Lung nodules classification in CT images using texture descriptors

Classificação de nódulos pulmonares em imagens de TC utilizando descritores de textura

Luís V. de Moura[1], Caroline M. Dartora[1,2], Ana Maria M. da Silva[1,2,3]

[1]PUCRS, School of Science, Laboratory of Medical Imaging, Porto Alegre, Brazil
[2]PUCRS, School of Medicine, Graduate Program in Biomedical Gerontology, Porto Alegre, Brazil
[3]PUCRS, Brain Institute – INSCER, Porto Alegre, Brazil

Abstract

CAD systems can be applied in several medical fields, but mostly in cancer detection and diagnosis. In the case of lung nodules, an early diagnosis of then can improve potentially the prognosis. Accurate interpretation of CT scans demands big efforts by radiologists due to a large number of images that is required, managed and analyzed every day. Therefore, the necessity of invasive pathological tests, like biopsy, remains necessary and this is why CAD for lung nodules is so used. One of the most used methods in CAD for lung abnormalities is feature extraction. The aim of this exploratory study is to investigate the performance of lung nodules classification in 2D and 3D CT images using Haralick texture feature analysis and binary logistic regression. Data used in this work came from a database of Hospital São Lucas da PUCRS and have 17 benign and 20 malignant nodules with anatomopathological results. Haralick feature extraction was used in segmented images by an expert radiologist. For 2D images, the largest nodule area slice was used. For 3D images, GLCM with all volume of nodule in 26 angles was constructed. For 3D images features calculation, the combination of zero-pairs in GLCM was removed. Dimensionality was reduced for two or three components by PCA. Binary logistic regression results showed an accuracy of ≥70% in lung nodules classification using two and three PC. The use of 3D images to GLCM and Haralick texture features extraction with 3 PC showed better sensitivity (80%) and specificity (73%) on nodule classification, although worse accuracy, than 2D image analysis.
Keywords: lung nodules; CT image; CAD; Haralick descriptors;

Resumo

Sistemas CAD são aplicados em diversas áreas médicas, principalmente na detecção e diagnóstico de câncer. No caso de nódulos pulmonares, um diagnóstico precoce do nódulo pode aumentar potencialmente o prognóstico. A acurada interpretação das imagens de TC demanda grandes esforços dos radiologistas devido ao grande número de exames que são solicitados, gerenciados e analisados diariamente. Assim, a necessidade de testes patológicos invasivos, como biopsia, permanece necessária e por isso sistemas CAD para nódulos pulmonares é tão utilizado. Um dos métodos mais utilizados em CAD para anormalidades no pulmão é a extração de características. O objetivo deste trabalho exploratório é investigar o desempenho na classificação de nódulos pulmonares em imagens 2D e 3D de TC de pulmão utilizando a análise de características de textura pelo método de Haralick e regressão logística binaria. Os dados utilizados nesse estudo são do banco de dados do Hospital São Lucas da PUCRS e possui 17 nódulos benignos e 20 malignos, com laudo dos testes anatomopatológicos. As imagens utilizadas foram segmentadas por um radiologista experiente. Para imagens 2D, o corte do nódulo de maior área foi utilizado. Para imagens 3D, a GLCM foi construída com todo o volume do nódulo considerando 26 ângulos. A dimensionalidade foi reduzida para dois e três componentes por PCA e os resultados da regressão logística binária mostraram acurácias ≥70% na classificação de nódulos. O uso de imagens 3D para a construção da GLCM e extração das características de textura da imagem pelo método de Haralick com 3 componentes principais na regressão logística binária apresentou melhor sensibilidade (80%) e especificidade (73%) na classificação de nódulos, apesar de não apresentar melhor acurácia do que a análise em imagens 2D.
Palavras-chave*: nódulos de pulmão; tomografia computadorizada, CAD, descritores de Haralick.*

1. Introduction

Computer-aided Diagnosis (CAD) is one of the most used tools in the diagnosis of lung cancer. CAD is defined as the use of clinical information by a trained radiologist and the quantitative analysis made by a computer. Its purpose is to improve diagnostic accuracy and reproducibility as the radiologist image interpretation is guided for a computer[1].

CAD can be divided into two subgroups: CADe and CADx. CADe is for computer-aided detection when the assistant of a computer is to indicate a suspected location of lesions or abnormalities. CADx is for the directly computer-aided diagnosis when the CAD

system helps in the characterization of a region or lesion, which have been previously located and indicated to the computer[2].

CAD systems can be applied in several diagnostic fields, but mostly in cancer detection and diagnosis, for example, in mammography, in images of colonoscopy and for the diagnostic of lung nodules[2].

In the case of lung nodules, an early diagnosis of then can improve potentially the prognosis. The term nodule is used for opacity with a diameter between 3 and 30 mm[3]. Some characteristics like location or position, neighboring structures, size, shape and internal density, as solid, subsolid or non-solid, can indicate if the nodules it is benign or malignant (cancerous)[4,5]. The likelihood that a nodule can be cancerous is about 40% but the risk with age varies. When a nodule is detected in an X-ray Computed Tomography (CT) scan, the radiologist usually compares with an early one. If there is a change in a nodule or a new one, as the first CT scan, a bronchoscopy or tissue biopsy is recommended to determine if it is a malignant nodule[6]. The accurate interpretation from a CT scan demands big efforts by the radiologists due to a large number of CT scan that is required, managed and analyzed every day. Therefore, the necessity of invasive pathological tests, like biopsy, remains necessary[4] and this is why CAD for lung nodules shows to be so used.

Since CAD was defined in 1999[1], your use for lung abnormalities has been widely studied, implemented and improved. One of the most used methods used in CADx for lung abnormalities is the feature extraction, which is a process to obtain higher-level of information of the image[7]. As the texture is one of the key components of human visual perception[7], statistical approaches of these features were widely used during the last years. In 1973, Haralick had presented in an IEEE transaction paper the development of new textural features for image classification[8]. Haralick presented the Gray Level Occurrence Matrix (GLCM) and a set of 14 textures features that compute local features at each point in the image and inferred a set of statistics from the distributions of the local features[7,8].

Texture features were first developed for two-dimensional (2D) images, as chest X-ray. However, with the advance of computers, three-dimensional (3D) methods to apply texture analysis were developed. In 2014, Han et al[9] realized a systematic investigation of three types of 2D texture features and extend to study the impact of expansion as a differentiation task to a 3D space. They had used an available database of CT images and compared 2D texture features as Haralick, Gabor and Local Binary Patterns. They have gain when calculating 2D features on all image slices as compared to a single largest slice. They had observed that Haralick features type is a better choice to differentiate malignant and benign lung nodules and the 3D extension reveals a potential gain in the results when an optimal number of directions to construct the GLCM can be found.

In 2018, Wei et al.[10] had implemented a content-based image retrieval scheme with a two steps similarity metric approach to classify lung nodules based on Mahalanobis and European distances. He had used three groups of texture descriptors (local binary pattern feature, Gabor features, and Haralick features) in 366 2D images of benign and malignant nodules of Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI). Results shows that classification accuracy using just Haralick texture descriptors was better than a combination of the other used methods.

CADx system has been developed in the last years using Artificial Intelligence (AI) methods[11–14]. Machine Learning, radiomics and neural networks are the most used algorithms for classification of lung nodules. They had shown excellent accuracy (more than 80%) using big databases, 2D and 3D images, and a combination of features, as texture, morphological, and demographic data. However, hospitals and clinics have no software/hardware support for this type of technology.

The aim of this exploratory study is to investigate the performance of lung nodules classification in 2D and 3D CT images using Haralick texture feature analysis and binary logistic regression.

2. Materials and Methods

2.1. Data of study

Data used in this project came from a database of Hospital São Lucas da PUCRS. Use of this data had ethical approvement (CAAE: 12385713.4.0000.5336, protocol: 339574). The database came with CT images of the lung, and data of clinical evaluation and anatomopathological tests of nodules. Anonymization of patient data was preserved.

The images were acquired in a Siemens Somatom Emotion 16 channels, 120 kV, 241 mAs, slice thickness as 2.0 mm, pitch 0.9 and with a matrix of 512 x 512 pixels with a resolution of 1.185 pixel/mm.

Patients were divided into two groups: (1) patients with malignant nodules and (2) patients with benign nodules, with 20 and 17 subjects, respectively. Malignant nodules involve adenocarcinoma of levels II and III and squamous cell carcinomas. Benign nodules include hamartoma, lipoma, and chondroadenoma.

The data is referred to 37 patients, being 20 females (8 benign and 12 malignant nodules) and 17 male (9 benign and 8 malignant nodules), with age varying between 41 and 83 years. Mean patients age for malignant nodules was 62.0 ± 9 and 62.8 ± 12 for benign nodules.

In this study, the anatomopathological test results were used as the gold standard of diagnosis for the classification model analysis.

2.2. Nodules images pre-processing

Segmented images of lung nodes by radiologist were used to avoid tissues of no interest in the quantitative analysis. Segmented images of lung nodules were organized in symmetric matrices depending on to the larger diameter of the nodule.

2.3. Features Extraction

An in-house code in MATLAB R2012b was developed to calculate GLCM and apply Haralick texture features in 2D (slice of the nodule with the largest area) and 3D images of lung nodules. The algorithm calculates GLCM based in 8 angles (directions), to 2D images, and 26 angles, for 3D images. Before applying Haralick descriptors, the zero-pair of numbers in GLCM was discarded to avoid high fluctuations in the features due to a large number of zero-pairs.

The Haralick texture features was based in the mathematical description of the original article of Haralick[8]. The calculated features used in this work was:

- Angular second moment (ASM);
- Contrast;
- Correlation;
- Dissimilarity;
- Energy;
- Entropy;
- Homogeneity;
- Variance;
- GLCM mean.

2.4. Classification model

Binary logistic regression with principal component analysis (PCA) was used to differentiate benign nodules of the malignant ones based in the Haralick features extracted of nodules image. SPSS® 17 was used for statistical analysis.

Data were normalized, and PCA was applied due to the small number of images compared to the number of features. Selected components to explain more than 75% of data variance were described. These components were used in binary logistic regression with $p<0.005$.

2.5. Performance analysis of the classification model

To a quantitative analysis of prediction model, we computed the sensitivity, specificity, and accuracy of each method.

Sensitivity was calculated following the equation 1:

$$Sensitivity = \frac{TP}{TP + FN} * 100\% \quad (1)$$

Specificity was calculated following the equation 2:

$$Specificity = \frac{TN}{TN + FN} * 100\% \quad (2)$$

Accuracy was calculated following the equation 3:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} * 100\% \quad (3)$$

Where True Positive (TP) occurs when the classification results in malignant nodules and it is really a malignant nodule and False Positive (FP) occurs when the classification gives a malignant nodule the result of a benign one. True Negative (TN) are the true classification of benign nodules in benign and False Negative (FN) occurs when malignant nodules are classified as benign ones[15].

3. Results

We had used Haralick texture features in segmented 2D and 3D CT images to investigate the performance of lung nodules classification using binary logistic regression. Our set of images were composed of 17 benign and 20 malignant nodules.

Mean (m) and standard deviation (std) of Haralick features data for the 2D and 3D nodules image analysis is shown in Table 1.

Table 1 - Mean and standard deviation (mean ± std)

Haralick features	2D		3D	
	Malignant	Benign	Malignant	Benign
ASM	0.29 ± 0.20	0.13 ± 0.07	0.008 ± 0.004	0.01 ± 0.01
Contrast	21416 ± 20196	21949 ± 16264	22988 ± 17008	39658 ± 27394
Correlation	0.96 ± 0.04	0.95 ± 0.04	0.69 ± 0.13	0.75 ± 0.12
Dissimilarity	43 ± 27	63 ± 31	73 ± 36	115 ± 52
Energy	0.5 ± 0.2	0.35 ± 0.10	0.09 ± 0.02	0.10 ± 0.05
Entropy	4.9 ± 1.5	6 ± 1	10.2 ± 0.9	10.4 ± 0.9
Homogeneity	0.5 ± 0.2	0.37 ± 0.08	0.05 ± 0.02	0.07 ± 0.07
Variance	233711 ± 62657	238983 ± 70077	44347 ± 31080	106384 ± 92548
GLCM mean	473 ± 169	545 ± 99	964 ± 92	821 ± 179

Source: The author (2019).

Principal Component Analysis (PCA) to reduce the dimensionality of features in data was applied due to the restricted number of images. PCA showed that the use of two and three components can explain, for 2D data analysis, 85.8% and 96.25% of data variance. For 3D data analysis, two and three components could explain 72.4% and 91.7% of data variance.

Sensitivity, specificity, and accuracy to evaluate the performance of the classification method were calculated. Performance analysis for binary logistic regression was calculated using two and three principal components as shown in Table 2.

Table 2 - Classification performance with 2 and 3 principal components (PC)

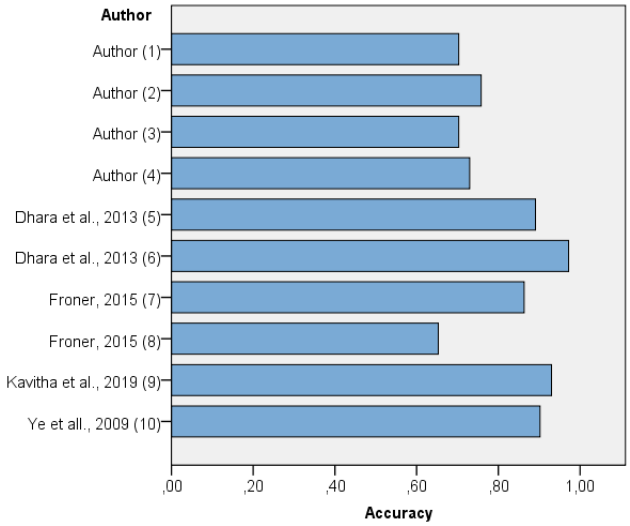
Image	# of PC	Sensitivity	Specificity	Accuracy
2D	2	75%	69%	70%

	3	75%	72%	76%
3D	2	75%	69%	70%
	3	80%	73%	73%

Source: The author (2019).

Our results were compared to other studies[5,12–14] to show the efficacy of proposed method. Figure 1 shows the comparison between our results of 2D and 3D data with other studies.

Figure 1 - Accuracy comparison between our results and literature.



Source: The author (2019).

Where
(1) Our data results for 2D images with 2 PC;
(2) Our data results for 2D images with 3 PC;
(3) Our data results for 3D images with 2 PC;
(4) Our data results for 3D images with 3 PC;
(5) Dhara et al., 2013[13], results for 2D images of nodules;
(6) Dhara et al., 2013[13], results for 3D images of nodule;
(7) Froner, 2015[5], multivariate logistic regression method
(8) Froner, 2015[5], accuracy results of an expert radiology physician in classification based just in images, within random sorting;
(9) Kavitha et al., 2019[12], which had applied Haralick texture features in 2D images
(10) Ye et al., 2009[14], which used features in 2D and 3D images.

4. Discussions

The aim of this study was to investigate the performance of lung nodules classification in 2D and 3D CT images using texture feature analysis. The used classification method was binary logistic regression and was used two and three components extracted by PCA in original data.

Our results show more than 70% of accuracy using 2D and 3D image data to calculate the GLCM and texture attributes based on Haralick method. Comparison in Figure 1 shows that our proposed method is less accurate than that in literature. However, literature results are based on AI algorithms, like Support Vector Machines[12,14] and Artificial Neural Networks[13], with bigger databases than ours are. It is worth mentioning that, within the models proposed in the literature, ours had a smaller database and a huge variety of benign and malignant type of nodules to the size of sample images. Information in literature shows more attributes to classification that texture, as of intensity, morphology, and demographic data of subjects.

Is important to emphasize that hospitals and clinics have no resources and support for software/hardware required by Artificial Intelligence algorithms in CAD systems. Binary logistic regression can be applied in any regular computer with a statistical software, as Matlab.

Interestingly, obtained results diverge from that founded in Dhara *et. al*. (2013)[13], which uses neural networks to classification nodules based on Haralick texture features of 2D and 3D image data. This is due to the bigger sample images that they used in the classification that our method.

Froner[5], in 2015, had used the same database of CT lung images than use. Her aim was to evaluate the use of patient data (demographic) and quantitative attributes of lung nodules of CT images to build a model of classification in terms of malignancy. She had analyzed a lot of morphological, demographical, and texture features in a model of multivariate logistic regression. Although the different number of sample images used in this work, as features information, our results showed better predictive values than visual evaluation by expert radiology physicians. Froner[5] results of physicians evaluations of CT images, not including the result by chance, was 65.3% where our results are above 70%.

Limitations of this study include small sample size and number of attributes and lack of morphological features. For future work a large sample of CT images, more attributes, as morphological and demographic ones will be necessary.

Although the limitations of this study, we showed an accuracy ≥70% in lung nodules classification using two and three PC in binary logistic regression. The use of 3D images to GLCM and Haralick texture features extraction with 3 PC in binary logistic regression showed better sensitivity (80%) and specificity (73%) on nodule classification, although worse accuracy, than 2D image analysis.

5. Conclusions

We investigated the performance of lung nodules classification in 2D and 3D lung CT images using texture feature analysis. The binary logistic regression classification method was used as well as PCA due to the reduced number of images.

Although the limitations of this study, we showed a better accuracy in lung nodules classification predictive values than visual evaluation by expert radiology physicians, using a binary logistic regression. The use of 3D images increased the sensitivity and specificity on lung nodule classification. Further studies are required to improve the classification method.

Acknowledgements

The authors would like to thank PUCRS and CAPES for the financial support.

References

1. Doi K, MacMahon H, Giger ML, Hoffmann KR. Computer-aided diagnosis and its potential impact on diagnostic radiology. Comput Diagnosis Med imaging Amsterdam, Netherlands Elsevier Sci. 1999;11–20.
2. Giger ML, Suzuki K. Computer-Aided Diagnosis. Biomed Inf Technol [Internet]. 2008 Jan 1 [cited 2019 Mar 18];359–XXII. Available from: https://www.sciencedirect.com/science/article/pii/B9780123735836500207
3. Zia ur Rehman M, Javaid M, Shah SIA, Gilani SO, Jamil M, Butt SI. An appraisal of nodules detection techniques for lung cancer in CT images. Biomed Signal Process Control [Internet]. 2018 Mar 1 [cited 2019 Mar 18];41:140–51. Available from: https://www.sciencedirect.com/science/article/pii/S1746809417302811
4. Rubin GD. Lung nodule and cancer detection in computed tomography screening. In: Journal of Thoracic Imaging [Internet]. 2015 [cited 2019 Mar 18]. p. 130–8. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4654704/pdf/nihms714352.pdf
5. Froner APP. Caracterização de nódulos pulmonares em imagens de tomografia computadorizada para fins de auxílio ao diagnóstico [Internet]. Pontifícia Universidade Católica do Rio Grande do Sul; 2015 [cited 2019 Apr 10]. Available from: http://tede2.pucrs.br/tede2/handle/tede/6385
6. Madero Orozco H, Vergara Villegas OO, Cruz Sánchez VG, Ochoa Domínguez H de J, Nandayapa Alfaro M de J. Automated system for lung nodules classification based on wavelet feature descriptor and support vector machine. Biomed Eng Online [Internet]. 2015 Dec 12 [cited 2019 Mar 18];14(1):9. Available from: http://www.biomedical-engineering-online.com/content/14/1/9
7. Zayed N, Elnemr HA. Statistical Analysis of Haralick Texture Features to Discriminate Lung Abnormalities. Int J Biomed Imaging [Internet]. 2015 Oct 8 [cited 2019 Apr 3];2015:1–7. Available from: http://www.hindawi.com/journals/ijbi/2015/267807/
8. Haralick RM, Shanmugam K, Dinstein I. Textural Features for Image Classification. IEEE Trans Syst Man Cybern [Internet]. 1973 Nov [cited 2019 Apr 10];SMC-3(6):610–21. Available from: http://ieeexplore.ieee.org/document/4309314/
9. Han F, Wang H, Zhang G, Han H, Song B, Li L, et al. Texture Feature Analysis for Computer-Aided Diagnosis on Pulmonary Nodules. J Digit Imaging [Internet]. 2014 [cited 2019 Mar 22];28(1):99–115. Available from: https://link.springer.com/content/pdf/10.1007%2Fs10278-014-9718-8.pdf
10. Wei G, Cao H, Ma H, Qi S, Qian W, Ma Z. Content-based image retrieval for Lung Nodule Classification Using Texture Features and Learned Distance Metric. J Med Syst [Internet]. 2018 [cited 2019 Apr 15];42(1). Available from: https://doi.org/10.1007/s10916-017-0874-5
11. Franco MLN, Nunes LM, Froner APP, Silva AMM, Patrocínio AC. REDES NEURAIS ARTIFICIAIS APLICADAS NA CLASSIFICAÇÃO DE TUMORES PULMONARES [Internet]. [cited 2019 Apr 15]. Available from: http://www2.inca.gov.br/wps/wcm/connect/tiposdec
12. Kavitha MS, Shanthini J, Sabitha R. ECM-CSD: An Efficient Classification Model for Cancer Stage Diagnosis in CT Lung Images Using FCM and SVM Techniques. J Med Syst [Internet]. 2019 Mar 12 [cited 2019 Apr 15];43(3):73. Available from: http://www.ncbi.nlm.nih.gov/pubmed/30746555
13. Dhara AK, Mukhopadhyay S, Khandelwal N. 3D texture analysis of solitary pulmonary nodules using co-occurrence matrix from volumetric lung CT images. In: Novak CL, Aylward S, editors. International Society for Optics and Photonics; 2013 [cited 2019 Apr 15]. p. 867039. Available from: http://proceedings.spiedigitallibrary.org/proceeding.aspx?doi=10.1117/12.2007016
14. Xujiong Ye, Xinyu Lin, Dehmeshki J, Slabaugh G, Beddoe G. Shape-Based Computer-Aided Detection of Lung Nodules in Thoracic CT Images. IEEE Trans Biomed Eng [Internet]. 2009 Jul [cited 2019 Apr 15];56(7):1810–20. Available from: http://ieeexplore.ieee.org/document/5073252/
15. Ramteke RJ, Y KM. Automatic Medical Image Classification and Abnormality Detection Using K-Nearest Neighbour. Int J Adv Comput Res [Internet]. 2012 [cited 2019 Apr 3];2(4):190–6. Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.300.4729&rep=rep1&type=pdf

Contact:

Luís Vinícius de Moura
PUCRS, Laboratory of Medical Imaging
Av. Ipiranga, 6681 Partenon - Porto Alegre/RS
CEP: 90619-900 - Phone: (51) 3320.7813
Email: luis.moura.001@acad.pucrs.br