Image Grayscale Density Distribution Calculation Model and Posterior for Lung Nodule Benign-Malignant Classification

Authors: VanbangLe, Zhu Yu, Zheng Bingbing, Yang Dawei, REN Xiaodong, Thiminhchinh Ngo

Date: 2018-11-29T00:00:00+00:00

Abstract

Computer-aided diagnosis of lung cancer is of significant importance for early detection and timely treatment. This study proposes a feature evaluation algorithm based on density distribution and introduces a pattern recognition model to evaluate the effectiveness of the proposed method. First, pixel patch sets are randomly extracted from lung tumor images and partitioned into 10 clusters via the K-means clustering algorithm. Based on the relationship between pixel values of lung nodules in CT images and cluster centers, a 10-dimensional feature vector is extracted, and a random forest classifier is employed for model training to classify the malignancy level of lung nodules. Experiments on the publicly available CT image dataset LIDC-IDRI demonstrate that the average classification accuracy reaches 0.9008. Comparative analysis of experimental results indicates that the proposed feature representation method achieves superior classification performance and higher robustness.

Full Text

Pulmonary Nodule Image Grey Density Distribution Feature Extraction Algorithm and Adenocarcinoma Benign/Malignant Classification

Vanbang Le[1], Zhu Yu[1]†, Zheng Bingbing[1], Yang Dawei[2], Ren Xiaodong[1], Thiminhchinh Ngo[3]

[1]School of Information Science & Engineering, East China University of Science & Technology, Shanghai 200237, China
[2]Zhongshan Hospital, Fudan University, Shanghai 200032, China
[3]Hitech Telecommunication Center, Hanoi, Vietnam

Abstract: Computer-aided lung cancer diagnosis plays a crucial role in early detection and treatment. This paper proposes a density distribution-based feature evaluation algorithm combined with pattern recognition models to assess classification efficacy. First, pixel block sets are randomly extracted from lung tumor images and clustered into 10 categories using the K-means algorithm. Based on the relationship between pulmonary nodule pixel values and cluster centers in CT images, a 10-dimensional feature vector is extracted and used to train a random forest classifier for determining benign/malignant levels. Experiments on the public CT image dataset LIDC-IDRI demonstrate an average classification accuracy of 0.9008. Comparative analysis shows that the proposed feature representation method achieves superior classification performance and robustness.

Key words: lung nodule classification; density distribution feature; K-means

0 Introduction

Lung cancer has become the leading cause of cancer mortality worldwide according to surveys from major cancer research centers and health organizations. With the expanding application of thoracic scanning technology, analyzing CT (computed tomography) images using digital image processing techniques has emerged as a prominent research direction. Computer-aided diagnosis systems for lung CT images primarily encompass pulmonary nodule detection, segmentation, and classification. Improving the diagnostic and recognition capabilities for small pulmonary adenocarcinoma nodules (lesion diameter < 30mm) can significantly enhance lung cancer diagnostic accuracy and provide clinicians with more precise diagnostic recommendations, making it a critical research topic in the image processing field.

Segmentation represents one of the most important steps in computer-aided diagnosis systems and substantially impacts benign/malignant classification of pulmonary nodules. Common segmentation methods include grayscale thresholding, graph cuts, level sets, and deep learning. Primary algorithm validation datasets include LIDC-IDRI, ELCAP, and NLST. Segmented nodule images enable assessment of subsequent growth trends and benign/malignant lesion levels to a certain extent.

Benign/malignant classification of pulmonary nodules aims to provide doctors with scientific, reliable auxiliary classification results, making the diagnostic process more accurate while effectively reducing workload. Classification relies on image feature extraction methods combined with classifiers for training and testing. Common classifiers include SVM, KNN, and random forest. Clinically, pulmonary nodules can be categorized from CT value distribution perspectives as ground-glass, semi-solid, and solid types, while risk levels can be classified as benign or malignant.

Han et al. studied the LIDC database, extracting 2D/3D texture (Harralick texture features) and geometric features (circularity, bounding rectangle solidity, etc.) to classify nodules as benign/malignant, achieving a maximum ROC index of 92.7%. Dhara classified LIDC-IDRI samples using geometric and Harralick texture features, reaching an optimal AUC of 0.9505. Reeves from Cornell University used 46-dimensional spatial features for benign/malignant classification on PLIB and NLST databases, achieving 70% accuracy under optimal parameters. The Mayo Clinic' s physiology and biomedical team introduced the CANARY system (computer-aided nodule assessment and risk yield), performing density cluster analysis on NLST with five-year patient follow-up to propose computer-aided classification and risk prediction. Maldonado proposed a pulmonary nodule image density distribution calculation method for CANARY' s classification module, characterizing HU (Hounsfield unit) value distribution as a highly informative feature. Pei Bo from Taiyuan University of Technology used a fuzzy support vector machine based on bidirectional membership functions, integrating grayscale, texture, and shape features to achieve 83% recognition accuracy with 10% misdiagnosis rate.

To improve benign/malignant classification performance, this paper proposes a grayscale density distribution feature extraction model for pulmonary nodule images based on image block sets. The approach first obtains a block set from pulmonary nodule images, calculates the dataset' s autocorrelation matrix, and applies unsupervised clustering to generate labels for image blocks. By finding the best-matching block for each target test pixel, the density distribution level is computed. Finally, grayscale density distribution features are statistically generated and combined with a random forest classifier for dataset classification.

1.1 Experimental Materials

The LIDC-IDRI (Lung Image Database Consortium and Image Database Resource Initiative) public CT database serves as a large-scale, commonly used pulmonary CT database. Downloaded from The Cancer Imaging Archive (TCIA), nodule edge coordinates and features can be extracted from accompanying *.XML files. In LIDC-IDRI, all lung CT images are 512×512 (in HU values). Key parameters include: Slice Thickness (slice thickness in millimeters) and Pixel Spacing (physical distance between pixel centers in millimeters). Precise nodule edge coordinates and feature annotations are typically provided by four radiologists, with inevitable inter-observer variability. When extracting nodule regions from XML files, the target annotation with maximum area is selected, i.e., $\text{nodule}_{\text{arcmaxmark}} = \max_{i=1\sim4}(\text{mark}_i)$.

This study selected a subset from the large-scale LIDC-IDRI pulmonary nodule samples, focusing on small nodules with maximum diameters under 20mm, totaling 1,285 samples. Pixel spacing ranged from 0.5 to 0.8mm, slice spacing from 0.6 to 5.0mm, and diameter range from 2.79mm to 15.77mm. The most im-

portant evaluation parameter—malignancy risk level—is divided into five ranks (1~5), with sample counts of 147, 390, 387, 250, and 119 respectively, totaling 1,285 nodules. Using Han' s benign/malignant classification scheme, three prior definition configurations were formed: Configuration 1, 2, and 3 (CF 1, 2, 3). CF1 comprises benign (ranks 1, 2) and malignant (ranks 4, 5) samples; CF2 uses ranks 1, 2, 3 as benign and ranks 4, 5 as malignant; CF3 defines rank 3 as malignant, i.e., ranks 1, 2 as benign and ranks 3, 4, 5 as malignant. Detailed LIDC-IDRI dataset statistics are shown in Table 1 .

For analyzing classification performance across different sizes, nodules were further divided into two subsets: Sub-set 1 and 2 (LIDC-IDRI Sub-set, LS). LS1 contains nodules from ~3mm to ~10mm, while LS2 contains nodules from ~11mm to ~20mm. By statistically analyzing performance differences between LS1 and LS2, this study summarizes how image features affect recognition of different-sized pulmonary nodules.

1.2 Pulmonary Nodule Image Grey Density Distribution Features

In lung CT images, the grayscale distribution in suspicious regions affects nodule localization and classification. Therefore, grayscale density distribution serves as an important criterion for assessing nodule malignancy risk. Grayscale density distribution refers to the relationship between pixel values and surrounding neighboring points, characterizing the intensity and magnitude of grayscale values appearing in any local region. Areas with densely occurring high grayscale values constitute high-density regions, while areas with sparse high grayscale values represent low-density regions.

1.2.1 Image Block Database Acquisition and Autocorrelation Matrix

The pulmonary CT image block database (IBD) represents the most critical component in the proposed feature extraction process. These blocks are randomly extracted from pulmonary nodule images, with their quantity and diversity determining feature representation accuracy. To ensure IBD diversity and balance, a balanced number of blocks are selected from each category across different databases (LIDC-IDRI and ZSDB). Regarding block size, since target nodules range from 3mm to 30mm, blocks cannot be too large or too small. Overly small blocks approach point processing and introduce noise, while overly large blocks cause significant errors for small nodules. Therefore, 5×5, 7×7, or 9×9 block sizes are typically adopted. Block size affects smoothing: larger blocks suit large nodules, while small nodules require smaller blocks. Balanced quantities of blocks are randomly extracted from each nodule image to construct the IBD.

For IBD clustering, pairwise distances between blocks are calculated to generate

distance matrix $h(i,j)$. With N blocks in IBD, $h(i,j)$ forms an $N \times N$ symmetric matrix where each row (or column) represents a distance vector between one block and all others. Common distance metrics include Euclidean (EU), Canberra (CA), Chebyshev (CH), and Bray-Curtis (BR). To maximize distance discriminability, this study uses Canberra distance. For equal-length vectors p and q, the distance formulas are:

$$
\begin{aligned}
d_{CA}(p,q) &= \sum_{i=1}^{n} \frac{|p_i - q_i|}{|p_i| + |q_i|} \\
d_{EU}(p,q) &= \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \\
d_{CH}(p,q) &= \max_i |p_i - q_i| \\
d_{BR}(p,q) &= \frac{\sum_{i=1}^{n} |p_i - q_i|}{\sum_{i=1}^{n} (|p_i| + |q_i|)}
\end{aligned}
$$

Unsupervised clustering categorizes the distance matrix into 10 clusters. Experiments show K-Means clustering yields optimal Silhouette (SIL) values. Clustering results are mapped to IBD to generate the Marked Image Block Database (MIBD), where each block receives a label corresponding to its grayscale density distribution level. During IBD clustering, the clustering result for each row (or column) of the distance matrix simultaneously represents the grayscale density distribution level of the corresponding image block. Figure 1 illustrates image blocks, distance matrices, and clustering statistics for different distance metrics (BR, CA, CH, EU). The visualization shows 1,600 blocks arranged in a $[(40 \times 7) \times (40 \times 7)]$ matrix, with blue and dark orange representing minimum and maximum grayscale values. The corresponding histograms (bins=256) and clustering label statistics demonstrate smooth distributions. To quantitatively evaluate clustering quality, the Silhouette parameter is calculated, ranging from -1 to 1, where larger values indicate better performance. Canberra distance achieves the highest SIL of 0.4310, making it the chosen metric for autocorrelation analysis. MIBD labels are sorted by cluster center values.

1.2.2 Image Block-Based Grayscale Density Distribution Feature Extraction

After IBD clustering, each non-background pixel's grayscale density distribution level is computed by traversing the nodule image. For a target pixel centered in a window matching IBD block size, Euclidean distance identifies the most similar block in IBD:

$$
I_{\text{matched}}(x,y) = \arg \min_{i \in [1,N]} d(I_{\text{test}}(x,y), IBD_i)
$$

where $I_{\text{test}}(x, y)$ is the test window and $I_{\text{matched}}(x, y)$ is the best-matching block. The density distribution level is then:

$$\text{Level}(I_{\text{test}}(x, y)) = \text{Label}(I_{\text{matched}}(x, y))$$

Repeating this for all non-zero pixels yields the nodule' s CT value density distribution image with 10 valid levels (1~10), where density levels replace original pixel values. This density distribution feature serves as the recognition feature for machine learning training. Figure 2 [Figure 2: see original paper] illustrates the feature extraction process: (a) original nodule image, (b) grayscale density distribution image, and (c) feature vector represented as a pie chart. The density distribution image reveals HU value distribution patterns, helping quantitatively estimate solid component locations and sizes to improve clinical classification accuracy.

In the density map, categories generally appear as rings around the center, with the outermost being the minimum value (k=1), making the first feature vector component non-zero. For LIDC-IDRI, low-density proportions decrease stably from rank 1 to rank 5, while high-density proportions increase. The ranking of high distribution level proportions is rank 1 $<$ rank 2 $<$ rank 3 $<$ rank 4 $<$ rank 5. These features show reliable statistical significance with clear, stable differences between classes, substantially improving benign/malignant classification performance. Figure 4 [Figure 4: see original paper] shows mean grayscale density distribution features for LIDC-IDRI samples.

1.3 Random Forest Classifier and Pattern Recognition Evaluation

This study employs Random Forest as the classifier, consisting of multiple decision trees. Evaluation parameters include Sensitivity, Specificity, ROC curve, and Accuracy. Sensitivity (or True Positive Rate, TPR) describes the correct classification rate for "positive" samples:

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

Specificity (or True Negative Rate, TNR) measures correct classification for "negative" samples:

$$\text{Specificity} = \frac{TN}{TN + FP}$$

where TP = true positive, FP = false positive, TN = true negative, FN = false negative.

The ROC curve graphically evaluates binary classification models using TPR and False Positive Rate (FPR). Area Under the Curve (AUC) quantifies ROC performance, where larger AUC indicates better classifier performance. Accuracy is calculated as:

$$\text{Accuracy} = \frac{TP + TN}{\sum \text{All Samples}}$$

For LIDC-IDRI, benign/malignant feature distributions for small and large nodules show some differentiation but limited differences, with similar average curve trends from rank 1 to rank 5, particularly for LS2 (>10mm). For the combined LS1+LS2, the sum of the first four components (density values < -663 HU) decreases from 51.3% to 32.01% across ranks 1-5, while the sum of the last four components (density values > -282 HU) increases from 11.19% to 34.83%. The proportion of high distribution levels (representing solid regions) consistently ranks as rank 1 < rank 2 < rank 3 < rank 4 < rank 5. Thus, observing nodule density distribution maps and statistical features can more intuitively express lesion structure, improving diagnostic efficiency and classification accuracy.

---

2 Experimental Results and Analysis

This section presents experimental validation of database classification performance. The classification model configuration is: Random Forest (RF) classifier with 100 estimators; 50:50 training/testing sample ratio for LIDC-IDRI; 100 experimental runs per subset with averaged performance metrics. Platform: Python 3.0 on Windows 10; Hardware: Intel Core i7-6700HQ 2.60GHz (8 CPUs), Geforce 960 GPU, 8GB RAM. The proposed system operates efficiently with real-time processing capability, achieving an average single-sample classification time of 35 ± 0.5 ms.

2.1 Pulmonary Nodule Image Feature Extraction and Analysis

LIDC-IDRI sample clustering results are shown in Figure 3 [Figure 3: see original paper]. Density map categories generally appear as concentric rings, with the outermost minimum value (k=1) ensuring non-zero first components in feature vectors. Low-density proportions decrease stably from rank 1 to rank 5, while high-density proportions increase. The consistent ranking of high distribution levels (rank 1 < rank 2 < rank 3 < rank 4 < rank 5) demonstrates reliable statistical significance with clear inter-class differences, substantially improving classification performance. Figure 4 [Figure 4: see original paper] displays mean feature vectors for LIDC-IDRI samples.

2.2 Pulmonary Nodule Image Classification

Experimental results show CF1 performs best on the complete sample set (LS1+LS2), while performance decreases when rank 3 nodules are assigned to benign or malignant categories (CF2 and CF3). Table 2 summarizes AUC, sensitivity, and specificity metrics.

Table 2 Evaluation Parameter Statistics of LIDC-IDRI Classification Performance

Configuration	LS1+LS2 (All Samples)	LS1 (Small Nodules)	LS2 (Large Nodules)
CF1	0.9681±0.0055	0.9820±0.0043	0.9273±0.0196
CF2	0.9405±0.0065	0.9702±0.0045	0.8203±0.0263
CF3	0.8070±0.0129	0.7681±0.0148	0.8941±0.0219

For LS1, performance ranks as CF1 > CF2 > CF3, while for LS2 it ranks CF1 > CF3 > CF2. This suggests rank 3 small nodules in LIDC-IDRI tend toward benignancy, while large ones tend toward malignancy. Since LS1 contains more samples (most LIDC-IDRI nodules are <10mm), rank 3 nodules overall exhibit more benign characteristics.

LIDC-IDRI test accuracies are 0.9008, 0.8782, and 0.7258 for CF1, CF2, and CF3 respectively. Confusion matrices are shown in Figure 5 [Figure 5: see original paper] (Bn = Benign, Ml = Malignant). CF1 demonstrates more stable performance than CF2 and CF3 due to rank 3 sample interference.

As rank 3 samples significantly impact LIDC-IDRI classification, larger rank 3 quantities theoretically reduce stability. Compared to Han et al. (172 samples) and Dhara (349 samples), this study uses 387 rank 3 samples, presenting greater classification challenges. Nevertheless, the proposed grayscale density distribution features achieve substantial evaluation metrics: average AUC values of 0.9681, 0.9405, and 0.8070 for CF1, CF2, and CF3 respectively. Table 3 compares performance with geometric and texture features, showing the proposed model outperforms previous methods despite larger rank 3 sample sizes. This confirms the effectiveness of grayscale density distribution features for pulmonary nodule benign/malignant classification.

Table 3 Comparison of Classification Performance

Method	AUC	Sensitivity	Specificity
Geometric & Texture*	0.8070±0.0129	0.7239±0.0297	0.7296±0.0476
Han et al. [9]	0.7681±0.0148	0.5958±0.0402	0.8047±0.0573
Dhara et al. [10]	0.7239±0.0297	0.7296±0.0476	-

*Using Dhara' s geometric and texture features for ROC calculation.

3 Conclusion

This paper proposes a grayscale density distribution feature calculation method for pulmonary nodules based on image block sets. The approach first constructs a block set by randomly selecting uniformly-sized patches from nodule images, then calculates the autocorrelation matrix and clusters it into 10 categories to obtain block labels. By traversing nodule images, each pixel' s surrounding window is matched against the block set, with the best-matching block' s label assigned as the pixel' s density distribution level. Finally, density distribution maps are statistically generated to extract features. Experimental results and comparative analysis demonstrate that the density distribution-based feature evaluation algorithm effectively classifies benign/malignant nodule levels. This research provides a novel method for clinical auxiliary diagnosis of pulmonary nodules and offers valuable insights for developing early lung cancer diagnosis systems in China and Asia.

References

[1] Zhang Man. Establishment of a mathematic model for predicting malignancy in solitary pulmonary nodules [D]. Guangzhou: Southern Medical University, 2016.

[2] Armato S G, McLennan G, et al. The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans [J]. Medical Physics, 2011, 38(2): 915-931.

[3] Welch H G, Woloshin S, Schwartz L M, et al. Overstating the evidence for lung cancer screening: the international early lung cancer action program (I-ELCAP) study [J]. Archives of Internal Medicine, 2007, 167(21): 2289-2295.

[4] National Lung Screening Trial Research Team. Reduced lung-cancer mortality with low-dose computed tomographic screening [J]. New England Journal of Medicine, 2011, 365(5): 395-409.

[5] Xing Qianqian. Research on irregular lung nodule automatic segmentation and spiculation detection [D]. Guangzhou: Southern Medical University, 2015.

[6] Tan Yongqiang, Schwartz L H, Zhao Binsheng. Segmentation of lung lesions on CT scans using watershed, active contours, and Markov random field [J]. Medical Physics, 2013, 40(4): 043502-043502.

[7] El-Baz A, Nitzken M, Elnakib A, et al. 3D shape analysis for early diagnosis of malignant lung nodules [C]//Proc of International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer-Verlag, 2011: 175-182.

[8] Breiman L. Random forests [J]. Machine Learning, 2001, 45(1): 5-32.

[9] Han Fangfang, Wang Huafeng, Zhang Guopeng, et al. Texture feature analysis for computer-aided diagnosis on pulmonary nodules [J]. Journal of Digital Imaging, 2015, 28(1): 99-115.

[10] Dhara A K, Mukhopadhyay S, Dutta A, et al. A combination of shape and texture features for classification of pulmonary nodules in lung CT images [J]. Journal of Digital Imaging, 2016, 29(4): 466-475.

[11] Reeves A P, Xie Yiting, Jirapatnakul A. Automated pulmonary nodule CT image characterization in lung cancer screening [J]. International Journal of Computer Assisted Radiology & Surgery, 2016, 11(1): 73-88.

[12] Maldonado F, Boland J M, Raghunath S, et al. Noninvasive characterization of the histopathologic features of pulmonary nodules of the lung adenocarcinoma spectrum using computer-aided nodule assessment and risk yield (CANARY)—pilot study [J]. Journal of Thoracic Oncology, 2013, 8(4): 371-378.

[13] Rousseeuw P J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis [J]. Journal of Computational & Applied Mathematics, 1987, 20: 53-65.

[14] Sculley D. Web-scale k-means clustering [C]//Proc of International Conference on World Wide Web. 2010: 1177-1178.

[15] Pei Bo, Qiang Yan, Zhao Juanjuan. A PET/CT-based prediction model for malignancy probability of solitary pulmonary nodules [J]. Computer Applications and Software, 2015, 32(12): 170-174.

Note: Figure translations are in progress. See original paper for figures.