Med Biol Eng Comput (2017) 55:1129–1146
DOI 10.1007/s11517-016-1577-7

ORIGINAL ARTICLE

Computer-aided diagnosis system for lung nodules based on computed tomography using shape analysis, a 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]

Received: 9 September 2015 / Accepted: 21 September 2016 / Published online: 3 October 2016

Abstract Lung cancer is the major cause of death among patients with cancer worldwide. This work is intended to develop a methodology for the 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. To differentiate the patterns of malignant and benign forms, we used a Minkowski functional, distance measures, representation of the vector of points measures, triangulation measures, and Feret diameters. Finally, we applied a genetic algorithm to select the best model and a support vector machine for classification. In the test stage, we applied the proposed methodology to 1405 (394 malignant and 1011 benign) nodules from the LIDC–IDRI database. The proposed methodology shows promising results for diagnosis of malignant and benign forms, achieving accuracy of 93.19 %, sensitivity of 92.75 %, and specificity of 93.33 %. The results are promising and demonstrate a good rate of correct detections using the shape features. Because early detection allows faster therapeutic intervention, and thus a more favorable prognosis for the patient, herein we propose a methodology that contributes to the area.

Keywords Lung cancer · Shape analysis · Genetic algorithm · Medical image

1 Introduction

It has been reported that, worldwide, the number of lung cancer cases has risen by 1.8 million cases yearly in estimation, which corresponds to 13 % of the total cancer cases. Moreover, lung cancer is becoming the leading cause of cancer-related deaths worldwide [5, 31, 37]. A lung nodule is defined as a nearly spherical opacity of up to 3 cm in diameter, surrounded by the pulmonary parenchyma [18]. Lesions larger than 3 cm are called masses and are often malignant [13, 15, 18]. Most lung cancer cases are related to smoking (representing approximately 80 % of the cases), but cases are also related to the aging of society, industrialization, urbanization, pollution, and poor lifestyles [5].

Presently, the most efficient method for defeating lung cancer is diagnosis and treatment at the initial stages, which elevates the post-diagnosis survival rate to approximately 90 % [22]. However, diagnosing nodules using computed tomography (CT) is a delicate task [23], because analysis by a specialist physician is a repetitive and time-consuming process given that CT examinations normally comprise a large number of images to be analyzed. This analysis can

✉ 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, São Luís, MA 65085-580, Brazil

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

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

be very tiring, which may lead to incorrect conclusions [23, 31].

Early detection of lung cancer allows faster therapeutic intervention, and thus a more favorable prognosis to the patient [9, 36]. Furthermore, the larger the amount of information that the specialist has available, the more precise will be the diagnosis. With the goal of increasing the precision of the diagnosis and aid the physician with a second opinion, computational tools that use digital image processing and pattern recognition techniques have been widely explored. Those techniques have been used together to develop computer-aided detection (CAD)/computer-aided diagnostic (CADx) systems [42]. Thus, we propose an automatic methodology to assist the experts that provide an analysis strategy through the analysis of lung nodules using image processing techniques and pattern recognition to help give a second opinion.

In most CADx methodologies, the feature extraction stage is based on: (a) geometry that measures, for example, how circular is the candidate, and (b) texture that describes aspects of the candidate based on its gray-level distribution. Because of the existence of shape patterns that are commonly used by specialists to identify benign and malignant lesions—for example, the degree of spicule, spiculated rate, and surface roughness properties of the nodule; that is, properties widely used by specialists to make a possible diagnosis [13, 15, 18]—we only used a descriptor based on the shape of the lung nodule for classification. To this end, we used a Minkowski functional (MF), distance measures, representation of the vector of points measures, triangulation measures, and Feret diameters.

MF is a morphometric measurement (volume, mean curvature, surface area, and Euler's number) capable of characterizing the nodule shape. Distance measurements allow verification of how regularly spherical is a nodule. The representation of a vector of points-based measurements allows extraction of information about the sphericity or elongation of a nodule based on its medial axis. In contrast, triangulation measurements allow finding deformations/variations in the contour of the nodule. Finally, the Feret diameters provide information about the dimensions of a nodule.

By marking the region that contains the lesion, most specialists are very careful and often end up marking regions of lung tissue that are not compromised. This could cause nodules of different classes to have the same shape behavior. To overcome this problem, we automatically defined the inner region of the lesion so that we could find tissue patterns inside the nodule; by individually analyzing or comparing the inner region with the original nodule, we were able to differentiate between the classes.

We contribute to the area in the following aspects: (a) use of MF to characterize the shape of the nodule; (b) development of new metrics based on the distance between objects, triangulation, representation of the vector of points, and Feret diameters; and finally, (c) creation of an inner region intended to find patterns in lesioned tissues.

Other studies have been conducted in the diagnosis of lung nodules as malignant or benign with the goal of increasing the accuracy rates of detection and diagnosis on CADx systems. Table 8 presents a review of the recent literature in the area of lung nodule diagnosis, and it lists the following details: work, techniques, database, and results. Because detection is a complex task, most of these works do not use shape descriptors only to classify the nodule as benign or malignant; they also combine other descriptors that analyze the texture of the nodule, thus complementing the analyses. Our goal is to examine only the lung nodule shape based on three basic factors: (1) most CAD systems that use segmentation techniques to generate regions through the similarity between the gray levels, i.e., regions with very similar texture. Thus, an analysis based on shape would be more appropriate. (2) Manual or semiautomatic segmentation can result in excess tissue; that is, the specialist might delimit an area that is not part of the lesion where the texture properties contained in this region might result in classification inconsistencies. (3) Because of their humans constraints, the expert evaluates most easily the shape of the nodule in order to classify it as benign or malignant. It is less complex for human eyes to evaluate the morphologic properties of the nodule compared with the complexity found through the analysis of texture. In short, our intention is not to prove that shape measures are better than texture because we believe that they complement each other. However, with the proposed approach, we intend to provide a possible solution to the problems presented above.

This paper is organized as follows. In Sect. 2, we present the methodology used to classify the nodules extracted from CT images into malignant and benign classes using the extraction of shape features, selection of the best model by means of a genetic algorithm (GA), and classification by means of a support vector machine (SVM). In Sects. 3 and 4, we show and discuss the results achieved by the proposed methodology. Finally, in Sect. 5, we present the final remarks about this work.

2 Methodology proposed

In this section, we describe the methodology used to classify lung nodules as either benign or malignant. In Fig. 1, we show all of the stages involved in this classification. The methodology comprises five major stages: The first is the acquisition of images from the Image Database Consortium and Image Database Resource Initiative (LIDC–IDRI) [3].

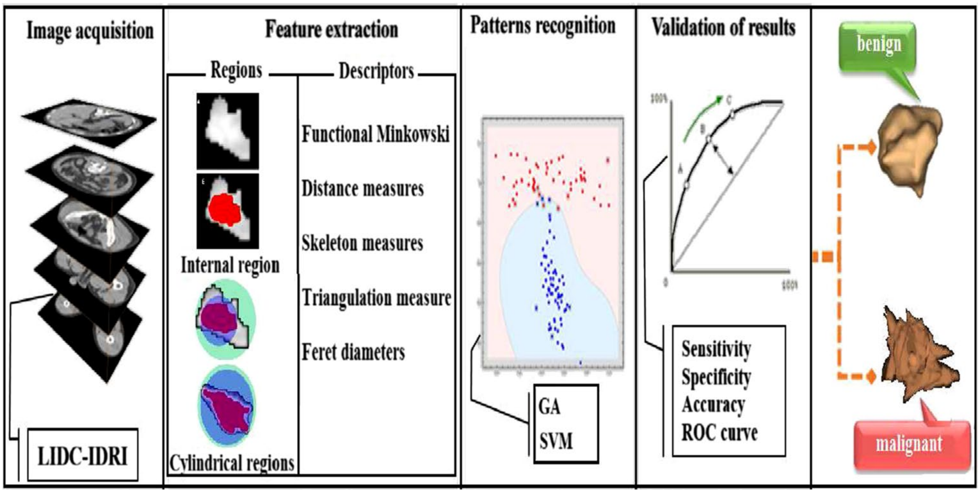

Fig. 1 Proposed methodology

In the second stage, we extract the nodules based on the specialists' markings. Then, we generate the internal region of the nodule. Next, we apply feature extraction from the nodule and its subregions. Finally, we select the best model, execute the classification, and validate the results.

2.1 Image acquisition

The image database used in this work is LIDC–IDRI [3], which is available on the Internet and is the result of an association between the Lung Image Database Consortium and the Image Database Resource Initiative; this database has 1018 CT scans. The CT images were acquired from different tomographs. This fact increases the difficulty in classifying the lung nodules. The database has an XML file that contains markings of nodules contained in each examination made by four experts, in addition to some characteristics, such as sphericity, texture, and malignancy (all indicated by a value from 1 to 5).

There is no imposition for consensus; all nodules indicated by the radiologists are considered and recorded. Therefore, it is possible to have different diagnoses for the same nodule. In this work, we consider only one instance per nodule with the objective of minimizing the impact of subjectivity in examinations. Classification with regard to malignancy or benignity is obtained first with the calculations presented in [19], which summarizes into one single value the nodular features made by up to four specialists by computing the mode or median. According to the result of this summary, in this work, we consider that malignant nodules are those cases that present malignancy semantic values of moderately or highly suspicious nodules, and benign nodules are those cases that present characteristics of highly or moderately benign nodules. Contour is adopted as the one that contains larger bounds. We obtained a total of 1405 nodules (1011 benign and 394 malignant).

2.2 Segmentation of nodules

To segment the nodules, we obtain information about the contour as supplied in an XML file that contains the coordinates of the nodules, along with the analysis of each specialist.

2.3 Internal region

Analyses based on the shape of the lung nodule are widely explored, regardless of whether they are intended to suggest a diagnosis (benign or malignant) [28, 29] or reduce false positives (nodule and non-nodule) [8, 27].

As mentioned in Sect. 2.1, each nodule contained in LIDC–IDRI can have up to four different markings, i.e., one per expert. This occurs because there is no enforcement to reach a consensus: All the nodes indicated by the review of radiologists are ascertained and saved. In order to resolve this issue, we follow the methodology proposed in [19].

Based on this, we only consider one instance per nodule with the goal of minimizing the impact of subjectivity in examinations. However, there is no indication in the annotation of radiologists (XML file) as to which information refers to the same nodule. To resolve this issue, we calculate the center point of the nodules after checking whether the coordinates of this point are in the region of a nodule as determined by another expert. Figure 2 illustrates the process.

In Fig. 2, the colored lines represent the markings set individually by experts. The green square in the center refers to the centroid calculated for the outline of the same color. If the coordinates of the centroid are found in the areas bounded by other experts, we consider in this paper that these areas represent the same nodule; therefore, only one instance should be accepted: that which is related to the marking of the nodule with the greater boundary area. In the simplified example of Fig. 2, the instance of the nodule

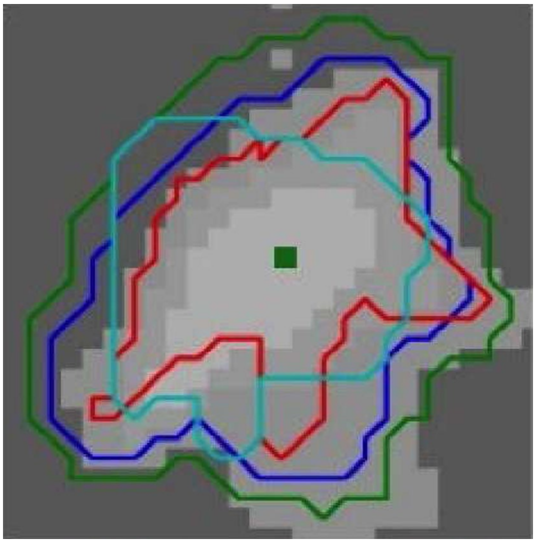

Fig. 2 Illustration summary of nodules [19]

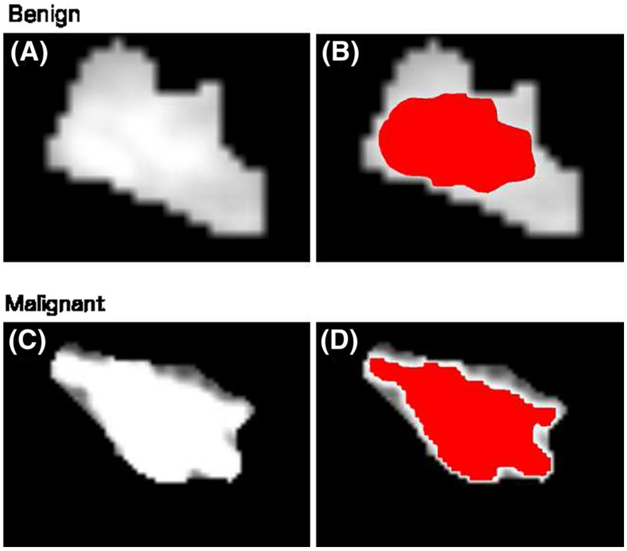

Fig. 3 Examples of benign and malignant nodules with the inner region in *red* (color figure online)

used is the one represented by the green outline because it meets the criteria.

Based on this, this greater region selected to represent the nodule can contain an excess of tissue not affected by cancer cells. Thus, the definition of the nodule is a few times larger than the area/volume of the nodule itself, i.e., the experts usually cautiously delimit a region so that it is larger than required. This fact can be observed in Fig. 3 (in this figure, the region that actually represents the lesion is shown in red). Such a procedure could hinder shape-based analysis, thus leading to errors. As shown in Fig. 3a, c, the nodules marked by the specialists have similar shapes, which can lead to errors in an eventual classification.

In an attempt to alleviate this problem, we propose automatic segmentation that selects only the region that can best represent what we believe to be the nodule, and eliminates or reduces possible excesses in the markings made by experts. To this end, we generate a region/subregion within the nodule by means of the Otsu algorithm [41].

The basic idea of the Otsu algorithm is closer to the histogram of an image of two Gaussian functions; in addition, we choose a threshold for minimizing intra-class variance. In other words, the Otsu algorithm finds the optimal threshold that separates the desired classes; that is, because our goal is simply to separate the region affected by the cancer cells from the healthy tissue, and these similarities are well defined, the Otsu method achieves good results.

Consider a digital image *f* with dimensions $M \times N$ quantized in L gray levels. The first step is to calculate the histogram p of the image given by

$$p_i = \frac{n_i}{MN} \tag{1}$$

where n_i is the amount of *pixels* in image I that has gray intensity i, for $i = 0, \ldots, L-1$. Thus, $MN = n + 0 + n_1 + \cdots + n_{L-1}$ and

$$\sum_{i=0}^{L-1} p_i = 1, p_i \geq 0 \tag{2}$$

Let k the gray level that partitions the histogram of the image into two classes C_1 and C_2, wherein the first and second classes include those pixels whose gray levels are in the interval [0, k] and $[k+1, L-1]$. Herefore, we can set the following probabilities:

- $P_1(k)$ is the gray level of probability k being in class C_1
- $P_2(k)$ is the gray level of probability k being in class C_2

$$P_1(k) = \sum_{i=0}^{k} p_i, \tag{3}$$

$$P_2(k) = \sum_{i=k+1}^{L-1} p_i \tag{4}$$

Because the histogram is approximated by two Gaussian functions,

$$m_1(k) = \sum_{i=0}^{k} iP(i|C_1) \tag{5}$$

and using Bayes rule, we have

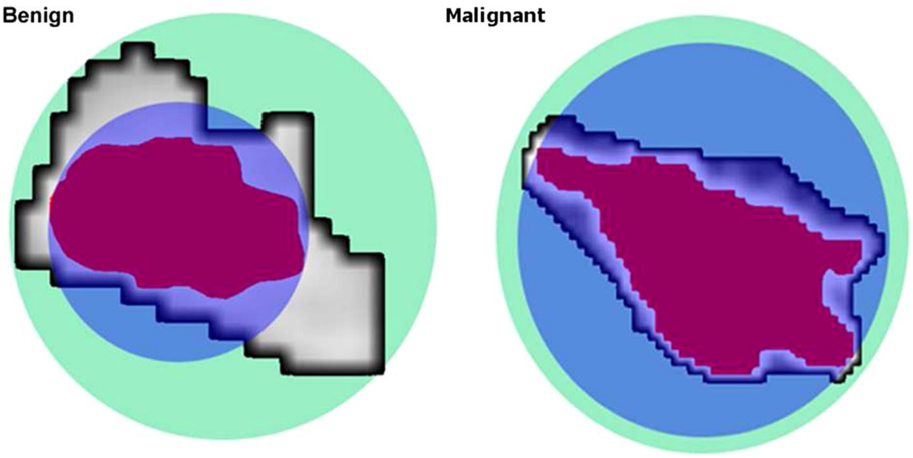

Fig. 4 Regions to be analyzed: *green* external cylinder; *blue* internal cylinder; *red* internal region generated by the Otsu algorithm; remainder: nodule marked by the specialist (color figure online)

$$m_1(k) = \sum_{i=0}^{k} i \frac{P(C_1|i)P(i)}{P(C_1)} \tag{6}$$

where $P(C_1) = P_1(k), P(i)$ is its own p_i and $P(C_1|i)$ is always equal to 1, because i is the gray interval of class C_1. Thus,

$$m_1(k) = \frac{1}{P_1(k)} \sum_{i=0}^{k} ip_i \tag{7}$$

Likewise,

$$m_2(k) = \frac{1}{P_2(k)} \sum_{i=k+1}^{L-1} ip_i \tag{8}$$

In turn, the variance for each probability distribution can be determined as

$$\sigma_1^2(k) = \frac{1}{P_1(k)} \sum_{i=0}^{k} (m_1(k) - p_i)^2 \tag{9}$$

$$\sigma_2^2(k) = \frac{1}{P_2(k)} \sum_{i=k+1}^{L-1} (m_2(k) - p_i)^2 \tag{10}$$

Finally, we determine the intra-class variance relative to gray level k, as follows:

$$\sigma_C^2(k) = \sigma_1^2(k)P_1(k) + \sigma_2^2(k)P_2(k) \tag{11}$$

After calculating σ_C^2 for all values of k, we determine the optimal threshold k^* according to Eq. (12), as follows:

$$k^* = \min_{0 \le k \le L-1} \sigma_C^2(k) \tag{12}$$

The Otsu algorithm was applied to the area inside the specialist's marking. We used a single threshold, thus segmenting the nodule into two regions: the one of interest (the lesion) and the one that includes the rest of the excess tissue delimited by the specialist, which might not contain cancer cells. Figure 3b, d shows the output of the algorithm. In addition to performing automatic delimitation of the inner region of the original marking, it is worthy to mention that the Otsu algorithm is efficient in terms of time because it could accomplish this task with an average of 1.186 s per nodule.

This procedure aims at reducing the possible excess of tissue that any manual marking performed by experts might add; furthermore, it seeks the behaviors of the texture form inside the nodule that can provide more detailed information for each pattern, malignant or benign. In other words, in this region, it is possible to ascertain the forms generated by the texture in the intratumoral region of the nodule, thus making possible the analysis of the form generated by the spread of the cells involved around the lesion, which characterizes the speculation commonly found in malignant nodules.

2.4 Cylindrical regions

In addition to the regions obtained by the internal delimiting performed by the Otsu algorithm, we propose the generation of other complementary regions (cylinders) over each nodule marked by the specialist and its inner region. In doing so, we project cylinders over each region with the objective of extracting the maximum amount of shape information from each separate region and then combine them. Figure 4 exhibits the results of the regions to be analyzed.

The length of the cylinder is the *Z*-axis for each nodule. The radius of the cylinder must be determined before projecting it onto the volume of interest (VOI). To this end, the center of mass of the first slice of the VOI is determined; from here, four line segments are projected: two on the *X*-axis and two on the *Y*-axis (see Fig. 5). The largest segment is then chosen for that slice. This procedure is repeated for all VOI slices. The cylinder is then created using the overall longest segment as the radius. Figure 5 shows the center of mass and the four segments, with the largest segment used to form the cylinder.

After computing the radius, the cylinder is created. Figure 6 shows the result of the cylinder on the object. Figure 6a, b shows examples of a benign and malignant nodule formed by three slices, respectively. The area of the cylinder appears in red.

This is the procedure adopted for the composition of the cylinders used in our methodology.

Our purpose in generating the cylinders is to propose a new perspective for nodule analysis and segmentation by region using the Otsu algorithm. This new approach provides information on the nodule by projecting it on a cylinder, and thus, it becomes possible to determine the shape properties of the nodule using the cylinder. Among them, we have sphericity, cost to form the cylinder starting from the center of mass and border of a nodule proportion measures by relating the nodule with the cylinder.

Thus, we analyze not only the original region of the nodule delimited by the experts, but also the new internal region (delimited by the Otsu algorithm) that presents more reliable information about the shape of the nodule and its cylinders. All the generated regions (Fig. 4) are used throughout the methodology.

2.5 Feature extraction

The delimited regions (Sects. 2.3 and 2.4) are subjected to the feature extraction process solely based on techniques that only describe the shape of the lung nodule.

2.5.1 Minkowski functionals

Minkowski functionals (MFs) are a set of extremely robust and efficient mathematical tools used in integral geometry [24], mathematical morphology [33], and image analysis

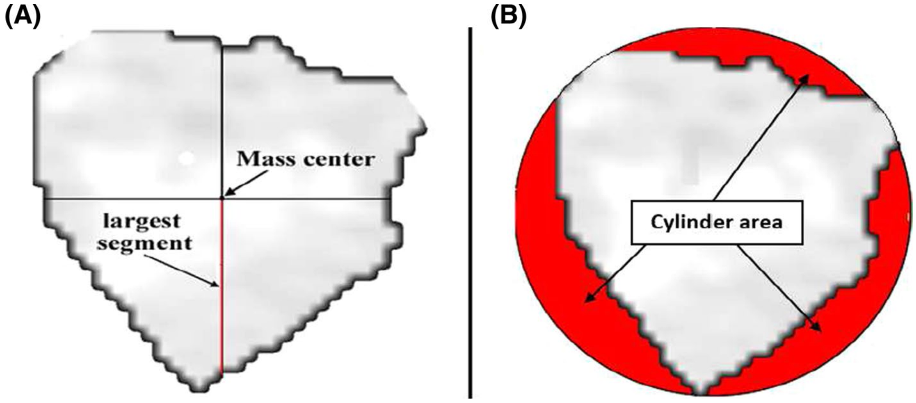

Fig. 5 Radius computation to form the cylinder: **a** nodule with center of mass and line segments and **b** nodule with the created cylinder

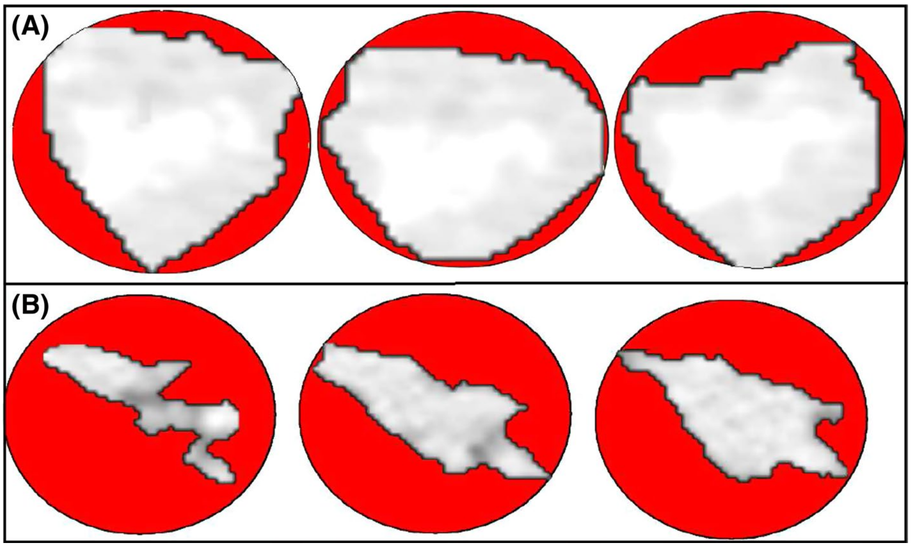

Fig. 6 Example of cylinders. The area of the cylinder is shown in *red*: **a** benign nodule and **b** malignant nodule (color figure online)

[7]. In the case of image analysis, its use is related to the quantitative analysis of morphological patterns [7, 25].

According to the Hugo Hadwiger theorem [16], MFs are applied to sets that present three main properties: additivity, invariance over movement, and continuity. Thus, in a d-dimensional space, the global morphological properties can be fully characterized by $d + 1$ MFs. The quantitative geometrical description is intended to reduce the complexity of the object to a limited number of relevant measurements [32]. These measurements must satisfy three requirements in order to maintain compatibility between the results obtained for different patterns and by different observers. These requirements are:

Additivity The result obtained by the union of two subregions, A and B, must be the same obtained by the sum of the results of two individual subregions minus the intersection between them. This property is important because we do not have to analyze the object as a whole; each region is analyzed separately. This property is defined by:

$$W(A \cup B) = W(A) + W(B) - W(A \cap B) \tag{13}$$

where $W(A)$ is a certain measurement of the set (object) A.

Invariance over movement The results must be constant, even if the object moves or rotates, regardless of the viewer's position.

Continuity Small changes in a certain object under analysis could lead to small changes in its measurements. However, this property must be sufficiently robust to prevent any alterations, such as noises, from interfering with the results.

According to the Hugo Hadwiger theorem [16], a morphological measurement is mathematically defined as a functional when A and B are sets (objects) that satisfy additivity, continuity, and invariance over movement. Thus, a large volume of information obtained by means of morphology patterns can be compressed into a finite number of relevant parameters. Therefore, any property of the type $\varphi(A)$, which depends on the shape of object A, can be written as a linear combination of $d + 1$ MFs [24, 32].

$$\varphi(A) = \sum_{k=0}^{d} c_k W_k(A) \tag{14}$$

where c_k are real-valued coefficients that depend on the $\varphi(A)$ property, but do not depend on object A.

Property 1 *Additivity is a good mathematical property because it allows MFs to be computed simply based on a counting configuration.*

As shown in [26], the MFs are defined for a compact and convex set $A \subset R^3$ by Steiner's formula. Thus, assuming that $A \oplus B_r$ is the dilatation of the set A in a close sphere of radius r with its center in the origin, the volume V of $A \oplus B_r$ can be written as a polynomial function of r as:

$$V(A \oplus B_r) = \sum_{k=0}^{3} W_k(A) r^3 \tag{15}$$

where W_k is the kth Minkowski functional. For example, let C be a cube of side a, then:

$$V(C \oplus B_r) = a^3 + 6a^2 r + 3a\pi r^2 + \frac{4\pi}{3} r^3 \tag{16}$$

With $W_0(C) = a^3$, $W_1(C) = 2a^2$, $W_2(C) = a\pi$, and $W_3(C) = 4\pi/3$, it is possible to notice the relationship between the MFs and the descriptors such as volume (V), surface area (S), mean curvature (B), and Euler's number (χ), respectively.

$$W_0(A) = V(A) \tag{17}$$

$$W_1(A) = \frac{1}{3} S(A) \tag{18}$$

$$W_2(A) = \frac{2}{3} \pi B(A) \tag{19}$$

$$W_3(A) = \frac{4}{3} \pi \chi(A) \tag{20}$$

Because there is a proportion relationship between MFs and the shape/connectivity descriptors (V,S,B, and χ), they are commonly referred to as MFs.

In 3D images, we compute the volume (V), surface area (S), mean curvature (B), and 3D Euler's number (χ). These parameters form the so-called Minkowski functional [7, 24]. Volume is related to the mass of a certain object of interest. Surface area may supply information about the circularity of the object, i.e., low values indicate a higher similarity to a spherical object. The mean curvature is related to the roughness of the surface, i.e., high values indicate a rougher surface and a smaller value indicates a smooth surface. Finally, Euler's number refers to the topology of the structure that is related to the connectivity, indicating the number of structures connected minus the number of holes in these structures.

Using Property 1, the computation of MFs in 3D images is reduced basically to a count of voxels, i.e., count the number of open cubes n_3, open faces n_2, open edges n_1, and vertices n_0 in such a way that, during the count, each object (voxel) is considered only once.

In a discrete structure, such as 3D images, computation of the MFs can be then represented by:

$$\begin{aligned} V &= n_3, \\ B &= \frac{3}{2} n_3 - n_2 + \frac{3}{2} n_1, \\ S &= -6n_3 - 2n_2, \\ \chi &= -n_3 + n_2 - n_1 + n_0 \end{aligned} \tag{21}$$

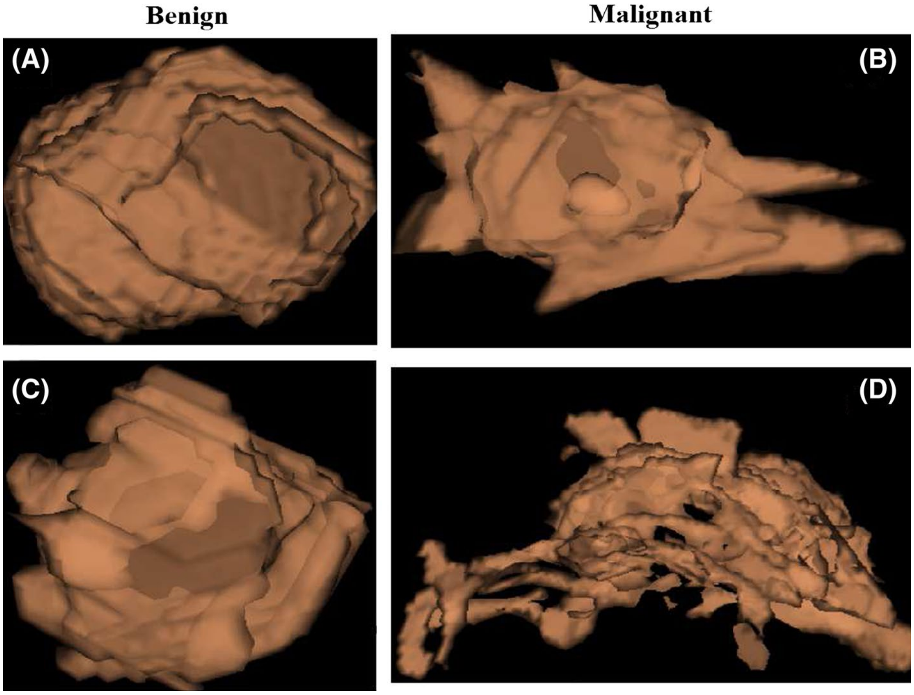

Fig. 7 **a**, **c** Benign patterns; **b**, **d** malignant patterns

2.5.2 *Distance measurements*

Malignant nodules present irregular behavior and various shapes, such as spiculated or branched (Fig. 7b, d), thus indicating a disordered growth of cancer cells. These irregularities are extremely important for methodologies that use the shape of the lung nodule to describe it. However, benign nodules present an opposite shape pattern (Fig. 7a, c): one with a less spiculated shape and fewer branches [13, 15, 18]. Nevertheless, in some cases, these patterns are lost or reduced, i.e., the marking made by the experts is often careful, and so the marked area contains additional tissue not compromised by the lesion (Sect. 2.3).

Distance measurements compute the mean distance cost necessary for objects to reach another closer point from a given starting point.

Equation 22 presents the mean distance cost (Md1) consumed by the internal object to reach the surface of the outer object from its surface.

$$\mathrm{Md1} = \frac{\sum_{i=1}^{n} \mathrm{dist}(A_i, B)}{n} \tag{22}$$

where n represents the number of points on the surface of the internal object (nodule marked by the expert or internal region), A represents the vector of points of the internal object, and B represents the vector of points on the external object (cylinders).

The second piece of information [Eq. (23)] verifies the mean distance cost necessary to reach a particular object starting from its center of mass (Md2).

$$\mathrm{Md2} = \frac{\sum_{i=1}^{n} \mathrm{dist}(P, A_i)}{n} \tag{23}$$

where n represents the number of points on the surface of the object, P represents the center of mass, and A represents the vector of points on the surface of the object.

To illustrate the behavior of Eqs. (22) and (23), we created Fig. 8, which presents a 2D example. In Fig. 8a, blue represents the cost for some points to reach the border of the circle, and in Fig. 8b, we illustrate the cost of reaching the border starting from the center of pass (p).

It is important to note that n does not affect the value of the equation, but the irregularities in the analyzed object, that is, the final values of the sum of Eqs. (22) and (23), will always be weighted with the total number of points on the surface. Thus, for nodules with well-defined spherical properties, this value tends to follow a pattern, regardless of nodule size.

2.5.3 *Skeleton-based measurements*

We believe that the extraction of measurements from an object, based on its mean forming axis, provides essential information because such measurements indicate irregularities that the object might have. For this, we propose measurements based on the information extracted from its representation of the vector of points.

For the representation of the vector of points, we extract the following measurements:

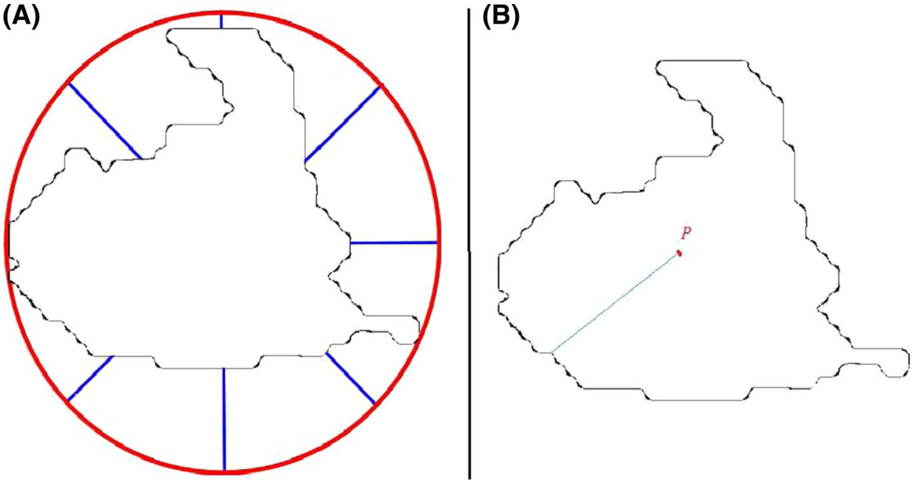

Fig. 8 2D example of the descriptors: **a** average expense required for an object to reach the edge of the outermost object (Md1) and **b** distance from point P to the border of the object (Md2)

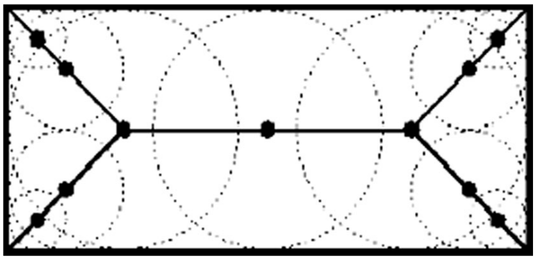

Fig. 9 Examples of circles generated by each skeleton point

1. Volume of the represents the vector of points.
2. Length of the *X*-axis.
3. Length of the *Y*-axis.
4. Length of the *Z*-axis.
5. In Eq. (24), the mean distance for the surface (MT).
6. The mean diameter (MD) of the spheres, obtained through Eq. (25).

$$\mathrm{MT} = \frac{\sum_{i=1}^{n} \mathrm{dist}(A, B)}{n} \tag{24}$$

where n represents the number of points of the representation of the vector of points, A represents the vector of points of the representation of the vector of points, and B represents the vector of points of the object extracted from the representation of the vector of points.

For each point on the skeleton, we compute the spheres, provided that they are inside the object that forms the skeleton. In Fig. 9, we show an example using a 2D object.

$$\mathrm{MD} = \frac{\sum_{i=1}^{n} \mathrm{diameter}(A)}{n} \tag{25}$$

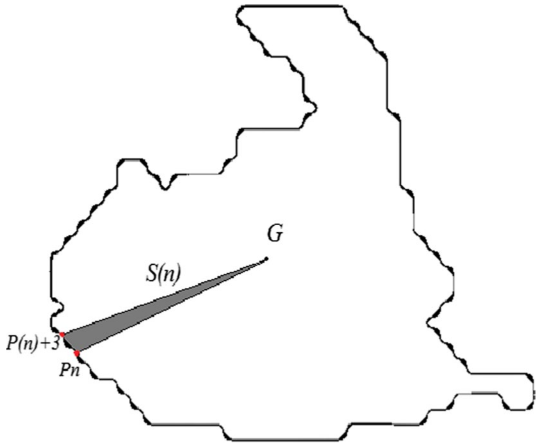

Fig. 10 Example of the triangle area formed from the center of mass (G) to two points on the border [Pn and P(n) + 3]

where n represents the number of points of the skeleton and A represents the vector of diameters computed from each point of the skeleton.

2.5.4 Triangulation-based measurements

These measurements are intended to check the irregularities of the object with respect to its center of mass and surface.

When the surface points change along the length of the shape, the area of the triangle formed by two points of the surface and the center of mass (Fig. 10) also changes. This gives us a function of the area that can be exploited as a representation of the shape. Where *S(n)* represents the area between successive boundary points and P_i, $P_i + 3$ and G is the center of mass. According to Eq. (26), the area of the triangle is measured for each object

$$S(n) = \frac{\sum_{i=1}^{n} \mathrm{area}(G, P_i, P_i + 3)}{n} \tag{26}$$

where n represents the number of points of the surface of the object.

We chose to use $P_i + 3$ because we intend to investigate the irregularities contained in the border of the nodule; this would not be possible if we opted to use $P_i + 1$. In other words, the triangle area followed by consecutive points would be almost equal for all cases, in view of the proximity of the border of points. When it uses interleaved points, this area features more appropriate values for the purpose of the technique because the irregularities start to be highlighted by calculating the triangle area.

A representation of TAR is computed based on the TARs formed by the points on the shape surface [1, 2]. The curvature of the contour point (P_n) with coordinates (x_n, y_n) is measured using TAR, as described below.

For each set of five consecutive points $P_n - t_s$ $(x_n - t_s, y_n - t_s)$, $P_n(x_n, y_n)$, and $P_n + t_s(x_n + t_s, y_n + t_s)$, where $n \in [1, n]$ and $ts = 2$. The area of the triangle formed by these points is given by:

$$\mathrm{TAR}(n, t_s) = \frac{1}{2}\begin{vmatrix} x_n - t_s & y_n - t_s & 1 \\ x_n & y_n & 1 \\ x_n + t_s & y_n + t_s & 1 \end{vmatrix} \tag{27}$$

Similar to $S(n)$, we intend to analyze the nodule's border for irregularities. Given the proximity of the points, we select five points where, from a given point P_n, any point can be calculated as $P_n - 2$ and $P_n + 2$. This way, more precise information about the irregularities contained in the border of the nodule can be obtained.

The contour is scanned in a clockwise direction. When the TAR value is positive, it indicates that the points are convex. When the TAR value is negative, the points are concave, and when TAR is zero, the points form a straight line. Figure 11 shows these three types of TARs.

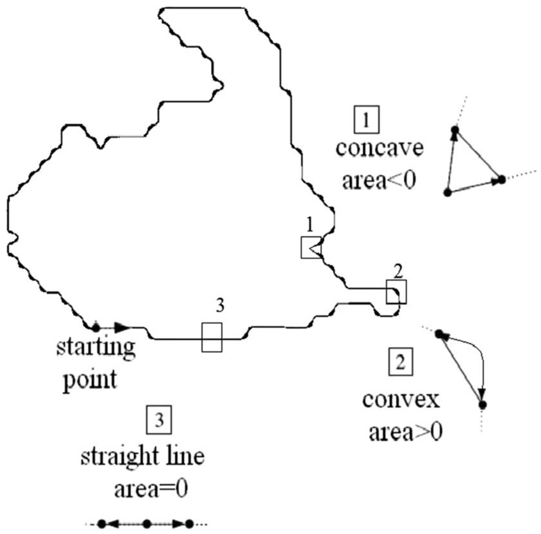

Fig. 11 Examples of TAR results

$$\mathrm{TARmean} = \frac{\sum_{i=1}^{n} \mathrm{TAR}_i}{n} \tag{28}$$

where n represents the number of points on the surface of the object.

2.5.5 *Measurements based on Feret diameters*

Feret diameters determine the biggest and smallest tangent of an object, i.e., the values of the tangents with respect to the X-, Y-, and Z-axes.

Measurements based on Feret diameters are intended to compute the elongation degree of an object. Thus, it is computed as a ratio between the longest Y-axis (d) and the shortest X-axis (w), defined by Eq. (32), and the proportions of each axis are analyzed for the internal object (nodule marked by the expert and internal region) and external object (cylinders), as shown in Eqs. (29)–(31).

$$PX = \frac{x_{\mathrm{internal}}}{x_{\mathrm{external}}} \tag{29}$$

where x_{internal} represents the size of the minimum X-axis internal object (nodule marked by the expert or internal region) and x_{external} indicates the size of the X-axis external object (cylinder).

$$PY = \frac{y_{\mathrm{internal}}}{y_{\mathrm{external}}} \tag{30}$$

where y_{internal} represents the size of the minimum Y-axis internal object (nodule marked by the expert or internal region) and y_{external} indicates the size of the Y-axis external object (cylinder).

$$PZ = \frac{z^2}{(PX + PY)} \tag{31}$$

$$FP = \frac{d}{w} \tag{32}$$

2.6 Selection of the best model

Finding methodologies in the literature that use techniques for selecting the most significant features in order to generate a training database is common. We opted to use GA as proposed by [8] to select the best individuals in order to generate the model used in the classification.

GA proposed by [8] for the selection of the best individuals can be summarized in the following steps:

1. From each nodule obtained during image acquisition (Sect. 2.1), the set of all feature vectors forms the analysis base (AB).

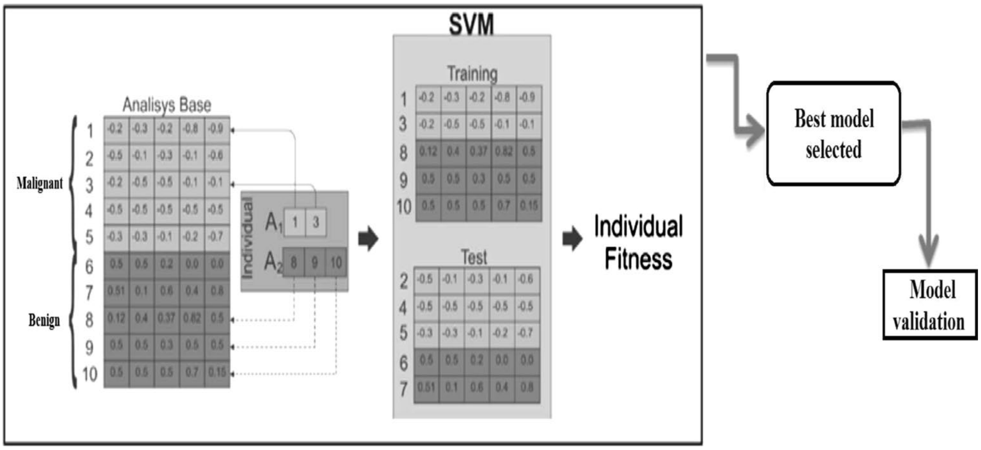

Fig. 12 Analysis for choosing the best training model [8]

2. Each element in array A1 contains the position of a feature vector extracted from a nodule in AB.
3. Each element in array A2 contains the position of a feature vector extracted from a non-nodule in AB.
4. The mutation and crossing genetic operators modify the values contained in A1 and A2. However, the values cannot be repeated. Elements of A1 cannot be present in A2 or vice versa. In particular, the mutation and crossing genetic operators do not combine data from the arrays.
5. For each individual of a generation, the feature vectors selected in A1 and A2 are trained by the SVM [10], the machine learning algorithm. The other feature vectors from AB, which were not selected, are also tested using SVM. The fitness of each individual is computed through the sum of specificity, sensitivity, and accuracy. This process is repeated until the fitness of the best individual is the same for 500 consecutive generations.
6. At the end of the evolution, the best training model is formed by all the feature vectors whose positions are contained in arrays A1 and A2 of the individual with the best fitness in the last generation.
7. Finally, the selected model is validated. This validation is conducted by means of a classification where all the remaining nodules in the analyzed base are used, thus observing the addition of three non-nodules to every nodule. In order to measure how good is the model, sensitivity, specificity, and accuracy are computed.

In summary, Fig. 12 presents a general scheme of how the choice is made by the GA model. First, there is the composition of the base to be analyzed and the creation of the initial individual. Subsequently, a sequence of changes is made, where genetic material is exchanged in the search for the individual α. This individual is chosen by best fitness. Finally, after a certain number of repetitions occur where the fitness does not improve, the best model is obtained. This model is then used for testing.

2.7 Pattern recognition

After finishing the feature extraction stage and selection of the fittest individuals, the 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 a radial base function (RBF) [34].

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 is 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 [10].

3 Results

In this section, we present the results obtained with the proposed methodology for the diagnosis of lung nodules described in Sect. 2. The analysis of the results follows this strategy: (a) acquisition of the images used to train and test the methodology; (b) description of how the process of feature extraction occurred; (c) tests with SVM for each technical group shown in Sect. 2.5, GA use as described in Sect. 2.6 to select the best training model, and computation of the accuracy, sensitivity, specificity, and receiver operating characteristic (ROC) value

Table 1 Results of the Minkowski functionals

Region under analysis	Accuracy	Sensitivity	Specificity	ROC
NME	83.87	79.74	85.5	0.804
IR	85.30	84.70	85.56	0.848

Nodule marked by the expert (NME) and internal region (IR)

for each test; and finally (d) a comparative analysis with other related works.

3.1 Result 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 the performance analysis of systems based on image processing, namely sensitivity, specificity, and accuracy [10]. In [39], another method for measuring the performance of computer-based detection techniques is used: ROC curves. A ROC curve indicates the true positive rate (sensitivity) as a function of the false positive rate (1-specificity).

3.2 Database separation

The LIDC–IDRI database contains 1018 CT images. However, two factors made 185 of them inappropriate for this methodology. The first factor is the presence of nodules with diameters equal to or below 3 mm because they do not include information that indicates their degree of malignancy. The second factor is the divergence between the information found in the marking file and that present in the DICOM header of the same examination, which invalidates the markings [8]. Therefore, the proposed methodology was applied to 833 examinations.

All the training bases were generated with GA. GA is responsible for selecting the best nodules, thus ensuring that only the most significant nodules are selected for generating a training model based on the mutation, selection, and crossing criteria of GA. Another important parameter that must be mentioned is the number of generations used until GA stops evolving, i.e., until the fitness is the same for 500 consecutive times, thus reaching the stop criterion for GA.

From the 833 examinations, we extracted 1405 nodules (1011 benign and 394 malignant), and they were divided into 80 % (1124) for training and validation, and 20 % (281) for tests; all were randomly selected. The 20 % were used to test the final model selected by GA. The 80 % from the training base were subjected to GA for selection of the best individuals (nodules). In order to select only the most significant individuals, we only passed 70 % of the training base, i.e., from the original 80 %, GA selected only the 70 % that best represented the benign and malignant classes. The remaining 30 % were used by GA to validate the best model. This procedure was applied to all the tests performed with our methodology.

To test our theory and justify the creation of the internal region, we opted to perform tests with the separate regions. In particular, we first made tests with the nodules that had the markings made by the expert and their cylinders. Then, we performed the tests using the internal region proposed in this work (Sect. 2.3) with the cylinders, thus allowing a comparison between the nodule marked by the expert and the internal region.

3.3 Classification

The tests were first performed with all the techniques shown separately in Sect. 2.5. Then, we executed new tests where all the features were pooled. It is important to stress that all these stages followed the training/test scheme are presented in Sect. 3.2.

3.3.1 Tests with the Minkowski functionals

With the tests using the Minkowski functionals [(V), surface area (S), mean curvature (B), and Euler's number (χ)], we extracted 5 measurements for the nodule with the markings from the expert and 5 for the internal region, namely 4 from the Minkowski functionals and 1 volume measurement of its cylinder.

The results for MFs listed in Table 1 prove that the internal region is capable of supplying more details, thus leading to a better result, compared with the nodule marked by the specialist.

3.3.2 Tests with distance measurements

In the tests that used distance measurements, we extracted four measurements from the nodule marked by the expert and its cylinder, and four from the internal region and its cylinder.For both types of regions analyzed, the distance-based descriptors have good results. However, we must emphasize that the results achieved with the internal region are relatively more expressive when compared with a nodule delimited by experts.

3.3.3 Tests with the skeleton-based measurements

For this scenario, the tests relied on the six measurements described in Sect. 2.5.3. For each region, six measurements were extracted for the nodule marked by the experts, and six for the internal region.

Table 2 Results of the distance measurements

Region under analysis	Accuracy	Sensitivity	Specificity	ROC
NME	87.45	86.90	87.69	0.839
IR	91.03	90.54	91.21	0.913

Nodule marked by the expert (NME) and internal region (IR)

Table 3 Results of the skeleton-based measurements

Region under analysis	Accuracy	Sensitivity	Specificity	ROC
NME	86.02	82.55	87.56	0.837
IR	85.30	80.59	86.79	0.829

Nodule marked by the expert (NME) and internal region (IR)

Table 4 Results of triangulation-based measurements

Region under analysis	Accuracy	Sensitivity	Specificity	ROC
NME	84.94	81.81	86.13	0.804
IR	88.53	85.91	89.42	0.874

Nodule marked by the expert (NME) and internal region (IR)

Table 3 indicates that the results found for the internal region are slightly smaller when compared with the region originally marked by the experts. This does not compromise the proposed methodology because these results are quite smaller than the rest.

We believe that the measurements based on the skeleton do not provide good results because this is a sensitive method; that is, some nodule properties might not have been achieved in the formation of the skeleton, or they might not have been as evident through these descriptors. Thus, the best results occur with the nodules marked by experts because, as previously stated, if the internal regions are compared, they have the ability to extract several details with the skeleton.

3.3.4 Tests with triangulation-based measurements

In the tests that used triangulation measurements, we extracted two measurements from the nodules marked by experts, and two from the internal region. In Table 4, we list the efficiency of the methodology that uses only two descriptors. In the best result, we again observe that the internal region leads to superior results when compared with the original region.

3.3.5 Tests with measurements based on Feret diameters

In the tests that used Feret diameters, we extracted four measurements for each region (Sect. 2.5.5) in the original nodule and internal region.

The results presented in Tables 1, 2, 3, 4, and 5 demonstrate the effectiveness of the proposed descriptors. Of the five tests performed, four (Tables 1, 2, 4, 5) demonstrate the efficiency of the analysis made by means of the internal region. The good results for these descriptors are related to the fact that the internal region supplied more precise details about the shape of the lung nodule when compared with the markings made by the experts.

The results presented in Table 3 indicate that the internal region has worse results. We believe that the skeleton-based descriptors are not capable of supplying good details because the medial axis of the internal region does not allow the extraction of good descriptors. However, this fact does not compromise the performance of the methodology because, in general, these are the worst results of the methodology.

To verify whether one technique is capable of identifying what the other cannot, we decided to perform a test with all the measurements together. For this, we used the five measurements extracted from MFs (Sect. 2.5.1), four extracted from the distance measurement (Sect. 2.5.2), six based on skeleton measurements (Sect. 2.5.3), two extracted using triangulation (Sect. 2.5.4), and four extracted based on the Feret diameters (Sect. 2.5.5), thus resulting in 21 measurements.

With the 21 measurements that describe the nodule marked by experts and 21 that describe the internal region, we performed the same test procedure of the previous sections. Table 6 lists the results.

To ascertain the potential of the descriptors, we made a feature selection using stepwise discriminant analysis. In the stepwise estimation, a variable is selected according to their significance, and after each step, the most significant variables are extracted, thus forming a data set for investigation. The process starts by choosing the best discriminatory variable. The initial variable is then paired with one of the other independent variables, one at a time, and the most appropriate variable for improving the discriminatory potential of the function in combination with the first is chosen. The other variables are chosen analogously. After each stage of incorporating a variable, there is a stage in which the previously selected variables can be discarded. The procedure is completed when no variable is added or dropped [17].

After the selection, we obtained only 12 features: four of distance measures, one of MF, two based on the skeleton, two of measures based on triangulation, and three of Feret diameters.

The results are very promising because only 12 measurements were used to characterize the nodules, and the results achieved are very close to the best case. Thus, we

Table 5 Results of measurements based on Feret diameters

Region under analysis	Accuracy	Sensitivity	Specificity	ROC
NME	86.73	89.70	85.78	0.884
IR	90.68	87.84	92.11	0.896

Nodule marked by the expert (NME) and internal region (IR)

Table 6 Results of all of the measurements together

Region under analysis	Accuracy	Sensitivity	Specificity	ROC
NME	89.24	87.5	89.94	0.879
IR	93.19	92.75	93.33	0.930

Nodule marked by the expert (NME) and internal region (IR)

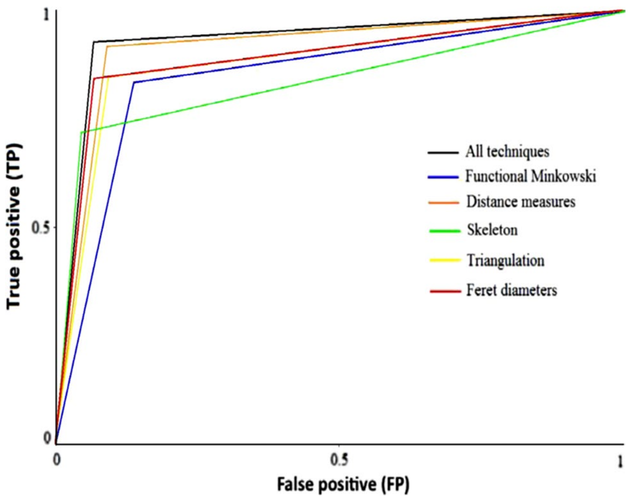

Fig. 13 ROC curves of the best results

Table 7 Results for the texture descriptors and shape

Region under analysis	Descriptors	Accuracy	Sensitivity	Specificity
NME	Texture	81.91	72.4	84.44
IR	Texture	84.26	75.54	86.18
NME	Shape	89.24	87.5	89.94
IR	Shape	93.19	92.75	93.33
NME	Texture and shape	91.34	90.55	92.56
IR	Texture and shape	93.92	93.05	94.26

Nodule marked by the expert (NME) and internal region (IR)

achieved an accuracy of 91.37 %, sensitivity of 91.01 %, specificity of 91.61 %, and ROC curve of 0.915. As can be seen, the result is slightly below the best case; however, this is promising because we were able to achieve an accuracy rate above 91 % with several significantly reduced descriptors.

Upon combining all the measurements, the results show the efficiency of the methodology for diagnosing lung cancer. Table 6 presents significant and promising results. Finally, Fig. 13 presents the results of the six ROC curves for the best results.

Of all the descriptors presented, only MFs were adapted for use in our methodology. The rest of the techniques were originally developed and proposed by our work.

3.3.6 Tests with measurements of texture and shape

Tests were performed with three classic texture analysis techniques: histogram (statistics of the first order) [38], gray-level co-occurrence matrix (GLCM) (statistics of the second order) [14], and gray-level run lengths (GLRL) (statistical high order) [14]. The tests were performed using the same logic employed in the previous experiments, i.e., we analyzed only the region selected by skilled experts, and another test used only the internal target region. Table 7 lists the results obtained using all the texture descriptors (histogram, GLCM, and GLRL). Subsequently, we combined all the texture descriptors and all the proposed shape descriptors into our method. Finally, we also present only the results obtained for all the measures in the order proposed by our method.

As indicated in Table 7, the results are better when the measurements of texture and shape are combined. Such improved results were expected because these measurements complement one another. It is noteworthy that our proposed definition for an internal region promotes improvement in all test cases. Another important point is the relationship where the measurements proposed allowed a significant improvement in the results; in addition, our results from using only the best shape descriptors can be as promising as the results obtained from combining the measurements for texture and shape.

3.4 Comparison with other related works

Comparison with other works in this area is challenging because examination details were not included in any of the references. The only piece of information provided is the database used. Thus, we were unable to perform a rigorous evaluation of our method with respect to other works.

Our objective with Table 8 is to provide an overview (examination database, complexity of the methodology, etc.) of the results found in related works and our work. We intend to show that our methodology is promising because, compared with other works, we achieved results above 93 % for various types of situations: (a) diagnosis using

Table 8 Comparison with other publications with respect to the classification of lung nodules into benign and malignant

Work	Techniques	Database	Ac (%)	Se (%)	Spe (%)	Sample
[28]	Texture features using diversity indexes of Shannon and Simpson, linear discriminant analysis (LDA), and SVM	LIDC–IDRI	83.75	82.95	84.58	73
[29]	Texture features, correlation-based feature selection and *k*-nearest neighbor, and SVM	NBIA-ELCAP	82.66	96.15	52.17	–
[20]	Texture features, correlation-based feature selection, and *k*-nearest neighbor	LIDC–IDRI	90.91	85.71	94.74	33
[6]	Texture features using matrix co-occurrence of gray levels, principal component analysis, and artificial neural network	Private	90.63	92.30	89.47	128
[30]	Texture features using matrix co-occurrence of gray levels, and SVM	Private	–	91.38	89.56	11
[40]	Shape features using gradient field and radius features, stepwise, simplex optimization, LDA, and SVM	Private	85	–	–	256
[42]	Shape features, fuzzy thresholding method, volumetric shape index map, Hessian matrix	Private	90.2	–	–	108
[21]	Shape features using bi-orthogonal wavelet, and fuzzy classifier.	Private	90	86	84	–
[11]	Shape features using spherical harmonics, mapping this model to the unit sphere, and *k*-nearest classification	Private	93.6	–	–	327
[12]	Shape and texture features and radial basis function neural network	Private	94.44	–	88.14	–
Our work	Shape descriptors, genetic algorithm, and SVM	LIDC–IDRI	93.19	92.75	93.33	833

Ac accuracy, *Se* sensitivity, and *Spe* specificity

only shape; (b) large and complex samples; and (c) several sample configurations for training and testing.

By comparing the best results achieved in our work with those presented in Table 8, it is possible to see that our results are promising. Table 8 presents the results of the works that used shape and texture features to characterize their nodules. The works by [11, 12] present accuracy that is slightly superior to ours; however, this is not an indication that they are more efficient because sensitivity and specificity, both of which suggest the efficiency of a given methodology, are quite inferior or uninformed. In addition, the number of cases analyzed by [11, 12] is less than that in our work.

It is important to stress that finding works in the literature that use only shape descriptors to diagnose lung nodules is challenging, which proves the difficulty of this analysis and the efficiency of our methodology.

4 Discussion

The proposed methodology was evaluated by applying a set of 1405 nodules (benign and malignant) from the LIDC–IDRI database, and these were divided into an 80/20 training/test set with the use of GA. The experiment results allow the formulation of the following conclusions:

1. The internal region led to good results because it was possible to extract shape information for the nodules, which was not possible before.
2. The use of GA showed the efficiency in the selection of the best individuals; this was used to create the training model.
3. Several methods for analyzing lung nodules were presented, making them applicable to other areas.
4. The combination of all the techniques led to better discrimination in the classification of the nodules, thus achieving an accuracy of 93.19, sensitivity of 92.75, and specificity of 93.33.
5. The difficulty of shape analysis provides reliability and robustness to our methodology because few works in the literature use only shape to characterize lung nodules to ease the diagnosis.
6. We believe that the results based on the skeletal analysis gave the worst results because of the limitations of the method itself, i.e., skeleton techniques are sensitive to noise; if there is some noise in the image, this could provide incorrect or imprecise skeletons. It was also noted that in some cases, the resulting skeletons did not match correctly with the original image. We believe that this was caused by noise or the very geometry of the nodule. In addition, in some methods, very smooth regions could also compromise the result of the skeleton, thus leading to skeletons that were not faithful to the actual shape of the image.
7. As shown by [35], up to 1/3 of all malignant nodules can have smooth surfaces, and thus, the presence of a smooth contour is not a very reliable signal. We believe that the good performance of our metrics with regard to this fact is the delimitation of the intratumoral region of the lesion because, based on the numbers of sensitivity that measure how good the method is for

guessing the malignant class, our 92.75 % success rate represents the values that exceed the 2/3 that theoretically would not show the smoothness of the contour of the lesion, as shown in [35]. In short, our method was robust even in these cases.

8. Finally, it is important to highlight that the LIDC–IDRI database is extremely complex and diversified: It contains countless different cases of lung nodules. This database has examinations extracted by various tomographs, thus making it more difficult to detect, classify, or even diagnose through CAD/CADx systems.

All of these items add value to our methodology. The shape properties of the lung nodule analyzed by the proposed techniques along with GA yielded good results. In addition, the complexity of the LIDC–IDRI database allowed us to form a more precise conclusion about the results.

5 Conclusions

High rates of deaths and records of lung cancer occurrences worldwide demonstrate the importance of developing research in order to produce resources for early diagnosis of the disease, thereby providing better treatments. Lung cancer stands out for presenting the highest mortality rate and one of the lowest survival rates after diagnosis (5 years for 14–20 % of the patients). Precocious diagnosis represents a considerable increase on the survival probability of the patients. Based on this, an automatic methodology that assists the expert in giving a second opinion on the diagnosis of lung nodules was developed in this work.

This paper presented an analysis method based on MF, distance measures, measures based on the skeleton, triangulation-based measures, measures based on the Feret diameters, GA, and SVM for classification of pulmonary nodules into malignant and benign. The methodology proved to be a useful tool for specialist physicians.

The results highlighted the promising performance of the techniques for analyzing lung nodule shapes. Another important factor was the combination of the proposed techniques and use of GA, which led to better results in differentiating between benign and malignant nodules.

At last, the methodology presented in this work could integrate a CADx tool to be applied in the detection and diagnosis of lung cancer in order to classify the nodules into malignant and benign, thus making examination analysis by specialists more agile and less exhaustive.

Acknowledgments The authors acknowledge Coordination for the Improvement of Higher Education Personnel (CAPES), 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

1. Alajlan N, Kamel MS, Freeman G (2006) Multi-object image retrieval based on shape and topology. Image Commun 1(10):904–918. doi:10.1016/j.image.2006.09.002
2. Alajlan N, Rube IE, Kamel MS, Freeman G (2007) Shape retrieval using triangle-area representation and dynamic space warping. Pattern Recognit 1(7):1911–1920. doi:10.1016/j.patcog.2006.12.005
3. Armato SG, McLennan G, Bidaut L, McNitt-Gray MF, Meyer CR, Reeves AP, Zhao B, Aberle DR, Henschke CI, Hoffman EA, Kazerooni EA, 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 (2011) The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. Med Phys 1(2):915–31. http://www.biomedsearch.com/nih/Lung-Image-Database-Consortium-LIDC/21452728.html
4. Chang CC, Lin CJ (2013) LIBSVM—a library for support vector machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm/
5. Chen W, Li Z, Bai L, Lin Y (2011) Nf-kappab in lung cancer, a carcinogenesis mediator and a prevention and therapy target. Front Biosci (Landmark edition) 1:1172–1185. doi:10.2741/3782
6. Dandil E, Cakiroglu M, Eksi Z, Ozkan M, Kurt O, Canan A (2014) Artificial neural network-based classification system for lung nodules on computed tomography scans. In: 2014 6th International conference of soft computing and pattern recognition (SoCPaR), p 382–386. doi:10.1109/SOCPAR.2014.7008037
7. David L, Kien K, Marie FD (2011) Computation of minkowski measures on 2d and 3d binary images. Image Anal Stereol 1(2):83–92. http://www.ias-iss.org/ojs/IAS/article/view/811
8. de Carvalho Filho AO, de Sampaio WB, Silva AC, de Paiva AC, Nunes RA, Gattass M (2014) Automatic detection of solitary lung nodules using quality threshold clustering, genetic algorithm and diversity index. Artif Intell Med 1(3):165–177. doi:10.1016/j.artmed.2013.11.002
9. de Valk J, Eijkman E (1984) Analysis of eye fixations during the diagnostic interpretation of chest radiographs. Med Biol Eng Comput 1(4):353–360. doi:10.1007/BF02442106
10. Duda RO, Hart PE (1973) Pattern Classif Scene Anal. Wiley-Interscience Publication, New York
11. El-Baz A, Nitzken M, Khalifa F, Elnakib A, Gimelfarb G, Falk R, El-Ghar M (2011) 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, vol 6801, p 772–783. Springer, Berlin. doi:10.1007/978-3-642-22092-0-63
12. Elizabeth D, Nehemiah H, Retmin Raj C, Kannan A (2012) Computer-aided diagnosis of lung cancer based on analysis of the significant slice of chest computed tomography image. Image Process IET 1(6):697–705. doi:10.1049/iet-ipr.2010.0521
13. Fujimoto J, Wistuba II (2014) Current concepts on the molecular pathology of non-small cell lung carcinoma. Semin Diagn Pathol

1(4):306–313. doi:10.1053/j.semdp.2014.06.008 (Lung Carcinoma: Beyond The WHO Classification)

14. Galloway MM (1975) Texture analysis using gray level run lengths. Comput Graphics Image Process 1(2):172–179. doi:10.1016/S0146-664X(75)80008-6
15. Gould M, Maclean C, Kuschner W, Rydzak C, Owens D (2001) Accuracy of positron emission tomography for diagnosis of pulmonary nodules and mass lesions: a meta-analysis. JAMA 285(7):914–924. doi:10.1001/jama.285.7.914
16. Hadwiger H (1957) Vorlesungen über Inhalt, Oberfläche und Isoperimetrie. Grundlehren der mathematischen Wissenschaften. Springer. https://books.google.com.br/books?id=-BWoAAAAIAAJ
17. Hair JF, Black WC, Babin BJ, Anderson RE, Tatham RL (2006) Multivariate data analysis, vol 6. Pearson Prentice Hall, Upper Saddle River
18. Hansell DM, Bankier AA, MacMahon H, McLoud TC, Muller NL, Remy J (2008) Fleischner society: glossary of terms for thoracic imaging. Radiology 246(3):697–722. doi:10.1148/radiol.2462070712
19. Jabon SA, Raicu DS, Furst JD (2009) Content-based versus semantic-based retrieval: an lidc case study. Proc SPIE 7263:72,631L–72,631L–8. doi:10.1117/12.812877
20. Krewer H, Geiger B, Hall L, Goldgof D, Gu Y, Tockman M, Gillies, R (2013) Effect of texture features in computer aided diagnosis of pulmonary nodules in low-dose computed tomography. In: 2013 IEEE international conference on systems, man, and cybernetics (SMC), pp 3887–3891. doi:10.1109/SMC.2013.663
21. Kumar S, Ramesh J, Vanathi P, Gunavathi K (2011) Robust and automated lung nodule diagnosis from ct images based on fuzzy systems. In: 2011 International conference on process automation, control and computing (PACC), pp 1–6. doi:10.1109/PACC.2011.5979050
22. Lederlin M, Revel MP, Khalil A, Ferretti G, Milleron B, Laurent F (2013) Management strategy of pulmonary nodule in 2013. Diagn Int Imaging 1(11):1081–1094. doi:10.1016/j.diii.2013.05.007
23. Leef 3rd J, Klein J (2002) The solitary pulmonary nodule. Radiol Clin N Am 1(1):123–143, ix. doi:10.1056/NEJMcp012290
24. Mecke K, Stoyan D (2002) Morphology of condensed matter: physics and geometry of spatially complex systems. Lecture notes in physics. Springer. https://books.google.com.br/books?id=0BeJY4ZwndsC
25. Mecke K, Buchert T, Wagner H (1993) Robust morphological measures for large-scale structure in the Universe. Max-Planck-Institut für Astrophysik Garching, München: MPA. Max-Planck-Inst. für Astrophysik. https://books.google.com.br/books?id=jLzgGwAACAAJ
26. Michal C, Elke T, Abhir B, David P (2008) Integral geometry descriptors for characterizing emphysema and lung fibrosis in HRCT images. First Int Workshop Pulm Image Anal 1:155–164
27. Miyake N, Kim H, Itai Y, Tan JK, Ishikawa S, Katsuragawa S (2009) Automatic detection of lung nodules in temporal subtraction image by use of shape and density features. In: 2009 Fourth international conference on innovative computing, information and control (ICICIC), pP 1288–1292. doi:10.1109/ICICIC.2009.118
28. Nascimento LB, de Paiva AC, Silva AC (2012) 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-Verlag, Berlin, pp 454–466
29. 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
30. Parveen SS, Kavitha C (2014) Article: classification of lung cancer nodules using svm kernels. Int J Comput Appl 1(25):25–28 (full text available)
31. Patil SS, Godoy MC, Sorensen JI, Marom EM (2014) Lung cancer imaging. Semin Diagn Pathol 1(4):293–305. doi:10.1053/j.semdp.2014.06.007 (Lung Carcinoma: Beyond The WHO Classification)
32. Roth K, Boike J, Vogel HJ (2005) Quantifying permafrost patterns using minkowski densities. Permafr Periglac Process 1(3):277–290. doi:10.1002/ppp.531
33. Santaló L (2004) Integral geometry and geometric probability. Cambridge Mathematical Library. Beijing World Publishing Corporation (BJWPC). https://books.google.com.br/books?id=xq1iiumncV4C
34. Schölkopf B, Smola A (2002) Learning with kernels: support vector machines, optimization, and beyond. MIT Press, Cambridge (Regularization)
35. Seemann MD, Seemann O, Luboldt W, Bonél H, Sittek H, Dienemann H, Staebler A (2000) Differentiation of malignant from benign solitary pulmonary lesions using chest radiography, spiral CT and HRCT. Lung Cancer 1(2):105–124. doi:10.1016/S0169-5002(00)00104-5
36. Sone S, Takashima S, Li F, Yang Z, Honda T, Maruyama Y, Hasegawa M, Yamanda T, Kubo K, Hanamura K, Asakura K (1998) Mass screening for lung cancer with mobile spiral computed tomography scanner. Lancet 351(9111):1242–1245. doi:10.1016/S0140-6736(97)08229-9
37. Stewart (2014) World Cancer Report 2014. IARC Nonserial Publication, New York
38. Umbaugh SE, Snyder J, Fedorovskaya E (2011) Digital image processing and analysis: human and computer vision applications with cviptools, second edition. J Electron Imaging 1(3):039, 901–039, 901–903. doi:10.1117/1.3628179
39. van Erkel A, Pattynama P (1998) Receiver operating characteristic (ROC) analysis: basic principles and applications in radiology. Eur J Radiol 1(2):88–94
40. Way TW, Sahiner B, Chan HP, Hadjiiski L, Cascade PN, Chughtai A, Bogot N, Kazerooni E (2009) Computer-aided diagnosis of pulmonary nodules on ct scans: improvement of classification performance with nodule surface features. Med Phys 1(7):3086–3098. doi:10.1118/1.3140589
41. Yang X, Shen X, Long J, Chen H (2012) An improved median-based otsu image thresholding algorithm. AASRI Proc 1:468–473. doi:10.1016/j.aasri.2012.11.074 (conference on modelling, identification and control)
42. Ye X, Lin X, Dehmeshki J, Slabaugh G, Beddoe G (2009) Shape-based computer-aided detection of lung nodules in thoracic CT images. IEEE Trans Biomed Eng 1(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).

Aristófanes Corrêa Silva received a Ph.D. degree in Informatics from Pontiphical Catholic University of Rio de Janeiro—Brazil, in 2 004. Currently he is a Professor at the Federal University of Maranhão (UFMA), Brazil.

Anselmo Cardoso de Paiva received B.Sc. in civil engineering from Maranhão State University—Brazil, in 1990, an MSc in civil engineering—Structures and a Ph.D. in Informatics from Pontiphical 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 Ph.D. 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.