Pattern Recognition 119 (2021) 108083

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

COVID-index: A texture-based approach to classifying lung lesions based on CT images

Vitória de Carvalho Brito[a,b], Patrick Ryan Sales dos Santos[a,b], Nonato Rodrigues de Sales Carvalho[a,b], Antonio Oseas de Carvalho Filho[a,b,*]

[a] Department of Information Systems, Federal University of Piauí R. Cícero Duarte, 905, Junco, Picos 64607-670, PI, Brazil
[b] Department of Electrical Engineering, Federal University of Piauí – PI, Teresina, Brazil

ARTICLE INFO

Article history:
Received 14 October 2020
Revised 22 May 2021
Accepted 27 May 2021
Available online 6 June 2021

Keywords:
COVID-19
Computed tomography
3D texture analysis
Phylogenetic diversity

ABSTRACT

COVID-19 is an infectious disease caused by a newly discovered type of coronavirus called SARS-CoV-2. Since the discovery of this disease in late 2019, COVID-19 has become a worldwide concern, mainly due to its high degree of contagion. As of April 2021, the number of confirmed cases of COVID-19 reported to the World Health Organization has already exceeded 135 million worldwide, while the number of deaths exceeds 2.9 million. Due to the impacts of the disease, efforts in the literature have intensified in terms of studying approaches aiming to detect COVID-19, with a focus on supporting and facilitating the process of disease diagnosis. This work proposes the application of texture descriptors based on phylogenetic relationships between species to characterize segmented CT volumes, and the subsequent classification of regions into COVID-19, solid lesion or healthy tissue. To evaluate our method, we use images from three different datasets. The results are promising, with an accuracy of 99.93%, a recall of 99.93%, a precision of 99.93%, an F1-score of 99.93%, and an AUC of 0.997. We present a robust, simple, and efficient method that can be easily applied to 2D and/or 3D images without limitations on their dimensionality.

1. Introduction

Since the discovery of a new coronavirus (COVID-19) in China in late 2019, the disease has become a global concern, mainly due to its rapid spread. As of April 2021, the number of confirmed cases notified to the World Health Organization (WHO) has already exceeded 135 million, while the number of deaths bas exceeded 2.9 million [1]. COVID-19 is an infectious disease caused by a recently discovered type of coronavirus called SARS-CoV-2. Although most people infected with COVID-19 recover without special treatment, older people and people with preexisting illnesses such as diabetes, cardiovascular disease, chronic respiratory disease, and cancer are more likely to be severely affected [1]. The early diagnosis of COVID-19 is therefore essential for the treatment of the disease. Real-time polymerase chain reaction (RT-PCR) or chest computed tomography (CT) examination are possible alternatives for early diagnosis.

Several computer-aided detection (CAD) systems for the early diagnosis of COVID-19 have been developed in this context. A CAD system typically consists of three steps: (i) image acquisition; (ii) segmentation of candidate regions; and (iii) characterization and classification of these regions. In a CAD system, the segmentation stage is typically automatic, and needs to be able to handle numerous regions with similar characteristics (shape, density, or texture). It is therefore essential to apply a stage that efficiently classifies all of these regions. Thus, the proposed method acts in the characterization and classification stages. Overall, CT images of cases of COVID-19 share certain specific features, such as the presence of ground-glass opacities (GGOs) in the early stages and lung consolidation in the advanced stages [2]. Pleural effusion may also occur in cases of COVID-19, but is less common than the other lesions. It is therefore important to point out some difficulties with this approach, as follows:

- Although the features of COVID-19 are found in most cases, CT images of some viral pneumonias also show these features, which can ultimately make diagnosis more difficult [2];
- In some cases of COVID-19, biopsies are needed [3];
- Correct classification is required between healthy and diseased regions, especially those with COVID-19 and other more serious diseases such as lung nodules; and

* Corresponding author at: Department of Information Systems, Federal University of Piauí R. Cícero Duarte, 905, Junco, 64607-670 Picos, PI, Brazil.
E-mail addresses: vitoriacarvalho@ufpi.edu.br (V. de Carvalho Brito), sales@ufpi.edu.br (P.R.S. dos Santos), nrdesales@ufpi.edu.br (N.R. de Sales Carvalho), antoniooseas@ufpi.edu.br (A.O. de Carvalho Filho).

https://doi.org/10.1016/j.patcog.2021.108083
0031-3203/

- According to [4,5], COVID-19 regions are generally more rounded in shape.

In view of the above, we can see that there is a need to provide specialists with a method that can enable an individual analysis of the types of lesions found in CT scans. Through correct classification, our method can provide the individual details of the lesions, which can help the specialist in making decisions regarding the need for biopsies. Furthermore, based on the work in [4,5], we believe that the techniques used in our method of texture characterization can provide more meaningful information for lesion classification, since the shape of the lesion is not considered in the analysis.

This work makes original contributions in the areas of both medicine and computer science, as follows:

- In the area of computer science:
  - In the context of COVID-19, we propose phylogenetic and taxonomic diversity indexes for the characterization of image textures;
  - We improve the efficiency of the index calculation by optimizing the phylogenetic tree assembly.
- In the medical field:
  - We present a method that can be applied in patient triage and can therefore assist in the efficient management of healthcare systems;
  - Our system can diagnose patients quickly, especially when the medical system is overloaded;
  - Our approach can reduce the burden on radiologists and assist underdeveloped areas in making an accurate and early diagnosis.

The rest of the article is organized as follows: Section 2 reviews related works in the literature; Section 3 describes the proposed method; Section 4 presents the results obtained from an implementation of our approach; Section 5 discusses some relevant aspects of this work; and, finally, Section 6 presents the conclusion.

2. Related works

Today, pattern recognition, and particularly intelligent analysis, is one of the most promising areas of computer science. Several studies are notable in this field, such as the method developed in [6], which used a 3D model to segment brain regions. In [7], intelligent solutions for expression recognition and landmark localization problems were presented, while the authors of [8] proposed a method based on adversary learning to improve the efficiency of deep learning approaches in object detection. In the area of optimization, we highlight the work in [9], which introduced a consensus-based technique for a new form of data clustering and representation. Finally, we note the study in [10], which presented a method for recognizing patterns in low-resolution images using super-resolution networks.

Based on the aforementioned studies and global trends in this area, pattern recognition techniques have been used in numerous ways to help combat COVID19 and to mitigate the damage caused by the pandemic. In this section, we note some relevant works on this topic and several works that have applied phylogenetic diversity to image texture analysis to find solutions to other problems.

2.1. Phylogenetic diversity indexes for feature extraction

There is a great variety of diversity indexes, each of which has particular properties according to their categorization, for example (i) those that exploit the richness of species and the abundance of individuals; (ii) those that explore the relationships between species; and (iii) those that explore the topology of the representations of common ancestors.

There are several notable studies in this area. For example, the work in [11] used the phylogenetic distance and taxonomic diversity with a support vector machine (SVM) to classify pulmonary nodules. Other approaches have also used these indexes in automatic methods for the detection of glaucoma [12] and for the classification of breast lesions [13].

As can be seen from the studies described above, the literature contains recent works that have used phylogenetic diversity indexes as texture descriptors and have applied this approach to diverse problems with promising results. Since these were carried out in contexts that were different from ours, the studies listed above will not be used for a comparison of our results, but solely to highlight the contributions of the descriptors in the literature.

2.2. COVID-19 image classification

In [14] a methodology based on X-ray images and deep learning techniques was presented for the classification of images into COVID-19 and non-COVID-19. The results were promising, with an accuracy of above 90%. The authors of [15] also analyzed X-ray images, and combined the texture and morphological features to classify them into COVID-19, bacterial pneumonia, non-COVID-19 viral pneumonia, and normal. The results showed an AUC of 0.87 for this multi-class classification scheme. Feature combination was also used in [16], where radiomic features and clinical details were used to describe CT images and a Random Forest algorithm was applied to classify the features into non-severe and severe COVID-19.

Due to a lack of availability of public CT data on COVID-19, the authors of [17] built a dataset called COVID-CT, composed of 349 CT images that showed COVID-19 and 463 that did not. In [18], two subsets were extracted from a set of 150 CT images containing COVID-19 and non-COVID-19 patches, and classification was carried out using the deep feature fusion and ranking technique. The studies in [2] and [19] also applied CT images and deep learning approaches to classify images into cases that were positive and negative for COVID-19. In the approach proposed in [20], a pre-trained model of the DenseNet201 architecture was used to classify CT images into COVID-19 and non-COVID-19, by applying the weights in ImageNet. In [21], an automatic framework was proposed that used CNNs to extract visual features for COVID-19 detection. It was evaluated on a set of 4,356 chest CT exams.

Unlike the previous approaches, the authors of [22–24] proposed methodologies for the detection of COVID-19 using 3D volumes of CT scans. In [22], each volume was segmented using a pre-trained UNet and later classified with a weakly-supervised 3D deep neural network called DeCoVNet. The authors of [23] also evaluated their proposed method using 3D CT regions, this time obtained from scans of 81 patients. Their scheme was a radiomic model that combined texture features with patients' clinical data to classify COVID-19 into common or severe types. Finally, the approach in [24] classified CT images into healthy, idiopathic pulmonary fibrosis (IPF) and COVID-19 using a 3D approach called three-dimensional multiscale fuzzy entropy.

It can be observed from the above studies that solving the problem of COVID-19 diagnosis is not a simple task. Despite the application of various CNN-based methods to image classification, in which a CNN is responsible for extracting and selecting representative features in its convolutional layers, these feature maps are not always efficient enough to allow for classification.

Although recent work on a diverse range of imaging applications has used CNNs, with results that have surpassed those of other methods, in the case where there are representative features of a specific problem, the use of these features may be more efficient than CNN methods. Another problem encountered when using CNNs, which was also noted in [17,18], is the large number of

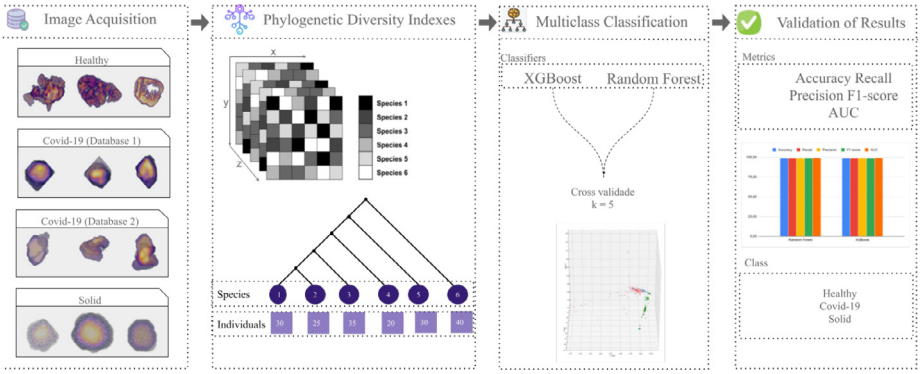

Fig. 1. Stages in the proposed methodology, consisting of image acquisition, extraction of characteristics through phylogenetic diversity indexes, data classification for the target classes, and validation of the results.

parameters required to create models for both the architecture and the parameters. The authors of these studies therefore proposed the use of transfer learning instead.

In addition, the training of a CNN requires considerable time in order to create a capable model, and several tests of architectures and parameters are required. Powerful machines are also needed to run these networks. Finally, the use of a CNN requires a large number of images, and data augmentation is often required to handle this issue. However, this is not a trivial task, as it requires countless tests and training of the whole network until satisfactory results are obtained.

We therefore propose to use phylogenetic diversity indexes for the feature extraction task of COVID-19, solid lesions and healthy tissue, in conjunction with the random forest and extreme gradient boost classifiers.

3. Materials and methods

This section describes our methodology for classifying CT volumes as COVID-19, solid lesion, or healthy tissue. The images used in this study were acquired from the Lung Image Database Consortium Image Collection (LIDC-IDRI) [25] and from the MedSeg [26] repository, the latter of which contained two datasets with COVID-19 images. At the feature extraction stage, we applied phylogenetic diversity indexes. The extraction and classification algorithms developed in this study are available from our GitHub repository. Fig. 1 illustrates the workflow of our methodology.

3.1. Image acquisition

We used three sets of volumes of interest (VOIs) extracted from the LIDC-IDRI and the MedSeg repositories to evaluate our method. These were as follows: (i) a set of images extracted from the LIDC-IDRI that contained VOIs showing solid lesions, for which we used the markings made by specialists for the base documentation; (ii) a set of healthy tissue VOIs that were extracted from LIDC-IDRI by applying the algorithm proposed in [27], to guarantee that the VOIs of healthy tissue did not intersect with those of solid-type lesions (this method was chosen as it could provide samples found in real scenarios); and (iii) a set of images acquired from the MedSeg repository [26] which contained some external datasets of various types of CT exams, including those diagnosed with COVID-19. Hence, we used two different sets of images containing lesions caused by COVID-19, i.e., regions with GGO lesions, consolidation, and pleural effusion. We used the specialists' markings that were available for the respective datasets to extract these lesions. Since MedSeg does not provide terminology for the COVID-19 datasets used here, we refer to these as COVID-19 (Dataset 1) and COVID-19 (Dataset 2).

Table 1
Distribution of images between datasets.

Dataset	Diagnosis	Number of images
LIDC	Solid lesions	1679
	Healthy tissue	17742
COVID-19 (1)	GGO, consolidation and pleural effusion	215
COVID-19 (2)	GGO, consolidation and pleural effusion	274

Table 1 shows the number of images in each dataset.

3.2. Feature extraction

In this step, we present the rationale for the proposed indexes for texture characterization. Each index corresponds to a certain characteristic, meaning that a total of eight characteristics are extracted from each analyzed image.

3.2.1. Phylogenetic diversity indexes

Phylogenetics is a branch of biology in which the evolutionary relationships between species are studied and the similarities between them described. In phylogenetic trees, leaves represent species and nodes represent common ancestors. The phylogenetic tree used in this work is called a cladogram. Fig. 2 illustrates an example of a cladogram that represents the genetic relationship between the monkey and human species; it can be observed that from a genetic perspective, humans and chimpanzees are closer than the other pairs of species in the tree.

A combination of phylogenetic trees and phylogenetic diversity indexes is used to analyze the evolutionary relationships between species and to measure the variation between species in a community. In order to be able to apply these concepts to the characterization of CT images, we need to define a correspondence between

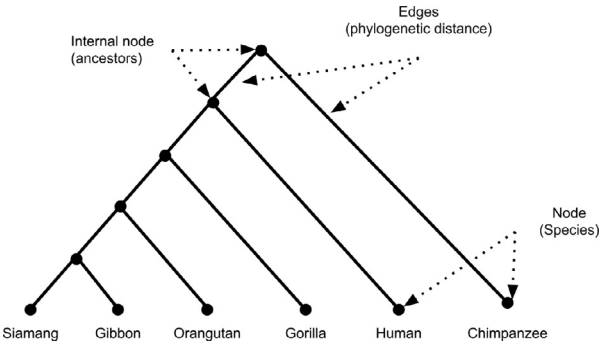

Fig. 2. Example of a cladogram for a set of primates.

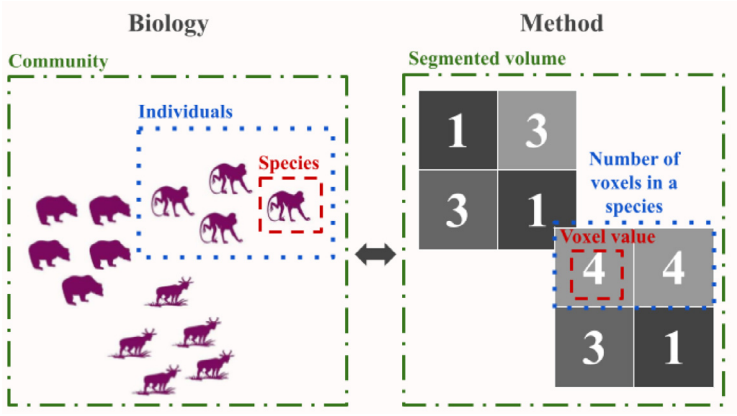

Fig. 3. Correspondence between biological concepts and the elements in the proposed method.

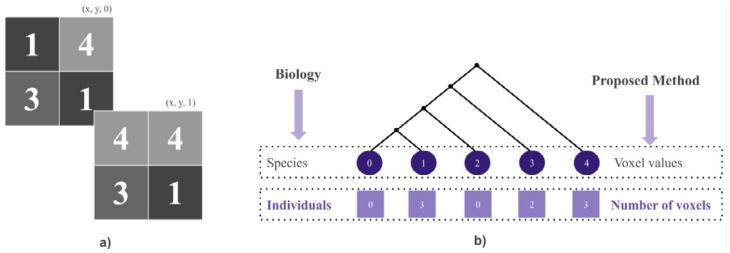

Fig. 4. (a) Example of an analyzed volume; (b) representation of the cladogram extracted from the example image.

the definitions used in biology and those used in this work. This is illustrated in Fig. 3.

Using the correspondence shown in Fig. 3 as a basis, we generate a cladogram for each study image, and an example of this is shown in Fig. 4(b). The proposed indexes are computed by applying the relevant equations from the cladogram. Each index represents a specific feature of the image, so we use eight indexes (i.e., eight features) to represent each image. The details of the proposed indexes are given below.

The phylogenetic diversity (PD) index [28] is a measure that gives the sum of the distances of phylogenetic branches in the tree. When the branch length is longer, the species become more distinct. Eq. (1) shows the formula for the PD, where B represents the number of branches in the tree, L_i is the extension of branch i (the number of edges in that branch), and A_i refers to the average abundance of the species that share branch i:

$$PD = B \times \frac{\sum_i^B L_i A_i}{\sum_i^B A_i}. \tag{1}$$

The sum of phylogenetic distances (SPD) is a phylogenetic index that gives the sum of the distances between the pairs of species present in the tree [29]. Eq. (2) is used to calculate this index, where S represents the number of species in the community, d_{ij} is the distance between species i and j, a_i and a_j correspond to the abundance of species i and j, respectively. The term $\sum_{i=1}^{S-1}\sum_{j=i+1}^{S}$ in the numerator represents the double sum of the products of the distances between all species in the tree based on their abundance; in the denominator, it corresponds to the double sum of the products of the abundances of the species.

$$SPD = \left(\frac{S(S-1)}{2}\right) * \frac{\sum_{i=1}^{S-1}\sum_{j=i+1}^{S} d_{ij} a_i a_j}{\sum_{i=1}^{S-1}\sum_{j=i+1}^{S} a_i a_j}. \tag{2}$$

The mean nearest neighbor distance (MNND) is a weighted average of the phylogenetic distance of the nearest neighbor of each species [30]. The weights represent the abundance of each species. Eq. (3) shows the formula used to calculate this index, where S represents the number of species in the community, $min(d_{ij})$ represents the distance between species i and j, and a_i corresponds to the abundance of species i. In the case of d_{ij}, j refers to the closest relative of the species i.

$$MNND = \sum_{i=1}^{S-1} min(d_{ij}) a_i, \tag{3}$$

The phylogenetic species variability (PSV) index measures the variation between two species in a community, and quantifies the phylogenetic relationship between them.The PSV is calculated using Eq. (4), where C is a matrix, trC is the sum of the diagonal values of this matrix, $\sum c$ represents the sum of all of the values in the matrix, and S is the total number of species.

$$PSV = \frac{n(trC) - \sum c}{S(S-1)}. \tag{4}$$

The phylogenetic species richness (PSR) index calculates the richness of the species present in a community based on their variability [29]. As shown in Eq. (5), this calculation is done by multiplying the number of species (S) by the PSV.

$$PSR = S * PSV. \tag{5}$$

The mean phylogenetic distance (MPD) represents the average phylogenetic distance, which is calculated by analyzing combinations of all pairs of species in the community [30]. The equation for this index uses the total number of species, indicated by S, the phylogenetic distance between each pair of species, denoted by d_{ij}, and a variable $p_i p_j$, which takes a value of one if the species is present and zero otherwise. The term $\sum_{i=1}^{S-1}\sum_{j=i+1}^{S}$ is a double sum of the products of the distances between all species in the tree and the value indicating the presence or absence of the species, i.e., one or zero Eq. (6) shows the formula used to calculate the MPD.

$$MPD = \frac{\sum_{i=1}^{S-1}\sum_{j=i+1}^{S} d_{ij} p_i p_j}{\sum_{i=1}^{S-1}\sum_{j=i+1}^{S} p_i p_j}. \tag{6}$$

Table 2
Distances calculated for the cladogram shown in Fig. 4.

i	j	D_{ij}
0	1	D_{01} = 1 - 0 + 1 = 2
0	2	D_{02} = 2 - 0 + 1 = 3
0	3	D_{03} = 3 - 0 + 1 = 4
0	4	D_{04} = 4 - 0 + 1 = 5
1	2	D_{12} = 2 - 1 + 2 = 3
1	3	D_{13} = 3 - 1 + 2 = 4
1	4	D_{14} = 4 - 1 + 2 = 5
2	3	D_{23} = 3 - 2 + 2 = 3
2	4	D_{24} = 4 - 2 + 2 = 4
3	4	D_{34} = 4 - 3 + 2 = 3

The taxonomic diversity index ($\triangle$) value represents the average phylogenetic distance between the individuals of the species [31]. This index takes into consideration the number of individuals of each species and the taxonomic relationships between them. The formula for calculating $\triangle$ is defined by Eq. (7), where a_i (i = 1,..., S) represents the abundance of species i, a_j (j = 1,..., S) represents the abundance of species j, S indicates the total number of species, n denotes the total number of individuals, and d_{ij} is the taxonomic distance between species i and j.

$$\triangle = \frac{\sum_{i=1}^{S-1}\sum_{j=i+1}^{S} d_{ij}a_i a_j}{[n(n-1)/2]}. \tag{7}$$

Finally, the taxonomic distinction index ($\triangle^*$), defined by Eq. (8), expresses the average taxonomic distance between two individuals of different species [31]. In this expression, a_i (i = 1,...,S) is the abundance of species i, a_j (j = 1,...,S) is the abundance of species j, S is the total number of species and d_{ij} is the taxonomic distance between species i and j.

$$\triangle^* = \frac{\sum_{i=1}^{S-1}\sum_{j=i+1}^{S} d_{ij}a_i a_j}{\sum_{i=1}^{S-1}\sum_{j=i+1}^{S} a_i a_j}, \tag{8}$$

3.2.2. Example of index calculation for an image

The equations in Section 3.2.1 are derived in relation to biological concepts. For a better understanding of how these indexes are calculated for images, we present an example of a three-dimensional image with two slices, from which we extract the cladogram and calculate the distances and the eight indexes. We used a small image so that the calculations were not extensive in the paper. Fig. 4 shows the example image and the extracted cladogram. We can see that the image has species diversity in relation to the voxels. It is should be noted that the image in Fig. 4 is simply an example, and we have not used the true voxel values for the colors in the figure.

To calculate the phylogenetic distances based on the cladogram, we use the following equations:

$D_{ij} = j - i + 1$, if $i = 0$, and

$D_{ij} = j - i + 2$, if $i \neq 0$, where i and j are two different species.

Table 2 shows the distances obtained using these equations for the cladogram in Fig. 4.

In our implementation of these indexes, we represent the cladogram as a histogram structure. Each position in the histogram represents a species (which are the intensities in the image), and each value refers to the abundance (which is the number of voxels with each intensity). We can then calculate the distances using the histogram. When constructing the cladogram, we apply a simple but effective optimization for faster extraction, in which we assemble the tree using only the range of levels of color variation in each image; this means that it is not necessary to choose a fixed, standardized size for the histogram, and a variable size for the analyzed image can therefore be used. Based on the image, the cladogram and the values of the distances, we can calculate the indexes described in Section 3.2.1.

Table 3
Calculation of the L_i and A_i terms of the PD index.

i	j	L_i	A_i	$L_i * A_i$
0	1	2	1.5	3
0	2	3	1	3
0	3	4	1.25	5
0	4	5	1.6	8
1	2	3	1.5	4.5
1	3	4	1.66	6.64
1	4	5	2	10
2	3	3	1	3
2	4	4	1.66	6.64
3	4	3	2.5	7.5
Sum			15.67	57.28

The PD defined in Eq. (1) considers the number of branches of the tree (B), the extension of each branch (L_i), and the average abundance of the species in each branch (A_i). In relation to the images, B is equivalent to the number of species minus one, L_i represents the distance between the species of branch i, and A_i is the average abundance of the species connected to branch i by the number of species also connected to that branch. The values of L_i are the same as in Table 2, while the abundance of species can be seen in Fig. 4. Since the calculation of the PD is more complex than the other indexes, we describe it using Table 3.

Based on the values in Table 3,the PD value for the image example was as follows:

$$PD = B * \frac{\sum_i^B L_i A_i}{\sum_i^B A_i} = 4 * \frac{57.28}{15.67} = 14.62$$

To calculate the SPD, as shown in Eq. (2), we need values for the distance between species (d_{ij}), the number of species (S) (which in this case is the number of intensities in the histogram), and the species abundance (a) (which represents the number of voxels of a given intensity). Substituting the values into each equation, we obtain:

$$\sum_{i=1}^{S-1}\sum_{j=i+1}^{S} d_{ij}a_i a_j = 87$$

$$\sum_{i=1}^{S-1}\sum_{j=i+1}^{S} a_i a_j = 21$$

$$\left(\frac{S*(S-1)}{2}\right) = 10$$

$$SPD = 10 * \frac{87}{21} = 42.42$$

MNND, as shown in Eq. (3), requires only the distance from one species to its closest relative. Here, S indicates the number of species (number of intensities in the histogram), i denotes the species in question (intensity i), j represents the closest relative of i (intensity j, which refers to the intensity following i), and a_i denotes the abundance of the species i (number of voxels with intensity i). Since our cladogram has only one path between one species and another, the minimum path is the only path between species. Thus, for our image, the MNND has the following value:

$$MNND = \sum_{i=1}^{S-1} min(d_{ij})a_i = 15.00$$

To calculate the PSV, as shown in Eq. (4),we consider operations applied to the image, which is the matrix C. Here, trC is the sum

of the diagonal image, $\sum c$ represents the sum of all of the voxels in the image, and $\bar{c}$ refers to the average of the voxels outside the diagonal. In addition, S represents the number of species in the community for which individuals are present, i.e., the number of intensities that are present in the image. We also consider a summation of the diagonals of the volume, using the x and y axes for each z traversed in the matrix.

$$PSV = \frac{n(trC) - \sum c}{S(S-1)} = \frac{3*7-21}{3*(3-1)} = 0$$

The PSR, as shown in Eq. (5), is simply the PSV multiplied by the number of species in the community for which individuals are present, i.e., the number of image intensities.

$$PSR = S * PSV = 3 * 0 = 0$$

To calculate the MPD, as defined in Eq. (6), we consider the sum of the distances between species (d_{ij}) multiplied by the variables p_i and p_j, which take on a value of zero if the species is not present, or one if the species is present. Thus, when a given intensity in the histogram exists in the image, p is set to one; otherwise, p is set to zero.

$$\sum_{i=1}^{S-1} \sum_{j=i+1}^{S} d_{ij} p_i p_j = 12$$

$$\sum_{i=1}^{S-1} \sum_{j=i+1}^{S} p_i p_j = 3$$

$$MPD = \frac{12}{3} = 4.00$$

In the calculation of $\triangle$, as defined in Eq. (7), d_{ij} represents the distance between the species (intensities) i and j, a refers to the abundance of the species (the number of voxels for each intensity), and n is the number of individuals in the community (number of voxels in the histogram). Hence, we can calculate $\triangle$ as follows:

$$\Sigma \Sigma_{i<j} d_{ij} a_i a_j = 87$$

$$[n * (n-1)/2] = 28$$

$$\triangle = \frac{110}{36} = 3.10$$

Finally, the calculation of $\triangle^*$, as defined in by Eq. (8), is similar to that of $\triangle$, with the difference that in the this case, of $\triangle^*$ the denominator is the sum of the multiplication products of species the abundances of species. Substituting in the values, we have:

$$\Sigma \Sigma_{i<j} d_{ij} a_i a_j = 87$$

$$\Sigma \Sigma_{i<j} a_i a_j = 21$$

$$\triangle^* = \frac{87}{21} = 4.14$$

3.3. Classification

For pattern recognition from the extracted characteristics, we applied two classifiers that are commonly used in the literature, the Random Forest (RF) [32] and Extreme Gradient Boost (XGBoost) [33] algorithms, both of which were used with the default parameters in the Sklearn library [34]. We applied the cross-validation technique to validate the models, which involved randomly dividing the set of images into k folds of approximately equal size. The first fold was treated as a validation set, and the method was trained using the other k - *1* folds. Each image in the data sample is assigned to an individual group, and stays in that group for the duration of the procedure. This means that each sample can be used in the validation set only once but is used to train the model $k-1$ times. For $k = 5$, each fold contains an average of 20% of the number of images in each class. Table 4 gives a brief description of the classifiers and the main parameters used in the models.

3.4. Validation metrics

To evaluate the classification models obtained in the previous step, we used the following metrics: accuracy, recall, precision, F1-score, and AUC. To calculate these metrics, we analyze the confusion matrix, which is constructed based on four values: true positive (TP), false positive (FP), false negative (FN), and true negative (TN). These represent the numbers of samples classified correctly and incorrectly.

3.5. Example of the proposed method for a real volume

For a better understanding of the proposed method, we present an example of the complete flow of our methodology, using a real volume from the dataset, and illustrate each of the steps described above. We select the image with the smallest distribution of voxel values from the dataset to facilitate the visualization of the results.

Fig. 5 shows this process. The input volume of the solid class (Fig. 5(a)) is passed to the feature extraction step (Fig. 5(b)), where the indexes are calculated using the cladogram created based on the image histogram. Finally, the extracted features are passed to the classifier (Fig. 5(c)), which predicts the input as one of three possible classes based on the rules defined during training.

4. Results

This section presents the results of tests performed on the datasets described in Section 3.1. The characterization was performed using the techniques described in Section 3.2 and classification was carried out as described in Section 3.3. This section is divided into three parts, as follows: (i) we present the results of our method; (ii) we carry out an extensive comparison with results from similar schemes; and finally, (iii) we explore the results obtained from our method by analyzing some cases of success and failure;

Our tests were divided into three types of experiments, as shown in Table 5. Each experiment included combinations of classification with the bases used, with COVID-19 (1) and COVID-19 (2) referring to the two VOI bases of lesions caused by COVID-19, provided by MedSeg. The three scenarios used in these experiments posed important challenges for the evaluation of the proposed method.

Table 6 shows the results achieved from the proposed method for the three experiments summarized in Table 5. These experiments were performed to illustrate the potential of the proposed method for different test scenarios and classifiers.

From Table 6, we can see that the proposed method achieves promising results for the classification of lung lesions. When we analyze the results for Experiments 1 and 3, we observe that the classifier that gave the best results was XGBoost, with AUCs of 0.997 and 0.990, respectively. In Experiment 2, the best results were achieved with RF, with an AUC of 0.996. We believe that the XGBoost and RF results were good because these are tree-based techniques and because the features are very representative, which improves the generalizability. Another important factor is the time required by the classifiers to perform data prediction. In addition to the use of the indexes to carry out characterization efficiently,

Table 4
Classifiers used in the proposed method.

Classifier	Description	Parameters
RF	RF is a regression and classification algorithm developed in [32]. In this method, predictions are made from decision trees. Each tree in the RF provides a class prediction, so the class with the most votes becomes the model's output for the sample in question.	number of estimators (number of trees) = 100, min samples split = 2, min samples leaf = 1, max number of features = "auto" (sqrt-number features) bootstrap = True, max depth = None (unlimited)
XGBoost	XGBoost is an optimized machine learning technique developed in [33], which is based on decision trees and uses a gradient-increasing structure. This algorithm was designed to be flexible and efficient, and its parameters can easily be changed [13]. XGBoost can be applied to regression and classification problems.	max depth = 6, learning rate = 0.1, number of estimators (number of trees) = 100, booster = "gbtree", objective = "binary:logistic", gamma = 0, max delta step = 0

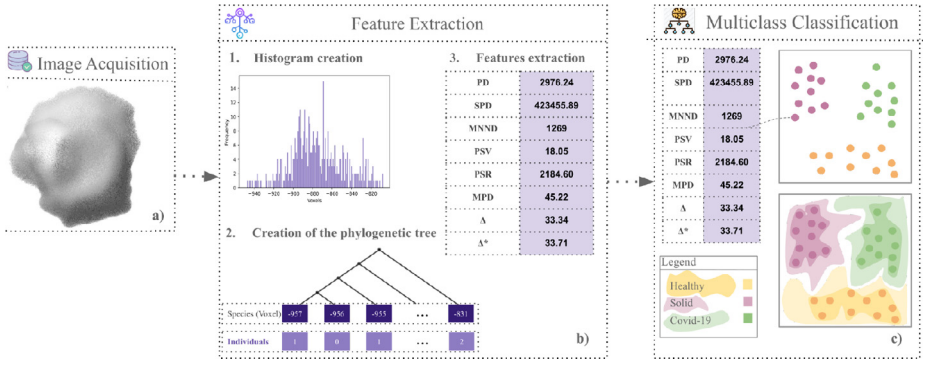

Fig. 5. Flow of the proposed method for a sample image from the dataset.

Table 5
Experiments performed at work.

Experiment	Description/classes	Number of images
1	Healthy tissue × solid × COVID-19 (1) and COVID-19 (2)	19.868
2	Healthy tissue × solid × COVID-19 (1)	19.624
3	Healthy tissue × solid × COVID-19 (2)	19.653

Table 6
Results of the proposed method.

Experiment	Classifier	Acc (%)	Rec (%)	Prec (%)	F1 (%)	AUC	Classification runtime (s)
1	RF	99.90 ± 0.02	99.90 ± 0.02	99.90 ± 0.02	99.90 ± 0.02	0.995 ± 0.001	3.1
	XGBoost	**99.93 ±0.02**	**99.93 ±0.02**	**99.93 ±0.02**	**99.93 ±0.02**	**0.997 ±0.001**	**3.6**
2	**RF**	**99.93 ±0.02**	**99.93 ±0.02**	**99.93 ±0.02**	**99.93 ±0.02**	**0.996 ±0.003**	**2.9**
	XGBoost	99.92 ± 0.02	99.92 ± 0.02	99.92 ± 0.02	99.92 ± 0.02	0.994 ± 0.003	2.5
3	RF	99.86 ± 0.06	99.86 ± 0.06	99.86 ± 0.06	99.86 ± 0.06	0.988 ± 0.004	2.9
	XGBoost	**99.90 ±0.04**	**99.90 ±0.04**	**99.90 ±0.04**	**99.90 ±0.04**	**0.990 ±0.004**	**3.3**

-The best result for each experiment is shown in bold.

the low number of features facilitates the classification task with the techniques used here.

Another important observation that can be made from the data in Table 6 relates to the good balance of the metrics. A good classification method should allow us to successfully classify not only sick cases but also healthy cases, thus enabling more efficient patient triage.

To analyze the impact of the results achieved by our method concerning other descriptors, we performed a series of comparisons to investigate the potential of the proposed method. For each experiment in Table 5, we evaluated the following descriptors:

- Traditional approaches
  - Histogram features: we extracted several descriptors (mean, variance, kurtosis, energy, entropy, and skewness), which were calculated from the histogram of the image;
  - GLCM: we calculated several attributes that are widely used in the literature (dissimilarity, homogeneity, contrast, energy, correlation, and second angular moment). To calculate the GLCM, we considered angles of 0°, 45°, 90°, and 135°, and a distance between pixels of 3;
- Deep learning architectures: we applied the TL technique, in which a model trained on one task can be used on a different one. We used the following pre-trained models on the ImageNet [35]
  - DenseNet-121, DenseNet-169, DenseNet-201, Inception-V3, VGG16, EfficientNet-B0, and EfficientNet-B1.

All of the results in Table 7 were obtained using k-fold cross-validation, with k=5. For clarity, we present only the best result for each descriptor in each experiment.

Based on the values in Table 7, we can conclude that the classification problem addressed in this work is challenging. A CAD sys-

Table 7
Comparison with other descriptors.

Experiment	Method	Classifier	Acc (%)	Rec (%)	Prec (%)	F1 (%)	AUC	Classification runtime (s)
1	Histogram	XGBoost	96.52 ± 0.20	96.52 ± 0.20	96.38 ± 0.23	96.28 ± 0.20	0.840 ± 0.007	7.5
	GLCM (angle 45)	XGBoost	98.17 ± 0.09	98.17 ± 0.09	98.09 ± 0.10	98.11 ± 0.10	0.904 ± 0.004	117.1
	DenseNet-121	XGBoost	98.81 ± 0.02	98.81 ± 0.02	98.73 ± 0.03	98.74 ± 0.03	0.863 ± 0.003	4018.4
	DenseNet-169	XGBoost	98.80 ± 0.05	98.80 ± 0.05	98.72 ± 0.06	98.72 ± 0.05	0.858 ± 0.005	5924.6
	DenseNet-201	XGBoost	98.83 ± 0.02	98.83 ± 0.02	98.76 ± 0.03	98.76 ± 0.03	0.866 ± 0.003	7420.2
	Inception-V3	XGBoost	98.71 ± 0.02	98.71 ± 0.02	98.61 ± 0.03	98.61 ± 0.03	0.844 ± 0.002	5228.4
	VGG16	XGBoost	98.95 ± 0.03	98.95 ± 0.03	98.90 ± 0.04	98.91 ± 0.04	0.888 ± 0.004	1249.3
	EfficientNet-B0	XGBoost	99.07 ± 0.02	99.07 ± 0.02	99.03 ± 0.02	99.04 ± 0.02	0.902 ± 0.002	6738.7
	EfficientNet-B1	XGBoost	99.02 ± 0.02	99.02 ± 0.02	98.98 ± 0.02	98.98 ± 0.02	0.897 ± 0.003	6589.5
	Proposed Method	**XGBoost**	**99.93 ±0.02**	**99.93 ±0.02**	**99.93 ±0.02**	**99.93 ±0.02**	**0.997 ±0.001**	**3.6**
2	Histogram	XGBoost	97.01 ± 0.24	97.01 ± 0.24	96.86 ± 0.29	96.70 ± 0.28	0.799 ± 0.015	7.3
	GLCM (angle 45)	XGBoost	98.28 ± 0.06	98.28 ± 0.06	98.21 ± 0.07	98.23 ± 0.07	0.907 ± 0.003	95.2
	DenseNet-121	XGBoost	98.96 ± 0.04	98.96 ± 0.04	98.89 ± 0.04	98.90 ± 0.04	0.869 ± 0.005	3833.0
	DenseNet-169	XGBoost	98.93 ± 0.03	98.93 ± 0.03	98.86 ± 0.03	98.87 ± 0.04	0.862 ± 0.006	7355.8
	DenseNet-201	XGBoost	98.96 ± 0.02	98.96 ± 0.02	98.90 ± 0.03	98.91 ± 0.03	0.869 ± 0.003	8617.4
	Inception-V3	XGBoost	98.86 ± 0.01	98.86 ± 0.01	98.79 ± 0.01	98.79 ± 0.02	0.852 ± 0.004	5056.9
	VGG16	XGBoost	99.02 ± 0.03	99.02 ± 0.03	98.97 ± 0.04	98.98 ± 0.04	0.881 ± 0.004	1165.5
	EfficientNet-B0	XGBoost	99.09 ± 0.02	99.09 ± 0.02	99.05 ± 0.02	99.06 ± 0.02	0.892 ± 0.001	6780.7
	EfficientNet-B1	XGBoost	99.06 ± 0.04	99.06 ± 0.04	99.02 ± 0.04	99.03 ± 0.04	0.887 ± 0.004	6601.9
	Proposed Method	**RF**	**99.93 ±0.02**	**99.93 ±0.02**	**99.93 ±0.02**	**99.93 ±0.02**	**0.996 ±0.003**	**2.9**
3	Histogram	XGBoost	97.15 ± 0.15	97.15 ± 0.15	97.09 ± 0.14	97.01 ± 0.17	0.863 ± 0.009	7.2
	GLCM (angle 0)	XGBoost	98.40 ± 0.06	98.40 ± 0.06	98.33 ± 0.06	98.34 ± 0.06	0.866 ± 0.005	109.3
	DenseNet-121	XGBoost	99.27 ± 0.01	99.27 ± 0.01	99.24 ± 0.01	99.22 ± 0.01	0.828 ± 0.003	3246.4
	DenseNet-169	XGBoost	99.25 ± 0.01	99.25 ± 0.01	99.22 ± 0.02	99.19 ± 0.01	0.822 ± 0.003	6607.8
	DenseNet-201	XGBoost	99.28 ± 0.02	99.28 ± 0.02	99.24 ± 0.02	99.23 ± 0.02	0.830 ± 0.004	8448.9
	Inception-V3	XGBoost	99.19 ± 0.02	99.19 ± 0.02	99.16 ± 0.02	99.13 ± 0.03	0.815 ± 0.004	4150.1
	VGG16	XGBoost	99.40 ± 0.01	99.40 ± 0.01	99.39 ± 0.01	99.38 ± 0.01	0.893 ± 0.003	996.6
	EfficientNet-B0	XGBoost	99.47 ± 0.01	99.47 ± 0.01	99.46 ± 0.01	99.46 ± 0.01	0.917 ± 0.004	5563.4
	EfficientNet-B1	XGBoost	99.45 ± 0.01	99.45 ± 0.01	99.45 ± 0.01	99.45 ± 0.01	0.920 ± 0.002	5519.4
	Proposed Method	**XGBoost**	**99.90 ±0.04**	**99.90 ±0.04**	**99.90 ±0.04**	**99.90 ±0.04**	**0.990 ±0.004**	**3.3**

tem must strike a good balance between the evaluation metrics, and the proposed method therefore achieved the most promising results in all of the experiments; it yielded an AUC value of above 0.99, indicating that it was able to classify both cases with lesions and healthy cases, enabling more effective triage of patients. s expected, the results achieved by CNN-based methods were promising, despite their disadvantages, for example (i) the need for class balancing to achieve good learning by the models; (ii) the need for adequate parameterization for the problem; (iii) the need for good hardware depending on the problem; (iv) the need for large amounts of data to identify the correct standards; and (v) the requirement for each architecture to have different dimensions from the input images. Although the use of TL mitigated these limitations to some degree, the results achieved by the proposed method were better. The same held true when our method was compared to traditional descriptors, as it gave superior results. We believe that this was due to the limitations on the characterization of these descriptors.

An analysis of the execution time for the classification shows that the proposed method required less time than the other approaches due to the small number of attributes generated by the indexes. As the results in Table 6 show, XGBoost was the classifier that achieved the best results. It appears that the characterization capabilities of the diversity indexes were sufficient to highlight the texture contained in each image class effectively. It should also be highlighted that the results were promising even for unbalanced classes, since the presence of unbalanced data is similar to the conditions in a real clinical environment.

Finally, we compared the results obtained in this work with those of the studies reviewed in Section 2. The purpose of this comparison was simply to compare the results achieved by the proposed method with the state of the art, and was not intended to disparage other methods. The first comparison is shown in Table 8, which summarizes the main similarities and differences between the proposed approach and other related works that have used phylogenetic diversity in their methodology (Section 2.1).

From Table 8, we can see that the diversity indexes have been applied in a wide range of problems in the context of medical imaging. We highlight the differences and the advantages of the proposed method over other works based on diversity indexes, as follows:

- Regarding the domain in the image, only the scheme in [11] applied indexes to 3D images, and these were different from the indexes proposed in this work. In addition, the phylogenetic tree used by our method was the cladogram, whereas the dendrogram was used in [11];
- The methods developed in [12,13] aimed to solve different problems using 2D images;
- In addition to their different goals, all of the methodologies shown in Table 8 used different strategies to achieve their goals; and
- We performed a simple but effective optimization of the cladogram that allowed for higher efficiency of the index calculation process.

Next, we compared our approach with those in the papers reviewed in Section 2.2. Table 9 shows a summary of these methods. Only the best results from our method are highlighted in the table.

From Table 9, several important points can be noted. Overall, our approach proved superior to the other methods; however, each study used a different dataset, and only [15,24] performed a multiclass classification. The types of images differed: only [22–24] used CT volumes in their method, whereas the others used 2D regions.

None of the other papers in Table 9 classified lung regions into COVID-19, solid lesions and healthy tissue. For this reason, we performed the comparison presented earlier (Table 7), where we used techniques from related papers to reproduce the experiments using our dataset; this allowed for a fairer comparison with the other approaches, and demonstrated the efficiency of the proposed descriptors in terms of characterizing the images.

Our method uses information from the image itself to perform the classification; since we extract features that can repre-

Table 8
Comparison with other approaches using diversity indexes.

Work	Goal	Image type	Methodology	Descriptors	Number of indexes in common with the proposed work	Classification	Results
[11]	Classification of lung nodules into benign and malignant	CT (3D)	Diversity indexes are adapted to generate a standardized entry for CNN	Topology-based phylogenetic diversity indexes: sum of basic taxic weights and sum of standardized taxic weights	0	k-fold cross-validation, with k = 10	Accuracy: 92.63% Sensitivity: 90.7% Specificity: 93.47% ROC: 0.934
[12]	Glaucoma classification	Retinal images (2D)	Generative Adversarial Network used in conjunction with taxonomic indexes	Taxonomic diversity indexes: $\triangle$ and $\triangle^*$	2	k-fold cross-validation, with k = 10	Accuracy: 100% Sensitivity: 100% Specificity: 100% ROC: 1
[13]	Breast cancer diagnosis	Histological images (2D)	Phylogenetic diversity indexes used for classification and content-based image retrieval	Phylogenetic diversity indexes: PD, SPD, MNND, PSV and PSR	5	k-fold cross-validation, with k = 10	Accuracy: 95.0% Precision: 96.0% AUC: 0.98
Proposed Method	Lung lesions classification for COVID-19 detection	CT - COVID-19 (3D)	Phylogenetic diversity indexes and cladogram optimization for classification of multiple lung lesions	Phylogenetic diversity indexes + taxonomic diversity indexes: PD, SPD, MNND, PSV, PSR, MPD, $\triangle$ and $\triangle^*$	-	k-fold cross-validation, with k = 5	Accuracy: 99.93% Recall: 99.93% Precision: 99.93% F1-score: 99.93% AUC: 0.997

Table 9
Comparison of our method with related works.

Work	Exam type	Number of images	Goal/Classes	Acc (%)	Recall (%)	Precision (%)	F1-score (%)	AUC
[14]	X-ray	206	COVID-19 and non-COVID-19	95.12	97.91	-	-	-
[15]	X-ray	558	COVID-19, bacterial pneumonia, non-COVID-19 viral pneumonia and normal	79.52	-	-	-	0.87
[16]	CT	118	Non-severe and severe COVID-19	89.00	-	-	-	0.98
[17]	CT	746	COVID-19 and non-COVID-19	86.00	-	-	85.00	0.94
[18]	CT	150	COVID-19 and non-COVID-19	98.27	98.93	97.63	98.28	-
[2]	CT	453	COVID-19 and non-COVID-19	82.90	84.00	-	-	-
[19]	CT	1396	COVID-19 and non-COVID-19	95.99	94.04	-	92.84	0.99
[20]	CT	2492	COVID-19 and non-COVID-19	96.25	96.29	96.29	96.29	-
[21]	CT	4356	COVID-19 and non-COVID-19	-	90.00	-	-	0.96
[22]	CT	540	COVID-19 and non-COVID-19	-	90.70	-	-	0.95
[23]	CT	81.00	Common and severe COVID-19	-	-	-	-	0.93
[24]	CT	103	Healthy, IPF and COVID-19 cases	89.60	96.10	-	-	-
Proposed Method Experiment 1 - XGBoost	**CT**	**19.868**	**COVID-19, solid lesion and healthy tissue**	**99.93**	**99.93**	**99.93**	**99.93**	**0.997**

sent the image well, there is no need for exhaustive training to search for the best classification architectures and parameters, and we achieved an AUC of higher than 0.98 for both RF and XGBoost.

4.1. Case study

In this section, we analyze our results based on some samples that were correctly classified by the method and others that were incorrectly classified. For this analysis, we chose the best result from Scenario 1.

We first present a plot of the features extracted with the proposed indexes. We applied principal component analysis (PCA) to reduce the dimensionality of the data while preserving the most important information. Fig. 6(a) shows both the PCA has three components of the data, and the generated confusion matrix is shown in Fig. 6(b). It can be seen that the healthy tissue class, for which the results were more accurate, was spatially distant from the other classes, which facilitated the work of the classifiers. In contrast, most of the samples of solid lesions and COVID-19 classes were spatially close, but it was possible to draw a decision boundary between them.

In the next analysis, we randomly selected a sample from each item in the confusion matrix for the test case. Fig. 7 shows selected images in the form of a confusion matrix. The red dashes represent

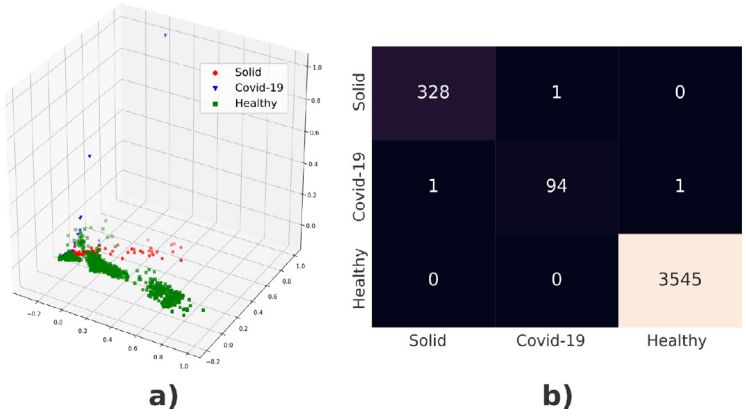

Fig. 6. a) PCA with three components representing the extracted characteristics; b) confusion matrix of results.

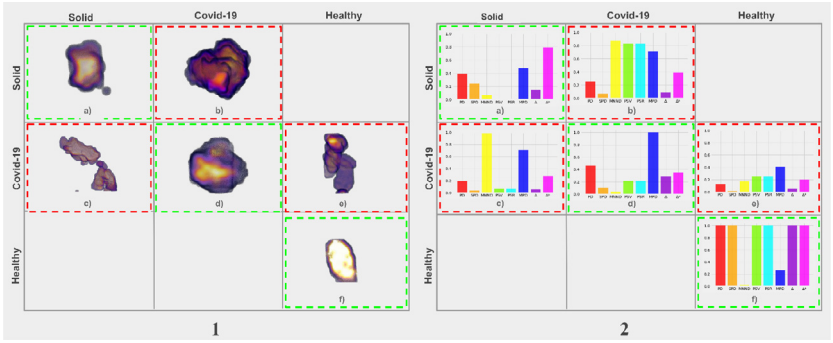

Fig. 7. (1) Confusion matrix with correctly and incorrectly classified regions; (2) graphs of the indexes extracted from the images in the confusion matrix.

the samples in which the classifier was wrong, while the green ones represent the correct answers.

We can see that the healthy tissue class has a markedly different presentation from the others, thus proving the PCA presented in the previous analysis and highlighted by the indexes in Fig. 7.2(b). Likewise, we can see that the tissue in Fig. 7.1(b) has a texture similar to that in Fig. 7.1(d), and a similar correspondence can be seen in the patterns in Fig. 7.2(b) and (d), thus justifying the erroneous classification of the class of solid lesions with that of COVID-19. The lesions in Figs. 7.1(c) and 7.2(c) both have a similar texture and index values closer to the class of solid lesions than the others.

Finally, for the lesion in Fig. 7.1(e), we can see that the graph in Fig. 7.2(e) differs from the pattern seen in the other graphs in Fig. 7.2, which is why the classifier identified it as healthy tissue.

5. Discussion

The proposed method extracts texture descriptors to enable the classification of images of lung tissue. Based on the results presented in Section 4, we point out some highlights and some aspects that need further investigation.

5.1. Contributions of the proposed methodology

1. The indexes are used to characterize the volumes in their original dimensions, i.e., 3D images;
2. The indexes achieve promising results without the need for data augmentation or pre-processing techniques;
3. The proposed descriptors do not require parameterization;
4. Promising results are obtained using two different classifiers, thus demonstrating the quality of the extracted features;
5. We performed three experiments combining the bases used in this work, with the aim of evaluating our method on images from different sources;
6. We performed several comparisons of results using different image characterization techniques, and the proposed method was shown to be as effective as the other techniques and to achieve slightly better results.

5.2. Limitations of the proposed methodology

1. Images with large differences between the higher and lower voxel intensities may require a longer time for the calculation of the indexes.

6. Conclusion

In this work, we propose a method for characterizing the texture of lung tissue that can be used to correctly classify lesions caused by COVID-19, solid lesions, and healthy tissue. The application of this methodology can assist in the early diagnosis of COVID-19 by a specialist or as part of a screening mechanism.

6.1. Evaluation of our approach

- The phylogenetic diversity indexes adopted in this work, when used with the XGBoost classifier, showed the best efficiency in terms of the characterization and classification of lung lesions, and yielded promising results in all of the experiments that were performed;
- The results were promising and consistent in all test scenarios and with all of the classifiers used, motivating the application of descriptors proposed in real environments;

- Of all the works considered here, the set of images used with our method was the most extensive; this is important, because it showed that the proposed descriptors achieved good results, even on a wide variety of images;
- The properties of COVID-19 lesions are also common to other lung diseases, and especially to viral forms of pneumonia. The individual classification of lesions is a strong advantage of our method;
- We believe that the promising results obtained by our method are related to the use of phylogenetic indexes for characterization, as each of these contributes to a particular attribute of the image, meaning that our approach can achieve good discrimination between tissue types associated with lung lesions.

6.2. Future work

1. Use other sets of COVID-19 images to improve the effectiveness of the prediction method;
2. Explore the use of image sets that include other types of lesions from patients with COVID-19, such as multifocal solid lesion irregular nodules and nodules with visible halo signs;
3. Adapt our method for the classification of COVID-19 variants.

We believe that the method presented here can be integrated into a CAD that can be applied in real situations to aid in the diagnosis of COVID-19. Our approach offers benefits both to specialists, who can obtain a second opinion at the diagnosis stage, and to patients, since early detection is important to allow them to receive treatment promptly, thus increasing the chances of a cure.

Declaration of Competing Interest

Authors certify that they have NO affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers' bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript.

Acknowledgement

The proposed method was supported by the following institutions: FAPEPI - www.fapepi.pi.gov.br (5492.UNI253.59248.15052018); CAPES - www.capes.gov.br; and the CNPq - www.cnpq.br (435244/2018-3).

References

[1] W. H. O. WHO, Coronavirus disease (COVID-19) outbreak situation, 2021, (Retrieved from https://www.who.int/emergencies/diseases/novel-coronavirus-2019), Accessed April 13, 2021.
[2] S. Wang, B. Kang, J. Ma, X. Zeng, M. Xiao, J. Guo, M. Cai, J. Yang, Y. Li, X. Meng, B. Xu, A deep learning algorithm using CT images to screen for corona virus disease (COVID-19), medRxiv (2020). 10.1101/2020.02.14.20023028
[3] H. Zhang, P. Zhou, Y. Wei, H. Yue, Y. Wang, M. Hu, S. Zhang, T. Cao, C. Yang, M. Li, et al., Histopathologic changes and SARS-cov-2 immunostaining in the lung of a patient with COVID-19, Ann. Intern. Med. 172 (9) (2020) 629–632, doi:10.7326/M20-0533.
[4] W. Li, L. Hu, J. Huang, F. Lv, B. Fu, Z. Chu, Outcome of pulmonary spherical ground-glass opacities on ct in patients with coronavirus disease 2019 (covid-19): a retrospective analysis (2020). 10.21203/rs.3.rs-30665/v1
[5] M. Gülbay, B.O. Özbay, B.A.R. Mendi, A. Baştuğ, H. Bodur, A CT radiomics analysis of COVID-19-related ground-glass opacities and consolidation: is it valuable in a differential diagnosis with other atypical pneumonias? PLoS One 16 (3) (2021) 1–21, doi:10.1371/journal.pone.0246582.
[6] J. Wu, X. Tang, Brain segmentation based on multi-atlas and diffeomorphism guided 3d fully convolutional network ensembles, Pattern Recognit. 115 (2021) 107904, doi:10.1016/j.patcog.2021.107904. URL https://www.sciencedirect.com/science/article/pii/S0031320321000911
[7] B. Chen, W. Guan, P. Li, N. Ikeda, K. Hirasawa, H. Lu, Residual multi-task learning for facial landmark localization and expression recognition, Pattern Recognit. 115 (2021) 107893, doi:10.1016/j.patcog.2021.107893. URL https://www.sciencedirect.com/science/article/pii/S0031320321000807
[8] Y. Xiao, C.-M. Pun, B. Liu, Fooling deep neural detection networks with adaptive object-oriented adversarial perturbation, Pattern Recognit. 115 (2021) 107903, doi:10.1016/j.patcog.2021.107903. URL https://www.sciencedirect.com/science/article/pii/S003132032100090X
[9] J. Liu, S. Teng, L. Fei, W. Zhang, X. Fang, Z. Zhang, N. Wu, A novel consensus learning approach to incomplete multi-view clustering, Pattern Recognit. 115 (2021) 107890, doi:10.1016/j.patcog.2021.107890. URL https://www.sciencedirect.com/science/article/pii/S0031320321000777
[10] Y. Jin, Y. Zhang, Y. Cen, Y. Li, V. Mladenovic, V. Voronin, Pedestrian detection with super-resolution reconstruction for low-quality image, Pattern Recognit. 115 (2021) 107846, doi:10.1016/j.patcog.2021.107846. URL https://www.sciencedirect.com/science/article/pii/S0031320321000339
[11] A.O. de Carvalho Filho, A.C. Silva, A.C. de Paiva, R.A. Nunes, M. Gattass, Classification of patterns of benignity and malignancy based on CT using topology-based phylogenetic diversity index and convolutional neural network, Pattern Recognit. 81 (2018) 200–212, doi:10.1016/j.patcog.2018.03.032. URL https://www.sciencedirect.com/science/article/pii/S0031320318301237
[12] T.R.V. Bisneto, A.O. de Carvalho Filho, D.M.V. Magalhães, Generative adversarial network and texture features applied to automatic glaucoma detection, Applied Soft Computing 90 (2020) 106165, doi:10.1016/j.asoc.2020.106165. URL https://www.sciencedirect.com/science/article/pii/S1568494620301058
[13] E.D. Carvalho, A.O.C. Filho, R.R.V. Silva, F.H.D. Araújo, J.O.B. Diniz, A.C. Silva, A.C. Paiva, M. Gattass, Breast cancer diagnosis from histopathological images using textural features and CBIR, Artif. Intell. Med. 105 (2020) 101845, doi:10.1016/j.artmed.2020.101845.
[14] A. Abbas, M. Abdelsamea, M. Medhat Gaber, Classification of COVID-19 in chest x-ray images using detrac deep convolutional neural network (2020). URL https://europepmc.org/article/PPR/PPR138002. 10.1101/2020.03.30.20047456
[15] L. Hussain, T. Nguyen, H. Li, A.A. Abbasi, K.J. Lone, Z. Zhao, M. Zaib, A. Chen, T.Q. Duong, Machine-learning classification of texture features of portable chest x-ray accurately classifies COVID-19 lung infection, BioMed. Eng. OnLine 19 (1) (2020) 1–18, doi:10.1186/s12938-020-00831-x.
[16] Z. Tang, W. Zhao, X. Xie, Z. Zhong, F. Shi, T. Ma, J. Liu, D. Shen, Severity assessment of COVID-19 using CT image features and laboratory indices, Phys. Med. Biol. 66 (3) (2021) 035015, doi:10.1088/1361-6560/abbf9e. URL https://pubmed.ncbi.nlm.nih.gov/33032267/
[17] X. He, X. Yang, S. Zhang, J. Zhao, Y. Zhang, E. Xing, P. Xie, Sample-efficient deep learning for COVID-19 diagnosis based on CT scans, medRxiv (2020). 10.1101/2020.04.13.20063941
[18] U. Özkaya, c. Öztürk, M. Barstugan, Coronavirus (COVID-19) classification using deep features fusion and ranking technique, in: Big Data Analytics and Artificial Intelligence Against COVID-19: Innovation Vision and Approach, Springer International Publishing, 2020, pp. 281–295, doi:10.1007/978-3-030-55258-9_17.
[19] H. Yasar, M. Ceylan, A novel comparative study for detection of covid-19 on CT lung images using texture analysis, machine learning, and deep learning methods, Multimed. Tools Appl. 80 (4) (2021) 5423–5447, doi:10.1007/s11042-020-09894-3.
[20] A. Jaiswal, N. Gianchandani, D. Singh, V. Kumar, M. Kaur, Classification of the COVID-19 infected patients using densenet201 based deep transfer learning, J. Biomol. Struct. Dyn. (2020) 1–8, doi:10.1080/07391102.2020.1788642. PMID: 32619398
[21] L. Li, L. Qin, Z. Xu, Y. Yin, X. Wang, B. Kong, J. Bai, Y. Lu, Z. Fang, Q. Song, K. Cao, D. Liu, G. Wang, Q. Xu, X. Fang, S. Zhang, J. Xia, J. Xia, Using artificial intelligence to detect COVID-19 and community-acquired pneumonia based on pulmonary CT: evaluation of the diagnostic accuracy, Radiology 296 (2) (2020) E65–E71, doi:10.1148/radiol.2020200905. PMID: 32191588
[22] C. Zheng, X. Deng, Q. Fu, Q. Zhou, J. Feng, H. Ma, W. Liu, X. Wang, Deep learning-based detection for COVID-19 from chest CT using weak label, medRxiv (2020). 10.1101/2020.03.12.20027185
[23] W. Wei, X.-W. Hu, Q. Cheng, Y.-M. Zhao, Y.-Q. Ge, Identification of common and severe COVID-19: the value of CT texture analysis and correlation with clinical characteristics, Eur. Radiol. 30 (12) (2020) 6788–6796, doi:10.1007/s00330-020-07012-3. URL https://europepmc.org/articles/PMC7327490
[24] A.S. Gaudêncio, P.G. Vaz, M. Hilal, G. Mahé, M. Lederlin, A. Humeau-Heurtier, J.M. Cardoso, Evaluation of COVID-19 chest computed tomography: A texture analysis based on three-dimensional entropy, Biomed. Signal Process. Control 68 (2021) 102582, doi:10.1016/j.bspc.2021.102582. URL https://www.sciencedirect.com/science/article/pii/S1746809421001798
[25] dataset, S.G. Armato III, G. McLennan, L. Bidaut, M.F. McNitt-Gray, C.R. Meyer, A.P. Reeves, B. Zhao, D.R. Aberle, C.I. Henschke, E.A. Hoffman, et al., The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans, Med. Phys. 38 (2) (2011) 915–931, doi:10.1118/1.3528204.
[26] dataset, MedSeg, Covid-19 ct segmentation dataset, 2020, (Retrieved from http://medicalsegmentation.com/covid19/), Accessed April 13, 2021.
[27] A.O. de Carvalho Filho, W.B. de Sampaio, A.C. Silva, A.C. de Paiva, R.A. Nunes, M. Gattass, Automatic detection of solitary lung nodules using quality threshold clustering, genetic algorithm and diversity index, Artif. Intell. Med. 60 (3) (2014) 165–177, doi:10.1016/j.artmed.2013.11.002. URL https://www.sciencedirect.com/science/article/pii/S0933365713001541

[28] D.P. Faith, Conservation evaluation and phylogenetic diversity, Biol. Conserv. 61 (1) (1992) 1–10, doi:10.1016/0006-3207(92)91201-3.
[29] M. Helmus, T. Bland, C. Williams, A. Ives, Phylogenetic measures of biodiversity, Am. Nat. 169 (3) (2007) E68–E83, doi:10.1086/511334. PMID: 17230400
[30] C. Webb, Exploring the phylogenetic structure of ecological communities: an example for rain forest trees, Am. Nat. 156 (2000) 145–155, doi:10.1086/303378. URL https://pubmed.ncbi.nlm.nih.gov/10856198/
[31] K. Clarke, R. Warwick, A taxonomic distinctness index and its statistical properties, J. Appl. Ecol. 35 (1998) 523–531, doi:10.1046/j.1365-2664.1998.3540523.x.
[32] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32, doi:10.1023/A:1010933404324.
[33] T. Chen, C. Guestrin, Xgboost: a scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, in: KDD '16, Association for Computing Machinery, New York, NY, USA, 2016, pp. 785–794, doi:10.1145/2939672.2939785.
[34] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[35] J. Deng, W. Dong, R. Socher, L. Li, K. Li, L. Fei-Fei, Imagenet: a large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255, doi:10.1109/CVPR.2009.5206848.

Vitória de Carvalho Brito graduated in Bachelor of Information Systems from the Federal University of Piauí (UFPI).

Patrick Ryan Sales dos Santos graduated in Bachelor of Information Systems from the Federal University of Piauí (UFPI).

Nonato Rodriges de Sales Carvalho master's degree student in Electrical Engineering from the Federal University of Piauí (UFPI). He is currently an information technician and head of the information technology division at UFPI/CSHNB, Brazil.

Antonio Oseas de Carvalho Filho received a Ph.D. in Electrical Engineering at Federal University of Maranhão - Brazil in 2016. Currently, he is a professor at the Federal University of Piaui (UFPI). His research interests include medical image processing, machine learning and deep learning.