Lung Cancer Detection from LDCT Images Using Deep Convolutional Neural Networks

Shahad Alghamdi, Mariam Alabkari, Fatima Aljishi, Ghazanfar Latif, and Abul Bashar

Abstract Lung cancer is second cancer common to men and women as well as it is one of the world's highest cause of death. Reports in recent years have shown that standard X-rays are not effective in diagnosing lung cancer. It has clinically established that low-dose computed tomography (LDCT)-based diagnosis helps to decreases mortality from lung cancer by 20% relative to normal chest X-rays. Deep learning is considered as one of the most beneficial techniques for lung cancer diagnosis. This technique used in many fields, including healthcare, which helps to facilitate complex tasks, analyze medical images, promote reliable diagnosis, and improve diagnostic accuracy. One of the deep learning algorithms is the convolutional neural network (CNN) and in this paper, different deep CNN based models are proposed for lungs cancer detection. The experiments are performed using dataset acquired from Data Science Bowl 2017 (KDSB17). The dataset consists of 6691 LDCT lung images. For testing the efficiency of the model, the accuracy is reckoned, which represents 91.75%. However, due to the sensitivity of this process, other techniques are also used to assess the model's performance including specificity, sensitivity, recall, precision, and f1-score.

Keywords Deep learning (DL) · Convolutional neural networks (CNNs) · Low-dose computed tomography (LDCT) · Lung image database consortium (LIDC)

1 Introduction

Lung cancer or lung carcinoma, regardless of race, is the principal cause of death from cancer among both men and women. More people die each year from lung cancer than from other popular cancers [1]. It has the lowest several rates among colorectal cancer (65%), prostate cancer (99%), and breast cancer (89%). In 2018, 1.76 million deaths have been estimated [2, 3]. For several decades, lung cancer was

S. Alghamdi · M. Alabkari · F. Aljishi · G. Latif (✉) · A. Bashar
College of Computer Engineering and Sciences, Prince Mohammad Bin Fahd University, Al Khobar, Saudi Arabia
e-mail: glatif@pmu.edu.sa

V. Bindhu et al. (eds.), *International Conference on Communication, Computing and Electronics Systems*, Lecture Notes in Electrical Engineering 733,
https://doi.org/10.1007/978-981-33-4909-4_27

the most prevalent and deadliest form of cancer. It is the most common cancer among men and is the third most common cancer among women worldwide as two million new cases have been reported in 2018 [4]. Statistics show that over 50% of patients with lung cancer die within one year of diagnosis. In 2014, lung carcinoma was Saudi men's fourth most frequent cancer and Saudi women's 7th frequent cancer [5].

Medical imaging generates an immense volume of data, and thousands of images are involved in each medical study. Deep learning is used in healthcare to improve diagnosis accuracy, resolution, and promote reliable and fast diagnosis [6–9]. It can extract and process medical images at a speed and scale that exceed human capabilities and analyze more efficiently. Even though lung cancer is the deadliest type of cancer, it is highly curable if diagnosed early. Catching lung cancer before spreading can add years to human life. It can increase the possibility of survival for five years or more by 55% [1], so it is important to invest in a system that assists in early lung cancer detection. England statistics show that 88% of lung cancer patients who diagnosed at the first stage survived for at least one year compared to 19% of those diagnosed at the fourth stage [2]. Even though it is the number one cause of death from cancer, less money is spent on research into lung cancer as compared with other rising cancers.

Deep learning algorithms consists of several layers which extract higher-level characteristics from the raw input. This consists of several layers for extracting characteristics of higher levels from the raw inputs [10, 11]. Each layer in the network transforms its input into a more complex and abstract representation. The first representational layers learn to detect simple feature filters such as edges and corners while the middle representational layers learn more complex feature detection filters such as part of an object. The last layer will learn to recognize the full object. The research puts in a profoundly convolutional neural network model (CNN) to diagnose lung cancer patients into two classes: have cancer and does not have cancer. The network comprises two convolution layers, two max-pooling layers, two drops out layers, two fully connected layers, and a flattened layer.

In recent years, doctors found that normal X-rays are not appropriate for the diagnosis of lung cancer. They found that low-dose computed tomography (LDCT) decreases mortality from lung cancer by 20% relative to normal chest X-rays. comparison to regular chest radiography [2]. LDCT can detect pulmonary cancer early on, which not measurable with a standard X-ray. Thus, a data set consisting of 6691 LDCT images have been used for the experiments in this research [12].

2 Literature Review

Serj, Lavi, Hoff, and Valls gave a new deep convolutional neural network. (dCNN) model for learning high-level image representation to achieve highly accurate results with low variance using low-dose computed tomography (LDCT) images [13]. The researchers' goal was building a binary classification model that learns discriminant compact features at the beginning of the network to detect lung cancer. The researchers proposed a new deep convolutional neural network architecture consists of three convolution layers, two max-pooling layers, a full-connected layer, and a binary soft-max layer. The proposed model commences with several sequential convolution layers to generate high-order convolutional features. The researchers used a data set from the Data Science Bowl 2017 (KDSB17) to confirm the model results. The dataset consists of 63,890 (LDCT) images of cancer patients and 171,345 images of non-cancers. The researchers divided the data set into testing, training, and cross-validate set; 50% for the training and 25% for the validation and the rest for the testing. The researchers used cross-entropy as a loss function to maximize the probability of having cancer by maximizing the multinomial logistic regression objective. The results of the model were impressive. The performance of the model is measured using specificity (0.991), sensitivity (0.87), and F1 score (0.95).

Chon and Balachandar have used the deep convolutional neural network for lung cancer detection of CT scans [14]. They used a data set from Date Science Bowl 2017 along with updated U-Net that trained on data set LUNA16, which represents CT scans with marked nodules. At first, they started to pass the CT scans into the 3D CNNs directly for classification, which gave them a poor result. Then, to insert only the regions of interest into the 3D CNNs, they had to perform further pre-processing. For this to be the case, U-Net that trained on the LUNA16 data set was used (CT scans with labeled nodules) for identification in CT scans of nodule candidates. This process produced many false positive predictions, so they used the CTs scans to specify where the nodules located as determined by the U-Net outputs were fed into 3D convolutional neural networks. That ultimately helped to identify CT scans for lung cancer either as positive or negative.

Several kinds of research have applied lung cancer diagnosis to manipulating images and machine learning approaches. Makaju, Prasad, Singh, and Alsadoon had two models, their best model was not reliable and failed to identify the findings of cancer found in the nodules [15]. Therefore, they also introduced a new method for detecting cancer nodule from the CT scan image using the watershed segmentation to identify along with SVM to classify a nodule as malignant or benign. This model has identified 92% accuracy of cancer, this is better than the previous model which had 86.6% accuracy. Regarding their data set, they used actual CT scans from the Lung Image Database Consortium (LIDC) archive of patients. This collection of pulmonary cancer CT images for computer-aided diagnostic methods for the identification and treatment of lung cancer, which was introduced by the National Cancer Institute. The dataset was composed of 1018 cases which supported 7 research centers and 8 medical imaging firms. Images were in DICOM format with 512 * 512 pixels in size, and as DICOM format was hard to process; those images have been converted to JPE grayscale images the aid of MicroDicom software, which converts the DICOM CT scan images to JPEG format.

Sharma and Jindal obtained their CT images from NIH/NCI Lung Image Database Consortium (LIDC) data set that is provided for research purposes [16]. They obtained an automated CAD program to diagnose lung cancer early they achieved that by examining the CT images in multiple steps, their method started by extracting pulmonary areas from the CT picture using multiple image processing strategy, including bit image slicing, erosion, and Weiner filter. Instead of using the thresholding technique, they used the bit plane slicing technique that was used to convert the CT image into a binary image during the first stage of the extraction process. This strategy is faster and user autonomous compared to the thresholding technique. Afterwards, the lung regions extracted were segmented using region-widening segmentation algorithms. Lastly, the field has been used to classify cancerous regions and to obtain an objective result with a high sensitivity level of 90%, with a fair number of false positives per picture with 0.05 false positives per picture.

Medial image processing gained more importance in last decade due to the availability of high computational power and advancement in the imaging techniques such as medical image enhancement [17], noise removal [18], medical image classification [19], and medical image segmentation [20].

3 Methodology

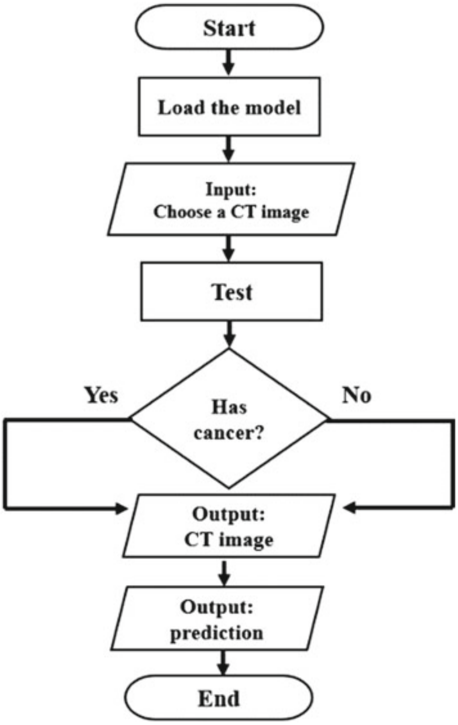
In this section, the models that have been tried will be discussed. The deep convolutional neural networks (dCNN) is generated by an input layer, hidden layers, and an output layer. It composes of two major parts: feature extraction and classification. The first layers learn basic detection filters while the complexity increases in the middle layers. The biggest advantage of dCNN, the developer is not expected to extract features manually from the image. During the training, the network learns to extract features. The classification decision will be guided in the last layers based on the features extracted from the preceding layers. Figure 1 represents the process of lung cancer detection system.

3.1 Classification Through Deep CNN

Medical images are often corrupted by noise, lighting, and affected by artifacts; this could have an impact on model accuracy. So, to limit these phenomena, the data set was divided by 255. 75% of the images were used for the training, and 25% for the testing. To balance the classes in each set, the data has been divided based on the labels, and the random state is set to 42. Before deciding our model, 4 deep convolutional neural network architectures were tried then compared the results and chosen the most accurate, which is the fourth model. Figure 2 shows the workflow of the deep convolutional neural network for lungs cancer detection.

The first model consists of three sequential convolution layers, two max-pooling layers, two fully connected layers, and a flattened layer. Firstly, the network begins with an input layer which is the first convolution layer; the first layer will take the image with an input size of 120 × 120 pixels. It consists of 50 filters and convolution kernel of 11 × 11. Secondly, another convolution layer has been added. The second layer consists of 120 filters and convolution kernel of 5 × 5. Thirdly, a max-pooling

Fig. 1 Process of lung cancer detection system



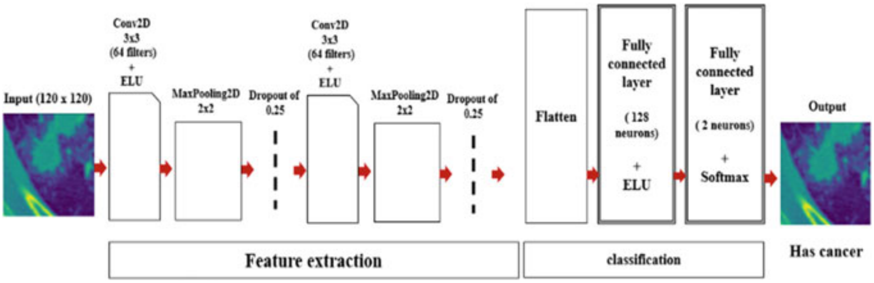


Fig. 2 Workflow of the deep convolutional neural network for lungs cancer detection

layer has been added with a pool size of 2 × 2 and strides of 2. Fourthly, a third convolution layer with 120 filters and convolution kernel of 3 × 3 is added. Fifthly, a max-pooling layer has been added with a pool size of 2 × 2 and strides of 2. RELU is used as an activation function in all the convolution layers. This model did not get the best results, this might be because the flattened layer is added after the first dense layer.

In the second model, the first convolution layer is the input layer. The input layer consists of 32 filters and with the convolution kernel of 2 × 2. After that, the batch normalization technique is used to speed up the training, reduce overfitting, and to put all our data on the same scale. Similar to batch normalization, the drop-out layer will help in reducing overfitting. After that, a convolution layer consists of 64 filters is added with the convolution kernel of 2 × 2. After that, the batch normalization is used and the drop-out layer again. After the feature extraction and normalization, the output of the final convolution layer will be the input of the flatten layer. The flatten layer will convert the data into a one-dimensional array and pass it to the first fully connected layer. After that, a drop-out layer is added between the dense layers because they have the largest number of parameters and could cause overfitting. Finally, the output layer will pass its output to the last fully connected layer, and the classification decision will be made.

The third model consists of an input layer that takes an image input (120, 120, 1); this convolution layer applies 50 kernels (filters) of size 50 with kernel size (11, 11). Then, the process of batch normalization is used to improve training and prevent overfitting. As with batch normalization, the drop-out layer helps to reduce overfitting. After that, a max-pooling layer is used with pool size 2 × 2 and strides of 2. A layer of convolution that consists of 80 filters with 3 × 3 kernel size is applied. Following that, batch normalization is used and the drop-out layer again. Afterwards, max-pooling layer with pool size 2 × 2 and strides of 2 is added. Finally, the output of the last layer will pass through the flatten layer which will convert the data into a 1-dimensional array and pass it to the first fully connected layer.

Because having one input and one output, the network begins with a sequential convolution layer consisting of 64 filters with a convolution kernel of 3 × 3. The layer uses the exponential linear unit (ELU) as an activation function. In addition, it takes the LDCT image of an input size of 120 × 120 pixels. Secondly, a max-pooling layer with a 2 × 2 pool size has been added. This layer was used to minimize the number of parameters along with minimizing the spatial size of the representation. Therefore, the computational cost and overfitting will be reduced [18]. Similarly, a dropout of 0.25 has been added to prevent the model from overfitting. Fourthly, a convolution layer consists of 64 filters a convolution kernel of 3 × 3 is added. The fourth layer uses ELU as an activation function. Fifthly, a max-pooling layer with a 2 × 2 pool size is added. Sixthly, a dropout of 0.25 is added. After that, two flatten layers is added to convert the pooled feature map to a single column to input it to the fully connected layer. Finally, two fully connected layers are added to take the results of the convolution and pooling process and use them to drive a classification decision. As shown in Fig. 1, the first fully connected layer uses ELU as an activation function while the second fully connected layer uses SoftMax. The SoftMax is usually applied in the last layer of the neural networks instead of using RELU, Tanh, or Sigmoid. It is used because it converts the input into values between 0 or 1, so they can be interpreted as probabilities [19]. After constructing the layers, the model is trained for 100 epochs. In addition, 10% of the training set is used for validation. Because Adam optimizer is used, the accuracy increased in each epoch.

3.2 Experimental Data

The data is acquired from the cancer imaging archive (TCIA) as a file of un-labeled scans. The file contains two datasets, ‘ct_slices’ which had 6691 slices of the original scans to help with the sharpness of the image when resizing it [12]. All the images are sorted in a hierarchical data format (HDF5). The second dataset ‘slice_class’ had the labels of each slice. The labels used were 0 and 1 to indicate the presence of cancer. The 6691 images consist of 2526 cancerous images, and 4165 non-cancerous images. Sample images of the CT lung cancer are shown in Fig. 3.

The original size of the images was 64 × 64 pixels, yet all the images were resized to 120 × 120 pixels by using resize function from cv2 library. This helped to obtain clear slices and detect cancer cells accurately while using the model. NumPy library which contains reshape function is used to reshape the images from [6691, 64, 64] to the new shape [6691, 120, 120, 1]. Reshaped it from a three-dimensional array to a four-dimensional array to use it with the model. 6691 represents the number of the images, 120 represents the height and width of the images, and 1 represents the number of channels in the images as all the images are in grayscale. Training the model by using grayscale images may increase the model performance because the model will focus on the shape of the images rather than the colors of each image [21]. Also, used to categorical function to convert the labels to a matrix to use it with the model since having two classes 1: cancerous and 0: non-cancerous. The dataset

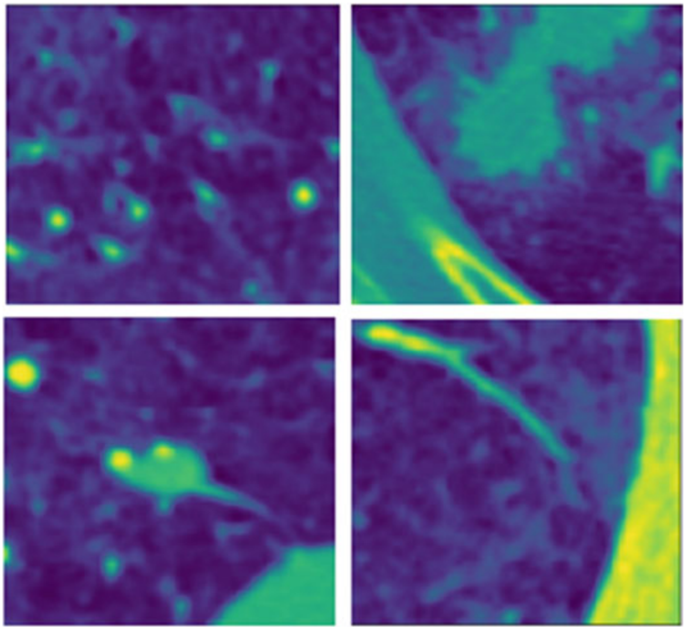


Fig. 3 Sample of low-dose computed tomography (LDCT) lung images

has been divided randomly into three categories: 65% of the images used for the training, 10% used to validate the model during the training process, and 25% used to test the model.

4 Experimental Results

The dataset has been divided randomly into three categories: 65% of the images used for the training, 10% used to validate the model during the training process, and 25% used to test the model. The dataset has been divided based on the labels to guarantee that each category includes images of both classes 0: non-cancer and 1: cancer. To achieve the objective of this research, four different experiments are conducted along with different features for the best results. In the first experiment, the model is trained with 200 epochs, which means 200 training cycles. In addition, the model was complicated, because it contained many layers with a high number of filters and big kernel size. Thus, it was the cause of overfitting. In the second and third experiments, tried to improve the model by decreasing the number of epochs from 200 to 100 cycles. The number of layers used is decreased along with reducing the number of kernels and kernel size. Different parameters are also applied to avoid the overfitting and increase the model performance. This helped to avoid overfitting; however, it affected the accuracy of the model. An accuracy of 91.75% is achieved in the fourth experiment, as long with high model performance.

The accuracy of the four experiments is calculated to evaluate the overall efficacy of the classification process. The accuracy is simply representing the percentage of the predictions that the model has been getting correct. Formally, the accuracy determined by using the formula below for binary classification:

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

Due to the sensitivity of this process, several techniques is used to evaluate our model's efficiency, including specificity, sensitivity, recall, precision, and f1-score. This will help us to calculate the percentage of the right and wrong predictions, which will help to avoid having wrong predictions since the software could have an impact on society as it could be used in medical diagnosis. A wrong diagnosis will affect individuals' health. As shown in the table below, a recall is determined showing the right positive rating average from all the real positive ratings. In addition, the precision is calculated from the cases predicted as positive that represents the rate of correct positive classification. The f1-score helped us to combine the recall and precision values to balance and see the overall results.

In medical image analysis, model performance usually measured using sensitivity and specificity, which is the percentage of true positive and true negative. The geometric mean rate shows a combination of sensitivity and specificity in a single matric.

Table 1 Comparison of the accuracies of different proposed DCNN Models

Epochs	Accuracy (%)
Proposed dCNN model M1	83.02
Proposed dCNN model M2	77.23
Proposed dCNN model M3	71.73
Proposed dCNN model M4	91.75

Different models were implemented to compare their performances and choose the best accuracy among them where the fourth model had the highest accuracy of 91.75% as shown in Table 1. The variance of the accuracy was because of the changes made in each model including, the type of layers added, number of kernels, kernel size, number of epochs (training cycles), and type of activation functions. In addition, more parameters have been added to boost model efficiencies such as kernel initializer, Adam optimizer, Elu activation, and some padding.

A sequential model is created to add multiple layers to the model, where each layer has only one input and one output. The proposed deep convolutional neural network model consists of two convolution layers, two max-pooling layers, two drop-out layers, a flatten layer, and two fully connected layers. Multiple layers are applied and repeated some of the layers multiple times for higher performance, along with extract more features from the images which may help in the classification process.

The two convolution layers applied 64 kernels of the size of 3 × 3. The 3 × 3 filter is the smallest and the most commonly used. In the fourth model, the size of the filter is tried to reduce from 11 × 11 to 3 × 3. This is because a smaller filter size of an odd number is preferred over large filter size of an even number. This will help simplify the image processing for better performance of the convolution layers, along with the number of kernels (filters) used to extract useful features from the images.

Elu activation function is used which more likely to converge cost to zero faster. In addition, Elu tends to produce more accurate results, and it combines the good features of ReLU and Leaky ReLU. Moreover, Elu does not have the main problems of ReLU. The same padding in both convolution layers is applied, which helped to keep the dimensions of the output the same as its input. That means the convolution layers output size is the same as the input. For instance, the input shape of the first convolution layer was (120, 120, 1) and the output shape was also (120, 120, 64). This helped to not reduce the features of the images and lose any important feature, which may help in the classification process.

Two drop-out layers are added after applied the Elu activation function inside the convolution layers; this technique helped us to avoid having overfitting. Each hidden unit (neuron) is set to 0 with a probability of 0.25 at passing 0.25. This means that there is a change of 25% forcing the neuron output to 0. The detailed proposed model's performance analysis is presented in Table 2 which compares different proposed models experimental results based on PrecisionRecall F1-score Specificity.

Adam optimizer is used from Keras library which helped to improve the model and increase the accuracy during the training process. This will guarantee that a

Table 2 In-depth proposed model's performance analysis

#		Precision	Recall	F1-score	Specificity
M1	Non-cancerous	0.87	0.86	0.86	0.79
	Cancerous	0.77	0.79	0.78	0.86
M2	Non-cancerous	0.74	0.98	0.84	0.42
	Cancerous	0.94	0.42	0.58	0.98
M3	Non-cancerous	0.70	0.97	0.81	0.30
	Cancerous	0.86	0.30	0.44	0.97
M4	Non-cancerous	0.92	0.95	0.93	0.86
	Cancerous	0.92	0.86	0.89	0.95

higher accuracy will get at each training cycle. In addition, Keras initializer is used to pass initializers to the layers. In addition, the number of dense is increased, which represents the number of neurons the full connection layer will connect to from 10 to 128 connections; this helped to get a perfect test error.

Lastly, the epoch number was reduced from 200 to 100 to prevent overfitting, because the first model is trained with 200 epochs, and the model started to memorize the images which gave us 1 accuracy during all the training cycles. During this process, the model has trained with only 6691 (LDCT) images, so no need to have a high number of epochs unless the number of images used is increased.

5 Conclusion

Studies have shown that computed tomography (LDCT) based diagnosis helps to decreases mortality from lung cancer by 20% relative to normal chest X-rays. The deep learning technique helped to analyze the (LDCT) medical images and promoted reliable diagnosis and improved the diagnosis accuracy. The paper proposed profoundly convolutional neural network architecture for binary lung cancer classification. Convolutional neural network (CNN) is one of the profound learning algorithms, which used to achieve high classification accuracy for tasks involving medical images. The convolutional neural network (CNN) model consists of two convolution layers, two max-pooling layers, two drop-out layers, a flatten layer, and two fully connected layers. CNN layers were used to extract features from the images, which permit the model to be trained using these features to be able to make the prediction. The models presented, predicted, and classified the cancerous and non-cancerous slices with 91.75% accuracy. However, due to the sensitivity of this process, a lot of methods are used to assess our model's performance, including specificity, sensitivity, recall, precision, and f1-score. These techniques helped to calculate the percentage of true positive, true negative, false positive, and false negative levels, which will help to avoid having wrong predictions since the software could have an impact on society as it could be used in medical diagnosis. This process will help to promote a reliable diagnosis along with improving the diagnosis accuracy.

References

1. Desai A, Gyawali B (2020) Fall in US cancer death rates: time to pop the champagne? EClinicalMedicine 19
2. Rindi G, Klimstra DS, Abedi-Ardekani B, Asa SL, Bosman FT, Brambilla E et al (2018) A common classification framework for neuroendocrine neoplasms: an International Agency for Research on Cancer (IARC) and World Health Organization (WHO) expert consensus proposal. Mod Pathol 31(12):1770–1786
3. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A (2018) Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 68(6):394–424
4. Latif G, Butt MM, Khan AH, Butt O, Iskandar DA (2017) Multiclass brain Glioma tumor classification using block-based 3D Wavelet features of MR images. In: 2017 4th ınternational conference on electrical and electronic engineering (ICEEE). IEEE, pp 333–337
5. Jazieh AR, AlGhamdi M, AlGhanem S, AlGarni M, AlKattan K, AlRujaib M et al (2018) Saudi lung cancer prevention and screening guidelines. Annals Thor Med 13(4):198
6. Kim M, Yun J, Cho Y, Shin K, Jang R, Bae HJ, Kim N (2019) Deep learning in medical imaging. Neurospine 16(4):657
7. Butt MM, Latif G, Iskandar DA, Alghazo J, Khan AH (2019) Multi-channel convolutions neural network based diabetic retinopathy detection from fundus images. Procedia Comput Sci 163:283–291
8. Latif G, Iskandar DA, Alghazo J, Butt MM (2020) Brain MR ımage classification for glioma tumor detection using deep convolutional neural network features Current Med Imag
9. Latif G, Iskandar DA, Alghazo J, Butt M, Khan AH (2018) Deep CNN based MR image denoising for tumor segmentation using watershed transform. Int J Eng Technol 7(2.3):37–42
10. Alghmgham DA, Latif G, Alghazo J, Alzubaidi L (2019) Autonomous traffic sign (ATSR) detection and recognition using deep CNN. Procedia Comput Sci 163:266–274
11. Latif G, Alghazo J, Alzubaidi L, Naseer MM, Alghazo Y (2018) Deep convolutional neural network for recognition of unified multi-language handwritten numerals. In: 2018 IEEE 2nd ınternational workshop on Arabic and derived script analysis and recognition (ASAR). IEEE, pp 90–95
12. Jadhav S (2020) Lung cancer detection using classification algorithms (Doctoral dissertation, Dublin, National College of Ireland)
13. Frank M, Drikakis D, Charissis V (2020) Machine-learning methods for computational science and engineering. Computation 8(1):15
14. Chon A, Balachandar N, Lu P (2017) Deep convolutional neural networks for lung cancer detection. Standford University
15. Makaju S, Prasad PWC, Alsadoon A, Singh AK, Elchouemi A (2018) Lung cancer detection using CT scan images. Procedia Comput Sci 125:107–114
16. Sharma D, Jindal G (2011) Identifying lung cancer using image processing techniques. In: International conference on computational techniques and artificial ıntelligence (ICCTAI), vol 17, pp 872–880
17. Khan AH, Al-Asad JF, Latif G (2017) Speckle suppression in medical ultrasound images through Schur decomposition. IET Image Proc 12(3):307–313
18. Al-Asad JF, Khan AH, Latif G, Hajji W (2019) QR based despeckling approach for medical ultrasound images. Current Med Imag 15(7):679–688
19. Latif G, Iskandar DA, Alghazo JM, Mohammad N (2018) Enhanced MR image classification using hybrid statistical and wavelets features. IEEE Access 7:9634–9644

20. Latif G, Iskandar DA, Jaffar A, Butt MM (2017) Multimodal brain tumor segmentation using neighboring image features. J Telecommun Electron Comput Eng (JTEC) 9(2–9):37–42
21. Joulin A, Cissé M, Grangier D, Jégou H (2017) Efficient softmax approximation for gpus. In: International conference on machine learning, pp 1302–1310