Medical & Biological Engineering & Computing (2018) 56:2125–2136
https://doi.org/10.1007/s11517-018-1841-0

ORIGINAL ARTICLE

Classification of malignant and benign lung nodules using taxonomic diversity index and phylogenetic distance

Robherson Wector de Sousa Costa[1] · Giovanni Lucca França da Silva[1] · Antonio Oseas de Carvalho Filho[2] · Aristófanes Corrêa Silva[1] · Anselmo Cardoso de Paiva[1] · Marcelo Gattass[3]

Received: 10 July 2017 / Accepted: 23 April 2018 / Published online: 23 May 2018

Abstract

Lung cancer presents the highest cause of death among patients around the world, in addition of being one of the smallest survival rates after diagnosis. Therefore, this study proposes a methodology for diagnosis of lung nodules in benign and malignant tumors based on image processing and pattern recognition techniques. Mean phylogenetic distance (MPD) and taxonomic diversity index (Δ) were used as texture descriptors. Finally, the genetic algorithm in conjunction with the support vector machine were applied to select the best training model. The proposed methodology was tested on computed tomography (CT) images from the Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI), with the best sensitivity of 93.42%, specificity of 91.21%, accuracy of 91.81%, and area under the ROC curve of 0.94. The results demonstrate the promising performance of texture extraction techniques using mean phylogenetic distance and taxonomic diversity index combined with phylogenetic trees.

Keywords Medical image · Lung nodules diagnosis · Phylogenetic tree · Mean phylogenetic distance · Taxonomic diversity index

✉ Robherson Wector de Sousa Costa
robhersonwector@gmail.com

Giovanni Lucca França da Silva
gioh.lucca@gmail.com

Antonio Oseas de Carvalho Filho
antoniooseas@gmail.com

Aristófanes Corrêa Silva
aricsilva@gmail.com

Anselmo Cardoso de Paiva
anselmo.c.paiva@gmail.com

Marcelo Gattass
mgattass@tecgraf.puc-rio.br

[1] Federal University of Maranhão - UFMA, Applied Computing Group - NCA, Av. dos Portugueses, SN, Campus do Bacanga, Bacanga, São Luís, MA, 65085-580, Brazil

[2] Federal University of Piauí - UFPI, Rua Cícero Duarte, SN, Campus de Picos, Junco, Picos, PI, 64600-000, Brazil

[3] Pontifical Catholic University of Rio de Janeiro - PUC-Rio, Rua São Vicente, 225, Gávea, Rio de Janeiro, RJ, 22453-900, Brazil

1 Introduction

Lung cancer is the most commonly occurring malignant tumor and is characterized by an annual incidence increase of 2%. It is strongly associated with tobacco use. Annually, the number of deaths from lung cancer exceeds the total number of deaths from colorectal, breast, and prostate cancers [1].

A lung nodule is characterized as a rounded opacity in the lung with a diameter less than 3 cm, surrounded by lung parenchyma [2]. Lung lesions with diameters exceeding 3 cm are considered to be malignant masses [3]. Early diagnosis and treatment of lung cancer increases the patient probability of survival by 90% [4]. Thus, medical images comprising mainly of computed tomography (CT) present important tools for precocious diagnosis [5]. However, the detection of nodules on the areas of CT images is not an easy task since the densities of the nodules may be similar to that of the other lung structures. Additionally, the nodules may be characterized by low contrast and small sizes in complex anatomic regions, and they could be close or joined to blood vessels or the lung border [6].

For these reasons, a precocious diagnosis increases the probability of curing the patient. Furthermore, an increase in the amount of information available to the expert physician increases the precision of the diagnosis. In the last decade, there was a considerable increase in research on developing and using digital image processing techniques in CT with the primary goal of increasing the accuracy of the diagnosis by providing the expert physician with a second opinion. Accordingly, various computer-aided diagnosis (CADx) systems have been developed.

This study used concepts from ecology and biology to process medical images. Diversity indexes such as the Simpson index, Shannon index, Mcintosh index, Taxonomic indexes, and Menhinick phylogenetic tree were used by previous studies [7–15]. This study investigated other representations and measures for texture analysis such as phylogenetic trees and mean phylogenetic distance.

Several CADx studies by others have been conducted in the area of lung nodule diagnosis [16–30]. These studies utilized texture features (e.g., co-occurrence matrix and wavelets) and/or shape features (e.g., sphericity and skeleton). Additionally, they used different machine learning techniques (e.g., self-organizing maps, artificial neural network, support vector machine and, deep learning models) to classify the nodules as benign or malignant. The mean accuracy was approximately 89%. However, all these studies used sample databases with small sizes (number of samples) and low complexity. Thus, there is a need for more tests and more complex exams to validate proposed methods.

The proposed methodology in this study used only texture features for lung nodules classification. The mean phylogenetic distance and taxonomic diversity index were used to describe the texture of benign and malignant nodules. These indexes were based on phylogenetic distance, which involves the architecture of a rooted tree in the form of an inclined cladogram. Particularly, this study contributed to the computer science field in the area of extracting texture features based on mean phylogenetic distance and taxonomic diversity index by characterizing benign and malignant nodules.

This paper is organized as follows. Section 2 presents the methodology used to classify the nodule extracted from CT as benign and malignant. In Section 3, the results achieved through the proposed methodology are presented and discussed in the Section 4. Finally, the Section 5 presents final remarks about this study.

2 Method

This section presents the methodology proposed in this study. The methodology is organized in five stages as described by Fig. 1. The first stage details the materials used as images of CT exams in the LIDC-IDRI database. Then, the marking made by experts were used to perform the nodules segmentation. The markings were available for each image. This is followed by the features extraction stage that uses mean phylogenetic distance (MPD) and taxonomic diversity index (Δ). The classification includes a stage to find a best training model by using a genetic algorithm and support vector machine. Finally, the results are evaluated.

2.1 Materials

The image database used in this work is the LIDC-IDRI [31], which is available on the internet. This database is the result of an association between the Lung Image Database Consortium and the Image Database Resource Initiative, and includes 1,018 CT exams. However, a few (i.e., 185) CT exams were inappropriate for this methodology because of two factors. The first factor involved exams that did not present nodules equivalent to or exceeding 3 mm. The second factor included the divergence of information found in the marking file of an exam versus the information present in the DICOM header of the same exam. Hence, this invalidated the marking [11]. Therefore, the proposed methodology was applied to 833 exams.

The CT exams were acquired from different tomographers. This difference will render difficult the work of lung nodules classification. For each exam, this database includes a file in XML format that contains markings of its nodules. The analyses in the XML files were made by four experts, which, apart from indicating the contour of each nodule, also include certain characteristics such as roundness, texture, malignancy, etc., indicated by a value from 1 to 5.

To compose the image database for tests of the proposed method, we follow the method proposed in [32]; this was because there is no imposition for consensus because all nodules indicated by expert's review are ascertained and saved. In this way, it is possible for the same nodule to have different diagnoses. Based on this, it is considered to compose this image database, only one instance per nodule, aiming to minimize the impact of subjectivity in the exams. However, there is no indication in the annotation of experts (XML file) about which information refer to the same nodule. To work around this issue, we calculate the center point of the nodules, subsequently checking that the coordinates of this point are in the region of a nodule determined by another expert. Figure 2 illustrates the process.

The colored lines in Fig. 2 represent tags individually defined by experts. The green square in the center refers to the centroid calculated for the contour of the same color. If the coordinates of this centroid are in the areas delimited

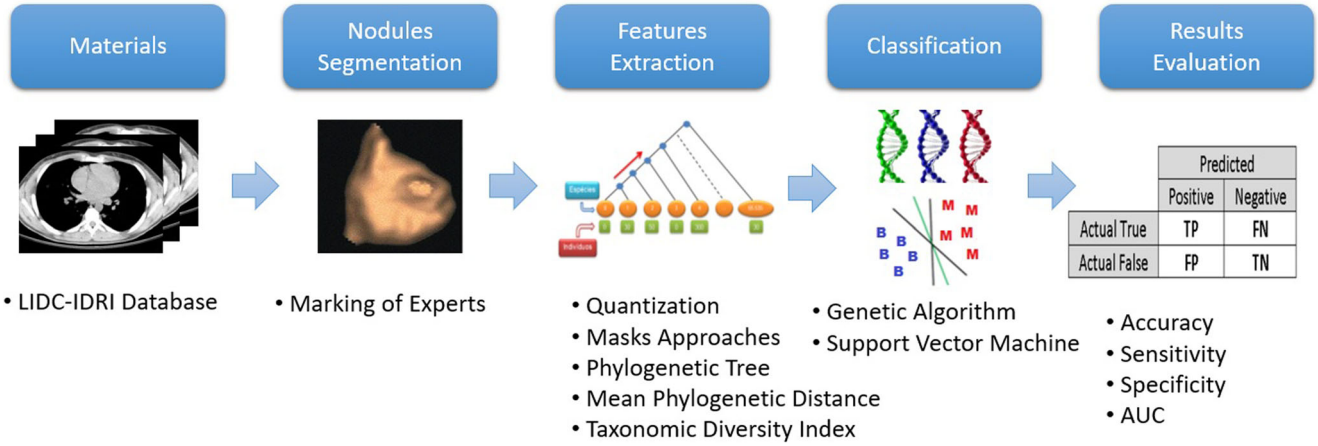


Fig. 1 Stages of the proposed methodology

by other expert, it is considered in this work that it is the same nodule, and therefore, only one instance should be accepted, which is that marking of the nodule that has the largest contour area. In the simplified example of Fig. 2, the instance of the nodule to be used will be that represented by the green contour, because it meets this criterion. After calculating which nodules have been noted by more than one expert and selecting which of the corresponding instances to be used, a summary of the diagnosis is made about malignancy or benignity. Diagnostic evaluation is already present for each nodule except for the LIDC-IDRI database, on a five-level subjective scale ranging from highly suspected benignity to highly suspected malignancy, i.e., the closer to 1, the more biased to benignity and the closer to 5, the greater the probability of malignancy. To summarize this information, the results of the previous step are used to calculate the results as presented in [32], where values of the characteristics belonging to the same nodule are reduced to a single value through the calculation of mode or median. The median is only used to summarize the value of the degree of malignancy when the method repeats itself. This process is illustrated in Fig. 3, highlighting with the rectangle the malignancy characteristic that will be used in the composition of the image database.

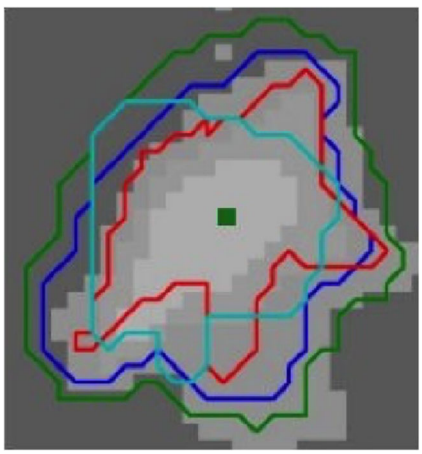


Fig. 2 Abstract illustration of nodules [32]

2.2 Nodules segmentation

In order to segment the nodules, information was obtained from the outline of a module from an XML file containing the nodule coordinates with criteria analysis from each expert. However, in the segmentation utilized in this study (as summarized in Section 2.1), only the bigger bound was used to represent the instance of the nodules described by a maximum of four markings by expert physicians.

2.3 Features extraction

The features extraction stage followed the nodules segmentation. This process involved four steps, namely, quantization, masks approaches, phylogenetic tree building, and diversity indexes computation.

The first step included the quantization of the volume of interest (VOI) by generating three versions of an image file with 16 (original), 12 , and 8 bits.

Then, in order to identify diversity patterns in the areas close to the border of the regions and in the inner areas, two approaches for sub-region definition referred to as internal mask and external mask were used [12]. These regions were generated through masks as binary images. The first internal mask was created with the binarization of the quantized volume of interest (VOI), and the posterior internal masks were based on successive reductions of the scale of the VOI with respect to the first one, while maintaining the center of mass. The successor masks were acquired from their previous mask following to the most internal. We defined a

Expert	Lob.	Mal.	Marg.	Spher.	Spic.	Subt.	Text.
A	4	5	4	2	4	3	4
B	3	4	4	4	3	5	5
C	2	4	3	4	3	4	5
D	3	4	2	2	4	3	3
Summarized	3	4	4	3	3	3	5

Fig. 3 Example of summary of nodule diagnosis

value of 20% for the diminution of scale, as it was verified in tests that the best results were achieved using five image masks with this scaling proportion.

The external masks were determined based on the difference between two internal masks. That is, the first external mask was determined by the difference between the first and the second internal masks, and so on. These approaches generated 9 masks (5 internal and 4 external) for each nodule. This allowed a more detailed diversity analysis with respect to each nodule. Figures 4 and 5 show examples of internal and external masks construction in 2D, respectively.

2.3.1 Phylogenetic tree

In the study of ecology communities, species variability is demonstrated by diversity and includes the relative significance of each species. Diversity is a key attribute in community studies, and therefore there are several methods to measure it. A more straightforward way to measure the diversity is by using species abundance, which consists of the number of species that exists in a specific community or interest area [33].

The phylogenetic diversity is a measure of community diversity that incorporates the phylogenetic relations between species. The main assumption of phylogenetic diversity is that the diversity is greater when species are phylogenetically more distinct [33]. A phylogenetic tree is a tree in which leaves represent the species and internal nodes represent the ancestors of the leaves. The tree edges linking nodes represent evolutionary relations [34]. In order to use these concepts to compare image patterns, it is necessary to map the biologic concepts to images as shown in Table 1.

A mean phylogenetic distance (MPD) is a value that describes the general structure of the phylogenetic community by analyzing all combinations of species pairs [35]. This is defined by Eq. 1 in the following expression:

$$\mathrm{MPD} = \frac{\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} d_{ij}\, p_i\, p_j}{\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} p_i\, p_j} \tag{1}$$

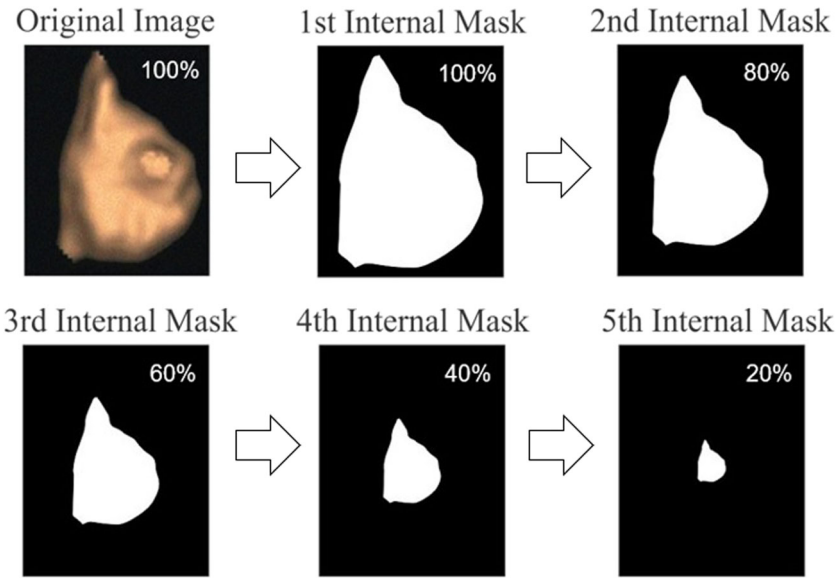


Fig. 4 Internal masks approach

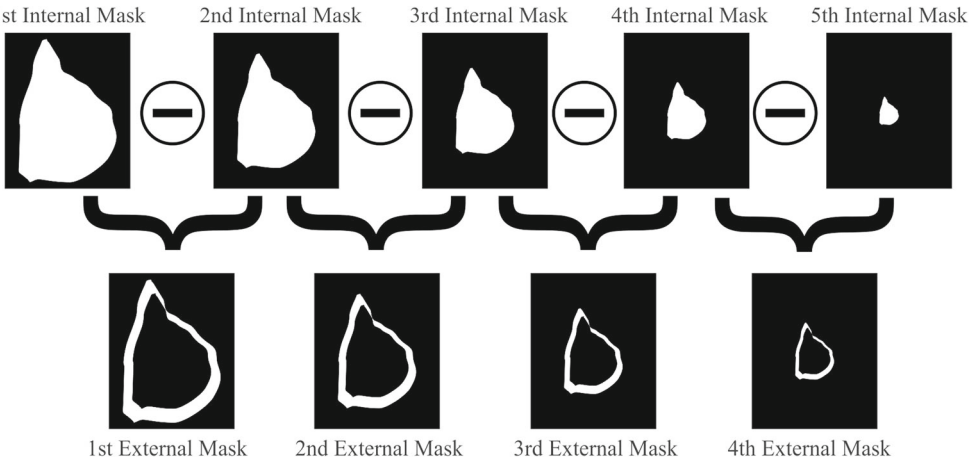


Fig. 5 External masks approach

where N denotes the total number of species in the community (maximum Hounsfield value found in a nodule), d_{ij} denotes phylogenetic distance from species i to species j in the taxonomic classification, and $p_i\,p_j$ is equal 0 when a species is absence and 1 if the species is present.

The taxonomic diversity index (Δ) considers the species abundance and the taxonomic relations between them. Its value represents the mean taxonomic distance between two individuals (voxel values), chosen in the sample [35]. This may be expressed by Eq. 2 given below:

$$\Delta = \frac{\sum\sum_{i<j} d_{ij}x_i x_j}{[n(n-1)/2]} \tag{2}$$

where $x_i (i = 0, ..., s)$ is the abundance (number of voxels) of the ith specie, $x_j (j = 0, ..., s)$ is the abundance of the jth specie, s represents the number of species, n is the total number of individuals and d_{ij} is the phylogenetic distance from species i to species j in the taxonomic classification.

2.3.2 Rooted tree as an inclined cladogram

The phylogenetic trees for each region were generated after the internal and external mask generation. The indexes (MPD and Δ) used in this study were based on phylogenetic trees with three essential aspects, namely, number of species, number of individuals, and species connections (edges number). The rooted tree was used as an inclined cladogram to represent the images.

Figure 6 illustrates the phylogenetic tree (Tree 1) in which the species correspond to the Hounsfield values in the interval $[-32, 768, +32, 768]$. This interval was translated to $[0, +65, 536]$ in the implementation, with the goal of making the index calculations simpler.

The relation between species in Tree 1 (Fig. 6) is considered from left to right as indicated by the arrow. Thus, the first relation is between species 0 and 1, which have two edges linking them (Fig. 7a). The second relation has three edges linking species 0 and 2 (Fig. 7b). The final relation has 65,536 edges between species 0 and 65,535 (Fig. 7d).

2.3.3 Rooted tree as an inclined cladogram excluding absent species

An additional tree architecture was developed in Tree 2, and it excluded species without individuals in the region. This generated a re-organization of the tree and of the edges. For

Table 1 Mapping biologic concepts to image processing

Biology	Image processing
Community	CT volume of interest
Species	Hounsfield value (voxel value)
Species richness: number of species found in a region	Species richness: number of distinct voxels values in a region
Individual	VOI voxel
Relative abundance: number of individuals of a specific species that exists in a region	Relative abundance: number of voxels with a specific value found in a region

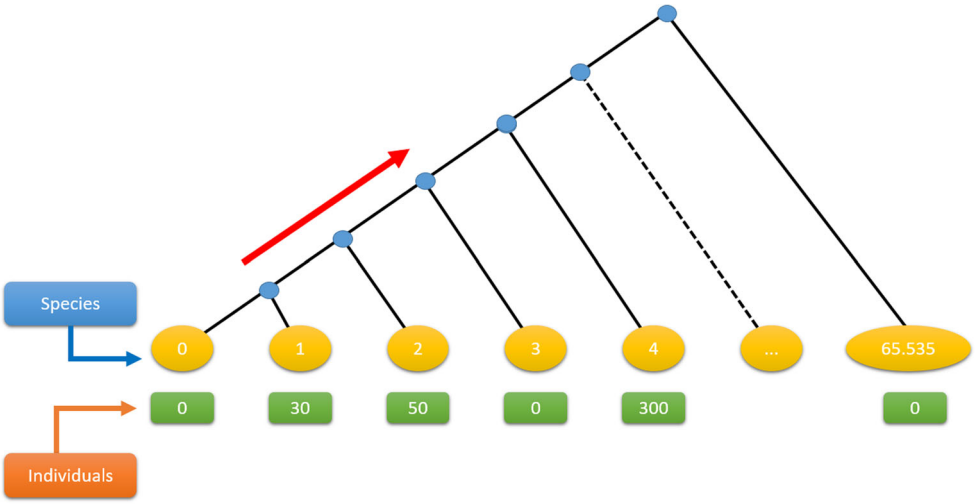


Fig. 6 Tree 1 – rooted tree as an inclined cladogram

example, the tree in Fig. 6 only had individuals from species 1, 2, and 4. The tree is represented as shown in Fig. 8.

2.3.4 Rooted tree as an inclined cladogram with modified edges

The third proposed tree (Tree 3) had the same form of the combination of species as Tree 1. The only difference between the trees is in the number of edges between the species. A weight was added to increase the distance between the phylogenetic species, and then more details was provided to differentiate the species. Figure 9 depicts the same procedure of combining species as that in Fig. 6.

2.4 Classification

It is common to find methodologies in the literature that use techniques to select the most significant individuals to generate a training database [11, 13, 36]. Therefore, we chose to use the genetic algorithm (GA) in conjunction with the support vector machine (SVM) proposed by [13] to select the best nodules to generate the model that will be used in the classification, thus making sure that only the most significant nodules among malignant and benign are selected to create the training model. The algorithm can be summarized in the following steps:

1. From each nodule obtained during nodules segmentation stage, the set of all feature vectors forms the analysis base (AB).
2. Each element in array A1 contains the position of a feature vector extracted from a benign nodule in AB. Likewise, each element in array A2 contains the position of a feature vector extracted from a malignant nodule in AB.
3. The mutation and crossing genetic operators modify the values contained in A1 and A2. However, the values cannot be repeated. Elements of A1 cannot be present in A2 or vice versa.
4. For each individual of a generation, the SVM [37] is trained using the feature vectors selected in A1 and A2. The other feature vectors from AB, which were not selected, are used to validate the training model.

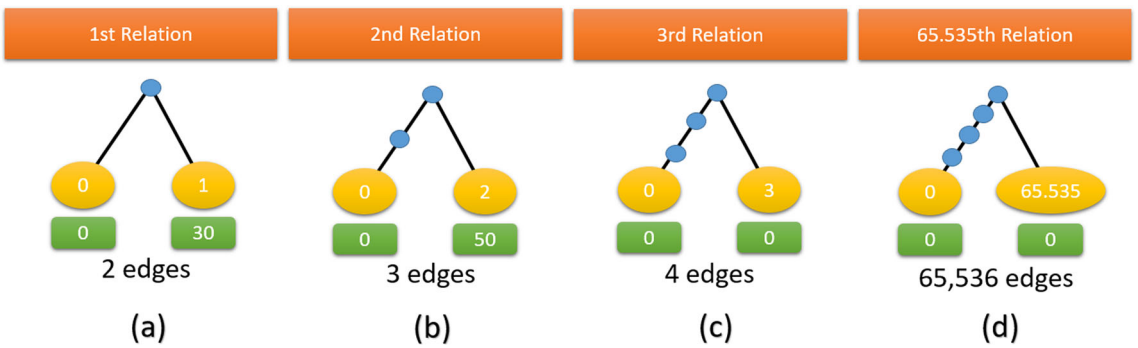


Fig. 7 Species relations

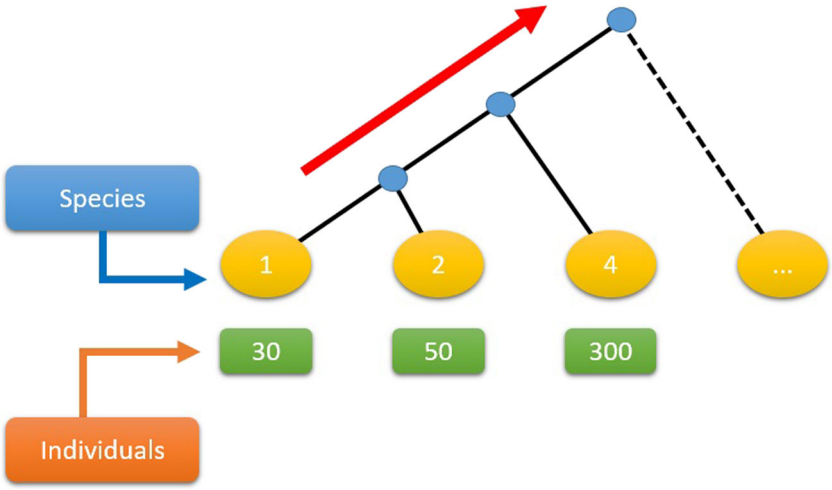


Fig. 8 Tree 2 — cladogram excluding absent species

5. The fitness of each individual is evaluated through the results obtained in the validation set. The fitness is defined by Eq. 3. A greater weight is given to the metric sensitivity, in order to get models with a high capacity to malignant nodules correctness.

$$Fitness = (3 * Sensitivity) + Specificity + Accuracy \quad (3)$$

This process is repeated until the fitness of the best individual is the same for 100 consecutive generations.
6. At the end of the evolution, the best training model is formed by all the feature vectors whose positions are contained in arrays A1 and A2 of the individual with the best fitness in the last generation.
7. Finally, the selected model is used for testing. In order to measure how good is the model, sensitivity, specificity, accuracy, and area under the ROC curve (AUC) are computed. Figure 10 presents a general scheme of how the choice is made by the GA in conjunction with the SVM.

3 Results

This section presents the results obtained with the proposed methodology with reference to the lung nodules diagnosis on CT scans. The methodology was implemented in C++ language and ITK software, running on a machine with an Intel Core i7 CPU at 3.07 GHz processor, 8 GB of RAM and Windows 7 operation system.

3.1 Database separation

There were 833 segmented exams corresponding to 1,405 nodules (1,011 benign and 394 malignant) from the LIDC-IDRI exams database. These nodules were randomly separated in two sets, with 80% (1,124) for training and

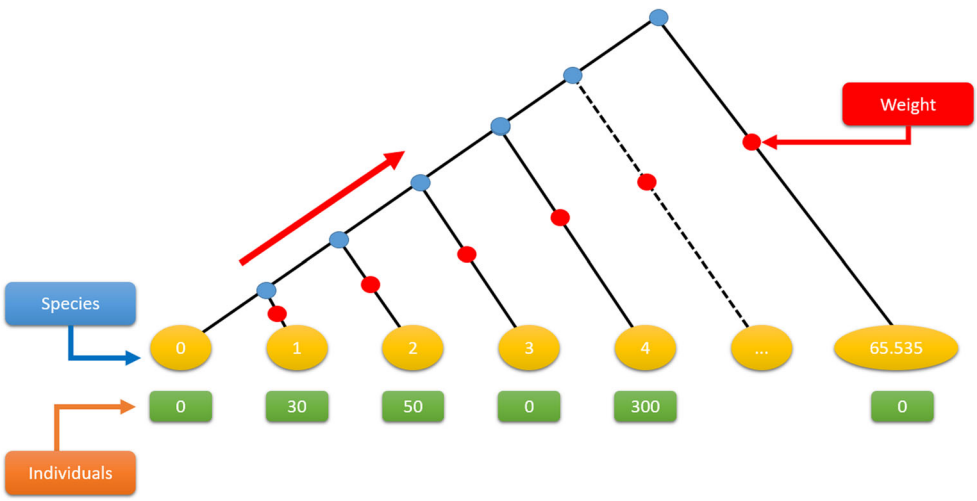


Fig. 9 Tree 3 — cladogram with modified edges

Fig. 10 Best model selection using GA and SVM

validation and 20% (281) for the testing. This division maintained the global proportion of malignant and benign nodules in the total set.

Table 2 shows the diameters of the nodules that composes the database. As can be seen in Table 2, the image database has several nodule sizes, allowing the proposed methodology to be evaluated based on the most diverse nodule stages.

3.2 Features extraction

Three representations of gray levels were used for each nodule (original image with 16 bits and two other versions with 8 and 12 bits). Furthermore, the nodules were subdivided in 9 masks based on the application of internal and external masks. Thus, 27 sub-volumes (3 representations × 9 masks) were used for each nodule to calculate two diversity indexes, MPD and Δ. Figure 11 details this process. The GA and SVM used a total of 54 features (27 sub-volumes × 2 indexes) for each tree architecture described in Section 2.3.

3.3 Classification

The training set (1,124) was submitted to GA for the selection of the best training model. This was performed by generating a balanced model with 500 nodules (250 malignant nodules and 250 benign nodules) for training and the rest of the individuals (65 malignant nodules and 559 benign nodules) to validate the selected model. This procedure was applied to all the experiments conducted in our methodology. Table 3 shows the results obtained in the selection of the best training model for each tree.

Table 4 presents the results obtained using the selected model by GA in conjunction with the SVM applied to the testing set (79 malignant nodules and 202 benign nodules).

As observed in the test experiments results, shown in Table 4, tree 2 architecture presented higher values for almost all the metrics, with all the values exceeding 91%. Accordingly to [33], these results corresponded to the findings of previous studies wherein prior knowledge with respect to the diversity of a specific region is fundamental in a better understanding of the region nature. This tree architecture only represented the species with individuals in the studied region. Thus, phylogenetic relationships with respect to the species that were absent were excluded from the tree. In contrast, the relationship was depicted for all possible species in tree 1 and 3 architectures. Tree 3 architecture presented a slightly worse result as it also added weights to the relations making the species more distinct from a phylogenetic point of view.

Table 2 Size of nodules

Nodule	Size/diameters		
	Up to 10 mm	Up to 20 mm	Up to 30 mm
Benign	359	488	164
Malignant	67	155	172

4 Discussion

Table 5 shows a comparison between the results found on this study and some of the related works. It is important to emphasize that to perform a reliable comparison with these previous works, it would be necessary to use the same image database, same training and test exams, and same settings for the classifiers, among other parameters.

Comparing the best results achieved in our study (Tree 2) with those presented in Table 5, it is possible to see that our results are promising. We achieved results above 91% for two types of situation: (1) diagnosis using only

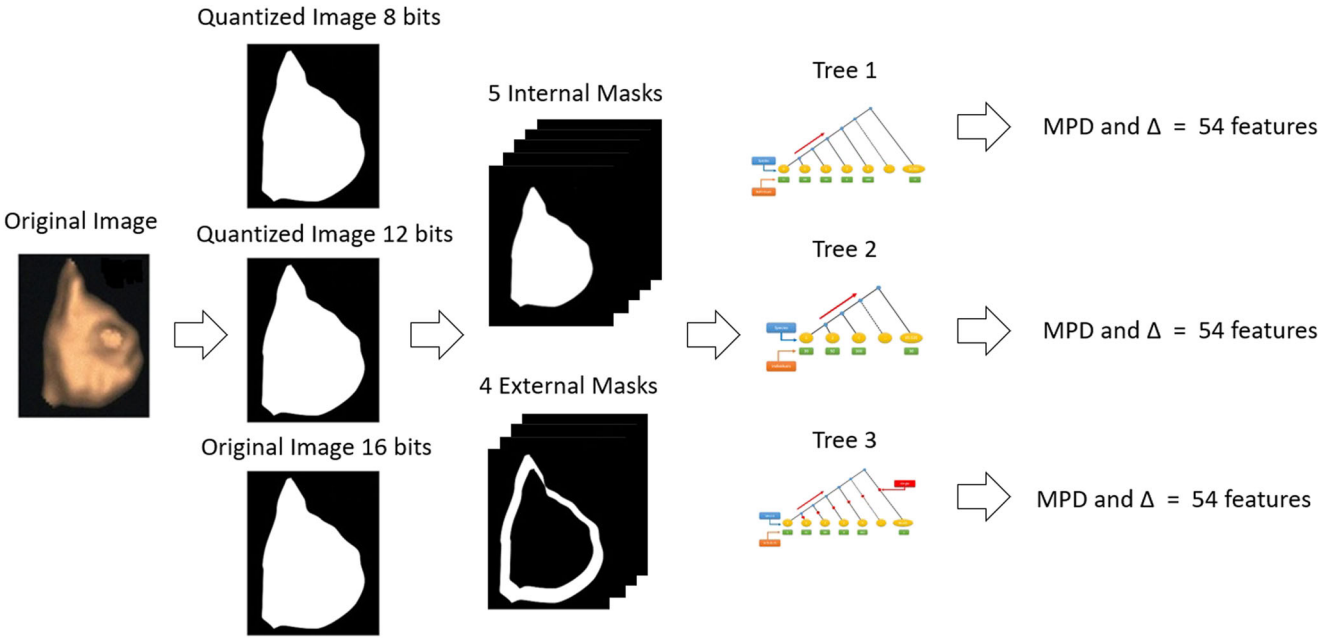


Fig. 11 Features extraction

texture features and (2) a large and complex database. The works of [9, 11, 17, 21, 28] present accuracy better than our research; however, our research presents a better sensitivity. In terms of CADx system, the sensitivity is the metric more important, because it shows the model performance to classify correctly malignant nodules, allowing the fast medical intervention. The deep learning-based methodologies [26, 27, 29] use an amount of samples superior to ours, since the nodules were analyzed by slice (2D). Nonetheless, our methodology presents results that surpass these works. The other works mentioned in Table 5 present lower results and smaller image database than ours. The conclusion obtained indicated the following:

1. The use of diversity indexes combined with phylogenetic trees were promising for the characterization of lung nodule textures. Previous studies [7, 8, 10–12, 14] indicated that the diversity index is a good texture descriptor. This study added to species diversity concepts by using phylogenetic trees as representations and the mean phylogenetic distance (MPD).
2. The use of uniform quantization to represent the image at different gray scale levels (8 and 12 bits, besides the original image) produced better results than using only the original image (16 bits).
3. Previous studies on lung nodules diagnosis used the analysis of shape and texture [38, 39] to achieve good results. This study achieved promising results by only analyzing the texture features. Thus, the addition of more shape features to the proposed methodology will yield better results.
4. Furthermore, this study illustrated that the concepts proposed by [12] wherein a region-based analysis could bring more information to the classification process and provide more subsidies to the classifier to obtain the correct decision.
5. The use of GA in conjunction with the SVM to select a set of individuals from a sample to generate the best training model was valid in keeping with the results of previous studies by [11] and [13]. Besides, the weight attributed to metric sensitivity obtained models with better performance to classify correctly malignant nodules.
6. Finally, it is important to highlight that the LIDC-IDRI database is extremely complex and diverse, containing

Table 3 Best result obtained by GA and SVM in the validation set

Tree architecture	Accuracy%	Sensitivity%	Specificity%	Fitness
1	95.51	100	94.91	490.43
2	94.71	100	94.06	488.77
3	86.53	95.77	85.35	459.21

Table 4 Results obtained by the proposed methodology in the testing set

Tree Architecture	Accuracy%	Sensitivity%	Specificity%	AUC
1	91.10	90.14	91.20	0.93
2	**91.81**	**93.42**	**91.21**	**0.94**
3	90.39	79.76	94.92	0.92

Table 5 Comparison with other researches with reference to the classification of lung nodules in whether malignant or benign

Work	Image database	Number of samples	Acc (%)	Sens (%)	Spec (%)	AUC
[25]	LIDC	—	74.1	—	—	0.69
[19]	Private	—	90	—	—	—
[22]	ELCAP/LI	—	—	86	97	—
[9]	LIDC	73	92.78	85.64	97.89	—
[17]	Private	—	94.44	—	—	—
[18]	LIDC	33	90.91	—	—	—
[20]	ELCAP and NBIA	128	84	—	—	—
[21]	JSRT	246	96	—	—	—
[23]	NBIA/ELCAP	—	82.66	96.15	52.16	—
[5]	Private	11	—	91.38	89.56	—
[16]	Private	128	90.63	92.3	89.47	—
[11]	LIDC-IDRI	—	97.55	85.91	97.70	—
[24]	LIDC	—	—	73.30	78.70	—
[26]	LIDC	2,545	—	73.30	78.70	—
[27]	LIDC	4,323	75.01	83.35	—	—
[28]	LIDC	10,133	95.6	92.4	98.9	0.989
[29]	LIDC	174,412	81.11	—	—	—
Our method	LIDC-IDRI					
Tree 1		**1,405**	91.10%	90.14%	91.20%	0.93
Tree 2			**91.81**%	**93.42**%	**91.21**%	**0.94**
Tree 3			90.39%	79.76%	94.92%	0.92

countless different cases of lung nodules. This database has exams that were extracted by various tomography methods, leading to difficulty in the classification through CADx systems.

5 Conclusion

The high number of deaths caused by lung cancer evidence the importance of developing research aimed at precocious diagnosis. This could lead to a more appropriate treatment for patients, thereby increasing their chances of survival. Based on this, computational tools that provide a second opinion to expert physicians could also be beneficial to patients.

This study presented a methodology for lung nodules diagnosis using the mean phylogenetic distance and the taxonomic diversity index. These features were used in conjunction with different phylogenetic tree architectures and SVM for the classification of lung nodule as benign or malignant. Hence, it is a useful tool for expert physicians.

The obtained results indicated the promising performance of the proposed texture extraction techniques. The creation of a phylogenetic tree was another important factor that led to good results. The use of this tree contributed considerably in discriminating between benign and malignant nodules. Although the image database used in this study was robust and ensured a high diversity of the analyzed nodules, additional tests with other databases are necessary to improve the proposed methodology and to make it more robust and generic.

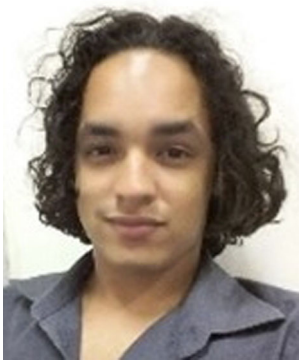
Acknowledgements The authors acknowledge the National Cancer Institute and the Foundation for the National Institutes of Health for their critical role in the creation of the free, publicly available LIDC-IDRI database used in this research.

Funding information This study is financially supported by CAPES and CNPq.

References

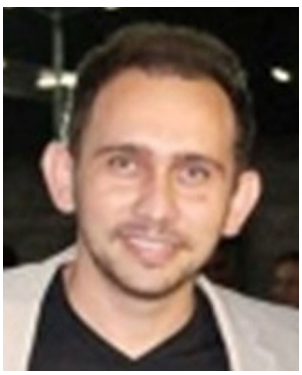
1. do Câncer (INCA) IN (2016) What is cancer? Available at http://www1.inca.gov.br/conteudo
2. Hansell DM, Bankier AA, MacMahon H, McLoud TC, Muller NL, Remy J (2008) Fleischner society: glossary of terms for thoracic imaging. Radiology 246:697–722. PMID: 18195376
3. Fujimoto J, Wistuba II (2014) Current concepts on the molecular pathology of non-small cell lung carcinoma. Seminars Diag Pathol 31:306–313. Lung Carcinoma: Beyond the {WHO} Classification

4. Lederlin M, Revel MP, Khalil A, Ferretti G, Milleron B, Laurent F (2013) Management strategy of pulmonary nodule in 2013. Diag Intervention Imag 94:1081–1094
5. Naidich DP, Webb RW, Muller NL et al (2001) Computed tomography and magnetic resonance of the thorax. Thorax 56:898
6. Leef JL, Klein JS (2002) The solitary pulmonary nodule. Radiol Clin North Am 40:123–143
7. Nunes AP, Silva AC, de Paiva AC (2009) Detection of masses in mammographic images using Simpson's diversity index in circular regions and svm. In: Perner P (ed) Proceedings of the 6th international conference on Machine learning and data mining in pattern recognition, MLDM 2009, Leipzig, Germany, July 23-25, 2009. Springer, Berlin, pp 540–553
8. de Sousa Carvalho PM, de Paiva AC, Silva AC (2012) Classification of breast tissues in mammographic images in mass and non-mass using mcintosh's diversity index and SVM. In: Proceedings of the 8th international conference on machine learning and data mining in pattern recognition, MLDM 2012, Berlin, Germany, July 13-20, 2012, pp 482–494
9. Nascimento LB (2012) Lung nodules classification in ct images using Shannon and Simpson diversity indices and svm. Mach Learn Data Min Pattern Recogn 7376:454–466
10. da Rocha SV, Junior GB, Silva AC, de Paiva AC (2014) Texture analysis of masses in digitized mammograms using Gleason and Menhinick diversity indexes. Revista Brasileira de Engenharia Biomédica 30:27–34
11. Filho AOC, de Sampaio WB, Silva AC, de Paiva AC, Nunes RA, Gattass M (2014) Automatic detection of solitary lung nodules using quality threshold clustering, genetic algorithm and diversity index. Artif Intell Med 60:165–177
12. de Oliveira FSS, Filho AOC, Silva AC, de Paiva AC, Gattass M (2015) Classification of breast regions as mass and non-mass based on digital mammograms using taxonomic indexes and SVM. Comput Biol Med 57:42–53
13. de Sampaio WB, Silva AC, de Paiva AC, Gattass M (2015) Detection of masses in mammograms with adaption to breast density using genetic algorithm, phylogenetic trees, LBP and SVM. Expert Syst Appl 42:8911–8928
14. Filho AOC, Corrêa A, de Paiva AC, Nunes RA, Gattass M (2016) Lung-nodule classification based on computed tomography using taxonomic diversity indexes and an svm. J Signal Process Sys, pp 1–18
15. Silva GLFd, Carvalho Filho AOd, Silva AC, Paiva ACd, Gattass M (2016) Taxonomic indexes for differentiating malignancy of lung nodules on ct images. Res Biomed Eng 32:263–272
16. Dandil E, Cakiroglu M, Eksi Z et al (2014) Artificial neural network-based classification system for lung nodules on computed tomography scans. In: international conference of soft computing and pattern recognition, pp 382–386
17. Kannan A (2012) Computer-aided diagnosis of lung cancer based on analysis of the significant slice of chest computed tomography image. IET Image Process 6:697–705(8)
18. Krewer H, Geiger B et al (2013) Effect of texture features in computer aided diagnosis of pulmonary nodules in low-dose computed tomography. In: IEEE International Conference on systems, man, and cybernetics, pp 3887–3891
19. Kumar SA, Ramesh DJ et al (2011) Robust and automated lung nodule diagnosis from ct images based on fuzzy systems. In: 2011 international conference on process automation, control and computing (PACC), pp 1–6
20. Orozco HM, Villegas OOV et al (2013) Lung nodule classification in ct thorax images using support vector machines. In: 2013 12th Mexican international conference on artificial intelligence (MICAI). IEEE, pp 277–283
21. Al-Absi HR, Samir BB, Shaban KB, Sulaiman S (2012) Computer aided diagnosis system based on machine learning techniques for lung cancer. In: 2012 international conference on computer & information science (ICCIS), vol 1. IEEE, pp 295–300
22. Farag A, Ali A, Graham J, Farag A, Elshazly S, Falk R (2011) Evaluation of geometric feature descriptors for detection and classification of lung nodules in low dose ct scans of the chest. In: 2011 IEEE international symposium on biomedical imaging: from nano to macro. IEEE, pp 169–172
23. Orozco HM, Villegas OOV, Maynez LO, Sánchez VG, De Jesus Ochoa Dominguez H (2012) Lung nodule classification in frequency domain using support vector machines. In: 2012 11th international conference on information science, signal processing and their applications (ISSPA). IEEE, pp 870–875
24. Akram S, Javed MY, Hussain A, Riaz F, Akram MU (2015) Intensity-based statistical features for classification of lungs ct scan nodules using artificial intelligence techniques. J Exper Theor Artif Intell 27:737–751
25. Zinovev D, Feigenbaum J, Furst J, Raicu D (2011) Probabilistic lung nodule classification with belief decision trees. In: Annual international conference of the IEEE engineering in medicine and biology society, EMBC, pp 4493–4498
26. Hua KL, Hsu CH, Hidayati SC, Cheng WH, Chen YJ (2015) Computer-aided classification of lung nodules on computed tomography images via deep learning technique. OncoTargets Therapy 8:2015–2022
27. Kumar D, Wong A, Clausi DA (2015) Lung nodule classification using deep features in ct images. In: 2015 12th conference on computer and robot vision (CRV). IEEE, pp 133–138
28. Cheng JZ, Ni D, Chou YH, Qin J, Tiu CM, Chang YC, Huang CS, Shen D, Chen CM (2016) Computer-aided diagnosis with deep learning architecture: applications to breast lesions in us images and pulmonary nodules in ct scans. Sci Rep, 6
29. Sun W, Zheng B, Qian W (2016) Computer aided lung cancer diagnosis with deep learning algorithms. In: SPIE medical imaging, international society for optics and photonics, pp 97850Z–97850Z
30. da Silva GL, da Silva Neto OP, Silva AC, de Paiva AC, Gattass M (2017) Lung nodules diagnosis based on evolutionary convolutional neural network. Multimed Tools Appl, pp 1–17
31. Armato SG, Geoffre M et al (2011) The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. Med Phys 38:915–31
32. Jabon SA, Raicu DS, Furst JD (2009) Content-based versus semantic-based retrieval: a lidc case study. SPIE Medical imaging. Int Soc Opt Photon, 72631L–72631L
33. Melo AS (2008) What do we win 'confounding' species richness and evenness in a diversity index? Biota Neotropica, 8
34. Clarke KR, Warwick RM (1998) A taxonomic distinctness index and its statistical properties. J Appl Ecol 35:523–531
35. Webb CO (2000) Exploring the phylogenetic structure of ecological communities: an example for rain forest trees. Am Nat 156:145–155
36. de Sousa JA, de Paiva AC, de Almeida JDS, Silva AC, Junior GB, Gattass M (2017) Texture based on geostatistic for glaucoma diagnosis from fundus eye image. Multimed Tools Appl, 1–18
37. Duda RO, Hart PE (1973) Pattern classification and scene analysis. Wiley-Interscience Publication, New York
38. Hardie RC, Rogers SK, Wilson T, Rogers A (2008) Performance analysis of a new computer aided detection system for identifying lung nodules on chest radiographs. Med Image Anal 12:240–258
39. Jing Z, Bin L, Lianfang T (2010) Lung nodule classification combining rule-based and svm. In: 2010 IEEE fifth international conference on bio-inspired computing: theories and applications (BIC-TA), pp 1033–1036



Robherson Wector de Sousa Costa currently, graduating in Computer Science from the Federal University of Maranhão. He works with image processing, specifically on medical imaging.

Giovanni Lucca França da Silva received MSc in Computer Science from Federal University of Maranhão, Brazil in 2017. Currently, a PhD student in electric engineering at Federal University of Maranhão. He works with image processing, specifically on medical imaging.



Antonio Oseas de Carvalho Filho received PhD in electric engineering from Federal University of Maranhão in 2016. Currently, he is a Professor at the Federal University of Piauí (UFPI), Brazil. He works with image processing, specifically on medical imaging.

Aristófanes Corrêa Silva received PhD in informatics from Pontifical Catholic University of Rio de Janeiro, Brazil in 2004. Currently, he is a Professor at the Federal University of Maranhão (UFMA), Brazil. He works with machine learning and image processing, specifically on medical imaging.

Anselmo Cardoso de Paiva received PhD in informatics from Pontifical Catholic University of Rio de Janeiro Brazil in 2002. Currently, he is a Professor at the Federal University of Maranhão (UFMA), Brazil. He works with computer graphics and image processing, specifically on medical imaging.

Marcelo Gattass received PhD in computer science from Cornell University - USA in 1982. Currently, he is a Professor at the Pontifical Catholic Universityof Rio de Janeiro. He works with computer graphics, 3D computer vision, geographic information systems, user interfaces, and web-based applications.