applied sciences

MDPI

Article

Effective Framework for Pulmonary Nodule Classification from CT Images Using the Modified Gradient Boosting Method

Harsha Vardhan Donga, Jaya Sai Aditya Nandan Karlapati, Harsha Sri Sumanth Desineedi, Prakasam Periasamy * and Sureshkumar TR

School of Electronics Engineering, Vellore Institute of Technology, Vellore 632 014, India
* Correspondence: prakasam.p@vit.ac.in

Abstract: Lung carcinoma, which is commonly known as lung cancer, is one of the most common cancers throughout the world. Mostly, it is not diagnosed until it has spread, and it is very difficult to treat. Hence, early diagnosis of benign and malignant pulmonary nodules can help in the risk assessment of lung cancer for patients, and with proper treatment can save their lives. In this study, a framework for the classification of pulmonary nodules from Computerized Tomography (CT) images using the machine learning-based modified gradient boosting method is proposed. Initially, the obtained CT scan images are preprocessed for better image quality. Next, a random walker method is used to segment the lung nodule boundaries based on seeds provided by the user. After that, the intensity and texture features are extracted using the Local Binary Pattern (LBP) filter and the coefficients of the Riesz wavelet transform. Finally, the proposed modified gradient boost classifier model is trained and tested using the extracted features to classify nodules as either benign or malignant. The proposed framework is verified and validated using the Lung Image Database Consortium (LIDC-IDRI) dataset. From the performance analysis, it was observed that the proposed method achieves a precision, recall, F1 score, and validation accuracy of 0.957, 0.91, 0.941, and 95.67%, respectively. The performance of the proposed method is compared with existing models and is found to be superior. It was found that the proposed classifier is able to efficiently classify pulmonary nodules as either benign or malignant.

Keywords: lung cancer; pulmonary nodules; machine learning; image processing; modified gradient boosting

check for updates

Citation: Donga, H.V.; Karlapati, J.S.A.N.; Desineedi, H.S.S.; Periasamy, P.; TR, S. Effective Framework for Pulmonary Nodule Classification from CT Images Using the Modified Gradient Boosting Method. *Appl. Sci.* **2022**, *12*, 8264. https://doi.org/10.3390/app12168264

Academic Editor: Jan Egger

Received: 7 July 2022
Accepted: 16 August 2022
Published: 18 August 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 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 (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Lung cancer is the second-most commonly diagnosed cancer, with 2.21 million cases. According to a World Health Organization report, lung cancer caused 1.80 million deaths in 2020 [1]. When the cells in the lung tissues divide and grow uncontrollably, they form a mass called a tumor, which can spread to other organs [2]. Small Cell Lung Cancer (SCLC) and Non-Small Cell Lung Cancer (NSCLC) are the two main diseases in cancer. As per the World Health Organization report, NSCLC accounts for up to 85% of cases, while SCLC accounts for the remaining 15%. The most common types of NSCLC are squamous cell cancer, adenocarcinoma, and large cell lung cancer. Additionally, non-smokers constitute 52.1% of lung cancer cases, with a division of 41.8% in men and 88% in women. Because the early stage of lung cancer appearance is very small or even has no symptoms, approximately 85% of lung cancers are diagnosed after the metastatic progression stage [3]. Lung cancer can be cured if it is identified early enough, and hence early diagnosis is the key requirement for reducing risk and improving survival rate.

Due to such a vital issue, many researchers have involved themselves in developing a novel classifier to segment the pulmonary nodules [4–7]. As the technology advances, many researchers have proposed classifiers using various machine learning-based methods, such

as Support Vector Machine (SVM) [8], random forest [9,10], and the gradient boosting algorithm [11]. Deep neural networks and deep residual networks have also been developed to classify lung cancer from Computerized Tomography (CT) scan images [12,13]. Because of its non-invasiveness and high spatial resolution, CT is affordable compared to other imaging techniques such as Magnetic Resonance Imaging (MRI) and Positron Emission Tomography (PET), etc. CT is one of the most often utilized imaging modalities for screening and diagnosing cancer [14]. For clinical purposes, the identification and classification of pulmonary nodules as either benign or malignant is performed by radiologists which is a time-consuming process and prone to errors. It is also a very challenging task due to variable sizes, low contrast, random positions, and the irregular shapes of lung nodules. With the advent of Computer-Aided Design (CAD)-based systems, medical diagnosis has become faster and more accurate, which helps radiologists to improve the sensitivity of their diagnosis by providing a second opinion.

1.1. Related Works

Today, CAD-based systems are regularly used to detect and diagnose numerous abnormalities in clinical practice. The efficiency of these systems is directly related to their disease prediction accuracy. Different image processing and machine learning algorithms are utilized in these systems to analyze the medical scans. Approximately 20% of the total CAD-based medical diagnostic systems are used for lung cancer diagnosis [15]. Hence, an effective framework to segment and classify pulmonary nodules from CT scan images seems to be a viable solution. It consists of multiple phases. Image pre-processing improves the quality of the image by removing artifacts, thus producing better results in the following steps. Segmentation makes the analysis of images easier by differentiating the subjects from the background. After that, extracting features such as the intensity, texture, and color of each pixel helps to represent every image uniquely in terms of mathematical parameters, which can be used to classify the image into different categories.

In a study conducted by Alves et al. [16], CT scan images were used with and without contrast medium and utilized texture features to train the SVM classifier. The SVM classifier was trained with the obtained features to classify the nodules into benign and malignant categories. Senthil Kumar et al. [17] explored various evolutionary image segmentation algorithms for the detection of lung cancer in CT scan images. The adaptive median filter had better pre-processing capabilities compared to the mean and median filters. The pre-processed image was segmented using four segmentation algorithms, i.e., k-means clustering, particle swarm optimization, inertia weighted PSO, and Guaranteed Convergence PSO (GCPSO), and classification accuracy was reported as 0.885, 0.89, 0.889, and 0.958, respectively.

A novel method to automatically detect and classify solitary lung nodules from CT scan images has been suggested [18], and the input CT scan images were denoised using a bilateral filter and thresholding, and morphology was applied for segmentation. The intensity and geometric features were extracted, and a Bayesian classifier, to reduce the time required for training, was employed to define the pulmonary nodules either as benign or malignant. Sharma et al. [19] used CT scan data stored in DICOM format to analyze lung nodules and detect the presence of lung cancer. Raw pixel data from DICOM files were processed by removing unnecessary portions using Thresholding and Connected component analysis. A Convolutional Neural Network (CNN) with an Adam optimizer and ReLU activation function was trained to extract features from the images and classify them as cancerous or non-cancerous. The model, which was used on the data of 20 patients, achieved an accuracy of 65%.

The MRI and CT scan images were fused together based on DWT for better details and pre-processed using a median filter and were segmented with the region of interest method and Otsu's thresholding [20]. The feature extraction and classification of lungs into normal and abnormal was carried out using K-Nearest Neighbor (KNN). Further, region growing and morphology were applied to extract the tumors in abnormal lung images, which were

classified as benign or malignant with the help of an Artificial Neural Network (ANN). The NSCLC history classification was analyzed using different classifiers, such as radionics with KNN and Support Vector Machine (SVM) classifiers, CNN classifiers, CNN with Long Short-Term Memory (LSTM) classifiers, and CNN with LSTM and radiomics classifiers [21]. In this method, the clinical decision was made only with the help of radiological findings. Dhara et al. [22] delineated a classification model that used a semi-automatic technique for segmentation by taking a single seed from the user, and different shapes, margins, and texture-based features were calculated to represent the pulmonary nodules. Three different classification schemes with SVM as a classifier were evaluated, and their area under the Receiver Operating Characteristic (ROC) curve was found to be 0.95, 0.88, and 0.84 with configuration 1, where the "1" and "2" classes were defined as benign and "4" and "5" as malignant.

Bhatia et al. [23] presented a pipeline of pre-processing mechanisms to identify the regions in Lung CT scan images that were more likely to be cancerous and obtained features using deep residual networks such as ResNet and UNet models, which were fed into random forest and XGBoost classifiers. An ensemble of the algorithms performed better than individual classifiers in predicting the probability of lung cancer, with an accuracy of 84%. Deep CNN was proposed for automatic lung cancer diagnosis, and it provided a small Area Under Receiver Operating Characteristic (AUC-ROC) curve of 0.913, a Disc Similarity Coefficient (DSC) of 0.982, and a Jaccard Similarity Coefficient (JSC) of 0.967 [24]. Residual Neural Networks (RNN) with convolutional operations to extract local features and transformer blocks with self-attention to capture global features were shown to be able to identify the pulmonary nodules from CT images [25]. They achieved an AUC and accuracy of 0.9628 and 0.9292, respectively. Bruntha et al. [26] proposed a hybrid classification model that integrates deep features from RNN and handcrafted features from the histogram to classify the lung nodules as benign or malignant.

1.2. Research Contributions

Despite the notable advancements in the technology, there are still a few disadvantages in the existing methods. Most of the existing methods classify the pulmonary nodules as either benign or malignant, but if we also identify the subclass, then this will assist the medical practitioner in diagnosing the pulmonary nodules in a better way. Therefore, it is important for an automatic classifier to classify the in-depth classifications of the lung nodules. The main contributions of the proposed research are as follows:

- The gradient boosting algorithm is modified in a sequential manner by feeding the wrong prediction as the input to the next classifier model, thus improving the efficacy of the classification.
- An effective framework for pulmonary nodule classification from CT images using the modified gradient boosting method for the classification of benign (class 1 and 2) and malignant (class 4 and 5) lung nodules are proposed.
- The random walker method is utilized as the boundary extraction method to segment the pulmonary nodule according to seeds governed by the user.
- Features such as the intensity and texture are extracted using Reisz wavelet coefficients and an LBP filter.
- The performance of the proposed segmentation and classifier model using the modified gradient boosting algorithm method is evaluated by comparing with the existing methods, and is found to be superior.

2. Materials and Methods

In this research work, an ensemble model based on semi-automatic image segmentation and classification of pulmonary nodules in lungs from CT images is proposed. The framework of the proposed pulmonary nodule classification using a modified gradient boosting algorithm is illustrated in Figure 1. This model includes four stages, which are pre-

processing, nodule boundary extraction, feature extraction, and modified gradient boosting classifier.

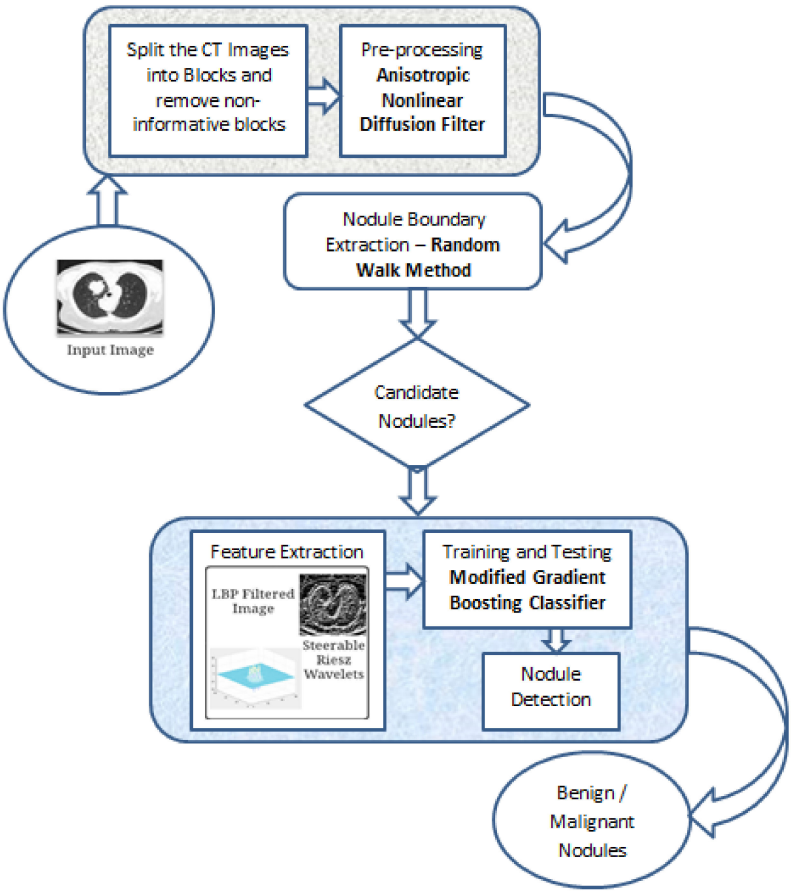


Figure 1. Pulmonary nodule detection and classification—proposed framework.

2.1. Dataset

The proposed machine learning-based modified gradient boosting classifier model is trained and validated using the Lung Image Database Consortium (LIDC-IDRI) (https://wiki.cancerimagingarchive.net/display/Public/LIDC-IDRI (accessed on 1 June 2022).) dataset. The LIDC-IDRI dataset is the largest lung image database and contains 1018 CT thoracic scans, with associated annotation in XML files. These were carried out by four experienced thoracic radiologists, in a two-step action. During the initial stage, the four radiologists sorted the findings into three groups (nodule $\geq$ 3 mm, nodule $\leq$ 3 mm, and non-nodule $\geq$ 3 mm). During the next process, the categorized data were reviewed by the remaining radiologists to ensure uniform evaluation. The LIDC-IDRI dataset contains 1018 CT scans from 1010 patients, with a total of 244,527 images. In this dataset, the lung nodules are divided and labelled as five classes. The lung nodules labelled as class 1 and 2 are considered as benign nodules, and those labelled as class 4 and 5 are considered as malignant nodules. Lung nodules labelled as class 3 are considered as non-nodules.

2.2. Preprocessing

The preprocessing plays a vital role in analyzing and extracting the desired features from the image. The obtained CT scan images are converted into jpg format for easy processing. The purpose of the proposed model is to segment and classify the pulmonary nodules. In order to achieve this, the unwanted artifacts and noise present in the CT scan images must be removed to improve the texture quality of the nodules. To meet these requirements, an anisotropic diffusion filtering technique is used.

The anisotropic nonlinear diffusion filter is used to remove the noise adaptively while preserving the details and edges. This helps the random walker algorithm to accurately segment the boundaries of the pulmonary nodules. This filter also enhances the texture quality of images, thereby helping in extracting the texture features precisely. The same traits are observed from the results after preprocessing the image.

2.3. Random Walker Method for Nodule Boundary Extraction

The random walker method is a semi-supervised learning method that is more suitable for image segmentation. The implementation of random walker for image segmentation has been presented in depth by Grady [27]. In this method, the image is represented as a graph instead of a matrix, and the pixel intensity values are assigned with each node (called seeds). Every node is connected to its adjacent node via an edge. In this technique, two seeds are given to the image in such a way that one seed is on the nodule (subject) and the other is on the rest of the lung area (background) [28]. Each node is associated with either of the two labels as background or foreground, depending on the probability that a random walker leaving the pixel will first arrive at each seed. The solution to this random walker problem can be found by solving an equivalent electrical circuit for the voltages at each node marked at one seed as '0′ and the other seed as '1′.

Let us assume there are two nodes, N_a and N_b, and the weight of the edge, w_{ab}, is given by:

$$w_{ab} = e^{-\beta|Na-Nb|} \tag{1}$$

where β is the conductance factor and w_{ab} represents the conductance of the edge. The higher the value, the easier it is to move on that edge.

After marking the seeds with "0" and "1", the values of the remaining nodes are calculated using:

$$N_a = \sum_b w_{ab} \tag{2}$$

where b is the number of nodes directly connected to the Node a. The incident matrix at the edge node is given by:

$$M_{ab} = \begin{cases} N_a, & if\ a = b \\ -w_{ab}, & if\ Na\ and\ Nb\ are\ adjacent \\ 0, & Otherwise \end{cases}$$

In general, the matrix, M_{ab}, can be represented by:

$$M = \begin{bmatrix} M_m & B \\ B' & M_u \end{bmatrix} \tag{3}$$

where M_m is the entries corresponding to marked nodes, B is the relation between one marked node and another unmarked node, B' is the relation between an unmarked node to a marked node, and M_u is the relation between all unmarked nodes.

When the potential of the node is $Q(N_i)$, then the critical points at node b is defined by:

$$l_b^w = \begin{cases} 1, & if\ Q(N_b) = w \\ 0, & if Q(N_b) \neq w \end{cases} \tag{4}$$

Therefore, for the label w, the solution to the conjunctional Dirichlet problem may be found by solving:

$$M_u X = -B'L \tag{5}$$

for one label and:

$$M_u X^* = -B'L \tag{6}$$

for the remaining labels. Therefore, the probability at node 'b' is given by:

$$p(w/x) = x_b^w \tag{7}$$

where $p(w \mid x)$ is the probability that node x belongs to class w.

After solving all these equations, one can easily classify the nodes belonging to their respective segments. After segmenting the image, the segmented nodule is used for further analysis. The major advantages of the random walker method are as follows:

- It can handle complex boundaries.
- As the model assumes the image as a graph where pixels are nodes or vertices, graph theory and popular graph algorithms can also be used.
- It can be easily employed to divide the image into multiple segments based on the number of seed labels provided.

2.4. Feature Extraction

In general, intensity-, geometry-, and texture-based features are extracted from images that are used as predictors for classification/regression purposes. Some of the features, such as intensity and texture, are extracted in the segmentation process for random walk. Furthermore, the texture features from the image are extracted by calculating the Reisz wavelet coefficients, which are helpful in the classification of the malignancy.

2.4.1. Local Binary Pattern (LBP) Filter

The Local Binary Pattern (LBP) filter is a well-known techniques used for image representation and classification in machine learning models. It is a simple but very efficient texture operator that labels the pixels of an image by thresholding the neighborhood of each pixel and assigning the result as a binary value. Because of its inequitable strength and computational efficient, it has become a suitable approach for different applications. Due to this, it is seen as a uniting approach and has replaced the conventional structural and divergent statistical texture analysis. Therefore, this method is more suitable for a wide range of applications, such as facial expression recognition, texture classification, and object detection. This filter applies a sliding window of size 3×3 and extracts the LBP code.

This LBP code is obtained by thresholding the neighborhood pixels with the center pixel of that window. The LBP code is given by the equation:

$$LBP = \sum_{n=0}^{7} s(i_n - i_c)2^n \tag{8}$$

where i_n is the neighborhood pixel value and i_c is the center pixel value

$$s(z) = \begin{cases} 1, & z \geq 0 \\ 0, & z < 0 \end{cases}$$

Thus, considering the binary pattern code as a number and replacing the center pixel with this number gives the output. The neighboring pixels are replaced with ones and zeros by comparing with the center pixel value.

2.4.2. Riesz Wavelet Coefficients

Furthermore, the texture features from the image are extracted by calculating the Reisz wavelet coefficients, which are helpful in the classification of the malignancy. Images may have either smooth regions interrupted by edges or abrupt changes, and these abrupt changes in an image are most crucial because they represent boundaries within the image. Fourier transform is a powerful tool for data analysis, but it does not represent those changes significantly because sine waves are not localized in time, nor in space. Therefore, in order to analyze these changes, wavelet transforms may be used [29]. A wavelet is a rapidly decaying wave-like oscillation that has zero mean; unlike sinusoids, which extend

to infinity, a wavelet exists for a finite duration. The availability of a wide range of wavelets is a key strength of wavelet analysis.

To generate Riesz coefficients [30], a 3D steerable wavelet transform composed of two operators is used in this proposed research. A 3D isotropic wavelet transforms will perform multiscale decomposition and yields vanishing moments, and then a 3D Riesz transform adds directionality and steerability [31,32]. The calculation of Riesz wavelet coefficients involves the wavelet frame, $\{\psi_{i,k}\}$ I ∈ Z, k ∈ Z^3, whose basis functions in $L_2(R^3)$ are generated by a single mother wavelet, $\psi(x)$. The normalized wavelets at scale i and translation parameter, k, are given by:

$$\psi_{i,k}(x) = \psi_i\left(x - 2^i k\right) \text{ with } \psi_i(x) = 2^{\frac{-3i}{2}}\ \psi\left(x/2^i\right). \tag{9}$$

The computation of Riesz wavelet coefficients is divided in two steps: (1) discrete convolution of the image with the sampled filters, (2) down sampling by a factor of T_i.

The set of Riesz wavelet coefficients, $s^n_{i,k} = \{\mathcal{R}^n \psi_i * f\}(x)\Big|_{x=T_i k}$, is computed as:

$$\begin{aligned} s^n_{i,k} &= \left\{ \mathcal{R}^n \sum_{k' \epsilon \mathbb{Z}^3} \omega_{i,k'} \frac{1}{T_i^3} sinc\left(\frac{-T_i k'}{T_i}\right) \right\}(x)\Big|_{x=T_i k} \\ &= \sum_{k' \epsilon \mathbb{Z}^3} \omega_{i,k'} \{\mathcal{R}^n sinc(k - k')\} \end{aligned} \tag{10}$$

where $\mathcal{R}^n$ denotes the N^{th} order Riesz transform with $n = (n_1, n_2, n_3) \in N^3$, such that $|n| = N$ with pulsation vector $\omega = (\omega_1, \omega_2, \omega_3) \in R^3$:

$$\hat{\mathcal{R}^n}(\omega) = (-j)^N \sqrt{\frac{N!}{n_1! n_2! n_3!}} \frac{\omega_1^{n_1} \omega_2^{n_2} \omega_3^{n_3}}{\left(\omega_1^2 + \omega_2^2 + \omega_3^2\right)^{N/2}} \tag{11}$$

From the above equations, the Riesz wavelet coefficients can be obtained by scaling eight times because each image is resized into 256 × 256 dimensions.

2.5. Proposed Modified Gradient Boosting Classifier

The segmented images from the random walker method are used to train the proposed modified gradient boosting algorithm-based classifier. The basic building blocks of the proposed classifier are shown in Figure 2. In order to classify robustly, the weak learners must be converted to strong learners by ensemble learning [33–35]. Regular ensemble methods, such as random forest, depend on the straightforward voting of learners in the ensemble. The group of boosting strategies is based on an alternate, helpful technique of ensemble arrangement. The boosting algorithm identifies the weak learner and generates a new weak prediction rule from that learner. This algorithm involves an iteration process, such that for every iteration, the boosting algorithm makes the weak prediction rules into a single strong prediction rule. The gradient boosting algorithm is a machine learning technique that works well with ensemble methods. This algorithm is used to reduce the error rate by applying the gradient descent technique.

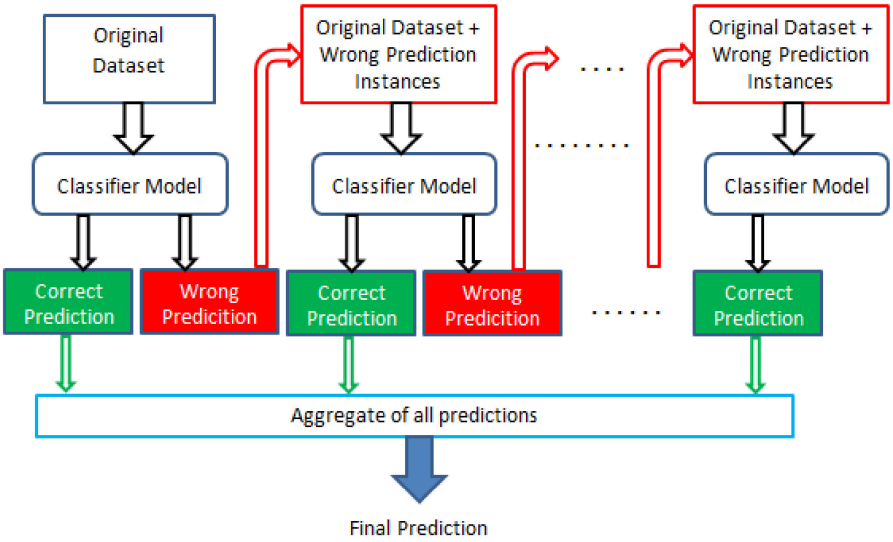


Figure 2. Basic building blocks of the proposed modified gradient boosting model.

The idea behind this technique is to create new base-learners that are maximally associated with the loss function's negative gradient, which is connected to the entire ensemble. The loss functions used can be arbitrary, but to give an intuition, if the error function is the typical squared-error loss, the learning technique will result in a series of error-fittings. Boosting algorithms are moderately easy to implement, which allows one to try different things with various model plans. The proposed modified gradient boosting algorithm used in the classifier model is explained in Algorithm 1.

Algorithm 1. Proposed Modified Gradient Boosting Algorithm

Begin
Inputs
Dataset (x_i y)
Number of iterations, M
Loss Functions, $L(y, F(x))$
Choose the base-learner function, a_m
Initalize $F_0(x)$
for m = 1 to M
Compute the negative gradient function $g_m(x_i)$.
Generate the new base-learner function $h(x; a_m)$.
$a_m = \text{arg}min_{a,\beta} \sum_{i=1}^{N} [g_m(x_i) - \beta h(x_i; a)]^2$, where $h(x; a)$ is a weak learner
Evaluate the best gradient descent step-size ρ_m.
$\rho_m = \text{arg}min_{\rho} \sum_{i=1}^{N} L(y_i, F_{m-1}(x_i) + \rho h(x_i; a_m))$
Update the estimation/prediction function
$F_m(x) = F_{m-1}(x) + {}_m\, h(x; a_m) = y$
End for
Output prediction, y
End

Let us suppose (x_i, y) to be the dataset, with x_i being the input parameters and y be the output parameter, which is a multi-class variable. The objective is to construct a functional reliance, $F^*(x)$, that maps from x to y, which minimizes the specified loss function, $L(y, F(x))$.

$$F^*(x) = \text{arg}min_F E_{y,\ x} L(y,\ F(x)) = \text{arg}min_F E_x \left[E_y (L(y,\ F(x))) \middle|\ x \right] \quad (12)$$

The initial value is obtained by:

$$\text{F}_0(x) = \text{arg}min_{\rho} \sum_{i=1}^{N} L(y_i,\ \rho) \quad (13)$$

where ρ is step size.

The negative gradient function is obtained by taking the derivative of the above function, and it is given as:

$$g_m(x_i) = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}, \; i = 1, \ldots N \tag{14}$$

In the end, at an estimator/prediction function is arrived at in order to classify the output class based on the input parameters.

3. Performance Metrics

The performance of the proposed modified gradient boosting classifier and random walker segmentation method is verified and validated using the disc coefficient, Jaccard Similar Coefficients, precision, recall, F1 score, validation accuracy, and Receiver Operating Characteristics (ROC) curve, based on the following assumptions:

The accuracy of the proposed segmentation method is evaluated using the Disc Coefficient (DSC) and the Jaccard Similar Coefficient (JSC) using the following equation:

$$DSC = \frac{2 \times (M \cap R)}{|M| + |R|} \tag{15}$$

$$JSC = \frac{(M \cap R)}{(M \cup R)} \tag{16}$$

where M and R are ground truth standard and segmented result of the proposed method, respectively.

The performance of the proposed modified gradient boosting classifier is evaluated using the following metrics:

$$Precision = \frac{TP}{TP + FP} \tag{17}$$

$$Recall = \frac{TP}{TP + FN} \tag{18}$$

$$F1\ Score = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{19}$$

where TP is True Positive—object is present at the goal location and detected as a correct class; TN is True Negative—object is not present at the goal location and not detected; FP is False Positive—object is not present at the goal location and detected as a wrong class; and FN is False Negative—object is present at the goal location and not detected.

4. Experimental Results

The proposed machine learning-based modified gradient boosting classifier model is trained and validated using the Lung Image Database Consortium (LIDC-IDRI) dataset. The computational setup consists of a CPU equipped with a 64-bit Windows10 operating system and an Intel i3-1005G1 processor with a clock frequency of 1.20 GHz and 16 GB RAM. The proposed lung nodule detection and classification framework is simulated using MATLAB R2020b.

4.1. Pre-Processing Result

The input CT images are divided into blocks, and the unwanted non-informative blocks are removed for further pre-processing. The unwanted artifacts and noise present in the CT scan images are removed in order to improve the texture quality of the nodules. The unwanted artifacts are filtered using the nonlinear anisotropic diffusion filter. The sample input image obtained from the dataset, the pre-processed image using the anisotropic nonlinear diffusion filter, and the extracted texture features for the input image using the proposed research are illustrated in Figure 3.

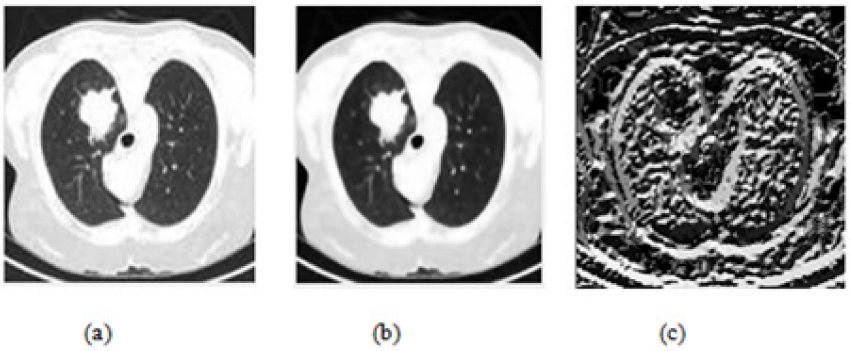


Figure 3. (**a**) Input CT image, (**b**) Preprocessed image using anisotropic nonlinear diffusion filter, (**c**) Texture features.

From Figure 3b, it can be seen that the anisotropic nonlinear diffusion filter reduces the noise and enhances the textures of the image. This helps in extracting the features in a better way for further analysis. The Peak Signal-to-Noise (PSNR) ratio is calculated to measure the effectiveness of the used pre-processing method. PSNR is the most commonly used metrics to measure the efficacy of the denoised/compressed images. The average PSNR obtained in this research is 34.56 dB.

4.2. Lung Nodule Segmentation Result

To examine whether the random walker method help to improve the segmentation performance, the boundaries of the pulmonary nodules are segmented and extracted using the random walker method. The pre-processed lung nodule CT images are fed into the segmentation block to extract the boundaries of the nodules. In this proposed research, the random walker method is used. The original CT image, the ground truth standard, and the result of segmentation is illustrated in Figure 4. It can be seen that the random walker method segments the small and micro nodules accurately for further processing. The masked portion indicates the nodule under consideration for analysis.

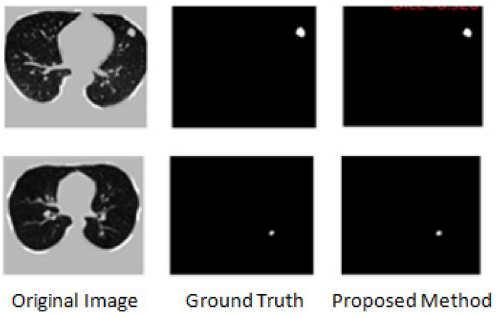


Figure 4. Lung nodule segmentation result using the random walker method.

The accuracy of the proposed segmentation method is validated using the DSC and JSC, and it is obtained as 0.979 ± 0.011 and 0.877 ± 0.008, respectively. From this result, is it can be seen that the proposed method provides better segmentation to extract the lung nodules. The various samples of benign class and malignant class are fed into the proposed model, and the boundaries are segmented. The segmented boundaries for various classes using the random walker method are illustrated in Figure 5.

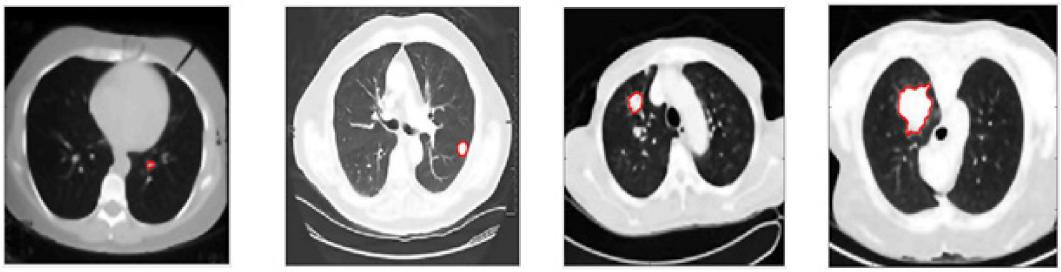


Figure 5. (**a**) Boundary of benign class 1 nodule, (**b**) Boundary of benign class 2 nodule, (**c**) Boundary of malignant class 4 nodule, (**d**) Boundary of malignant class 5 nodule.

4.3. Feature Extraction Result

A 3D steerable wavelet transform to calculate the Riesz coefficients, which consist of the composition of two operators, is executed. The first one is the 3D isotropic wavelet transform, which performs multiscale decomposition and gives vanishing moments. The next is the 3D Riesz transform, which brings directionality and steerability. The texture features of the segmented images are extracted using the steerable Riesz wavelet transforms and are shown in Figure 6. The steerable Riesz wavelet functions are steering in the direction of the local gradient energy, which is more suitable for analyzing and processing the texture of the CT images.

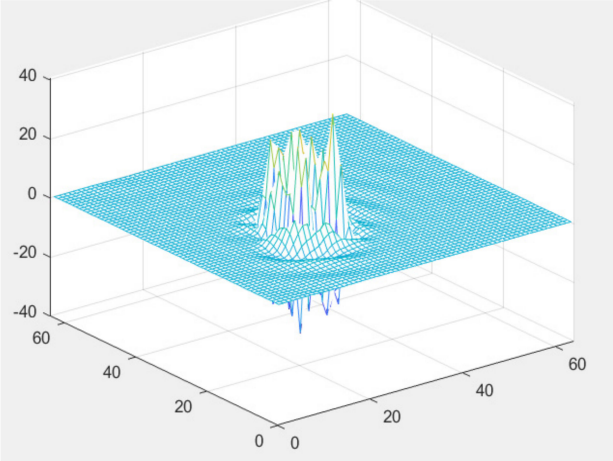


Figure 6. Steerable Riesz wavelets for texture analysis.

The proposed classifier is trained using the features extracted and validated to classify the pulmonary nodules as either benign or malignant.

4.4. Lung Nodule Classification Result

The proposed modified gradient boosting classifier is trained and tested with the given dataset. The entire dataset is divided into 75:25 for training and validating the proposed classifier. Out of 1018 CT scan images, 764 have been used to train the proposed classifier model, and the remaining 254 images have been used to validate the proposed model. A confusion matrix is generated for classifying the test dataset as benign class 1 nodule, benign class 2 nodule, non-nodules class 3, malignant class 4 nodule, and malignant class 5 nodule for the proposed method and is shown in Figure 7.

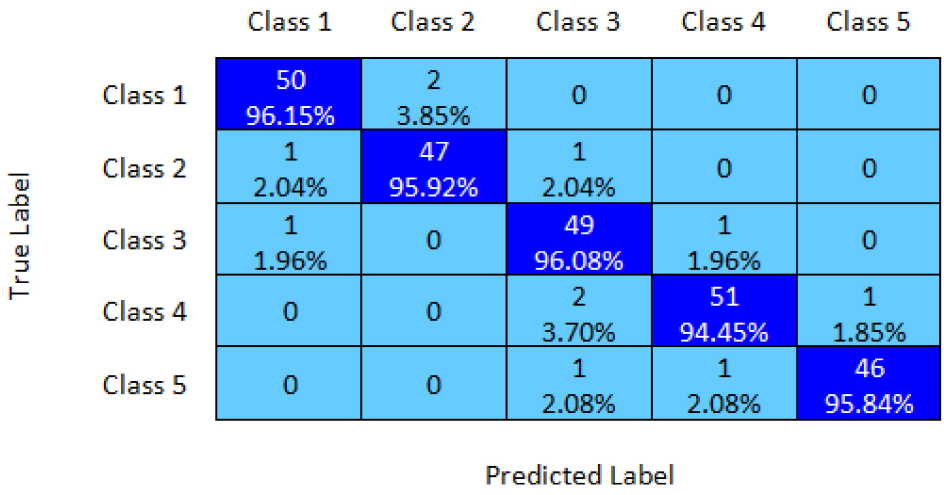


Figure 7. Confusion matrix of the proposed modified gradient boosting classifier.

From Figure 7, it can be seen that the proposed classifier identifies the lung nodules in a correct class more accurately. Due to the similarity in the nature of the boundaries, the proposed system classifies malignant nodules as non-nodules, with 5.78%, and benign nodules as non-nodule, with 2.04%.

The precision vs. recall and the ROC curve are generated based on the confusion matrix for the proposed modified gradient boosting classifier and are shown in Figure 8. From Figure 8a, it can be seen that the proposed modified gradient boosting classifier curve is steeper to the right corner of the graph, which indicates the better performance of the proposed method.

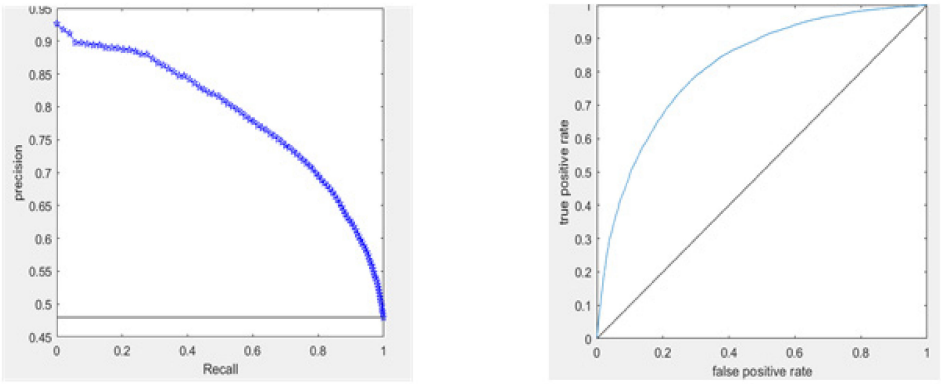


Figure 8. (**a**) Precision vs. recall curve and (**b**) ROC curve for the proposed modified gradient boosting classifier model.

The ROC curve, which is illustrated in Figure 8b, is the graph between the True Positive Rate (TPR) and the False Positive Rate (FPR). The blue line indicates the ROC curve of the classifier and diagonal line indicates the arbitrary classifier. From Figure 8b, it can be seen that the proposed modified gradient boosting classifier curve is steeper toward the left corner of the graph. From the performance analysis, it can be seen that the proposed classifier achieves a precision, recall, F1 score, and validation accuracy of 0.957, 0.91, 0.941, and 95.67%, respectively.

5. Discussion

In this research work, a modified gradient boosting classifier was proposed to identify pulmonary nodules as either benign, malignant, or non-nodule. The CT scan images

obtained from the LIDC-IDRI dataset were pre-processed to remove artifacts using an anisotropic nonlinear diffusion filter, and it was observed that this provides a PSNR of 34.56 dB. This proves the efficacy of the experimented pre-processing method. The random walker method was used to extract the boundaries of the pulmonary nodules from the pre-processed CT scan images. The ground truth standard and the segmented result obtained using the random walker method was analyzed using DSC and JSC, which were found to be 0.979 ± 0.011 and 0.877 ± 0.008, respectively. The accuracy of the proposed research was compared with existing methods, such as the graph cut method, watershed method, 3D U-Net [36], and 3D FCN [37]. From Table 1, it can be seen that the pulmonary segmentation using the random walker method achieved the higher DSC and JSC, compared with other existing methods. The major contribution of the proposed research could segment the appropriate region of the pulmonary nodule for different cases.

Table 1. Image segmentation performance comparison with existing methods.

Methods/Model	DSC	JSC
Graph Cut Method	0.566 ± 0.025	0.414 ± 0.021
Watershed Method	0.628 ± 0.027	0.494 ± 0.025
3D U-Net	0.911 ± 0.009	0.811 ± 0.011
3D FCN	0.845 ± 0.008	0.738 ± 0.011
Proposed Method	0.979 ± 0.011	0.877 ± 0.008

Various texture features were extracted from the segmented image to train and test the proposed modified gradient boosting classifier using the LBP filter and steerable Riesz wavelet coefficients. The steerable Riesz wavelet functions are steering in the direction of the local gradient energy, which is more suitable for analyzing and processing the texture of the CT images. The ROC accuracy, precision, recall, and F1 score were used to measure the performance of the proposed classifier.

To confirm the superiority of the proposed modified gradient boosting classifier model, the performance was compared with the existing methods/models. The various methods/models considered were random forest classifier [38], the deep RNN model [39], the SVM classifier [16], and Res U-Net [36]. The performance comparison is shown in Table 2. From Table 1, it can be seen that the proposed modified gradient boosting classifier achieved a validation accuracy of 95.67%, which is higher than other reporting methods. The main reason for this is that the modified gradient boosting classifier algorithm uses a sequential approach, making it more accurate and efficient. Additionally, it achieved a higher precision and F1 score, which indicates that it is suitable for more efficient pulmonary nodule classification. However, from Figure 7, it can be seen that the proposed classifier drops in accuracy when a higher number of malignant nodules are presents in the testing image dataset. This is identified as one of the major limitation of the proposed classifier.

Table 2. Performance comparison of the proposed modified gradient boosting classifier with existing models.

Methods/Model	Dataset Used	Accuracy (%)	Precision	Recall	F1 Score
Random Forest [38]	LIDC-IDRI	85.00	0.7962	0.7810	0.7918
Deep RNN [39]	LIDC-IDRI/Luna 16	85.71	0.879	0.892	0.882
SVM [16]	LIDC-IDRI	93.0	0.87	0.86	0.84
Res U-Net [36]	LIDC-IDRI	90.70%	0.916	0..902	0.911
Proposed Modified Gradient Boosting Classifier	LIDC-IDRI	95.67	0.957	0.91	0.941

6. Conclusions

In this research article, an efficient and effective framework model to classify the pulmonary nodules in CT scan images using a modified gradient boosting algorithm is implemented. The quality of the CT scan images is considerably improved during the pre-processing stage. The random walker method produces the desired level of segmentation in the CT scan image, and the Riesz wavelets provide a good basis for the representation of the images to extract features. The proposed classifier is used to properly classify the pulmonary nodules as either benign or malignant. From the validation, it is observed that the proposed modified gradient boosting classifier is able to classify the pulmonary nodules with 95.67% accuracy. From comparison with existing methods, it is proved that it can outperform them. In the current model, a semiautomatic approach for the segmentation of lung nodules has been used. In our future studies, it is planned to overcome the limitations of the proposed classifier by adding more feature sets to train and test the proposed classifier. Additionally, implementation of a fully automatic technique to identify lung nodules and explore 3D-based deep networks for feature extraction is planned. The performance of a hybrid pipeline with an ensemble of classifiers on various datasets for the detection of lung cancer will be studied and cross-validated with experts to find a use for clinical diagnosis.

Author Contributions: P.P. and S.T. devised the main conceptual ideas and proofread them. H.V.D. conceived the majority of the technical details, devised the model, and undertook data collection and experimentation. P.P. performed the evaluation metrics to validate the proposed model. S.T., H.V.D., J.S.A.N.K. and H.S.S.D. prepared the entire manuscript, and it was verified by P.P. All authors have read and agreed to the published version of the manuscript.

Funding: The Article Processing Charge is funded by the Vellore Institute of Technology, Vellore, India.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Custom Code—will be submitted based on the request.

Conflicts of Interest: We hereby declared that there is no conflict of interest in this research work/paper.

References

1. World Health Organization. Cancer Report. Available online: https://www.who.int/news-room/fact-sheets/detail/cancer (accessed on 1 June 2022).
2. Li, X.; Li, B.; Tian, L.; Zhang, L. Automatic benign and malignant classification of pulmonary nodules in thoracic computed tomography based on RF algorithm. *IET Image Process.* **2018**, *12*, 1253–1264. [CrossRef]
3. Zhang, G.; Yang, Z.; Gong, L.; Jiang, S.; Wang, L. Classification of benign and malignant lung nodules from CT images based on hybrid features. *Phys. Med. Biol.* **2019**, *64*, 125011. [CrossRef] [PubMed]
4. Abdillah, B.; Bustamam, A.; Sarwinda, D. Image processing based detection of lung cancer on CT scan images. *J. Phys. Conf. Ser.* **2017**, *893*, 012063. [CrossRef]
5. Tanabe, Y.; Ishida, T. Development of a novel detection method for changes in lung conditions during radiotherapy using a temporal subtraction technique. *Phys. Eng. Sci. Med.* **2021**, *44*, 1341–1350. [CrossRef] [PubMed]
6. Tartar, A.; Kilic, N.; Akan, A. Classification of Pulmonary Nodules by Using Hybrid Features. *Comput. Math. Methods Med.* **2013**, *2013*, 1–11. [CrossRef] [PubMed]
7. Wang, X.; Mao, K.; Wang, L.; Yang, P.; Lu, D.; He, P. An Appraisal of Lung Nodules Automatic Classification Algorithms for CT Images. *Sensors* **2019**, *19*, 194. [CrossRef] [PubMed]
8. Abaszade, M.; Effati, S. A New Method for Classifying Random Variables Based on Support Vector Machine. *J. Classif.* **2018**, *36*, 152–174. [CrossRef]
9. Zhou, J.; Li, E.; Wei, H.; Li, C.; Qiao, Q.; Armaghani, D.J. Random Forests and Cubist Algorithms for Predicting Shear Strengths of Rockfill Materials. *Appl. Sci.* **2019**, *9*, 1621. [CrossRef]
10. El-Askary, N.S.; Salem, M.A.M.; Roushdy, M.I. Lung Nodule Detection and Classification using Random Forest: A Review. In Proceedings of the 9th International Conference on Intelligent Computing and Information Systems (ICICIS), Cairo, Egypt, 8–10 December 2019; pp. 105–111. [CrossRef]
11. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
12. Jakimovski, G.; Davcev, D. Using Double Convolution Neural Network for Lung Cancer Stage Detection. *Appl. Sci.* **2019**, *9*, 427. [CrossRef]

13. Song, Q.; Zhao, L.; Luo, X.; Dou, X. Using Deep Learning for Classification of Lung Nodules on Computed Tomography Images. *J. Healthcare Eng.* **2017**, *2017*, 1–7. [CrossRef] [PubMed]
14. Khan, S.A.; Hussain, S.; Yang, S.; Iqbal, K. Effective and Reliable Framework for Lung Nodules Detection from CT Scan Images. *Sci. Rep.* **2019**, *9*, 4989. [CrossRef] [PubMed]
15. Tomassini, S.; Falcionelli, N.; Sernani, P.; Burattini, L.; Dragoni, A.F. Lung nodule diagnosis and cancer histology classification from computed tomography data by convolutional neural networks: A survey. *Comput. Biol. Med.* **2022**, *146*, 105691. [CrossRef] [PubMed]
16. Alves, A.F.F.; Souza, S.A.; Ruiz, R.L.; Reis, T.A.; Ximenes, A.M.G.; Hasimoto, E.N.; Lima, R.P.S.; Miranda, J.R.A.; Pina, D.R. Combining machine learning and texture analysis to differentiate mediastinal lymph nodes in lung cancer patients. *Phys. Eng. Sci. Med.* **2021**, *44*, 387–394. [CrossRef] [PubMed]
17. Kumar, K.S.; Venkatalakshmi, K.; Karthikeyan, K. Lung Cancer Detection Using Image Segmentation by means of Various Evolutionary Algorithms. *Comput. Math. Methods Med.* **2019**, *2019*, 1–16. [CrossRef]
18. Mukherjee, J.; Chakrabarti, A.; Shaikh, S.H.; Kar, M. Automatic Detection and Classification of Solitary Pulmonary Nodules from Lung CT Images. In Proceedings of the Fourth International Conference of Emerging Applications of Information Technology, Kolkata, India, 19–21 December 2014; pp. 294–299. [CrossRef]
19. Sharma, K.; Soni, H.; Agarwal, K. Lung Cancer Detection in CT Scans of Patients Using Image Processing and Machine Learning Technique. In *Advanced Computational and Communication Paradigms. Lecture Notes in Electrical Engineering*; Bhattacharyya, S., Gandhi, T., Sharma, K., Dutta, P., Eds.; Springer: Singapore, 2018; Volume 475, pp. 336–344. [CrossRef]
20. Veeraprathap, V.; Harish, G.S.; Narendra Kumar, G. Lung Cancer detection and multi-level classification using discrete Wavelet Transform approach. *Int. J. Biomed. Biol. Eng.* **2020**, *14*, 17–23. [CrossRef]
21. Marentakis, P.; Karaiskos, P.; Kouloulias, V.; Kelekis, N.; Argentos, S.; Oikonomopoulos, N.; Loukas, C. Lung cancer histology classification from CT images based on radiomics and deep learning models. *Med Biol. Eng. Comput.* **2021**, *59*, 215–226. [CrossRef]
22. Dhara, A.K.; Mukhopadhyay, S.; Dutta, A.; Garg, M.; Khandelwal, N. A Combination of Shape and Texture Features for Classification of Pulmonary Nodules in Lung CT Images. *J. Digit. Imaging* **2016**, *29*, 466–475. [CrossRef] [PubMed]
23. Pabón, O.S.; Torrente, M.; Provencio, M.; Rodríguez-Gonzalez, A.; Menasalvas, E. Integrating Speculation Detection and Deep Learning to Extract Lung Cancer Diagnosis from Clinical Notes. *Appl. Sci.* **2021**, *11*, 865. [CrossRef]
24. Maity, A.; Nair, T.R.; Mehta, S.; Prakasam, P. Automatic lung parenchyma segmentation using a deep convolutional neural network from chest X-rays. *Biomed. Signal Process. Control* **2021**, *73*, 103398. [CrossRef]
25. Liu, D.; Liu, F.; Tie, Y.; Qi, L.; Wang, F. Res-trans networks for lung nodule classification. *Int. J. Comput. Assist. Radiol. Surg.* **2022**, *17*, 1059–1068. [CrossRef]
26. Bruntha, P.M.; Pandian, S.I.; Anitha, J.; Abraham, S.S.; Kumar, S.N. A novel hybridized feature extraction approach for lung nodule classification based on transfer learning technique. *J. Med. Phys.* **2022**, *47*, 1–9. [CrossRef] [PubMed]
27. Grady, L. Random Walks for Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2006**, *28*, 1768–1783. [CrossRef]
28. Eslami, A.; Karamalis, A.; Katouzian, A.; Navab, N. Segmentation by retrieval with guided random walks: Application to left ventricle segmentation in MRI. *Med Image Anal.* **2013**, *17*, 236–253. [CrossRef] [PubMed]
29. Dogra, A.; Goyal, B.; Agrawal, S. Bone vessel image fusion via generalized reisz wavelet transform using averaging fusion rule. *J. Comput. Sci.* **2017**, *21*, 371–378. [CrossRef]
30. Unser, M.; Chenouard, N. A Unifying Parametric Framework for 2D Steerable Wavelet Transforms. *SIAM J. Imaging Sci.* **2013**, *6*, 102–135. [CrossRef]
31. Chenouard, N.; Unser, M. 3D Steerable Wavelets in Practice. *IEEE Trans. Image Process.* **2012**, *21*, 4522–4533. [CrossRef]
32. Cirujeda, P.; Muller, H.; Rubin, D.; Aguilera, T.A.; Loo, B.W.; Diehn, M.; Binefa, X.; Depeursinge, A. 3D Riesz-wavelet based Covariance descriptors for texture classification of lung nodule tissue in CT. In Proceedings of the 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Milan, Italy, 25–29 August 2015; pp. 7909–7912. [CrossRef]
33. Bentéjac, C.; Csörgő, A.; Martínez-Muñoz, G. A comparative analysis of gradient boosting algorithms. *Artif. Intell. Rev.* **2020**, *54*, 1937–1967. [CrossRef]
34. Birunda, S.S.; Devi, R.K. A Novel Score-Based Multi-Source Fake News Detection using Gradient Boosting Algorithm. In Proceedings of the International Conference on Artificial Intelligence and Smart Systems (ICAIS), Coimbatore, India, 25–27 March 2021; pp. 406–414. [CrossRef]
35. Shin, Y. Application of Stochastic Gradient Boosting Approach to Early Prediction of Safety Accidents at Construction Site. *Adv. Civ. Eng.* **2019**, *2019*, 1–9. [CrossRef]
36. Yu, H.; Li, J.; Zhang, L.; Cao, Y.; Yu, X.; Sun, J. Design of lung nodules segmentation and recognition algorithm based on deep learning. *BMC Bioinform.* **2021**, *22*, 1–21. [CrossRef]
37. Shoji, K.; Shunske, K.; Yasushi, H.; Shingo, M.; Tohru, K.; Nobuyuki, T.; Yuki, S.; Masahiro, Y.; Noriyuki, T. Segmentation of Lung Nodules on CT Images Using a Nested Three-Dimensional Fully Connected Convolutional Network. *Front. Artif. Intell.* **2022**, *5*, 782225. Available online: https://www.frontiersin.org/articles/10.3389/frai.2022.782225 (accessed on 10 March 2022).

38. Nada, S.; El-Askary, M.; Salem, A.-M.; Roushdy, M.I. Feature Extraction and Analysis for Lung Nodule Classification using Random Forest. In Proceedings of the 8th International Conference on Software and Information Engineering, Cairo, Egypt, 9–12 April 2019; pp. 248–252. [CrossRef]
39. Wang, S.; Dong, L.; Wang, X.; Wang, X. Classification of pathological types of lung cancer from CT images by deep residual neural networks with transfer learning strategy. *Open Med.* **2020**, *15*, 190–197. [CrossRef] [PubMed]