J Digit Imaging (2017) 30:812–822
DOI 10.1007/s10278-017-9973-6

Computer-Aided Diagnosis of Lung Nodules in Computed Tomography by Using Phylogenetic Diversity, Genetic Algorithm, and SVM

Antonio Oseas de Carvalho Filho[1] · Aristófanes Corrêa Silva[1] · Anselmo Cardoso de Paiva[1] · Rodolfo Acatauassú Nunes[2] · Marcelo Gattass[3]

Published online: 19 May 2017

Abstract Lung cancer is pointed as the major cause of death among patients with cancer throughout the world. This work is intended to develop a methodology for diagnosis of lung nodules using images from the Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI). The proposed methodology uses image processing and pattern recognition techniques. In order to differentiate between the patterns of malignant and benign nodules, we used phylogenetic diversity by means of particular indexes, that are: intensive quadratic entropy, extensive quadratic entropy, average taxonomic distinctness, total taxonomic distinctness, and pure diversity indexes. After that, we applied the genetic algorithm for selection of the best model. In the tests' stage, we applied the proposed methodology to 1405 (394 malignant and 1011 benign) nodules. The proposed work presents promising results at the classification into malignant and benign, achieving accuracy of 92.52%, sensitivity of 93.1% and specificity of 92.26%. The results demonstrated a good rate of correct detections using texture features. Since a precocious detection allows a faster therapeutic intervention, thus a more favorable prognostic to the patient, we propose herein a methodology that contributes to the area in this aspect.

✉ Antonio Oseas de Carvalho Filho
antoniooseas@gmail.com

Aristófanes Corrêa Silva
ari@dee.ufma.br

Anselmo Cardoso de Paiva
paiva@deinf.ufma.br

Rodolfo Acatauassú Nunes
rodolfoacatauassu@yahoo.com.br

Marcelo Gattass
mgattass@tecgraf.puc-rio.br

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

[2] Sao Francisco de Xavier, State University of Rio de Janeiro, 524, Maracana, 20550-900, Rio de Janeiro, RJ, Brazil

[3] Department of Computer Science, Pontifical Catholic University of Rio de Janeiro - PUC-Rio, R. Marquês de São Vicente, 225, Gávea, 22453-900, Rio de Janeiro, RJ, Brazil

Keywords Lung cancer · Phylogenetic diversity index · Genetic algorithm · Medical image

Introduction

It is estimated a yearly increase of 1.8 million new cases of lung cancer throughout the world, and lung cancer already corresponds to 13% of all the cancer cases all over the world. Out of the more than 100 types of cancer, lung cancer is responsible for the highest mortality rate [5, 32, 41]. A lung nodule is defined as a nearly spherical opacity with up to 3 cm in diameter, surrounded by the pulmonary parenchyma [18]. Lesions larger than 3 cm are called masses and are often malignant [15, 17, 18]. Most of lung cancer cases are related to smoking (representing around 80% of the cases), by aging of society, industrialization, urbanization, pollution, and bad lifestyle [5].

The most effective manner to defeat lung cancer is the early diagnosis and treatment. If precociously treated, the post-diagnosis survival rate increases about 90% [25]. One of the most effective diagnosing methods is the image exam. The computed tomography (CT) is an affordable exam which provides good-quality images, and which is used in the analysis of several lesion types, including lung lesions. However, the analysis of a CT is a sensitive task which demands too much of the expert, since it is a repetitive and tiring process, with high possibility of errors due to the large number of images to be analyzed [26, 26, 32].

The early detection of lung cancer allows an anticipated and faster therapeutic intervention, providing the patient with a more favorable prognosis [40]. Various computational tools that employ digital image processing and pattern recognition techniques have been constantly explored. Those techniques have been used together to develop computer-aided detection (CAD)/computer-aided diagnostic (CADx) systems [48]. Such tools aim at increasing the precision of diagnosis, providing the expert with a second opinion since additional information results in a more precise diagnosis.

In most CADx methodologies, the feature extraction stage is based on: (1) geometry, which measures, for example, how circular the candidate is, and (2) on texture, that describes aspects of the candidate based on its gray levels distribution. In order to characterize the lung nodules, we only used texture descriptors. For such, we employed phylogenetic diversity indexes.

Diversity is a term frequently used in ecology. The objective of a diversity index is to describe the variety of species present in a community or area [27]. Phylogeny is a branch of biology concerned with studying the evolutionary relationships between species, by verifying the relationships among them, in order to determine possible common ancestors. A phylogenetic tree, or simply a phylogeny, is a tree in which the leaves represent the organisms and the internal nodes represent possible ancestors. The edges of the tree denote the evolutionary relationships [6].

In order to characterize texture, we used five diversity indexes, namely: (1) the intensive quadratic entropy index, which represents the phylogenetic difference between two randomly chosen species; (2) the extensive quadratic entropy, which indicates the sum of all the distances between the pairs of species; (3) the average taxonomic distinctness, in which we have the distance between two randomly chosen species; (4) the total taxonomic distinctness, which allows us to verify the sum of the mean phylogenetic distinction between all the species; and finally (5) the pure diversity, which is computed as the distance between neighbor species. These indexes are based on the phylogenetic diversity (number of edges) from the structure of a tree rooted as a dendrogram.

Our work brings direct contributions to some fields. In the medical field, our contribution is the development of an automatic system for pulmonary diagnosis through the analysis of the texture of the nodule. In the computer science field, we contribute in the following aspects: (a) in the use of texture measures based on the intensive quadratic entropy, extensive quadratic entropy, average taxonomic distinctness, total taxonomic distinctness, and pure diversity indexes, and (b) in the use of phylogenetic trees for characterization of lung nodules.

Others studies [8, 10, 11, 23, 24, 28, 28, 30, 31, 44], have been conducted concerning the diagnosis of lung nodules as malignant and benign, with the goal of increasing accuracy rates of CADx systems for detection and diagnosis of lung cancer. Tables 4 and 5 present a recent literature review in the area of diagnosis of lung nodules, showing the following details: work, techniques, database, and results. Since it is a complex task, most of these works do not only use texture descriptors to classify lung nodules into malignant and benign. Instead, they combine texture and shape descriptors, using them as complementary analyses [10, 11, 24, 44]. In most of the nodules and non-nodules manually classified by the experts, delimiting a region is quite bigger than the real area/volume of the nodule or non-nodule, or does not match its real shape. Another problem is that most CAD systems uses segmentation techniques that generate candidate regions with very similar shapes [3, 29], which may lead to incorrect classifications. Besides these works, there are some other studies that the intratumor heterogeneity of the lung nodule, which allows, for example, based on the analysis of the texture of the lesion, to know which chemotherapeutic agent is more suitable [16]. For these reasons, our methodology employs only texture descriptors to characterize each lung nodule.

This paper is organized as follows. In "Materials and Methods" we present the methodology used to classify nodules extracted from CT into malignant and benign, using the texture-based extraction of features, selection of the best model by genetic algorithm and classification by of the Support Vector Machine. In "Results and Discussion" and "Discussion," we show and discuss the results achieved by the proposed methodology. Finally, in "Conclusion," we present the final remarks about this work.

Materials and Methods

In this section, we describe the methodology used to classify lung nodules into malignant or benign. Figure 1 summarizes the four stages followed by our methodology. In the first stage, we perform the acquisition of the images from the LIDC-IDRI database [1]. In the second stage, we have the extraction of nodules based on the experts' markings. Then, the feature extraction is applied to the nodule. At the end, we have the selection of the best model, classification, and validation of results.

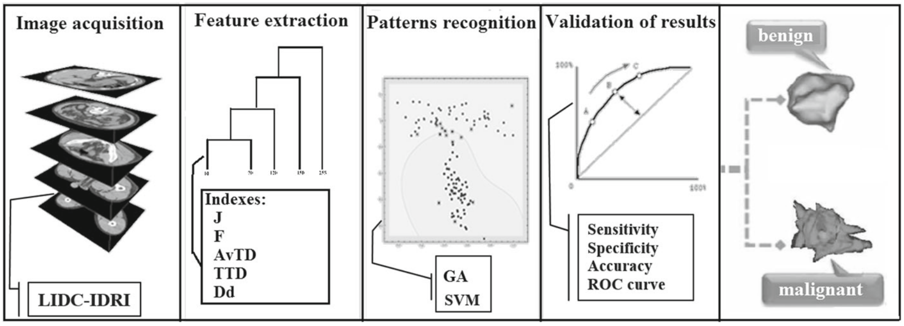


Fig. 1 Proposed methodology

Image Acquisition

The image database used in this work is the LIDC-IDRI [1], which is available on the internet as a result of an association between the Lung Image Database Consortium and the Image Database Resource Initiative with 1018 CT scans. The CT was acquired in different tomographies. This increases the difficulty for the classification of the lung nodules. The database has XML format files that contains markings of nodules contained in each exam made by four experts, besides some characteristics such as sphericity, texture, malignancy, etc., indicated by values from 1 to 5 representing the diagnosis given by each specialist regarding to that characteristic of the nodule. For example, in the case of the malignancy, the closer to 5 the specialist has noted, the greater the probability of the nodule is malignant, and the closer it is to 1, the greater the likelihood of being benign.

There is no imposition for consensus, all nodules indicated by the radiologists revision are taken into account and recorded. Therefore, it is possible to have different diagnosis for the same nodule. In this work, it is considered only one instance per nodule, with the objective of minimizing the impact of subjectivity in exams. The classification regarding malignancy or benignity is obtained first with the calculations presented in [22], which summarizes into one single value the nodular features made by up to four specialists through computing the mode or the median, i.e, when mode is repeated, the median is used to select which region is used for analysis. According to the result of this summary, in this work it is considered that malignant nodules are those cases which present malignancy semantic values of moderately suspicious or highly suspicious, and benign nodules are those cases which present characteristics of highly or moderately indicated benignity. As contour, it adopted the one that contains larger bounds. As a total, there were 1405 nodules obtained (1011 benign and 394 malignant).

Segmentation of Nodules

In order to segment nodules, we obtained contour information supplied by an XML file, which contains the coordinates of the nodules together with the analysis of each specialist. However, the segmentation used in this work is a summary, as presented in “Image Acquisition,” in which only the larger bound is chosen to represent the instance of the nodules described through markings made by up to four experts.

Feature Extraction

After the acquisition, the nodules are subjected to the texture-based feature extraction stage. In order to describe the texture of the nodules, we used indexes that compute the distances between pairs of species. These indexes consider the phylogenetic distance, computed from the tree architecture presented in Fig. 3. In order to build and organize the tree, we need to be aware of which species are present in the nodules. The species are represented by the Hounsfield units (HU) present in each nodule.

Next, we describe the fundamentals of how the tree was organized and of the diversity indexes.

Phylogenetic Tree - Dendrogram

Phylogenetic trees are used in Biology to describe the evolutionary phylogenetic relations between species. In these trees, the leaves represent the species and the nodes represent the common ancestors. So, it is possible to establish an evolutionary connection between the species under study. The dendrogram is a graphical representation which can be used to describe the phylogenetic relation between species and their ancestors [43].

These trees allow the extraction of indexes that connect diversity and parenthood between species [38]. Figure 3 presents an example of the phylogenetic tree, represented by an inclined dendrogram. In this tree, the leaf nodes are the analyzed species, the internal nodes correspond to some common ancestor, and the edges indicate the phylogenetic distance between two species. By means of phylogenetic trees, we can compute taxonomic indexes that connect the species of a community.

In order to establish the division of the species in the dendrogram, we first adopted a strategy based on the texture of each nodule, that is, establish a form of parenthood

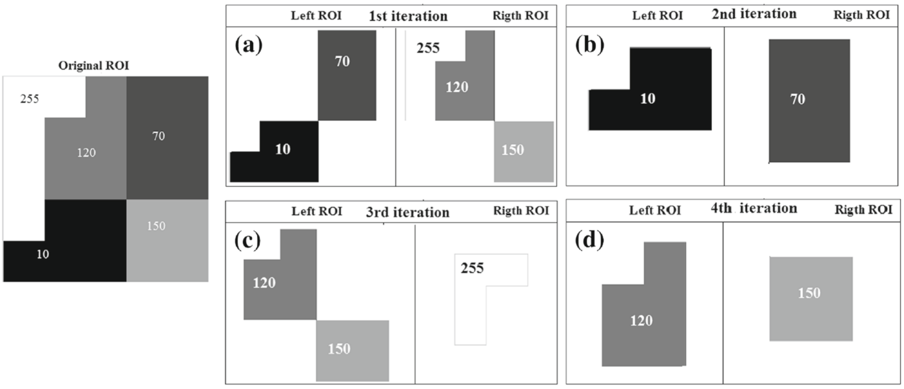


Fig. 2 Four iterations of the Otsu algorithm to separate the original ROI until there is only one species on each leaf. In **a**, the first iteration; in **b**, the second iteration; in **c**, **d**, the third and fourth ones

based on the similarity between species. For such, we used an automatic segmentation strategy based on the Otsu algorithm [47].

First, the Otsu algorithm segments the nodule into two regions of interest (ROI) based on their texture properties. Based on the thresholds generated by the Otsu algorithm, we have two ROIs: left ROI, which contains all the species (voxels) smaller than or equal to the threshold; and right ROI, which contains all the species greater than the threshold. The left ROI is then segmented once more by the Otsu algorithm, producing two new ROIs (right and left). This procedure is recursively repeated for every resulting ROI, both left and right, until the number of species present in each ROI is 1. Figure 2 exemplifies this division.

With the procedure of separation of the species contained in the ROI, such as show in Fig. 2, we obtain the dendrogram (Fig. 3) with its species and phylogenetic features (Fig. 4).

Having the dendrogram established, the phylogenetic diversity indexes can be computed in order to measure the phylogenetic relations between the species in the community. Next, we present the diversity indexes used to describe the texture of the lung nodule.

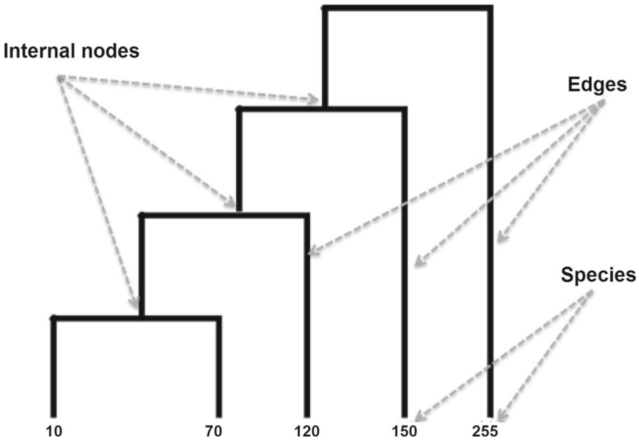


Fig. 3 Dendrogram generated from the ROI is shown in Fig. 2

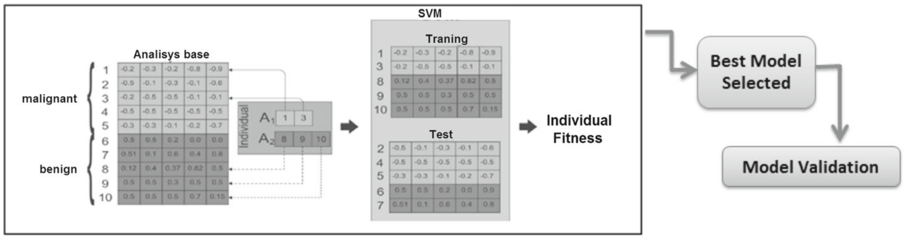
Distance-Based Phylogenetic Diversity Index

Several studies in ecology, specially the large-scale ones, depend on the richness of species as a measure of biodiversity. However, richness of species by itself can be limited as an indicator of biodiversity, since it treats every species as if they were equal, and it does not take the phylogenetic relationships into account [43].

In [37], it was shown that biodiversity points throughout the world present much more evolutionary history than one could find by just using richness of species. The work by [19] demonstrate that the phylogenetic relationships are among the most important factors the determine the extinction of species, and, in [42], they showed that the degree of phylogenetic distance can determine the success of the invasion of exotic species and sub-species. For this reason, phylogenetic information can represent a better indicator of preservation than using richness of species by itself. The application of information obtained from phylogenetic relationships is therefore a promising approach [45].

The studies that verify the distance relationships between pairs of species are based on a matrix that contains all the species of a community. The distances can be based on morphological or functional differences [21], on the length of the branches of the phylogenetic relationships based on molecular data [33, 39] or, if the lengths of the branches are unknown, on the number of nodes that separate each pair of species [13]. The values inside the distance matrix can be interpreted as the distinction between each pair of species or between a particular species and all the others [20, 34]. In our work, the distances will be represented by the number of nodes between the species.

The five indexes used to describe the texture of lung nodules are based on the distances between the pairs of species. These indexes are the intensive quadratic entropy [20], the extensive quadratic entropy [20], the average taxonomic distinctness [6], the total taxonomic distinctness [7], and the pure diversity measure [14, 46].



Analisys base

1	-0.2	-0.3	-0.2	-0.8	-0.9
2	-0.5	-0.1	-0.3	-0.1	-0.6
3	-0.2	-0.5	-0.5	-0.1	-0.1
4	-0.5	-0.5	-0.5	-0.5	-0.5
5	-0.3	-0.3	-0.1	-0.2	-0.7
6	0.5	0.5	0.2	0.0	0.0
7	0.51	0.1	0.6	0.4	0.8
8	0.12	0.4	0.37	0.82	0.5
9	0.5	0.5	0.3	0.5	0.5
10	0.5	0.5	0.5	0.7	0.15

A1	1	3	
A2	8	9	10

Traning

1	-0.2	-0.3	-0.2	-0.8	-0.9
3	-0.2	-0.5	-0.5	-0.1	-0.1
8	0.12	0.4	0.37	0.82	0.5
9	0.5	0.5	0.3	0.5	0.5
10	0.5	0.5	0.5	0.7	0.15

Test

2	-0.5	-0.1	-0.3	-0.1	-0.6
4	-0.5	-0.5	-0.5	-0.5	-0.5
5	-0.3	-0.3	-0.1	-0.2	-0.7
6	0.5	0.5	0.2	0.0	0.0
7	0.51	0.1	0.6	0.4	0.8

Fig. 4 Ilustration of how to choose the best training model [3]

The intensive quadratic entropy (J) was firstly proposed by [20] in order in establish a possible connection between the diversity indexes and biodiversity measurement indexes. When we have the same values of abundance (hypothetical or formal), the index J (Eq. 1) will be a function that represents the number of species and their taxonomic relationships. This way, it expresses the mean taxonomic distance between two randomly chosen species and, so, the relationships between the species affect the value of J, which does not happen in other diversity indexes [20].

$$J = \left[\sum d_{i,j}\right]/s^2 \tag{1}$$

where $d_{i,j}$ represents the distance between the species i, j, and s represents the number of species.

Concerning the properties of the index, monotonicity is a property which is generally necessary for diversity measurements. So, if we represent a certain measure by I, this property means that: $I(A \cup \{x\}) > I(A)$. This makes sure that the value of the index will raise if we add a new species x to a set of species A. The index J does not meet this requirement. So, it is not an ideal index. To overcome this problem, based on the index J, we apply the extensive quadratic entropy, F [20], which represents the sum of the differences between the species. The monotonicity is applied to the index F since, for any set of species A and a new species x, the value will change. Eq. 2 defines the index F.

$$F = \sum d_{i,j} \tag{2}$$

where $d_{i,j}$ represents the distance between the species i and j.

The average taxonomic distinctness ($AvTD$) [6] and total taxonomic distinctness indexes (TTD) [7] were originally developed based on taxonomic relationships. However, they can be easily adapted to phylogenetic information [36]. $AvTD$ is the mean taxonomic distance between any two species randomly chosen [6]. TTD, in turn, represents the sum of average phylogenetic distinctness for all the species. Equations 3 and 4 refer to these indexes, respectively.

$$AvTD = \left[\sum\sum_{i<j} d_{ij}\right] / [s(s-1))/2] \tag{3}$$

$$\mathrm{TTD} = \sum_i \left[\left(\sum_{j\neq i} d_{ij}\right)/(s-1)\right] \tag{4}$$

In both Eqs. 3 and 4, $d_{i,j}$ represents the distance between the species i and j, and s represents the number of species.

Finally, we have the pure diversity measure (D_D) [14, 46], which checks the distance from a species to its closer neighbor.

$$D_D = \sum d_i\_\mathrm{min} \tag{5}$$

where $d_i_$min is the smallest distance from a neighbor species i to all other species.

The phylogenetic tree pooled with intensive quadratic entropy, extensive quadratic entropy, average taxonomic distinctness, total taxonomic distinctness, and pure diversity indexes are used in biology to compare behavior patterns of species in different areas. In order to implement this idea, the first step is to make a correspondence between the terms used in biology and those used in our methodology. Table 1 shows this correspondence.

The indexes presented were used as texture descriptors to characterize each lung nodule.

Selection of the Best Model

It is common to find methodologies in the literature that use techniques to select the most significant features to generate a training database. We chose to use the genetic algorithm (GA) proposed by [3] to select the best individuals to generate the model that will be used in the classification.

Table 1 Correspondence between biology and our methodology terms

Biology	Our methodology
Community	Region of interest (nodule) of the CT image
Species	Maximum number of HU in the region
Ancestors	Number of internal nodes in the dendrogram
Phylogenetic distance	Number of edges between two species

The GA proposed by [3] to select the best individuals can be summarized in the following steps:

1. Form the basis of analysis (AB) starting from the set of all feature vectors of each nodule obtained in the image acquisition ("Image Acquisition").
2. Each element in the A1 matrix contains the position of a features vector extracted from a benign nodule in the AB.
3. Each element in the A2 matrix contains the position of a features vector extracted from a malignant nodule in the AB.
4. The values contained in the A1 and A2 matrices are modified by mutation and the crossover genetic operators, whereas the values cannot be repeated. This means that A1 elements may not be present in A2, or vice versa. The genetic operators of crossover and mutation do not match data matrices.
5. The feature vectors selected in A1 and A2 for each individual in a generation are trained by the support vector machine (SVM) [9]. The SVM is also used to test the feature vectors in the AB that were not selected. The sum of specificity, sensitivity and accuracy is used to measure the *fitness* of each individual. This process is repeated for 500 consecutive generations of that best individual until the *fitness* be the same.
6. At the end of evolution all feature vectors, whose positions are contained in the A1 and A2 matrices of the individual with the best fitness in the last generation, form the best training model.
7. The last step is to validate the selected model by means of a classification using the remaining nodules of the base. And then, calculation of sensitivity, specificity, and accuracy is made. Doing so, it is possible to measure the quality of the model.

Pattern Recognition

After finishing the feature extraction stage and the selection of the fittest individuals, nodules are classified as malignant or benign. The feature vectors are obtained by means of the proposed shape analysis. These values are used by the SVM classifier with the radial base function (RBF) [35].

SVM is a powerful, state-of-the-art algorithm with strong theoretical foundations based on the Vapnik-Chervonenkis theory. SVM has strong regularization properties. Regularization refers to the generalization of the model to new data. This characteristic was the main reason for choosing this classifier in our work. The accuracy of an SVM model is highly dependent on the selection of kernel parameters such as C and λ. We used the LibSVM software [4] to estimate both these parameters. All values of the sample were normalized between -1 and 1 to improve the performance of the SVM to guarantee a shorter processing time without mischaracterizing the original value of the feature [9].

Results Validation

After the conclusion of the pattern recognition stage, it is necessary to validate and discuss the results. This methodology uses metrics commonly applied in CAD/CADx systems for performance analysis of systems based on image processing, namely: sensitivity, specificity, and accuracy [9]. In [12], another way of measuring the performance of computer-based detection techniques is used, receiver operating characteristic (ROC) curves. A ROC curve indicates the true positive rate (sensitivity) as a function of the false positive rate (1−specificity).

Results and Discussion

In this section, we present the results achieved by the proposed methodology at the diagnosis of lung nodules, as described in "Materials and Methods." The strategy for analysis of the results is the following: (1) acquisition of the images used to train and test the methodology; (2) description of how the process of feature extraction occurred; (3) execution of tests with the SVM for each index shown in "Feature Extraction." We used the genetic algorithm described in "Selection of the Best Model" to select the best training model, and then computed the accuracy, sensitivity and specificity for each test and the ROC value; (4) finally, a comparative analysis with other related works.

Database Separation

The LIDC-IDRI database contains 1018 exams, but two factors made 185 of them unsuitable for this methodology. The first factor concerns exams that do not have nodules larger than or equal to 3 mm in diameter, since these cases do not include information that indicate the degree of malignancy of the nodule. The second factor is the divergence between information found in the marking file of the exam and information contained in the DICOM header for the same exam, which hinders the coherent use of the markings [3]. Therefore, we applied the proposed methodology to 833 exams.

All of the training bases were generated by means of the genetic algorithm (GA) ("Selection of the Best Model"). This algorithm is responsible for selecting the best nodules, thus making sure that only the most significant nodules among malignant and benign are selected to create the training model. An important factor that must be mentioned is the number of generations used until the GA stops evolving.

Table 2 Size of nodules

Nodule	Size/diameters		
	Up to 10 mm	Up to 20 mm	Up to 30 mm
Benign	359	487	163
Malignant	67	155	172

When the fitness value repeats for 500 consecutive times (generations), the GA reaches the stop criterion.

From the 833 exams, we extracted 1405 nodules (1011 benign and 394 malignant), which were split into 80% (1124) for training and validation and 20% (281) for tests, all being randomly selected. The 20% will be used to test the final model selected by the GA. The 80% in the training base were subjected to the GA for selection of the fittest individuals (nodules). In order to select just the most significant individuals and balance the training base, the population size given to the GA was of just 70% of the base of the class with fewer elements, that is, the malignant class. This way, from the original 80% of the training base, the GA chooses only the 70% that best represent the malignant class, and this same quantity is used to perform the selection in the benign class. The rest of the nodules in the training base were used by the GA to validate the model. This procedure was applied to all the experiments conducted in our methodology.

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

Classification

The tests were carried out for each separate index, that is, each nodule was characterized by a single index (one feature per nodule). After that, another test was applied, in which each nodule was represented by all the indexes together (five features per nodule). All the tests followed the training/test scheme shown in "Database Separation." Table 3 presents all the results for these tests, as well as the parameters C and λ estimated for use in the RBF kernel of the SVM in the classification stage.

The ideal CADx system has a good balance between the three metrics used for evaluation (accuracy, sensitivity, and specificity), since a good methodology must be capable of successfully classifying both malignant and benign cases. Analyzing the results in Table 3, based on this criterion, the best result is found for the combination of all indexes, achieving accuracy of 92.52, sensitivity of 93.10, specificity of 92.26, and an area under the ROC curve of 0.921. As the worst case, we highlight the results for the *AvTD* index, with accuracy of 83.90, sensitivity of 91.95, specificity of 79.38 and, finally, a ROC of 0.856.

The number of support vectors is related to the generalization capability of the methodology; the smallest the number of vectors, more generic the classification model is. Our work managed reach very significant values in all experiments. For the best result, we achieved 0.273%, that is, below 30%, representing the generalization capability of the methodology. For the worst case, we have the *AvTD* index, which achieved 0.485%.

Since we believe an index can identify something that another one cannot, we decided to test them all together. As one may observe in Table 3, the combined indexes present promising values. These results prove that the indexes are complementary. Therefore, when all measures are used together, the results prove the efficacy of the methodology for diagnosing lung nodules.

The average time spent on the extraction of the characteristics was 4.5 s, with emphasis on the stage of dendrogram assembly ("Phylogenetic Tree - Dendrogram"), which consumes on average 3.8 s. The step of selecting the best model by GA consumed 2100 s (35 min), and this step is performed only once, i.e., after the generation of the model, it is only necessary to use it by SVM. The entire methodology was developed using Microsoft Windows operating system, with the C++ programming language and hardware composed of: Intel i7 processor and 8GB of RAM memory.

Table 3 Results for all the indexes individually tested

Index	Ac(%)	Se(%)	Spe(%)	ROC	C	λ	Sv (%)
J	83.98	93.10	79.89	0.867	8	0.5	0.441
F	88.61	86.20	89.69	0.837	24	0.23441	0.338
AvTD	83.27	91.95	79.38	0.856	512	0.187453	0.485
TTD	87.90	89.65	87.11	0.881	128	0.34234	0.322
D_D	88.61	93.10	86.59	0.898	256	0.98343	0.308
All index	92.52	93.10	92.26	0.921	1024	0.125	0.273

AC accuracy, *SE* sensitivity, *Spe* specificity, and number of *SV* support vectors

Comparison With Other Related Works

The comparison with other works in the area is a hard task because no one of the works cited in this article supplied the exams used. The only peace of information provided is the database used. So, we were unable to perform a rigorous evaluation of our method with respect to other works.

Our objective with Tables 4 and 5 is to provide an overview (exam database, complexity of the methodology, etc.) of the results found in the related works and in our work. So, we intend to show that our methodology is promising, since, compared to other works, we achieved results above 93% for various types of situation: (1) diagnosis using only shape; (2) large and complex sample; and (3) several configurations of the sample for training and test.

The comparison was performed in two manners. First, in Table 4, we compared our methodology with works that only used texture measures to form their feature vectors. In Table 5, in turn, we present the comparison with works that used both texture and shape measures to characterize their nodules.

Comparing the best result found in our work with those presented in Table 4, one may notice that our results are very promising, since they are either superior or equal to those which use shape measures only. Furthermore, if we take into account the number of cases analyzed, our methodology analyzed a superior quantity of nodules.

Table 5 presents the results of works that used texture and shape measures for the characterization of their nodules. The works by [10, 11] present slightly superior accuracy to ours, which does not mean that they are superior, since the values of sensitivity and specificity, which tell how efficient the methodology is at detecting the presence or absence of the illness, are quite lower or even not informed. Besides, the number of cases they analyze is considerably inferior.

Discussion

The proposed methodology was evaluated by applying a set of 1405 nodules (benign and malignant) from the LIDC-IDRI database. The training set contained 80% of the base, and the remaining 20% were used for the tests with the genetic algorithm. The experimental results allowed the formulation of the following conclusions:

1. The use of the diversity indexes J, F, $AvTD$, TTD, and D_D combined with phylogenetic trees led to good results.
2. The use of the genetic algorithm showed efficiency at the selection of the fittest individuals for creation of the best training model and to balance the classes [2].

Table 4 Comparison with other publications with respect to the classification of lung nodules in benign and malignant using only features of texture

Work	Techniques	Database	Ac(%)	Se(%)	Spe(%)
[28]	Texture features using diversity indexes of Shannon and Simpson, linear discriminant dnalysis (LDA), and SVM	LIDC	92.78	85.64	97.89
[28]	Texture features using diversity indexes of Shannon and Simpson, LDA, and SVM	LIDC-IDRI	83.75	82.95	84.58
[30]	Texture features, correlation-based feature selection, k-nearest neighbor, and SVM	NBIA-ELCAP	82.66	96.15	52.17
[23]	Texture features, correlation-based feature selection, and k-nearest neighbor	LIDC-IDRI	90.91	85.71	94.74
[8]	Texture features using matrix co-occurrence of gray levels, principal component analysis, and artificial neural network	Private	90.63	92.30	89.47
[31]	Texture features using matrix co-occurrence of gray levels and SVM	Private	–	91.38	89.56
Our work		LIDC-IDRI	92.52	93.10	92.26

AC accuracy, *SE* sensitivity, and *Spe* specificity

Table 5 Comparison with other publications with respect to the classification of lung nodules in benign and malignant using only features of shape and texture

Work	Techniques	Database	Ac(%)	Se(%)	Spe(%)
[44]	Shape features using gradient field and radius features, stepwise, simplex optimization, LDA, and SVM.	Private	85	–	–
[24]	Shape features using bi-orthogonal wavelet and fuzzy classifier.	Private	90	86	84
[10]	Shape features using spherical harmonics, mapping this model to the unit sphere, and k-nearest classification	Private	93.6	–	–
[11]	Shape and texture features using and radial basis function neural network		94.44	–	88.14
Our work		LIDC-IDRI	92.52	93.10	92.26

AC accuracy, *SE* sensitivity, and *Spe* specificity

3. Regardless of the shape analyzed, we could achieve good results just by using texture for the characterization of the lung nodules, combining the diversity indexes *J*, *F*, *AvTD*, TTD, and phylogenetic tree.
4. The combination of all the techniques allowed a better discrimination at the classification of the nodules, reaching accuracy of 92.52, sensitivity of 93.10, and specificity of 92.26.
5. Finally, it is important to highlight that the LIDC-IDRI database is extremely complex and diversified, this is, contains countless different cases of lung nodules. This database has exams that were extracted by various tomographers, making it harder to detect, classify or even diagnose them through CAD/CADx systems.

All those items aggregate value to this methodology. The texture features of the lung nodule analyzed by the indexes *J*, *F*, *AvTD*, TTD, and D_D combined with phylogenetic trees pooled with the genetic algorithm are behind the good results. Besides that, the complexity of the LIDC-IDRI database allows us a more reliable conclusion about the results.

Conclusion

High rates of deaths and records of lung cancer occurrences around the world demonstrate the importance of developing research in order to produce resources for early diagnosis of the disease, thereby providing a better treatment. This article presented a methodology for classification of lung nodules into malignant and benign. For such, we used the following distance-based phylogenetic diversity indexes as texture descriptors: intensive quadratic entropy, the extensive quadratic entropy, the average taxonomic distinctness, the total taxonomic distinctness, and pure diversity measure. At the end, genetic algorithm and support vector machine for classification of lung nodules into malignant and benign. The methodology proved to be a useful tool for specialist physicians.

The results demonstrate the promising performance of the techniques for analysis of texture of lung nodules based on phylogenetic diversity indexes. Another important factor is the combination of the proposed techniques and the use of the genetic algorithm, since they allow a better result at the differentiation between malignant and benign.

Finally, the methodology presented in this work might integrate a CADx tool to be applied to the detection and diagnosis of lung cancer, in order to classify the nodules in malignant and benign, thus making the analysis of exams by the specialist more agile and less exhaustive.

Acknowledgments The authors acknowledge the Coordination for the Improvement of Higher Education Personnel (CAPES), the National Council for Scientific and Technological Development (CNPq), the Foundation for the Protection of Research and Scientific, the Technological Development of the State of Maranhão(FAPEMA), and the Foundation for Research Support of the State of Piauí (FAPEPI) for financial support.

References

1. Armato SG, McLennan G, Bidaut L, McNitt-Gray MF, Meyer CR, Reeves AP, Zhao B, Aberle DR, Henschke CI, Hoffman EA, Kazerooni E, MacMahon H, Van Beeke EJR, Yankelevitz D, Biancardi AM, Bland PH, Brown MS, Engelmann RM, Laderach GE, Max D, Pais, RC, Qing, DPY, Roberts RY, Smith

AR, Starkey A, Batrah P, Caligiuri P, Farooqi, A, Gladish GW, Jude CM, Munden RF, Petkovska I, Quint LE, Schwartz LH, Sundaram B, Dodd LE, Fenimore C, Gur D, Petrick N, Freymann J, Kirby J, Hughes B, Casteele AV, Gupte S, Sallamm M, Heath MD, Kuhn MH, Dharaiya E, Burns R, Fryd DS, Salganicoff M, Anand V, Shreter U, Vastagh S, Croft BY: The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. Med Phys 2:915–31, 2011. http://www.biomedsearch.com/nih/Lung-Image-Database-Consortium-LIDC/21452728.html.

2. Ben-Hur A, Weston J: A user's guide to support vector machines. In Carugo O, Eisenhaber, F Eds. Data Mining Techniques for the Life Sciences, Methods in Molecular Biology, vol 609, Humana Press, 2010, pp 223–239. doi:10.1007/978-1-60327-241-4-13.
3. de 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 3:165–177, 2014. doi:10.1016/j.artmed.2013.11.002, http://www.sciencedirect.com/science/article/pii/S0933365713001541.
4. Chang CC, Lin CJ: LIBSVM — a library for support vector machines, 2013. http://www.csie.ntu.edu.tw/cjlin/libsvm/.
5. Chen W, Li Z, Bai L, Lin Y: Nf-kappab in lung cancer, a carcinogenesis mediator and a prevention and therapy target. Front Biosci (Landmark edition) 16:1172–85, 2011. doi:10.2741/3782.
6. Clarke KR, Warwick RM: A taxonomic distinctness index and its statistical properties. J Appl Ecol 35(4):523–531, 1998. http://www.jstor.org/stable/2405167.
7. Clarke KR, Warwick RMRM, Laboratory PM: Change in marine communities: an approach to statistical analysis and interpretation, 2nd edition. Plymouth, U.K.: PRIMER-E Ltd, 2001. Includes bibliographical references (p. A3-1-A3-5).
8. Dandil E, Cakiroglu M, Eksi Z, Ozkan M, Kurt O, Canan A: Artificial neural network-based classification system for lung nodules on computed tomography scans. In: Soft computing and pattern recognition (soCPar), 2014 6th international conference of, 2014, pp 382–386. doi:10.1109/SOCPAR.2014.7008037.
9. Duda RO, Hart PE: Pattern classification and scene analysis. Wiley-Interscience Publication: New York, 1973.
10. El-Baz A, Nitzken M, Khalifa F, Elnakib A, Gimelfarb G, Falk R, El-ghar M: 3D shape analysis for early diagnosis of malignant lung nodules. In: Szekely G, Hahn H Eds. Information Processing in Medical Imaging, Lecture Notes in Computer Science, 6801. Springer: Berlin, 2011, pp 772–783. doi:10.1007/978-3-642-22092-0-63.
11. Elizabeth D, Nehemiah H, Retmin Raj C, Kannan A: Computer-aided diagnosis of lung cancer based on analysis of the significant slice of chest computed tomography image. IET Image Process 6(6):697–705, 2012. doi:10.1049/iet-ipr.2010.0521.
12. van Erkel A, Pattynama P: Receiver operating characteristic (ROC) analysis: basic principles and applications in radiology. Eur J Radiol 27(2):88–94, 1998.
13. Faith DP: Conservation evaluation and phylogenetic diversity. Biol Conserv 61(1):1–10, 1992. doi:10.1016/0006-3207(92)91201-3.
14. Faith DP: Phylogenetic pattern and the quantification of organismal biodiversity Philos Trans: Biol Sci 345(1311):45–58, 1994. http://www.jstor.org/stable/56137.
15. Fujimoto J, Wistuba II: Current concepts on the molecular pathology of non-small cell lung carcinoma. Semin Diagn Pathol 31(4):306–313, 2014. doi:10.1053/j.semdp.2014.06.008, http://www.sciencedirect.com/science/article/pii/S0740257014000616. Lung Carcinoma: Beyond The {WHO} Classification.
16. Gerlinger M, Rowan AJ, Horswell S, Larkin J, Endesfelder D, Gronroos E, Martinez P, Matthews N, Stewart A, Tarpey P, Varela I, Phillimore B, Begum S, McDonald NQ, Butler A, Jones D, Raine K, Latimer C, Santos CR, Nohadani M, Eklund AC, Spencer-Dene B, Clark G, Pickering L, Stamp G, Gore M, Szallasi Z, Downward J, Futreal PA, Swanton C: Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. N Engl J Med 366(10):883–892, 2012. doi:10.1056/NEJMoa1113205. PMID: 22397650.
17. Gould M, Maclean C, Kuschner W, Rydzak C, Owens D: Accuracy of positron emission tomography for diagnosis of pulmonary nodules and mass lesions: a meta-analysis. JAMA 285(7):914–924, 2001. doi:10.1001/jama.285.7.914.
18. Hansell DM, Bankier AA, MacMahon H, McLoud TC, Muller NL, Remy J: Fleischner society: glossary of terms for thoracic imaging. Radiology 246(3):697–722, 2008. doi:10.1148/radiol.2462070712. PMID: 18195376.
19. Heard SB, Mooers AO: Phylogenetically patterned speciation rates and extinction risks change the loss of evolutionary history during extinctions. Proc Biol Sci 267(1443):613–620, 2000. http://www.jstor.org/stable/2665984.
20. Izsák J, Papp L: A link between ecological diversity indices and measures of biodiversity. Ecol Model 130(1–3):151–156, 2000. doi:10.1016/S0304-3800(00)00203-9, http://www.sciencedirect.com/science/article/pii/S0304380000002039.
21. Izsáki J, Papp L: Application of the quadratic entropy indices for diversity studies of drosophilid assemblages. Environ Ecol Stat 2(3):213–224, 1995. doi:10.1007/BF00456668.
22. Jabon SA, Raicu DS, Furst JD: Content-based versus semantic-based retrieval: an LIDC case study. Proc SPIE 7263:72,631L–72,631L–8, 2009. doi:10.1117/12.812877.
23. Krewer H, Geiger B, Hall L, Goldgof D, Gu Y, Tockman M, Gillies R: Effect of texture features in computer aided diagnosis of pulmonary nodules in low-dose computed tomography. In: Systems, man, and cybernetics (SMC), 2013 IEEE international conference on, 2013, pp 3887–3891. doi:10.1109/SMC.2013.663.
24. Kumar S, Ramesh J, Vanathi P, Gunavathi K: Robust and automated lung nodule diagnosis from ct images based on fuzzy systems. In: Process automation, control and computing (PACC), 2011 international conference on, 2011, pp 1–6. doi:10.1109/PACC.2011.5979050.
25. Lederlin M, Revel MP, Khalil A, Ferretti G, Milleron B, Laurent F: Management strategy of pulmonary nodule in 2013. Diagn Interv Imaging 94(11):1081–1094, 2013. doi:10.1016/j.diii.2013.05.007, http://www.sciencedirect.com/science/article/pii/S2211568413001964.
26. Leef 3rd J, Klein J: The solitary pulmonary nodule. Radiol Clin N Am 40(1):123–43, ix, 2002. doi:10.1056/NEJMcp012290.
27. Magurran AE: Measuring biological diversity. Afr J Aquat Sci 29(2):285–286, 2004.
28. Nascimento LB, de Paiva AC, Silva AC: Lung nodules classification in CT images using Shannon and Simpson diversity indices and SVM. In: Proceedings of the 8th international conference on machine learning and data mining in pattern recognition, 12. Springer: Berlin, 2012, pp 454–466.
29. Netto SMB, Silva AC, Nunes RA, Gattass M: Automatic segmentation of lung nodules with growing neural gas and support vector machine. Comp in Bio and Med 42(11):1110–1121, 2012.
30. Orozco H, Osiris Vergara Villegas O, Maynez L, Sanchez V, De Jesus Ochoa Dominguez H: Lung nodule classification in frequency domain using support vector machines. In: Information science, signal processing and their applications (ISSPA),

2012 11th international conference on, 2012, pp 870–875. doi:10.1109/ISSPA.2012.6310676.

31. Parveen SS, Kavitha C: Classification of lung cancer nodules using SVM Kernels. Int J Comput Appl 95(25):25–28, 2014. Full text available.
32. Patil SS, Godoy MC, Sorensen JI, Marom EM: Lung cancer imaging. Semin Diagn Pathol 31(4):293–305, 2014. doi:10.1053/j.semdp.2014.06.007, http://www.sciencedirect.com/science/article/pii/S0740257014000604. Lung Carcinoma: Beyond The who Classification.
33. Pavoine S, Ollier S, Dufour AB: Is the originality of a species measurable? Ecol Lett 8:579–586, 2005. https://hal.archives-ouvertes.fr/hal-00427764.
34. Rao C: Diversity and dissimilarity coefficients: a unified approach. Theor Popul Biol 21(1):24–43, 1982. doi:10.1016/0040-5809(82)90004-1, http://www.sciencedirect.com/science/article/pii/0040580982900041.
35. Schölkopf B, Smola A: Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT Press, 2002.
36. Schweiger O, Klotz S, Durka W, Kühn I: A comparative test of phylogenetic diversity indices. Oecologia 157(3):485–495, 2008. doi:10.1007/s00442-008-1082-2.
37. Sechrest W, Brooks TM, Fonseca GABD, Konstant WR, Mittermeier RA, Purvis A, Rylands AB, Gittleman JL: Hotspots and the conservation of evolutionary history. Proc Natl Acad Sci USA 99(4):2067–2071, 2002. http://www.jstor.org/stable/3057919.
38. da Silva IA, Batalha MA: Taxonomic distinctness and diversity of a hyperseasonal savanna in central Brazil. Divers Distrib 12(6):725–730, 2006. doi:10.1111/j.1472-4642.2006.00264.x.
39. Solow A, Polasky S, Broadus J: On the measurement of biological diversity. J Environ Econ Manag 24(1):60–68, 1993. doi:10.1006/jeem.1993.1004, http://www.sciencedirect.com/science/article/pii/S0095069683710041.
40. Sone S, Takashima S, Li F, Yang Z, Honda T, Maruyama Y, Hasegawa M, Yamanda T, Kubo K, Hanamura K, Asakura K: Mass screening for lung cancer with mobile spiral computed tomography scanner. Lancet 351(9111):1242–1245, 1998. doi:10.1016/S0140-6736(97)08229-9, http://www.sciencedirect.com/science/article/pii/S0140673697082299.
41. Stewart: World cancer report 2014. IARC Nonserial Publication: New York, 2014.
42. Strauss SY, Webb CO, Salamin N: Exotic taxa less related to native species are more invasive. Proc Natl Acad Sci 103:5841–5845, 2006. doi:10.1073/pnas.0508073103.
43. Vane-Wright R, Humphries C, Williams P: What to protect?—systematics and the agony of choice. Biol Conserv 55(3):235–254, 1991. doi:10.1016/0006-3207(91)90030-D, http://www.sciencedirect.com/science/article/pii/000632079190030D.
44. Way TW, Sahiner B, Chan HP, Hadjiiski L, Cascade PN, Chughtai A, Bogot N, Kazerooni E: Computer-aided diagnosis of pulmonary nodules on CT scans: improvement of classification performance with nodule surface features. Med Phys 36(7):3086–3098, 2009. doi:10.1118/1.3140589, http://scitation.aip.org/content/aapm/journal/medphys/36/7/10.1118/1.3140589.
45. Webb CO, Ackerly DD, McPeek MA, Donoghue MJ: Phylogenies and community ecology. Annu Rev Ecol Syst 33(1):475–505, 2002. doi:10.1146/annurev.ecolsys.33.010802.150448.
46. Weitzman M: On diversity. Q J Econ 107:363–405, 1992.
47. Yang X, Shen X, Long J, Chen H: An improved median-based Otsu image thresholding algorithm. AASRI Procedia 3:468–473, 2012. doi:10.1016/j.aasri.2012.11.074, http://www.sciencedirect.com/science/article/pii/S2212671612002338. Conference on Modelling, Identification and Control.
48. Ye X, Lin X, Dehmeshki J, Slabaugh G, Beddoe G: Shape-based computer-aided detection of lung nodules in thoracic CT images IEEE Trans Biomed Eng 56(7):1810–1820, 2009. doi:10.1109/TBME.2009.2017027.