

Lung-Nodule Classification Based on Computed Tomography Using Taxonomic Diversity Indexes and an SVM

Antonio Oseas de Carvalho Filho¹ · Aristófanés Corrêa Silva¹ ·
Anselmo Cardoso de Paiva¹ · Rodolfo Acatauassú Nunes² · Marcelo Gattass³

Received: 4 March 2015 / Revised: 2 March 2016 / Accepted: 22 March 2016 / Published online: 6 April 2016
© Springer Science+Business Media New York 2016

Abstract The present work aims to develop a methodology for classifying lung nodules using the LIDC-IDRI image database. The proposed methodology is based on image-processing and pattern-recognition techniques. To describe the texture of nodule and non-nodule candidates, we use the Taxonomic Diversity and Taxonomic Distinctness Indexes from ecology. The calculation of these indexes is based on phylogenetic trees, which, in this work, are applied to the candidate characterization. Finally, we apply a Support Vector Machine (SVM) as a classifier. In the testing stage, we used 833 exams from the LIDC-IDRI image database. To apply the methodology, we divided the complete database

into two groups for training and testing. We used training and testing partitions of 20/80 %, 40/60 %, 60/40 %, and 80/20 %. The division was repeated five times at random. The presented methodology shows promising results for classifying nodules and non-nodules, presenting a mean accuracy of 98.11 %. Lung cancer presents the highest mortality rate and has one of the lowest survival rates after diagnosis. Therefore, the earlier the diagnosis, the higher the chances of a cure for the patient. In addition, the more information available to the specialist, the more precise the diagnosis will be. The methodology proposed here contributes to this.

Keywords Lung cancer · Phylogenetic trees · Taxonomic diversity index · Taxonomic distinctness · Medical image

✉ Antonio Oseas de Carvalho Filho
antoniooseas@gmail.com

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

² State University of Rio de Janeiro, Sao Francisco de Xavier, 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

1 Introduction

Lung nodules are a potential occurrence of lung cancer, and early detection is essential for survival. A crucial factor that contributes to the occurrence of this type of cancer is a high exposure to smoking. The majority of lung-cancer cases (around 80 %) are related to smoking. On average, smokers have a 20 to 30 times higher risk of developing lung cancer [9]. Moreover, it is also known as one of the cancers with a lower survival rate [6].

The detection of lung nodules is extremely important, because they have a high chance of turning into cancer [16]. The detection of such nodules using Computerized Tomography (CT) is not a simple task, since they can have a contrast similar to other structures, low density, small size in a complex area (e.g., connected to blood vessels or in the borders of the lung), etc. [19]. Another factor that makes detection difficult is the fact that specialists have

a large number of CTs to analyze. This exhausting and repetitive process can lead to distractions and result in a misunderstanding of the analysis, especially when the image simultaneously presents other anomalies. Thus, this type of analysis is frequently vulnerable to errors [20].

For these reasons, the earlier the diagnosis, the higher the chances of cure for the patient. Furthermore, the more information available to the specialist, the more precise the diagnosis will be. In the last decades, the development and usage of digital image-processing techniques in CT have attracted a great deal of interest, with the main goal of increasing the diagnosis accuracy and providing the specialist with a second opinion. These techniques have been combined to develop Computer-Aided Detection (CAD) and Computer-Aided Diagnostic (CADx) systems [34].

In a CAD system, most often the segmentation stage is automatic. This stage usually segments many structures with features (shape, density, or texture) similar to lung nodules. Therefore, it is essential that there is a stage that removes as many non-lung nodules as possible and preserves the lung nodules. This stage is called reduction of false positives. The more non-nodules are removed, the better the performance of the CAD system.

To illustrate the problem, in the work of [7], in the automatic segmentation step for 833 exams, 607 nodules and 67067 non-nodules were segmented. As can be seen, there is a very high number of non-nodules. Thus, the false-positive reduction step was necessary. By applying this stage, were removed around 97 % of the non-nodules.

Importantly, the CAD systems are not interested in specifying whether the structures left over from the false-positive reduction step (nodules and a few non-nodules) have benign or malignant features. This discrimination is performed by a CADx system, where only the nodules are analyzed by shape, density, and/or texture, and classified as benign or malignant.

Usually the detection process consists of four stages: 1) image acquisition; 2) segmentation of nodule candidates; 3) extraction of features from the candidates; and 4) reduction of false positives, and classifying candidates as nodules and non-nodules. All stages are fundamental to the success of a CAD system. However, in this paper, we will emphasize the feature-extraction and classification stages. In these stages, we obtain information to reduce the rate of false positives and false negatives. Thus, our work will contribute to the area.

Various studies are frequently conducted with the goal of increasing the accuracy rates of lung-cancer detection on CAD systems. Three important common points of these works are a high number of false positives, a high rate of false negatives, and a reduced number of evaluation cases. Therefore, there is a continuous need for developing CAD systems to support the classification of lung nodules. At the

end of the paper, we present a summary of each work (see Table 13).

This paper is organized as follows. In Section 2, work related to the proposed methodology is presented. In Section 3, we present the methodology used to classify the nodule candidates extracted from CT as nodules and non-nodules. We extract the features using the taxonomic indexes and classify them with a support vector machine (SVM). In Section 4, we show and discuss the results achieved through the proposed methodology. Finally, in Section 5, we present our final remarks about this work.

2 Related Works

After segmentation, incorrect nodule candidates may be generated, known as false positives. One important challenge is to reduce the false positives without losing actual nodules. Determining which texture and/or form measurements should be used in combination with a computational-intelligence system is a major difficulty in any system.

Over the years, researchers have tried to develop methodologies that overcome these challenges. We briefly present some works that contribute solutions to these problems. In this section, we present some works that are closely related to our methodology.

In [15], a CAD-system mechanism for lung-nodule detection is shown. Tests were performed on a data set containing 167 chest radiographs with 181 lung nodules. The system used an adaptive-threshold algorithm based on the distance between points, applied to the nodule-segmentation step. Immediately thereafter, measures were taken based on the shape, intensity, and gradient to characterize the nodule candidates. In addition, in the classification stage, a Fisher linear discriminant classifier was used, reaching a sensitivity of 78.1 % and a rate of four false positives (FP) per image.

The combination of techniques presented by [13] sets this work apart. The techniques are scale-invariant feature transform (SIFT), local binary pattern (LBP), principal component analysis (PCA), and linear discriminant analysis (LDA). The first combinations were PCA-SIFT and PCA-LBP, which presented a sensitivity of 85 %. The combination of LDA-SIFT and LDA-LBP achieved the same sensitivity.

In [19], the classification is aided by a cluster-based method. Experiments were conducted using the examinations of 32 patients, including 5721 images, with the nodules previously identified by experts. As its best result, the method achieved a sensitivity of 97.33 %, a specificity of 97.11 %, and an area under the Receiver Operating Characteristic (ROC) curve of 0.9786.

The methodology presented in [2] shows a combination of machine-learning techniques that make up a CAD

system. It consists of three major phases: (1) feature extraction, (2) feature selection, and (3) classification. The methodology is applied to a set of images that have 154 nodule regions and 92 non-nodule regions, reaching an accuracy of 96.58 %.

In [10], the authors present a method called genetic-algorithm template matching for automatic detection of lung nodules. The computation of the fitness function is based on the geometric shape of the voxel, and then combined with the global distribution of the nodule's intensity. The authors report a rate of 14 false positives per exam. The methodology was performed with a set of 70 CT images with 178 nodes. It achieved a sensitivity of 85 %.

To perform lung-nodule classification, the methodology proposed in [17] presents an approach that combines the rule-based and SVM methods. First, the region of interest (ROI) candidates' measurements are extracted based on shape, facilitating the withdrawal of some blood vessels. Thereafter, further measurements of the remaining candidates are extracted based on texture; finally, the aforementioned candidates are used as inputs to the SVM classifier. The methodology was applied to a set of tests containing 50 slices, with 50 nodules and 204 non-nodules; it reached a specificity of 92 % and an accuracy of 84.39 %.

The work proposed by [21] presents a methodology composed of the following stages: lung segmentation, lung-nodule candidate enhancement, feature extraction, and classification using an SVM with a radial basis function (RBF). An image database, containing information from 32 patients with lung nodules, was used to validate the results. They achieved a sensitivity of 93.75 %, a specificity of 87.6 %, an accuracy of 87.8 %, and an FP rate of 4.6 per exam.

A methodology to classify lung nodules is presented in [25]. The regions of interest were selected manually. To characterize a candidate, measurements were extracted from its histogram. In the classification stage, an SVM with an RBF was used. The methodology was validated on 75 tests. They achieved results of 10 false negatives (FN) and 2 FP, sensitivity and specificity of 96.15 % and 52.17 %, respectively, and an accuracy of 82.66 %.

The methodology developed by [29] included a lung-nodule detection system using segmentation, with fuzzy clustering models and SVM classification. This methodology uses three types of kernels (linear, polynomial, and RBF) for the SVM. The RBF kernel presented better results, with 80.36 % accuracy, 76.47 % specificity, and 82.05 % sensitivity.

The CAD developed by [7] presents an automatic methodology for lung-nodule detection and classification. It can be summed up in three major stages: 1) extraction and reconstruction of the lung parenchyma, which thereafter highlights their structures; 2) nodule candidates are segmented; and 3) shape and texture features are extracted, and

then classified using an SVM. The results achieved a sensitivity of 85.91 %, a specificity of 97.70 %, and an accuracy of 97.55 %.

The methodology proposed in [31] presents a classification method based on hybrid descriptors. The measures-extraction step was performed using two-dimensional Principal Component Analysis (2D-PCA) and morphological image processing. In addition, at the classification stage, methods for selecting the most significant descriptors were used, and finally, Artificial Neural Network (ANN), Random Forest (RF), Bagging, and AdaBoost were used to classify the candidates. The results achieved 90.7 % accuracy, 89.6 % sensitivity, and 87.5 % specificity.

The authors of [1] aim to detect and classify lung nodules. To that end, texture descriptors based on statistics and an SVM are used. The lung volume is extracted from a lung CT using thresholding, background removal, hole-filling, and contour correction of the lung lobe. Candidate nodules are extracted based on experts (specialists)' notes. After segmentation, the nodule candidates' features are extracted, based on statistical techniques, and finally, the candidates are classified using an SVM. A sensitivity of 96.31 % was reached.

These are examples of systems that have been developed for the detection/classification of lung nodules; Table 13 summarizes the approaches. Three common important points of these systems are a high number of false positives, a high false-negative rate, and a small number of cases for evaluation, which allows for a better conclusion. Additionally, the sensitivity and specificity results were unbalanced because of either the low number of cases or the testing methods. We explore and improve on these weaknesses as described in the following section.

3 Materials and Methods

In this section, we show our methodology for classifying lung nodules. In Fig. 1, we present the stages involved, from the image acquisition to the final stage of classifying the nodule candidates. In the first stage, we acquire the images from the LIDC-IDRI image base [3]. Using their exams, we extract the nodule and non-nodule candidates. After this, we extract the features from each nodule candidate using taxonomic indexes, and classify them with SVM. Finally, we validate the results.

3.1 Image Acquisition

The image database used in this work is the LIDC-IDRI [3], available on the Internet because of an association between the Lung-Image Database Consortium and the Image-Database Resource Initiative. The CTs were acquired in

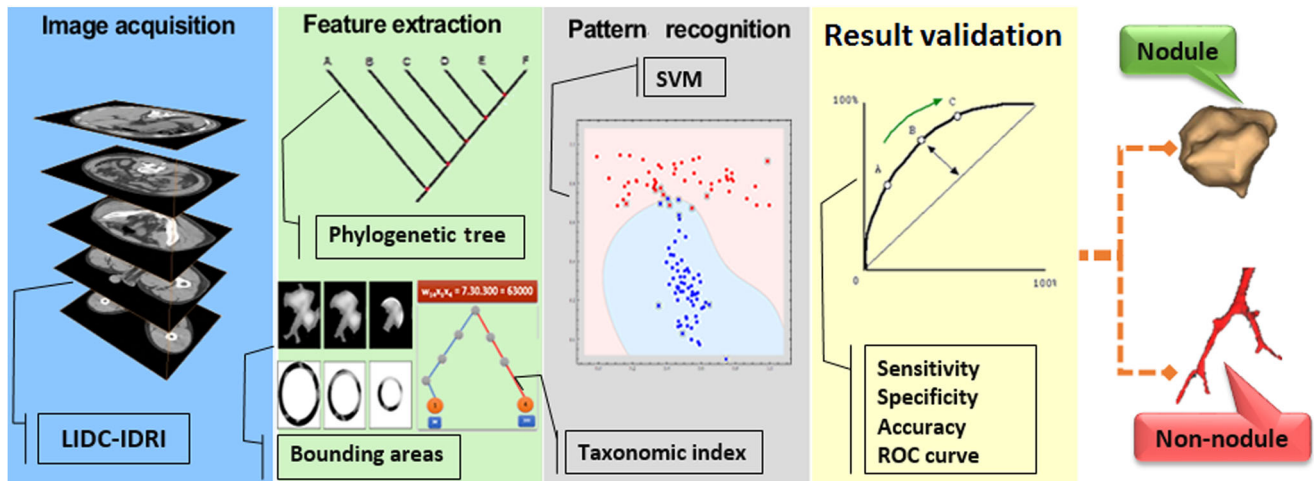


Figure 1 Proposed methodology.

different tomographs, which increases the difficulty in classifying the lung nodules. We believe that the main characteristic of this image database is the diversity of acquisition protocols. Therefore, using it makes generalization a more difficult task for any methodology. However, a methodology that achieves good results with this database will probably achieve good results with other different protocols.

The samples used in this work have two types of Volumes of Interest (VOIs): nodules and non-nodules. The nodules were automatically segmented based on markings provided by the LIDC's specialists. Figure 2 shows an example of a specialist's marking in a CT image.

Once the nodule is properly extracted, it is necessary to acquire the non-nodules. To do so, we consider important factors for a non-nodule database, namely:

1. The manual-extraction process is unfeasible, since we need a large number of non-nodules to better evaluate our methodology;

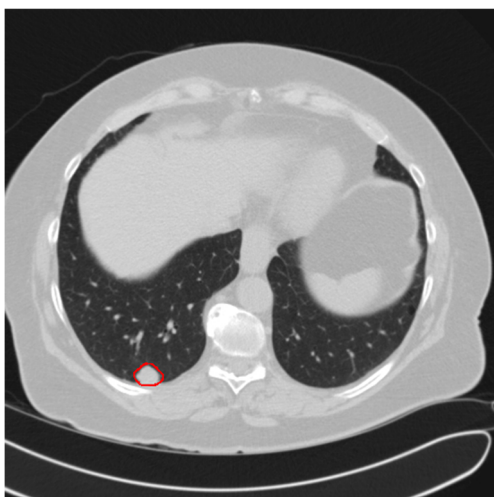


Figure 2 Example of a marking on a CT slice.

2. An automatic extraction form is needed, without any induction or human/manual intervention; therewith, we can guarantee a large number of non-nodules of various types.

Based on the items above, we chose the automatic-segmentation methodology presented in [7]. Thus, we use non-nodule candidates that we could find in a real scenario, since our goal is to integrate our methodology with a CAD system.

The non-nodule database that will be used to test our methodology was based on the segmented structures from the work of [7]. We opted to use this strategy because: 1) no non-nodule databases are available; and 2) we want structures with features similar to the nodules. Therefore, our test will be applied to a complex database.

The methodology of [7] automatically detects lung nodules. The methodology is divided into three phases: 1) extraction of lung parenchyma and improvement of the internal structures, 2) segmentation of lung-nodule-like structures; and 3) reduction of false positives. In Phase 2, 17781 structures were segmented with nodules and non-nodules.

With the aforementioned structures, we selected the non-nodules (17231) for our image base, excluding nodules (550). We know which ones are nodules from the specialist marking. Thus, if any voxel intersects with a structure segmented by [7] and nodules marked by an expert, we delete it from the base. We repeat the process for all nodules marked by the specialist and all 17781 structures. At the end of the process, we will only have non-nodules (Fig. 3).

We used non-nodule candidates from [7] because their work is consistent. Non-nodules are excluded from the detection step and can be reused in the proposed methodology. Their nodule-candidate database is not employed in the present work because the database is small compared to the

Figure 3 2D view of non-nodule candidates.



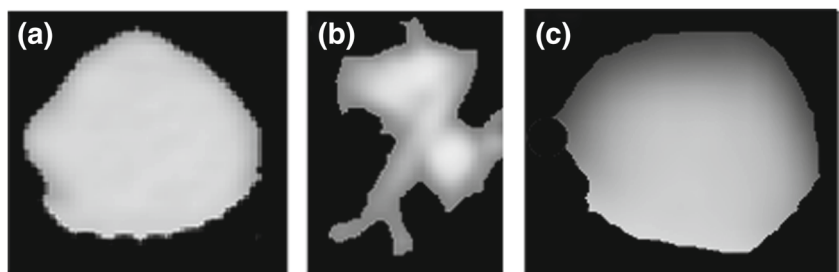
specialist markers from LIDC-IRDI, and may thus result in misclassification.

We analyzed a total of 21415 VOIs, and these, 4184 VOIs of nodules recovered from the markings and 17231 VOIs non-nodules. In Figure, we present three examples of cases of structures analyzed by our methodology. Figure 4a shows a nodule case, and Fig. 4b presents a non-nodule case. Although Fig. 4c presents a shape similar to a lung nodule (round), it is just a blood vessel. Blood vessels are responsible for most of the errors found in methodologies that use shape descriptors only, and this is one more reason for using texture descriptors.

3.2 Feature Extraction

After acquiring the nodule and non-nodule candidates, they are submitted to the feature-extraction stage, based only on texture. For the texture description of the objects, the indexes Δ and Δ^* were used. These indexes are based on the phylogenetic distance (counting the number of edges) based on the proposed tree architecture. With this goal in mind, we represent the individuals by voxels, and their Hounsfield Units (HU) represent the set of species. We used a spatial subdivision to apply these indexes, allowing a detailed analysis of each extracted region.

Figure 4 Three example candidates, **a** nodule; **b** and **c** non-nodules.



3.2.1 Approach by Spheres and Rings

Before starting the feature-extraction stage, each nodule/non-nodule candidate undergoes a stage that generates annular and spherical regions for each candidate.

These approaches allow a higher number of details to be analyzed separately in different parts of the VOIs. Thus, we are able to analyze texture-behavior patterns starting from the edges and working toward the center.

These approaches already show their effectiveness in determining regions that characterize nodule candidates, as shown in [7, 24]. We extracted six regions (rings and spheres), with increasing radii, looking for texture details for each candidate.

The size of each i and radius is defined by Eq. 1:

$$R_i = \frac{i}{6R_0} \tag{1}$$

where R_0 is the size of the radius that circumscribes the entire sample, the full extent of the region of interest; R_i are the smallest radii for $i = 1, 2, \dots, n$. In this work, the best results were obtained with three circles ($n = 2$), starting from radius R_0 .

The ring-formation process uses the radius calculated in Eq. 1. From each radius, another radius is generated with a

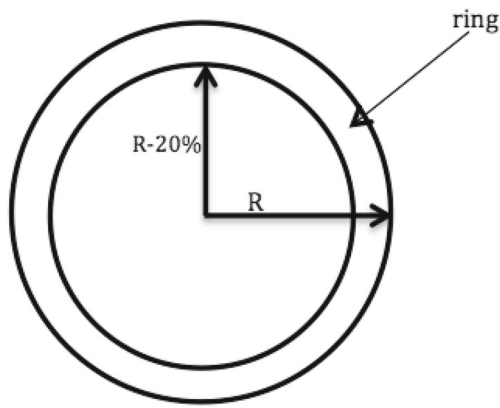


Figure 5 Examples of rings.

value 20 % smaller; thus, we have two radii. Moreover, the space formed between these two radii is equivalent to the ring that will be generated. Figure 5 shows an example of the ring formation.

This approach was applied to each candidate region. This generates three spheres and three rings for each candidate, which are used to extract texture information via the taxonomic indexes.

3.2.2 Phylogenetic Tree

Phylogenetic trees are used in biology to describe the evolutionary relations among species, as well as verifying the relationships among them in order to determine possible common ancestors. In these trees, the leaves represent the species and the internal nodes represent common ancestors to the species. Therefore, it is possible to make an evolutionary connection between the studied species. The inclined cladogram is a graphical representation used to describe the phylogenetic relation between ancestor species [4].

These trees allow the extraction of indexes that connect diversity, richness, and parenthood between species [28]. Figure 6 presents an example of the apes' phylogenetic tree, represented by an inclined cladogram, where one may notice that a chimpanzee has a higher phylogenetic proximity to humans than it does to a siamang. 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. Using phylogenetic trees, we can compute the taxonomic indexes that connect the species of a community.

Let the community of primates be the lung nodule. The chimpanzee, gorilla, and human voxel intensities would be 0, 1, and 2, respectively. The primate community has six species and the lung nodule has 65536, as each voxel may have 65536 (16 bits) voxel-intensity values. Supposedly, humankind has about 7 billion individuals (earth's population), and species 0 of the lung nodule community has on average 100 individuals (number of voxels with intensity of 0).

The relationship between two randomly chosen organisms in a phylogeny existing in a community is presented by taxonomic-diversity (Δ) and taxonomic-distinctness (Δ^*) indexes [26]. These indexes have three essential factors for application: number of species, number of individuals, and the connection structure of the species (number of edges). In this work, we use these two indexes to discriminate between nodule and non-nodule regions.

The taxonomic diversity index (Δ) considers the abundance of the species and the taxonomic relation between them. Therefore, its value expresses the mean taxonomic distance between any two individuals, randomly picked from a sample [26]. This index is defined by

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

where $x_i (i = 0, \dots, s)$ is the abundance (number of voxels) of the i th species, $x_j (j = 0, \dots, s)$ is the abundance of the j th, s represents the number of species, n is the total number of individuals and w_{ij} is the distance from species i to species j in the taxonomic classification.

The taxonomic-distinctness index (Δ^*), in turn, represents the mean taxonomic distance between two individuals, with the constraint that they belong to different species [26]. This index is defined by

$$\Delta^* = \frac{\sum \sum_{i < j} w_{ij} x_i x_j}{\sum \sum_{i < j} x_i x_j} \tag{3}$$

where $x_i (i = 0, \dots, s)$ is the abundance (number of voxels) of the i th species, $x_j (j = 0, \dots, s)$ is the abundance of the j th, s represents the number of species and w_{ij} is

Figure 6 Example of ape phylogenetic tree. Source: [4].

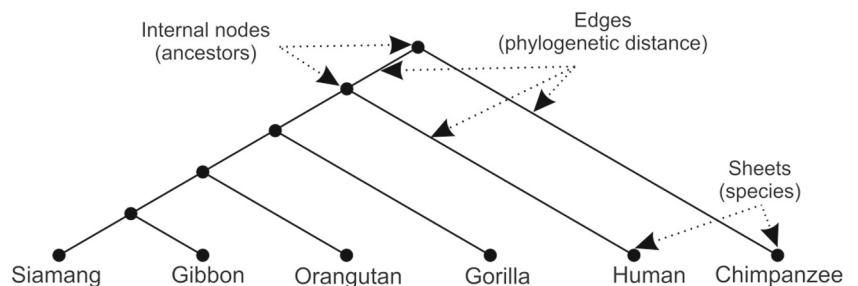


Table 1 Correspondence between biology terms and our work.

Biology	Our methodology
Community	Region (volume) of interest (VOI) of the CT image
Species	Number of HU in the VOI
Richness of species: number of species found in a certain region	Number of voxels found in the VOI
Individuals	Number of voxels of a particular species contained in the VOI
Relative abundance: number of individuals of a certain species existing in a given area	Number of voxels found in the VOI that have the same HU value

the distance from species i to species j in the taxonomic classification.

Many architectures in the literature represent species through trees, such as the architecture called "rooted tree in the shape of an inclined cladogram" [23]. In the present work, we adapt this architecture to find a higher discrimination between the nodule and non-nodule classes, where, according to [22], a community in which the species are distributed in many kinds must present a higher diversity than a community where most species belong to the same kind.

Phylogenetic trees, pooled with the taxonomic diversity and distinctness indexes, are used in biology to compare behavior patterns of species in different areas. 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.

3.2.3 Tree 1—Rooted Tree Shaped as an Inclined Cladogram

With the candidate region extracted (Section 3.2.1), the tree is created. In Fig. 7, a tree is shown where the species are represented by Hounsfield Units (HU), which can vary between +32768 and -32768. We apply a simple change to

make every value positive, with the goal of making the index calculations simpler: We move the lowest negative value so it starts from zero, after which 65536 species are possible.

The relation between species is considered from left to right. The relation between a species i and j has $w_{ij} = (j - i) + 1$ edges, for $i = 0$, and $w_{ij} = (j - i) + 2$ edges, for $i > 0$.

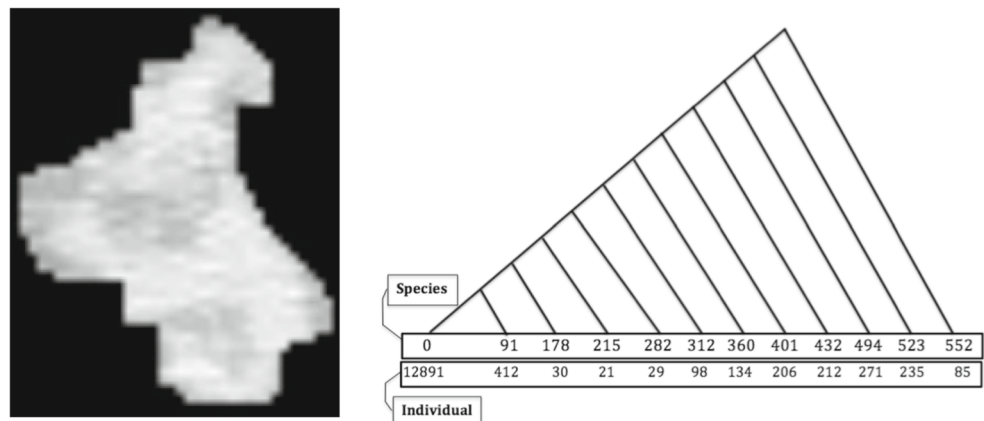
3.2.4 Tree 2—Rooted Tree Shaped as an Inclined Cladogram, Excluding Species with No Individuals

Following the same logic as the calculation of the indexes for the previous tree, we developed another architecture, which has the characteristic of eliminating species with no individuals, thus resulting in reorganizing the edges for the remaining species. The distances between species (w_{ij}) are computed traversing this modified structure.

3.2.5 Tree 3—Rooted Tree Shaped as an Inclined Cladogram, Modifying the Edges

The third proposed tree has the same combination process between species as Tree 1, where the only difference is in the computation of the number of edges, adding a weight

Figure 7 2D image of VOI with its corresponding phylogenetic tree. Species are HU and individuals are the number of voxels in the species.



for the more distant species pair. Thus, w_{ij} is computed by: $w_{ij} = 2 * (j - i)$ edges, for $i = 0$, and $w_{ij} = 2 * (j - i) + 1$ edges, for $i > 0$.

3.3 Pattern Recognition

The last stage of the methodology is classifying the candidates into nodules and non-nodules. Feature vectors were obtained in the feature-extraction step by computing the taxonomic indexes Δ and Δ^* based on the phylogenetic tree, and considering two spatial approaches: spheres and rings. These values are used by the SVM classifier with the radial base function (RBF) [27].

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 for an RBF. We used the LibSVM software [8] to estimate both of these parameters. All of the sample values were normalized between -1 and 1 to improve the SVM performance. Thus, we can guarantee improved performance without mischaracterizing the original value of the feature.

3.4 Result Validation

After concluding the-pattern recognition stage, it is necessary to validate and discuss the results. This methodology uses metrics commonly applied in CAD/CADx systems for the performance analysis of systems based on image processing. These metrics are sensitivity, specificity, and accuracy [11]. To analyze our results in more detail, we apply the variation coefficient, which calculates the dispersion of a sample with respect to its mean [30], and the chi-square test (χ^2), which is responsible for testing hypotheses [5]. In [12], the authors use Receiver Operating Characteristic (ROC) curves as another way to measure the performance of computer-based detection techniques. An ROC curve indicates the true-positive rate (sensitivity) as a function of the false-positive rate ($1 - \text{specificity}$). Finally, we use a dispersion graphic of the mean accuracies with their respective standard deviations (calculated from the mean accuracy of the five random tests).

Equations 4, 5, 6 represent the formulas used to calculate the sensitivity, specificity and accuracy, respectively.

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (4)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (5)$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

where TP is true positive, FN is false negative, TN is true negative, and FP is false positive.

Equations 7 and 8 represent the formulas for the variation coefficient and the chi-square test, respectively.

$$CV = \frac{S}{x} \quad (7)$$

where S is the standard deviation and x the average.

$$\chi^2 = \sum [(o - e)^2 / e] \quad (8)$$

where o the observed frequency for each class and e is the expected frequency for that class.

4 Results and Discussion

In this section, we present the results achieved with the lung-nodule classification methodology described in Section 3. The analysis of the results follows this strategy:

- 1) Acquisition of the images used to train and test the methodology;
- 2) Evaluation of the feature-extraction process;
- 3) Evaluation of the classification results for all test/training proportions of the tests performed with the methodology. For each proportion (e.g., 20/80 %), we calculate the means of the sensitivity, specificity, accuracy, variation coefficient, and ROC curve [12] for each tree. Concluding the analysis of the results by performing tests in parallel for all trees, we constructed two more methods of evaluating our results, the χ^2 test and the dispersion graph, based on the standard deviation of the mean accuracy;
- 4) Finally, we perform a comparative analysis with other works.

Figure 8 allows a better visualization and understanding of the flow of all the tests.

4.1 Images Acquisition

The images used to test and validate our methodology were native from the LIDC-IDRI image base [3]. A total of 833 exams were used for the application. We divided our image database into four groups, with training/testing percentages of 20/80 %, 40/60 %, 60/40 %, and 80/20 %. For each group, the individuals were randomly chosen for training and testing. The SVM performed five classifications, which were evaluated in terms of sensitivity, specificity, and accuracy. At the end of the process, we obtained the mean of each measurement. Our goal is to show that our methodology is robust to diverse and complex situations. Therefore, we

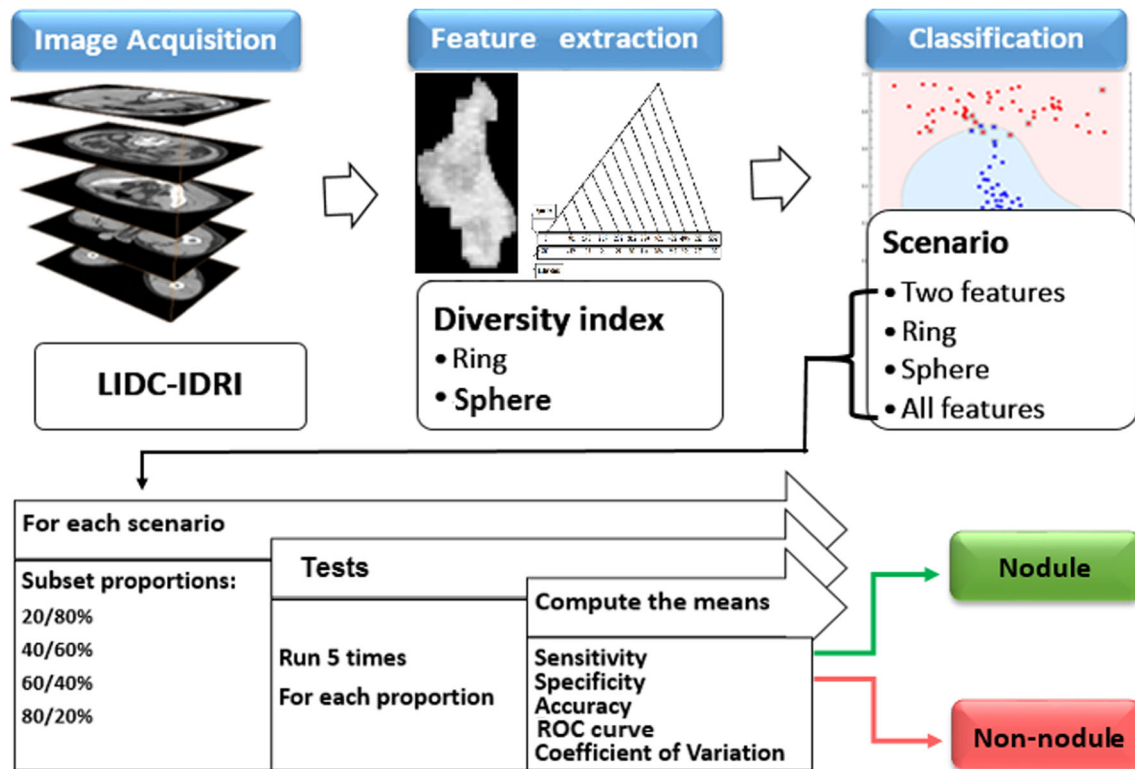


Figure 8 Results analysis flow.

divide the samples in random proportions for training/testing, going from the worst case (20/80 %) to the best case (80/20 %). From the methodology’s validation strategy, we showed that the methodology had good performance in all tests.

We are aware that the 80/20 % group, for example, may have a risk of overfitting; however, our purpose with these groups is to show that the methodology performs well with the best and the worst training/testing cases.

This image database has a file containing information about the nodule marking, as performed by four specialists. The nodule contours and characteristics are only marked for nodules between 3 mm and 30 mm. For nodules smaller than 3 mm, the only markings refer to their center of mass. In addition to the markings, the files also provide information about the properties of the lung nodules: smoothness, internal structure, calcification, sphericity, spiculation, texture, and malignancy. All nodules, with no distinction, were used in the training and testing of the proposed methodology. Since the method is based solely on the candidate’s texture information, details concerning shape and size make no difference, because the taxonomic indexes (Δ and Δ^*) are invariant to spatial properties [22].

4.2 Feature Extraction

In all experiments, we applied the techniques and approaches described in Section 3. We applied the indexes

Δ e Δ^* for each candidate, as well as for each approach (ring and sphere). Thus, the composition of the feature base is two features for each candidate and two for each ring and sphere, resulting in fourteen features.

4.3 Classification

In this section, we describe how each feature base is formed for classification, according to the three types of tree described in Section 3. For each tree, we extract 14 features from each nodule/non-nodule candidate. These features are divided into four scenarios:

1. In the first scenario, each candidate is represented by two features extracted using the indexes Δ and Δ^* (two features per candidate).
2. In the second scenario, each candidate is represented by eight features. Two of these features are the same as described for scenario 1, which are joined to six other features extracted using the indexes in the annular regions. For each ring (there are three), we extract two features using the two indexes, (Δ and Δ^*), thus resulting in a total of eight features.
3. In the third scenario, each candidate has eight features. Two of them are as described for scenario 1. The other six are extracted using the indexes in the spherical regions. For each sphere (from a total of three), we

Table 2 Results of experiments for the tree 1.

	MA	CV	MS	MSP	MA	CV	MS	MSP
Training/test	Scenario 1				Scenario 2			
20/80	94.82	0.31 %	89.86	93.72	96.46	0.34 %	91.28	95.25
40/60	95.09	0.25 %	89.58	93.84	96.84	0.21 %	92.72	95.89
60/40	95.26	0.17 %	90.24	94.12	96.92	0.13 %	93.84	96.2
80/20	95.45	0.46 %	90.06	94.24	97.32	0.37 %	93.81	96.5
CV-m	0.28 %	–	0.31 %	0.26 %	0.36 %	–	1.30 %	0.56 %
	Scenario 3				Scenario 4			
20/80	96.1	0.10 %	93	95.40	96.68	0.22 %	93.04	95.84
40/60	96.85	0.18 %	94.11	96.20	97.25	0.27 %	94.27	96.55
60/40	97.01	0.25 %	94.9	96.47	97.46	0.34 %	95.48	97.00
80/20	97.08	0.12 %	94.97	96.60	97.30	0.21 %	95.59	96.91
CV-m	0.47 %	–	0.97 %	0.56 %	0.35 %	–	1.27 %	0.55 %

Mean Accuracy (MA), Coefficient of Variation in each ratio (CV), Mean Sensitivity (MS), Mean Specificity (MSP) and Coefficient of Variation measurements in the proportions (CV-m).

extracted two features using the two indexes (Δ and Δ^*), thus resulting in eight measurements.

- In the fourth and last scenario, we join all the features. Thus, each candidate is now represented by 14 features. They include two features extracted in the first scenario, six extracted from the rings (scenario 2), and six extracted from the spheres (scenario 3), which results in 14 measurements.

After the composition of each scenario, we made the divisions according to the proportions described in Section 4.1. Five tests were carried out for each proportion. For the results analysis, only the means of each proportion are presented.

In the next sections, we present the results for the three trees, highlighting in bold the best and the worst results of each tree.

4.3.1 Tree 1

For the experiments of tree 1 (Table 2), we obtained the best mean accuracy of 97.46 % for the 60/40 % proportion present in scenario 4, with a variation coefficient around 0.34 %; meaning that the dispersion had little variation with respect to the mean accuracy. The mean accuracy and specificity were, respectively, 95.48 % and 97 %, with a mean area under the ROC curve of 0.952. In the worst case for the experiments of this tree, we have scenario 1 with the 20/80 % configuration presenting a 94.82 % mean accuracy, with a mean sensitivity of 89.86 %, and a mean specificity of 93.72 %. The coefficient of variation was 0.31 %. In addition to good average accuracy in the best and worst cases,

the low values of the coefficients of variation of the CV-m allow us to affirm that this tree presents good results, regardless of the proportion and scenario used.

4.3.2 Tree 2

The data presented in Table 3 show the results for the means of accuracy, sensitivity, and specificity for the five tests performed on each scenario and each proportion, as well as the coefficients of variation for the mean accuracy values.

For tree 2, the best mean accuracy was 99.22 % in the 80/20 % proportion present in scenario 4. The coefficient of variation was of 0.19 %, showing that the dispersion had very little variation with respect to the mean accuracy. The mean sensitivity and specificity were 98 % and 98.82 %, respectively, with an area under the ROC curve of 0.985. As the worst result for the tree 2 experiments, we found scenario 1 in the 60/40 % configuration, presenting 89.72 % mean accuracy, with a mean sensitivity and specificity of 84.61 % and 88.95 %, respectively, and, finally, a coefficient of variation of 0.4 %. We obtained expressive results with this tree, with a high average accuracy in all scenarios and proportions, as well as low average coefficients of variation (CV-m).

4.3.3 Tree 3

Table 4 contains the results of the five tests performed for each scenario and proportion, calculated as the mean values of sensitivity, specificity, accuracy, and coefficients of variation.

Table 3 Results of experiments for the tree 2.

	MA	CV	MS	MSp	MA	DP	MS	MSp
Training/test	Scenario 1				Scenario 2			
20/80	90.93	0.95 %	81.93	88.61	99.43	0.16 %	91.34	97.25
40/60	90.32	0.86 %	84.54	88.81	99.55	0.09 %	91.74	97.43
60/40	89.72	0.40 %	84.61	88.72	99.59	0.11 %	91.9	97.56
80/20	89.96	0.90 %	84.62	88.95	99.56	0.17 %	92.51	97.73
CV-m	0.58 %	–	1.59 %	0.16 %	0.07 %	–	0.53 %	0.21 %
	Scenario 3				Scenario 4			
20/80	99.17	0.27 %	91.56	97.18	99.36	0.13 %	93.51	97.89
40/60	98.99	0.05 %	94.76	97.94	99.12	0.12 %	96.57	98.49
60/40	99.17	0.11 %	95.38	98.2	99.33	0.17 %	96.79	98.64
80/20	99.2	0.28 %	95.2	98.43	99.22	0.19 %	98	98.92
CV-m	0.10 %	–	1.91 %	0.55 %	0.11 %	–	1.99 %	0.44 %

Mean Accuracy (MA), Coefficient of Variation in each ratio (CV), Mean Sensitivity (MS), Mean Specificity (MSp) and Coefficient of Variation measurements in the proportions (CV-m).

Tree 3 presents its best mean accuracy of 97.65 % using the 60/40 % proportion of scenario 4; the coefficient of variation is around 0.17 %, indicating a little dispersion with respect to the mean accuracy. The mean sensitivity and specificity are 95.24 % and 97.10 %, respectively, with an area under the ROC curve of 0.952. As the worst results of tree 3, we highlight the one present on scenario 1, using the proportion 20/80 %, with a 94.88 % mean accuracy, a mean sensitivity and specificity of 89.35 % and 93.64 %, respectively, and a coefficient of variation of 0.23 %. In summary, this tree shows its effectiveness by the elevated values of the mean accuracy and the low variation in its mean coefficients (CV-m). Therefore, good results can be obtained regardless of the scenario or proportion used.

Figure 9 presents a dispersion graph with the best and worst results of each tree, where we can observe the variation of the standard deviation (indicated by the bar) with respect to the mean accuracy (red dot). The experiments with the best results (“b”), except t2-b-e15, had little variation. This indicates little difference on the accuracy found in each experiment, obtaining a mean standard deviation of 0.142. As a negative characteristic, we note experiments t2-w-e2 and t2-w-e4, both on tree 2, which resulted in a mean standard deviation of 0.790 and an oscillation of almost 1 %.

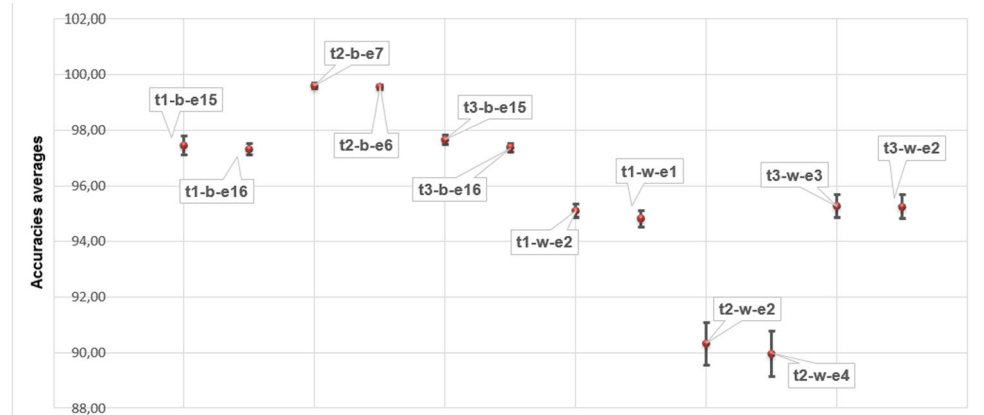
Table 5 presents the SVM parameters for the best results of each tree; i.e., parameters C and λ of the five tests comprising each experiment performed. To ensure that the RBF kernel of the SVM used by our methodology showed the

Table 4 Results of experiments for the tree 3.

	MA	CV	MS	MSp	MA	CV	MS	MSp
Training/test	Scenario 1				Scenario 2			
20/80	94.88	0.23 %	89.35	93.64	96.27	0.37 %	91.45	95.15
40/60	95.25	0.44 %	89.64	93.96	96.91	0.46 %	92.55	95.87
60/40	95.27	0.44 %	89.62	93.97	96.94	0.27 %	93.78	96.21
80/20	95.53	0.34 %	89.8	94.2	97.20	0.22 %	94.37	96.54
CV-m	0.28 %	–	0.21 %	0.24 %	0.41 %	–	1.40 %	0.62 %
	Scenario 3				Scenario 4			
20/80	95.88	0.30 %	92.75	95.17	96.67	0.20 %	92.51	95.71
40/60	96.58	0.35 %	93.54	95.88	97.19	0.21 %	94.32	96.52
60/40	96.86	0.20 %	95.00	96.43	97.65	0.17 %	95.24	97.10
80/20	97.08	0.31 %	95.15	96.65	97.37	0.15 %	95.55	96.95
CV-m	0.54 %	–	1.23 %	0.69 %	0.42 %	–	1.45 %	0.65 %

Mean Accuracy (MA), Coefficient of Variation in each ratio (CV), Mean Sensitivity (MS), Mean Specificity (MSp) and Coefficient of Variation measurements in the proportions (CV-m).

Figure 9 Dispersion graph. Legend: “t”? = tree. “b”? = best or “w” = worst. “e” = experiment number.



best results, we performed the tests with the best result of each tree. Table 6 shows the average results for the accuracy, sensitivity, and specificity of each kernel.

To reduce the errors and increase the generalizability of our methodology, we conducted another test procedure using the *k-fold* test methodology, assuming $k = 10$. Table 7 presents the results for the three trees.

To ensure that the result of our approach is promising. We performed tests with three classic texture-analysis techniques: histogram (first-order statistics) [18], Gray-Level Co-Occurrence Matrix (second-order statistics) [14, 33], and Gray-Level Run Lengths (high-order statistics) [18]. As can be seen in the Table 8, the proposed methodology had better results than the other techniques.

To show that our methodology is promising and has good generalization ability, new experiments were performed with all the features representing each VOI; i.e., 14 measures per VOI. Nonetheless, only the most significant features were selected for inclusion in the database. The selection was done using the stepwise technique [32]. Table 9 shows the descriptors that were selected by stepwise for each tree. Moreover, Table 10 shows the results for the performed tests, taking into account the selected measures using SVM with the RBF kernel.

As can be seen in Table 10, after selecting the best measures, the results can reach significant values; therefore, we were very close to the best results of each tree. We believe that the results, together with all the measures, showed a better performance because one feature might highlight (emphasize) something that another might not.

Concerning with χ^2 test, we will assume that H_0 is: *Regardless of the tree used, we will have a good result; i.e., we can obtain similar or approximated values.* For a level of significance (α of 0.05), we have two degrees of freedom (G), and the value of $\chi^2 E$ is 0.3. We mapped the value of $\chi^2 E$ established, and found it as 5.99. $\chi^2 E$ is lower than $\chi^2 T$; therefore, we accept hypothesis H_0 as true [5].

In the experiments described in Tables 2, 3 and 4, we found that the best results presented values above 97 % for the average rate of accuracy, 95 % for mean sensitivity, and 96 % for mean specificity. The highest value obtained among all the experiments was 99.22 % for the mean accuracy in tree 2.

We believe that tree 2 presents the best result because of the elimination of species with no individuals. Therefore, in a community in which the species actually have individuals and are organized according to them, the diversity among the species becomes higher, as shown in [22].

Table 5 SVM parameters for the best results of the three trees.

Tree	1		2		3	
	C	λ	C	λ	C	λ
1	64	0.012332	256	0.0129283	128	0.2432132
2	128	0.201239	512	0.10109874	512	0.18896578
3	1024	0.21203	128	0.0982348	512	0.121353
4	512	0.112234	1024	0.12123923	256	0.012
5	256	0.0102309	512	0.123762	2048	0.0223723

Table 6 Results for other SVM kernels.

Tree	1			2			3		
	MA	MS	MSP	MA	MS	MSP	MA	MS	MSP
Kernel	MA	MS	MSP	MA	MS	MSP	MA	MS	MSP
Linear	87.11	59.22	96.05	92.47	79.22	97.01	87.78	60.25	96.21
Polynomial	81.78	30.11	98.33	90.83	68.08	98.62	82.68	31.55	98.42
Sigmoidal	66.66	40.11	75.17	80.38	42.25	93.44	66.64	39.07	75.14
RBF	97.46	95.48	97	99.22	98	98.92	97.65	95.24	97.10

Mean Accuracy (MA), Mean Sensitivity (MS), and Mean Specificity (MSP).

To reliably compare the three trees described by our methodology, we performed a final test with the same database using the 80/20 % proportion. To do so, we prepared five random bases respecting the established proportion, after which we performed the experiments in each base with each tree.

Table 11 presents the means of sensitivity, specificity, accuracy, and coefficients of variation and, finally, the result of the χ^2 test for $\alpha = 0.05$ and $G = 2$, considering that H_0 is: *Regardless of the tree, the result will be the same or nearly the same.* If we compare the obtained values according to [5], we have: $\chi^2 E$ calculated = 0.5 and the $\chi^2 T$ established = 5.99. By doing so, the value $\chi^2 E$ is lower than $\chi^2 T$, it shows that the hypothesis H_0 is true.

To ascertain our results more rigorously, we performed the χ^2 test for all proportions of each tree. In other words, from the five tests that composed each mean of the results, we analyzed only the best one for each proportion of the three trees. Table 12 shows the values calculated for χ^2 . These values allow us to ascertain even more our hypothesis H_0 , for an α of 0.05 and degree 2.

The promising results presented in Tables 2, 3, 4 and 11 show the high rates of correct detections achieved by the indexes. One reason for this is related to the distribution of the HU of the non-nodules, which is more heterogeneous

(high diversity, many species) than the nodules. That is, a nodule region has lower diversity, since there is a uniformity in the HU that forms the carcinogenic region, which is an aspect not found in non-nodule regions. These differences between the number of species present in the nodule and non-nodule regions are strongly highlighted by the computation of the indexes, enabling the SVM to successfully converge in the separation of classes.

4.4 Comparison with Other Related Works

The comparison with other works in the area was difficult, since none of the works cited in this article supplied the exams used. The only piece of information provided was the database used. Therefore, we were unable to perform a rigorous evaluation of our method with respect to other works.

Our objective with Table 13 is to provide an overview (exam database, complexity of the methodology, etc.) of the results found in the related works and in our work. Thus, we intend to show that our methodology is promising since, compared to other works, we achieved results above 97 % for various types of situations: 1) classification using only texture; 2) large and complex samples; and 3) several sample configurations for training and testing.

Table 7 Results for k-fold tests.

Tree	MA	MS	MSP
1	94.84	95.5	96.27
2	95.16	96.18	97.07
3	94.91	96.01	96.89

Mean Accuracy (MA), Mean Sensitivity (MS), and Mean Specificity (MSP).

Table 8 Comparison of results with other texture descriptors.

Technique	MA	MS	MSP
Histogram	81.91	72.4	84.44
Gray-Level Co-Occurrence Matrix	84.26	75.54	86.18
Gray Level Run Lengths	85.34	81.55	86.56
Proposed method	99.2	98	97.69

Mean Accuracy (MA), Mean Sensitivity (MS), and Mean Specificity (MSP).

Table 9 Results of stepwise feature selection.

Features	Tree 1	Tree 2	Tree 3
Δ	X	X	X
Δ^*	X	X	X
Ring 1 + Δ			X
Ring 2 + Δ	X		
Ring 3 + Δ		X	
Ring 1 + Δ^*			
Ring 2 + Δ^*		X	
Ring 3 + Δ^*	X		X
Sphere 1 + Δ	X	X	
Sphere 2 + Δ			X
Sphere 3 + Δ			
Sphere 1 + Δ^*		X	X
Sphere 2 + Δ^*	X		
Sphere 3 + Δ^*			

The best result of the means of each tree can be analyzed in Table 13, which shows our improvements in comparison with most of the other works. Even if we compared the same number of exams used on the methodologies, only studies [7, 29] used the same set of images. For comparison with those that developed CADs, we will refer only to the classification stage. The CAD developed by [29] shows a value inferior to those presented here for the three trees for sensitivity, specificity, and accuracy. However, for the methodology presented by [7], we highlight the comparison for the results obtained in tree 2, which presents superior values for the tree measures. In summary, the proposed methodology reached a mean accuracy comparable to the best results found in recent literature for the classification of lung-nodule candidates.

4.5 Discussion

The proposed methodology was evaluated by applying a set of 833 exams from the LIDC-IDRI database, divided into training/testing proportions of 20/80 %, 40/60 %, 60/40 %, and 80/20 %.

Table 10 Results for tests after feature selection.

Tree	MA	MS	MSP
1	94.24	95.5	95.97
2	97.09	97.87	98.12
3	94.94	94.97	96.29

Mean Accuracy (MA), Mean Sensitivity (MS), and Mean Specificity (MSP).

Table 11 Experimental results for the three trees.

Tree	MA	CV	MS	MSP
1	97.79	0.27 %	99	99.39
2	98.84	0.23 %	99.24	99.37
3	98	0.37 %	98.98	99.27
χ^2			0.5	

Mean Accuracy (MA), Mean Sensitivity (MS), Mean Specificity (MSP), Coefficient of Variation of mean Accuracy (CV), and value of χ^2 .

and 80/20 %. The experimental results allow the formulation of the following conclusions:

1. The use of taxonomic indexes Δ and Δ^* combined with phylogenetic trees lead to good results in lung-nodule classification. We believe that this kind of indexes deserve diversified studies and tests.
2. The use of regions extracted based on rings and spheres allowed good individual results, but, when combined, presented the best result among all the trees. In other words, with these approaches, we were able to analyze individually, with higher detail, each nodule and non-nodule region.
3. Using only texture for nodule and non-nodule characterization, the taxonomic indexes Δ and Δ^* in combination with the phylogenetic trees, presented a good result, independently of the analyzed form.
4. The large number of individuals found does not compromise the methodology, since the taxonomic indexes Δ and Δ^* have the advantage of being independent of the sample effort (number of individuals) [22].
5. Finally, it is important to highlight that the LIDC-IDRI database is extremely complex and diversified; i.e., contains countless different cases of lung nodules. This database has exams that were extracted by various tomographs, making it harder to detect, classify, or even diagnose through CAD/CADx systems [7, 29].

All these items add value to our methodology. The texture-analysis properties, through the taxonomic indexes

Table 12 Result of the χ^2 for all proportions.

Proportion	χ^2
20/80	0.86
40/60	0.78
60/40	0.48
80/20	0.42

Table 13 Comparison of results with related works.

Work	Descriptors/classifiers	Database	Ac %	Se%	Sp%
[15]	Adaptive threshold algorithm, features of shape, intensity, and gradient, Fisher Discriminant Linear classifier.	JSRT	–	78.1	–
[19]	Random forest, clustering.	LIDC	–	97.33	97.11
[2]	Machine learning, Wavelet Transform, K-Near neighbor, K-means.	JSRT	96.58	–	–
[10]	Genetic-algorithm template matching, geometric shape, distribution of intensity.	–	–	85	–
[13]	Scale-Invariant Feature Transform, Local Binary Pattern, Principal Component Analysis, Linear Discriminant Analysis.	ELCAP/LI	–	86	97
[17]	Feature based on shape and texture, combines rule-based, SVM.	3AHG	84.39	–	92
[25]	Measurements based on texture, histogram, SVM.	NBIA/ELCAP	82.66	96.15	52.17
[21]	Threshold method, voxel feature, SVM.	NSRCT-LUNG/LIDC	87.8	93.75	87.6
[29]	Fuzzy clustering models, SVM with three kernels (linear, polynomial, and RBF)	LIDC-IDRI	80.36	82.05	76.47
[7]	Quality threshold clustering, genetic algorithm, diversity index, SVM.	LIDC-IDRI	97.55	85.91	97.7
[31]	2D-PCA, minimum redundancy maximum relevance, ANN, RF, Bagging, and AdaBoost.	University of Istanbul	90.7	89.6	87.5
[1]	Texture features based on statistical techniques and SVM.	LIDC	–	96.31	–
Tree 1	Phylogenetic tree, Δ, Δ^*, texture features		97.4	95.48	97
Tree 2	With spheres and rings, SVM. These techniques	LIDC-IDRI	99.2	98	98.92
Tree 3	Were used for all trees.		97.6	95.24	97.1

Accuracy (Ac), Sensitivity (Se), Specificity (Sp).

of diversity (Δ) and distinction (Δ^*) combined with the phylogenetic trees, showed a good response to the experiments. Moreover, the complexity of the LIDC-IDRI database allows a more precise conclusion to the results.

5 Conclusion

Lung cancer stands out as having the highest mortality rate, as well as one of the lowest survival rates after the diagnosis

(5 years, for 14 % to 20 % of the patients). Early diagnosis represents a considerable increase in the patient's survival probability. The present work showed a methodology for classifying lung nodules based on the taxonomic diversity (Δ) and distinction (Δ^*) indexes in conjunction with phylogenetic trees, and used a Support Vector Machine to classify lung-nodule candidates into nodules and non-nodules; thus, proving to be a useful tool for specialists.

The attained results demonstrate the promising performance of the texture-extraction techniques by the indexes presented herein. Another important factor with respect to the good results was the creation of the phylogenetic tree. This tree contributes greatly to the discrimination between nodules and non-nodules. Although the database used is robust and ensures a great diversity of nodules to be analyzed, more tests with other databases are needed to improve our methodology, making it more robust and generic.

Lastly, the methodology presented in this work can integrate a CAD/CADx tool for the detection and diagnosis of lung cancer, with the intent of classifying segments suspected to be nodules and non-nodules, and thus make the analysis of exams by the specialist more flexible and less exhausting.

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

References

- Akram, S., Javed, M.Y., Hussain, A., Riaz, F., & Akram, M.U. (2015). Intensity-based statistical features for classification of lungs ct scan nodules using artificial intelligence techniques. *Journal of Experimental & Theoretical Artificial Intelligence*, 27(6), 737–751. doi:10.1080/0952813X.2015.1020526.
- Al-Absi, H., Samir, B., Shaban, K., & Sulaiman, S. (2012). Computer aided diagnosis system based on machine learning techniques for lung cancer. In *2012 International conference on computer information science (ICIS)* (Vol. 1, pp. 295–300). doi:10.1109/ICISci.2012.6297257.
- Armato, S.G., McLennan, G., Bidaut, L., McNitt-Gray, M.F., Meyer, C.R., Reeves, A.P., Zhao, B., Aberle, D.R., Henschke, C.I., Hoffman, E.A., Kazerooni, E.A., MacMahon, H., Van Beeke, E.J.R., Yankelevitz, D., Biancardi, A.M., Bland, P.H., Brown, M.S., Engelmann, R.M., Laderach, G.E., Max, D., Pais, R.C., Qing, D.P.Y., Roberts, R.Y., Smith, A.R., Starkey, A., Batrah, P., Caligiuri, P., Farooqi, A., Gladish, G.W., Jude, C.M., Munden, R.F., Petkovska, I., Quint, L.E., Schwartz, L.H., Sundaram, B., Dodd, L.E., Fenimore, C., Gur, D., Petrick, N., Freymann, J., Kirby, J., Hughes, B., Castele, A.V., Gupte, S., Sallamm, M., Heath, M.D., Kuhn, M.H., Dharaia, E., Burns, R., Fryd, D.S., Salganicoff, M., Anand, V., Shreter, U., Vastagh, S., & Croft, B.Y. (2011). The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. *Medical Physiology*, 38(2), 915–931. <http://www.biomedsearch.com/nih/Lung-Image-Database-Consortium-LIDC/21452728.html>.
- Baxeavanis, A.D., & Ouellette, B.F.F. (2004). *Bioinformatics: a practical guide to the analysis of genes and proteins. Methods of biochemical analysis*. Wiley. <http://books.google.com.br/books?id=ghUZaEAdHUC>.
- Bolboaca, S.D., Jantschi, L., Sestraa, A.F., Sestra, R.E., & Pamfil, D.C. (2011). Pearson-fisher chi-square statistic revisited. *Information*, 2(3), 528–545. doi:10.3390/info2030528. <http://www.mdpi.com/2078-2489/2/3/528>.
- Câncer, I.N. (2014). Estimativas da incidência e mortalidade por câncer no brasil. Available: <http://www.inca.gov.br/estimativa/2012/>. (Accessed: 1 January 2014).
- de Carvalho Filho, A.O., de Sampaio, W.B., Silva, A.C., de Paiva, A.C., Nunes, R.A., & Gattass, M. (2013). Automatic detection of solitary lung nodules using quality threshold clustering, genetic algorithm and diversity index. *Artificial Intelligence in Medicine*. doi:10.1016/j.artmed.2013.11.002. <http://www.sciencedirect.com/science/article/pii/S0933365713001541>.
- Chang, C.C., & Lin, C.J. LIBSVM—a library for support vector machines (2013). Available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>.
- Chen, W., Li, Z., Bai, L., & Lin, Y. (2011). Nf-kappab in lung cancer, a carcinogenesis mediator and a prevention and therapy target. *Frontiers in Bioscience (Landmark edition)*, 16, 1172–1185. doi:10.2741/3782.
- Dehmeshki, J., Ye, X., Casique, M.V., & Lin, X. (2006). A hybrid approach for automated detection of lung nodules in ct images. In *ISBI* (pp. 506–509). IEEE. <http://dblp.uni-trier.de/db/conf/isbi/isbi2006.html>.
- Duda, R.O., & Hart, P.E. (1973). *Pattern classification and scene analysis*. New York: Wiley-Interscience Publication.
- van Erkel, A., & Pattynama, P. (1998). Receiver operating characteristic (ROC) analysis: basic principles and applications in radiology. *European Journal of Radiology*, 27(2), 88–94.
- 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* (pp. 169–172). doi:10.1109/ISBI.2011.5872380.
- Galloway, M.M. (1975). Texture analysis using gray level run lengths. *Computer Graphics and Image Processing*, 4(2), 172–179. doi:10.1016/S0146-664X(75)80008-6. <http://www.sciencedirect.com/science/article/pii/S0146664X75800086>.
- Hardie, R.C., Rogers, S.K., Wilson, T.A., & Rogers, A. (2008). Performance analysis of a new computer aided detection system for identifying lung nodules on chest radiographs. *Medical Image Analysis*, 12(3), 240–258. <http://dblp.uni-trier.de/db/journals/mia/mia12.html#HardieRWR08>.
- Huang, P.W., Lin, P.L., Lee, C.H., & Kuo, C. (2013). A classification system of lung nodules in ct images based on fractional brownian motion model. In *2013 International conference on system science and engineering (ICSSE)* (pp. 37–40). doi:10.1109/ICSSE.2013.6614710.

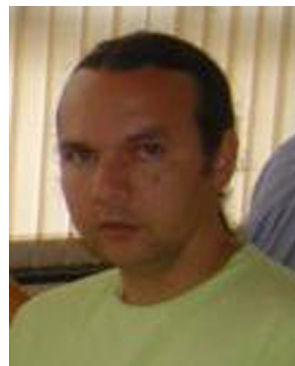
17. Jing, Z., Bin, L., & Lianfang, T. (2010). Lung nodule classification combining rule-based and svm. In *2010 IEEE fifth international conference on bio-inspired computing: theories and applications (BIC-TA)* (pp. 1033–1036). doi:10.1109/BICTA.2010.5645114.
18. King, P.H. (2012). Digital image processing and analysis: Human and computer applications with cviptools, 2nd edition (umbaugh, s.; 2011) [book reviews]. *IEEE Pulse*, 3(4), 84–85. doi:10.1109/MPUL.2012.2196843.
19. Lee, S., Kouzani, A., & Hu, E. (2010). Random forest based lung nodule classification aided by clustering. *Computerized Medical Imaging and Graphics*, 34(7), 535–542. doi:10.1016/j.compmedimag.2010.03.006. <http://www.sciencedirect.com/science/article/pii/S0895611110000418>.
20. Leef, J. 3rd, & Klein, J. (2002). The solitary pulmonary nodule. *Radiologic Clinics of North America*, 40(1), 123–143, ix. doi:10.1056/NEJMcp012290.
21. Liu, Y., Yang, J., Zhao, D., & Liu, J. (2009). Computer aided detection of lung nodules based on voxel analysis utilizing support vector machines. In *International conference on future biomedical information engineering, 2009. FBIE 2009* (pp. 90–93).
22. Magurran, A.E. (2004). Measuring biological diversity. *African Journal of Aquatic Science*, 29(2), 285–286.
23. Moura, H., & Viana, G. (2011). Phylogenetic trees drawing web service. In *BIOTECHNO 2011, the third international conference on bioinformatics, biocomputational systems and biotechnologies* (pp. 73–77).
24. Netto, S.M.B., Silva, A.C., Nunes, R.A., & Gattass, M. (2012). Automatic segmentation of lung nodules with growing neural gas and support vector machine. *Computers in Biology and Medicine*, 42(11), 1110–1121. doi:10.1016/j.combiomed.2012.09.003.
25. Orozco, H., Osiris Vergara Villegas, O., Maynez, L., Sanchez, V., & 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)* (pp. 870–875). doi:10.1109/ISSPA.2012.6310676.
26. Pienkowski, M.W., Watkinson, A.R., Kerby, G., Clarke, K.R., & Warwick, R.M. (1998). A taxonomic distinctness index and its statistical properties. *Journal of Applied Ecology*, 35(4), 523–531. doi:10.1046/j.1365-2664.1998.3540523.x.
27. Schölkopf, B., & Smola, A. (2002). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. Cambridge: MIT Press.
28. da Silva, I.A., & Batalha, M.A. (2006). Taxonomic distinctness and diversity of a hyperseasonal savanna in central brazil. *Diversity and Distributions*, 12(6), 725–730. doi:10.1111/j.1472-4642.2006.00264.x.
29. Sivakumar, S., & Chandrasekar, C. (2013). Lung nodule detection using fuzzy clustering and support vector machines. *International Journal of Engineering and Technology (IJET)*, 5(11), 179–185.
30. Soliman, A.A., Abd Ellah, A.H., Abou-Elheggag, N.A., & Modhesh, A.A. (2012). Estimation of the coefficient of variation for non-normal model using progressive first-failure-censoring data. *Journal of Applied Statistics*, 39(12), 2741–2758. <http://EconPapers.repec.org/RePEc:taf:japsta:v:39:y:2012:i:12:p:2741-2758>.
31. Tartar, A., Kilic, N., & Akan, A. (2013). Classification of pulmonary nodules by using hybrid features. *Computational and Mathematical Methods in Medicine*, 2013, 148363. doi:10.1155/2013/148363. <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3708407/>.
32. Wagner, J.M., & Shimshak, D.G. (2007). Stepwise selection of variables in data envelopment analysis: Procedures and managerial perspectives. *European Journal of Operational Research*, 180(1), 57–67. doi:10.1016/j.ejor.2006.02.048. <http://www.sciencedirect.com/science/article/pii/S0377221706002839>.
33. Walker, R.F., Jackway, P.T., & Longstaff, I.D. (1997). Recent developments in the use of the co-occurrence matrix for texture recognition. In *1997 13th international conference on digital signal processing proceedings, 1997. DSP 97* (Vol. 1, pp. 63–65). doi:10.1109/ICDSP.1997.627968.
34. Ye, X., Lin, X., Dehmeshki, J., Slabaugh, G., & Beddoe, G. (2009). Shape-based computer-aided detection of lung nodules in thoracic ct images. *IEEE Transactions on Biomedical Engineering*, 56(7), 1810–1820. doi:10.1109/TBME.2009.2017027.



Antonio Oseas de Carvalho Filho received the Master degree in Science of Computing at Federal University of Maranhão -Brazil in 2013. Currently he is a Professor at the Federal University of Maranhão (UFMA).



Aristófaes Corrêa Silva received a PhD degree 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.



Anselmo Cardoso de Paiva received BSc in civil engineering from Maranhão State University -Brazil in 1990, a MSc in civil engineering-Structures and a PhD in Informatics from Pontifical Catholic University of Rio de Janeiro - Brazil in 1993 and 2002. Currently he is a Professor at the Federal University of Maranhão (UFMA), Brazil.



Rodolfo Acatauassú Nunes received a PhD degree in General Surgery - Thoracic Area, from Federal University of Rio de Janeiro in 1995. Currently he is Professor of General Surgery Department at Universidade do Estado do Rio de Janeiro (UERJ).



Marcelo Gattass took his Ph.D. in 1982 from Cornell University and is a full professor at PUC-Rio's Computer Science Department. He is also the Director of Tecgraf/PUC-Rio - Computer Graphics Technology Laboratory.