Contents lists available at ScienceDirect

Healthcare Analytics

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

An intelligent algorithm for lung cancer diagnosis using extracted features from Computerized Tomography images

Negar Maleki [a], Seyed Taghi Akhavan Niaki [b,*]

[a] *School of Information Systems and Management, Muma College of Business, University of South Florida, Tampa, FL, USA*

[b] *Department of Industrial Engineering, Sharif University of Technology, PO Box 11155-9414 Azadi Ave., Tehran 1458889694, Iran*

ARTICLE INFO

Keywords:
Machine learning
Dimension reduction
Genetic algorithm
Lung cancer
Image processing
Feature extraction

ABSTRACT

According to the World Health Organization, lung cancer is a leading cause of death worldwide. This research aims to process the Computerized Tomography (CT) images of lung cancer patients for the early diagnosis of the disease. The images are processed using Convolutional Neural Network (CNN) in the first approach, where Artificial Neural Network (ANN) is employed to classify the images. In the second approach, the images are pre-processed and segmented before utilizing CNN and ANN. In the third method, all the pre-processed and segmented images are converted to numerical data via specific feature extraction algorithms in the last step. Besides, dimensional reduction and feature selection algorithms are employed to classify with three machine learning techniques, i.e., Gradient Boosting (GB), Random Forest (RF), and Support Vector Machine (SVM). An extensive comparative analysis is made to come up with the best technique. The comparisons are made by evaluating the methodologies on a set of lung CT scan images collected from a medical center. The results show that when either SVM or RF classification techniques are used, a 95% accuracy is obtained in diagnosing lung cancer.

1. Introduction

According to the World Health Organization (WHO), cancer is a leading cause of death worldwide. In 2020, nearly 10 million people died of various cancers. Financially, about 70% of cancer deaths occur in low and middle-income countries. The economic impact of cancer is significant and increasing, with the total annual cost of cancer treatment in 2010 being about $1.16 trillion. According to WHO statistics, the most common causes of cancer death in 2020 were lung (1.8 million deaths), colon and rectum (935,000 deaths), liver (830,000 deaths), stomach (769,000 deaths), and breast (686,000 deaths) [1].

Given that lung cancer is at the top of this list, we tried to review what causes the disease to get a checklist of the factors that affect it. Although diagnosing the disease in its early stages is complicated, its symptoms are similar to respiratory infections, even though there may be no symptoms at first. Although the disease can affect anyone, lung cancer is more likely to occur in smokers. Diagnosis in the early stages can save a patient's life, as lung cancer cells can spread to other organs before a doctor can diagnose them. Cancer metastasis makes treatment much more difficult.

Research on lung cancer cases claims that smoking is the most crucial cause of this disease, which is more common in women today than in the past. In history, due to the lower consumption of women's cigarettes, a lower incidence rate of this type of disease was recorded than in men [2,3]. Other contributing factors include age, gender, race, socioeconomic status, exposure to occupational and environmental factors, chronic lung disease, air pollution, individual genetics, obesity, exposure to secondhand smoke, dry cough, alcohol consumption, and diet. Even people's lifestyles help to spread the disease [2,4]. Considering these, one can take a big step toward the early detection of this disease.

With increased health data, management, analysis, and decision-making have become very challenging in recent years. Moreover, as the population increases, the medical community faces many problems dealing with and diagnosing various diseases. Thus, conducting experiments imposes enormous costs on the relevant organizations [5].

Given the sheer volume of data and the various occupations of physicians, the possibility of errors in their decisions is very high. Thus, data mining algorithms will greatly help the medical community and patients. However, it should be noted that these methods confirm the doctor's opinion and have little reliability alone [6]. Many tests to diagnose disease have devastating effects on the patient's body and cost a lot of money, which can be a solid reason to use data mining techniques to diagnose the disease [7]. Besides, data mining methods allow hidden inter-dependencies between the data, which sometimes take years to make through classical methods.

While an increasing body of work in the literature is investigating lung cancer, this paper is distinguished, firstly, as the Computerized

[*] Corresponding author.
E-mail addresses: negarmaleki@usf.edu (N. Maleki), Niaki@Sharif.edu, niaki@sharif.edu (S.T.A. Niaki).

https://doi.org/10.1016/j.health.2023.100150
Received 5 December 2022; Received in revised form 9 February 2023; Accepted 17 February 2023

Tomography (CT) scan images are collected from a hospital in Tehran, Iran considering a critical difference. In previous studies [8–13], researchers used healthy lung versus cancerous lung CT images to diagnose lung cancer. However, we gather our data from a different perspective. In our dataset, CT scan images are divided into two main categories: Cancerous CT images and noncancerous ones. As the name of the cancerous category shows, patients who struggle with lung cancer are placed in this category. However, the noncancerous class includes ***healthy lungs*** and ***lungs with other diseases***, e.g., COVID-19, except lung cancer.

Second, we implement three approaches to determine which would help physicians diagnose the disease early. The images are processed using Convolutional Neural Network (CNN) in the first approach, where Artificial Neural Network (ANN) is employed to classify the images. In the second approach, the images are pre-processed and segmented before utilizing CNN and ANN. In the third method, all the pre-processed and segmented images are converted to numerical data via specific feature extraction algorithms in the last step. Besides, dimensional reduction and feature selection algorithms are employed to classify with three machine learning techniques, i.e., Gradient Boosting (GB), Random Forest (RF), and Support Vector Machine (SVM). To our knowledge, this is the first time CT images are converted to numerical features, and dimensional reduction and feature selection are applied to them. The results show that our proposed framework outperforms and reaches 95% accuracy in diagnosing lung cancer.

The rest of this paper is organized as follows. Section 2 will review what has been done so far to diagnose diseases by machine learning algorithms to achieve the importance of using this field in medical science. Section 3 explains three methodologies and their phases in detail. In Section 4, the methods are implemented step by step to illustrate their results. Methods' comparison, methods' sensitivity analysis, and strengths and weaknesses of each technique will be discussed in Section 5. Finally, the best approach will be introduced, and future works will be presented in Section 6.

2. Literature review

Previous research works in this field are divided into two subsections for better insight into the difference between machine learning and deep learning algorithms on lung and other cancers. Moreover, it shows the importance of data mining in medical science and determining the research gap.

2.1. Cancer diagnosis using machine learning algorithms

Maleki et al. [14] used the k-Nearest Neighbor (kNN) algorithm on the lung cancer dataset. They applied feature selection on the dataset and came up with the six most essential features among the dataset features. This paper uses the same feature selection algorithms in our proposed framework to develop the important features [14,15]. While an increasing body of work applied SVM and decision tree algorithms to diagnose different types of cancer [5,16–19]; there is little work used Gradient Boosting algorithm for this purpose. To this end, we use two traditional algorithms (SVM and Decision Tree) in literature along with GB in our framework.

The healthcare system is quite different from other industries, so it has a higher priority and customers in this area. In other words, regardless of its costs, patients expect the highest level of treatment and service. Since machine and deep learning have been successful in different areas and have provided precise solutions, they are considered fundamental methods for solving health problems [20].

2.2. Cancer diagnosis using deep learning algorithms

Rapid progress in using machine learning algorithms in medicine has been seen in the literature. Compared to deep learning algorithms, machine learning algorithms are time-consuming and require expert knowledge to adjust their characteristics. However, deep learning algorithms can capture raw data, automatically adjust the features, and analyze and learn the data more quickly [20]. Although deep learning algorithms are powerful, they need an adequate number of data to show their power. While there is no specific rule to say how much data is enough to achieve good results, a rule of thumb says the more data, the better result. Looking at the literature shows that depending on the data source, researchers utilized different amounts of data in their analysis. For example, KL et al. [21] used 100 CT images from an online source, Chaunzwa et al. [22] employed 331 CT images from Massachusetts General Hospital (MGH), Khan et al. [23] utilized 2101 CT images from Kaggle website, and Toğaçar et al. [24] used 100 CT images from Cancer Imaging Archive (CIA). Therefore, collecting real-life data is not an easy task, and it requires lots of effort. In this paper, we collect 364 real-life CT images from a hospital in Iran, so comparing the number of images we analyzed with previous works concerning the data source, we would say, we used a medium to large-scale data in our analysis.

Radiologists take pulmonary CT images to diagnose and evaluate tumor growth. Therefore, the visual interpretation of these data leads to the tumor's identification in the final stages of tumor growth. Treatment at this stage only increases the mortality rate in this type of cancer. As a result, diagnosing the tumor in the early stages of the disease is vital, which can be done by analyzing the images [12]. In [8], 1018 images of pulmonary CT were examined. In this study, researchers used a convolutional neural network method without processing or segmenting images, with a sensitivity of 78.9% with twenty false positives per scan and 71.2% with ten false positives per scan. Therefore, in this paper, we considered pursuing this approach and comparing the result with our framework. The result shows that our framework reaches 95% accuracy.

There is a large body of work using a combination of CNN with other machine learning or deep learning algorithms. For instance, Bonavita et al. [25] and Moitra and Mandal [26] used CNN alone to detect lung cancer, Saleh et al. [27] and Nanglia et al. [28] employed CNN and SVM, Onishi et al. [29] and Huang et al. [30] applied Deep CNN (DCNN) to detect or classify lung cancer in Ct images. Hence, CNNs play a massive role in detecting or classifying lung cancer. Based on the previous studies' limitations, there are still opportunities to develop a combination of CNN with other machine learning or deep learning algorithms. In this paper, we use a combination of CNN with ANN as a deep learning method and GB, RF, and SVM as machine learning methods.

Despite problems in the image segmentation phase, many researchers [9–11,13,31] used image segmentation after pre-processing step to identify the touching objects in the image. One of these segmentation algorithms is watershed segmentation which was used in [32] study after reviewing many different approaches. Therefore, in this study, we use the watershed segmentation algorithm to identify the most critical objects in the CT images.

According to the literature shown in Table 1, a small number of studies used pre-processing methods and data dimensional reduction. However, using these techniques can significantly improve the efficiency of diagnostic and predictive algorithms. Moreover, powerful algorithms such as GB or RF have not been used for classifying diseases. However, these methods are potent and may produce better results.

One of the most critical gaps in the literature is the inability to compare the methods used to diagnose lung cancer. The most reliable way to compare the methods is to use the same data. Therefore, this paper compares three different methods and shows that our proposed framework outperforms the other two in diagnosing lung cancer. In the first method, unprocessed CT images are considered input to the CNN and ANN structural architecture. In the second method, the same tasks are applied with the difference that pre-processed and segmented CT images are used as input. Finally, in the third method, the proposed

Table 1
The comparison table of literature review.

Authors	Year	Pre-processing	Feature selection	Dimensional reduction		Techniques												Disease	Data type
			GA	PCA	LDA	KNN	NBs	C4.5	Decision tree	Random forest	Gradient boosting	SVM	K-means	GA - SVM	Deep learning	ANN	CNN		
Chen and Yang [16]	2013													✓				Breast Cancer	Numerical
Zheng et al. [19]	2013											✓	✓					Breast Cancer	Numerical
Odajima and Pawlovsky [33]	2014					✓												Breast Cancer	Numerical
Lynch et al. [5]	2017								✓			✓						Lung Cancer	Numerical
Septiani et al. [34]	2017					✓	✓											Breast Cancer	Numerical
Cherif [17]	2018					✓	✓					✓				✓		Breast Cancer	Numerical
Kr and Aradhya [18]	2018						✓	✓				✓						Lung Cancer	Numerical
Maleki et al. [14]	2021		✓			✓												Lung Cancer	Numerical
Zayed and Elnemr [13]	2015														✓			Breast Cancer	Image
Miah and Yousuf [10]	2015	✓													✓	✓		Lung Cancer	Image
Golan et al. [8]	2016														✓		✓	Lung Cancer	Image
Kaucha et al. [9]	2017	✓										✓	✓		✓			Lung Cancer	Image
Makaju et al. [32]	2018	✓										✓			✓			Lung Cancer	Image
Shakeel et al. [11]	2019	✓													✓	✓		Lung Cancer	Image
Onishi et al. [29]	2020	✓													✓		✓	Lung Cancer	Image
Saleh et al. [27]	2021	✓										✓					✓	Lung Cancer	Image
Nanglia et al. [28]	2021	✓										✓					✓	Lung Cancer	Image
Huang et al. [30]	2022	✓													✓		✓	Lung Cancer	Image
This paper		✓	✓	✓	✓					✓	✓	✓			✓	✓	✓	Lung Cancer	Image/ Numerical

framework, we extract numerical data with the help of 40 features from each pixel in segmented CT images to form a data frame to apply two different dimensional reduction algorithms (Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA)), feature selection algorithm (Genetic Algorithm (GA)), and machine learning algorithms (GB, RF, and SVM).

3. The proposed lung cancer diagnosis methodologies

The methodology adopted in this paper is carried out in three different methods, shown in Fig. 1. This study's problem is diagnosing whether the patient has cancer in the early stage. As is clear from the research purpose, the target variable is defined as discrete, so we need to use classification algorithms to identify the target variable.

According to Fig. 1, the first method comprises one fundamental building phase called image classification. It means, in this method, the raw CT images were given to CNN followed by ANN without any preprocessing (Raw CT images went through the third phase – the blue rectangular – immediately). The second method includes three primary building phases: image pre-processing, image segmentation, and image classification (According to Fig. 1, Raw CT images went through first (the yellow rectangular), second (the green rectangular), and third (the blue rectangular) phases, respectively). Finally, the third method comprises seven fundamental phases: image pre-processing, image segmentation, image feature extraction, building a numerical dataset, dimensional reduction, feature selection, and classification (According to Fig. 1, Raw CT images went through the first to eighth phases, respectively). Although these methods have fundamental phases in common, they are entirely different methods implemented on the same lung CT scan images.

The pre-processing image phase of the study itself is composed of two parts: image resizing and image denoising. Initially, raw lung CT images are resized, and subsequently, the median filter is applied to denoise them. The watershed segmentation algorithm identified the most critical objects in the CT images in the image segmentation phase to make the following steps more reliable. In the image classification phase, raw CT images in method one and segmented CT images in method two are used as input. The CNN and ANN algorithms are applied to the CT images to classify whether the images belong to a cancerous or noncancerous patient.

So far, the fundamental phases that are implemented in methods one and two are described. In the third method, the image feature extraction phase is applied to segmented lung CT images to extract possible numerical features from each image's pixels. The extracted statistical data for each image is stored in a dataset, so building a numerical dataset phase is completed. We obtained a vast dataset as we extracted any possible mathematical features from the CT images to improve diagnosing cancerous patients from noncancerous ones. Therefore, the dimensional reduction phase, composed of two different algorithms: PCA and LDA, is utilized to reduce dataset dimensions and produce two separate datasets, one from the PCA algorithm and the other by the LDA algorithm. From now on, the experiment is continued in two parts. In part one, classification algorithms – GB, RF, and SVM – are applied to both LDA and PCA datasets separately to classify the data into two groups based on extracted features. In part two, the genetic algorithm is initially applied in the feature selection phase to select the essential components in the PCA dataset. Subsequently, the same classification algorithms are used to classify the data into two groups based on the selected features. In the following subsections, we will describe each fundamental phase in detail.

3.1. Image pre-processing

The term pre-processing belongs to a series of tasks needed for enhancing the quality of raw images, increasing the performance of the subsequent phases such as image segmentation, image feature extraction, and classification. The primary objectives in this study phase are to apply image resizing and denoising.

3.1.1. Image resizing

In the image resizing step, pixels are either added to the image or removed. Since medical images have many details that may be effective, no pixel removal is performed in the current research. On the other hand, the CNN algorithm's input images should preferably be in square sizes for a better diagnostic process. Therefore, all images used in this study have dimensions of 512 × 512 pixels.

3.1.2. Image denoising

Image denoising is a process of applying filter(s) to reduce image noise. It should be noted that CT images have the lowest noise level among the medical images, and it is practically unnecessary to perform this step. In any case, to prevent image distortion, the median filter is applied to cancel any possible noise of lung CT.

3.2. Image segmentation

The image segmentation tries to help the algorithm to diagnose as best it can by removing unnecessary parts. In fact, at this point, the algorithm can only focus on the lung region, which will improve the classification performance. Watershed segmentation is used in the

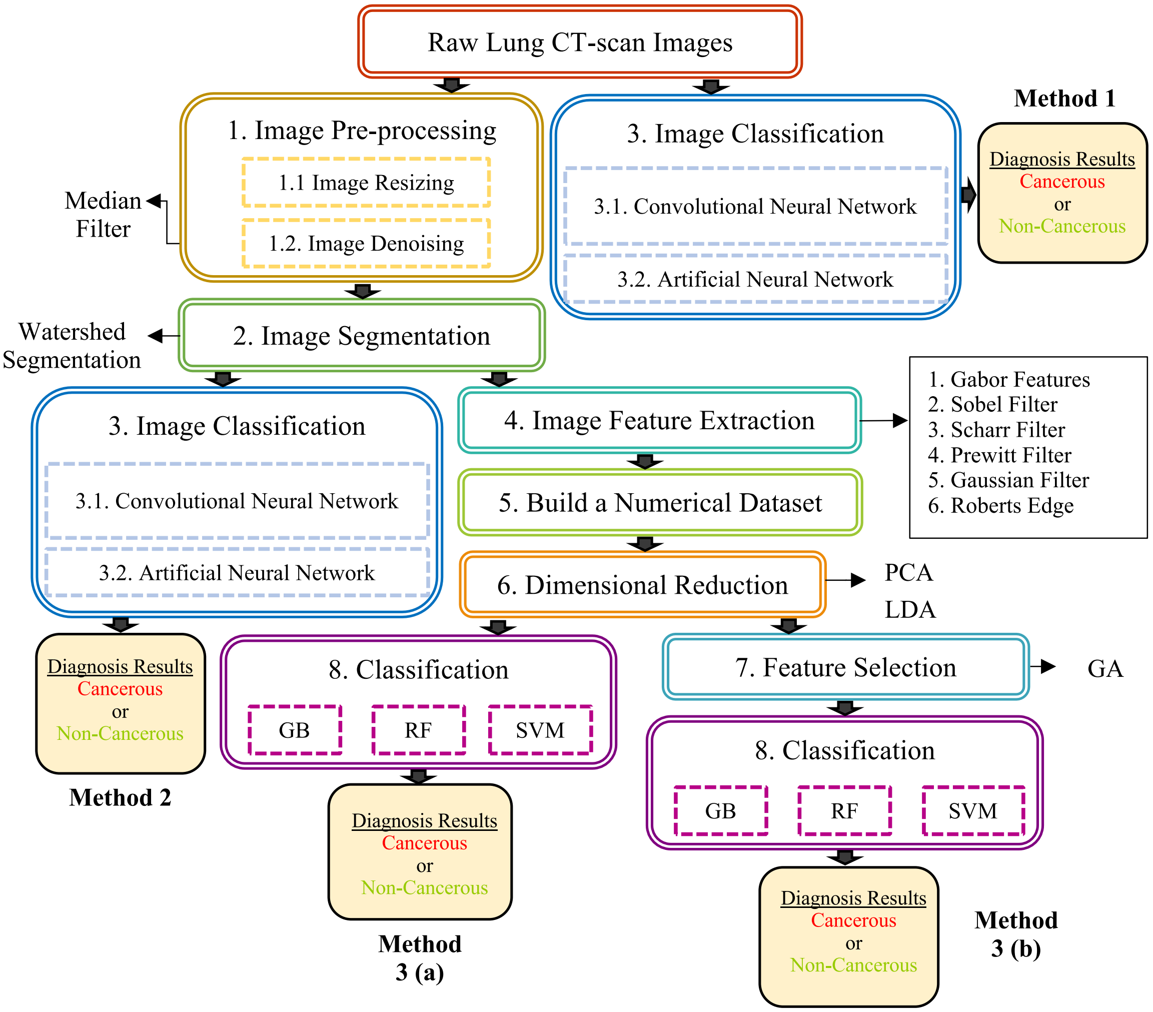


Fig. 1. The proposed framework, in comparison to other methods. . (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

suggested model [32]. The watershed technique is utilized when segmenting complicated pictures since basic thresholding and contour detection will not produce accurate results. The watershed method is built on capturing specific background and foreground information. Markers are then used to run watersheds and determine the precise borders. Markers can be defined by users, e.g., manually, or defined by some algorithms, e.g., thresholding operation — we used thresholding operation in our analysis.

3.3. *Image classification*

Image classification is the primary domain in which deep neural networks play the most critical role in medical image analysis. The image classification accepts the given input images and produces output classification to identify whether the disease is present [35]. The image classification phase is composed of two parts: CNN and ANN.

3.4. *Image feature extraction*

The following phases are performed in the third proposed method. After the raw CT scan images are processed and segmented, several numerical features are extracted from the images in the image feature extraction phase. Each feature will be obtained from each pixel in a single image and then stored in a dataset.

3.4.1. *Gabor features*

The Gabor feature is a linear filter used to analyze tissue in image processing. It explains whether the content of a particular frequency in the image is in specific directions in a local area around the point [36].

3.4.2. *Sobel filter*

The Sobel operator is used in image processing and computer vision, especially in edge recognition algorithms. It creates a marginal image [37].

3.4.3. *Scharr filter*

This is a filter used to identify and highlight edges/slope features using derivative 1. It has a function like the Sobel filter and is used to detect edges/changes in the pixel intensity [38].

3.4.4. *Prewitt filter*

The Prewitt filter is like the Sobel in the sense that it uses two cores. One to change the horizontal direction and the other to change the vertical direction. The two centers are wrapped with the original image (meaning the same convolution operation) to calculate the derivatives roughly [38].

3.4.5. *Gaussian filter*

The Gaussian filter is linear. This filter is usually used to blur the image or reduce the noise [39].

3.4.6. Roberts edge

Crossover operator Roberts measures the slope of a simple, fast space to calculate two-dimensional on an image. It, therefore, highlights areas of high frequency that often correspond to the edges [40].

3.5. Building a numerical dataset

In the previous step, many features were extracted from each pixel in a CT image. For example, an image with 256 × 256 dimensions has 65,536 pixels, so if we extract 40 features from each and store them in a dataset, we will have 1 row for the image, and 40 × 65,536 columns. On the other hand, adding a target column should not be forgotten to determine whether the input image belonged to a cancerous patient or a noncancerous one. This procedure is continued until all the images' features are extracted and stored in a dataset.

3.6. Dimensional reduction

In the previous phase, an extensive dataset consisting of many features was obtained. However, implementing classification on this extensive dataset is time-consuming and not efficient. Implementing dimensional reduction algorithms on large data is one of the most critical steps. PCA and LDA are two-dimensional reduction algorithms used in this paper.

3.7. Feature selection

Feature selection methods have become an unavoidable part of the machine learning process to deal with high-dimensional data. Feature selection can identify related features and eliminate unrelated and repetitive ones to observe a subset of attributes that best describe the problem.

The first goal of the proposed feature selection method is to reach the same accuracy rate as the exclusive features. The second goal is to improve the accuracy rate. Here, gathering extensive information on the features costs too much, both in time and money, and new information is wasted in classifying and diagnosis. Reducing the dimension in terms of the number of features is recommended to get a better response and find a better correlation between the features and the outcomes.

A GA is a technique to select the best features. This technique generates a binary random vector consisting of the features using Eq. (1).

$$Vector_{(s_j)}: \quad s_j = Y_i; \quad Y_i = \begin{cases} 1 & if\ Vector_{s_j}\ contains\ feature\ i \\ 0 & otherwise \end{cases} \tag{1}$$

An objective function based on the misclassification performance criterion is defined for any selected combination of the features. This objective function is a penalty function that should be minimized to find the best features. Here, the misclassification rate is simple and is obtained using Eq. (2).

$$mcr = {}^{\sum a_{ij} - [\sum a_{ij};(i=j)]}/_{\sum a_{ij}}; \quad i, j = 1, 2, \dots, m \tag{2}$$

The number of classification targets is the number of cases, the target is classified as the target using the classification method. The elements that construct the matrix in (3) form the so-called confusion matrix that depends on the problem as well as the dataset.

$$\begin{pmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mm} \end{pmatrix}_{m\times m} \tag{3}$$

Now, the objective function to be minimized is a weighted sum of the mcr and n_f (number of selected features) defined as:

$$Min\ Z = w_1 * mcr + w_2 * n_f \tag{4}$$

Dividing both sides of Eq. (4) by w_1, we will have:

$$Min\ Z = mcr + {}^{w_2}/_{w_1} * n_f. \tag{5}$$

Assuming ${}^{w_2}/_{w_1} = W$, the objective function becomes:

$$Min\ Z = mcr + W * n_f. \tag{6}$$

Now, W is defined as:

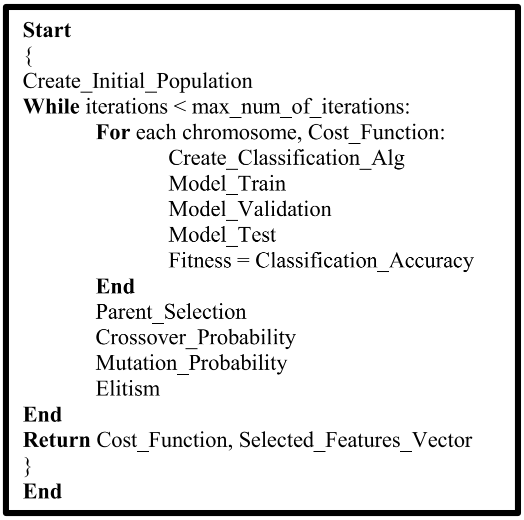
$$W \propto mcr \rightarrow W = \beta * mcr \rightarrow Min\ Z = mcr + \beta * mcr * n_f \tag{7}$$

Finally, the objective function will be:

$$Min\ Z = mcr(1 + \beta * n_f), \tag{8}$$

where β is defined as a penalty for having an additional feature ($0 \leq \beta \leq 1$). Using this objective function, the GA finds the best combination of the features with the minimum number of features that minimize both the cost and the misclassification rate. Here, the stopping criterion to end the iterations in GA is chosen to be a predefined number of iterations.

The Pseudocode of the GA-based feature selection algorithm is:



```
Start
{
Create_Initial_Population
While iterations < max_num_of_iterations:
        For each chromosome, Cost_Function:
                Create_Classification_Alg
                Model_Train
                Model_Validation
                Model_Test
                Fitness = Classification_Accuracy
        End
        Parent_Selection
        Crossover_Probability
        Mutation_Probability
        Elitism
End
Return Cost_Function, Selected_Features_Vector
}
End
```

3.8. Classification

Three different classifiers – GB, RF, and SVM – are implemented on three other datasets separately. The first dataset, the PCA dataset, is derived from the PCA dimensional reduction algorithm. The following dataset, called the LDA dataset, is obtained using the LDA dimensional reduction algorithm. The third one is constructed employing the GA, which is applied to the PCA dataset. Each dataset has several rows of images and has various columns depending on the method of execution alongside a binary target feature that defines the samples as noncancerous or cancerous (0 or 1).

4. Experimental results and analysis

This section discusses the way the data is collected, the implementation results of the proposed three methods on the data, and the analysis of the results.

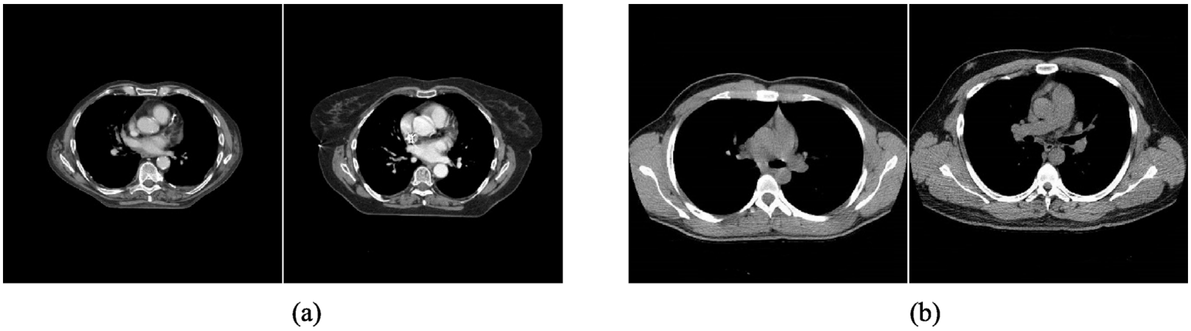


Fig. 2. (a) medical CT images of lung cancer patients, (b) medical CT images of lung patients other than lung cancer.

4.1. Data collection

Images are collected from a hospital situated in Tehran, Iran. The images used in this study are provided at https://data.mendeley.com/datasets. Part of these CT scan images of lungs belongs to lung cancer patients and are classified as cancerous images. The rest belongs to other lung diseases, such as patients who caught COVID-19, classified as noncancerous images. As lung cancer symptoms are rare, all the possible lung diseases are considered noncancerous images to improve lung cancer diagnosis efficiency. For this reason, lung cancer disease is not detectable in the early stages. Most physicians and doctors in the early stages of tumor growth diagnose a disease other than cancer, which causes this type of cancer progression in the infected person.

The total number of CT scan images used in this paper is 364, of which 238 are cancerous images, and the rest (126) belong to noncancerous images. All these images are collected with the help of a pulmonologist to skip any probable error in classifying images. Some of the CT images of the lungs acquired from the hospital database are shown in Fig. 2.

4.2. The implementation results of the first method

As seen in Fig. 2, applying any pre-processing or segmentation on the raw images is not needed when implementing CNN and ANN. In other words, the raw lung CT scan images are fed as inputs to the CNN and ANN architecture in the first method. Several different structures are evaluated to obtain the best structure to distinguish cancerous CT images from noncancerous ones. However, the best structure consists of three convolution layers with 64, 64, and 128 feature maps, respectively, in the first, second, and third layers in the convolutional neural network section. The artificial neural network also contains two hidden layers, each containing 128 neurons. This study uses max pooling with dimensions of 2 × 2 after each convolution layer to maintain the feature maps. Fig. 3 shows the graphical structure of the best approach.

The structural model has 63,109,441 trainable parameters, which would have been quadrupled if the max-pooling layers had not been used. This would have increased the algorithm's execution time. Another case in point is the number of images used to learn and evaluate the algorithm, of which 324 images were for training, and 40 images were for algorithm testing. Finally, implementing this structure on raw (unprocessed) images, an accuracy of 65.81% was obtained with the training loss function value of 5.4368 and the testing loss function value of 4.0295 with 62.50% accuracy. The difference between the two accuracies shows that the model is not overfitted.

4.3. The implementation results of the second method

As shown in Fig. 1, image pre-processing and segmentation are used in the second method before running the CNN and the ANN algorithms. This method aims to determine whether performing image pre-processing and image segmentation affects the performance of the first method. In the first step of image pre-processing, all the image sizes are set to 512 × 512 so that all the pixels in an image remain intact. The next step applies the median filter to remove any possible noise of resized lung CT images. Fig. 4 shows the image before and after the median filter is performed on both the cancerous and noncancerous lung CT scans. As seen in this figure, the images after the filter are not much different from the images without the filter, which is a characteristic of CT scan images.

The image segmentation tries to help the algorithm to diagnose as best it can by removing unnecessary parts. In fact, at this point, the algorithm can only focus on the lung, which will improve the algorithm's performance. For better understanding, all the steps are applied to the two filtered images in Fig. 4. The masks that cover unnecessary parts of the images are shown in Fig. 5. Then, by placing these masks on the filtered images, the lung will be visible in Fig. 6.

As shown in Fig. 6, unnecessary parts are removed. After completing all the above steps, 364 processed and segmented images are given to the CNN and ANN algorithms. All the first method steps are repeated from this point on, except that the images are not raw. Similar to the first method, several different structures are evaluated in the second method. However, the architecture produced the best result consisting of three convolution layers with 64, 64, and 128 feature maps, respectively, in the first, second, and third layers in the CNN section. The ANN also included two hidden layers, each containing 128 neurons. As mentioned, max pooling with dimensions of 2 × 2 is used after each convolution layer to maintain the image features. Fig. 7 shows the graphical structure of the best structure.

The second method's model has 63,109,441 trainable parameters. It would have quadrupled the number of parameters and increased the time of the algorithm's execution if max-pooling layers were not applied. In this method, the same as in the first method, 324 lung CT images are considered for training, and 40 lung CT images are used to evaluate the algorithm's performance. Finally, by implementing this structure on the processed and segmented images, a 100% accuracy with a loss function of 3.0576 × 10-5 for the train and 93% accuracy with a loss function of 1.096 × 10-7 for the test is obtained. The difference between the two accuracies shows that the model is not overfitted. Besides, performing image pre-processing and segmentation affects the first method's performance significantly.

4.4. The implementation results of the third method

As masks are placed on filtered images in the image segmentation phase, images' unnecessary parts are covered. Therefore, to reduce the number of columns, the segmented images are resized to 256 × 256. By these dimensions for each segmented image, 2,621,440 data are generated for each image when the features are extracted.

The filters/features are applied alone to each image's pixel and store each pixel's calculations in a data frame. To create a numeric dataset, the number of pixels in each image is 65,536, and the total number of filters executed on each pixel is 40. The data values of the original

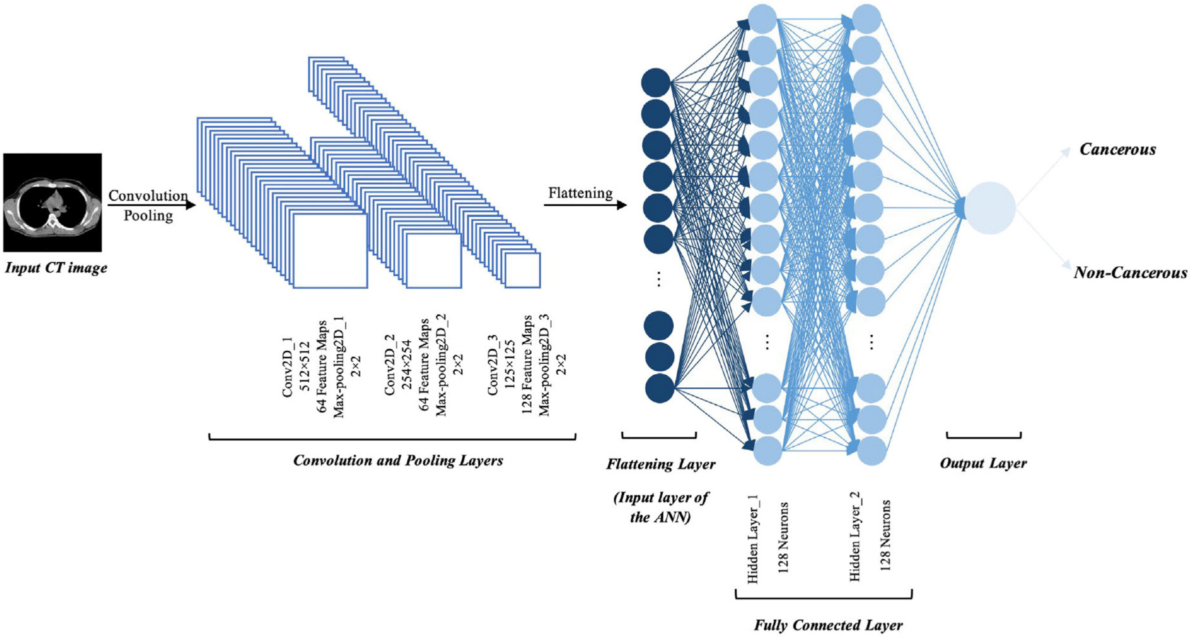


Fig. 3. Best result graphical structure of the first method.

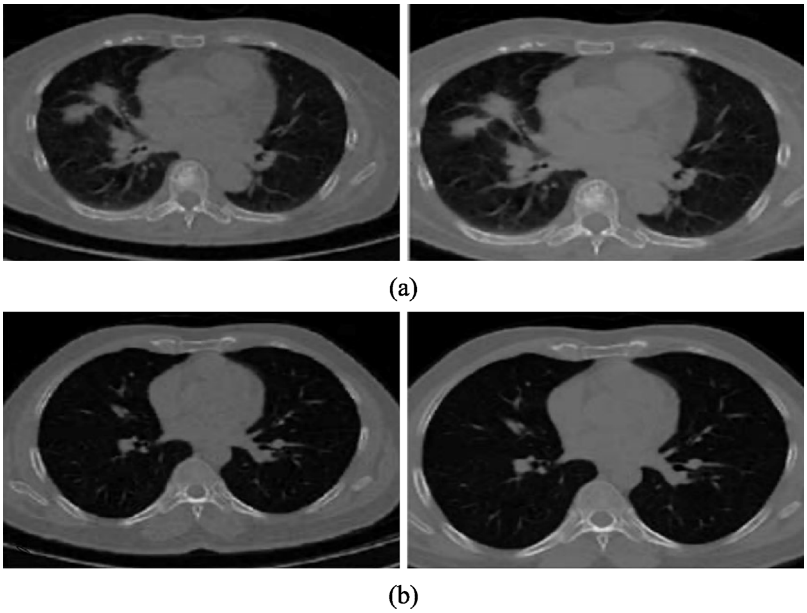


Fig. 4. (a) On the left side, the cancerous lung CT image is shown before the median filter, and on the opposite side, the after median-filter result is shown, (b) On the left side, the noncancerous lung CT image is shown before median-filter, and on the opposite side, the after median-filter is shown.

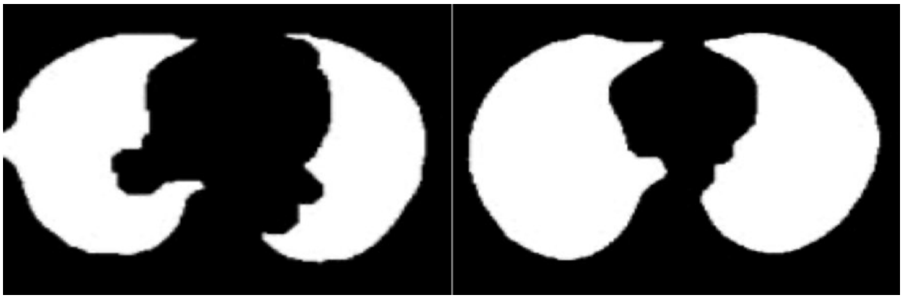


Fig. 5. The image on the right is a mask for a cancerous lung, and the image on the left is a mask for a noncancerous lung.

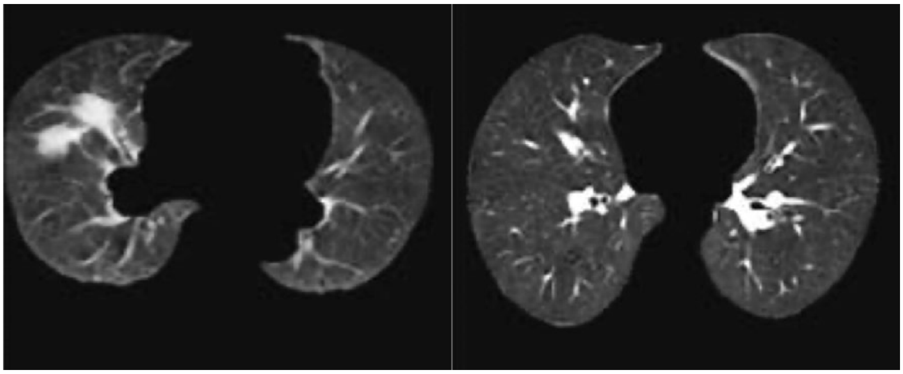


Fig. 6. The image on the right shows lung cancer after applying the mask, and the image on the left shows noncancerous lung after using the mask.

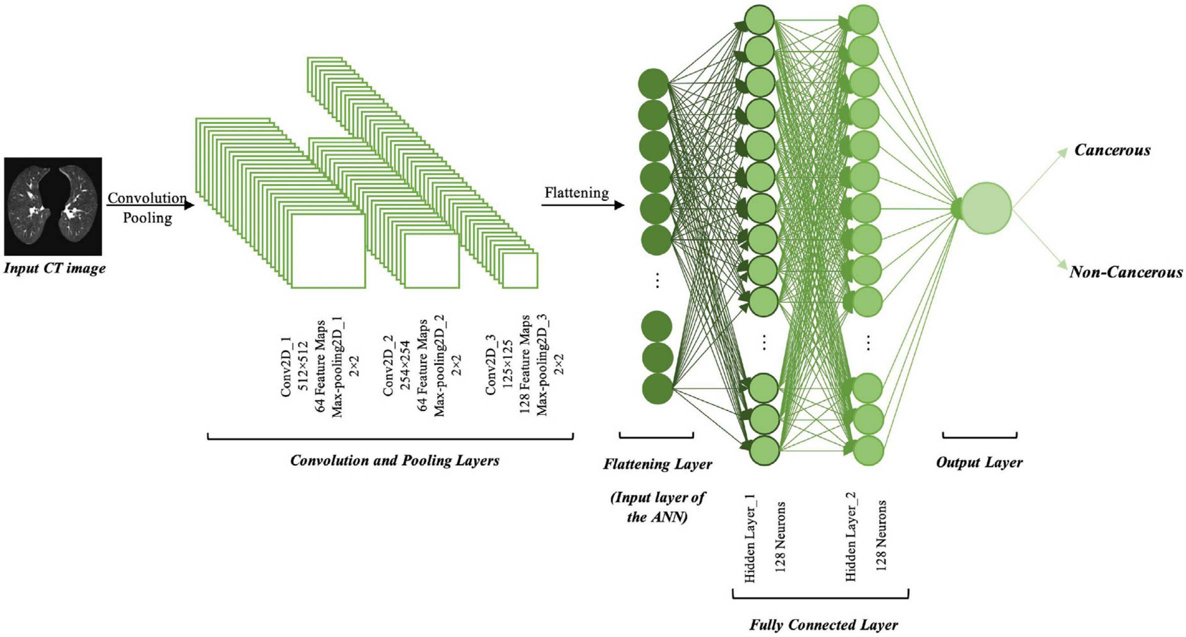


Fig. 7. Best result graphical structure of the second method.

Table 2
Dimension of the created dataset.

Dataset dimensions	Number of rows	Number of columns
	364	2,621,442

image's pixels and the label of being cancerous (1) or noncancerous (0) are also given in this dataset. Thus, each picture contains 1 row and 40 × 65,536 feature columns plus a label column and the original pixel's value. The dimensions of the dataset are shown in Table 2, considering all the images.

PCA and LDA were separately applied to the dataset to produce the PCA and LDA datasets, respectively. Given that all the following steps must be performed together on the entire dataset, a robust system is required to read it. Therefore, all subsequent steps are performed on High-Performance Computing (HPC) with 80 cores and 500 GB of RAM.

Before the PCA and LDA can be applied in the dimensional reduction phase, it is necessary to determine the number of components required for each algorithm. Therefore, a loop is used to obtain the optimal number of components. Some of the results of this loop for the PCA method are shown in Table 3. The numbers shown in the number of components' output required by the PCA in Table 3 show that the conversion of 2,621,441 to how many can explain the total data variance.

According to Table 3, by forming 363 columns in the PCA algorithm, the total data variance can be fully explained. In other words, PCA can combine 2,621,441 feature columns into 363 columns so that this number of columns can explain 100% of the variance of the entire dataset. In Fig. 8, the number of components from 1 to 364 is drawn with the percentage of the corresponding variance explanation for each.

The same procedure is done with the LDA algorithm, where the number of columns must be equal to the minimum number of rows and the number of objective functions. Since there are 364 rows and two objective functions (0 for noncancerous and 1 for cancerous), a maximum of 2 components is possible. The HPC output of this number is set to 1 component.

After reducing the dimensions, it is time to perform the feature selection phase using the genetic algorithm. However, before implementing the genetic algorithm, all the following steps are taken once without implementing the feature selection algorithm so that one can finally compare whether the feature selection improves the performance of the third method.

As mentioned above, the LDA algorithm, by reducing the size of the dataset, eventually reaches 364 rows and two columns, named the LDA dataset. The first column contains 2,621,441 feature columns, and the second column is the objective function column. In the PCA algorithm, after reducing the dimensions of the dataset, it finally reaches 364 rows and 325 columns, named the PCA dataset. The first 324 columns are

Table 3
Part of the number of component output.

n_components	Variance description	n_components	Variance description	n_components	Variance description
332	0.99108	343	0.99486	354	0.99810
333	0.99145	344	0.99518	355	0.99836
334	0.99181	345	0.99549	356	0.99861
335	0.99217	346	0.99580	357	0.99886
336	0.99252	347	0.99610	358	0.99910
337	0.99287	348	0.99640	359	0.99934
338	0.99322	349	0.99670	360	0.99956
339	0.99356	350	0.99700	361	0.99974
340	0.99389	351	0.99728	362	0.99989
341	0.99422	352	0.99756	363	1
342	0.99454	353	0.99784	364	1

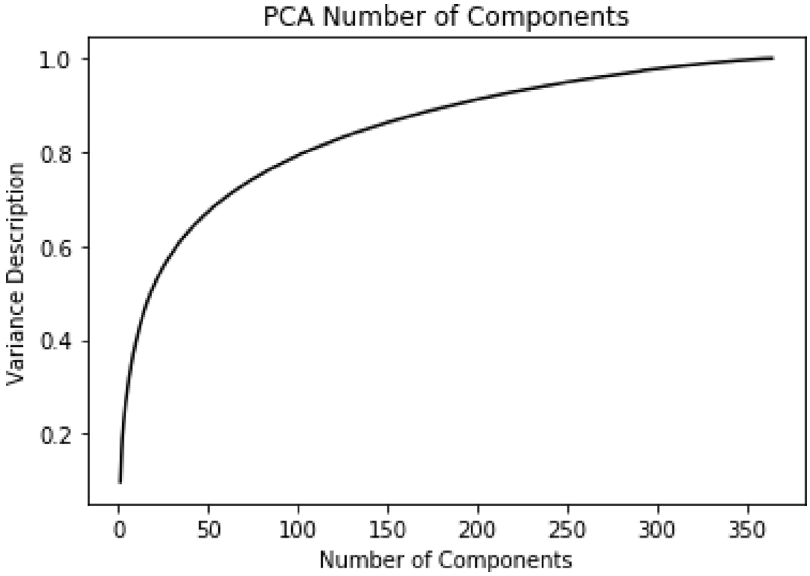


Fig. 8. The number of components corresponding to their variance explanation in PCA.

Table 4
Performance measurements of GBC, RFC, and SVC on the PCA dataset.

Methods	Accuracy	Precision	Recall	F1-score
GBC	0.95	0.95	0.95	0.95
RFC	0.82	0.86	0.82	0.79
SVC	0.73	0.53	0.72	0.61

related to the combination of 2,621,441 feature columns, and the last column is related to the objective function column. With these two sets of data (PCA and LDA), supervised machine learning algorithms, including GB classification, RF classification, and SVM classification, are implemented. Tables 4 and 6 show the performance of these three classification algorithms on the PCA and LDA datasets, respectively. Tables 5 and 7 present the confusion matrix of the three classification algorithms on the PCA and LDA datasets. In addition, Figs. 9 and 10 demonstrate a performance comparison among these three classification algorithms on the PCA and LDA datasets, respectively.

As shown in Table 4, the GB classification (GBC) algorithm creates the most accuracy after execution on the PCA dataset. The confusion matrix in Table 5 shows that the GBC algorithm correctly recognizes 37 data from 40 test data. In all parts of the employed machine learning algorithms, 40 data for the test and 324 data for the train are considered.

The results in Table 6 show that the SVC algorithm creates the most accuracy after running on the LDA dataset. As demonstrated by its confusion matrix in Table 7, the SVC algorithm correctly identifies 38 data from 40 test data.

A receiver operating characteristic (ROC) curve is a graphical tool that examines a binary classification performance in statistical analysis. The information contained in ROC curves is beneficial in choosing an appropriate classifier under specific criteria. The curve is plotted using the true-positive rate (TPR), also known as sensitivity, against the false-positive rate (FPR), equivalent to one minus specificity at various cut-off points of a parameter. High values of TPR and low amounts of FPR indicate an improvement in ROC curves; these values cause the points to move towards the upper left corner of ROC and make a desirable decision. The area under the curve is calculated to evaluate a given classifier's performance on a set of data and classifier consistency analysis. In this research, the three classifiers' ROC curves are depicted in Figs. 11 and 12 for the PCA and LDA datasets, respectively. The graphs demonstrate that the GBC technique's classification results in Fig. 11 and the SVC technique in Fig. 12 are more accurate and reliable than the other algorithms. In Fig. 12, the ROC graphs of both GBC and RFC are stacked on top of each other.

After implementing the step reduction scale to perform the genetic feature selection algorithm and then the machine learning algorithms, let us go back to the next step. As mentioned earlier, the LDA dataset consists of two columns, so it is virtually impossible to run GA on it. Therefore, one can only execute the feature selection algorithm on the PCA dataset. While the maximum number of iterations was set to 10, after the sixth iteration, the cost function value converges to 0.12464, as shown in Fig. 13, and the rest iterations remain the

Table 5
Confusion matrix of GBC, RFC, and SVC on the PCA dataset.

GBC	Actual value		RFC	Actual value		SVC	Actual value	
Predicted value	9	2	Predicted value	4	7	Predicted value	0	11
	0	29		0	29		0	29

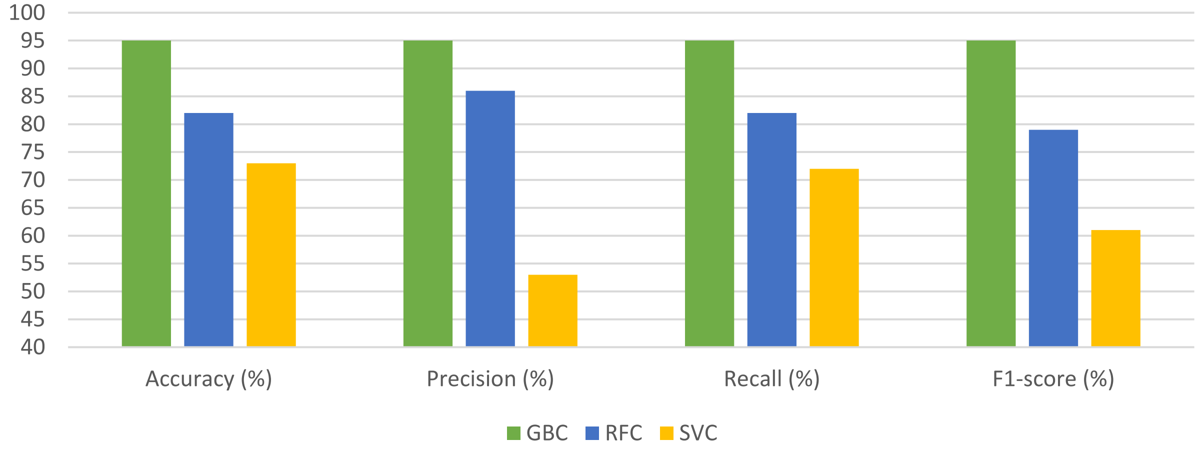


Fig. 9. The performance of GBC, RFC, and SVC on the PCA dataset.

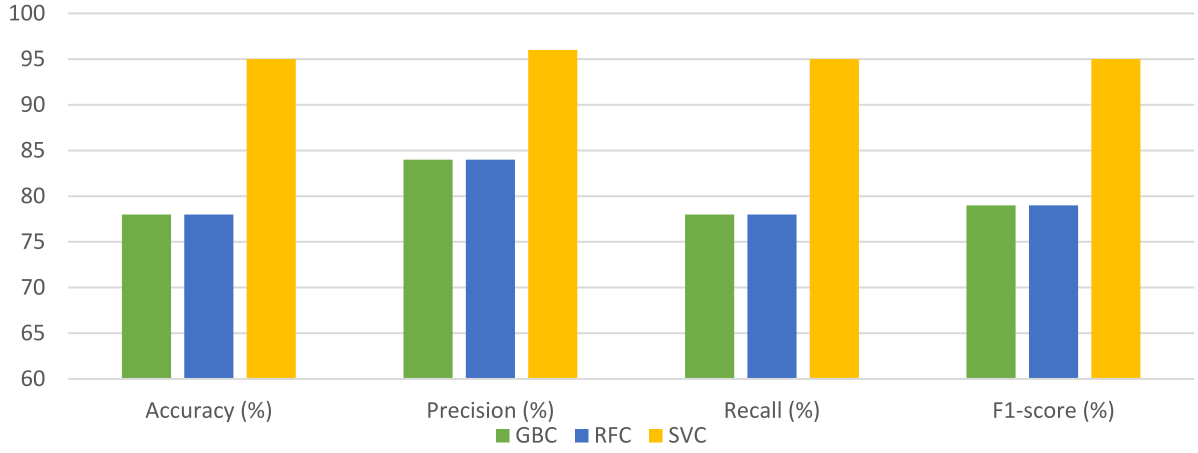


Fig. 10. The performance of GBC, RFC, and SVC on the LDA dataset.

Table 6
Performance measurements of GBC, RFC, and SVC on the LDA dataset.

Methods	Accuracy	Precision	Recall	F1-score
GBC	0.78	0.84	0.78	0.79
RFC	0.78	0.84	0.78	0.79
SVC	0.95	0.96	0.95	0.95

same. The population size was 50, the crossover operator probability was 0.7, and the mutation probability was 0.3. Moreover, the roulette wheel method selects the parents in all operations. Table 8 shows GA's hyperparameters tuning along with their cost functions.

After selecting 172 columns as practical ones, the rest of the columns are first deleted, and the three previously mentioned machine learning algorithms are executed. Tables 9 and 10 show the performance and the confusion matrix of the three classification algorithms on the dataset derived from the feature selection algorithm, respectively. Besides, Figs. 14 and 15 demonstrate a performance comparison among these three classification algorithms and the ROC curves of these three classifiers on the derived dataset from the GA, respectively.

As shown in Table 9, the RFC algorithm creates the most accuracy after execution on the dataset obtained from the GA. As is clear from its confusion matrix in Table 10, this algorithm correctly recognized 34 data from 40 test data.

5. Sensitivity analyses and comparative study

In the previous section, the three proposed methods were examined, based on which the best results of each method were described. In this section, the sensitivity of each method to its parameters is first analyzed. Then, they are compared to shed light on their strength and weaknesses.

5.1. Sensitivity analysis

According to the previous section's explanations given on CNN and ANN, each network includes parameters that must be adjusted to achieve the optimal result. To this aim, different structures, from the simplest to the most complex ones, were examined until there was no improvement. It should be noted that the simplicity of the structure

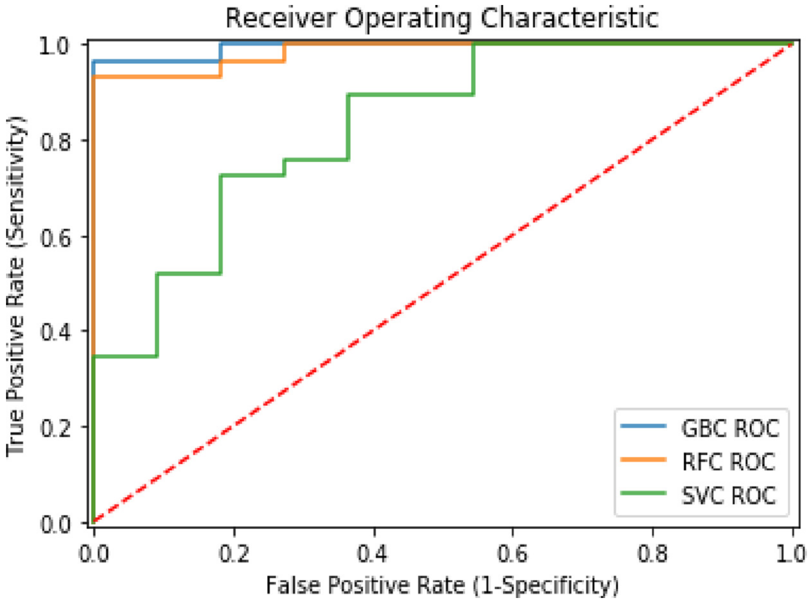


Fig. 11. ROC curves for three classifiers on the PCA dataset.

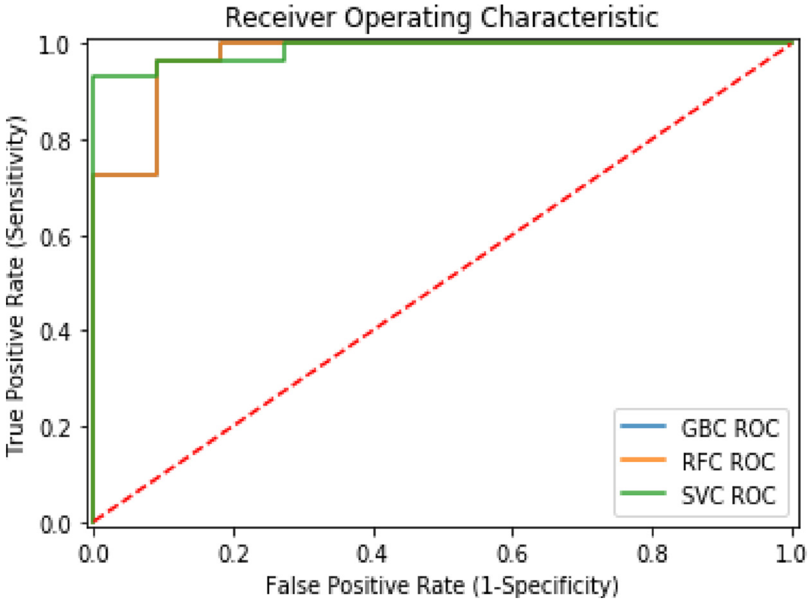


Fig. 12. ROC curves for three classifiers on the LDA dataset.

Table 7
Confusion matrix of GBC, RFC, and SVC on the LDA dataset.

GBC	Actual value		RFC	Actual value		SVC	Actual value	
Predicted value	10 8	1 21	Predicted value	10 8	1 21	Predicted value	11 2	0 27

or its complexity does not necessarily affect the result. However, the difference in implementing complex structures over simple ones is that the smaller the number of parameters, the higher the effort to achieve the best possible response. This implies shorter the algorithm's execution time due to the lower number of parameters.

Tables 11–14 show the sensitivity analyses of the first two methods, the convolutional and artificial neural network on raw images, and the convolutional and artificial neural network on the pre-processed and segmented images. As mentioned, the complexity of the structures has continued to the point where there is no improvement anymore. It is necessary to consider the accuracy and loss function of the test and train simultaneously to achieve the desired result.

As shown in Table 11, various convolution layers with different filter detections are utilized. In each of them, multiple nodes in different hidden layers are also used to build a structural architecture. For example, as seen in the last row of Table 11, three convolution layers with 64, 64, and 128 detection filters are used to build the CNN structural architecture. The ANN structural architecture uses three hidden layers containing 128, 128, and 256 neurons. In conclusion, as the accuracy rate decreased and the loss function for both train and

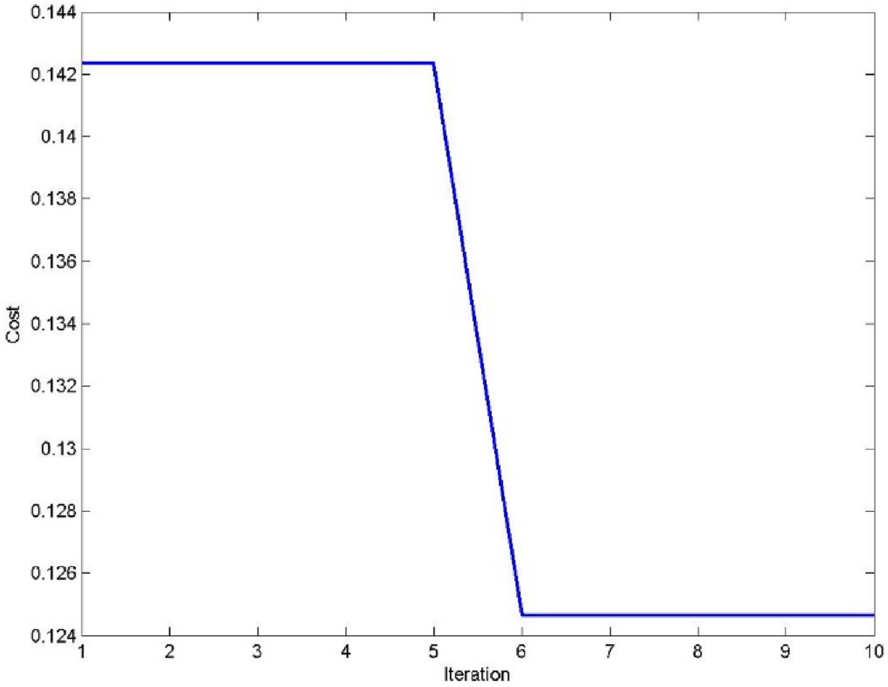


Fig. 13. The GA cost function for each iteration.

Table 8
Hyperparameters tuning process (the **bold** values show the best hyperparameters concerning the cost function value, and the **underlining** values show the second best hyperparameters).

Max iteration	Population	% Mutation	% Crossover	Time (s)	Cost function	# of selected features
10	**20**	**0.3**	**0.7**	48,235.780	**0.12890**	**172**
10	20	0.4	0.7	27,149.952	0.16745	194
10	20	0.5	0.7	21,853.446	0.16745	194
10	20	0.3	0.8	22,557.235	0.14843	187
10	20	0.4	0.8	32,168.301	0.16738	191
10	20	0.5	0.8	21,063.200	0.16745	194
10	20	0.3	0.9	21,245.173	0.14843	187
10	20	0.4	0.9	47,362.354	0.16745	194
10	20	0.5	0.9	42,417.132	0.16745	194
10	**50**	**0.3**	**0.7**	40,130.765	**0.12464**	**172**
10	50	0.4	0.7	85,919.442	0.16738	191
10	50	0.5	0.7	103,398.083	0.16738	191
10	50	0.3	0.8	52,323.353	0.16738	191
10	50	0.4	0.8	91,157.160	0.16738	191
10	50	0.5	0.8	63,885.141	0.16738	191
10	50	0.3	0.9	66,239.779	0.13741	178
10	50	0.4	0.9	10,9868.125	0.15745	187
10	50	0.5	0.9	10,6193.759	0.15745	187
10	80	0.3	0.7	86,121.910	0.16738	191
10	80	0.4	0.7	91,011.980	0.16738	191
10	80	0.5	0.7	120,447.483	0.16738	191
10	80	0.3	0.8	74,341.422	0.16738	191
10	80	0.4	0.8	150,442.208	0.16738	191
10	80	0.5	0.8	104,947.371	0.16738	191
10	80	0.3	0.9	111,308.229	0.16738	191
10	80	0.4	0.9	114,145.831	0.16738	191
10	80	0.5	0.9	122,187.983	0.16738	191

test data increased, the implementation of more complex structures stopped. Consequently, the fifth structure in Table 11 is determined as the best answer to this problem.

As presented in Table 12, various convolution layers with different filter detections are utilized. In each of them, multiple neurons in a different number of hidden layers are also used to build a structural architecture. In the end, the implementation of a more complex structure is canceled, and the fifth row of Table 12 is determined as the best solution to the problems at hand.

Regarding the third method's sensitivity analysis, the whole process of this method is described in Section 4. Still, a brief explanation of this method's sensitivity analysis is not without merit here. The goal of the third method was to analyze numerical data obtained after extracting features from images. After extracting the features, two-dimensional

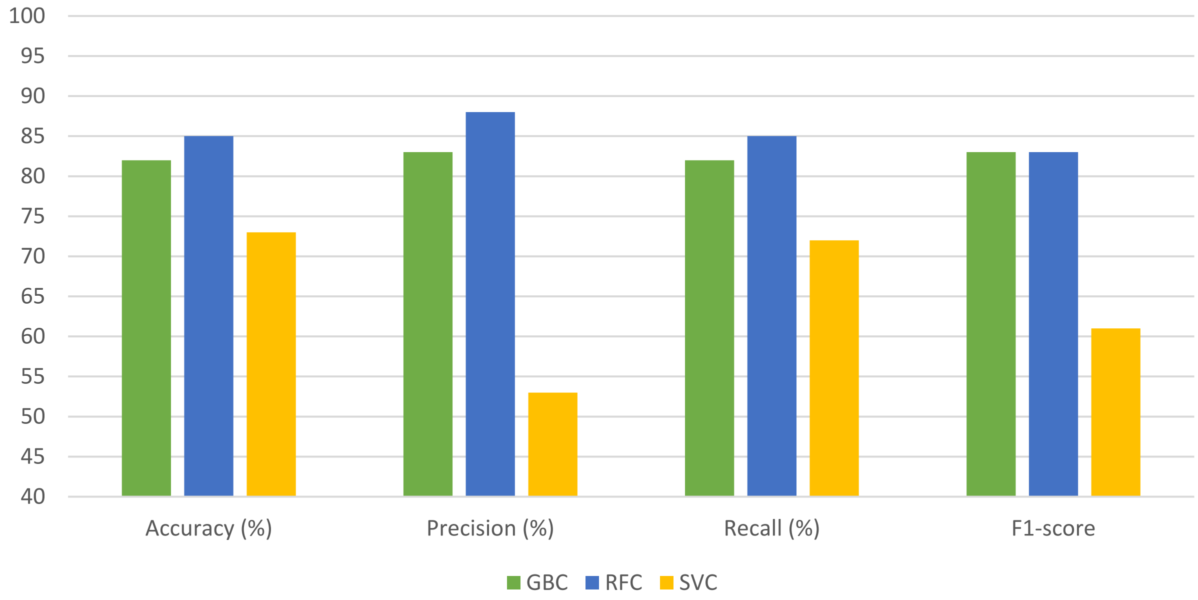


Fig. 14. The performance measurements criteria of GBC, RFC, and SVC on the derived dataset from the GA.

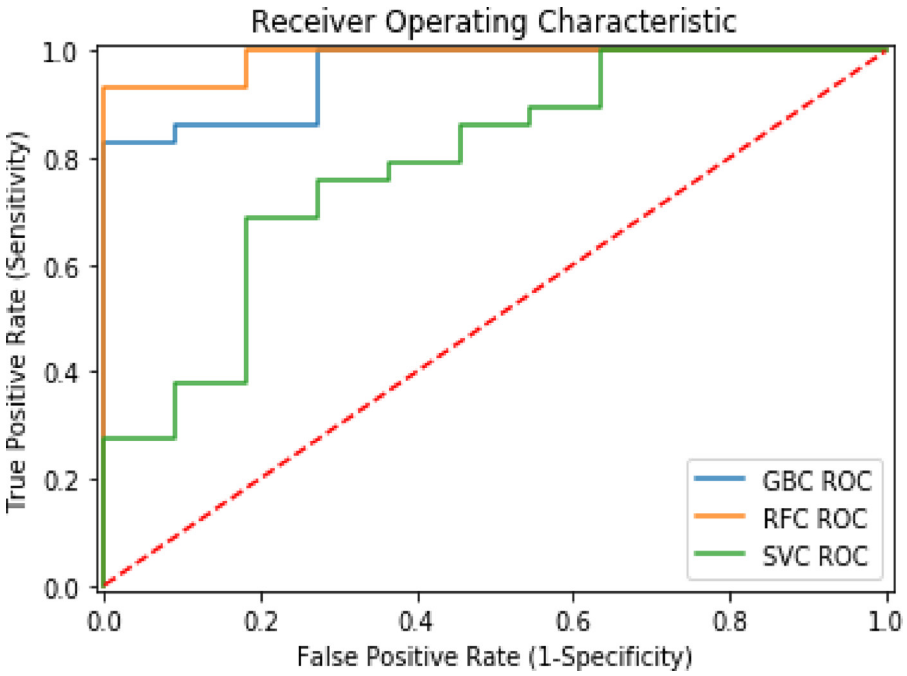


Fig. 15. ROC curves for three classifiers on the derived dataset from the GA.

Table 9
Performance measurements of GBC, RFC, and SVC on the dataset derived from the GA.

Methods	Accuracy	Precision	Recall	F1-score
GBC	0.82	0.83	0.82	0.83
RFC	0.85	0.88	0.85	0.83
SVC	0.73	0.53	0.72	0.61

reduction methods (PCA and LDA) are used to examine which one performs better. Following each of them, supervised machine learning algorithms (GBC, RFC, and SVC) are implemented separately to see how performing only three feature extraction steps reduces the dimensions and what results are obtained using machine learning algorithms. Table 13 presents the results of these three steps and the accuracy of each algorithm. It should be noted that other evaluation criteria are given in Tables 4–7, discussed previously in Section 4.

As shown in Table 13, by applying the above three steps, the SVC on the LDA dataset and GBC on the PCA dataset offers the best performance in diagnosing lung cancer. However, in the third method, we did not suffice with these results and tried to use different ways to improve lung cancer diagnosis. Thus, the genetic feature selection algorithm is implemented to see if there is an improvement in the diagnostic determination of the disease. As mentioned in Section 4, it is impossible to implement the genetic feature selection algorithm on the LDA dimensional reduction method. One can only use it on the dataset obtained from the PCA dimensional reduction method. Table 14

Table 10
Confusion matrix of GBC, RFC, and SVC on the dataset derived from the GA.

GBC	Actual value		RFC	Actual value		SVC	Actual value	
Predicted value	8	3	Predicted value	5	6	Predicted value	0	11
	4	25		0	29		0	29

Table 11
Sensitivity analysis of performing CNN and ANN methods on raw images.

Num. of convolution layers	Num. of filter detections[a]	Num. of hidden layers	Num. of nodes in hidden L.[b]	Num. of total params	Train		Test		Time (s)
					Accuracy (%)	Loss function	Accuracy (%)	Loss function	
1	64	1	128	532,686,849	65.71	5.5623	62.5	4.0295	312,418.58
2	64-64	1	128	130,095,169	65.75	5.5452	62.5	4.0295	76,300.26
2	64-64	2	128-128	130,111,681	65.77	5.5204	62.5	2.0148	76,309.95
3	64-64-128	1	128	63,092,929	65.8	5.4398	62.5	4.0295	37,003.74
3	64-64-128	2	128-128	63,109,441	65.81	5.4368	62.5	4.0295	37,013.42
3	64-64-128	3	128-128-256	63,142,593	65.8	5.5018	62.5	12.0886	37,032.86

[a] The number of feature detectors is listed in layers.
[b] The number of neurons is listed in layers.

Table 12
Sensitivity analysis of performing CNN and ANN methods on processed and segmented images.

Num. of convolution layers	Num. of filter detections[a]	Num. of hidden layers	Num. of nodes in hidden L.[b]	Num. of total params	Train		Test		Time (s)
					Accuracy (%)	Loss function	Accuracy (%)	Loss function	
1	64	1	128	532,686,849	99.52	0.019	92.5	0.754	307,915.06
2	64-64	1	128	130,095,169	99.58	0.023	92.5	1.5115	75,200.39
2	64-64	2	128-128	130,111,681	99.92	0.0024	93	1.12E−07	75,209.94
3	64-64-128	1	128	63,092,929	99.95	0.0012	93	1.84E−07	36,470.32
3	64-64-128	2	128-128	63,109,441	100	0.00003	93	1.09E−07	36,479.87
3	64-64-128	3	128-128-256	63,142,593	100	0.00004	93	5.13E−07	36,499.03

[a] The number of feature detectors is listed in layers.
[b] The number of neurons is listed in layers.

Table 13
Sensitivity analysis of two-dimensional reduction methods (PCA and LDA).

Feature extraction		
LDA dimensional reduction	PCA dimensional reduction	Algorithms
78%	95%	GBC
78%	82%	RFC
95%	73%	SVC

Table 14
Sensitivity analysis of the implementation of machine learning algorithms before and after the implementation of the genetic feature selection algorithm.

PCA dimensional reduction		
After GA feature selection	Before GA feature selection	Algorithms
82%	95%	GBC
85%	82%	RFC
73%	73%	SVC

contains the accuracy of the algorithms after execution on the dataset of the genetic feature selection algorithm and before its implementation. It is recommended to refer to Tables 4–5 and 9 and 10 to see the other evaluation criteria.

According to the results in Table 14, the implementation of machine learning algorithms on the genetic feature selection algorithm's dataset has improved the RFC method's results. However, this improvement is not valid for GBC, as this method's accuracy has been severely reduced. Moreover, the performance of the SVC method did not change.

5.2. *Comparison*

This section is devoted to the comparison of the methods.

5.2.1. *Method 1: CNN and ANN on raw images*

As the name implies, all CT scans enter the CNN system without pre-processing or modification. After passing this stage, they enter the artificial neural network to classify the images into cancerous and non-cancerous categories. Implementing this method on raw (unprocessed) images resulted in an accuracy of 65.81% with a 5.4368 loss function value for the train and 62.50% accuracy with a 4.0295 loss function value for the test.

5.2.2. *Method 2: CNN and ANN on the pre-processed and segmented images*

In this method, the images are pre-processed and segmented before the CT images enter the CNN system. The image size is changed in the pre-processing phase, and the median filter is applied to reduce possible image noise. By implementing this structure on the pre-processed and segmented images, an accuracy of 100% is achieved with a loss function value of 3.0576×10^{-5} for the training set and 93% accuracy and a loss function value of 1.096×10^{-7} for the test set.

5.2.3. Method 3: Numerical features of the pre-processed and segmented images and employing machine learning algorithms

In this method, all the pre-processing and segmentation phases are the same as in the second method. After these two phases, the numerical features are extracted from the images and converted to numerical data in a data frame. This dataset's dimensions are 364 × 2,621,442, which is referred to as large or big data. As the size of this dataset is large, dimensional reduction algorithms are first used, and then a genetic feature selection algorithm is utilized to increase the processing speed. Next, machine learning algorithms are employed to classify cancer and non-cancer images. The GBC algorithm obtains the best results with 95% accuracy when implemented on the PCA dataset, the SVC algorithm with 95% accuracy when implemented on the LDA dataset, and the RFC algorithm with 85% accuracy when performed on a dataset obtained from the genetic feature selection.

6. Discussion and conclusion

In this study, three different methods were used to diagnose lung cancer in its early stages. One of the most significant achievements of this research was the comparability of all three methods. Methods were comparable when all input images were the same in each method, as in this paper. Another comparable feature was the use of the same amount of data for training and testing sets; 324 images for training and the remaining 40 images for the test in all investigations. Another contribution of this research includes the implementation of the third method. This method extracts numerical features from the pre-processed and segmented lung CT images. The dimensional reduction algorithms (PCA and LDA) are applied to the obtained dataset, GA feature selection is utilized, and supervised machine learning algorithms (GB, RF, SVC) are performed.

The comparison analysis showed that the third method had the best performance (95% accuracy for the testing set). Since the third method is one of the main contributions of this paper, we can see that it has the best accuracy in using two different machine learning classification algorithms. Therefore, the results illustrated that separately applying GB on the PCA dataset and SVC in the LDA has the best performance with 95% accuracy in both.

Future works can improve the accuracy of the cancer diagnosis by executing GA feature selection before dimensional reduction. Using different algorithms for feature selection can probably improve the accuracy of cancer detection. Moreover, extracting more different features in the feature extraction step may positively impact the system's accuracy. Moreover, other cancers kill countless people every year. Therefore, given that this research method's implementation has yielded promising results, we will try to perform the best approach on various cancers and diseases in future work.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] WHO, Cancer, 2021, Retrieved from https://www.who.int/news-room/fact-sheets/detail/cancer.

[2] B.C. Bade, C.S.D. Cruz, Lung cancer 2020: epidemiology, etiology, and prevention, Clin. Chest Med. 41 (1) (2020) 1–24.

[3] C.R. MacRosty, M.P. Rivera, Lung cancer in women: A modern epidemic, Clin. Chest Med. 41 (1) (2020) 53–65.

[4] A.S. Ahmad, A.M. Mayya, A new tool to predict lung cancer based on risk factors, Heliyon 6 (2) (2020) e03402.

[5] C. Lynch, B. Abdollahi, J. Fuqua, A. deCarlo, J. Bartholomai, R. Balgemann, H. . Frieboes, Prediction of lung cancer patient survival via supervised machine learning classification techniques, Int. J. Med. Inform. 108 (2017) 1–8, http://dx.doi.org/10.1016/j.ijmedinf.2017.09.013.

[6] W. Raghupathi, Data mining in health care, Healthc. Inform.: Improv. Effic. Prod. 211 (2010) 223.

[7] D. Tomar, A survey on data mining approaches for healthcare, Int. J. Bio - Sci. Bio - Technol. 5 (2013) 241–266, http://dx.doi.org/10.14257/ijbsbt.2013.5.5.25.

[8] R. Golan, C. Jacob, J. Denzinger, Lung nodule detection in CT images using deep convolutional neural networks, in: Paper Presented at the 2016 International Joint Conference on Neural Networks, IJCNN, 2016.

[9] D.P. Kaucha, P.W. Prasad, A. Alsadoon, A. Elchouemi, S. Sreedharan, Early detection of lung cancer using SVM classifier in biomedical image processing, in: Paper Presented At the 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering, ICPCSI, 2017.

[10] M.B. Miah, M. Yousuf, Detection of lung cancer from CT image using image processing and neural network, in: Paper Presented at the International Conference on Electrical Engineering and Information Communication Technology ICEEICT, JU, Savar, Dhaka, Bangladesh, 2015.

[11] P.M. Shakeel, M.A. Burhanuddin, M.I. Desa, Lung cancer detection from CT image using improved profuse clustering and deep learning instantaneously trained neural networks, Measurement 145 (2019) 702–712.

[12] M. Vas, A. Dessai, Lung cancer detection system using lung CT image processing, in: Paper Presented At the 2017 International Conference on Computing, Communication, Control and Automation, ICCUBEA, 2017.

[13] N. Zayed, H. Elnemr, Statistical analysis of haralick texture features to discriminate lung abnormalities, Int. J. Biomed. Imaging 2015 (2015) 1–7, http://dx.doi.org/10.1155/2015/267807.

[14] N. Maleki, Y. Zeinali, S.T.A. Niaki, A k-NN method for lung cancer prognosis with the use of a genetic algorithm for feature selection, Expert Syst. Appl. 164 (2021) 113981.

[15] Y. Zeinali, S.T.A. Niaki, Heart sound classification using signal processing and machine learning algorithms, Mach. Learn. Appl. 7 (2022) 100206, http://dx.doi.org/10.1016/j.mlwa.2021.100206.

[16] A.H. Chen, C. Yang, The improvement of breast cancer prognosis accuracy from integrated gene expression and clinical data, Expert Syst. Appl. 39 (5) (2012) 4785–4795, http://dx.doi.org/10.1016/j.eswa.2011.09.144.

[17] W. Cherif, Optimization of K-NN algorithm by clustering and reliability coefficients: application to breast-cancer diagnosis, Procedia Comput. Sci. 127 (2018) 293–299, http://dx.doi.org/10.1016/j.procs.2018.01.125.

[18] P. Kr, N. Aradhya, Lung cancer survivability prediction based on performance using classification techniques of support vector machines, C4.5 and naive Bayes algorithms for healthcare analytics, Procedia Comput. Sci. 132 (2018) 412–420, http://dx.doi.org/10.1016/j.procs.2018.05.162.

[19] B. Zheng, S.W. Yoon, S. Lam, Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms, Expert Syst. Appl. 41 (4) (2013) 1476–1482, http://dx.doi.org/10.1016/j.eswa.2013.08.044.

[20] M.I. Razzak, S. Naz, A. Zaib, Deep learning for medical image processing: Overview, challenges and the future, in: Classification in BioApps, Springer, 2018, pp. 323–350.

[21] S. KL, S.N. Mohanty, K. S, N. A, G. Ramirez, Optimal deep learning model for classification of lung cancer on CT images, Future Gener. Comput. Syst. 92 (2019) 374–382, http://dx.doi.org/10.1016/j.future.2018.10.009.

[22] T.L. Chaunzwa, A. Hosny, Y. Xu, A. Shafer, N. Diao, M. Lanuti, H.J. . Aerts, Deep learning classification of lung cancer histology using CT images, Sci. Rep. 11 (1) (2021) 1–12.

[23] M.A. Khan, S. Rubab, A. Kashif, M.I. Sharif, N. Muhammad, J.H. Shah, S.C. . Satapathy, Lungs cancer classification from CT images: An integrated design of contrast based classical features fusion and selection, Pattern Recognit. Lett. 129 (2020) 77–85, http://dx.doi.org/10.1016/j.patrec.2019.11.014.

[24] M. Toğaçar, B. Ergen, Z. Cömert, Detection of lung cancer on chest CT images using minimum redundancy maximum relevance feature selection method with convolutional neural networks, Biocybern. Biomed. Eng. 40 (1) (2020) 23–39, http://dx.doi.org/10.1016/j.bbe.2019.11.004.

[25] I. Bonavita, X. Rafael-Palou, M. Ceresa, G. Piella, V. Ribas, M.A.G. Ballester, Integration of convolutional neural networks for pulmonary nodule malignancy assessment in a lung cancer classification pipeline, Comput. Methods Programs Biomed. 185 (2020) 105172.

[26] D. Moitra, R.K. Mandal, Classification of non-small cell lung cancer using one-dimensional convolutional neural network, Expert Syst. Appl. 159 (2020) 113564.

[27] A.Y. Saleh, C.K. Chin, V. Penshie, H.R.H. Al-Absi, Lung cancer medical images classification using hybrid CNN-SVM, Int. J. Adv. Intell. Inform. 7 (2) (2021) 151–162.

[28] P. Nanglia, S. Kumar, A.N. Mahajan, P. Singh, D. Rathee, A hybrid algorithm for lung cancer classification using SVM and neural networks, ICT Express 7 (3) (2021) 335–341.

[29] Y. Onishi, A. Teramoto, M. Tsujimoto, T. Tsukamoto, K. Saito, H. Toyama, H. . Fujita, Multiplanar analysis for pulmonary nodule classification in CT images using deep convolutional neural network and generative adversarial networks, Int. J. Comput. Assist. Radiol. Surg. 15 (2020) 173–178.

[30] Y.-S. Huang, P.-R. Chou, H.-M. Chen, Y.-C. Chang, R.-F. Chang, One-stage pulmonary nodule detection using 3-D DCNN with feature fusion and attention mechanism in CT image, Comput. Methods Programs Biomed. 220 (2022) 106786.

[31] C. Chen, R. Xiao, T. Zhang, Y. Lu, X. Guo, J. Wang, Z. Wang, Pathological lung segmentation in chest CT images based on improved random walker, Comput. Methods Programs Biomed. 200 (2021) 105864.

[32] S. Makaju, P.W.C. Prasad, A. Alsadoon, A.K. Singh, A. Elchouemi, Lung cancer detection using CT scan images, Procedia Comput. Sci. 125 (2018) 107–114, http://dx.doi.org/10.1016/j.procs.2017.12.016.

[33] K. Odajima, A. Pawlovsky, A detailed description of the use of the kNN method for breast cancer diagnosis, in: Paper Presented at the 2014 7th International Conference on Biomedical Engineering and Informatics, 2014.

[34] N.W.P. Septiani, R. Wulan, M. Lestari, Breast cancer detection using data mining classification methods, Proc. ICMETA 1 (1) (2017) 185–191.

[35] K. Balaji, K. Lavanya, Chapter 5 - medical image analysis with deep neural networks, in: A.K. Sangaiah (Ed.), Deep Learning and Parallel Computing Environment for Bioengineering Systems, Academic Press, 2019, pp. 75–97.

[36] S. Marčelja, Mathematical description of the response of simple cortical cells, J. Opt. Soc. Amer. 70 (1980) 1297–1300, http://dx.doi.org/10.1364/JOSA.70.001297.

[37] I. Sobel, G. Feldman, A 3 ×3 isotropic gradient operator for image processing, Pattern Classif. Scene Anal. 27 (1973) 1–272.

[38] I. Sobel, An isotropic 3 ×3 image gradient operater, Mach. Vis. Three-Dimens. Scenes (1990) 376–379.

[39] R.A. Haddad, A.N. Akansu, A class of fast Gaussian binomial filters for speech and image processing, IEEE Trans. Signal Process. 39 (3) (1991) 723–727.

[40] L.S. Davis, A survey of edge detection techniques, Comput. Graph. Image Process. 4 (3) (1975) 248–270, http://dx.doi.org/10.1016/0146-664X(75)90012-X.