

Article

Automatic Detection of Pulmonary Nodules using Three-dimensional Chain Coding and Optimized Random Forest

May Phu Paing ¹, Kazuhiko Hamamoto ², Supan Tungjitkusolmun ¹, Sarinporn Visitsattapongse ¹ and Chuchart Pintavirooj ^{1,*}

- ¹ Faculty of Engineering, King Mongkut's Institute of Technology Ladkrabang, Bangkok 10520, Thailand; 59601030@kmitl.ac.th (M.P.P.); supan.tu@kmitl.ac.th (S.T.); sarinporn.vi@kmitl.ac.th (S.V.)
- ² School of Information and Telecommunication Engineering, Tokai University, Tokyo 108-8619, Japan; hama@keyaki.cc.u-tokai.ac.jp
- * Correspondence: chuchart.pi@kmitl.ac.th; Tel.: +66-83-449-7543

Received: 7 February 2020; Accepted: 25 March 2020; Published: 29 March 2020



Abstract: The detection of pulmonary nodules on computed tomography scans provides a clue for the early diagnosis of lung cancer. Manual detection mandates a heavy radiological workload as it identifies nodules slice-by-slice. This paper presents a fully automated nodule detection with three significant contributions. First, an automated seeded region growing is designed to segment the lung regions from the tomography scans. Second, a three-dimensional chain code algorithm is implemented to refine the border of the segmented lungs. Lastly, nodules inside the lungs are detected using an optimized random forest classifier. The experiments for our proposed detection are conducted using 888 scans from a public dataset, and achieves a favorable result of 93.11% accuracy, 94.86% sensitivity, and 91.37% specificity, with only 0.0863 false positives per exam.

Keywords: automated seeded region growing; 3D chain code; firefly; lung cancer; pulmonary nodule; random forest

1. Introduction

Lung cancer is a disease that mainly affects the lungs, which is the principal organ of the respiratory system and performs vital activities for human survival. It usually starts once the epithelial cells of the bronchioles or alveoli grow abnormally. If the lungs do not carry out their functions properly, due to lung cancer, other parts of the body can also be affected, which can eventually lead to death. This fact makes the mortality rate of lung cancer higher than those of other types of cancer. The global cancer data by the World Health Organization (WHO) stated that lung cancer reached the most significant number of deaths, 1.8 million, which was 18.4% of the total cancer deaths in 2018 [1]. Effective ways to reduce the mortality of lung cancer include preventions, early detections, and precise treatments. However, most patients usually notice lung cancer at the advanced and uncurable stages because lung cancer is asymptomatic at the immature stage of the disease.

Importantly, since the advent of medical imaging technologies, computed tomography (CT) became an effective modality for the early diagnosis of lung cancer. According to the National Lung Screening Trial (NLST), the mortality rate of lung cancer reduced up to 15%–20% in patients who performed low-dose CT screening [2]. CT can provide detail morphologic information about lung cancer by detecting abnormal lesions inside the lungs, called pulmonary nodules. They are the round or oval-shaped opacities on the CT scans that have well-defined or irregular margins. Nodules are generally regarded as an early signal of lung cancer. They form due to the inflammation of pulmonary structures, and often react to infection and diseases. However, not all the nodules discovered in



the lungs are cancerous. They can be categorized into two types, namely, benign and malignant. Benign nodules are commonly found in almost all CT examinations, but they are not cancer. However, malignant nodules are cancerous and can even metastasize to other parts of the body. There are several clinical factors to distinguish benign and malignant pulmonary nodules, such as size, location, calcification, margin-types, and texture [3]. With the increasing number of lung cancer cases, manual detection of nodules has become a challenging task. It is effort-intensive and error-prone because the interpreter has to read the massive amount of CT scans slice-by-slice. Moreover, this task requires a high proficiency by the interpreters as an incorrect or miss-detection of the potential malignant nodule is highly-risky for patients. For these reasons, automated nodule detection schemes were strongly demanded as supportive tools. They were developed with aim of providing second options for better, more confident, and reliable diagnosis results. Additionally, they can make the interpretation process faster and more convenient by stacking the multiple sequential CT slices together as volumetric images.

Different research on automated nodule detections have emerged over the years. Generally, most of these works are based on two fundamental steps: (i) suspicious nodule detection and (ii) false reduction. The first step aims to extract the lesions that have a high possibility of being a nodule. In this step, the lung regions from the input CT scan are initially segmented. Afterwards, potential nodules scattered in the segmented lungs are extracted. The existing segmentation methods for this step can be categorized into four major groups: intensity-based, region-based, template-based, and machine-learning-based methods. Intensity-based segmentations are the fastest and simplest methods. They segment the region of interest (ROI) by grouping image pixels based on one or more predefined grey-value. Nodule detections using fixed thresholding [4,5], Otsu's thresholding [6], and optimal thresholding [7–9] are examples of the intensity-based methods. While they are easy to use, the effectiveness of these methods strongly depends on the predefined or initial threshold values. Moreover, they are less sensitive to noises and blur boundaries because they neglect spatial information about the image. Unlike the intensity-based methods, region-based segmentation schemes [10–13] consider not only spatial information, but also the relationship between the pixels. To segment the ROI, they identify an initial region and determine the homogeneity between neighbouring pixels. Regrettably, these methods require a proper choice of the initial region and homogeneity criteria for accurate segmentation. Indeed, it is difficult due to the fuzzy nature of the lungs and nodules. Next, several template-based segmentations, for example, genetic-algorithm-based template matching [14,15] and 3D template matching [16], create template models to detect the target position of the nodules. Nonetheless, template-based methods are computationally intensive and may have flaws for the affine variant of the nodules. With the recent emergence of machine learning techniques, segmentation using machine-learning-algorithms became famous for lung cancer diagnosis. Multi-class pixel-wise segmentation using deep neural networks, such as SegNet [17] and CNN [18], are examples of deep-learning-based segmentation schemes. Due to the use of multiple network layers, these segmentations are in favor of unique feature learning [19]. However, these methods usually consume a massive amount of data and training efforts.

Generally, nodules are solitary inside the lungs, but sometimes they may attach to the boundary of the lungs. These nodules are radiologically termed as juxta-pleural nodules. In most cases, such nodules are usually under-segmented since their intensity is substantially similar to the intensity of the lungs. For this reason, a follow-up task, called border reconstruction, is usually performed after the lung segmentation process. Morphological closing, which uses a structuring element (SE) [7,20,21], can help to reconstruct the boundary of the lungs. Similarly, a rolling ball algorithm [4,10], which uses a round-shaped SE, is also a kind of border reconstruction scheme. Instead of using a round-shaped SE, the border reconstruction scheme in [12] used a 45-degree rotated window. However, all these methods have weaknesses in adjusting the optimal size of the SE or window. As the size and shape of the nodules are fuzzy, a small-sized SE or window may cause under-segmentation; conversely, a big-sized SE may cause over-segmentation. Another popular border reconstruction method is the Freeman chain code [5]. This method traces the directional changes of the pixels along the boundary of the segmented

lungs. If concave regions along the border are detected, they are assumed to be the boundary-attached nodules and enclosing them will be attempted. The bidirectional chain code in [6] is an improved version of the original Freeman chain code. Unlike the original one, it traces the boundary pixels in both the horizontal and vertical directions using two code-word systems. Compared to other border reconstruction schemes, chain coding algorithms have more of a guarantee to detect variant sizes and shapes of the nodules because they do not depend on the SE or window. Nevertheless, they need extra effort to enclose the concave regions. For example, [6] used SVM to detect the critical point pairs in order to enclose the concave areas along the lung boundary. Since SVM is a supervised learning method, it needs historical data to train the model. As a negative consequence, it may suffer training related problems, such as overfitting and underfitting. An alternative border reconstruction method, known as contour-marching (CM), was proposed in [22]. It corrects the lung boundary by checking the texture features. If a region suspected to be a border-attached nodule is detected, CM encloses that region into the lungs using the region growing method. Indeed, only the texture features are insufficient to distinguish the border-attached nodules.

The second step, false reduction, aims to discriminate the nodules from a high number of suspicious structures. This task can be accomplished using a classification algorithm. For example, rule-based methods [23], k-nearest neighbour (KNN) [20] and support vector machine (SVM) [6–10,12,13], are some popular algorithms for false reduction. Similarly, the neural network-based classifiers, such as the artificial neural network (ANN) [11,21] and convolutional neural network (CNN) [24,25], are also found in this step. Besides, the ensemble classifiers like the integrated ANN [26] and random forest (RF) [27] are also applicable for false reduction. As there are several classification algorithms, it is difficult to pick the best one because the performance of each classifier depends on the segmentation methods and the dataset applied. However, the recent trend of automated lung nodule detection aims to achieve better performance. Moreover, a de facto nodule detection scheme which is fully automatic, increases sensitivity, reduces false positives, and ensures the detection of variant nodules, is still demanding for practical use.

Related Works

As automated nodule detection has been an active research area since 1980, numerous research papers have been published on this topic. Those papers proposed different algorithms using different datasets with different focuses. Among them, we studied some related research that presented the same objective and processes. As our experiments are based on the Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) [28–30], we reviewed some previous works using the same dataset. An extensive review of nodule detections, working on different datasets, can be referred to [31].

In 2015, Demir and Çamurcu [9] proposed computer-aided lung nodule detection using outer surface extraction. For lung region segmentation, they applied multiple thresholding, which uses two threshold values: maximum and minimum. Afterward, possible nodules were detected using labeling and rule-based methods. From each possible nodule, four different feature groups, including (i) morphological features, (ii) statistical and histogram features, (iii) outer surface statistical and histogram features, and (iv) outer surface textural features, are extracted. Based on these features, false reduction was conducted using a support vector machine (SVM), optimized by the particle swarm algorithm. Their experiments were performed using 200 CT exams of the LIDC-IDRI dataset. Their findings revealed that using outer surface features can provide a higher performance up to 98.03% sensitivity, 87.71% selectivity, 90.12% accuracy, and 2.45 false positives per scan. As a weakness, their work could not detect the nodules that had intensities outside the maximum and minimum threshold range. Moreover, they did not take into account the detection of nodules that were attached to the lung boundary.

In addition, Arindra et al. [24] also implemented novel nodule detection using multi-view convolutional networks (ConvNets). They initiated suspicious nodule detection by combining three

existing candidate detectors, which are designated for solid, sub-solid, and large solid nodules, respectively. Subsequently, multiple streams of 2D ConvNets were created using 2D patches of the detected nodule candidates. ConvNets can learn the discriminative features automatically from the training data in order to perform the false reduction. Their detection is evaluated using three datasets, and the highest performance (sensitivity 90.1% with four false positives per scan) has resulted in the LIDC-IDRI dataset. Even though ConvNets are powerful for feature learning and classification, they demand very high computation costs, especially the graphical processing unit (GPU).

Recently, Zhang et al. [13] also developed a CAD for nodule detection that applied 3D skeleton features. For lung region segmentation, they used the global optimal active contour model, which can enclose both solitary and boundary attached nodules. Then, the suspicious nodules scattered in the lungs are segmented using thresholding, morphological operations, and labeling methods. Ten features, including the 3D skeleton features, are extracted, and invalid nodule candidates are eliminated, based on prior anatomical knowledge. Finally, the real nodules are segregated by the SVM-based false reduction. Their CAD was also tested on 71 exams of LIDC-IDRI and reported a 89.3% sensitivity, 93.6% accuracy, and 2.1 false positives per subject. Nonetheless, their CAD did not describe the detection of small nodules surrounded by the vessels.

Moreover, Yuan et al. [25] also presented a lung nodule classification scheme, based on hybrid features. Instead of using a segmentation method to detect suspicious nodules, they created icosahedrons to sample the spherical structures. From these samples, they estimated nodule radii based on threshold values and created a 3D volume of nodule candidates. Then, convolutional neural networks (CNNs)-based statistical features and Fisher vector (FV)-based geometrical features were extracted from the volume candidates. These features were fused and fed into the multi-class support vector machine (SVM) for the classification of different nodule types. Their scheme is evaluated using two datasets, namely LIDC-IDRI and ELCAP. They achieved the classification rate of 93.1% on LIDC-IDRI and 93.9% on ELCAP, respectively. However, their nodule voxels selection was based on the radii threshold; thus, it is challenging to make it a scalable scheme for the detection of invariant and very tiny nodules.

The primary objective of this paper is to present a fully automated scheme that can detect various types of nodules providing a high accuracy. Unlike the previous works, we present three significant contributions. For the first contribution, we propose an automated seeded region growing algorithm for the segmentation of lung regions. It is an advance of the original region-growing method and can overcome the weakness in the conventional one, in particular, the needs of the proper initial region and homogeneity criteria. For the second contribution, we present a three-dimensional chain coding algorithm to detect the boundary-attached nodules. This algorithm traces the voxelized edge of the lungs to find the juxta-pleural nodules. As it is based on three-dimensional codeword systems, it can provide more specific information about the nodules, as compared to the two-dimensional methods. Moreover, it does not need an extra algorithm to find the start and endpoints of the boundary gap. As an advantage, we can eliminate the training related problems as well as saving on computational effort. For the third contribution, we develop an optimized random forest (RF) for false reduction. The original RF is a successful ensemble method for nodule detection, and we further improve its performance by combining the attribute selection algorithms in order to get more accurate diagnosis results. The rest of this paper is organized as follows. Section 2 describes the details of the materials and methods used in this work. Section 3 designs the proposed nodule detection and evaluates the experimental results. Lastly, Section 4 summarizes the paper by discussing the scope of future works.

2. Materials and Methods

2.1. LIDC-IDRI Dataset

Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) [28–30] is a well-known public dataset of thoracic CT scans for lung cancer. This dataset is comprised of

1018 exams, and each exam contains sequential, two-dimensional (2D) CT slices in DICOM format. Each slice has a 512×512 pixels dimension, and the number of slices per exam varied from 65 to 764. Moreover, each exam in the dataset contained an XML file that annotated the diagnosis results. These results are accurate because four experienced thoracic radiologists confirmed them through a two-phased reading process.

In our proposed nodule detection, we applied 888 exams of LIDC-IDRI because some exams contained missing and inconsistent slices; thus, we excluded such exams from our experiments. We applied the annotation files as the ground truths to train and evaluate our method. Instead of detecting the pulmonary nodules directly from the 2D scans, we reconstructed three-dimensional volumetric images by stacking all the 2D slices in a single exam, i.e., each exam gave a volume image, and it was used as an input for our nodule detection.

2.2. Methods

Our proposed nodule detection has two fundamental steps: (i) suspicious nodules detection and (ii) false reduction, as illustrated in Figure 1. The first step is conducted using three methods, namely, (i) automated seeded region growing, (ii) three-directional chain code, and (iii) preliminary screening. In the second step, there are two methods: (i) 2D and 3D feature extraction and (ii) an improved random forest classifier. We discuss the technical details of these five methods in the following sub-sections.



Figure 1. Processing steps and methods used in the proposed nodule detection.

2.2.1. Automated Seeded Region Growing

The automated seeded region growing is an enhanced version of the original seeded region growing (SRG) algorithm [32]. SRG is famous for segmentation due to its rapid, robust, and easy-to-use

procedure. Unlike other region-based algorithms, SRG does not require homogeneity criterion to determine the similarity between the voxels. Instead, it starts the segmentation by assigning sets of voxels called the initial seed points, where each set represents a separate region in the image. In each step of the algorithm, SRG assigns an unlabelled voxel $v_{x,y,z}$ into one of the initial regions. The algorithm terminates when all the voxels in the image are labelled into different regions. For example, assume a volume image *I* has initial seed regions (S_1 , S_2 , S_3 , ..., S_n) and a set of unlabelled voxels *V*. Then, we want to label each element of *V* to each S_i , where i = 1, 2, 3, ..., n.

$$S_1, S_2, S_3, \dots S_n \subset I, V \in I$$
 (1)

$$V = \left\{ v_{x,y,z} \notin \bigcup_{i=1}^{n} S_i \mid N(v_{x,y,z}) \cap \bigcup_{i=1}^{n} S_i \neq \varnothing \right\},$$
(2)

where $N(v_{x,y,z})$ is the 26 connected neighbors of $v_{x,y,z}$. After the *m* step of the algorithm, $N(v_{x,y,z})$ meets just one of the initial seed regions S_i , and $N(p_{x,y}) \cap S_i(p_{x,y}) \neq \emptyset$. Then, we calculate the difference grey value between $v_{x,y,z}$ and its adjacent region by using the following equation.

$$\partial \left(v_{x,y,z} \right) = \left| g \left(v_{x,y,z} \right) - mean_{r \in S_i} \left[g(r) \right] \right|, \tag{3}$$

where $g(v_{x,y,z})$ is the grey value of the unlabelled pixel, and the latter part of the equation is the mean grey value of all the voxels in the adjacent region *r*. If $v_{x,y,z}$ is located on the boundary of two or more regions, we assign $v_{x,y,z}$ into the region that has the minimum ∂ value.

$$\partial(z) = \min_{v_{x,y,z} \in P} \left(\partial(v_{x,y,z}) \right). \tag{4}$$

In this way, SRG allocates all the unlabelled voxels into separate regions. SRG is easy to use, even for the unskilled user who does not have detailed knowledge about the input image. Nevertheless, the segmentation result of SRG strongly depends on the initial seed regions. If the seed regions fall on the noisy or outlier voxels, it will result in poor segmentation. Moreover, user-oriented seed selection of SRG is also a flaw for fully automated detection.

Therefore, we propose an enhanced version of SRG, which can identify the initial seed points automatically. Seed point selection is a vital task for SRG because correct segmentation results are strongly correlated to the seed points. Each seed point should represent a typical grey value of its region, but it should not be a noise or an outlier voxel. Fortunately, on a CT scan, specific organs exhibit a range of grey values, depending on the radiographic density, which is measured by the Hounsfield unit (HU). Based on this fact, we applied the HU value of the input CT scan for automatic seed point selection.

As one instance, Figure 2a illustrates a histogram that shows the distribution of HU values in the input CT volume. In this histogram, HU values are ranging from –1024 to 2513. Then, we can calculate the mean HU value of the image to find the seed points. The mean HU of the CT volume in this example is 147.78. Based on this mean HU value, we can assign the seed points automatically. In our research, we identify two initial seed points. The first seed point is selected by finding the very first voxel, which is less than the mean HU value. In Figure 2b, the first seed point is denoted by a blue square, and it represents black regions (low-density regions). Likewise, the second seed point is set as the very first voxel, which is higher than the mean HU (denoted by a red square in Figure 2b), and it represents the grey region (high-density regions). Based on these two seed points, the segmentation of the lungs can be performed by following the steps for SRG.



Figure 2. Automatic seed points selection using HU values of input CT volume (**a**) Histogram showing the distribution of HU values (**b**) two seed points selected for seeded region growing.

2.2.2. Three Dimensional Chain Coding

The segmentation of the lung regions usually fails when the nodules adhere to the boundary of the lungs. Such nodules are known as the juxta-pleural nodules and are exposed the intensities that are similar to the lung wall. Thus, during segmentation, they are usually excluded, together with the lung wall, and this produces an under-segmentation result. As an instance, Figure 3a demonstrates a 2D CT slice containing a juxta-pleural nodule (denoted by a red rectangle), and Figure 3b shows the segmentation result of the lungs in 3D, applying the automated SRG. From this image, we can notice that the juxta-pleural nodule is under-segmented (denoted by a red rectangle).



Figure 3. Unsuccessful segmentation of a Juxta-pleural nodule. (**a**) A two-dimensional computer tomography (CT) scan with a juxta-pleural nodule, and (**b**) an under-segmentation of the juxta-pleural nodule in a three-dimensional view.

Chain coding [4,5] is the most frequently used algorithm to solve the problem of juxta-pleural nodule segmentation. It traces the pixels/voxels along the contour of the lungs and determines whether there is a discrete path or not. If a discrete region is found, it is assumed to be the place of

the juxta-pleural nodule and it is enclosed by connecting the lung border. The bi-directional chain coding method in [6] is an advancement to the original chain coding in [5]. It uses two-dimensional coordinate systems to trace the boundary pixels in both the horizontal and vertical directions. It works well for juxta-pleural nodule segmentation, but it needs to operate on each 2D slice of the input CT exams. As a negative consequence, it consumes a lot of processing time and effort. Additionally, it needs a supplementary algorithm, for example, the support vector machine (SVM) [6], to determine and connect the start-point and end-point of the discrete path.

Unlike the chain codes in [4,5], we propose a three-dimensional chain code algorithm [33,34] for juxta-pleural nodule detection. The main difference is that the chain coding in [4,5] works in 2D segments of lung regions, but our proposed 3D chain coding works in the 3D segments of lung volume. Firstly, we extract the voxelized edge *V*, i.e., a single voxel outline, of the segmented lungs. On that boundary or edge, *V*, suppose there are *N* number of voxels. V(i) is the *i*th voxel of the boundary where $i = \{1, 2, ..., N\}$. Then, Freeman's F26 code word system [33] is applied to trace the directional changes between V(i) and its adjacent voxel V(i+1).

As illustrated in Figure 4, F26 views a boundary voxel as a cubic lattice, and there are three types of adjacencies between neighborhood voxels. A cube is comprised of six faces, twelve edges, and eight vertices; thus, there are twenty-six possible directions. The associated codeword values of these directions are also described in Figure 4, using a to z characters.



Figure 4. Codeword values for 26 possible directions of the three dimensional chain code [30].

The codeword values are assigned, depending on the direction of the movement between a voxel to its adjacent one. A movement between two voxels is defined as a set of codeword values c(i), and each c(i) contains three elements that represent a face, edge and vertex connection, such that $c(i) \in \{0, 1, -1\}$. For a visual explanation, Figure 5 demonstrates a path of the voxelized edge, and its associated codeword values can be assigned based on the adjacencies and directions between voxels along that path. The path contains seven voxels; thus, there are six directional movements between its voxels. Then, the associated codeword values of these movements can be defined by six sets, where each set contains three elements, as described below.



Code word values = <1, 0, 0>, <1, 0, 1>, <1, 1, 0>, <1, 0, 1>, <0,-1,-1>, <1,0,-1>

Figure 5. Example of assigning codeword values, based on the F26 chain code.

After assigning the codeword values for all voxels along the boundary, we need to detect the points that have convexity changes. These points are called the inflection points (denoted by the red stars in Figure 6b. They can be detected by calculating the differential operations of the codeword values.

$$f(i) \leftarrow i \ if \ \nabla c(i) \neq 0, \tag{5}$$

where *i* is the *i*th voxel of the lung boundary and c(i) is the corresponding codeword value of that voxel. The non-zero points after the differential operation of c(i) can then be defined as inflection points f(i). Once inflection points are detected, we calculate and fill the volume of the gap surrounded by the inflection points, as demonstrated in Figure 6. Unlike the hole filling using SE [4,7,10,12,20,21], our border reconstruction using 3D chain coding can accurately fill the volume of the gap, i.e., the exact region of juxta-pleural nodules is filled without resulting in under- or over-filling.



Figure 6. Border reconstruction. (**a**) Under-segmented juxta-pleural nodule, (**b**) inflection points along the boundary voxels, and (**c**) segmented lung after boundary reconstruction.

2.2.3. Preliminary Screening

Preliminary screening is a strategy to remove potential false positives (FPs) from the image. Generally, the automatic detection of nodules produces a relatively high number of non-nodule lesions, called FPs, because most of the lung structures such as airways, bronchi, and blood vessels possess a grey level that is similar to that of the real nodules. As a negative impact, these structures are usually misclassified as nodules and results in a decline in classification accuracy. Additionally, being prone to FPs can also increase the computational workload. Therefore, it is crucial to reduce FPs before the classification process.

In this paper, we initially performed simple thresholding to segment all the structures inside the lungs, such as nodules, airways, blood vessels, and fissures. These structures expose the attenuation ranging from -910 HU to -500 HU [35] on the CT scans. We choose the low-density value of -910 HU as a threshold and segment all the objects higher than that value. This helps to ensure the detection of nodules with ground glass opacities that have very faint intensities.

As an illustration, Figure 7a depicts the volumetric structure of the segmented lung regions using the automated seeded region growing and 3D chain-coding methods. Figure 7b describes the initial segmented lung structures using thresholding, and it contains a high number of FPs. To reduce these FPs, we initially applied 3D connected component labeling [36] and analyzed the shape of each labeled object. Generally, nodules are circular, while other lesions, especially vessels and fissures, are tubular in their structure. Based on this fact, we calculated three criteria, namely, surface area, eccentricity, and voxel remove rate (VRR) [13] in order to check the shape of each candidate object.



Figure 7. Preliminary screening. (**a**) Volumetric structure of segmented lungs, (**b**) roughly segmented lung structures, and (**c**) candidate nodules after removing FPs by three criteria.

Let *l* be a labeled object of a roughly segmented lung structure, and its surface area *A* can be calculated by

$$A(l) = \sum l(x, y, z), \tag{6}$$

where l(x, y, z) is a voxel value of l at the three-dimensional coordinate point (x, y, z). By checking the surface area, we can exclude huge and tiny objects that show a high probability of being non-nodules. We empirically set two threshold values, T1 = 10 and T2 = 3700, to eliminate the FPs outside that range.

$$FP = l \begin{cases} if A(l) < T1\\ if A(l) > T2 \end{cases}$$

$$\tag{7}$$

Similarly, we calculated the eccentricity *Ecc* to check the roundness of each labeled object. The *Ecc* value ranged from 0 (a circular object) to 1 (a tubular object), and we empirically applied the *Ecc* threshold as 0.9 to remove the tubular objects.

$$Ecc (l) = \frac{majoraxis (l)}{minoraxis (l)}$$
(8)

$$FP = l \quad if \ Ecc(l) > 0.9 \tag{9}$$

Then, another criterion *VRR* was also calculated, based on the skeleton structure of the labeled object. Skeletonization gives a topological structure by transforming the labeled object into a set in the interconnected single-voxel centerline. By calculating the *VRR*, we could exclude the elongated skeletal structures, such as the vessels and bronchioles.

$$VRR(l) = 1 - \frac{A(l)}{A(ls)}$$
(10)

where *ls* is the skeleton of *l*. Then, we empirically set the *VRR* threshold as 0.9 to remove the vessels and bronchioles.

$$FP = l \quad if \ VRR(l) > 0.9 \tag{11}$$

After a preliminary screening using these three criteria, the FPs could be eliminated, as illustrated in Figure 7c. However, it could not guarantee the removal of all FPs because the nodules might have attached to the adjacent pulmonary structures, particularly the vessels, in some cases. This type of nodule is called a juxta-vascular nodule, and it may be excluded, together with the tubular structures during preliminary screening. To prevent this, we set very high thresholds (0.9) for *Ecc* and *VRR*. Nonetheless, it was not a perfect way to detect all the vessel-attached nodules. A more effective way would be to cut the nodules from the attached vessels.

Fortunately, the nodules were in blob shapes in a three-dimensional structure, and hence, a Hessian-based blob enhancement filter could help to discriminate them from the vessels. Let *V* be the volumetric image, and *H* be its Hessian matrix, which can be calculated by the second-order partial derivatives of the voxels in *V*.

$$H\left(V\right) = \begin{bmatrix} \frac{\partial^2 V}{\partial x^2} & \frac{\partial^2 V}{\partial x \partial y} & \frac{\partial^2 V}{\partial x \partial z} \\ \frac{\partial^2 V}{\partial y \partial x} & \frac{\partial^2 V}{\partial y^2} & \frac{\partial^2 V}{\partial y \partial z} \\ \frac{\partial^2 V}{\partial z \partial x} & \frac{\partial^2 V}{\partial z \partial y} & \frac{\partial^2 V}{\partial z^2} \end{bmatrix}$$
(12)

$$\nabla \left(H - I * \lambda\right) = 0,\tag{13}$$

where *I* is the identity matrix and $\lambda = {\lambda_1, \lambda_2, \lambda_3}$ is the associated eigenvalues of *H*. These three eigenvalues can provide information about the shape of the objects as follows:

$$V = Blob \ if \ \lambda_1 < 0, \ \lambda_2 < 0, \ \lambda_3 < 0 \tag{14}$$

$$V = Tube \ if \ \lambda_2 < 0 \ , \ \lambda_3 < 0 \tag{15}$$

$$V = Plate \ if \ \lambda_3 < 0. \tag{16}$$

Depending on the shape that we want to enhance, different mathematical computations can be done using these three eigenvalues. In our proposed method, we intend to detect pulmonary nodules. Hence, we enhance the blob-like structures and suppress tube-like objects using the following equations [37].

$$Blob = \lambda_1^2 \lambda_p \left[\frac{3}{2\lambda_1 + \lambda_p} \right]^3 \tag{17}$$

$$\lambda_p = \begin{cases} \lambda_3 & \text{if } \lambda_3 < T * \min_x \lambda_3(x,s), \\ T * \min_x \lambda_3(x,s) & \text{otherwise,} \end{cases}$$
(18)

where *T* is the cutoff threshold value (between 0 and 1); x is the coordinate points of the *V* and *s* is the gaussian scales. Using these equations, we can discriminate the juxta-vascular nodules from the attached vessel, as illustrated in Figure 8.



Figure 8. Juxta-vascular nodule detection. (**a**) Juxta-vascular nodule in 2D structure, (**b**) juxta-vascular nodule in the volumetric structure, and (**c**) segmented juxta-vascular nodule using the Hessian method.

2.2.4. Feature Extraction

The features are the fundamental factors to determine the nodules and non-nodules. Thus, they must contain sufficient and relevant information in order to produce correct diagnosis results. In clinical practice, radiologists manually detect and classify the nodules, depending on several radiological criteria such as size, shape, margin, calcification, contrast, and density. Based on these characteristics, we extract image features from each suspicious nodule. We assume that the features such as size, shape, margin, and density are associated with the geometric image features. Similarly, the calcification and contrast are associated with the intensity and contrast feature of the image.

We extracted 25 features in total from both 2D and 3D structures of the nodule candidates. The extracted features and their mathematical calculations can be seen in Table 1. In the formulae, N represents a suspicious nodule in a 2D or 3D structure. For 2D feature extraction, the median slice of the suspicious nodule is used because it has the maximum number of pixels, as compared to the other slices. Assume that the coordinate points are only x and y in 2D and x, y and z in 3D. dx, dy, and dz represent the diameters of N; G_N represents the grey value of N and n represents the total number of pixels (voxels) in N.

Features	Name	Formula				
	Surface area	surface area $= \sum N(x, y, z)$				
-	Perimeter	$perimeter = \sum edge(N)$				
Geometric features (2D,3D)	Effective radius	$radius (2D) = perimeter * \sqrt[2]{\frac{Surface area}{\pi}}$ $radius (3D) = perimeter * \sqrt[3]{\frac{3*surface area}{4\pi}}$				
-	Centroid	centroid = mean (N(x, y, z))				
-	Minor axis length	minor axis length = min $(N(dx, dy, dz))$				
-	Minor axis length	major axis length = max $(N(dx, dy, dz))$				
_	Elongation	$elongation = rac{minor \ axis \ length}{major \ axis \ length}$				
Geometric	Volume	volume = surface area * xspacing * yspacing * zspacing				
features (3D)	Sphericity	$sphericity = rac{volume}{surface area}$				
-	Compactness	$compactness = \frac{volume}{number of elements(N)}$				
Intensity and	Mean	$mean = \frac{1}{n} \sum_{(x,y,x) \in \mathbf{N}} G_{\mathbf{N}}(x,y,z)$				
contrast features (2D, 3D)	Standard deviation	$\sigma = \sqrt{\frac{1}{n} \sum_{(x, y, z) \in \mathbb{N}} [G_N(x, y, z) - mean]^2}$				
_	Skewness	skewness = $\frac{\sum_{(x, y, z) \in \mathbb{N}} [G_N(x, y, z) - mean]^3}{n\sigma^3}$				
-	Kurtosis	$kurtosis = \frac{\sum_{(x, y, z) \in \mathbb{N}} [G_N(x, y, z) - mean]^4}{no^4}$				

Table 1. Features extracted and the mathematical formulae.

2.2.5. Improved Random Forest

Several authors have proved that the random forest (RF) [38] is a significant machine learning algorithm for the classification of pulmonary nodules. RF is computationally effective in solving problems with noisy data and overfitting. It is a combination of multiple decision trees $h_1(x)$, $h_2(x)$, ... $h_K(x)$, where x is a group of input samples, and K is the total number of trees in the forest. Each tree $h_i(x)$, $i = \{1, 2, ..., K\}$ works in parallel and produces its own prediction p_i independently. The final prediction result of RF is decided by major voting and finding the winning class among p_i . Individual tree $h_i(x)$ is grown on a random subset of samples R_i , $i = \{1, 2, ..., K\}$ from the training data T with replacements. The number of sample n in each random subset R_i is the same. RF does not use all the samples n in R_i for the training of individual tree. It only uses 36.8% of n for training and the rest are used for internal validation.

Suppose $R_i(x)$ is a random subset that has a group of samples x. Each sample in x has attributes A_i , $i = \{1, 2, ..., a\}$, where a is the total number of attributes. However, RF does not use all the attributes A_i to grow the trees. Instead, it selects random attributes A_r by calculating the Gini values of A_i .

$$Gini (A_i) = -\sum_{i=1}^{c} p(y_i)^2 + \sum_{j=1}^{m_i} p(v_{i,j}) \sum_{i=1}^{c} p(y_i | v_{i,j})^2$$
(19)

where *c* is the total number of classes and $p(y_i)$ is the probability of the class y_i . Each attribute A_i has values v_j where m_i is the number of values and $p(v_{i,j})$ is the probability that the attribute A_i has the value v_j . Then, $p(y_i | v_{i,j})$ is the probability of the class y_i conditioned by the attribute A_i having value v_j .

The Gini calculation is easy and fast because it evaluates the attributes separately without considering the conditional dependencies and interactions with other attributes. As a negative impact,

the performance of the RF may degrade when the attributes are strongly dependent on each other. Moreover, using random attributes to grow each tree also makes the individual trees weaker because random selection cannot be guaranteed to obtain the relevant and non-redundant attributes [39].

For this reasons, we aim to improve the conventional RF by combining it with attribute selection algorithms. Instead of selecting random attributes by calculating the Gini value, we select optimal ones to grow each decision tree. We apply four state-of-the-art attribute selection algorithms, namely RELIEFF [40], genetic algorithm (GA) [41], particle swarm optimization (PSO) [42], and firefly algorithm (FA) [43], to measure the quality of the attributes. Put simply, we use these four feature selection algorithms as a pre-processing step before growing individual trees. Each algorithm estimates the quality of the attributes and selects *r* number of attributes that have the highest quality. The decision trees in the RF are then grown using these selected attributes. Figure 9 demonstrates the architecture of our optimized RF, with *i* decision trees ($i = 1, 2, 3 \dots, K$). Unlike the random attribute selection of the original RF, the attribute selection algorithms in our optimized RF take into account the conditional dependencies between the attributes.



Figure 9. The architecture of the optimized random forest (RF).

3. Results and Discussions

To design and evaluate the performance of our proposed nodule detection, we randomly divided the input materials (888 CT exams of LIDC-IDRI) into k subsets where k = 5 and performed k-fold cross-validation. Unlike other validation methods that only use one training and one testing subset, k-fold cross-validation makes our proposed nodule detection more robust. Among k subsets of samples, it uses a single subset for testing and the remaining k-1 subsets for training. In this way, cross-validation iterates the training and testing processes k times so that each of the k subsets have appeared as a test set exactly once in order to validate the performance of the proposed method. Then, the final assessment of cross-validation can be obtained by calculating the average. This idea helps to provide more accurate assessments on the predictability of the proposed nodule detection on unknown samples. Moreover, k-fold cross-validation helps our proposed nodule detection prevent overfitting.

After preparing data for cross-validation, we try to set the hyperparameters of our proposed RF. As the performance of RF strongly depends on the hyperparameters used, selecting appropriate values of the hyperparameters is very important. Hyperparameters are the values that are predefined before running the model. Generally, these values are manually set by the developer because they cannot be estimated from the data. The common hyperparameters that are applied in RF are (i) the number of decision trees in the forest, (ii) the maximum number of features selected to develop an individual tree, (iii) the maximum number of depths of an individual tree, (iv) the minimum number of samples required to split an internal node, and (v) the minimum number of samples at the leaf node. Proper adjustment of these hyperparameter values can help the RF's prediction results be more reliable and robust.

In this experiment, we define the optimal hyperparameter values using the "trials and errors." We initially set a possible range of values that correspond to a specific hyperparameter and perform cross-validation. The accuracy score generated by RF, using each value of the hyperparameter, is measured, and the value that can provide the highest accuracy is picked as an optimal hyperparameter.

For the first hyperparameter, in regards to the number of decision trees used in RF, we define a range from 100 to 500. Then, the accuracy scores, using each number of decision trees from the range, is calculated. Subsequently, a validation curve is created to show and compare the accuracy scores that are produced using the different number of decision trees, as illustrated in Figure 10a. From this curve, we can see that the highest accuracy value (0.9304) is reached using (230) decision trees. Hence, we choose (230) as an optimal hyperparameter value for the number of decision trees in RF.

In general, most RFs set the second hyperparameter, which is the maximum number of features, as the square root of the total number of features. However, in our research, we tested using a range from 1 to 25 (total number of extracted features) and chose 15 as a hyperparameter because it could ensure the highest accuracy (0.9232), as demonstrated in Figure 10b. As well as, the third hyperparameter, the maximum number of depths of an individual tree is selected by finding the highest accuracy using a range from 2 to 20. As we can see in Figure 10c, the FR reached the highest accuracy (0.9285) at the depths (18); thus, we selected that number as the third hypermeter.

In the same way, the next hyperparameter, the minimum number of samples to split an internal leaf node is selected as (5) by testing a range from 2 to 20, as depicted in Figure 10d. Finally, the last hyperparameter, the minimum number of samples at the leaf node, is chosen as (2) as it could give the highest accuracy score (0.9178), as shown in Figure 10e.



Figure 10. Cont.



Figure 10. Adjusting the hyperparameter values for RF. (a) Number of decision trees in the forest, (b) maximum number of features selected for each tree, (c) maximum number of depths of an individual tree, (d) minimum number of samples required to split an internal node, and (e) minimum number of samples at the leaf node.

To summarize the hyperparameter adjustments, Table 2 lists the optimal values of each hyperparameter used in our proposed RF. All of these hyperparameters are selected using an original version of RF before combining any attribute selection algorithm.

Hyperparameter	Range	Accuracy	Selected Value	
Number of decision trees	100 to 500	0.9304	230	
Maximum number of features	1 to 25	0.9232	15	
Maximum number of depth	2 to 20	0.9285	18	
Minimum number of samples to split	2 to 20	0.9267	5	
Minimum number of samples at the leaf node	1 to 5	0.9178	2	

Table 2. Hyperparameter adjusting for RF.

Using these hyperparameter values, we developed the five RFs (original and four improved RFs). All of these RFs are implemented using the same setting of parameters and evaluated by k-fold cross-validation. In each iteration of k, we assessed the following statistical measurements and calculated the average for the final evaluation results.

$$Accuracy (Acc) = \frac{TP + TN}{TP + TN + FP + FN}$$
(20)

$$Sensitivity (Sen) = \frac{TP}{TP + FN}$$
(21)

$$Fasle \ positive \ (FP) = \frac{FP}{Number \ of \ exams}$$
(23)

$$F1 \ score \ (F1) = \frac{2TP}{2TP + FP + FN}$$
(24)

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

In addition to these measurements, we also provide a graphical method called the receiver operating characteristics (ROC) curve to evaluate the performance of the proposed nodule detection quickly. ROC is a two-dimensional graph showing the false-positive rate (FPR) (1-Specificity) in the X-coordinate and the true-positive rate (TPR) (Sensitivity) in the Y-coordinate [44]. The continuous outputs of the RFs, i.e., the scores that represent the probability of a sample belonging to a specific class, are applied to generate the ROC curve. Firstly, the samples are sorted depending on the descending order of their corresponding scores. Then, a threshold value is set as the maximum (α), and it decreases until the minimum (- α). In every decrease, the threshold value is compared to the prediction scores. If the score of a sample is equal to or higher than the threshold, that sample can be classified as a positive class (1); otherwise, it is defined as a negative class (0).

At the maximum threshold value (α), the scores of all samples are less than α ; therefore, all the samples are classified as negative classes. As a result, the (FPR, TPR) at that threshold is (0,0), and we plot it on the graph. Inversely, if the threshold value is minimum (- α), then the coordinate points of (FPR, TPR) reach (1,1), meaning that all the samples are classified as positives. In this way, the ROC curve is generated by plotting (FPR, TPR) for every threshold value.

Figure 11 demonstrates an example of ROC plotting based on different threshold values. Suppose we have ten samples of two classes (five for positives and the rest for negatives), and they are arranged according to their prediction scores. At first, the threshold value is set as α and calculates the FPR and TPR. Then, the threshold value is reduced until it equals the score of the second sample, and (FPR, TPR) is recomputed by comparing that threshold with the scores. In this way, the threshold is decreased, and (FPR, TPR) is computed for all the samples. Finally, we can generate the ROC curve by plotting the resulting (FPR, TPR) values on a 2D coordinate system (as shown in the right side of Figure 11). The ROC in the figure exposes a step function because only ten samples are tested in this example. The real ROC curves in practice are smoother and are more complex, as compared to the curve in this example [44]. Based on this ROC, we can know that the highest performance of the classification model in this example is reached at the point (0.2,0.8), which is the minimum FPR and maximum TPR. If a ROC curve reaches the point (0,1), it means that the model produces an ideal performance because it can correctly predict all positive and negative samples.



Figure 11. Example of threshold setting and receiver operating characteristics (ROC) plotting.

Our experiment uses five classification models (original RF and four improved RFs), and each iterates the training and testing process five times for five-fold cross-validations. Therefore, in

every iteration, we generate an ROC of each RF. After that, the mean ROC curves are generated by interpolating the single ROCs, as illustrated in Figure 12a–e. Finally, the comparison of the mean ROCs of five RFs is shown in Figure 12f.



Figure 12. Comparison of ROC curves for five RFs. (**a**) Original RF, (**b**) improved RF using RELIEFF attribute selection algorithm, (**c**) improved RF using genetic algorithm (GA) attribute selection algorithm, (**d**) improved RF using particle swarm optimization (PSO) attribute selection algorithm, (**e**) improved RF using firefly (FA) attribute selection algorithm, and (**f**) comparison of the mean ROCs of five RFs with respected to five-fold cross-validation.

However, comparing the ROC curves directly from the graph is very difficult as they are not scalar values. For this reason, we add another assessment measurement called the area under the curve (AUC) to ascertain the specific differences between the ROC curves. As the name implies, AUC can be measured by calculating the area of the trapezoid region under the ROC curve. Table 3 summarizes all the assessment values of each RF with respect to the five-fold cross-validation. From this table, we can know that the optimized RF that applies the firefly attribute selection (RF+FA) can provide superior performance, as compared to its counterparts.

	Assessments	Five-fold Cross-Validation						
Classifiers		Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	Mean	
Original RF	Accuracy	0.8986	0.9128	0.8899	0.9263	0.9037	0.9063	
	Sensitivity	0.8991	0.9358	0.9174	0.9444	0.8899	0.9173	
	Specificity	0.8981	0.8899	0.8624	0.9083	0.9174	0.8952	
	False Positive	0.1019	0.1101	0.1376	0.0917	0.0826	0.1048	
	F1-score	0.8991	0.9148	0.8929	0.9273	0.9023	0.9073	
	AUC	0.9413	0.9279	0.9249	0.9529	0.9383	0.9371	
	Accuracy	0.9124	0.9174	0.9171	0.9358	0.9220	0.9209	
	Sensitivity	0.9352	0.9358	0.9174	0.9541	0.9358	0.9357	
RELIEFF +	Specificity	0.8899	0.8991	0.9167	0.9174	0.9083	0.9063	
RF	False Positive	0.1101	0.1009	0.0833	0.0826	0.0917	0.0937	
	F1-score	0.9140	0.9189	0.9174	0.9369	0.9231	0.9221	
	AUC	0.9760	0.9632	0.9805	0.9786	0.9822	0.9761	
	Accuracy	0.9174	0.9174	0.9124	0.9263	0.8945	0.9136	
	Sensitivity	0.9725	0.9725	0.9444	0.9358	0.9174	0.9485	
CA + PE	Specificity	0.8624	0.8624	0.8807	0.9167	0.8716	0.8787	
GA + Kr	False Positive	0.1376	0.1376	0.1193	0.0833	0.1284	0.1213	
	F1-score	0.9217	0.9217	0.9148	0.9273	0.8969	0.9165	
	AUC	0.9646	0.9690	0.9822	0.9761	0.9729	0.9730	
	Accuracy	0.8853	0.9312	0.9174	0.8986	0.9217	0.9108	
	Sensitivity	0.9266	0.9633	0.9633	0.9266	0.9259	0.9411	
$PSO \pm PE$	Specificity	0.8440	0.8991	0.8716	0.8704	0.9174	0.8805	
130 + Kr	False Positive	0.1560	0.1009	0.1284	0.1296	0.0826	0.1195	
	F1-score	0.8899	0.9333	0.9211	0.9018	0.9217	0.9135	
	AUC	0.9532	0.9736	0.9697	0.9397	0.9718	0.9616	
FA + RF	Accuracy	0.9078	0.9539	0.9266	0.9404	0.9266	0.9311	
	Sensitivity	0.9630	0.9633	0.9174	0.9817	0.9174	0.9486	
	Specificity	0.8532	0.9444	0.9358	0.8991	0.9358	0.9137	
	False Positive	0.1468	0.0556	0.0642	0.1009	0.0642	0.0863	
	F1-score	0.9123	0.9545	0.9259	0.9427	0.9259	0.9323	
	AUC	0.9862	0.9859	0.9838	0.9826	0.9649	0.9807	

Table 3. Performance assessment measurements of original RF and four improved RFs with respect to five-fold cross-validation.

Some example of outputs of our proposed nodule detection are demonstrated in Figure 13. Figure 13a shows the detection of a sharp solid nodule that resides inside the lung without touching other lesions. Such nodules are easy to detect, and our proposed method well-detected them. However, Figure 13b shows the results of an isolated ground glass nodule (GGO) that had a very faint intensity. This type of nodules is hard to detect due to their hazy structures. However, the proposed method can successfully detect them because we used a very low-intensity value (–910 HU) during nodule candidate detection. Moreover, our method ensures the detection of juxta-pleural nodules that are connected to the lung boundary. Figure 13c represents the detection of a small juxta-pleural nodule, and Figure 13d represents the detection of a very big one, respectively. Our method can ensure

the detection of juxta-pleural nodules, regardless of the size, which is beneficial to the use of the 3D chain coding method for border reconstruction.



(e)

Figure 13. Results of the proposed nodule detection. (**a**) Solid isolated nodule, (**b**) isolated ground glass nodule, (**c**) small juxta-pleural nodule, (**d**) big juxta-pleural nodule, and (**e**) juxta-vascular nodule.

Lastly, Figure 13e shows the detection of a juxta-vascular nodule, which is also challenging to detect due to nearby pulmonary structures. Our method can successfully detect them with the help of the blob enhancement filter in preliminary screening. Therefore, it is evident that our proposed method can detect any variants of nodules.

Moreover, we compute execution time for our proposed nodule detection because it is an essential factor in evaluating the performance of an algorithm. Table 4 describes the profile summary, which shows the time complexity for each function to detect the pulmonary nodules from a single exam with a maximum number of slices. From this table, we can see that our automated nodule detection takes 144.4 seconds (2.4 minutes) in maximum to predict a single CT exam on a standard processor with Intel(R) Core (TM) i7-6500U CPU @2.6 GHz.

Function Name	Total Time
Reading input images (Maximum: 764 slices)	1.53 s
Lung regions segmentation (Automated seeded region growing)	11.49 s
Border reconstruction (3D chain coding)	99.68 s
Preliminary screening	18.56 s
Feature extraction	10.40 s
False reduction (RF+FA)	2.74 s
Entire program	144.40 s

Table 4. The total execution time of our proposed nodule detection.

Lastly, we conduct the performance comparison of the proposed nodule detection with some of the state-of-the-art techniques discussed in Section 2.1. While these techniques worked on the same dataset (LIDC-IDRI), they used a different number of samples and different performance assessment measurements. For a fair comparison, we quantitatively compared each method based on the same assessment measurements, as demonstrated in Table 5. From the table, we can analyze that the techniques in [13] and [25] achieved higher and the same accuracy values, compared to our proposed method, but their performances were assessed using fewer samples. Moreover, the performance of [24] is also quite high, using the same number of samples. However, [24] applied the CNN, which consumes a high computation cost and is applicable in some specialized computers that have good GPUs. Unlike their methods, our proposed method applied the improved random forest by firefly (RF + FA), which is preferable and applicable in any standard computers. Moreover, the proposed automated nodule detection is promising for the detection of variant size and type of nodules in a high detection accuracy.

Table 5. Performance comparison of proposed nodule detection with previous methods.

References	Method	Exams	Acc (%)	Sen (%)	Spec (%)	FP (per exam)
[9]	 Multiple thresholding Labeling Rule-based SVM (optimized by particle swarm) 	200	90.12	98.03	87.71	2.45

References	Method	Exams	Acc (%)	Sen (%)	Spec (%)	FP (per exam)
[13]	 Global optimal active contour, Thresholding Morphological operations Labeling SVM 	71	93.6	89.3	-	2.1
[24]	Candidate detectorsCNN	888	-	90.1	-	4
[25]	 Icosahedron sampling Thresholding based on nodule radii CNN + FV features SVM 	744	93.1	-	-	-
Proposed Method	 Automated seeded region growing 3D chain code Preliminary screening RF + FA 	888	93.11	94.86	91.37	0.0863

Table 5. Cont.

4. Conclusions

In this paper, we presented a fully automated nodule detection for lung cancer diagnosis. Firstly, we proposed an automated seeded region growing (SRG) to segment the lungs from the input CT volume images. The strength of the proposed SRG is that it can select initial seed regions automatically and perform the segmentation without requiring homogeneity criteria to determine the similarity between the pixels. Secondly, we designed a 3D chain code algorithm for the detection of juxta-pleural nodules. It traces the voxelized boundary of the segmented lungs to identify the concave regions along the border and enclose them by detecting the inflection points. Compared to the 2D border reconstruction methods, the use of 3D chain coding takes less effort and provides more accurate results. Moreover, it does not require training data or other classification algorithms to find the start-point and end-point of the nodule gap. Finally, we developed an optimized random forest (RF) classifier using four attribute selection algorithms, namely, RELIEFF, genetic algorithm (GA), particle swarm optimization (PSO), and firefly algorithm (FA), for false reduction. Our proposed nodule detection was evaluated using 888 exams of LIDC-IDRI, and a favorable performance of 93.11% accuracy, 94.86% sensitivity, and 91.37% specificity with only 0.0863 false positives per scan was achieved using optimized random forest by firefly algorithm. Moreover, our proposed nodule detection ensures the detection of variant type and size of nodules using a standard machine. Therefore, it can be applied as a supportive tool for physicians and radiologists to diagnose lung cancer. However, as a limitation, the preliminary screening of our nodule detection depends on some empirical threshold values to reduce false positives. In future work, the pulmonary nodules detected by our proposed method can be further classified into different clinical stages to determine the severity of lung cancer.

Author Contributions: M.P.P. and C.P. conceived and designed this study; M.P.P. performed the experiments, simulations, and draft preparation of the paper; C.P., K.H., S.T. and S.V. reviewed and edited the paper. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: The authors' sincere thanks go to the National Cancer Institute, the Foundation for the National Institutes of Health, and The Cancer Imaging Archive (TCIA) for publicly sharing of thoracic CT scans that are applied in this research. We do thank ASEAN University Network/Southeast Asia Engineering Education Development Network (AUN/SEED-Net) for providing academic funding. Moreover, the authors would like to acknowledge to Elsevier copyrights team for giving the grant to resue Figure 4.

Conflicts of Interest: The authors declare no conflict of interest.

References

- Bray, F.; Ferlay, J.; Soerjomataram, I.; Siegel, R.L.; Torre, L.A.; Jemal, A. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* 2018, 68, 394–424. [CrossRef] [PubMed]
- 2. The National Lung Screening Trial Research Team. Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening. *N. Engl. J. Med.* **2011**, *365*, 395–409. [CrossRef] [PubMed]
- 3. Paing, M.P.; Pintavirooj, C.; Hamamoto, K.; Tungjitkusolmun, S. A study of ground-glass opacity (GGO) nodules in the automated detection of lung cancer. In *Lung Cancer and Imaging*; IOP Publishing: Bristol, UK, 2019; ISBN 978-0-7503-2540-0.
- 4. Messay, T.; Hardie, R.C.; Rogers, S.K. A new computationally efficient CAD system for pulmonary nodule detection in CT imagery. *Med. Image Anal.* **2010**, *14*, 390–406. [CrossRef] [PubMed]
- 5. Choi, W.-J.; Choi, T.-S. Genetic programming-based feature transform and classification for the automatic detection of pulmonary nodules on computed tomography images. *Inf. Sci.* **2012**, *212*, 57–78. [CrossRef]
- 6. Shen, S.; Bui, A.A.T.; Cong, J.; Hsu, W. An automated lung segmentation approach using bidirectional chain codes to improve nodule detection accuracy. *Comput. Biol. Med.* **2015**, *57*, 139–149. [CrossRef] [PubMed]
- 7. John, J.; Mini, M.G. Multilevel Thresholding Based Segmentation and Feature Extraction for Pulmonary Nodule Detection. *Procedia Technol.* **2016**, *24*, 957–963. [CrossRef]
- 8. Choi, W.-J.; Choi, T.-S. Automated pulmonary nodule detection based on three-dimensional shape-based feature descriptor. *Comput. Methods Programs Biomed.* **2014**, *113*, 37–54. [CrossRef]
- 9. Demir, Ö.; Yılmaz Çamurcu, A. Computer-aided detection of lung nodules using outer surface features. *Biomed. Mater. Eng.* **2015**, *26*, S1213–S1222. [CrossRef]
- 10. da Silva Sousa, J.R.F.; Silva, A.C.; de Paiva, A.C.; Nunes, R.A. Methodology for automatic detection of lung nodules in computerized tomography images. *Comput. Methods Programs Biomed.* **2010**, *98*, 1–14. [CrossRef]
- 11. Cascio, D.; Magro, R.; Fauci, F.; Iacomi, M.; Raso, G. Automatic detection of lung nodules in CT datasets based on stable 3D mass–spring models. *Comput. Biol. Med.* **2012**, *42*, 1098–1109. [CrossRef]
- Keshani, M.; Azimifar, Z.; Tajeripour, F.; Boostani, R. Lung nodule segmentation and recognition using SVM classifier and active contour modeling: A complete intelligent system. *Comput. Biol. Med.* 2013, 43, 287–300. [CrossRef] [PubMed]
- 13. Zhang, W.; Wang, X.; Li, X.; Chen, J. 3D skeletonization feature based computer-aided detection system for pulmonary nodules in CT datasets. *Comput. Biol. Med.* **2018**, *92*, 64–72. [CrossRef] [PubMed]
- Yongbum, L.; Hara, T.; Fujita, H.; Itoh, S.; Ishigaki, T. Automated detection of pulmonary nodules in helical CT images based on an improved template-matching technique. *IEEE Trans. Med. Imaging* 2001, 20, 595–604. [CrossRef] [PubMed]
- 15. El-Baz, A.; Elnakib, A.; Abou El-Ghar, M.; Gimel'farb, G.; Falk, R.; Farag, A. Automatic Detection of 2D and 3D Lung Nodules in Chest Spiral CT Scans. *Int. J. Biomed. Imaging* **2013**, 2013, 1–11. [CrossRef]
- Ozekes, S.; Osman, O.; Ucan, O.N. Nodule Detection in a Lung Region that's Segmented with Using Genetic Cellular Neural Networks and 3D Template Matching with Fuzzy Rule Based Thresholding. *Korean J. Radiol.* 2008, 9, 1. [CrossRef]
- 17. Omar, A. Lung CT Parenchyma Segmentation using VGG-16 based SegNet Model. *IJCA* **2019**, *178*, 10–13. [CrossRef]
- 18. Xu, M.; Qi, S.; Yue, Y.; Teng, Y.; Xu, L.; Yao, Y.; Qian, W. Segmentation of lung parenchyma in CT images using CNN trained with the clustering algorithm generated dataset. *Biomed. Eng. Online* **2019**, *18*, 2. [CrossRef]

- 19. Zhang, G.; Jiang, S.; Yang, Z.; Gong, L.; Ma, X.; Zhou, Z.; Bao, C.; Liu, Q. Automatic nodule detection for lung cancer in CT images: A review. *Comput. Biol. Med.* **2018**, *103*, 287–300. [CrossRef]
- 20. Murphy, K.; van Ginneken, B.; Schilham, A.M.R.; de Hoop, B.J.; Gietema, H.A.; Prokop, M. A large-scale evaluation of automatic pulmonary nodule detection in chest CT using local image features and k-nearest-neighbour classification. *Med. Image Anal.* **2009**, *13*, 757–770. [CrossRef]
- 21. Tan, M.; Deklerck, R.; Jansen, B.; Bister, M. A novel computer-aided lung nodule detection system for CT images. *Med. Phys.* **2011**, *38*, 5630–5645. [CrossRef]
- Kim, D.-Y.; Kim, J.-H.; Noh, S.-M.; Park, J.-W. Pulmonary Nodule Detection using Chest Ct Images. *Acta Radiol.* 2003, 44, 252–257. [CrossRef] [PubMed]
- 23. Matsumoto, S.; Ohno, Y.; Yamagata, H.; Takenaka, D.; Sugimura, K. Computer-aided detection of lung nodules on multidetector row computed tomography using three-dimensional analysis of nodule candidates and their surroundings. *Radiat. Med.* **2008**, *26*, 562–569. [CrossRef] [PubMed]
- Setio, A.A.A.; Ciompi, F.; Litjens, G.; Gerke, P.; Jacobs, C.; van Riel, S.J.; Wille, M.M.W.; Naqibullah, M.; Sanchez, C.I.; van Ginneken, B. Pulmonary Nodule Detection in CT Images: False Positive Reduction Using Multi-View Convolutional Networks. *IEEE Trans. Med. Imaging* 2016, 35, 1160–1169. [CrossRef] [PubMed]
- 25. Yuan, J.; Liu, X.; Hou, F.; Qin, H.; Hao, A. Hybrid-feature-guided lung nodule type classification on CT images. *Comput. Graph.* **2018**, *70*, 288–299. [CrossRef]
- Suzuki, K.; Feng, L.; Sone, S.; Doi, K. Computer-aided diagnostic scheme for distinction between benign and malignant nodules in thoracic low-dose CT by use of massive training artificial neural network. *IEEE Trans. Med. Imaging* 2005, 24, 1138–1150. [CrossRef] [PubMed]
- Lee, S.L.A.; Kouzani, A.Z.; Hu, E.J. Random forest based lung nodule classification aided by clustering. Comput. Med. Imaging Graph. 2010, 34, 535–542. [CrossRef]
- Armato, S.G.; McLennan, G.; Bidaut, L.; McNitt-Gray, M.F.; Meyer, C.R.; Reeves, A.P.; Clarke, L.P. Data From LIDC-IDRI. The Cancer Imaging Archive. 2015. Available online: https://wiki.cancerimagingarchive.net/ display/Public/LIDC-IDRI (accessed on 4 February 2020).
- 29. Armato, S.G.; McLennan, G.; Bidaut, L.; McNitt-Gray, M.F.; Meyer, C.R.; Reeves, A.P.; Zhao, B.; Aberle, D.R.; Henschke, C.I.; Hoffman, E.A.; et al. The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A Completed Reference Database of Lung Nodules on CT Scans: The LIDC/IDRI thoracic CT database of lung nodules. *Med. Phys.* 2011, *38*, 915–931. [CrossRef]
- Clark, K.; Vendt, B.; Smith, K.; Freymann, J.; Kirby, J.; Koppel, P.; Moore, S.; Phillips, S.; Maffitt, D.; Pringle, M.; et al. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. *J. Digit. Imaging* 2013, 26, 1045–1057. [CrossRef]
- 31. Valente, I.R.S.; Cortez, P.C.; Neto, E.C.; Soares, J.M.; de Albuquerque, V.H.C.; Tavares, J.M.R.S. Automatic 3D pulmonary nodule detection in CT images: A survey. *Comput. Methods Programs Biomed.* **2016**, 124, 91–107. [CrossRef]
- 32. Adams, R.; Bischof, L. Seeded region growing. *IEEE Trans. Pattern Anal. Mach. Intell.* **1994**, *16*, 641–647. [CrossRef]
- 33. Freeman, H. Computer Processing of Line-Drawing Images. ACM Comput. Surv. 1974, 6, 57–97. [CrossRef]
- 34. Sánchez-Cruz, H.; López-Valdez, H.H.; Cuevas, F.J. A new relative chain code in 3D. *Pattern Recognit.* 2014, 47, 769–788. [CrossRef]
- Magdy, E.; Zayed, N.; Fakhr, M. Automatic Classification of Normal and Cancer Lung CT Images Using Multiscale AM-FM Features. *Int. J. Biomed. Imaging* 2015, 2015, 1–7. [CrossRef] [PubMed]
- 36. Samet, H.; Tamminen, M. Efficient component labeling of images of arbitrary dimension represented by linear bintrees. *IEEE Trans. Pattern Anal. Mach. Intell.* **1988**, *10*, 579–586. [CrossRef]
- Jerman, T.; Pernuš, F.; Likar, B.; Špiclin, Ž. Beyond Frangi: An improved multiscale vesselness filter. In Proceedings of the SPIE 9413, Medical Imaging 2015: Image Processing, Orlando, FL, USA, 20 March 2015; pp. 24–26.
- 38. Breiman, L. Random Forests. Mach. Learn. 2001, 45, 5–32. [CrossRef]
- Robnik-Šikonja, M. Improving Random Forests. In *Proceedings of the Machine Learning: ECML 2004;* Boulicaut, J.-F., Esposito, F., Eds.; Springer: Berlin/Heidelberg, Germany, 2004; pp. 359–370.
- 40. Kononenko, I.; Imec, E.S.; Ikonja, M.R.-S. Overcoming the Myopia of Inductive Learning Algorithms with RELIEFF. *Appl. Intell.* **1997**, *7*, 39–55. [CrossRef]

- 41. Oluleye, B.; Leisa, A.; Leng, J.; Dean, D. A Genetic Algorithm-Based Feature Selection. *Int. J. Electron. Commun. Comput. Eng.* **2014**, *5*, 899–905.
- 42. Kennedy, J.; Eberhart, R. Particle swarm optimization. In Proceedings of the ICNN'95—International Conference on Neural Networks, Perth, WA, Australia, 27 November–1 December 1995; Volume 4, pp. 1942–1948.
- 43. Yang, X.-S. *Nature-Inspired Metaheuristic Algorithms*, 2nd ed.; Luniver Press: Frome, UK, 2010; ISBN 978-1-905986-28-6.
- 44. Tharwat, A. Classification assessment methods. Appl. Comput. Inform. 2018, in press. [CrossRef]



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).