JHI JOURNAL OF HEALTH INFORMATICS

Shape Analysis 3D to false positive reduction

Análise de Forma 3D para redução de falsos positivos

Análisis de Forma 3D para la reducción de falsos positivos

Alexandre Ribeiro Cajazeira Ramos[1], Antonino Calisto dos Santos Neto[1], Antonio Oseas de Carvalho Filho[2]

ABSTRACT

Keywords: False positive reactions; Anatomy regional; Lung neoplasms

Objective: The present work describes the development of a methodology for the reduction of false positives, so that it helps in the consistent distinction between pulmonary nodules and other structures present in CT examinations. **Methods:** The proposed method is based on the exploration of characteristics of form related to compactness, spherical disproportion, Feret diameters, skeleton and statistical measures. These characteristics were submitted to a set of classifiers, highlighting the MLP algorithm. The method was evaluated in a set of 24156 regions from 833 examinations of the LIDC-IDRI base. **Results:** The method was promising, reaching 91% of sensitivity, 92.8% of specificity and 92.3% of accuracy. **Conclusion:** The results showed that the morphological analysis is satisfactory and presents significant contributions to the CAD systems, directly aiding in a more favorable prognosis to the patient.

RESUMO

Descritores: Reações falso-positivas; Anatomia regional; Neoplasias pulmonares

Objetivo: O presente trabalho descreve o desenvolvimento uma metodologia para a redução de falsos positivos, de modo que auxilie na distinção consistente entre nódulos pulmonares e outras estruturas presentes em exames de tomografia computadorizada. **Métodos:** O método proposto baseia-se na exploração de características de forma relacionadas à compacidade, desproporção esférica, diâmetros de Feret, esqueleto e medidas estatísticas. Estas características foram submetidas a um conjunto de classificadores, com destaque para o algoritmo MLP. O método foi avaliado em um conjunto de 24156 regiões oriundas de 833 exames da base de imagens LIDC-IDRI. **Resultados:** O método mostrou-se promissor, alcançando 91% de sensibilidade, 92,8% de especificidade e 92,3% de acurácia. **Conclusão:** Os resultados mostraram que a análise morfológica é satisfatória e apresenta contribuições significativas para os sistemas CAD, auxiliando diretamente em um prognóstico mais favorável ao paciente.

RESUMEN

Descriptores: Reacciones Falso Positivas; Anatomía regional; Neoplasias pulmonares

Objetivo: El presente trabajo describe el desarrollo de una metodología para la reducción de falsos positivos, de modo que auxilie en la distinción consistente entre nódulos pulmonares y otras estructuras presentes en exámenes de tomografía computarizada. **Métodos:** El método propuesto se basa en la exploración de características de forma relacionadas con la compacidad, desproporción esférica, diámetros de Feret, esqueleto y medidas estadísticas. Estas características fueron sometidas a un conjunto de clasificadores, con destaque para el algoritmo MLP. El método fue evaluado en un conjunto de 24156 regiones oriundas de 833 exámenes de la base de imágenes LIDC-IDRI. **Resultados:** El método se mostró prometedor, alcanzando el 91% de sensibilidad, el 92,8% de especificidad y el 92,3% de exactitud. **Conclusión:** Los resultados mostraron que el análisis morfológico es satisfactorio y presenta contribuciones significativas para los sistemas CAD, auxiliando directamente en un pronóstico más favorable al paciente.

[1] *Graduado em Sistemas de Informação, pela Universidade Federal do Piauí - UFPI, Teresina (PI), Brasil.*
[2] *Professor Doutor da Universidade Federal do Piauí - UFPI, Teresina (PI), Brasil.*

Autor Coorespondente: **Alexandre Ribeiro Cajazeira Ramos**
e-mail: **alexandre.cajamos@gmail.com**

Artigo recebido: **01/12/2016**
Aprovado: **14/06/2017**

INTRODUCTION

Cancer is a set of diseases with one aspect in common: the uncontrolled growth of cells that invade tissues and organs causing substantial damage. This is an aggressive and uncontrollable process, determining the formation of malignant tumors that may spread to other places[1].

The factors that lead to this condition are varied and are usually associated with environmental variables, lifestyle and contact with substances with carcinogenic potential, as it occurs at the contamination, smoking habit or pesticide intake[1].

The impact caused by late detection of cancer cases is devastating. Among the diagnosis of the most common and lethal malignant tumors, there is the lung cancer, which at the end of the twentieth century became one of the leading causes of preventable deaths, with average five-year survival ranging from 13 and 21% in developed countries and between 7 and 10% in developing countries[2].

The annual forecast increase in its worldwide incidence is 2%, and the latest estimate pointed to 1.92 million new cases of lung cancer in 2012. In Brazil, for example, the INCA's forecast is 28,220 new cases, with 17,330 in men and 10,890 in women, with a death rate of 24,490 people, of which 14,811 are men and 9,675 are women[2].

The diagnosis of this type of cancer is made by experts through the interpretation of a computed tomography (CT) scan of the chest. However, because the analysis of the components of these tests should be exhaustive and the implications of the diagnosis are high risk, there is an increase in the number of studies that aim to contribute to the work of these experts, reducing errors that in this context are unacceptable.

From the problems generated by these diseases world wide, a range of techniques that help the detection and diagnosis has been studied and implemented in different areas of knowledge. In the literature it's possible to find Computer-Aided Detection (CAD), a set of tools that translate the peculiar information of pathologies in computing solutions, using techniques related to digital image processing, artificial intelligence, machine learning, among others, with significant responses to demands in question[3].

The early detection of lesions that can become cancer has a substantial contribution on the survival of patients, and CAD systems, in this context, associated shape and texture information of these lesions to distinguish nodules from other tissues found on CT scans, like blood vessels for example.

The objective of this paper is to develop a methodology to reduce false positives reactions, through the extraction of morphological features. The features of compactness, spherical disproportion, Feret diameters, skeleton, statistical measures and several geometric analysis equations were studied.

This paper will compose a very important step in a CAD system, assisting the consistent distinction between nodules and other structures present in computed tomography. In addition to evaluating the generalization capacity of the implemented descriptors as well as their results in several classifiers. In addition, this paper show advances in relation to other analyses under this focus, like methodology proposed in Carvalho Filho et. al.[4] which fails in cases where the cylindrical approach tends to characterize a blood vessel, once, this is a perfect cylinder, this flaw was overcome in the operationalization of the method described in this paper.

METHOD

The methodology developed in this work includes the steps of acquiring an image base, the extraction of characteristics of the regions of interest by implementing form lung nodule descriptors, the feature selection, classification and validation of results.

We divided our approach in the following steps: Image Acquisition; Characteristics extraction and Classification. Each of these steps are described below.

Image Acquisition

We used the public image base LIDC-IDRI, provided by the National Cancer Institute (NCI) in the USA, resulting from the junction of the Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI) bases. These images will be used for training and validation of the method[5].

As the input of the feature extraction process, we used the results of the segmentation step of nodule candidates proposed by Carvalho Filho et. al.[6]. Given that: 1 - The base LIDC-IDRI images comprises of a set of CT of entire lung parenchyma; 2 - a lot of examples of solitary lung-nodules and no nodule is required for testing and validation of the proposed methodology; 3 - the segmentation process, responsible for separating parenchymal tissues, is a complex step and can not be done manually; 4 – The result of the segmentation step of nodule candidates proposed by Carvalho Filho et. al.[6] is consistent when compared to the notes of four LIDC-IDRI specialists.

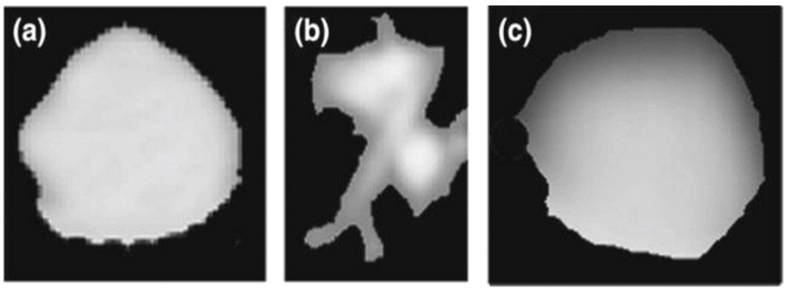

Figure 1 - Examples of lung-nodule candidates. Wherein (a) it is a nodule, (b) and (c) no nodule, adapted from Carvalho Filho et al.[6].

Thus, 24156 segmented images belonging to 833 tests from public base images LIDC-IDRI, including 6414 nodules and 17742 no-nodules, are analyzed and subjected to a set of classifiers.

Figure 1 shows examples of candidate nodule segmented by Carvalho Filho et. al.[6] and how these structures are difficult to differentiate.

Feature Extraction

A common characteristic of cancer, for example, is the uncontrolled growth of diseased cells, making a surface with atypical and irregularly shape, but which describes the development of disease and, hence, is essential for its understanding.

The CT scans contains some objects that can be lungnodules or other healthy tissues. One of the criteria used by specialists in the differentiation between these structures is shape that they assume, such as blood vessels which are elongated, while nodules are rounded. Therefore the extraction of characteristics used in this paper seeks the representation of morphological peculiarities that are capable of differentiating lung-nodules from other structures presents in the CT scans.

For the analysis of the features of structures present in medical images, uses the mathematical morphology, a reasoned concept in set theory that is the analysis of geometric and topographical characteristics of a region.

Among the studied descriptors are 2.2.1 Spherical Disproportion; 2.2.2 Compactness; 2.2.3) Statistical Measures Related to Edge Distances to the Object's Mass Center; and 2.2.4 Relations based on geometric features, skeleton and Feret diameters of the analyzed structures. Thus, all the morphological characteristics analyzed constitute the set of descriptors that will provide important information for the classifiers in consistent differentiation between nodules and other structures.

Spherical disproportion

The spherical disproportion measures how irregular is the surface relative to a perfectly spherical surface so that it can differentiate lung-nodules from blood vessel or other tissue, considering that nodules tend to be spherical, whereas vessels tend to be elongated[7].

It's possible measure the spherical disproportion (DespEsf) of an object from the following formula:

$$DespEsf = \frac{A}{4\,\pi\,R^2} \quad (1)$$

where *A* is the area of the object and *R* is the estimated distance, calculated using the Equation 2:

$$R = \sqrt[3]{\frac{3V}{4\pi}} \quad (2)$$

where *V* is the object volume.

Compactness

The compactness (Comp) is the object density relative to a perfectly dense object such as a circle. Thus, it shows how a structure is dense, and consequently, the internal behavior of its shape. We can calculate by the formula below.

$$Comp = \frac{p^2}{4\pi A} \quad (3)$$

where *p* is the perimeter and *A* the area of the object.

A Figure 2 shows comparison of objects through compactness.

Statistical Measures

In order to score, with precision, the variation between the distances from the edge points to the center of the object, we calculated the mean and standard deviation of these distances, which are statistical measures that punctuate the degree of dispersion values, and its variation[9].

$$M = \frac{\int_i^n Xi}{n} \quad (4)$$

$$DP = \sqrt{\frac{\int_i^n (Xi - M)^2}{n}} \quad (5)$$

where *M* is the average *DP* standard deviation, *n* the number of analyzed values (Number of object border points), *Xi* is the distance of each point from the edge to the center of mass of the object.

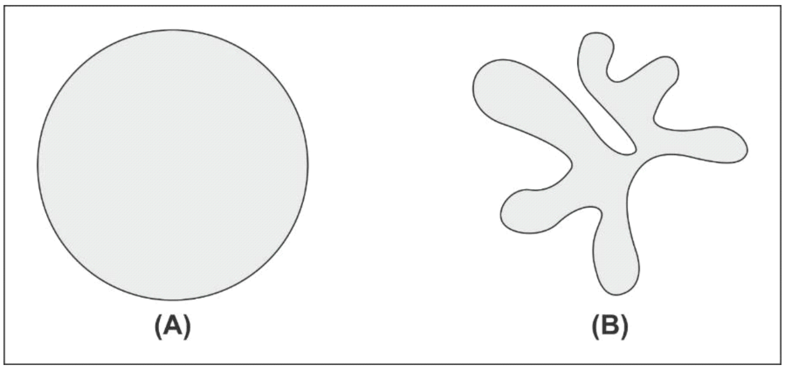

Figura 2 - Compare (a) a compact and (b) a not compact object. Adapted from Sampaio[8].

Relations based on geometric features, skeleton and Feret diameters

The Feret diameters measure the distances between opposite straight lines that touch the edge of the object, demonstrating its length with respect to a specific direction. We can see these parameters in Figure 3, which analyzes a two-dimensional object. All these measures are used as descriptors and contribute to the characterization of lung nodules[8].

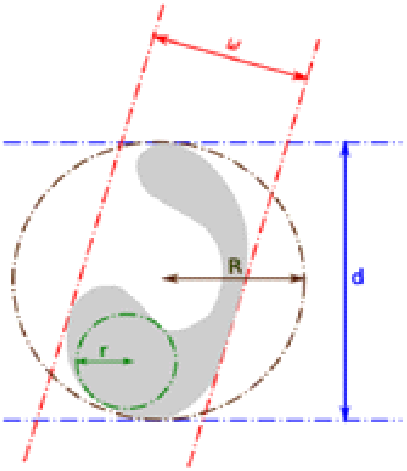

Figura 3 - Ilustration of Feret Diameters.

Where w is the minimum Feret Diameter, d the maximum Feret Diameter, R the radius of the smallest outer circumference and r the radius of the largest inner circumference of the object Adapted from Sampaio[8].

The skeletons of a geometric shape are constituted by the coordinates representing the centers of the circles that touch at least two points on the edge of an object[7], as shown in Figure 4, disposed below.

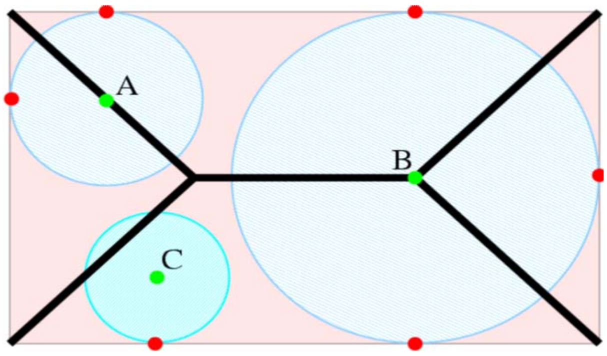

Figure 4 - Skeleton illustration of a rectangle, where the *A* and *B* belong to the skeleton and *C* does not. Adapted from Sampaio[8].

Therefore, the skeleton of an object represents the characteristics of its shape and surface in a reduced set of coordinates of points, allowed quickly understand the object volume and the behavior of the surface for each region of your skeleton. We can then extract as characteristic the size of the skeleton of a particular object, but also its relationship to other implemented features such as area, maximum Feret, minimum Feret, among others.

Associating the variables present in the Feret diameters, skeleton and geometric information such as volume and perimeter, It is possible to extract features that are invariant to rotation, translation and scale[8]. We can observe, in table 1, the equations that present such characteristics and that were implemented and used in this paper.

Table 1 - Other proposed descriptors.

	Formulas obtained through the relationship between descriptors
1	$r\,/\,R$
2	$w\,/\,(2R)$
3	$A/2\pi R^2$
4	$2r/d$
5	w/d
6	$4A/\pi d^2$
7	$R^2\sqrt{3}/d$
8	$2\pi r/P$
9	$\pi w/P$
10	$4\pi A/P^2$
11	$2d/P$
12	$4R/P$
13	$\pi r2/A$
14	$2r/w$
15	$d\,/\,E$
16	$w\,/\,E$

Where *A, P, r, R, w, d* and *E* are the volume, surface area, radius of the largest inner circumference, radius of the lower outer circumference, minimum Feret diameter, maximum Feret diameter, and skeletal size, respectively.

Attributes Selection

After extraction was performed to select the features that best discriminate the nodule and non-nodule classes. For this, we used the selection method "Principal Component Analysis" - PCA with the ranking technique and default standards, both available in the WEKA software[10].

The descriptors proposed in this paper produces a feature vector for each analyzed structure. The PCA aims scoring determining factors in differentiating classes, propitiated by specific features. Thus select the best descriptors that help in differentiating between lung nodules other structures[11].

Classification

The classification process involves subjecting the features extracted from tissues found in exams to algorithms able to learn and infer about your distinction, so that, during analysis from peculiar characteristics to the classes studied (nodules and non-nodules) these algorithms are capable of classify a new object, forming a new information to be analyzed by an expert or a new condition for reduction of false positives in their stage from some CAD system.

The algorithms of which the characteristics set were sent will be presented in the results list, all followed the method of k-fold cross validation, which divides the data set 10 parts, uses the first nine for algorithm training and the latter to tests, repeating this process 10 times (k = 10) and presenting the results.

For this, we used the Waikato Environment tool for Knowledge Analysis (WEKA), developed by researchers at the University of Waikato, New Zealand, that consists of a tangle of machine learning techniques and data mining. With this, WEKA provides tools for pre-processing of data, classification, regression, clustering, feature selection, among others[10].

RESULTS AND DISCUSSION

The results of the classifiers are evaluated through the area under the Receiver-Operating Characteristic (ROC) curve, sensitivity (S), specificity (Sp) and accuracy (A)[12].

The sensitivity demonstrates how good is the methodology in the detection of disease, in this case the proportion of correctly classified nodules. Specificity shows how good the method is in the healthy image classification, or non-nodules, in this case. The Accuracy measures how good is the methodology to distinguish between two classes, which were true positives, measured by specificity, and false positives, measured by sensitivity.

The extracted features were submitted to the evaluation criteria through the following classifiers: Support Vector Machine - SVM; Multilayer Perceptrom - MLP; Randon Forest - RF; J48; IBK; and Simple Logistic - SL. All presents in the WEKA software.

It is possible to observe the results of multiple classifiers in Table 2.

Table 2 - Results using multiple classifiers

Classifiers	Results using multiple classifiers			
	ROC curve	S (%)	Sp(%)	A(%)
SVM	0.761	85.3	85.8	85.7
MLP	0.978	84.9	96.7	93.3
RF	**0.985**	**88.5**	**96.8**	**94.5**
J48	0.929	86.2	95.6	93.1
IBK	0.891	82.5	94.4	91.1
SL	0.975	85.5	96.1	93.1

The results presented in Table 2 demonstrate that a set of classifiers had satisfactory results, with all measurements of "S", "Sp" and "A" above 80%. Highlighting the RF, which obtained the best result in the classification context between lung nodes and non-nudes. This classifier, combined with implemented descriptors, had 88.5% sensitivity, 96.8% specificity, 94.5% of accuracy, 0.8616 Kappa index and 0.985 area under the ROC curve. In contrast, the MVS performed the worst, with 85.3% sensitivity, 85.8% specificity, 85.7% of accuracy and 0.761 area under the ROC curve.

After the selection of attributes with the PCA, the set of classifiers had their optimized results. Table 3 shows this result.

The results presented in Table 3 show that the selection of attributes optimized the results of the proposed methodology, and highlight the SVM and MLP that after the feature selection obtained results superiors to 90% on all valuation metrics. The best result was presented by the MLP, this classifier, combined with selection descriptors, had 91% sensitivity, 92.8% specificity, 92.3% of accuracy and 0.970 area under the ROC curve.

Table 3 - Results after attributes selection

Classifiers	Results after attributes selection			
	ROC curve	S (%)	Sp(%)	A(%)
SVM	0.928	90.9	94.7	93.7
MLP	**0.970**	**91.0**	**92.8**	**92.3**
RF	0.979	88.2	95.1	93.3
J48	0.934	82.2	94.9	91.4
IBK	0.888	84.1	93.1	90.7
SL	0.967	83.7	94.6	91.7

The comparative analysis between the proposed methodology results and related works analyzing characteristics of shape and/or texture for description of lung nodules presented in Table 4.

Table 4 - Comparative results between methods.

	Comparative results between methods			
	Image bases	S (%)	Sp(%)	A(%)
Fernandes et. al.[12]	LIDC-IDRI	87.94	94.32	91.05
Santos et. al.[13]	LIDC-IDRI	90.6	85	88.4
Carvalho Filho et. al.[6]	LIDC-IDRI	85.91	97.7	97.55
Proposed methodology	**LIDC-IDRI**	**91**	**92.8**	**92.3**

The methodology proposed by Fernandes et. al.[13], who used only characteristics of the form of solitary lung nodules, obtained lower results in the sensitivity, specificity and accuracy evaluation metrics, among the related work. This work, in turn, achieved significant results compared to Fernandes et. al.[13], bringing important contributions to the study of the subject.

This work focused on shape analysis of lung nodules to false positive reduction, demonstrating the exclusive differentiation capacity of these descriptors. This methodology makes it difficult to extract features, because encompasses parameters and measurements of great complexity calculation.

Given the comparative analysis shown in Table 4, we can score the methodology proposed as satisfactory as it brings significant results, with balanced metrics evaluation, besides the success when compared to other studies that too use only shape analysis, as Fernades et. al.[13] for example.

CONCLUSION

In this work, a method was proposed to reduce false positive in CT scans, using shape descriptors which relate to geometrical features of volume, surface area, spherical disproportion, Feret diameters and skeleton, as well as statistical measures related to the distances of the points the surface to the center of mass of an object.

Another relevant aspect of this work is the basis of images used, the LIDC-IDRI. Through it a great collection of exams and medical notes are able to provide the implementation, testing and validation of various techniques to aid in the detection and diagnosis of lung cancer.

The results showed that the morphological analysis is satisfactory and may have significant contributions to the CAD systems and consequently the lives of people facing these problems. For performing well in the characterization of lung nodules peculiarities and other tissues found in the CT scans.

For future works, we intend to test the generalizability of the proposed descriptors, submitting them to other bases of images, other contexts and other diseases or stages of CAD system. Furthermore, to enrich the set of features extracted, it is intended to add other morphological attributes. All additions and changes may optimize results, contributing to the effectiveness of the proposed methodology.

REFERENCES

1. National Cancer Institute - NCI. What is cancer? Bethesda, MD: NCI; 2015. Available from: https://www.cancer.gov/about-cancer/understanding/what-is-cancer
2. Instituto Nacional do Câncer José Alencar Gomes da Silva–INCA. Tipos de câncer: câncer de pulmão. Rio de Janeiro: INCA; 2016. Disponível em: http://www2.inca.gov.br/wps/wcm/connect/tiposdecancer/site/home/pulmao
3. Miranda GHB, Marques PMA, Felipe JC. Aplicação de conceitos da lógica nebulosa à classificação BI-RADS de nódulos de mama. J. Health Inform. 2009; 1(1):7-16.
4. Carvalho Filho AO, Silva AC, de Paiva AC, Nunes RA, Gattas M. 3D shape analysis to reduce false positives for lung nodule detection systems. Med Biol Eng Comput. 2017; 55(8):1199-1213.
5. Armato SG, McLennan G, Bidaut L, McNitt-Gray MF, Meyer CR, Reeves AP, et al. The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scan. Med Phys. 2011;38(2):915-31.
6. Carvalho Filho AO, Silva AC, Paiva AC, Nunes RA, Gattas M. Lung-nodule classification based on computed tomography using taxonomic diversity indexes and an SVM. J Sign Process Syst. 2017;87(2):179-96.
7. Carvalho Filho AO, de Sampaio WB, Silva AC, de Paiva AC, Nunes RA, Gattass M. Automatic detection of solitary lung nodules using quality threshold clustering, genetic algorithm and diversity index. Artif Intell Med. 2014;60(3):165-77.
8. Sampaio WB. Detecção de massas em imagens mamográficas usando uma metodologia adaptada à densidade da mama [dissertação]. Maranhão (MA): Universidade Federal do Maranhão – Programa de Pós-Graduação em Engenharia Eletrica; 2015.
9. Brito NM, Amarante Junior OP, Polese L, Ribeiro ML. Validação de métodos analíticos: estratégia e discussão. Pesticidas: R.Ecotoxicol. e Meio Ambiente 2003;13:129-46.
10. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann BPP, Witten IH. The WEKA data mining software: an update. ACM SIGKDD explorations newsletter. 2009;11(1):10-8.
11. Westad F, Hersleth M, Lea P, Martens H. Variable selection in PCA in sensory descriptive and consumer data. Food Quality and Preference, 2003; (14):463-72.
12. Chimieski BF, Fagundes RDR. Association and classification data mining algorithms comparison over medical datasets. J. Health Inform. 2013;5(2):44-51.
13. Fernandes VPM, Kanehisa RFA, Braz Júnior G, Silva AC, Paiva AC. Lung nodule classification based on shape distributions. In: Proceedings of the 31st Annual ACM Symposium on Applied Computing; 2016 Apr 4-8; Pisa, Italy: Association for Computing Machinery; 2016.
14. Santos AM, Carvalho Filho AO, Silva AC, Paiva AC, Nunes RA, Gattass M. Automatic detection of small lung nodules in 3D CT data using Gaussian mixture models, Tsallis entropy and SVM. Eng Appl Artif Intel 2014; 36:27-39.