Automatic lung cancer detection from CT image using improved deep neural network and ensemble classifier

P. Mohamed Shakeel[1] · M. A. Burhanuddin[2] · Mohammad Ishak Desa[1]

Abstract

The development of the computer-aided detection system placed an important role in the clinical analysis for making the decision about the human disease. Among the various disease examination processes, lung cancer needs more attention because it affects both men and women, which leads to increase the mortality rate. Traditional lung cancer prediction techniques failed to manage the accuracy because of low-quality image that affects the segmentation process. So, in this paper new optimized image processing and machine learning technique is introduced to predict the lung cancer. For recognizing lung cancer, non-small cell lung cancer CT scan dataset images are collected. The gathered images are examined by applying the multilevel brightness-preserving approach which effectively examines each pixel, eliminates the noise and also increase the quality of the lung image. From the noise-removed lung CT image, affected region is segmented by using improved deep neural network that segments region in terms of using layers of network and various features are extracted. Then the effective features are selected with the help of hybrid spiral optimization intelligent-generalized rough set approach, and those features are classified using ensemble classifier. The discussed method increases the lung cancer prediction rate which is examined using MATLAB-based results such as logarithmic loss, mean absolute error, precision, recall and F-score.

Keywords Computer-aided detection (CAD) · Improved deep neural network (IDNN) · Hybrid swarm intelligent rough set approach · Ensemble classifier

1 Introduction

In the developing technologies, most of the people affected by genetic problem [1] due to the false mutations completely changes human life style. The false mutation entirely changes the structure and function of DNA. The generated wrong mutated DNA cell replaces old DNA cell that creates the abnormal growth of the DNA cell. The abnormal mutation [2] is happened due to the various external factors such as population air breathing, alcohol habits, chemical gas exposure and so on. Mostly, the abnormal cell (DNA) mutation [3] creates tumors that may be occurred in any places such as lung, skin, breast and brain in human body. Among the several tumors, lung cancer [4] is one of the most affected diseases because of the external factors that generally affect respiratory system. From the study in 2005, the number of deaths is increased up to 159,292 that is increased up to 25% in 2018. From the US [5] report of North American association of central cancer registries, it is declared that 234,030 people are influenced by lung cancer in 2018. Further, the American cancer society conducts the survey [6] in the USA in 2019, 228,150 new peoples are affected by lung cancer in which 111,710 peoples are women and 116,440 are men. From the analysis, 142,670 people died due to the lung cancer (66,020 are women and 76,650 are men). According to the survey, it is finally concluded that lung cancer-affected people ratio is increased gradually in the last 5 years. Based on the analysis, lung cancer is the most common considered diseases in medical field to

✉ P. Mohamed Shakeel
  shakeelji@ieee.org

1 Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka, Melaka, Malaysia

2 Director of UTeM International Centre and Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka, Melaka, Malaysia

🖄 Springer

diagnosis [7] the disease in earlier stage. Normally, the lung cancer is manually predicted with the help of the disease symptoms [8] such as blood coughing, chest pain, shortage of breath, fatigue, weight loss, memory loss, bone fracture, joint pains, headache, neurological problem, bleeding, facial swelling, voice change and change of sputum color. Once the patient has been affected by these technologies, different screening procedures [9] like genetic testing, scopy bronchi, reflex testing, fluid biopsy, biopsy and blood testing have been used continuously for evaluation. From the screening methodologies, national institute for health and care department provides the general guidelines to predict the lung cancer and stages of lung cancer effectively. The discussed screening methodologies successfully examine lung cells and deviations in cells that used to predict the lung cancer, but the prediction accuracy is difficult to sustain. Among the screening methodologies, computerized tomography [10] is one of the effective screening processes that effectively examines the deviations and changes present in human body that is detected by passing X-rays on human body. The 30-min passing of X-rays successfully examines internal organ function, and tissues and affected part details are collected effectively compared to PET and MRI screening process. By using CT images, automatic lung cancer prediction [11] system is created to detect the disease by using several traditional steps such as image noise removal, region segmentation, cancer feature extraction, feature selection and cancer classification [12]. From the discussed step, region segmentation and highlight selection process is one of the crucial roles because the successful prediction-affected region determines the deviation in normal and cancer cell effectively. Moreover, the segmented region helps to extract the meaningful cancer features which reduce the complexity of the system. Then the feature selection process [13] minimizes computation time to predict the cancer that also reduces the involvement of data overfitting. There are several segmentation techniques [14] such as k-means clustering, distributed clustering, canny edge detection, sobel detection, fuzzy c-means, fuzzy k-means clustering, self-organizing map and Hopfield neural network that are used to extract meaningful region from the captured x-ray images. After that, different features are obtained from the extracted region and with the assistance of different feature selection techniques optimal features are selected [15] such as wrapper methods, ant colony approach, particle swarm optimization, genetic algorithm, fireflies and bacterial swarm optimization that are used to select effective features from set of features. These selected lung cancer features are classified using defined classifiers such as K-nearest neighboring, support vector machine and other intelligent classifiers. Although the traditional automatic

system successfully predicts the lung cancer, they still handle the recognition accuracy [16] and also consume more time to process large volume of data. Further the system fails to process minimum quality of CT images that may cause to false lung feature that leads to create more misclassification rate [17]. Then the different authors proposed their opinion about the lung cancer detection process because their thoughts help to get the idea for developing intelligent cancer prediction system. In [18], detecting lung cancer from computer tomography screening process uses different optimization algorithms. During the analysis, process 5 approaches median clustering, mean clustering, particle swarm optimization and convergence particle approaches that are used to examine the tumor present in the lung CT image. The captured CT image noise is removed using adaptive median filter, and the histogram analysis is applied to improve the image appearances. Then the different features are extracted, and the affected regions are identified using above-mentioned algorithm. Thus, the author-introduced system effectively recognizes the lung cancer up to 95.89%. In [19], recognizing CT image lung cancer is with the assistance of the approach to convolution. First, the CT images are collected by stack encoder from the LIDC IDRI dataset. The network extracts various features that are learned using deep neural network. The successful utilization of multiple layers recognizes the abnormal features up to 84.2% of accuracy. Even though several techniques are used, still misclassification and large dimension of data handling are crucial issues. For overcoming above discussed problems in this paper, introduce the intelligent techniques to resolve and improve overall lung cancer detection process. The captured CT lung images are examined continuously using multilevel brightness-preserving approach that eliminates noise from image and enhance quality of the lung image. Due to the importance of segmentation process, in this work enhanced deep neural network approach utilizes multiple layers to extract the affected region effectively. From the derived region, effective and optimized features are selected using hybrid spiral optimization intelligent-generalized rough set approach and those features are classified using ensemble classifier. Finally, the defined intelligent technique-based cancer detection system is developed using MATLAB tool and efficiency of the system is determined using different performance metrics. The rest of the paper is structured as follows on the basis of the above analysis. Section 2 discusses the development procedure of improved deep neural network and ensemble classifier-based lung cancer detection process. Section 3 analyzes the efficiency of improved deep neural network and ensemble classifier and concludes in Sect. 4.

2 Improved deep neural network and ensemble classifier-based lung cancer detection system

2.1 Materials and methods

As discussed in Sect. 1, different authors utilize the optimized techniques and methods to predict the lung cancer. According to their opinions, in this section effective intelligent technique is called improved deep neural network and ensemble classifier to detect the lung cancer effectively. During the cancer prediction process, lung cancer images are gathered from cancer imaging archive (CIA) dataset [20]. The database consists of several lung CT images which are obtained from national cancer institute that correlated with proteomic and genomic clinical data. The dataset consists of 5043 CT images that are gathered from 48 series in which 3500 images are used for training, and remaining 2543 images are used for testing purpose to determine the efficiency of cancer prediction system [21]. Then the sample lung image [22] is depicted in Fig. 1.

According to the discussion, the processing structure of the cancer prediction system is depicted in Fig. 2.

As shown in Fig. 1, lung cancer is identified from lung CT images that are collected from the dataset of the cancer imaging archive (CIA). The images collected are processed by implementing several preceding detection processing steps mentioned in the section below.

2.1.1 CT image preprocessing using multilevel brightness-preserving approach

The first level of the work is to removing noise [22] from the captured CT lung image because image capturing process consists of several unwanted information, radiation processing information and patient details that occupy the captured X-ray image. The unwanted details reduce the efficiency lung cancer prediction system. Then the noise present in the image is eliminated with the help of multilevel brightness-preserving approach that examines each and every pixel present in the captured X-ray images effectively. The introduced method effectively examines each image and its pixel for enhancing the quality of image effectively. During this process, the method enhances the image intensity by computing the mean value [23] of the pixels in the image. Once the image brightness is lower to the neighboring pixels, the pixel value is replaced by utilizing the mean value of pixel. While doing this process, image is divided or partition into different sub-images, each sub-image is analyzed separately and normalizes the image effectively. Considering the input lung CT image I, which is decomposed into two different sub-images according to the image brightness value μ_I, the decomposed images are named as the foreground image and background image that are denoted as, I_f and I_b. From the initialization, the image is represented as follows,

$$I = I_b \cup I_f \tag{1}$$

In Eq. (1), the image pixels are represented as m and n that are further expressed as follows,

$$I_b(m,n) = \{I((m,n)|I(m,n)) < \mu_I, \forall I(m,n) \in I\} \tag{2}$$

The foreground sub-images are represented as follows,

$$I_f(m,n) = \{I((m,n)|I(m,n)) \geq \mu_I, \forall I(m,n) \in I\} \tag{3}$$

Based on Eqs. (2, 3), images are divided according to the pixel mean value. After dividing the images, the sub-images are analyzed continuously, the intensity of the pixel is estimated and if any changes are occurred that is replaced with the help of mean value to improve the
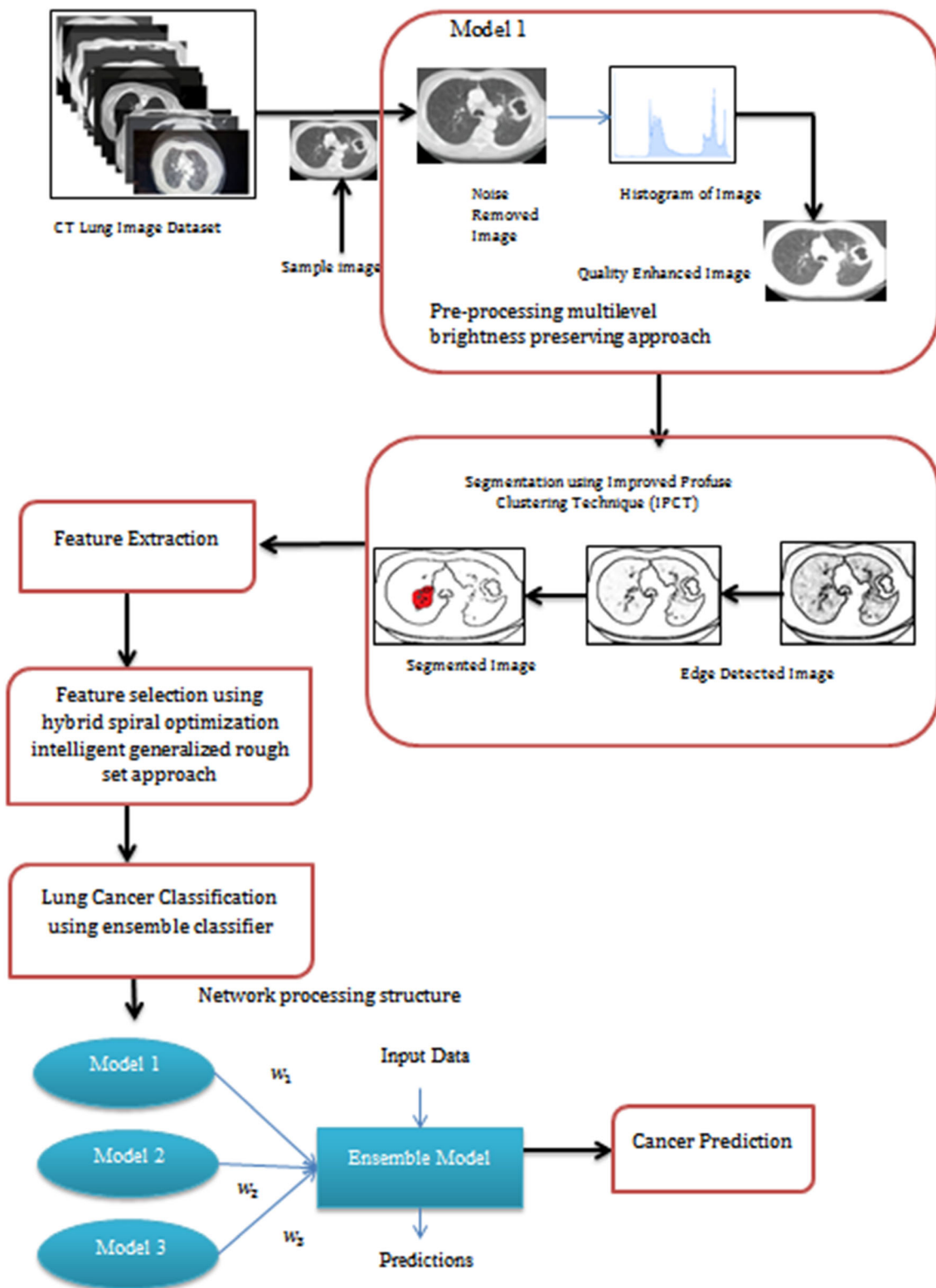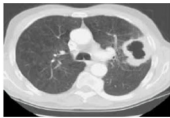
Fig. 1 CT lung images
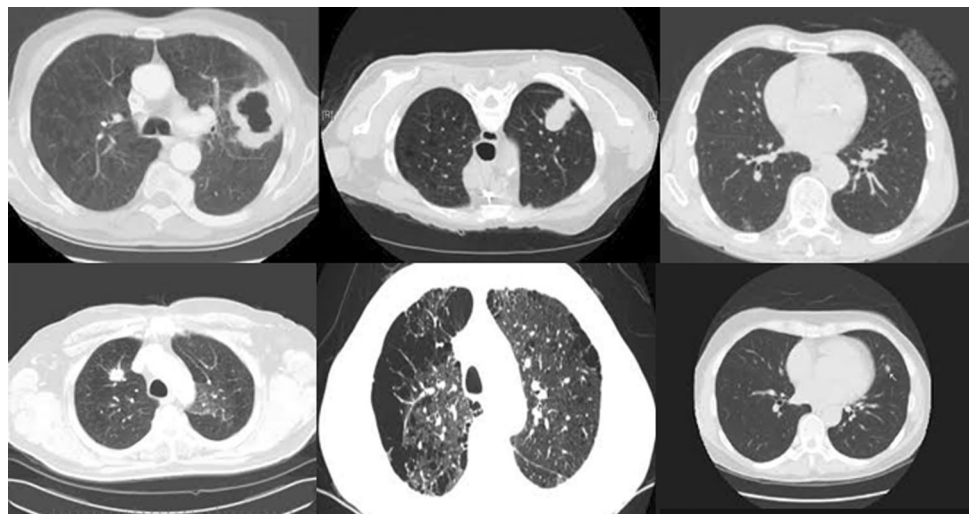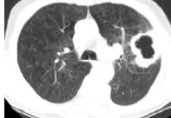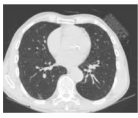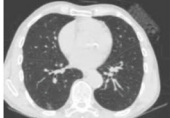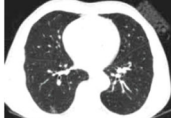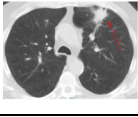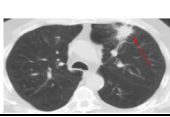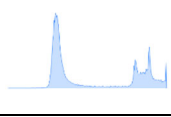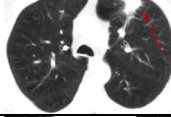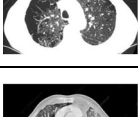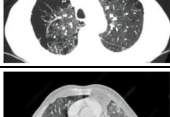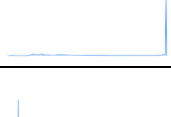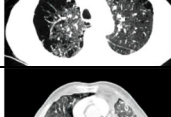
Fig. 2 Improved deep neural network and ensemble classifier-based lung cancer detection system structure

Table 1 Noise-removed CT image

S. no	CT lung image	Noise removed image	Histogram image	Enhanced image
1				
2				
3				
4				
5				

brightness of the image effectively. Then the pixel probability density value [24] is estimated as follows,

$$p(I_k) = \frac{n^k}{n} \tag{4}$$

In Eq. (4), k having value from 0 to $L - 1$ where L is the number of pixels in image, n^k is total times of particular level in image I and n is amount of input lung cancer sample images. After that, cumulative density value is computed as follows,
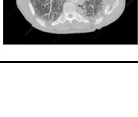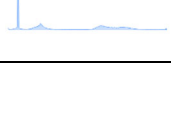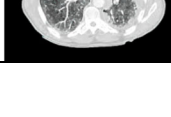
$$C(I) = \sum_{j=0}^{k} p(I_j) \tag{5}$$

The computed probability density and cumulative density functions are mapped with the histogram of the lung CT image. According to the computation, the histogram of the image is determined as follows,

$$f(I) = I_0 + (I_{L-1} - L_0)C(I) \tag{6}$$

In Eq. (6), $f(I)$ is defined as the transformation function of the input image and $C(I)$ is cumulative function of image. Based on the discussion, the output of the preprocessed image is defined as,

$$Y = f(I) = \{(f(I(m, n)|\forall I(m, n) \in I)\} \tag{7}$$

This above process is applied on both foreground and background divided sub-images effectively. This process also gradually improves the quality of the image, eliminating the noise from the captured CT image. Based on the above discussion, the quality-enhanced image eliminated from the noise is shown in Table 1.

Table 1 demonstrates sample noise-removed and quality-enhanced lung cancer CT image details. The successful examination of pixel intensity, density and cumulative density helps to determine the deviation in the image pixel. Based on the analysis, the mean computation and mapping of pixel density and cumulative density value improve the quality of pixel successfully. The noise-removed CT lung images are further examined using effective segmentation techniques which are discussed in the following Sect. 2.1.2.

2.1.2 Lung image segmentation using improved deep neural network (IDNN)

The second step is to extract the affected region from noise-removed lung CT image that is commonly known as segmentation. In this work, the segmentation [25] is done by applying improved deep neural network (IDNN)

Fig. 3 VGG net deep learning network structure
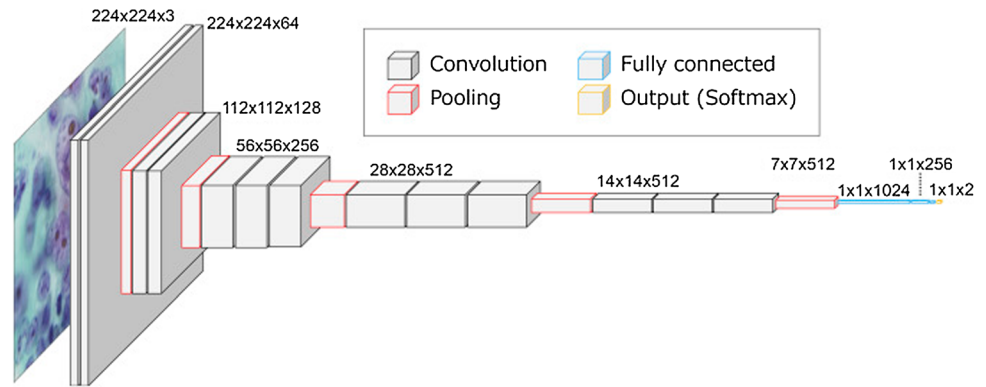
because it utilizes the multiple layers while processing the image. Moreover, the network has large volume data that are collected from previous analysis which helps to identify the affected region successfully that also consumes minimum computation time. During the sematic segmentation analysis, the neural network follows several steps such as pixel classification because it effectively examines each pixel and predicts whole inputs present in the lung CT image. After classifying the pixel, localization step is applied to the image for predicting the more information because this information helps to determine normal and abnormal pixel details. And finally, semantic segmentation in which each pixel is labeled correctly whether the pixel belongs to normal and abnormal region. The deep network [26]-based segmentation process includes several steps such as loading the pre-trained network for making segmentation process, selecting input lung cancer image from the defined lung dataset, pixel-labeled (normal and abnormal) images are load and classes are defined and plot the segmentation images using trained network information. Based on the mentioned steps, initially neural network structure is defined with number of layers; in this segmentation process, multiple hidden layers are used to analyze the input lung CT images. After that, weights value is defined for every node for improving overall segmentation accuracy. According to the discussion, in this work VGG deep learning network structure is used to lung image segmentation process [27]. So, the basic structure of the VGG deep learning network is depicted in Fig. 3.

As shown in Fig. 3, in this work VGG network is used for image segmentation because the network consists of 19 layers in which three layers are fully connected and 16 layers are convolution layers. The large number of intermediate layer helps to predict the exact affected region from noise-removed image effectively. In addition to this, the network having the 3*3 stride and 1 padding with 2 * 2 maxpooling filter is used to improve the model performance. According to the network definition, network consists [28] of several layers such as convolution, ReLU, pooling and fully connected layers. First, the selected lung

images are taken to perform the segmentation process. After that, cross-correlation value of image is analyzed by that and is represented in the matrix format. The input images are considered as 200 * 200 pixels that are processed by defining the kernel value. Further, the ReLu activation function is applied to the network to eliminate the false information from the lung image effectively. In this work,

$$f(x) = \log(1 + e^x) \tag{8}$$

In Eq. (8), $f(x)$ is defined as the activation function of particular image and x is the input pixel value. Along with ReLU value, gradient activation function is further defined as follows,

$$J(\theta) = E_\theta[f(x)] = \int f(x)\pi(x|\theta)\mathrm{d}x \tag{9}$$
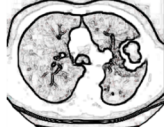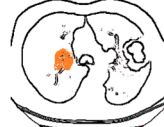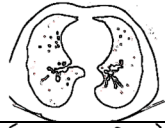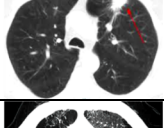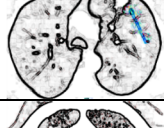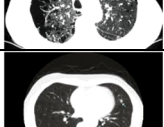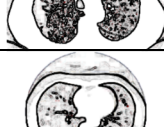
In Eq. (9), search distribution $\pi(x|\theta)$ parameter is defined as θ and $F(x)$ is defined as the maximum fitness value at x. Based on the discussion, the pixel distribution value identification process is further improved and updated as follows,

$$\theta \leftarrow \theta + \eta\nabla_\theta J(\theta) \tag{10}$$

This process effectively trains the incoming lung image that is stored in the testing process and also plots the segmentation effectively. The output of ReLU activation function is applied into pooling layer which successfully analyzes the nonlinear transformation of given input lung pixels. Afterwards, the maxpooling function is applied to get the deviated pixels in the image successfully. Finally, the fully connected layer predicts the affected region by comparing the previous layer activation function with the computed layer output value. This segmentation process [29]applied noise-removed lung image to get the affected region successfully. Based on the discussion, the segmented edge relevant information is shown in Table 2.

Table 2 clearly depicts that introduced method successfully identifies the affected and normal region from input lung image. From the derived region, various features

Table 2 Segmented lung image

S. no	Enhanced image	Edge detected image	Segmented image	
1				
2				
3				
4				
5				

are extracted that mostly helps to predict the abnormal region perfectly.

2.1.3 Lung feature extraction

Third step of the work is to derive the meaningful lung features [30]. During the extraction process, various features such as variance, third-moment skewness, entropy, mean, standard deviation and fourth-moment kurtosis are extracted and are shown in Table 3.

The defined features are extracted from segmented region, in which the each patient has several perceptive of images; these features are derived from entire captured image. Due to the analysis, the dimensionality of the image is improved. So, the dimensionality of dataset needs to be reduced for improving the overall lung cancer recognition process. For this purpose, feature selection approach is applied to get the optimized lung features from the collection of features.

2.1.4 Lung feature selection

Fourth step of the work is lung feature selection [31] which is done with the help of hybrid intelligent spiral optimization-based generalized rough set approach. The introduced method utilizes minimum control variable, fast selection result, local searching process, simple structure,

easiest optimization process and hidden patterns that are easy to identify, easy to make decision effectively, easy to understand and easy to interpret with the results. Due to these advantages, in this work hybrid intelligent spiral optimization-based generalized rough set approach is used to select optimized features from selected features. The spiral optimization algorithm [32] works according to the spiral phenomena which help to resolve the unconstrained optimization problem while selecting features. The method works in the n-dimensional spiral model by using effective setup such as convergence and periodic descent direction setting. With the help of the settings, the method predicts optimization features according to the exploration (global solution) and exploitation (good solution). When selecting optimization process, the method does not have single gradient function instead of this method that utilizes the multiple spiral points [33] which helps to choose current best point. From the collection of points, better solution is selected by processing the features up to reach maximum iteration. Based on the analysis, the algorithm steps of spiral optimization process are mentioned as follows,

Algorithm steps of intelligent spiral model

Initialization: m ($m \geq 2$), k_{max} // number of search points and maximum number of iteration.
Step 1: Initial search points are arranged in the search space as follows,
$x_i(0) \in R^n (i = 1, \dots . m)$ and estimate the center value
$$x^*(0) = x_{ib}(0)$$
$i_b = argmin_{i=1,\dots,m}\{f(x_i(0))\}$ then put k=0.
Step 2: after that rate $r(k)$ rule is decided
Step 3: searching points need to be updated as follows,
$$x_i(k + 1) = x^*(k) + r(k)R(\theta)(x_i(k) - x^*(k))i = 1, \dots . m$$
Step 4: center point of spiral must be updated as follows,
$$x^*(k + 1) = \begin{cases} x_{ib}(k + 1) & (if \ f(x_{ib}(k + 1)) < f(x^*(k))) \\ x^*(k) & otherwise \end{cases}$$
$$i_b = argmin_{i=1,\dots,m}\{f(x_i(k + 1))\}$$
Step 5: then set new k value as,
$$k = k + 1$$
Step 6: check condition $k = k_{max}$ satisfied the process is stopped and backed to step 2.
Output is $x^*(k)$.

According to defined algorithm steps, the search point, center point is defined and output is computed that is considered as selected features. The performance of the searching point needs to be improved using the search settings. The defined settings depend on the rotation matrix $R(\theta)$, rank r(k) and mentioned initial points. Based on the initial setup, the descent direction setting needs to be performed initially. The $R(\theta)$ is defined as follows,

$$R(\theta) = \begin{bmatrix} 0^T_{n-1} & -1 \\ I_{n-1} & 0_{n-1} \end{bmatrix} \tag{11}$$

In above (11) matrix, I_{n-1} is estimated as follows,

$$I_{n-1} = (n - 1) * (n - 1) \tag{12}$$

Equations (11, 12) are defined as the identity matrix and zero vector is computed as follows,

$$0_{n-1} = (n - 1) * 1 \tag{13}$$

After defining this $R(\theta)$ mentioned initial point must be satisfied, Eq. (14), with mentioned condition

$$\min_{i=1,\dots m} \left\{ \max_{j=1,2,\dots m} \left\{ rank\left[d_{j,i}(0), R(\theta)d_{j,i}(0), \dots R(\theta)^{2n-1}d_{j,i}(0)\right]\right\}\right\} \tag{14}$$

In Eq. (14), $d_{j,i}(0)$ is estimated as follows,

$$d_{j,i}(0) = x_j(0) - x_i(0) \tag{15}$$

Then the rank $r(k)$ is computed as follows,

$$r(k) = r = \sqrt[k_{max}]{\delta} \tag{16}$$

$$\delta > 0 \quad and \quad \delta = 1/k_{max} \quad or \quad \delta = 10^{-3} \tag{17}$$

After defining the descent direction setting of searching points, convergence setting need to be defined to improve the performance of searching process. This process is performed until to reach $k_{max} = \infty$ maximum iteration. Further the $r(k)$ is defined as follows,

$$r(k) = \begin{cases} 1 & (k^* \leq k \leq k^* + 2n - 1) \\ h & (k \geq k^* + 2n) \end{cases} \tag{18}$$

The h value present in Eq. (18) is computed as,

$$h = \sqrt[2n]{\delta}, \ \delta \in (0, 1) \delta = 0.5 \tag{19}$$

Based on the above settings, feature selection process is performed continuously in search space for improving the feature selection process. This process chose the global solution from the collection of features. The selected features are arranged in the search space which needs to be analyzed using generalized rough set process to determine whether the selected features accurately help to detect lung cancer effectively. The main benefits of utilizing the rough set [34] are to select the optimized subset from the feature set with effective manner. Also it reduces the number unwanted features and reduces attributes with minimum time. The selected lung features are treated as universe set U, and the rough set apply both lower and upper approximation concept to predict the optimized features effectively. Then the defined approximations are mentioned as follows,

$$\underline{X}\phi = \{X \in U : \phi(x) \neq \phi(y), \quad \text{for all } y \in x^c \tag{20}$$

where $x^c = U - x$ and its upper MSR approximation are defined as

$$\bar{X}\phi = \{X \in U : \phi(x) = \phi(y), \quad \text{for some } y \in x \tag{21}$$

Based on the above process, the successful lung cancer feature is selected which is repeated until to reach the maximum iteration. The selected features are fed into the

next step to classify the normal and abnormal lung features which is explained in the following section.

2.1.5 Lung cancer classification using ensemble classifier

The final step of the work is to classification of lung cancer which is done with the help of the ensemble classifier. The introduced method enhances the overall performance of the machine learning techniques that also maximize the cancer

In addition to this, the applied inputs are analyzed using boosting ensemble classifiers to change the weaker one into strong features [36]. Along with this, weaker features help to fit into the searching space for making the effective decision while classifying the lung features. Then the process of boosting ensemble classifier working process algorithm is defined as follows.

Ensemble classifiers algorithm steps

Step 1: Initialize the input weights $\{w_n\}$ to $1/N$

Step 2: for every inputs m=1 to M and the boosting process is performed as follows,
$\quad$ Classifier $y_m(x)$ need to be minimize the weight value J that is defined as,

Step 3: $\qquad J_m = \sum_{n=1}^{N} w_n^{(m)} 1[y_m(x) \neq t_n]$

Compute the ensemble value of the input as follows,

Step 4: $\epsilon_m = \left. \sum_{n=1}^{N} w_n^{(m)} 1[y_m(x) \neq t_n] \middle/ \sum_{n=1}^{N} w_n^{(m)} \right.$

Step 5: then evaluate the boosting process,

$$\alpha_m = log\left(\frac{1 - \epsilon_m}{\epsilon_m}\right)$$

Step 6: then update the weight value of data as follows,

$$w_n^{(m+1)} = w_n^{(m)} \exp\{\propto_m 1[y_m(x) \neq t_n]\}$$

Step 7: repeat this process continuously to reach end.

---

prediction rate. In addition to this, the ensemble classifiers [35] work with any neural network or machine learning model that used to estimate the output of particular input. The network consists of several layers that use the selected lung features. The network has much number of hidden layers that consume the lung image-related features from previous analysis. These deep learning concepts train the features and stored in the database for comparing process.

After converting the weak classifier into strong classifier, the classification process is performed to get the output value. The output is computed as follows,

$$Y_M(x) = \text{sign}\left(\sum_{m=1}^{M} \propto_m y_m(x)\right) \qquad (22)$$

Based on the above process, the output is computed for given input effectively. This ensemble classifier process minimizes the misclassification rate and also improves the overall cancer recognition rate effectively. At last, the efficiency of the system is analyzed using experimental Results and discussion.

3 Results and discussion

The excellence of improved deep neural network and ensemble classifier-based lung cancer detection system is evaluated in this section. As discussed earlier, during the implementation process system uses the cancer imaging archive (CIA) dataset. From the collected data, 3500 images are used for training features and remaining 2543 images are used as testing one. The divided features are analyzed according to the above discussed methods in Sects. 2.1.1–2.1.5 using MATLAB tool. The developed system successfully segments the affected part by using

Table 3 Spectral features

Features	Formula
Mean	$\mu = \frac{1}{N}\sum_{i=1}^{N} S_i$ N is total amount of pixel present in the segmented region
Standard deviation	$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(S_i - \mu)^2}$
Third-moment skewness	$sk = \left(\frac{1}{N*\sigma^3} * \sum_{i=1}^{N}(s_i - \mu)^3\right)^{1/3}$
Fourth-moment kurtosis	$ku = \left(\frac{1}{N*\sigma^4} * \sum_{i=1}^{N}(s_i - \mu)^4\right)^{1/4}$
Entropy	$\sum_{i,j=0}^{n-1} -\ln(P_{ij})P_{ij}$
Variance	$\sum_{i=0}^{n-1}\sum_{j=0}^{n-1}(i - \mu)^2.p(i,j)$

Table 4 Lung image region derivation efficient

Methods	Accuracy	Specificity	Precision	Recall	F1 score
Region growing (RG)	0.723	0.865	0.873	0.846	0.838
Global threshold (GT)	0.843	**0.972**	0.879	0.875	0.825
Fuzzy c-means (FCM)	**0.89**	0.951	0.875	0.869	0.872
Canny method (CM)	0.878	0.90	0.869	0.877	0.873
Sobel method (SM)	0.882	0.941	0.872	**0.884**	0.87
Watershed approach (WSA)	0.835	0.878	0.88	0.891	**0.894**
Improved deep neural network (IDNN)	**0.962**	**0.984**	**0.974**	**0.98**	**0.984**

Bold values indicate the better results than other filtering methods
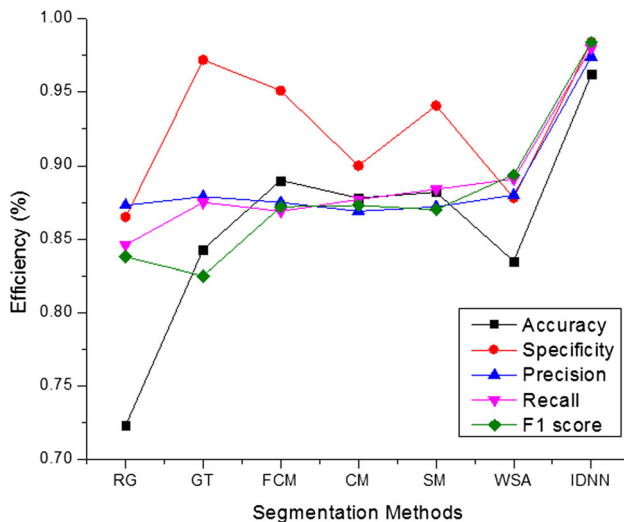
Fig. 4 IDNN segmentation efficiency

multiple layers of neural network. The accuracy of segmented region must be evaluated because the region mostly used to derive the effective features. Then the obtained results are depicted in Table 4. In Table 4, metrics are computed using following equations

$$\text{Accuracy} = (TP + TN)/(TP + TN + FP + FN) * 100\%$$
(23)

$$\text{Specificity} = TN/(TN + FN) * 100\%$$
(24)

$$\text{Precision} = \frac{TP}{TP + FP}$$
(25)

$$\text{Recall} = \frac{TP}{TP + FN}$$
(26)

$$F1 \text{ score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$
(27)

Based on the defined metrics, the accuracy of segmentation is mentioned in Table 4.

As appeared in Table 4, improved deep neural network (IDNN) method attains high segmentation accuracy

(accuracy—96%, specificity—98%, precision—97%, recall—98% and F1-score—98%) compared to the other approaches such as region growing (RG) (accuracy—72%, specificity—86%, precision—87%, recall—84% and F1-score—83%), global threshold (GT) (accuracy—84%, specificity—97%, precision—87%, recall—87% and F1-score—82%), fuzzy c-means (FCM) (accuracy—89%, specificity—95%, precision—87%, recall—86% and F1-score—87%), canny method (CM) (accuracy—87%, specificity—90%, precision—86%, recall—87% and F1-score—87%), sobel method (SM) (accuracy—88%, specificity—94%, precision—87%, recall—88% and F1-score—87%) and watershed approach (WSA) (accuracy—83%, specificity—87%, precision—88%, recall—89% and F1-score—89%). From the obtained result, the pictorial representation is shown in Fig. 4.

Figure 4 depicts that the efficiency of segmentation in which the IDNN approach attains maximum accuracy indicates that introduced method successfully extracts the affected region. From the affected region, different features are extracted which are processed by defined selection approach. The successfully selected features improve overall cancer classification rate. So, the introduced hybrid intelligent spiral optimization-based generalized rough set approach (HSOGR)-selected features are depicted in Table 5.

Table 5 depicts the number of lung features involved in the classification process. In the analysis, totally 50 features are extracted from the lung image in which each method selects the effective number of features for minimizing the computation complexity, in which the hybrid intelligent spiral optimization-based generalized rough set approach (HSOGR) selects totally 14 features by utilizing number of spiral center and settings. Moreover, the approximation criteria used to predict the optimized number of features. The selected features are minimum compared to other selection approaches such as genetic algorithm with wrapper approach (GAWA) [37], particle swarm optimization-based multiobjective selection (PSOMS) [38] and ant colony optimization (ACO) [39]. Even though the HSOGR approach selects low features, it

Table 5 Number of selected features

| Techniques	Number of features	Total no. of selected features
Genetic algorithm with wrapper approach (GAWA)	50	43
Particle swarm optimization-based multiobjective selection (PSOMS)	50	38
Ant colony optimization (ACO)	50	27
Hybrid intelligent spiral optimization-based generalized rough set approach (HSOGR)	50	14

Table 6 Results of the framework and other FS and classification algorithms and efficiency analysis

| Classification technique	FS algorithm	Kulczynski index	Folkes–Mallows index	Rand index	Russel–Rao index	Precision	Recall	F-measure	Accuracy
Naïve Bayes	GAWA	0.832	0.841	0.862	0.851	0.893	0.879	0.828	0.818
	ACO	0.793	0.7833	0.793	0.763	0.796	0.756	0.76	0.786
	PSOMS	0.759	0.729	0.709	0.719	0.725	0.73	0.748	0.738
	HSOGR	0.697	0.656	0.643	0.615	0.662	0.693	0.6581	0.681
IBK	GAWA	0.867	0.873	0.882	0.891	0.91	0.889	0.88	0.889
	ACO	0.81	0.793	0.803	0.823	0.836	0.76	0.786	0.796
	PSOMS	0.772	0.789	0.769	0.789	0.765	0.773	0.768	0.78
	HSOGR	0.701	0.726	0.743	0.715	0.702	0.713	0.721	0.731
J48	GAWA	0.882	0.884	0.882	0.891	0.903	0.889	0.89	0.88
	ACO	0.83	0.81	0.803	0.823	0.836	0.78	0.79	0.80
	PSOMS	0.784	0.792	0.789	0.792	0.772	0.78	0.782	0.793
	HSOGR	0.726	0.732	0.763	0.715	0.714	0.72	0.754	0.762
Random forest	GAWA	0.891	0.893	0.887	0.893	0.913	0.89	0.90	0.902
	ACO	0.834	0.82	0.812	0.834	0.82	0.80	0.82	0.834
	PSOMS	0.79	0.80	0.792	0.80	0.78	0.793	0.792	0.802
	HSOGR	0.735	0.746	0.77	0.743	0.74	0.735	0.743	0.773
JRip	GAWA	0.947	0.932	0.946	0.928	0.945	0.948	0.947	0.94
	ACO	0.927	0.912	0.907	0.927	0.925	0.929	0.928	0.92
	PSOMS	0.887	0.881	0.887	0.889	0.888	0.882	0.883	0.883
	HSOGR	0.875	0.878	0.870	0.872	0.879	0.871	0.871	0.873
Rough set classifier (RSC)	GAWA	0.948	0.94	0.94	0.94	0.956	0.942	0.941	0.943
	ACO	0.938	0.922	0.932	0.939	0.925	0.934	0.938	0.939
	PSOMS	0.917	0.916	0.916	0.917	0.905	0.926	0.914	0.904
	HSOGR	0.89	0.895	0.885	0.883	0.87	0.873	0.881	0.891
Ensemble classifier (EC)	GAWA	0.978	0.972	0.974	0.975	0.972	0.9753	0.9754	0.9761
	ACO	0.952	0.9594	0.952	0.959	0.935	0.944	0.948	0.948
	PSOMS	0.947	0.946	0.947	0.936	0.9252	0.936	0.934	0.932
	HSOGR	0.936	0.925	0.9215	0.925	0.913	0.923	0.921	0.9211

improves the entire cancer recognition process. Based on the discussion, the obtained features selection accuracy is shown in Table 6.

Table 6 depicts the accuracy of various features selection with relevant classifier. From the analysis, obtained results relevant graphical representation is shown in the following Fig. 5.

Figure 5a–g depicts the accuracy, F-measure, recall and precision value of different feature selection approaches such as genetic algorithm with wrapper approach (GAWA), particle swarm optimization-based multiobjective selection (PSOMS) and ant colony optimization (ACO). The selected features are processed with the help of various classifiers such as naïve Bayes (NB), IBK, RF, RJip, RSC, J48 and ensemble classifier to recognize the
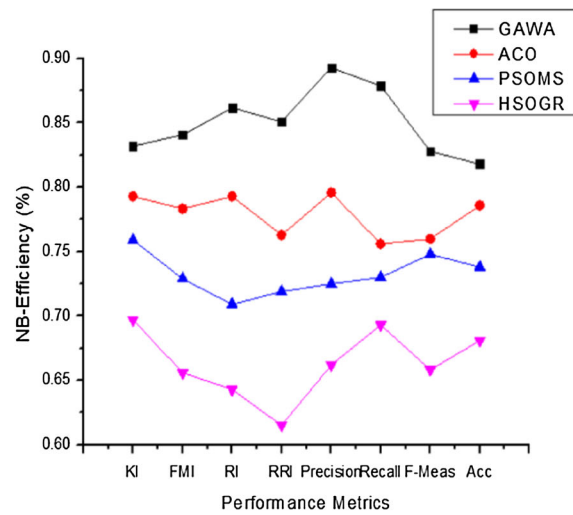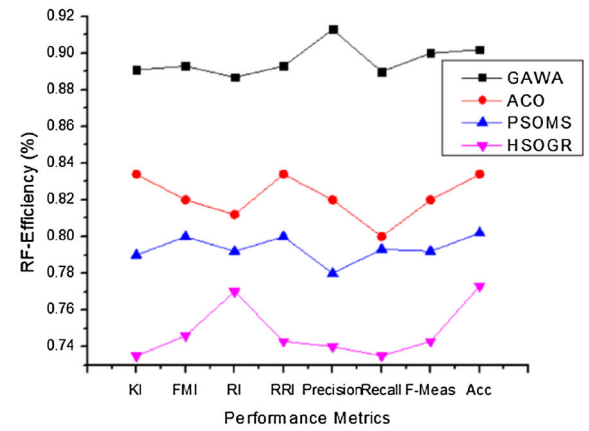
(a) Efficiency of NB classifier
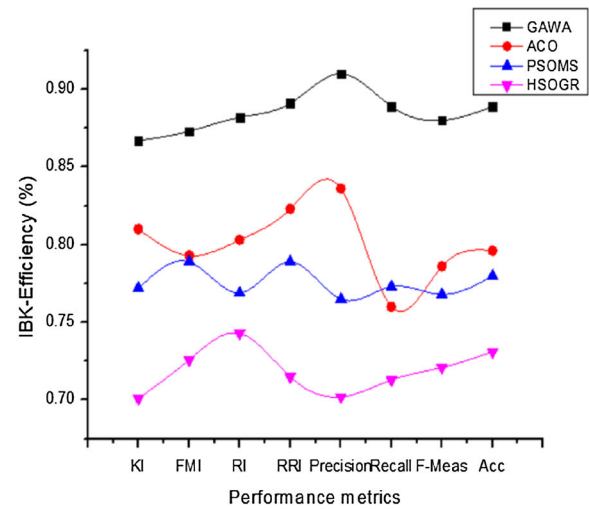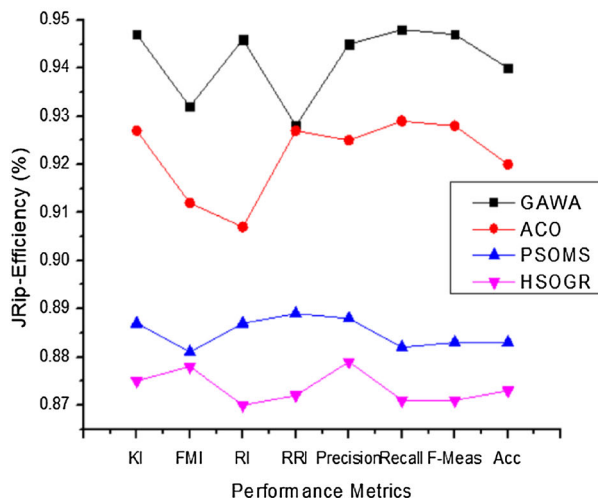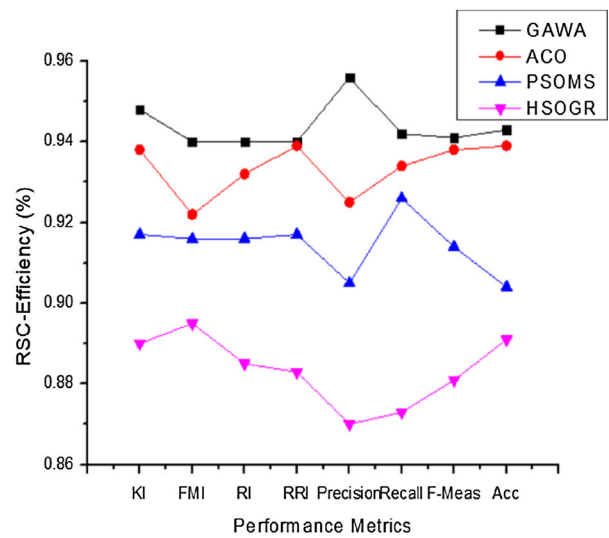

(b) Efficiency of IBK Classifier
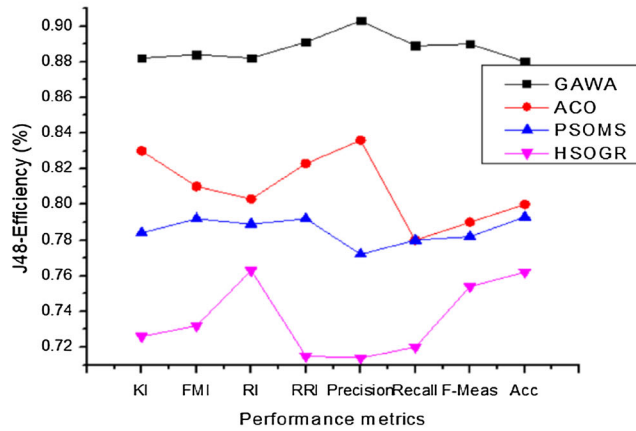
(c) Efficiency of J48 classifier

(d) Efficiency of RF Classifier

(e) Efficiency of RJipclassifier

(f) Efficiency of RSC Classifier
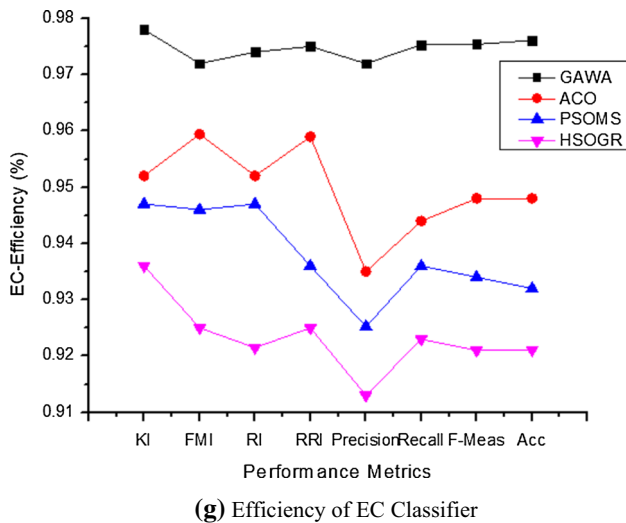
Fig. 5 Efficiency of feature selection versus classifiers

(g) Efficiency of EC Classifier

Fig. 5 continued

lung cancer. According to the analysis, the overall HSOGR with EC attains high accuracy compared to other approaches, and the HSOGR-based selected features attain maximum recognition accuracy with other classifiers. Not only these general metrics, the efficiency of the system is evaluated with some other metrics Kulczynski index (KI), folks mallows index(KMI), rand index (RI) and Russel–Rao index (RRI) of different feature selection approaches such as genetic algorithm with wrapper approach (GAWA), particle swarm optimization-based multiobjective selection (PSOMS) and ant colony optimization (ACO). The selected features are processed with the help of various classifiers such as naïve Bayes (NB), IBK, RF, RJip, RSC, J48 and ensemble classifier to recognize the lung cancer. According to the analysis, the overall HSOGR with EC attains maximum efficiency and minimum deviation compared to other approaches, and the HSOGR-based selected features attain maximum recognition accuracy with other classifiers. Thus, the HSOGR and EC classifier effectively recognize the lung cancer with maximum accuracy compared to other classification methods.

4 Conclusion

Thus, the paper analyzes the lung cancer using improved deep neural network and ensemble classifier. The system collects the cancer image from cancer imaging archive (CIA) dataset and divided the images into testing (2543) and training (3500). Then the collected image intensity level is examined to improve brightness level and eliminate the noise present in CT lung image. After that, each pixel is examined using multiple layer of network for segmenting affected region from lung image. The segmented region is

analyzed effectively, and various features are extracted which are huge in dimension that also consumes more time to recognize cancer. So, the dimensionality of the system is reduced by applying spiral settings and approximation concept that effectively selects optimized features. The features are boosted with the help of ensemble classifier which effectively classifies the abnormal cancer features. The efficiency of the system is evaluated using experimental results, and system recognizes the cancer with maximum accuracy.

Compliance with ethical standards

Conflict of interest The authors report no financial interests or potential conflicts of interest.

References

1. Hong Y, Hong SH, Oh YM et al (2018) Identification of lung cancer specific differentially methylated regions using genome-wide DNA methylation study. Mol Cell Toxicol 14:315. https://doi.org/10.1007/s13273-018-0034-0

2. Nair SS et al (2011) Comparison of methyl-DNA immunoprecipitation (MeDIP) and methyl-CpG binding domain (MBD) protein capture for genome-wide DNA methylation analysis reveal CpGsequence coverage bias. Epigenetics 6:34–44

3. Gaudet F et al (2003) Induction of tumors in mice by genomic hypomethylation. Science 300:489–492

4. Tang M, Xu W, Wang Q, Xiao W, Xu R (2009) Potential of DNMT and its epigenetic regulation for lung cancer therapy. Curr Genomics 10:336–352

5. Liu Z, Wang J, Yuan Z, Zhang B, Gong L, Zhao L, Wang P (2018) Preliminary results about application of intensity-modulated radiotherapy to reduce prophylactic radiation dose in limited-stage small cell lung cancer. J Cancer 9(15):2625–2630. https://doi.org/10.7150/jca.24976

6. Balmelli C, Railic N, Siano M, Feuerlein K, Cathomas R, Cristina V, Güthner C, Zimmermann S, Weidner S, Pless M, Stenner F, Rothschild SI (2018) "Lenvatinib in advanced radioiodine-refractory thyroid cancer: a retrospective analysis of the swiss lenvatinib named patient program. J Cancer 9(2):250–255. https://doi.org/10.7150/jca.22318

7. Manser R, Lethaby A, Irving LB, Stone C, Byrnes G, Abramson MJ, Campbell D (2013) Screening for lung cancer. Cochrane Database of System Rev 6(6):CD001991. https://doi.org/10.1002/14651858.cd001991.pub3

8. Brock MV et al (2008) DNA methylation markers and early recurrence in stage I lung cancer. N Engl J Med 358:1118–1128

9. Wang CC et al (2015) HOXA5 inhibits metastasis via regulating cytoskeletal remodelling and associates with prolonged survival in non-small-cell lung carcinoma. PLoS ONE 10:e0124191

10. Hulbert A, Jusue-Torres, I, Stark A, Chen C, Rodgers K, Lee B, Belinsky SA (2017) Early detection of lung cancer using DNA promoter hypermethylation in plasma and sputum. Clin Cancer Res 23(8):1998–2005

11. Lee HY et al (2015) Differential expression of microRNAs and their target genes in non-small-cell lung cancer. Mol Med Rep 11:2034–2040

12. Manogaran G, Shakeel PM, Hassanein AS, Priyan MK, Gokulnath C (2018) Machine-learning approach based gamma distribution for brain abnormalities detection and data sample imbalance analysis. IEEE Access. https://doi.org/10.1109/ACCESS.2018.2878276

13. Shakeel PM, Tolba A, Al-Makhadmeh Z, Jaber MM (2019) Automatic detection of lung cancer from biomedical data set using discrete AdaBoost optimized ensemble learning generalized neural networks. Neural Comput Appl. https://doi.org/10.1007/s00521-018-03972-2

14. Gambino O, Conti V, Galdino S, Fabio Valenti C, dos Santos WP (2019) Image segmentation techniques for healthcare systems. J Healthc Eng 2019:2723419. https://doi.org/10.1155/2019/2723419

15. Pratiwi AI, Adiwijaya (2018) On the feature selection and classification based on information gain for document sentiment analysis. Appl Comput Intell Soft Comput 2018:1407817. https://doi.org/10.1155/2018/1407817

16. Sridhar KP, Baskar S, Shakeel PM, Dhulipala VS (2018) Developing brain abnormality recognize system using multi-objective pattern producing neural network. J Ambient Intell Humaniz Comput. https://doi.org/10.1007/s12652-018-1058-y

17. Shakeel PM, Tobely TEE, Al-Feel H, Manogaran G, Baskar S (2019) Neural network based brain tumor detection using wireless infrared imaging sensor. IEEE Access 7:5577–5588

18. Senthil Kumar K, Venkatalakshmi K, Karthikeyan K (2019) Lung cancer detection using image segmentation by means of various evolutionary algorithms. Comput Math Methods Med 2019:4909846. https://doi.org/10.1155/2019/4909846

19. Song QZ, Zhao L, Luo XK, Dou XC (2017) Using deep learning for classification of lung nodules on computed tomography images. J Healthc Eng 7:8314740. https://doi.org/10.1155/2017/8314740

20. National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (CPTAC) (2018) Radiology data from the clinical proteomic tumor analysis consortium lung squamous cell carcinoma [CPTAC-LSCC] collection [data set]. Cancer Imaging Arch. https://doi.org/10.7937/k9/tcia.2018.6emub5l2

21. Bhuvaneswari P, Therese AB (2014) Detection of cancer in lung with K-NN classification using genetic algorithm. In: International conference on nanomaterials and technologies, vol 10, pp 433–440

22. Venkatalakshmi K, Mercyshalinie S (2005) Classification of multispectral images using support vector machines based on PSO and k-means clustering. In: Proceedings of IEEE international conference on intelligent sensing and information processing, pp 127–133, Bangalore, India, Dec 2005

23. Zhang X, Wang S (2012) Efficient data hiding with histogram preserving property. Telecommun Syst 49:179–185

24. Sengee N, Choi H (2015) A novel filter ed Bi-histogram equalization method. J Korea Multimed Soc 18(6):691–700

25. Raoof K, Kamoona K, Budayan C (2019) Implementation of genetic algorithm integrated with the deep neural network for estimating at completion simulation. Adv Civ Eng. https://doi.org/10.1155/2019/7081073

26. Kong Z, Li T, Luo J, Xu S (2019) Automatic tissue image segmentation based on image processing and deep learning. J Healthc Eng. https://doi.org/10.1155/2019/2912458

27. Havaei M, Davy A, Warde-Farley D et al (2017) Brain tumor segmentation with deep neural networks. Med Image Anal 35:18–31

28. Akkus Z, Galimzianova A, Hoogi A, Rubin DL, Erickson BJ (2017) Deep learning for brain MRI segmentation: state of the art and future directions. J Digit Imaging 30:449–459

29. Srhoj-Egekher V, Benders MJ, Viergever MA, Isgum I (2013) Automatic neonatal brain tissue segmentation with MRI. In: Proceedings of SPIE medical imaging, international society for optics and photonics, Bellingham, WA, USA, 2013

30. Farabet C, Couprie C, Najman L, LeCun Y (2013) Learning hierarchical features for scene labeling. IEEE Trans Pattern Anal Mach Intell 35(8):1915–1929

31. Golmohammadi D, Creese RC, Valian H, Kolassa J (2009) Supplier selection based on a neural network model using genetic algorithm. IEEE Trans Neural Netw 20(9):1504–1519

32. Tsai CW, Huang BC, Chiang MC (2014) A novel spiral optimization for clustering. In: Park J, Adeli H, Park N, Woungang I (eds) Mobile, ubiquitous, and intelligent computing. Lecture notes in electrical engineering, vol 274. Springer, Berlin

33. Tamura K, Yasuda K (2011) Spiral multipoint search for global optimization. In: International conference on machine learning and applications, vol 1, pp 470–475

34. Zeng K, Jing S (2018) Kernel neighborhood rough sets model and its application. Complexity. https://doi.org/10.1155/2018/1342562

35. Akhand MAH, Murase K (2007) Neural network ensemble training by sequential interaction. In: de Sá JM, Alexandre LA, Duch W, Mandic D (eds) Artificial neural networks: ICANN 2007. ICANN 2007. Lecture notes in computer science, vol 4668. Springer, Berlin

36. Tsymbal A, Pechenizkiy M, Cunningham P (2005) Diversity in search strategies for ensemble feature selection. Inf Fusion 6:83–98

37. Bouaguel W (2016) A new approach for wrapper feature selection using genetic algorithm for big data. In: Lavangnananda K, Phon-Amnuaisuk S, Engchuan W, Chan J (eds) Intelligent and evolutionary systems. Proceedings in adaptation, learning and optimization, vol 5. Springer, Cham

38. Xue B, Cervante L, Shang L, Zhang M (2012) A particle swarm optimisation based multi-objective filter approach to feature selection for classification. In: Anthony P, Ishizuka M, Lukose D (eds) PRICAI 2012: trends in artificial intelligence. PRICAI 2012. Lecture notes in computer science, vol 7458. Springer, Berlin

39. Kanan HR, Faez K, Taheri SM (2007) Feature selection using ant colony optimization (ACO): a new method and comparative study in the application of face recognition system. In: Perner P (ed) Advances in data mining. Theoretical aspects and applications. ICDM 2007. Lecture notes in computer science, vol 4597. Springer, Berlin