Hindawi
Computational and Mathematical Methods in Medicine
Volume 2022, Article ID 2733965, 14 pages
https://doi.org/10.1155/2022/2733965

Research Article

Ensemble Learning Framework with GLCM Texture Extraction for Early Detection of Lung Cancer on CT Images

Sara A. Althubiti,[1] **Sanchita Paul**,[2] **Rajanikanta Mohanty**,[3] **Sachi Nandan Mohanty**,[4] **Fayadh Alenezi**,[5] **and Kemal Polat**[6]

[1]*Department of Computer Science, College of Computer and Information Sciences, Majmaah University, Al-Majmaah, Saudi Arabia*
[2]*Department of Computer Science & Engineering, Birla Institute of Technology, Mesra, Ranchi, India*
[3]*Department of Computer Science & Engineering, Specialisation Program, Faculty of Engineering and /Technology, Jain University, Bangalore, India*
[4]*Department of Computer Science & Engineering, Vardhaman College of Engineering (Autonomous), Hyderabad, India*
[5]*Department of Electrical Engineering, College of Engineering, Jouf University, Saudi Arabia*
[6]*Department of Electrical and Electronics Engineering, Bolu Abant Izzet Baysal University, Faculty of Engineering, Bolu, Turkey*

Correspondence should be addressed to Kemal Polat; kpolat@ibu.edu.tr

Received 8 April 2022; Revised 29 April 2022; Accepted 10 May 2022; Published 2 June 2022

Academic Editor: Naeem Jan

Lung cancer has emerged as a major cause of death among all demographics worldwide, largely caused by a proliferation of smoking habits. However, early detection and diagnosis of lung cancer through technological improvements can save the lives of millions of individuals affected globally. Computerized tomography (CT) scan imaging is a proven and popular technique in the medical field, but diagnosing cancer with only CT scans is a difficult task even for doctors and experts. This is why computer-assisted diagnosis has revolutionized disease diagnosis, especially cancer detection. This study looks at 20 CT scan images of lungs. In a preprocessing step, we chose the best filter to be applied to medical CT images between median, Gaussian, 2D convolution, and mean. From there, it was established that the median filter is the most appropriate. Next, we improved image contrast by applying adaptive histogram equalization. Finally, the preprocessed image with better quality is subjected to two optimization algorithms, fuzzy c-means and k-means clustering. The performance of these algorithms was then compared. Fuzzy c-means showed the highest accuracy of 98%. The feature was extracted using Gray Level Cooccurrence Matrix (GLCM). In classification, a comparison between three algorithms—bagging, gradient boosting, and ensemble (SVM, MLPNN, DT, logistic regression, and KNN)—was performed. Gradient boosting performed the best among these three, having an accuracy of 90.9%.

1. Introduction

Carcinoma is the leading cause of death in the world. Carcinomas are cancers that start in cells that make up the skin or the tissue lining organs, such as the lungs or kidneys. Lung cancer, also known as carcinoma of the lungs, is characterized by an unrestricted growth of cells in lung tissue and distinguished by a specific growth pattern. Lung cancer is dangerous to leave untreated, as it may propagate to other body parts. Small-cell lung carcinoma and nonsmall-cell lung carcinoma are the two major categories, and the primary cause is smoking. Lung cancer has also been found in people with no smoking history but with exposure to air pollution, secondary smoking, and sometimes toxic gasses. Before the 12th century, occurrence of lung cancer was actually very rare. But nowadays, it is widespread. Many patients consult a doctor only when their disease and symptoms become extreme, thereby making these disease and symptoms very difficult to diagnose and cure. Thus, early-stage treatment of lung cancer is crucial in saving lives. One way

to detect the distinctive abnormal growth of cells is through X-ray. Another method of cancer detection is sputum cytology. If the lungs produce sputum, cancer can be seen by looking at the sputum through a microscope. Tissue sampling, also called biopsy, is another method for early detection of lung cancer. The conventional and most widespread method of detecting lung cancer is by using computer tomography (CT) and radiographs. CT scan uses X-ray and a computer to deliver a clear image of the lungs, giving better results than an X-ray alone. The CT scan image gives much more detail than a plain image, and the doctors can view a particular organ from different angles [1–33]. In this study, 20 lung image samples are taken for analysis. The image is denoised; then, the image is enhanced. Afterwards, features are extracted using GLCM. Lastly, classification is done. Integration of median filter, adaptive histogram equalization, and fuzzy c-means clustering for segmentation showed more accurate results. After applying feature extraction using GLCM (Haralick features), the accuracy of the ensemble classifier consisting of MLPNN, DT, SVM, and KNN classifiers was computed and confirmed to be highly effective. Thus, the study has great potential to advance the early detection of lung cancer.

2. Related Works

Senthil Kumar et al. [34] used a segmentation algorithm (k-means) on computer tomography (CT) scan images to detect lung cancer. Image segmentation was achieved by applying fuzzy c-means and k-means algorithms. Fuzzy c-means delivered enhanced performance in comparison to k-means. Using guaranteed convergence particle swarm optimization (GCPSO), an accuracy of 95.89% was achieved for the detection of lung cancer. Using a novel Multicrop Convolutional Neural Network (MC-CNN), an accuracy of 86.24% was achieved in identifying the lung module malignancy. In MC-CNN, features are extracted from the nodules by trimming distinct areas from convolution feature maps and applying max-pooling several times [35]. Sensitivity of 70%-90% was achieved using random forest and principal component analysis by extracting features using local shape analysis [36]. Using two successive k-nearest neighbor classifiers, a sensitivity of 80% was achieved using the curvedness and shape feature of the local image [37]. Accuracy of 95.91% was achieved using a probabilistic neural network (PNN) by extracting lung volume, and reduction was done using principal component analysis (PCA) [38]. Accuracy of 95.62% was achieved using texture, volumetric, intensity, and geometric features, and Fuzzy Particle Swarm Optimization (FPSO) was used for feature selection, with deep learning being applied for classification [39]. Sensitivity of 93.02% was achieved in detection detecting ground-glass opacity (GGO) using Support Vector Machine (SVM) twice and using four 2-dimensional features and 11 3-dimensional features [40]. Classification accuracy of 96% was achieved using speed up robust feature (SURF) along with genetic algorithms (GA) for optimization and a neural network (NN) for classification [41]. 97.61% accuracy was achieved using a genetic algorithm with wrapper approach (GAWA) using a multilevel brightness-preserving approach and segmentation using a deep neural network. Features are derived from the segment and selected using a generalized rough set (hybrid spiral optimization intelligent) [42]. An accuracy of 89.29% was obtained using two 3D deep learning models [43]. Using 2D and 3D shape and texture features and histogram, k-means clustering (autocenter) provided a sensitivity of 88.88%. [44]. Using volumetric CT data, sensitivity reached more than 90% using a 3D convolution neural network. [45–52].

3. Materials and Methods

Firstly, a filtering technique is used to filter out the noise from the 20 images. In this study, 4 filters were used for the purpose of comparison. The filters used were mean, median, Gaussian, and 2D convolution. Afterwards, adaptive histogram equalization was applied so that images became clear. A segmentation algorithm was applied for the proper segmentation of images. This step used k-means clustering and fuzzy c-means clustering for segmentation. After segmentation, with the help of GLCM (Gray Level Cooccurrence Matrix), 8 features, i.e., contrast, energy, entropy, homogeneity, sum of entropy, sum of variance, dissimilarity, and sum of average, were extracted from the images to form the dataset of 41 CT scan images (20 were from [34] and 20 were from a different paper: Abnormalities Detection in CT Scan Lung Images Using GLCM [37]) where 28 are lung cancer patients and 13 are patients not affected by cancer. The use of two datasets makes the results more generalized. Ensemble learning was used for the classification of the dataset. Bagging and gradient boosting (a part of ensemble learning) were used for classification. Figure 1 shows the block diagram of framework for detection of lung cancer.

3.1. Filtering

3.1.1. Mean Filter. It blurs the image to reduce noise to a minimum. It involves calculating the mean values of pixels in the $m \times m$ kernel. The mean will replace the intensity of the center element's pixel. This results in smoothing and removal of noise up to a certain extent. This can be implemented using the OpenCV library. For color images, it is necessary to convert the images from RGB to HSV, as the dimensions of RGB are interdependent, and the dimensions of HSV are independent separately.

3.1.2. Gaussian Filter. This filter is similar to the mean filter, but it calculates the weighted mean of the neighboring pixels having a parameter sigma with a discrete approximation. The kernel represents the value of the Gaussian distribution. Although it blurs edges like a standard filter, it is good at protecting edges compared to similar-sized filters. This can also be implemented using the OpenCV package. It allows us to specify the kernel's size.

3.1.3. Median Filter. This filter calculates the median of neighboring pixels to the center in the $m \times m$ kernel. The median then changes the center pixel. It does an excellent job in removing slight noises compared to mean and Gaussian filters. It also preserves the edges of the image but fails to

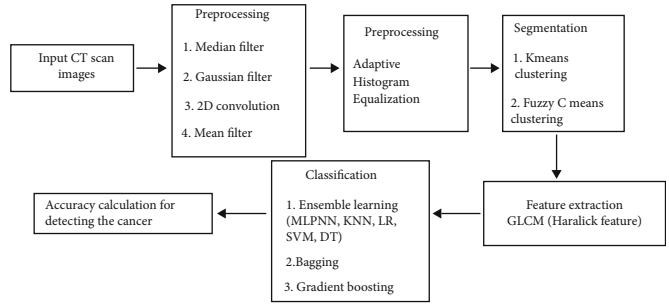

Figure 1: Block diagram of framework for detection of lung cancer.

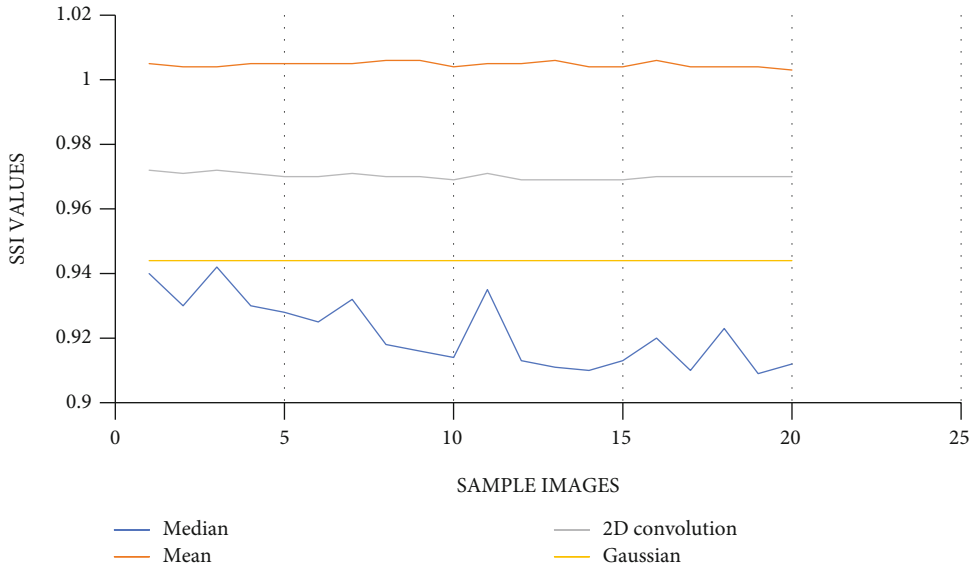

Figure 2: SSI comparison of filters using graph.

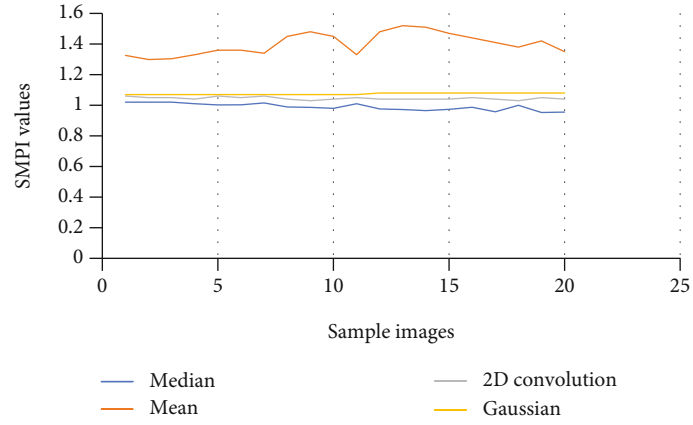

Figure 3: SMPI comparison of filters using graph.

deal with speckle noise. This can also be implemented using the OpenCV library.

3.1.4. 2D Convolution Filter. When applying a 2D Convolution filter, images are filtered utilizing Low Pass Filters (LPF) and High Pass Filters (HPF). Low Pass Filter blurs the image and removes noise. High Pass filters detect edges. For each pixel, a 3×3 window is centered on this pixel. All pixels falling within this frame are added, and then, the result is divided by 9. It is equivalent to computing the average pixel value inside that frame. This is performed for all image pixel values to give an output filtered image.

(1) Performance Measure. Performance measure of all the four filters, i.e., mean, median, Gaussian, and 2D convolution, is done by comparing SMPI (Speckle Suppression and Mean Preservation Index) and SSI (Speckle Suppression Index) metrics. Per these indices, a lower value represents better performance of filters for mean preservation and noise reduction. Figure 2 shows the SSI comparison of filters using graph. Figure 3 gives the SMPI comparison of filters using graph.

$$\mathrm{SSI} = \frac{\sqrt{\mathrm{Variance\ (final\ Image)}}}{\mathrm{mean\ (final\ image)}} \times \frac{\mathrm{mean\ (Initial\ Image)}}{\sqrt{\mathrm{Variance\ (Initial\ Image)}}},$$

$$\mathrm{SMPI} = Q \times \frac{\sqrt{\mathrm{Variance\ (Final\ Image)}}}{\sqrt{\mathrm{Variance\ (Initial\ Image)}}},$$

$$Q = 1 + |\mathrm{mean\ (initial\ Image)} - \mathrm{mean\ (final\ Image)}|. \quad (1)$$

In Table 1, the SSI value of the 4 filters (mean, median, Gaussian, and 2D convolution) is provided with their corresponding graphical comparisons in Figure 2. In Table 2, SMPI values of 4 filters are compared, with their graphical comparisons in Figure 3. The lower values of SSI and SMPI denote better preservation of the image after filtering. From the comparison of different filters, as shown in Figures 2 and 3 and Tables 1 and 2, it can be concluded that the median filter is the best and has more accurate characteristics than the remaining filters. Thus, we use median filtered images for image segmentation.

3.2. Adaptive Histogram Equalization. The color histogram in image processing addresses the number of pixels in each sort of colored part. Because the histogram equation causes a substantial change in the image's color balance, it cannot be applied independently for an image's red, green, and blue components. However, the algorithm can be applied to the luminance or value channel due to changes in the image's color and saturation if the image is first converted to another color space, such as the HSL/HSV color space. The primary difference between an adaptive histogram and ordinary histogram is that the adaptive approach generates numerous histograms for each image region and utilizes them to redistribute the image's lightness value. Therefore, it is appropriate for refining local contrast in each region of an image and increasing the definition of edges. This step enhances the image, and edges will become sharper and clearer which is necessary for medical image segmentation. Figure 4 shows the resultant image (1 to 20) after preprocessing.

3.3. Image Segmentation. Image segmentation is defined as the method by which a digital image is separated into several different regions, each a set of pixels with distinct objects or similar characteristics. Locating objects and boundaries in images is the main function of image segmentation. It can be divided into several methods. With this strategy, the distinct shapes of cancer cell clusters play an important role in determining how severe the cancer is. In our case, two clustering algorithms were used to perform segmentation of images—k-means clustering and fuzzy-c means clustering.

Table 1: SSI values of different filters.

Images	Median	Mean	2D convolution	Gaussian
1	0.94	1.005	0.972	0.944
2	0.93	1.004	0.971	0.944
3	0.942	1.004	0.972	0.944
4	0.93	1.005	0.971	0.944
5	0.928	1.005	0.97	0.944
6	0.925	1.005	0.97	0.944
7	0.932	1.005	0.971	0.944
8	0.918	1.006	0.97	0.944
9	0.916	1.006	0.97	0.944
10	0.914	1.004	0.969	0.944
11	0.935	1.005	0.971	0.944
12	0.913	1.005	0.969	0.944
13	0.911	1.006	0.969	0.944
14	0.91	1.004	0.969	0.944
15	0.913	1.004	0.969	0.944
16	0.92	1.006	0.97	0.944
17	0.91	1.004	0.97	0.944
18	0.923	1.004	0.97	0.944
19	0.909	1.004	0.97	0.944
20	0.912	1.003	0.97	0.944

Table 2: Comparison of SMPI values of 4 filters.

Images	Median	Mean	2D convolution	Gaussian
1	1.02	1.326	1.06	1.07
2	1.02	1.299	1.05	1.07
3	1.02	1.304	1.05	1.07
4	1.01	1.33	1.04	1.07
5	1.002	1.36	1.06	1.07
6	1.003	1.36	1.05	1.07
7	1.015	1.34	1.06	1.07
8	0.989	1.45	1.04	1.07
9	0.986	1.48	1.03	1.07
10	0.98	1.45	1.04	1.07
11	1.01	1.33	1.05	1.07
12	0.976	1.48	1.04	1.08
13	0.972	1.52	1.04	1.08
14	0.965	1.51	1.04	1.08
15	0.973	1.47	1.04	1.08
16	0.987	1.44	1.05	1.08
17	0.957	1.41	1.04	1.08
18	1	1.38	1.03	1.08
19	0.953	1.42	1.05	1.08
20	0.955	1.35	1.04	1.08

3.3.1. K-Means Clustering Algorithm. The k-means clustering algorithm is the most basic and classical form of cluster analysis. We apply k-means to separate the given dataset into two or more groups. The method's accuracy is measured by evaluating each cluster center produced by the

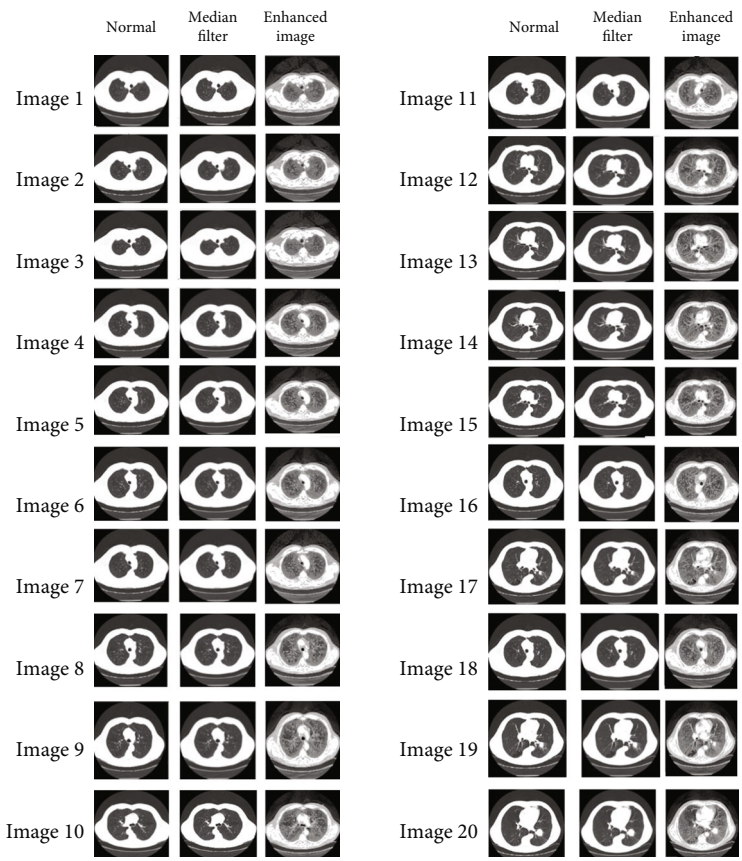

Figure 4: Resultant image (1 to 20) after preprocessing.

Step 1: Find cluster center - let it be "c".
Step 2: Compute Euclidean distance.
Step 3: Assign every pixel to the appropriate pixel by checking the minimum Euclidean distance between pixel and cluster.
Step 4: If all pixel segregation is done, then again calculate the new cluster center using the k-means formula.
Step 5: Repeat steps 2 to 4 until the end condition is encountered.

Algorithm 1.

Step 1: Find the cluster center, let it be "c" randomly select the cluster center.
Step 2: Compute Fuzzy belonging using Equation (3).
Step 3: Compute new fuzzy cluster center using Equation (4).
Step 4: Repeat steps 2 to 3 until the end condition is encountered or the objective function is achieved.

Algorithm 2.

algorithm, as selecting the proper cluster center is essential for getting the best results. A very simple method to separate the dataset is by using Euclidean distance, which we use to assign pixels to an individual cluster. The following function is used in this algorithm:

$$J = \sum_{i=1}^{m} \sum_{k=1}^{K} W_{ik} \left\| x^i - \mu_k \right\|^2, \quad (2)$$

where x_i is the pixels, v_j is the cluster centers, |xi − vj| is the Euclidean distance between x_i and v_j, Ci is the number of data points for the i^{th} cluster, and C_i is the number of cluster centers. Approach k-m to solve the problem is called expectation-maximization. The expectation phase assigns data points to the nearest cluster. The maximization phase calculates the nucleus of each cluster. Below is how we solve it mathematically.

3.3.2. Fuzzy C-Means Clustering Algorithm. Fuzzy clustering (also known as soft clustering or soft k-means) is a clustering method by which each data point can be assigned to

multiple clusters. This clustering or cluster analysis includes grouping data points into clusters such that items in the same cluster are as similar as possible, while points in different clusters are as dissimilar as possible. Groups are distinguished through similarity metrics such as distance, connectivity, and intensity. Depending on the data or application, different similarity measures can be employed. The membership of each data point relating to each cluster center is determined by the distance between the cluster center and the data point. The more data in the cluster center, the more membership towards the special cluster center. The membership magnitude of each data point must sum to one, after updating each recursive membership and cluster center principle:

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{c} \left(d_{ij}/d_{ik}\right)^{(2/m-1)}}, \tag{3}$$

$$V_j = \frac{\left(\sum_{i=1}^{n} \left(\mu_{ij}\right)^m x_{\mathrm{i}}\right)}{\left(\sum_{i=1}^{n} \left(\mu_{ij}\right)^m\right)}, \forall j = 1, 2, 3..c, \tag{4}$$

where

"μ_{ij}" represents the membership of i^{th} data to j^{th} cluster center. "c" represents the number of cluster centers. "d_{ij}" represents the Euclidean distance between i^{th} data and j^{th} cluster center, and "n" is the number of the data point. "m " is the fuzziness index $\mathrm{m} \in [1, \infty]$. "$\mathrm{v_j}$" represents the j^{th} cluster center.

Performance measure: Here, we do the accuracy measure of both clustering algorithms, i.e., k-means and Fuzzy c-means, with a median filter for the segmentation of the image

Accuracy: a performance measure that gives information about the correctness of any process

True positive (TP): foreground pixels are correctly segmented

True negative (TN): background pixels are correctly detected

False positive (FP): foreground pixels are incorrectly segmented

False negative (FN): background pixels are incorrectly detected

The above Tables 3 and 4 show the true positive rate, true negative rate, false positive rate, false negative rate, and accuracy of k-means clustering algorithm (Table 3) and fuzzy c-means clustering algorithm (Table 4). Figure 5 shows a graphical comparison of TPR between k-means and fuzzy c-means. Similarly, Figure 6 shows an FPR comparison. Figure 7 shows the TNR comparison. Figure 8 shows the FNR comparison. Figure 9 shows the accuracy comparison between k-means and fuzzy c-means using a graph.

Edge detection in an image is a crucial technique for determining the limits of various distinctive objects. It can be implemented by looking for discontinuities in the brightness. Masks can be used for edge detection. Some of them are Laplacian operators, Sobel, and Canny. They are calculated using dissimilarity between adjacent pixels of the image.

Table 3: Performance measure of fuzzy c-means clustering.

Images	TPR	FPR	TNR	FNR	Accuracy
1	96.4	0	100	3.5	98.71
2	96.4	0	100	3.5	98.74
3	96.6	0	100	3.3	98.78
4	95.9	0	100	4	98.63
5	95.7	0	100	4	98.61
6	95.5	0	100	4	98.57
7	96.1	0	100	3.8	98.68
8	95.1	0	100	4	98.51
9	94.6	0	100	5	98.37
10	94.4	0	100	5	98.28
11	96.3	0	100	3	98.74
12	94.4	0	100	5	98.9
13	94.4	0	100	5	98.29
14	93.9	0	100	6	98.15
15	94.6	0	100	5	98.34
16	95.3	0	100	4	98.53
17	91.6	0	100	8.3	97.39
18	95.5	0	100	4	98.56
19	91.1	0	100	8	97.2
20	94.3	0	100	5	98.27

Table 4: Performance measure of k-means clustering.

Images	TPR	FPR	TNR	FNR	Accuracy
1	84.5	1.2	98.7	15.4	93.07
2	82.2	1.1	98.8	17.7	92.11
3	85.8	1.1	98.8	14.1	93.65
4	75.1	1	98.9	24.8	89.04
5	74.3	1.2	98.7	25.6	88.71
6	75.3	1.1	98.8	24.6	89.39
7	76.4	1.1	98.8	23.5	89.49
8	68.7	1	98.9	31.2	86.19
9	68	1.1	98.8	31.9	86.02
10	67.4	1.3	98.6	32.5	85.54
11	79.8	1.1	98.8	20.1	91.08
12	70.6	1.4	98.5	29.3	87.35
13	66.9	1.2	98.7	33	85.25
4	67.3	1.2	98.7	32.6	85.65
15	70.7	1.3	98.6	29.2	87.40
16	67.7	1.1	98.8	32.2	85.35
17	68.9	1.4	98.5	31	86.27
18	74.6	1.15	98.8	25.3	89.17
19	69.8	1.5	98.4	30.1	86.71
20	69.7	1.4	98.5	30.2	86.96

3.4. Feature Extraction. Feature extractions from a segmented image yield several important properties that are utilized in defining the segmented image's characteristics. The crucial information of the presence of nodules (or lack thereof), which is used to detect or distinguish between

Figure 5: TPR comparison of k-means and fuzzy c-means.

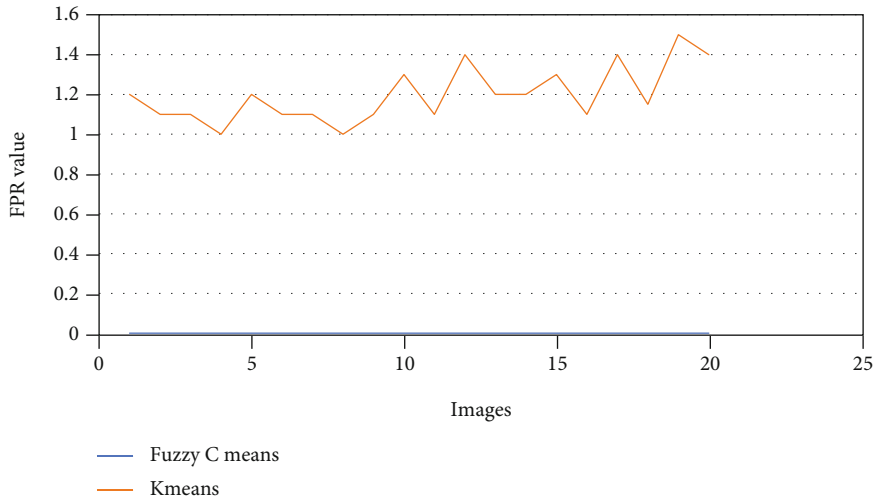

Figure 6: FPR comparison of k-means and fuzzy c-means.

Figure 7: TNR comparison of k-means and fuzzy c-means.

malignant and nonmalignant images, can be diagnosed using the extracted features. 8 Haralick features, namely, contrast, energy, entropy, homogeneity, sum of entropy, sum of variance, dissimilarity, and sum, as shown in Table 5, were extracted by finding GLCM (Gray Level Cooccurrence Matrix). These 8 features of the images were used in the analysis in this study.

3.4.1. Gray Level Cooccurrence Matrix (GLCM). GLCM is an image analysis technique. It is a statistical method for examining the shape of the pixels of an image as a gray-scale matrix, also known as the gray-scale spatial cooccurrence matrix. It is a classification technique, the final step of which is to train the classifier. Its main function is to extract the texture feature from the image. The GLCM function

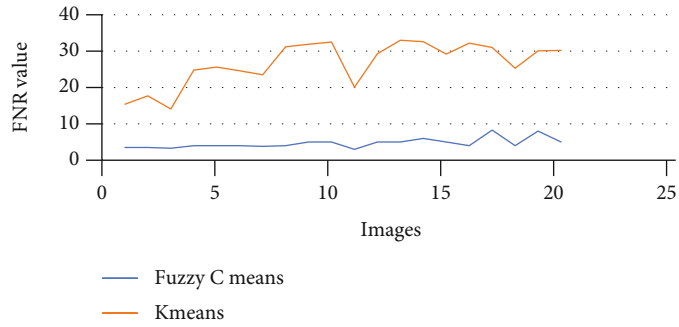

Figure 8: FNR comparison of k-means and fuzzy c-means.

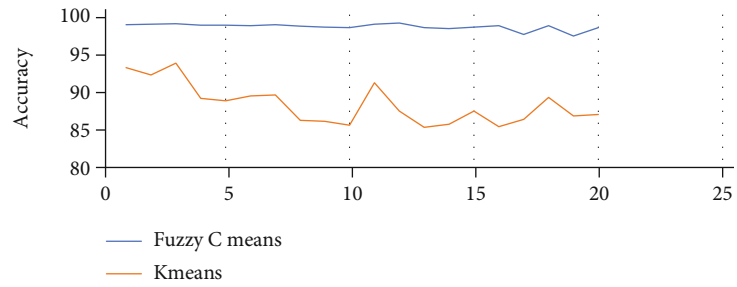

Figure 9: Accuracy comparison of k-means and fuzzy c-means.

Table 5: Haralick features extracted from GLCM.

1	Contrast	$\sum_i \sum_j (i-j)^2 p_d(i,j)$
2	Energy	$Energy = \sqrt{\text{ASM}}$ $\text{ASM} = \sum_i \sum_j {p_d}^2(i,j)$
3	Entropy	$-\sum_i \sum_j p_d(i,j) lnln\, p_d(i,j)$
4	Homogeneity	$\sum_i \sum_j \frac{1}{1+(i-j)^2} p_d(i,j)$
5	Sum of entropy	$-\sum_{i=2}^{2N_g} p_{x+y}(i) \text{loglog}\left\{p_{x+y}(i)\right\} = f_8$
6	Sum of variance	$\sum_{i=2}^{2N_g} (i-f_8)^2 p_{x+y}(i)$
7	Dissimilarity	$\sum_{j=1}^{N} \lvert i-j \rvert . p(i,j)$
8	Sum of average	$\sum_{i=2}^{2N_g} i p_{x+y}(i)$

generates a GLCM and then extracts the statistical functions from this matrix with the specified values and spatial relationship of the shape of an image. The gray-coefficient matrix is derived from the gray-scale coefficient matrix. Gray-level cooccurring grids are also called gray-level spatial dependence grids. The gray-cum-matrix is used to generate the GLCM by computation, but i, which usually represents gray-level (gray-level probability), is a valuable, horizontal neighbor to j. Each part of the GLCM (i, j) represents the sum of the image element. The figure below shows the gray-scale coherence grid-matrix (GLCM) of the gray-scale image (i and j = image element).

Haralick Features:

3.5. Classification

3.5.1. Ensemble Learning. Ensemble learning is a method for systematically building and combining a large number of machine learning models in tandem to solve a specific problem. By merging different models, machine learning outcomes can be dramatically improved. This method outperforms a single model in terms of prediction accuracy. Here, 5 models are considered for ensemble learning: decision tree classifier, multilayer perceptron classifier, Support Vector Machine, K-nearest neighbor classifier, and logistic regression classifier. For meta outcome evaluation, we use the maximum voting technique to find optimal accuracy among all 5 models.

3.5.2. Bagging. Bagging is a strategy used to boost the accuracy of a machine learning algorithm. The main goal is the creation of multiple different subsets of data from randomly chosen training samples, and then, substitution is done. The decision trees are trained by different subsets of data. This results in a collection of various models, which oftentimes multiplies the power of a model.

Bagging steps are as follows:

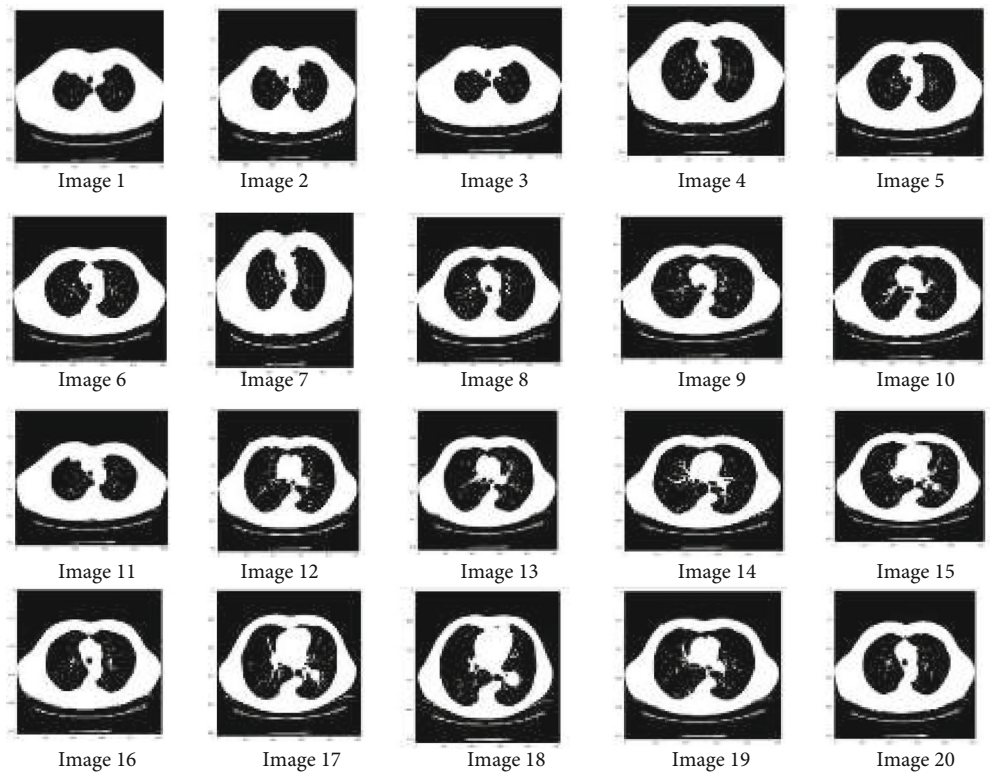

Figure 10: Resultant image (1 to 20) after segmentation.

Figure 11: Thresholding.

Figure 12: Masking.

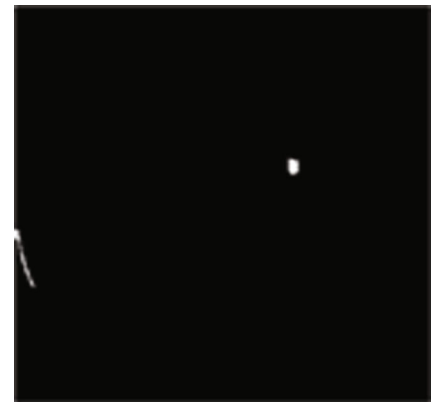

Figure 13: Extraction.

(i) Suppose that the training dataset has n observations and m characteristics. With substitution, one sample is randomly selected from the training dataset

(ii) A subset of L features is chosen randomly, and the best features are used to iterate over the partition node

(iii) The tree becomes the largest

(iv) Repetition of the above steps is carried out n times, and the prediction is built on the sum of predictions by the number of n trees

3.5.3. Boosting. Boosting is used to convert weak learners to strong learners. It is one of the most used algorithms in data

Table 6: Confusion matrix of various classification algorithms.

	True positive	True negative	False positive	False negative
Ensemble	2	0	2	7
Bagging	1	1	2	7
Boosting	1	1	0	9

Table 7: Comparison of performance measure of various classification algorithms.

	Ensemble learning	Bagging	Boosting
	DT, logistic regression, MLPNN, SVM, KNN	Decision tree	Gradient boosting
Accuracy	81.81%	72.72%	90.90%
Error	18.18%	27.27%	9.09%
Sensitivity	50%	33.33%	100%
Prediction	100%	50%	50%

science. In this method, learners are sequentially trained with early learners to fit simple models to the data, after which, the data is analyzed to detect the errors. In order to achieve a progressively higher accuracy in each step from the preceding tree, successive trees are fitted. When a hypothesis implies an input, its weight is increased, making the next hypothesis more likely to be categorized correctly. This technique transforms low-performing learners into high-performing models.

Boosting steps are as follows:

(i) Weak learner W is trained by drawing a random subset of training sample T without replacement from training set P

(ii) In order to train the weak learner W2, a second random training subset P2 is drawn without replacement from the training set, then 50 percent of the earlier incorrect classified/miscall sample is added

(iii) In order to train the third weak learner W3, training samples P are found in training set P3, on which there is a disagreement between W1 and W2

(iv) All the weak learners are mixed through majority voting

(v) In order to train the weak learner W2 again, a second random training subset T2 is drawn without replacement from the training set and 50 percent of the earlier incorrect classified/miscall sample is added

(vi) W3, the third weak learner, is trained by finding a training sample P in training set T3 where there is a disagreement between W1 and W2

(vii) Weak learners are again mixed through majority voting

3.5.4. Gradient boosting. The gradient boosting machine (GBM) is a machine learning technique for boosting, regression, and classification problems that generates weak prediction models, usually a prediction model combined with a decision tree. It is an ensemble learning method where the weak models used are decision trees. It defines a loss function and minimizes it. It builds step-by-step models just like other boosting methods and simplifies them by allowing optimization of the arbitrary differential loss function. Gradient boosting can be understood more easily with the basic idea of AdaBoost. Gradient boosting is a proven powerful algorithm to build a predictive model, which is why we tested and selected it here.

4. Results and Discussions

A confusion matrix is a table that shows how well a classification model (or "classifier") performs on a set of test data for which the true values are known. This enables the performance of an algorithm to be visualized.

In the preprocessing step, the performance of the median filter was the best among all the other tested filters—mean, Gaussian, and 2D convolution. From the SMPI and SSI values as shown in Tables 1 and 2 and Figures 2 and 3, it can be found that the image segmentation using a median filter has better performance than a mean filter—Gaussian and 2D convolution. True positive rate, true negative rate, false positive rate, and false negative rate were used to calculate the segmentation accuracy. For segmentation, the accuracy of fuzzy c-means clustering is higher than the k-means clustering algorithm. Fuzzy c-means achieves 97% accuracy. All the results are shown in Tables 3 and 4. All the comparisons of TPR, TNR, FNR, and FPR are shown in Figures 5–8. The accuracy comparison between k-means and fuzzy c-means was shown in Figure 9. The results show that the fuzzy c-means clustering algorithm outperforms k-means for lung cancer CT image segmentation. After that, the dataset was obtained by extracting Haralick features of 41 CT scan images (21 were from [34], and 20 were from abnormalities detection in CT scan lung images using GLCM [37]) and was classified using an ensemble learning algorithm. The resultant image of all 20 images after segmentation is shown in Figure 10. The output after

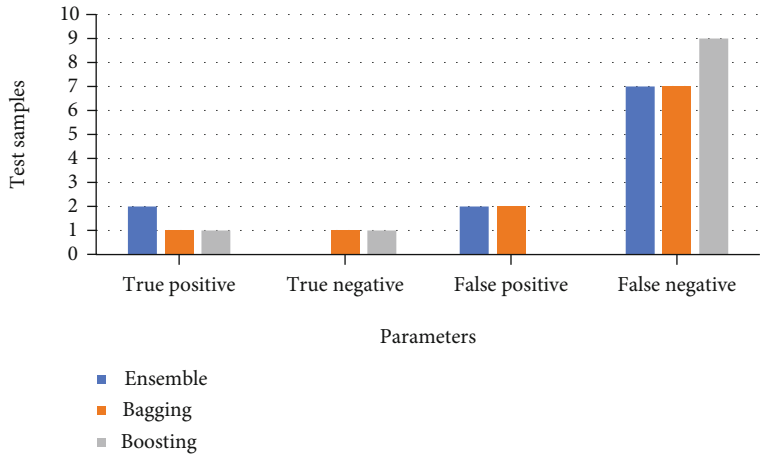

Figure 14: Confusion matrix of various classification algorithms.

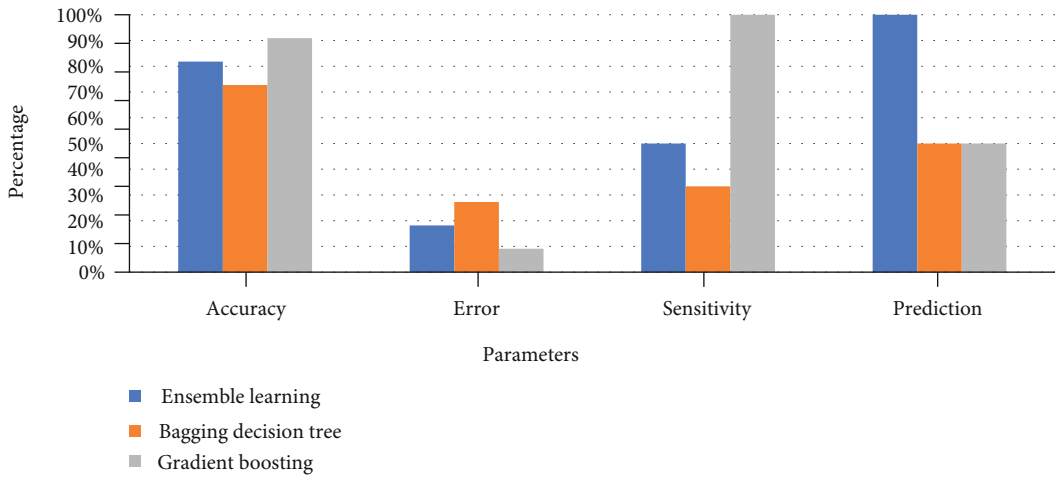

Figure 15: Performance measure of various classification algorithms.

Table 8

Paper name	Lung cancer detection using image segmentation by means of various evolutionary algorithms [34]	Lung cancer detection using image processing and classification techniques
Objective	To find a fast image segmentation algorithm for medical images to reduce the time it takes doctors to evaluate computer tomography (CT) scan images.	(i) Classification of lung cancer using extracted Haralick features (ii)Comparing the accuracy of various image segmentation algorithms
Features used	No features used.	Haralick features like contrast, energy, entropy, homogeneity, etc.
Segmentation also used	k-median, -means, particle swarm optimization, guaranteed convergence particle swarm optimization. Inertia-weighted particle swarm optimization, guaranteed convergence particle swarm optimization.	k-means, fuzzy c-means
Results	The highest accuracy is achieved in guaranteed convergence particle swarm optimization, i.e., 95.81%, and the average accuracy is above 90%.	The highest accuracy is achieved in fuzzy c-means, i.e., 98.78%, and the average accuracy is above 95%.

thresholding, masking, and extraction is shown in Figures 11–13.

The dataset was trained under 8 features and split into 75% for training the model and 25% for testing the model. The classifiers used in ensemble learning are DT, KNN, MLPNN, SVM, and logistic regression, with bagging using decision tree and gradient boosting. The performance measure of ensemble learning, bagging, and gradient boosting represented through a confusion matrix is shown in Table 6, and classification accuracy is compared in Table 7. The comparison of TP, TN, TP, and FP is shown in Figure 14, and a comparison of accuracy, sensitivity, and specificity is shown in Figure 15. Table 7 shows that the accuracy measure of gradient boosting was 90.9% which was found to be the highest.

A comparison between the proposed study and [34] was performed. The analysis was done using the same dataset. Table 8 shows that the proposed work achieved a higher accuracy of 98.78% using Fuzzy c-means.

A comparative study between existing and proposed methods is shown below in Table 8.

By combining two datasets [34, 37, 53, 54] into one, the study provided results that could be generalized. The limitation of this study is that the analysis and modeling are not powerful enough for even larger datasets.

5. Conclusions

In this paper, we performed image detection for lung cancer by combining the different strategies of GLCM texture extraction and ensemble learning for model-building. The first step, before undertaking any statistical analysis, was preprocessing the medical images. The median filter performed the best as shown by the result's superior SSI and SMPI metric values. Afterwards, clustering was implemented to achieve image segmentation for the cancer specimens. The fuzzy c-map clustering algorithm yielded the best results with a maximum accuracy of 98.78% and accuracy across all images of at least 95%. The classification of cancer was performed by implementing ensemble learning, which is the strategy of aggregating multiple models to reach a more generalized consensus. Developing the model also integrated the techniques of maximum voting, bagging, and gradient boosting. Gradient boosting helped improve the accuracy to 90.9%. Overall, the proposed framework achieved very high performance, with 98.78% accuracy in segmentation and 90.9% accuracy in classification. Thus, this proposed framework can assist medical practitioners and augment modern techniques in medical computer-aided diagnosis of lung cancer.

Data Availability

We can send the datasets at the request of the authors.

Ethical Approval

This article does not contain any studies with human participants. No animal studies were involved in this review.

Conflicts of Interest

The authors declare no conflict of interest.

Authors' Contributions

All authors contributed equally to this work. In addition, all authors have read and approved the final manuscript and given their consent to publish the article.

References

[1] A. El-Baz and J. S. Suri, *Lung Imaging and Computer Aided Diagnosis*, CRC Press, 2012.

[2] V. Krishnan, A. Praisy, and M. Shalinie, "A customized particle swarm optimization for classification of multispectral imagery based on feature fusion," *International Arab Journal of Information Technology*, vol. 5, no. 4, pp. 71–78, 2008.

[3] K. Venkatalakshmi, P. Praisy, R. Maragathavalli, and S. Shalinie, "Multispectral image clustering using enhanced genetic k-means algorithm," *Information Technology Journal.*, vol. 6, no. 4, pp. 554–560, 2007.

[4] G. Gupta, "Algorithm for image processing using improved median filter and comparison of mean, median and improved median filter," *International Journal of Soft Computing*, vol. 5, pp. 304–311, 2011, http://ijsce.org/attachments/File/Vol-1_Issue-5/E0234101511.pdf.

[5] B. Rani, A. K. Goel, and R. Kaur, "A modified approach for lung cancer detection using bacterial forging optimization algorithm," *International Journal of Scientific Research Engineering and Technology*, vol. 5, no. 1, 2016.

[6] K. Venkatalakshmi and S. S. Mercy, "Classification of multispectral images using support vector machines based on PSO and K-means clustering," in *Proceedings -2005 International Conference on Intelligent Sensing and Information Processing, ICISIP'05*, pp. 127–133, Chennai, India, 2005.

[7] K. Venkatalakshmi and S. Shalinie, "Multispectral image classification using modified k-means clustering," *Neural Network World*, vol. 17, no. 2, pp. 113–120, 2007.

[8] P. I. Dalatu, "Time complexity of K-means and K-medians clustering algorithms in outliers detection," *Global Journal of Pure and Applied Mathematics*, vol. 12, no. 5, pp. 4405–4418, 2016, http://www.ripublication.com/gjpam.htm.

[9] P. Bhuvaneswari and A. B. Therese, "Detection of cancer in lung with K-NN classification using genetic algorithm. Procedia," *Materials Science*, vol. 10, pp. 433–440, 2015.

[10] X. Wang, L. Ge, and L. Xiaojing, "Evaluation of filters for ENVISAT ASAR speckle suppression in pasture area," *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 7, pp. 341–346, 2012.

[11] M. S. Al-Tarawneh, "Lung cancer detection using image processing techniques," *Leonardo Electronic Journal of Practices and Technologies*, vol. 11, no. 20, pp. 147–158, 2012.

[12] A. K. Mohanty, S. Beberta, and S. K. Lenka, "Classifying benign and malignant mass using GLCM and GLRLM based texture features from mammogram," *International Journal of Engineering Research and Applications (IJERA)*, vol. 1, no. 3, pp. 687–693, 2011.

[13] J. C. Bezdek, R. Ehrlich, and W. Full, "FCM: the fuzzy c-means clustering algorithm," *Computers and Geosciences*, vol. 10, no. 2–3, pp. 191–203, 1984.

[14] H. Rao, X. Shi, A. K. Rodrigue et al., "Feature selection based on artificial bee colony and gradient boosting decision tree," *Applied Soft Computing Journal*, vol. 74, pp. 634–642, 2019.

[15] U. Pastorino, M. Bellomi, C. Landoni et al., "Early lung-cancer detection with spiral CT and positron emission tomography in heavy smokers: 2-year results," *Lancet*, vol. 362, no. 9384, pp. 593–597, 2003.

[16] G. Bastarrika, M. J. García-Velloso, M. D. Lozano et al., "Early lung cancer detection using spiral computed tomography and positron emission tomography," *American Journal of Respiratory and Critical Care Medicine*, vol. 171, no. 12, pp. 1378–1383, 2005.

[17] J. A. Howington, M. G. Blum, A. C. Chang, A. A. Balekian, and S. C. Murthy, "Treatment of stage I and II non-small cell lung cancer: diagnosis and management of lung cancer. 3rd American college of chest physicians evidence-based clinical practice guidelines," *Chest*, vol. 143, 5 Supplement, 2013.

[18] L. Mao, R. H. Hruban, J. O. Boyle, M. Tockman, and D. Sidransky, "Detection of oncogene mutations in sputum precedes diagnosis of lung cancer," *Cancer Research*, vol. 11, no. 5-6, pp. 429-430, 1994.

[19] A. Gajdhane and L. M. Deshpande, "Detection of lung cancer stages on CT scan images by using various image processing techniques," *IOSR Journal of Computer Engineering*, vol. 16, no. 5, pp. 28–35, 2014.

[20] S. K. Anand, "Segmentation coupled textural feature classification for lung tumor prediction," in *International Conference On Communication Control And Computing Technologies*, pp. 518–524, Nagercoil, India, 2010.

[21] M. S. Uzer, N. Yilmaz, and O. Inan, "Feature selection method based on artificial bee colony algorithm and support vector machines for medical datasets classification," *The Scientific World Journal*, vol. 2013, 10 pages, 2013.

[22] W. Sun, X. Huang, T. L. B. Tseng, and W. Qian, "Automatic lung nodule graph cuts segmentation with deep learning false positive reduction," *Medical Imaging 2017: Computer-Aided Diagnosis International Society for Optics and Photonics*, vol. 10134, 2017.

[23] K. Verma, S. B. Kumar, and A. S. Thokec, "An enhancement in adaptive median filter for edge preservation," *Procedia Computer Science*, vol. 48, pp. 29–36, 2015.

[24] S. Makaju, P. W. C. Prasad, A. Alsadoon, A. K. Singh, and A. Elchouemi, "Lung cancer detection using CT scan images," *Procedia Computer Science*, vol. 125, no. 2009, pp. 107–114, 2018.

[25] E. Magdy, N. Zayed, and M. Fakhr, "Automatic classification of normal and cancer lung CT images using multiscale AM-FM features," *International Journal of Biomedical Imaging*, vol. 2015, 7 pages, 2015.

[26] J. Kuruvilla and K. Gunavathi, "Lung cancer classification using neural networks for CT images," *Computer Methods and Programs in Biomedicine*, vol. 113, no. 1, pp. 202–209, 2014.

[27] M. A. Hussain, T. M. Ansari, P. S. Gawas, and N. N. Chowdhury, "Lung cancer detection using artificial neural network & fuzzy clustering," *Ijarcce*, vol. 4, no. 3, pp. 360–363, 2015.

[28] D. P. Tian, "A review on image feature extraction and representation techniques," *International Journal of Multimedia and Ubiquitous Engineering*, vol. 8, no. 4, pp. 385–395, 2013.

[29] Q. Z. Song, L. Zhao, X. K. Luo, and X. C. Dou, "Using deep learning for classification of lung nodules on computed tomography images," *Journal of Healthcare Engineering*, vol. 2017, 7 pages, 2017.

[30] D. B. Larkins and W. Harvey, "Introductory computational science using MATLAB and image processing," *Procedia Computer Science*, vol. 1, no. 1, pp. 913–919, 2010.

[31] O. Grove, A. E. Berglund, M. B. Schabath et al., "Data from: quantitative computed tomographic descriptors associate tumor shape complexity and intratumor heterogeneity with prognosis in lung adenocarcinoma," *PloS One*, vol. 10, no. 3, 2015.

[32] https://wiki.cancerimagingarchive.net/display/Public/LungCT-Diagnosis#19039728024ac253cffe4f7a9fb53e03368d83e3.

[33] Q. Dou, H. Chen, L. Yu, J. Qin, and P. A. Heng, "Multilevel contextual 3-D CNNs for false positive reduction in pulmonary nodule detection," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 7, pp. 1558–1567, 2017.

[34] K. Senthil Kumar, K. Venkatalakshmi, and K. Karthikeyan, "Lung cancer detection using image segmentation by means of various evolutionary algorithms," *Computational and Mathematical Methods in Medicine*, vol. 2019, 16 pages, 2019.

[35] W. Shen, M. Zhou, F. Yang et al., "Multi-crop convolutional neural networks for lung nodule malignancy suspiciousness classification," *Pattern Recognition*, vol. 61, pp. 663–673, 2017.

[36] A. Chaudhary and S. S. Singh, "Lung cancer detection on CT images by using image processing," in *In 2012 International Conference on Computing Sciences*, pp. 142–146, Phagwara, India, 2012.

[37] K. Murphy, B. van Ginneken, A. M. R. Schilham, B. J. de Hoop, H. A. Gietema, and M. Prokop, "A large-scale evaluation of automatic pulmonary nodule detection in chest CT using local image features and k-nearest-neighbour classification," *Medical Image Analysis. Elsevier BV*, vol. 13, no. 5, pp. 757–770, 2009.

[38] E. Dandil, "A computer-aided pipeline for automatic lung cancer classification on computed tomography scans," *Journal of Healthcare Engineering*, vol. 2018, 12 pages, 2018.

[39] N. Kalaivani, N. Manimaran, S. Sophia, and D. D. Devi, "Deep learning based lung cancer detection and classification," in *IOP Conference Series: Materials Science and Engineering. Multimedia Tools and Applications*, Tamil Nadu, India, 2020.

[40] C. F. J. Kuo, C. C. Huang, J. J. Siao et al., "Automatic lung nodule detection system using image processing techniques in computed tomography," *Biomedical Signal Processing and Control*, vol. 56, p. 101659, 2020.

[41] P. Nanglia, A. N. Mahajan, D. S. Rathee, and S. Kumar, "Lung cancer classification using feed forward back propagation neural network for CT images," *International Journal of Medical Engineering and Informatics*, vol. 12, no. 5, pp. 447–456, 2020.

[42] P. M. Shakeel, M. A. Burhanuddin, and M. I. Desa, "Automatic lung cancer detection from CT image using improved deep neural network and ensemble classifier," *Neural Computing and Applications. Springer London*, vol. 6, 2022.

[43] R. Gruetzemacher, A. Gupta, and D. Paradice, "3D deep learning for detecting pulmonary nodules in CT scans," *Journal of the American Medical Informatics Association*, vol. 25, no. 10, pp. 1301–1310, 2018.

[44] S. Krishnamurthy, G. Narasimhan, and U. Rengasamy, "An automatic computerized model for cancerous lung nodule detection from computed tomography images with reduced false positives," *Communications in Computer and Information Science*, vol. 709, pp. 343–355, 2017.

[45] "Optimal deep learning model for classification of lung cancer on CT images," *Future Generation Computer Systems*, vol. 92, no. 1, pp. 374–382, 2019.

[46] F. Alenezi, "Image dehazing based on pixel guided CNN with PAM via graph cut," *CMC-Computers, Materials & Continua*, vol. 71, no. 2, pp. 3425–3443, 2022.

[47] F. Alenezi, A. Armghan, S. N. Mohanty, R. H. Jhaveri, and P. Tiwari, "Block-greedy and CNN based underwater image dehazing for novel depth estimation and optimal ambient light," *Water*, vol. 13, no. 23, p. 3470, 2021.

[48] G. P. Joshi, F. Alenezi, G. Thirumoorthy, A. K. Dutta, and J. You, "Ensemble of deep learning-based multimodal remote sensing image classification model on unmanned aerial vehicle networks," *Mathematics*, vol. 9, no. 22, p. 2984, 2021.

[49] F. Alenezi and K. C. Santosh, "Geometric regularized hopfield neural network for medical image enhancement," *International Journal of Biomedical Imaging*, vol. 2021, Article ID 6664569, 2021.

[50] F. Alenezi and E. Salari, "A fuzzy-based medical image fusion using a combination of maximum selection and Gabor filters," *International Journal of Engineering Science*, vol. 9, pp. 118–129, 2018.

[51] F. S. Alenezi and S. Ganesan, "Geometric-pixel guided single-pass convolution neural network with graph cut for image dehazing," *IEEE Access*, vol. 9, pp. 29380–29391, 2021.

[52] S. Majid, F. Alenezi, S. Masood, M. Ahmad, E. S. Gündüz, and K. Polat, "Attention based CNN model for fire detection and localization in real-world images," *Expert Systems with Applications*, vol. 189, article 116114, 2022.

[53] A. Asuntha et al., "Lung cancer detection using SVM algorithm and optimization techniques," *Journal of Chemical and Pharmaceutical Sciences*, vol. 9, no. 4, pp. 3198–3203, 2016.

[54] S. K. Bandyopadhyay, "Edge detection from CT images of lung," *International Journal of Engineering Science & Advanced Technology*, vol. 2, no. 1, pp. 34–37, 2012.