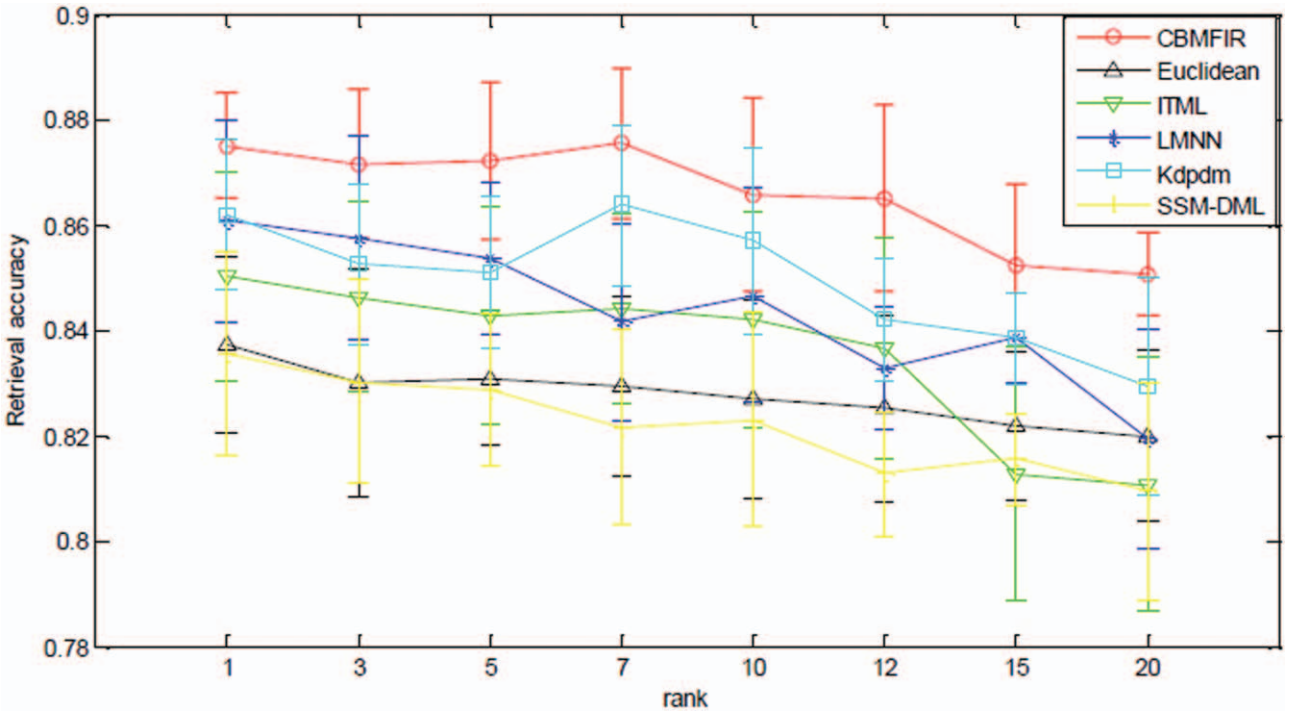


**Figure 3.** Retrieval accuracy of distance metric algorithms. 'rank' is the number of retrieved nodules.

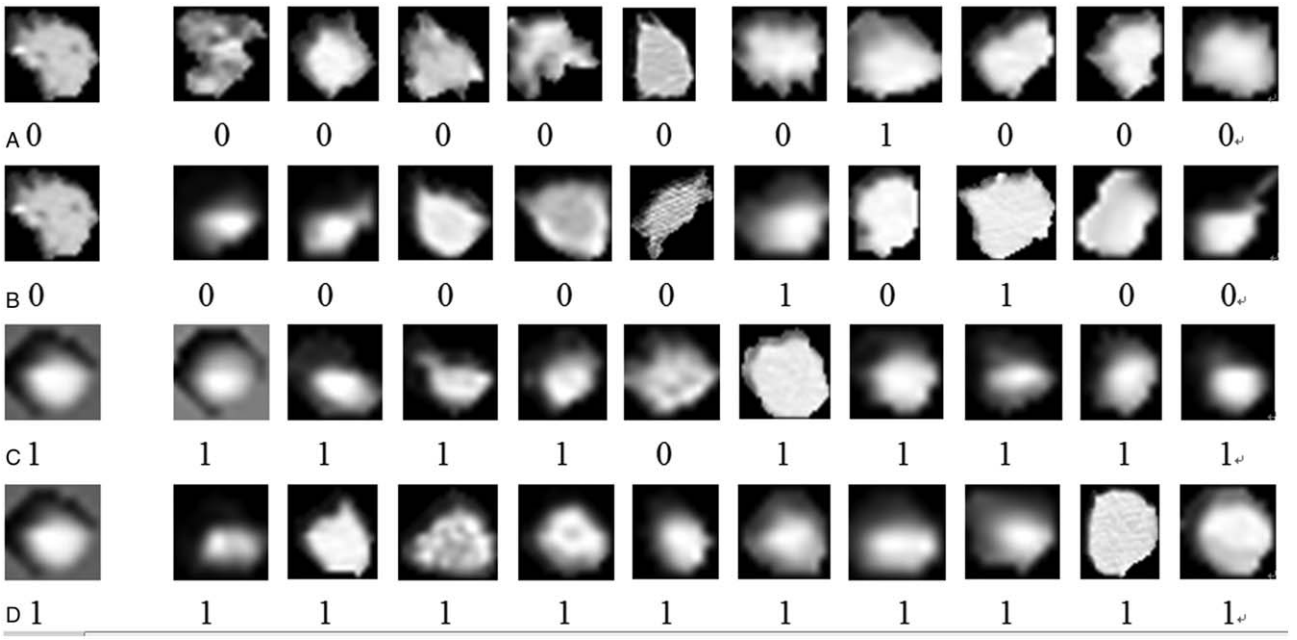


**Figure 4.** The query nodule (left) and their top 10 retrieval nodule set. For each nodule, its class is listed below the nodule. "1" indicates that the nodule is benign and "0" represents that the nodule is malignant. All query nodules are correctly identified based on a weighted majority vote of the retrieved reference nodule sets.

viewpoint to discriminate pulmonary nodules. They are complementary to each other.

Third, multiple types of features can better differentiate pulmonary nodules benign or malignant. A multi-feature fusion problem is investigated to propose a multi-feature distance metric learning algorithm for nodule similarity measurement. Our algorithm combines different types of features to avoid curse-of-dimensionality and over-fitting problems.

In addition to the promising results discussed above, this work also has some limitations. First, the nodule set only has 746 pulmonary nodules. A larger nodule set will be assembled in the future work. Second, this study only investigates texture features and density related features. However, there are many other types of features, which can be studied to represent pulmonary nodules. Therefore, in the subsequent work, fusion with other complement types of features would be explored to improve the diagnosis accuracy. Third, it is not enough to simply analyze image features in cancer research. Comprehensive genetic data and pathological reports of cancer diagnosis methods need to be studied in the future. Fourth, there are many empirically determined parameters in the proposed algorithm. However, we only use a search method to choose best parameters in the systematic experiments. Consequently, a more adaptively optimization method is investigated to study the optimal parameters.

## 5. Conclusions

In this paper, we investigate the feasibility of developing a multi-feature distance metric to measure the similarity of the query nodule and pulmonary nodule dataset for pulmonary nodule medical image retrieval. This multi-feature distance metric could combine multiple types of features of pulmonary nodules. The proposed retrieval scheme provides a reference for doctor's diagnosis. Experimental evaluations based on the proposed scheme suggest the effectiveness in the diagnosis of pulmonary nodules.

## Author contributions

GW, MQ and ZW conceived and designed the project, DW, PL revised the manuscript. KZ, FY and MX analyzed the lung nodule data set. YL and ML collected data and provided expert knowledge. All authors edited the manuscript.

Guohui Wei orcid: 0000-0002-2585-282X.

## References

[1] Siegel RL, Miller KD, Jemal A. Cancer statistics, 2018. Ca-Cancer J Clin 2018;68:7–30.
[2] Agarwal R, Shankhadhar A, Sagar RK, Detection of lung cancer using content based medical image retrieval. 2015 Fifth International Conference on Advanced Computing & Communication Technologies. IEEE, 2015.
[3] Xie Y, Zhang J, Xia Y, et al. Fusing texture, shape and deep model-learned information at decision level for automated classification of lung nodules on chest CT. Inform Fusion 2017;42:102–10.
[4] Kawagishi M, Chen B, Furukawa D, et al. A study of computer-aided diagnosis for pulmonary nodule: comparison between classification accuracies using calculated image features and imaging findings annotated by radiologists. Int J Comput Ass Rad 2017;12:1–0.
[5] Way TW, Sahiner B, Chan HP, et al. Computer-aided diagnosis of pulmonary nodules on CT scans: improvement of classification performance with nodule surface features. Med Phys 2009;36:3086–98.
[6] Han F, Wang H, Zhang G, et al. Texture feature analysis for computer-aided diagnosis on pulmonary nodules. J Digit Imaging 2015;28: 99–115.
[7] Orozco HM, Villegas O, Sánchez V, et al. Automated system for lung nodules classification based on wavelet feature descriptor and support vector machine. Biomed Eng Online 2015;14:9.
[8] Filho A, Silva A, Paiva A, et al. Computer-aided diagnosis system for lung nodules based on computed tomography using shape analysis, a genetic algorithm, and SVM. Med Biol Eng Comput 2016;55:1129–46.

[9] Lu C, Zhu Z, Gu X. An intelligent system for lung cancer diagnosis using a new genetic algorithm based feature selection method. J Med Syst 2014;38:97–105.
[10] Tian J, Dong D, Liu Z, et al. Radiomics in medical imaging-detection, extraction and segmentation. Artificial Intelligence in Decision Support Systems for Diagnosis in Medical Imaging. 1 2018;267–333.
[11] Ma J, Wang Q, Ren Y, et al. Automatic lung nodule classification with radiomics approach. Spie Med Imaging 2016;978906. https://doi.org/10.1117/12.2220768.
[12] Wang C, Elazab A, Wu J, et al. Lung nodule classification using deep feature fusion in chest radiography. Comput Med Imag Grap 2017; 57:10–8.
[13] Froz B, Filho A, Silva A, et al. Lung nodule classification using artificial crawlers, directional texture and support vector machine. Expert Syst Appl 2017;69:176–88.
[14] Huang P, Park S, Yan R, et al. Added value of computer-aided CT image features for early lung cancer diagnosis with small pulmonary nodules: a matched case-control study. Radiology 2017;286:286–95.
[15] Shen W, Zhou M, Yang F, et al. Multi-crop convolutional neural networks for lung nodule malignancy suspiciousness classification. Pattern Recogn 2017;61:663–73.
[16] Liu X, Hou F, Qin H, et al. Multi-view multi-scale CNNs for lung nodule type classification from CT images. Pattern Recogn 2018; 77:262–75.
[17] Gundreddy RR, Tan M, Qiu Y, et al. Assessment of performance and reproducibility of applying a content-based image retrieval scheme for classification of breast lesions. Med Phys 2015;42:4241–9.
[18] Jiang M, Zhang S, Li H, et al. Computer-aided diagnosis of mammographic masses using scalable image retrieval. IEEE T Biomed Eng 2015;62:783–91.
[19] Tsochatzidis L, Zagoris K, Arikidis N, et al. Computer-aided diagnosis of mammographic masses based on a supervised content based image retrieval approach. Pattern Recogn 2017;71:106–17.
[20] Dubey SR, Singh SK, Singh RK, et al. Local wavelet pattern: a new feature descriptor for image retrieval in medical CT databases. IEEE T Image Process 2015;24:5892–903.
[21] Ma L, Liu X, Gao Y, et al. A new method of content based medical image retrieval and its applications to CT imaging sign retrieval. J Biomed Inform 2017;66:148–58.
[22] Wei G, Ma H, Qian W, et al. Similarity measurement of lung masses for medical image retrieval using kernel based semisupervised distance metric. Med Phys 2016;43:6259–69.
[23] Wei G, Ma H, Qian W, et al. A content-based image retrieval scheme for lung nodule classification. Curr Med Imaging Rev 2017;13:210–6.
[24] Wei G, Cao H, Ma H, et al. Content-based image retrieval for lung nodule classification using texture features and learned distance metric. J Med Syst 2018;42:13.
[25] Yu J, Wang M, Tao D. Semisupervised multiview distance metric learning for cartoon synthesis. IEEE T Image Process 2012;21:4636–48.
[26] Yang L, Jin R, Mummert L, et al. A boosting framework for visuality-preserving distance metric learning and its application to medical image retrieval. IEEE Trans Pattern Anal Mach Intell 2010;32:30–44.
[27] Weinberger KQ, Blitzer J, Saul LK. Distance metric learning for large margin nearest neighbor classification. J Mach Learn Res 2009;10:207–44.
[28] Tao D, Li X, Wu X, et al. General tensor discriminant analysis and gabor features for gait recognition. IEEE Trans Pattern Anal Mach Intell 2007;29:1700–15.
[29] Davis JV, Kulis B, Jain PP, et al. Information theoretic metric learning. *The International Conference on Machine Learning* (ACM Press, Corvallis, OR, 2007), 209–16, 2007.