Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 7.056

IJCSMC, Vol. 10, Issue. 2, February 2021, pg.57 – 67

Deep Learning-Based System for Detection of Lung Cancer Using Fusion of Features

Mohamad Shady Alrahhal[1]; Eftkhar Alqhtani[2]
[1]Department of Computer Science, King Abdulaziz University, KSA
[2]Department of Information Systems, Bisha University, KSA
[1] shady.rahal1986@gmail.com; [2] Eftkhar.Alqhtani@gmail.com
DOI: 10.47760/ijcsmc.2021.v10i02.009

Abstract— The cancer detection is doing with the aid of the skilled expert docs and earlier tiers it may be helpful. The opportunity of human error must be there. Employing Deep Learning (DL) in the medical sector is very crucial due to the sensitivity of this field. This means that the low accuracy of the classification methods used for lung detection is a critical issue. This problem is accentuated when it comes to blurry medical images. Moreover, the low accuracy problem is accentuated when DL-based detection systems cannot manipulate the tumour from different angels of views. This paper presents the Adoptive Lung Cancer Detection (ALCD) system, which is built based on the Convolutional Neural Networks (CNN). The ALCD system uses an effective pre-processing phase, to ensure the quality of the medical images, depending on histogram equalization technique. In addition, the CNN is fed by features extracted using Scale Invariant Feature Transform (SIFT). Compared to the state-of-arte, the ALCD system shows better performance in terms of accuracy, sensitivity, and error rate.

Keywords— Lung Cancer, Deep Learning, SIFT, CNN, Detection, Feature Extraction, Accuracy

I. INTRODUCTION

Cancer is the disease in which cells in the body grows out of control. When cancer stats in the lungs it is called as lung cancer. Lung cancer is the leading cause of cancer death and second most diagnosed cancer in both men and women, especially in developing countries. Cigarette smoking, air pollution, and poor healthy nutrition are considered the root cause of lung cancer. Figure 1 illustrates a statistical survey about lung cancer due to smoking.

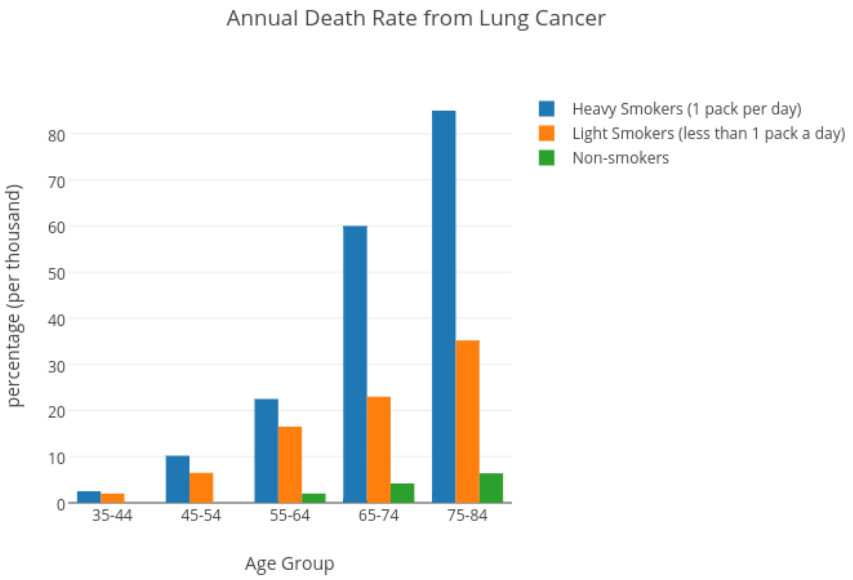

Fig. 1 A statistical survey about lung cancer due to smoking [1].

Statement of problem. Computer science researchers have employed DL for lung cancer detection and other illnesses, and many approaches have been proposed, such as [2-6]. However, the quality of any proposed approach for lung cancer detection is evaluated based on its accuracy. This is to say, due to the sensitivity of diagnosis in the medical sector, a high error rate is critical and may lead to death. In the context of DL, the error rate is represented by the false positive (FP) cases that the intelligent machine fails to classify correctly. In real life, it is harmful for the patient to provide a diagnosis result of negative while the diagnosis is actually positive. Figure 2 illustrates the problem being addressed in this work.

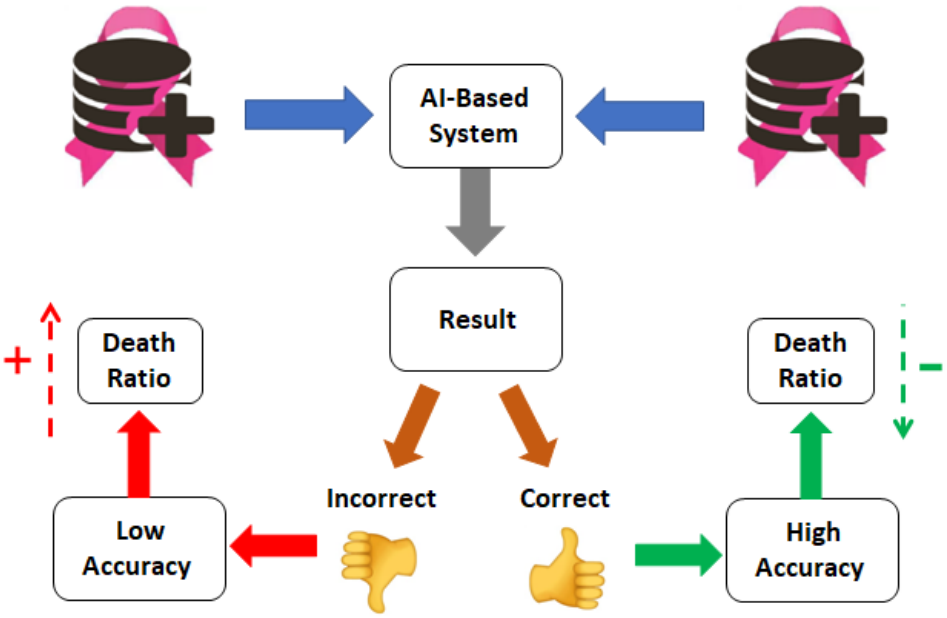

Fig. 2 Problem of DL-based lung cancer detection systems.

Using CNNs can contribute to solve this issue and lead to better accuracy. That is because we can exploit the architecture of the CNN to control the shifting problem based on conventional layers, while the scaling problem can be absorbed by pooling functions.

Contribution. The scientific contribution of this work can be listed as follows:

1. To enhance the quality of the medical images (CT-scan), a histogram equalization based pre-processing step is used. This leads to train the diagnosing medical system on clear and clean medical images, which in turn enhances the level of accuracy.

2. The diagnosing medical system relies establishing a CNN, where features are extracted using advanced technique (SIFT). The importance of this technique comes from its ability to handle C-scan images at the level of pixels. To speed up the training time and lower the computational cost, a fusion is made between the features extracted by the SIFT and the features extracted by the CNN.
3. Extensive experiments are conducted to show the effectiveness of the proposed methods and compare with similar approaches.

Organization of paper. The rest of this work is organized as follows. Section II reviews the related work. Section III provides the proposed CNN classification model. In section V, presents the experiments and discusses. Finally, the paper is concluded in section VI.

II. RELATED WORK

Several medical image detection approaches have been proposed by researchers in general and for lung cancer detection in particular. We categorize them into two major approaches. These are discriminative models and handcrafted models. Discriminative models construct detail information about lung anatomy. It digs out low level features and automatically models the relation of original features and its corresponding label. On the other side, different from the discriminative methods, the hand crafted methods are immensely focus on domain specific understanding about the identification of labels. Nodule presence is challenging to describe, and previous handcrafted methods commonly misscreening the true labels, i.e., the predicted label is not akin with the ground truth [7,8].

In [9] considers the implementation of the lung cancer prediction system by utilizing the convolution neural network that overcomes the issues relating to manual cancer prediction. During this process CT scan images are collected and processed using the layer of neural network that makes automatic extraction of the image features, which are processed using the deep learning process for prediction of the cancer related features making use of the utilization of huge volume of images. The authors have created a system that helps making the decision while analysing the patient CT scan report. In [10] have predicted lung nodules from CT scan images using the convolution neural network method. During this process LIDC IDRI database images are collected and fed into the stack encoder (SAE), convolution neural network (CNN) and deep neural network (DNN) for effective classification of the lung cancer related feature as benign and malignant. The author has introduced a system that ensures up to 84.32% accuracy.

In CNN, medical images are directly processed, most often without segmenting the lung fields, throughout several convolutional layers that work as spatially localized filters and fully connected layers. These CNN architectures involve a very large number of parameters or weights in order to encode the images into a compact high-level feature space. To adjust the CNN parameters, training should be performed using a very large number of data images to avoid under-fitting. The extracted deep learned feature vector has shown a significant capability to describe precisely the training data and distinguish between normal and cancerous lung nodules [11]. For example, Jin et al. [12] trained a 3D CNN architecture, composed of eleven layers, using the segmented lung regions for the task of lung nodule detection. Their method achieved a detection accuracy of 87.5%.

III. PROPOSED DL-BASED SYSTEM

Building the proposed DL-based detection system goes through 6 steps, as shown in Figure 3.

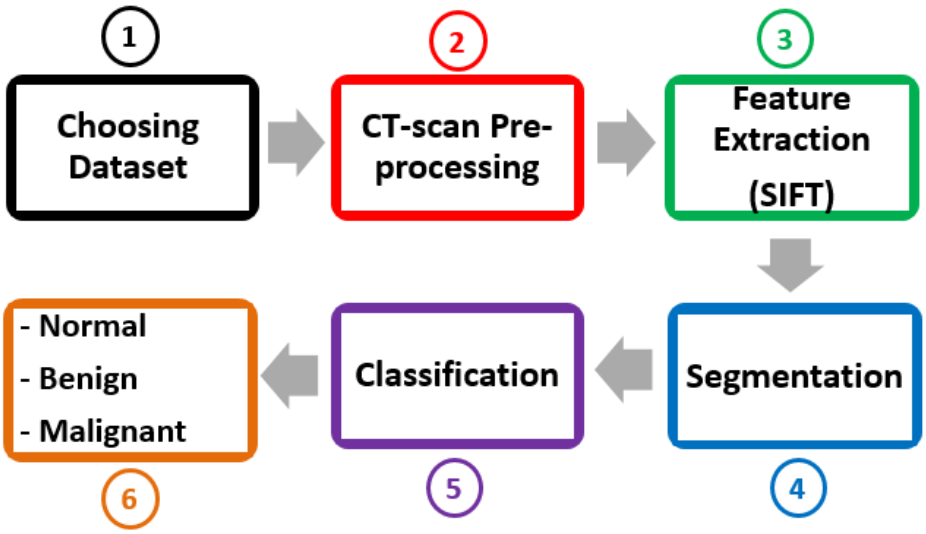

Fig. 3 Steps of DL-based detection systems.

A. Choosing Dataset

The datasets used are taken from Kaggle Data Science Bowl 2017 challenge (KDSB) and LUNA 16 [13]. These datasets are a collection of Computed Tomography scan images, in DICOM format. For each individual patient, in both datasets, there are about 100–400 axial slice of size 512 height and 512 widths. The two datasets are not similarly labelled. The Kaggle Data Science Bowl dataset contains 2,101 labelled data. It is labelled as 0 and 1.

B. CT-scan Pre-processing

The objective of this step is to enhance the quality of the CT-scan lung images. This in turn leads to higher classification accuracy. The reason behind this is that, the clearer edges and details are, the greater the ability to identify the area of the tumour is, and consequently the more accurate diagnosis is. Histogram equalization technique is used to end this step. Histogram is a graphical representation of the intensity distribution of an image. In simple terms, it represents the number of pixels for each intensity value considered. Histogram Equalization is a computer image processing technique used to improve contrast in images [14]. It accomplishes this by effectively spreading out the most frequent intensity values, i.e. stretching out the intensity range of the image. This method usually increases the global contrast of images when its usable data is represented by close contrast values. This allows for areas of lower local contrast to gain a higher contrast. Mathematically, the histogram equalized image $(img = x_{image}^{c-Scan})$ will be defined by

$$img_{y.t} = floor((y - 1) \times \sum_{i=0}^{f_{y.t}} p_i) \quad (1)$$

$$p_i = \frac{number\ of\ pixels\ with\ intensity\ i}{total\ number\ of\ pixels} \qquad i = 0.1. \ldots . y - 1 \quad (2)$$

Figure 4 shows the impact of histogram equalization in terms of enhancing the quality of medical images.

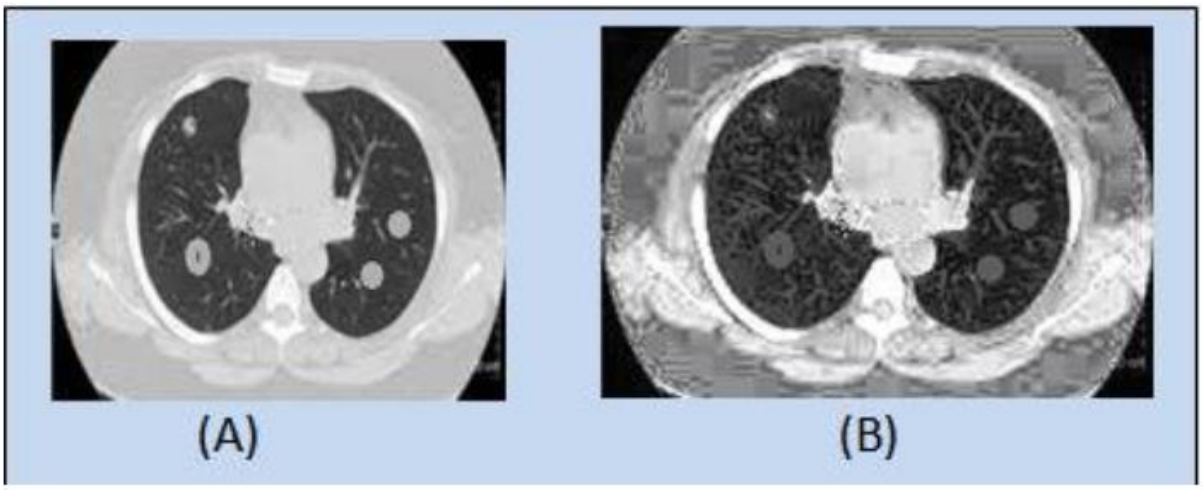

Fig. 4 lung CT-scan image before (A) and after (B) histogram equalization.

This step also contributes to make the proposed system adobtive against noisy medical images. the reason behind this is the medical images will not be allowed to enter the training phase until cleaning process if finished. Therfore, the noise is removed and the poor quality of the medical images is enahnced.

C. Features Extraction (SIFT)

The objective of the SIFT method is to extraction of features that are stationary even under change in rotation or scale of an image. In general, depending on some interesting points, the rotation invariance is guaranteed. This can be achieved by manipulation both the gradient orientations and the magnitudes of the pixels that are located as a neighbors to the interesting points. As for the scale invariance, it is guaranteed by utilizing a scale space based method [15]. The SIFT has five steps, as illustrated in Figure 5.

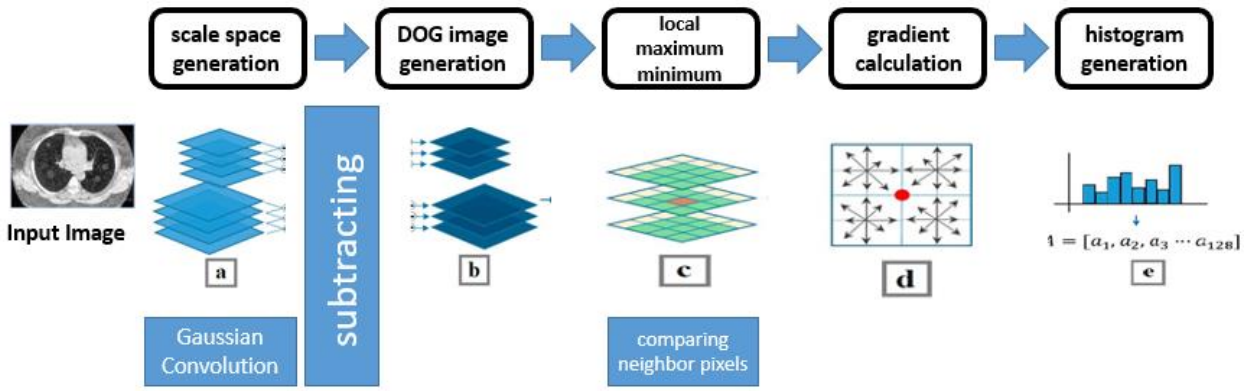

Fig. 5 steps of SIFT method.

The SIFT method used in this work can be seen as asupportive layer to the DL-based system. The DL-based system uses CNN to establish the classiifer. The features extracted from the CNN and the features extracted from the SIFT are fused to generate the final features that the calssifier will train on. Figure 6 shown the structue of the CNN used in this work.

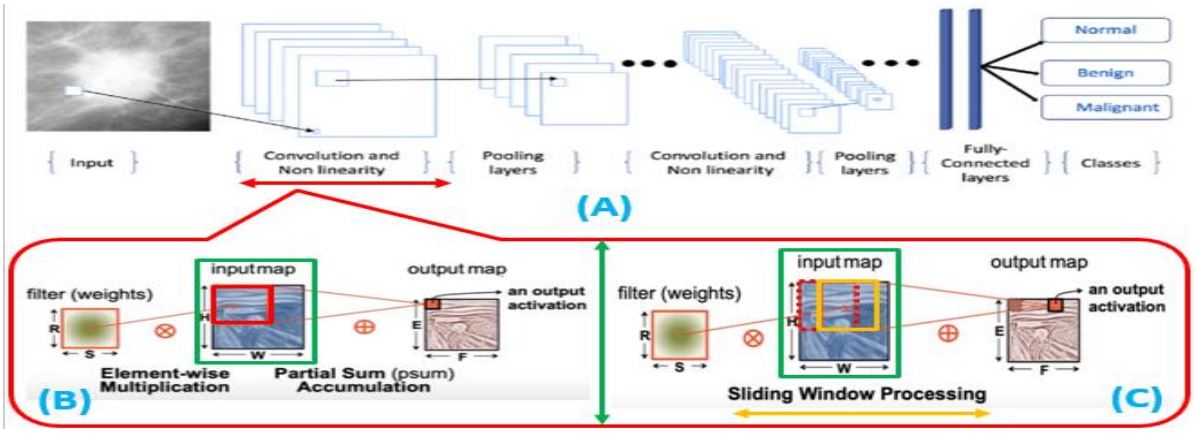

Fig. 6 structure of the CNN.

D. Segmentation

The objective of this step is to identify the edges as well as the features included within the edges (i.e. the pooled features extracted from the previous step). To this end, the Gaussian mixture model (GMM) is used [16]. figure 7 shows the result of the segmentation step on one C-scan images located in the selected data set.

Fig. 7 result of segmentation step.

E. Classification

The objective of this step is to predict the class of the medical image. This means that the classifier should train firstly on the features that are extracted from both the CNN and the SIFT. The features extracted from the both methods are too extensive. This leads to high computational cost as well as long training time [17]. To solve this issue, a fusion is performed among the features obtained from the CNN and from the SIFT, as shown in Figure 8.

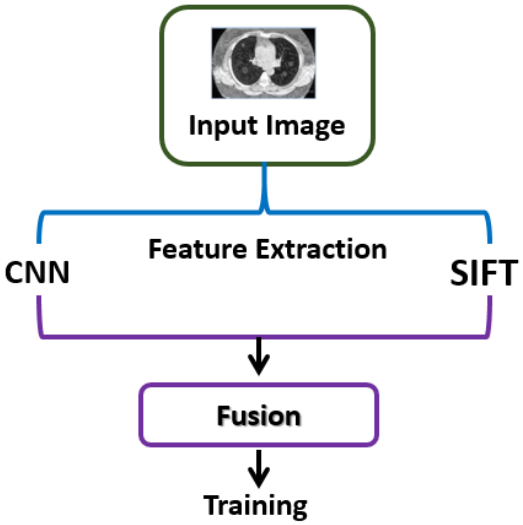

Fig. 8 Fusion of features.

The mechanism of fusion relies on the degree of similarity between tow given features. To end this, the correlation metric is employed to measure the level of similarity.

After training on the fused features, the fully connected layer in the CNN structure is built. The fully connected layer is linked with the Softmax activation function to perform the classification (prediction) process. Figure 9 illustrates the classification process using Softmax function.

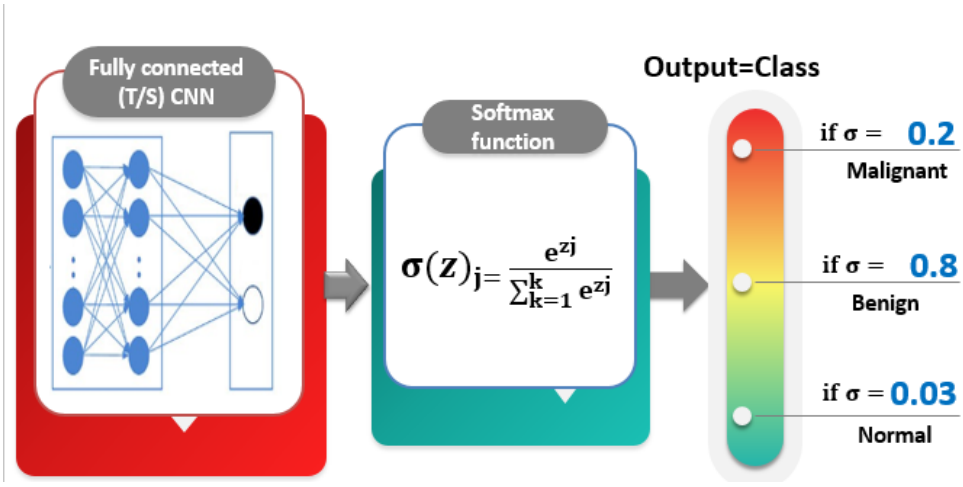

Fig. 9 Using Softmax function in the classification process.

F. Results Generation

In this step, the accuracy of the proposed detection system is calculated. This requires conducting many experiments to obtain the results. This step is described in details in the next section.

It is worth mentioning that building any system requires determining the software technology that is used to define the architecture of the system. Agent based software technology has many befits [18] and can be used to build such detection systems. However, the main issue related to build medical systems using agent software technology is related to the security issue [19, 20] and privacy challenge [21-24]. In addition, security of medical images, through steganography as an example [25], and ensuring high performance are required [26]. Such issues are considered as a future work.

IV. Results and Discussion

This section is organized so that it starts with setup, followed by presenting the systems that are involved in the comparison. Then the used metrics are provided for evaluation purpose. Finally, the results and discussions are presented.

A. Setup

The proposed DL-based system is executed on a laptop that has the specifications organized in Table I, where Matlab programming language is used for implementation.

TABLE I
SPECIFICATIONS

Item	Details (value)
Operating system	Microsoft Windows 10
System type	x64-based PC
System model	Dell Laptop 15-bs0xx
Processor	Intel(R) Core(TM) i7-7200U CPU @ 3.00 GHz
RAM	6 GB

Figure 10 shows the implemented interface of the DL-based system.

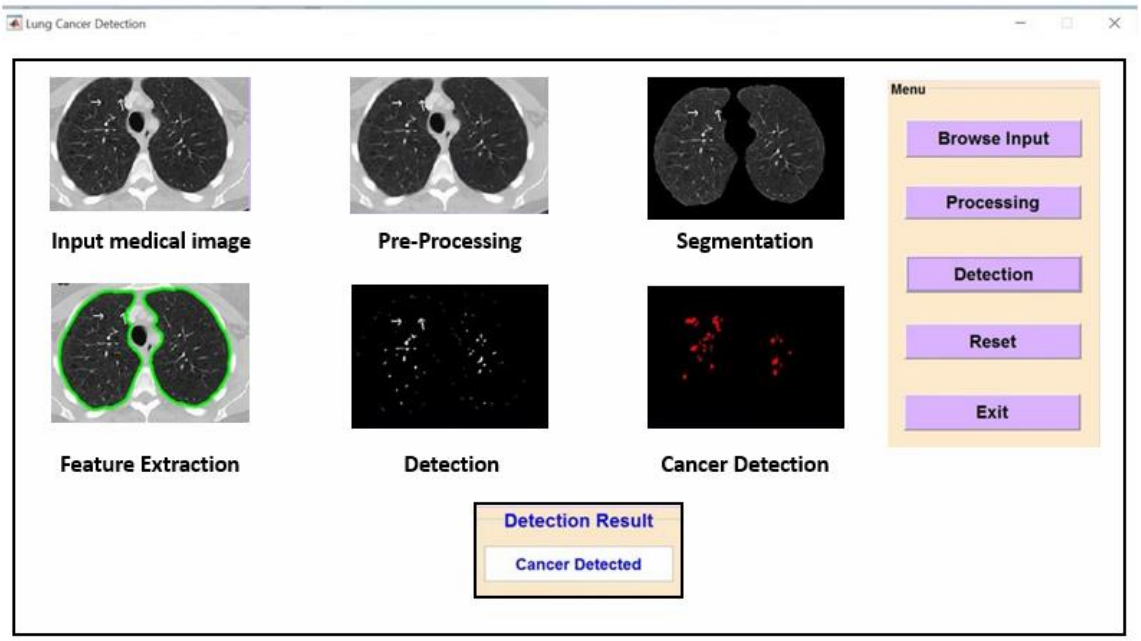

Fig. 10 Interface of the proposed DL-based detection system.

B. Systems Involved in Comparison

We compare the proposed system with two systems. The first system is built based on the decision tree technique, while the second one is built based on logistic regression technique. Table II summarizes the two systems.

TABLE II
SYSTEMS FOR COMPARISON

System name	Used technique for classification	Pre-processing	Features extraction method
DT system	Random forest	Smoothing	Only CNN
LG system	Logistic regression	Contrast enhancing	Only CNN

C. Used Metrics

Three types of metrics are presented for use in the evaluation process. They are DL-based metrics, performance-based metrics, and quality of image metrics.

For the DL-based metrics, confusion matrix is employed to draw the accuracy, sensitivity, precision, and recall metrics. The confusion matrix is given by Table III.

TABLE IIII
CONFUSION MATRIX

Actual class (Predicted class)	**Confusion matrix**		
	X1	***¬ X1***	***Total***
X1	True Positives (TP)	False Negatives (FN)	TP + FN = P
¬ X1	False Positives (FP)	True Negatives (TN)	FP + TN = N

Accuracy is the percentage of the test set images that are correctly classified. The accuracy is defined as:

$$Acc = \frac{(TP+TN)}{number\ of\ all\ records} \quad (3)$$

Sensitivity refers to the true positive recognition rate. It is given by:

$$Sen = \frac{TP}{P} \quad (4)$$

For precision, it refers to the exactness (what % of tuples that the classifier labelled as positive that are actually positive). It is given by:

$$Pre = \frac{TP}{TP+FP} \quad (5)$$

Recall refers to the completeness (what % of positive tuples did the classifier label as positive?). It is given by:

$$Rec = \frac{TP}{TP+FN} \quad (6)$$

Time of response (δ) is used to evaluate approaches in terms of performance. The δ is calculated based on the total time of the four main steps that are illustrated in Figure 6 above (i.e. pre-processing, feature extraction, segmentation, and classification steps). δ is given as:

$$\delta = \Gamma_{pre} + \Gamma_{fext} + \Gamma_{seg} + \Gamma_{clas} \quad (7)$$

Where $\Gamma_{pre}, \Gamma_{fext}, \Gamma_{seg}, and\ \Gamma_{clas}$ denote the time consumed by the pre-processing, feature extraction, segmentation, and classification steps, respectively.

For quality of images metrics, we utilize the mean square error (*MSE*) and peak signal-to-noise ratio (*PSNR*) metrics. These metrics provides numerical values to measure the amount of distortion. The PSNR measures the percentage of the signal to the noise. If its value is high, the quality of the image is good. It is given by:

$$PSNR = 20\ log_{10}\ (\frac{max_i}{\sqrt{MSE}}) \quad (8)$$

Where the MSE is given by:

$$MSE = \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}\left\|G_{prep}^{next} - G_{prep}^{prev}\right\|^2 \quad (18)$$

Where G_{prep}^{next} and G_{prep}^{prev} denote the medical CT-scan lung images after the pre-processing step and before the pre-processing step, respectively.

V. EXPERIMENTAL RESULTS

1) Evaluation from DL-based View: In this type of evaluation, 200 CT-scan lung images are utilized. To obtain the values of DL-based metrics, the confusion matrix is filled during conducting the experiments. Table IV shows the values of the confusion matrix.

TABLE IV
VALUES OF CONFUSION MATRIX

200 lung images 131 (P) 69 (N)	**Values**			
	TP	TN	FP	FN
RF-based system	107	57	25	17
LG-based system	100	54	37	23
ALCD (proposed system)	133	69	19	7

Table V presents the values of the metrics along with the systems involved in the comparison.

TABLE V
VALUES OF DL-BASED METRICS

System	**Metric**			
	Acc	Sen	Rec	Pre
RF-based system	73%	84%	63%	75%
LG-based system	75%	87%	68%	77%
ALCD (proposed system)	96%	98%	92%	96%

The proposed ALCD system achieves the highest values among the others. That is because of the fusion step that helps to use the most useful features extracted from both the CNN and the SIFT. The LG-based system shows better values when compared to the RF-based system. The reason behind this is related to the confusion that may happen in the decision trees due to the large number of branches in the constructed trees.

2) *Evaluation from Performance-based View:* Figure 11 shows the time consumed during the execution of the three systems, where a time counter is used to represent the time of response.

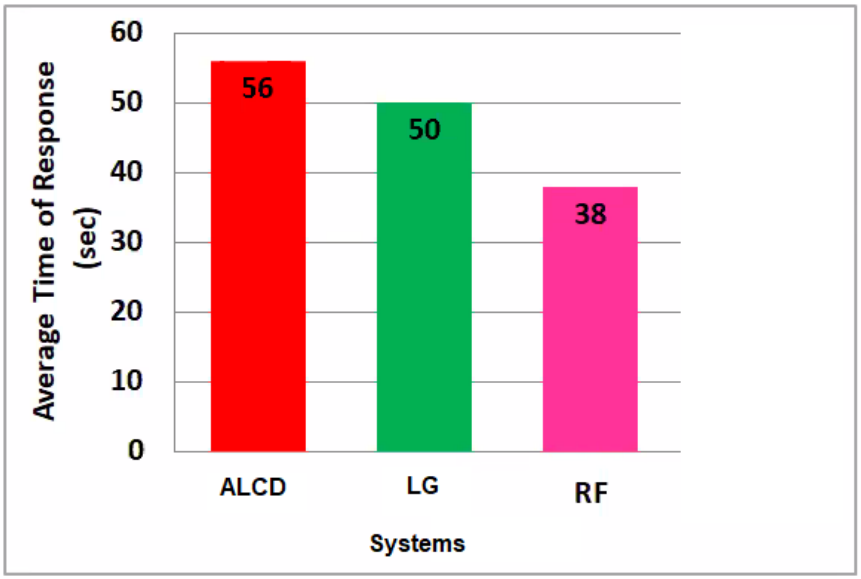

Fig. 11 Value of response time for the three systems.

As shown in Figure 11, the worst system is the proposed system (ALCD). This can be justified by the additional steps that are used to support the system (i.e., to enhance the accuracy level). The RF-based system is ranked in the top because of the fast construction of the decision trees. The LG-based system takes more time when compared to the RF-based system since it depends on mathematical functions, which requires more tome to be conducted.

3) *Evaluation from Quality of Image-based View:* Table VI shows the values of both the PSNR and the MSE for the methods used in each system for the purpose of enhancing the quality of the CT-scan lung images.

TABLE VI
VALUES OF MSE AND PSNR METRICS

Noise Cleaning Method	Metric	
	MSE	PSNR
Contrast enhancing	26.94469	12.463
Smoothing	30.6663	12.1618
Histogram equalization	11.70	28.98

Table VI shows that the histogram equalization used, for enhancing the quality of the medical images, in the ALCD system hits the highest values of PSNR and the lowest values of MSRE. This reflects the highest level of quality. That is because histogram equalization technique includes operations related to soothing and contrast improving.

VI. CONCLUSIONS

Recently, detecting lung cancer using deep learning has received wide attention from researchers. Employing Deep Learning (DL) in the medical sector is very crucial due to the sensitivity of this field. This means that the low accuracy of the classification methods used for lung detection is a critical issue. This paper presents the Adoptive Lung Cancer Detection (ALCD) system, which is built based on the Convolutional Neural Networks (CNN). The ALCD system uses an effective pre-processing phase, to ensure the quality of the medical images, depending on histogram equalization technique. In addition, the CNN is fed by features extracted using Scale Invariant Feature Transform (SIFT). The proposed system shows better performance in terms of accuracy of detection (96%) when compared to the random forest and logistic regression based systems.

Limitation. The proposed system suffers from the longtime of processing due to the use of histogram equalization in the pre-processing step.

Future work. In future work, a parallel based approaches can be used to enhance the response time depending on distributing the training phase on many threads.

REFERENCES

[1] Brain, Kate, et al. "Impact of low-dose CT screening on smoking cessation among high-risk participants in the UK Lung Cancer Screening Trial." Thorax 72.10 (2017): 912-918.
[2] Ardila, Diego, et al. "End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography." Nature medicine 25.6 (2019): 954-961.
[3] Alrahhal, Mohamad Shady, and Majed Abdullah Albarrk. "A Survey of the COVID-19 Epidemic Through the Eyes of Artificial Intelligence and Deep Learning: Challenges and Research Questions." (2020).
[4] Ciompi, Francesco, et al. "Towards automatic pulmonary nodule management in lung cancer screening with deep learning." Scientific reports 7.1 (2017): 1-11.
[5] Coudray, Nicolas, et al. "Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning." Nature medicine 24.10 (2018): 1559-1567.
[6] Lakshmanaprabu, S. K., et al. "Optimal deep learning model for classification of lung cancer on CT images." Future Generation Computer Systems 92 (2019): 374-382.
[7] Clark M C, Hall L O, Goldgof D B, Velthuizen R, Murtagh F R, Silbiger M S. Automatic tumor segmentation using knowledge-based clustering. IEEE Transaction on Medical Imaging, 1998, 17(2): 187–201
[8] Lin D T, Yan C R. Lung nodules identification rules extraction with neural fuzzy network. In: Proceedings of the 9th International Conference on Neural Information Processing. 2002, 2049–2053

[9] Yu Guo, Yuanming Feng, Jian Sun, et al., Automatic lung tumor segmentation on PET/CT images using fuzzy markov random field model, Comput. Math. Methods Med. 2014 (2014) 6, https://doi.org/10.1155/2014/401201, Article ID 401201.
[10] Ayman El-Baz, Garth M. Beache, Georgy Gimel'farb, et al., Computer-aided diagnosis systems for lung cancer: challenges and methodologies, Int. J. Biomed. Imaging 2013 (2013) 46, https://doi.org/10.1155/2013/942353, Article ID 942353.
[11] Setio, A. A. A., Traverso, A., De Bel, T., Berens, M. S., van den Bogaard, C., Cerello, P., Chen, H., Dou, Q., Fantacci, M.E., Geurts, B. and van der Gugten, R. (2017). Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge. Medical image analysis, 42:1-13. https://doi.org/10.1016/j.media.2017.06.015
[12] Jin, T., Cui, H., Zeng, S., and Wang, X. (2017). Learning deep spatial lung features by 3D convolutional neural network for early cancer detection. International Conference on Digital Image Computing: Techniques and Applications (DICTA). pp. 1-6 https://doi.org/10.110 9/DICTA.2017.8227454
[13] Sori, Worku J., et al. "DFD-Net: lung cancer detection from denoised CT scan image using deep learning." Frontiers of Computer Science 15.2 (2021): 1-13.
[14] Fu, Xueyang, and Xiangyong Cao. "Underwater image enhancement with global–local networks and compressed-histogram equalization." Signal Processing: Image Communication 86 (2020): 115892.
[15] He, Yi, et al. "Optimization of SIFT algorithm for fast-image feature extraction in line-scanning ophthalmoscope." Optik 152 (2018): 21-28.
[16] Riaz, Farhan, et al. "Gaussian mixture model based probabilistic modeling of images for medical image segmentation." IEEE Access 8 (2020): 16846-16856.
[17] Alrahhal, Mohamad Shady, and Adnan Abi Sen. "Data mining, big data, and artificial intelligence: An overview, challenges, and research questions." (2018).
[18] Alluhaybi, Bandar, et al. "A Survey: Agent-based Software Technology Under the Eyes of Cyber Security, Security Controls, Attacks and Challenges." International Journal of Advanced Computer Science and Applications (IJACSA) 10.8 (2019).
[19] Alluhaybi, Bandar, et al. "Dummy-Based Approach for Protecting Mobile Agents Against Malicious Destination Machines." IEEE Access 8 (2020): 129320-129337.
[20] Alluhaybi, Bandar, et al. "Achieving self-protection and self-communication features for security of agent-based systems." (2020).
[21] Alrahhal, Mohamad Shady, Maher Khemakhem, and Kamal Jambi. "Agent-Based System for Efficient kNN Query Processing with Comprehensive Privacy Protection." International Journal Of Advanced Computer Science And ApplicationS 9 (2018): 52-66.
[22] Alrahhal, Mohamad Shady, Maher Khemakhem, and Kamal Jambi. "A SURVEY ON PRIVACY OF LOCATION-BASED SERVICES: CLASSIFICATION, INFERENCE ATTACKS, AND CHALLENGES." Journal of Theoretical & Applied Information Technology 95.24 (2017).
[23] Alrahhal, Mohamad Shady, Maher Khemekhem, and Kamal Jambi. "Achieving load balancing between privacy protection level and power consumption in location based services." (2018).
[24] Alrahhal, Mohamad Shady, et al. "AES-route server model for location based services in road networks." International Journal Of Advanced Computer Science And Applications 8.8 (2017): 361-368.
[25] Al-Rahal, M. Shady, Adnan Abi Sen, and Abdullah Ahmad Basuhil. "High level security based steganoraphy in image and audio files." Journal of theoretical and applied information technology 87.1 (2016): 29.
[26] Fouz, Fadi, et al. "Optimizing Communication And Cooling Costs In Hpc Data Center." Journal of Theoretical and Applied Information Technology 85.2 (2016): 112.

