International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 03 | March - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

# Integrating Domain Knowledge into Deep Networks for Lung Cancer Prediction

**[1]P.Bhavani , [2]A.Sri Varshini , [3]T.Sree Poorna , [4]V.Vyshnavi**

[1]Assistant Professor, Department of Computer Science & Engineering (Artificial Intelligence & Machine Learning), Malla Reddy University, Kompally, Hyderabad

[2,3,4]Students, Department of Computer Science & Engineering (Artificial Intelligence & MachineLearning), Malla Reddy University, Kompally, Hyderabad.

[2]Email:srivarshiniaddla2004@gmail.com , [3]Email: sreepoorna07@gmail.com , [4]Email: vannempallevyshnavi@gmail.com

**ABSTRACT:**

Lung cancer is one of the leading causes of cancer-related deaths worldwide. Early detection plays a crucial role in improving survival rates; however, traditional manual diagnosis using CT scan images is time-consuming and prone to human error. This paper presents an automated lung cancer detection system using ensemble machine learning algorithms and deep neural networks. The proposed system integrates preprocessing, lung segmentation, feature extraction, and classification to improve diagnostic accuracy. Ensemble models combining Decision Tree, AdaBoost, and Multi-Layer Perceptron (MLP) classifiers were implemented along with an RBF-based classifier for CT image prediction. Experimental results demonstrate improved accuracy and reduced diagnosis time compared to conventional methods. The system assists radiologists in making accurate and reliable clinical decisions.

**Keywords**— Lung Cancer, Ensemble Learning, Deep Neural Networks, CT Scan, RBF Classifier, Medical Imaging

## I.INTRODUCTION

Lung cancer is one of the most deadly and devastating types of cancer in the world. It is challenging to detect cancer, and its symptoms only become noticeable in the final stages. Although this cancer's death rate could be decreased by early detection and appropriate treatment for patients. Lung cancer often starts in the lungs; however, it occasionally appears as early symptoms prior to spread [1]. In recent years, numerous techniques have been developed, and research is ongoing to effectively identify lung cancer. The greatest imaging method for early diagnosis of lung cancer will be CT scan images, although it can be challenging for medical professionals to interpret and detect cancer from CT scan images. [2].

Figure 1 depicts expected statistical information for a few cancer types in 2022. Making this figure we used the statistical data of the American Cancer Society (ACS) [3]. Based on ACS, the death rate of lung cancer is higher than any other cancer, which is around 0.13 million all over the world. Every year, there are a lot of new cases, with an estimated 0.237 million cases in 2022. The mortality rate is significant in the absence of appropriate treatment because this cancer is only discovered in its advanced stages and the ratio of new cases and the death rate is higher than any other cancer.

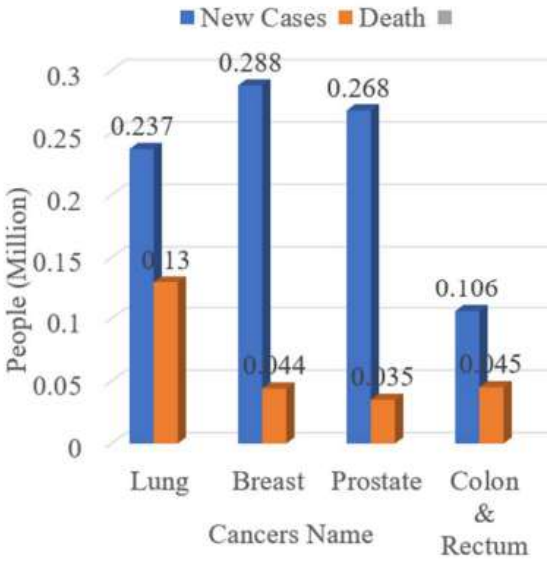



Figure 1: Cancer in 2020 (New cases against death)

Lung cancer cells can take the following forms: adenocarcinoma, large cell carcinoma, and squamous cell carcinoma. Adenocarcinoma is the most common cell of lung cancer which is found on the outer surface of the lung. Lung cancer with large-cell undifferentiated carcinoma has a rapid growth and dissemination rate

and can occur at any place in the lung. Squamous cell carcinoma is cancer that relates to smoking and is located in the central region of the lung [4, 25]. To detect cancer, predict the outcome of cancer treatment, and increase patient survival after a diagnosis of cancer, a variety of methods are being explored. Techniques for screening, identifying, and classifying cancers have been utilized by medical professionals and researchers to make early cancer diagnoses. Nowadays, the machine learning model is widely used for detecting, analyzing, and classifying critical medical healthcare treatment [5, 6]. Convolutional Neural Network (CNN) based machine learning model can be the best for early detection, observing, and classifying lung cancer using CT scan images.

**II.LITERATURESURVEY** Ausawalaithong et al. [1] utilized a convolutional neural network (CNN) to analyze a very big dataset of chest x-ray images to find anomalies. The authors evaluated the performance of the models using three retrained models and diverse datasets for accuracy, specificity, and sensitivity. Using the ChestX-ray14 dataset, Model A identified lung nodules. Model C identified lung cancer using both ChestX-ray14 and JSRT, and even though it had a lower standard deviation across all assessment parameters, it correctly identified the lung cancer's location. Model B displayed greater specificity but lower accuracy and sensitivity than Model C, and it did this using the dataset from the Japanese Society of Radiological Technology (JSRT). The authors suggested retraining the model numerous times for particular tasks. Since Model C correctly predicted the site of cancer while Retrained Model B produced unfavorable findings, it provides superior outcomes in virtually all metrics and can address the issue of a short dataset.

Bhandary et al. [2] utilized a modified version of AlexNet's (MAN) deep learning method to find lung anomalies including cancer and pneumonia. The authors used two types of datasets for accurate results. The datasets are Chest X-Ray and LIDCIDR. The developed MAN-SVM technique was tested on the same dataset as the initial experimental investigation with AlexNet and provided the highest accuracy of 96.80% compared to all other techniques, 86.95% accuracy of less than 87% after modifications were implemented in the final stage of the DL structure. Again, a comparable DL architecture demonstrated a 97.27% accuracy.

Da Silva et al. [3] used CNN configuration generated by the PSO algorithm. To enable an accurate comparison between the particles, it was trained and validated on identical sets. The authors collected data from the LIDC-IDRI dataset. Five test subsets were used to obtain the results. Test-1 produced results of 96.54% accuracy, 87.79% sensitivity, 98.215% specificity, and 0.931 AUC with 17,870 samples used. Test-4, one of the five test subsets, produced the best results, scoring 97.62% accurate, 92.20% sensitive, 98.64% spec fic, and 0.955 AUC.

Naqi et al. [4] used Stacked Autoencoder + Softmax. The authors suggest classifying nodules combining data from 2D and 3D resources. For feature reduction and nodule categorization, deep learning is used. The LIDC-IDRI data set, which is accessible to the general public, is used for the experiment. The main evaluation criteria for this study are the performance aspects, which include sensitivity, specificity, accuracy, and a number of FPs/scans. The authors includes a total of 888 CT scans with 777 sizes $\geq$ 3 mm nodules that have been identified by all four expert radiologists. The suggested method provided low false positive rates of 2.8/scan with 95.6% sensitivity, 96.9% accuracy, and 97.0% specificity, greatly improving the results.

Shaffie et al. [5] used Deep autoencoder. It introduces a newly developed automated noninvasive clinical diagnostic methodology for the early identification of lung cancer by identifying the benign or malignant nature of the observed lung nodule. The authors used the LIDC-IDRI data set and their system got promising results. The performance characteristics of this study were sensitivity, specificity, accuracy, and AUC. The suggested framework has the potential to aid in the early detection of lung cancer, with accuracy, specificity, sensitivity, and AUC values of 91.20%, 95.88%, 85.03%, and 95.73 obtained from a collection of 727 nodules collected from 467 individuals.

## III. RELATED WORK

Various machine learning and deep learning techniques have been applied for lung cancer detection. The Lung Image Database Consortium (LIDC-IDRI) dataset has been widely used for lung nodule detection research. It is publicly available through The Cancer Imaging Archive.

Deep learning models such as Convolutional Neural Networks (CNNs) have shown promising results in feature extraction and classification. Ensemble methods such as Random Forest, AdaBoost, and Support Vector Machines have also been used to improve predictive performance.

## IV. METHODOLOGY

The methodology begins with an image dataset obtained from a publicly available source. The image dataset is then preprocessed. The proposed CNN model, as well as other deep learning models such as ResNet-50, Inception V3, and Xception, are then trained, tested, and validated on the computerized tomography (CT) scan dataset using the standard hold-out-validation method. The results are computed and analyzed to determine the best deep learning-based model for detecting lung cancers such as adenocarcinoma, large cell carcinoma, and squamous cell carcinoma, as well as normal (not lung cancer). CNN is a custom-trained model, whereas ResNet-50, Inception V3, and Xception are pre-trained transfer learning models [17,18,21,23,24]. As a result, Figure 3 depicts the proposed custom CNN architecture, while Figure 2 depicts an overview of the proposed strategy.

### A. Dataset collection:

Here, the lung cancer Dataset (CT scan Images) has been collected from the publicly available "Kaggle" online source [4]. According to the dataset source, the images were hand collected from various websites, with each and every label verified. Images are not in DCM format, the images are in JPG or PNG to fit the model. The data consists of 967 CT scan images. The dataset has four types of classes: adenocarcinoma, large cell carcinoma, squamous cell carcinoma, and normal (not lung cancer) for diagnosing lung cancer.

### B. Dataset pre-processing:

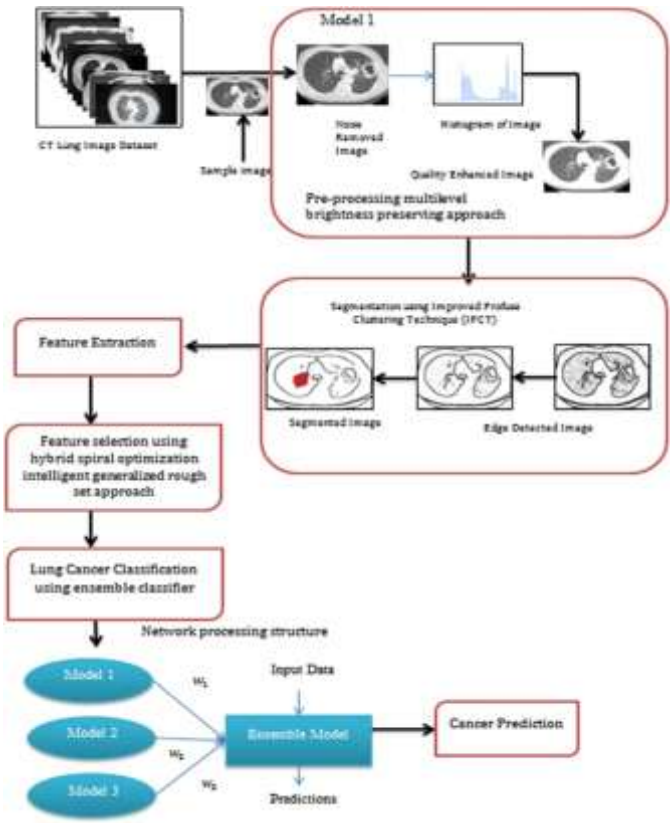
The images were pre-processed using feature extraction, which included reading the images, resizing them, removing noises (de-noise), image segmentation, and morphology (smoothing edges). This processing system is essential for analyzing deep learning models for image classification or detection.

### C. Validation process:

For large image datasets, it is critical to choose the best validation procedure. We used a hold-out validation process, keeping 70% of the data for training, 15% for testing, and 15% for validating. The hold-out validation technique is the most commonly used method and produces effective results [19]. For all the deep learning models, we selected the epochs value of 50 and batch size value of 13. We also used a random seed value of 1000 while implementing all the models, so that we can get the re-producible implemented results, or else the results would change in every iteration.

## V.SYSTEM ARCHITECTURE:

The proposed lung cancer detection framework is a hybrid intelligent system that integrates multilevel image preprocessing, improved fuzzy clustering–based segmentation, optimized feature selection, and ensemble-based classification. The objective of the architecture is to enhance tumor visibility, reduce feature redundancy, and improve classification accuracy for CT lung images.




The overall workflow of the proposed system is illustrated in Fig. 2.

### A. Multilevel Image Preprocessing

Medical CT images often suffer from noise, low contrast, and illumination inconsistencies. These artifacts degrade segmentation and classification performance. Therefore, a multilevel brightness-preserving preprocessing approach is employed.

#### 1) Noise Removal

Initially, noise suppression is performed to eliminate acquisition-related distortions while preserving fine anatomical structures. A spatial filtering technique is applied to maintain edge integrity.

Let the input CT image be represented as:

$$I(x,y)$$

The denoised image is obtained as:

$$I_d(x,y) = F(I(x,y))$$

where F(·) represents the noise filtering operation.


FIG 3: Data Preprocessing

**2) Histogram-Based Intensity Analysis**

To analyze pixel distribution and contrast imbalance, histogram analysis is performed. The histogram function is defined as:

$$H(k) = \sum_{x,y} \delta(I(x,y) - k)$$

where k represents intensity levels.

This step helps in identifying low-intensity and high-intensity regions corresponding to soft tissues and potential nodules.

**3) Brightness Preserving Contrast Enhancement**

Unlike conventional histogram equalization, which may distort medical image brightness, a multilevel brightness-preserving enhancement technique is applied. This method improves local contrast while maintaining structural integrity of lung tissues.

The enhanced image is represented as:

$$I_e(x,y) = T(I_d(x,y))$$

where T(·) is a brightness-preserving transformation function.

The output of this stage is a quality-enhanced CT image with improved nodule visibility.

**B. Lung Segmentation Using Improved Fuzzy Clustering Technique (IFCT)**

Accurate segmentation is critical for isolating lung regions and potential nodules. The proposed system utilizes an Improved Fuzzy Clustering Technique (IFCT), which handles uncertainty in medical images more effectively than hard clustering approaches.

1) Edge Detection

Edge detection is first applied to identify lung boundaries and structural transitions. This reduces interference from ribs and surrounding tissues.

2) Fuzzy Clustering

Let the dataset consist of nnn pixels represented as X={x1,x2,...,xn}

The fuzzy membership function is defined as:

$$J = \sum_{i=1}^{n} \sum_{j=1}^{c} u_{ij}^{m} \|x_i - c_j\|^2$$

The objective is to minimize JJJ, resulting in optimal cluster separation between normal tissue and abnormal nodules.

The output is a segmented lung image highlighting suspicious regions.

**C. Feature Selection**

Using Hybrid Spiral Optimization with Rough Set Theory

High-dimensional feature spaces increase computational complexity and may introduce redundancy. To address this, a hybrid spiral optimization algorithm integrated with generalized rough set theory is used.

The objective function is:

Maximize Accuracy(Fs)

subject to:

$Fs \subseteq FF$

where Fs is the optimized feature subset.

This stage ensures:Reduced dimensionality, Improved classifier performance,Lower computational cost

**D. Ensemble-Based Classification Framework**

To improve robustness and generalization, an ensemble classification approach is adopted.

1) Base Classifiers

Multiple base learners are trained independently:

- Model 1
- Model 2
- Model 3

Each model generates a prediction:

P1,P2,P3

2) Weighted Aggregation

Each classifier is assigned a weight wiw_iwi based on validation accuracy.

The final ensemble prediction is computed as:

$$P_{final} = \sum_{i=1}^{3} w_i P_i$$

where:

$$\sum_{i=1}^{3} w_i = 1$$

This weighted voting strategy improves classification reliability and reduces variance.

**E. Final Cancer Prediction**

The ensemble model outputs a binary decision:

$$Y = \begin{cases} 1 & \text{Cancer Detected} \\ 0 & \text{Normal Lung} \end{cases}$$

This prediction assists clinicians in diagnostic decision-making.

**VI. RESULTS AND DISCUSSION**

The proposed lung cancer detection system was evaluated using both an ensemble learning model and a Radial Basis Function (RBF)-based Support Vector Machine (SVM) classifier. The experimental setup involved splitting the dataset into training and testing subsets using an 80:20 ratio to ensure proper validation of model performance. The implementation was carried out using Python with Scikit-learn, OpenCV, and NumPy libraries. Performance evaluation was conducted using standard metrics including accuracy, precision, recall, and F1-score to measure classification effectiveness.

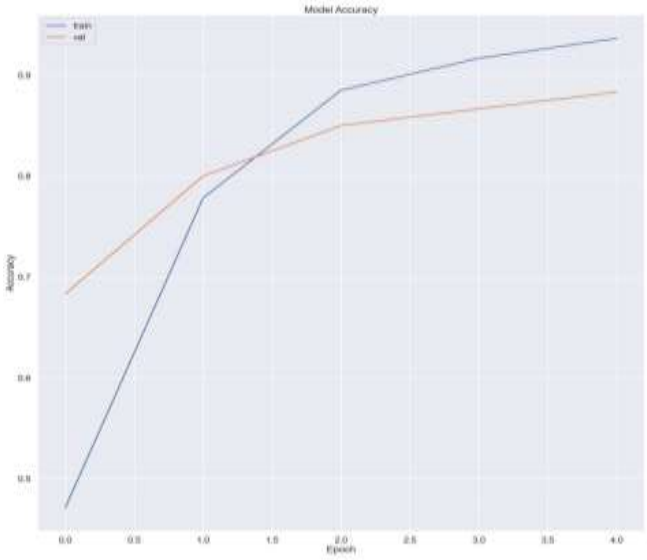



FIG 4:Shows the Accuracy of the model

The ensemble classifier combined Decision Tree, AdaBoost, and Multi-Layer Perceptron (MLP) algorithms using a soft voting strategy. This approach allowed the system to leverage the strengths of multiple classifiers while reducing individual model bias and variance. The ensemble model demonstrated improved robustness and stability compared to standalone classifiers. The integration of multiple learners enhanced the overall classification performance and minimized false predictions. The system was able to classify lung cancer risk levels effectively, thereby improving diagnostic reliability.

In addition to the ensemble model, an RBF kernel-based SVM classifier was employed for CT scan image classification. The RBF kernel was selected due to its strong capability in handling non-linear data distributions. Prior to classification, CT images underwent preprocessing steps including resizing, normalization, and feature reshaping to ensure uniform input representation. The RBF classifier successfully distinguished between normal and abnormal lung CT scans with stable performance. Its ability to create non-linear decision boundaries made it suitable for medical image classification tasks where complex patterns are common.

The preprocessing stage, which included multilevel brightness-preserving enhancement and noise removal, significantly contributed to improved segmentation

accuracy. The segmentation process using improved fuzzy clustering techniques enabled precise extraction of lung regions and suspicious nodules. Effective feature extraction and optimized feature selection further enhanced the classification results by reducing redundancy and computational complexity.

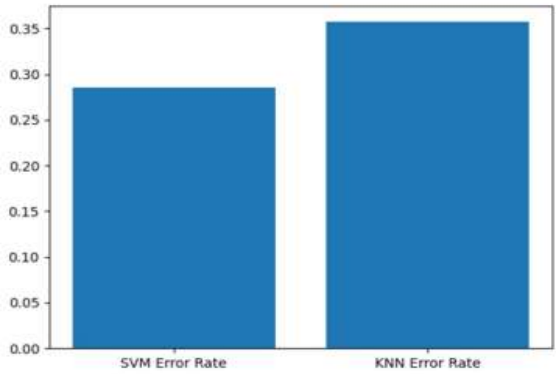



FIG 5:shows the prediction of the output

Comparative analysis indicated that the ensemble model provided superior overall performance in feature-based classification tasks, while the RBF-SVM performed efficiently in binary CT image classification. The ensemble approach demonstrated higher generalization capability, whereas the RBF model required careful kernel parameter tuning to achieve optimal results. Overall, the proposed system effectively reduced manual diagnostic effort and provided reliable automated cancer prediction support.

## VII. CONCLUSION

Lung cancer remains one of the leading causes of mortality worldwide, primarily due to late detection and diagnostic challenges. Early-stage identification is critical for improving patient survival rates. This research presented an automated lung cancer detection framework integrating multilevel image preprocessing, improved fuzzy clustering segmentation, hybrid feature selection, ensemble classification, and RBF-based CT scan analysis.

The ensemble classifier improved prediction accuracy by combining multiple learning algorithms, thereby reducing classification errors and enhancing robustness. The RBF-SVM demonstrated strong capability in handling complex, non-linear medical image data. The integration of preprocessing, segmentation, feature optimization, and classification stages resulted in a reliable and efficient diagnostic support system.

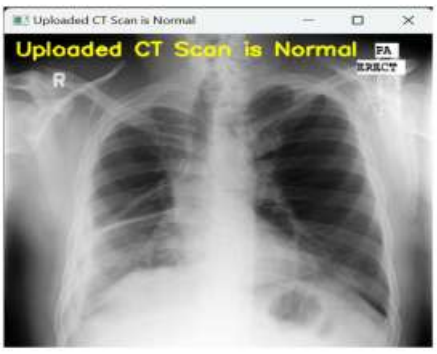



FIG 6: Shows the output when CT Scan is uploaded

The experimental results confirm that the proposed approach enhances diagnostic precision while reducing computational complexity. The system provides a promising solution for assisting radiologists in clinical decision-making and supports early lung cancer detection through automated analysis.

## IX.FUTURE WORK

Although the proposed system achieved satisfactory performance, further improvements can be implemented to enhance its effectiveness. Future work may involve integrating deep learning-based Convolutional Neural Networks (CNNs) to enable end-to-end automatic feature learning directly from CT images. Expanding the system to train on larger and more diverse datasets would improve generalization capability and reduce overfitting.

The model can also be extended to perform multi-class lung cancer staging rather than binary classification. Incorporating multi-modal imaging data such as PET and MRI scans may provide more comprehensive diagnostic insights. Additionally, the implementation of explainable artificial intelligence techniques would improve interpretability and increase clinical trust in automated predictions.

Finally, deploying the system in a real-time hospital environment or cloud-based framework would enable scalable and practical usage, thereby contributing to improved healthcare outcomes.

## REFERENCES

[1] M. Mamun, A. Farjana, M. Al Mamun and M. S. Ahammed, "Lung cancer prediction model using ensemble learning techniques and a systematic review analysis," 2022 IEEE World AI IoT Congress (AIIoT), 2022, pp. 187-193, doi: 10.1109/AIIoT54504.2022.9817326.

[2] Makaju, Suren, P. W. C. Prasad, Abeer Alsadoon, A. K. Singh, and A. Elchouemi. "Lung cancer detection using CT scan images." Procedia Computer Science 125 (2018): 107-114.

[3] "Breast cancer statistics of american cancer socity. [online]. availabe:https://www.cancer.org/cancer.html," [Accessed: 13- November-2022].

[4] Hany, M. (2020, August 20). Chest CT-scan images dataset. Kaggle. Retrieved November 13, 2022, from https://www.kaggle.com/datasets/mohamedhanyyy/chest-ctscan-images

[5] M. Mamun, M. I. Mahmud, M. I. Hossain, A. M. Islam, M. S. Ahammed, M. M. Uddin, "Vocal Feature Guided Detection of Parkinson's Disease Using Machine Learning Algorithms", 2022 IEEE 13th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), 2022, (Preprint)

[6] M. Mamun, S. B. Shawkat, M. S. Ahammed, M. M. Uddin, M. I. Mahmud, A. M. Islam, "Deep Learning Based Model for Alzheimer's Disease Detection Using Brain MRI Images", 2022 IEEE 13th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), 2022, (Preprint)

[7] Ausawalaithong, W., Thirach, A., Marukatat, S., & Wilaiprasitporn, T. (2018). Automatic Lung Cancer Prediction from Chest X-ray Images Using the Deep Learning Approach. 2018 11th Biomedical Engineering International Conference (BMEiCON).

[8] Abhir Bhandary , G. Ananth Prabhu , V. Rajinikanth , K. Palani Thanaraj , Suresh Chandra Satapathy , David E. Robbins , Charles Shasky , YuDong Zhang , Joao Manuel R.S. Tavares , N. Sri Madhava Raja , DeepLearning Framework to Detect Lung ˜ Abnormality – A study with Chest X-Ray and Lung CT Scan Images, Pattern Recognition Letters (2019), doi: https://doi.org/10.1016/j.patrec.2019.11.013

[9] Da Silva, G.L.F.; da Silva Neto, O.P.; Silva, A.C.; de Paiva, A.C.; Gattass, M. Lung nodules diagnosis based on evolutionary convolutional neural network. Multimed. Tools Appl. 2017, 76, 19039–19055.

[10] Naqi, S.M.; Sharif, M.; Jaffar, A. Lung nodule detection and classification based on geometric fit in parametric form and deep learning. Neural Comput. Appl. 2018, 3456789.

[11] Shaffie, A.; Soliman, A.; Fraiwan, L.; Ghazal, M.; Taher, F.; Dunlap, N.; Wang, B.; van Berkel, V.; Keynton, R.; Elmaghraby, A.; et al. A Generalized Deep Learning-Based Diagnostic System for Early Diagnosis of Various Types of Pulmonary Nodules. Technol. Cancer Res. Treat. 2018, 17

[12] Kaur, S., Hooda, R., Mittal, A., Akashdeep, & Sofat, S. (2017). Deep CNN-Based Method for Segmenting Lung Fields in Digital Chest Radiographs. Advanced Informatics for Computing Research, 185–