

Optimized Hyper parameter Approach for Lung Cancer Prediction and classification of types Using GLCM

Prakasha Rajee Urs M¹, Dr. G N K Suresh Babu²

Submitted:14/03/2024 Revised: 29/04/2024 Accepted: 06/05/2024

Abstract: Cancer is one of the most lethal illnesses, causing an uncountable number of deaths worldwide. Diagnosing lung cancer at an earlier stage has attracted significant interest from medical professionals. This research introduces a novel method for detecting lung cancer by processing images obtained from CT scans. Utilizing cases from the Lung Imaging Database Consortium (LIDC) database, this study evaluates the feasibility of applying algorithms to detect lung cancer. The primary aim of this study is to determine whether the tumors found in the lung are malignant or benign. This is accomplished using the GLCM feature extractor and the SVM classifier. The research proposes a hyper parameter optimization approach for identifying lung cancer and classification of its stages. The results of the hyper parameter optimization show an overall prediction accuracy of 95.86%, with more than 90% accuracy in classifying tumors as malignant, benign, or normal. The area under the curve (AUC) values are more than 90% for the malignant, benign, and normal classes, respectively. In the classification of lung cancer and its types including 91.1% for Adenocarcinoma, 91.5% for Large cell carcinoma, 92.9% for Squamous cell carcinoma, and 99.2% for Normal lung tissue.

Keywords: Area Under Curve, Benign, Hyper parameter, Lung Cancer, Lung Imaging Database Consortium Malignant, Optimization.

1. Introduction

According to the World Health Organization (WHO), lung cancer accounts for about 14 percent of all cancers. Therefore, early detection and treatment are crucial [1]. Computed Tomography (CT) scans provide valuable information for diagnosing lung diseases. Lung cancer is a fatal illness affecting a significant number of people worldwide. Each year, lung cancer claims more lives than all other types of cancer combined, including brain, prostate, and breast cancers [2]. In the U.S., lung cancer is one of the most common cancers, with over 225,000 cases, 150,000 deaths, and \$12 billion in healthcare costs annually. It is also one of the deadliest cancers, with only 17% of diagnosed individuals in the U.S. surviving five years post-diagnosis, and survival rates are even lower in developing countries [3-4]. The stage of cancer indicates how extensively it has metastasized. Stages 1 and 2 refer to cancers localized to the lungs, while latter stages indicate cancers that have spread to other organs. Current diagnostic methods include biopsies and imaging, such as CT scans. Early detection of lung cancer significantly improves survival chances but is challenging due to fewer symptoms in early stages. Lung cancer causes more deaths annually than colon, breast, and prostate cancers combined, accounting for about 25% of cancer-related fatalities [5-6]. Sophisticated diagnostic methods,

including chest radiography (X-ray), computed tomography (CT), sputum cytology, and magnetic resonance imaging (MRI), are employed to detect lung cancer in early [7]. Therefore, advancements are needed for the early identification of lung cancer. Individuals over the age of 40 are disproportionately affected [8]. According to the WHO, lung cancer claims over 7.6 million lives annually. Compared to other cancers, lung cancer has the highest mortality rate. Lung cancer accounts for around one-quarter of all cancer-related deaths [9]. There are two types of lung cancer: non-small cell and small-cell lung cancer. It progresses through four stages, starting in the lungs during the first stage and spreading to the chest in the second and third stages [10-11]. By stage 4, it has metastasized throughout the body. Smoking significantly increases the risk of lung cancer. Various procedures, such as X-rays, MRI scans, and CT scans, are used in facilities to detect cancer early [12]. The main downside of these methods is their high cost and the labor involved. Additionally, doctors often assess cancer based on symptoms that typically appear in the later stages of the disease [13]. Lung cancer has a long-term survival rate of 15%, making early detection crucial for improving survival chances. Convolution neural networks (CNNs) have shown better performance than Deep Belief Networks in recent studies on benchmark computer vision datasets [14]. CNNs have garnered significant interest in machine learning due to their strong representation abilities in learning useful features from input data. In this paper, we apply an extensive pre-processing technique to accurately identify nodules, enhancing the accuracy of lung cancer detection. Additionally, we perform end-to-end training of CNNs

¹Research Scholar, Srishti College of Commerce and Management, Banashankari, Bengaluru, Affiliated to University of Mysore, India, 560085.

²Professor, Department of Computer Science, Srishti College of Commerce and Management, Banashankari, Bengaluru, Affiliated to University of Mysore, India, 560085.

from scratch to fully realize their potential in learning discriminative features. Extensive experimental evaluations are conducted on a dataset of lung nodules from over 1,390 low-dose CT scans [15]. Diagnosing lung cancer based on microscopic examination of lung tissue samples has several limitations. One limitation is that doctors often rely on their interpretations. Accurate diagnosis of lung cancer requires careful observation and precise analysis by medical practitioners. Thus, there is a need for an autonomous system capable of diagnosing lung cancer from microscopic biopsy images. This study aims to develop a method for the early diagnosis of lung cancer using digital image processing and analysis of microscopic images obtained from biopsies.

2. Materials and Methods

CAD systems for lung cancer follow this pipeline: image pre-processing, detection of cancerous nodules, reduction of false positives, malignancy prediction for each nodule, and overall malignancy prediction for the CT scan. The proposed methodology for lung cancer classification consists of four phases [16-17]. In phase one, the necessary data is collected from the database at <https://www.kaggle.com/datasets>. In phase two, the images are resized. In phase three, features are extracted using the Gray Level Co-occurrence Matrix (GLCM). In phase four, these extracted features are used for classification, which is performed by a Convolution Neural Network (CNN).

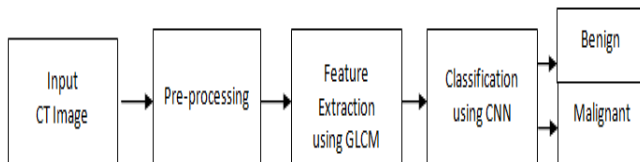


Fig 1: Block diagram of Lung cancer prediction

chest CT images and CNN. The first stage involves extracting the lung regions from the CT images and segmenting each slice within these regions to identify malignancies. The CNN architecture is trained with the segmented tumor regions. The dataset is divided into three sets for training, validation, and testing in a 70%:20%:10% ratio once the images are prepared in their binary matrix format. The patient images are then evaluated using the CNN [18].

2.1 Feature Extraction

Feature extraction is applied to extract features like energy, entropy, variance. Normality and abnormality of an image can be determined in this stage [19]. Various features of an image can be Contrast, Energy, Homogeneity, Mean, Standard Deviation, Entropy, RMS, Variance, Smoothness etc. Each object is extracted for its features based on certain parameters and is then assigned a certain class. The

GLCM feature extraction method is a matrix that describes the occurrence frequency of two pixels.

2.2 SVM classifier

Although Support Vector Machines (SVM) is typically thought of as a classification tool, it can also be used to solve regression problems. Many continuous and categorical variables can be handled with ease. To divide various classes, SVM creates a hyper plane in multidimensional space [20].

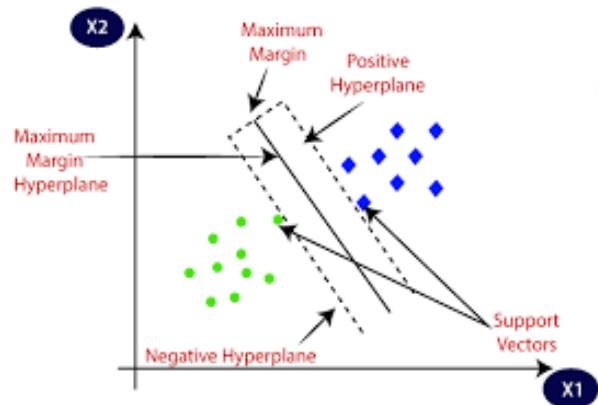


Fig 2: Block diagram of Lung cancer prediction

SVM generates ideal hyper plane in an iterative manner, which is utilised to minimise an error. Finding a maximum marginal hyper plane (MMH) that optimally classifies the dataset is the central goal of SVM. SVM's primary benefit is that it can be applied to both classification and regression issues. Here, we have developed SVM classifier for multiclass application. In the proposed method SVM must classify the lung CT images into benign, malignant and Normal class. The above-mentioned approach ensures that with larger margins, the classifier yields a lower generalization error. Separation hyper planes as shown in Figure 2, in which H1 does not divide the two classes; H2 divides them but with a small separating margin while H3 separates the classes with the best margin.

3. Results and Discussions

For the computation processes we have considered the Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) database, provide training data set. The size of CT image will be 512x512x3. At the time of feature extraction this image of both the sets are resized to 224x224x3. Here, 500 images have been used to conduct the experiment. Out of the available data, 70% of image data for training the model and other 30% for verifying the result and for checking accuracy of the network. Our research has carried out with and without hyper parameter tuning process to predict the possibilities of the disease in the lung images and classification of its types using hyper parameter. furthermore, statistical parameters like True Positive Rate (TPR), False Positive Rate, Accuracy and Confusion matrix have been computed

and tabulated to measure the performance of the system. Here, class 1 indicates benign, class 2 indicates malignant and class 3 as normal case condition.

3.1 Without Hyper parameter Tuning:

The proposed method has carried out using the combination of GLCM and SVM with and without hyper parameter tuning. The extracted features from GLCM are serve as inputs for the SVM classifier, whereas the outputs are categories of Benign (1), Malignant (2) and Normal (3). In Table 1 prediction of these classes for first 10 cases has been tabulated and it shows that in few cases class 2 which malignant is wrongly predicted as class 1 which is Benign and class 3 which is Normal.

Table 1: Prediction chart Without Hyper parameter Tuning

TrueLabel	PredLabel	Posterior		
2	{ '1' }	0.50919	0.32321	0.1676
2	{ '2' }	0.10901	0.87188	0.019109
2	{ '1' }	0.50919	0.32321	0.1676
2	{ '2' }	0.11797	0.87249	0.0095412
2	{ '2' }	0.11885	0.8716	0.0095529
3	{ '3' }	0.4124	0.1327	0.4549
2	{ '3' }	0.31966	0.22883	0.45152
2	{ '2' }	0.0151	0.96814	0.016756
2	{ '2' }	0.21438	0.68082	0.10479
2	{ '2' }	0.20042	0.68341	0.11617

The proposed model overall accuracy has been calculated and tabulated here considering benign, malignant, and normal class. In developing any system, k-fold cross-validation helps us to build the model as a generalized one. Here, 5-fold cross validation has been achieved and results corresponding to each fold is evaluated and tabulated as in Table 2.

Table 2: Lung cancer detection - without Hyper parameter

Cla	FPR	TPR	Preci	Rec	F1-Scor	Accu
ss			sion	all	e	racy
1	0.09	0.90		0.87	0.87	
	6	4	0.871	8	4	0.894
	0.02	0.97		0.84	0.88	
2	5	5	0.926	0	1	0.938
	0.04	0.95		0.96	0.99	
3	7	3	0.898	3	4	0.956

Here, figure 3 and figure 4 represents result pertaining to one of the fold out of 5-fold validation outcome. Confusion matrix evaluates the performance of the classification models, when they make predictions on test data, and tells how good our classification model is. It not only tells the error made by the classifiers but also the type of errors such as it is either benign (B), malignant (M) and Normal (N). Here, our model is providing an highest accuracy of 95.6% in detecting the levels like Normal and with a lowest accuracy of 89.4% while detecting Benign. overall

accuracy of the system will results in 92.93% towards detection of cancer

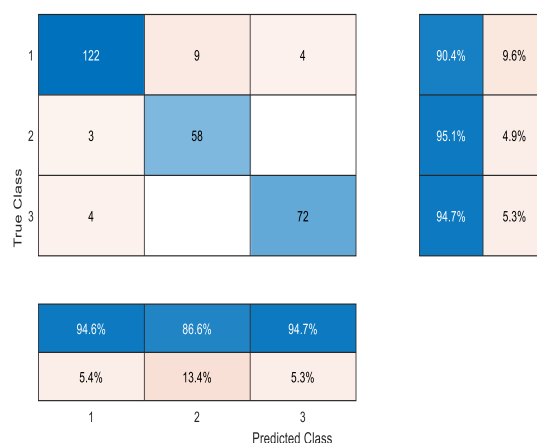


Fig 3: Confusion Matrix Without Hyper parameter

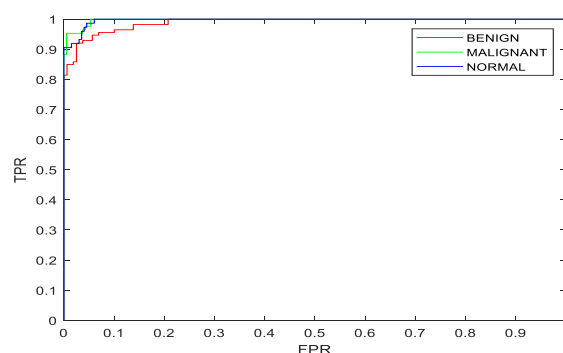


Fig 4: ROC plot for Without Hyper parameter

The ROC curve and the corresponding values for each class is as given in Figure 4. 5- Fold cross validation results have been recorded to exhibit AUC value for the each of the class. From the Table 3 it concludes that maximum AUC value 0.997 has obtained for class malignant as well as for normal.

Table 3: AUC Without Hyper-parameter Tuning

Class	Benign	Malignant	Normal
AUC	0.989	0.997	0.996

3.2 With Hyper parameter Tuning:

Hyper parameter minimizes the five-fold cross-validation loss by using automatic hyper parameter optimization. Here, max Objective Evaluations of 30 reached. The optimization minimizes the cross-validation loss (error) using by varying the parameters. A typical loss curve corresponding to hyper parameter tuning is as in figure 5.

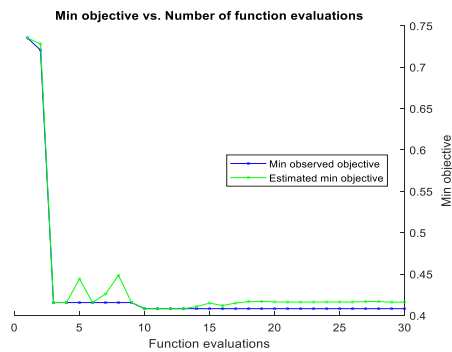


Fig 5: Loss curve for With Hyper parameter Tuning

Figure 5 indicates chart of 30 evaluation objectives for which it shown best or acceptable as evaluation result. Predictions from all learners have been computed. Loss for all observations has been computed and recorded as in figure 5. Computing posterior probabilities and predict the training-sample labels and class posterior probabilities are presented as in Table 4. Prediction of these classes for 10 cases has been tabulated and it shows that for only one of the cases of class 2 which malignant is wrongly predicted as class 1 which is Benign and class 3 which is Normal. For remaining all other cases it is predicting correctly. Using hyper parameter, overall accuracy of the proposed model has been calculated and tabulated as in table 5. Since hyper parameter auto tunes itself during prediction, k-fold cross-validation is not essential in this approach. Using this approach overall accuracy of 95.86% towards determining lung cancer and for all the 3 cases accuracy has been found to be more than 90%.

Table 4: AUC for each class Without Hyper parameter Tuning

TrueLabel	PredLabel	Posterior		
1	{'1'}	0.53565	0.11699	0.34736
1	{'1'}	0.65081	0.17032	0.17886
1	{'1'}	0.54663	0.3052	0.14816
3	{'3'}	0.25524	0.15997	0.58479
2	{'2'}	0.00080288	0.99901	0.00018583
2	{'2'}	0.46499	0.28175	0.25327
3	{'3'}	0.32718	0.11621	0.5566
2	{'1'}	0.61381	0.29768	0.088507
3	{'3'}	0.50549	0.40929	0.08522
3	{'3'}	0.48022	0.22754	0.29224

Table 5: Statistical Parameter -with Hyper-parameter Tuning

	FP R	TP R	Precisi on	Rec all	F1- Sco re	AU C	Accura cy
1	0.0	0.9	0.912	0.95	0.93	0.9	0.945
	61	39		4	2	93	
2	0.0	0.9	0.975	0.88	0.93	0.9	0.956
	11	89		9		96	
3	0.0	0.9	0.935	0.97	0.9	0.9	0.975
	25	75		3	93	97	

Using these statistical parameters from table 5 confusion matrix and ROC curve has been plotted as in figure 6 and figure 7 respectively. The results out of the plot indicates, the performance of the model is moderately good

True Class	1	106	8	2	91.4%	8.6%	
	2	4	72		94.7%	5.3%	
	3			80	100.0%		
					96.4%	90.0%	97.6%
					3.6%	10.0%	2.4%
					1	2	3
					Predicted Class		

Fig 6: Confusion Matrix for With Hyper parameter

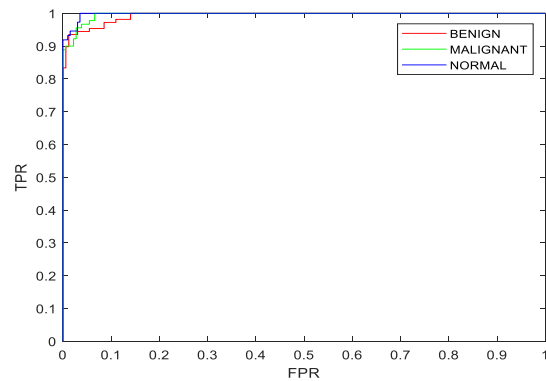


Fig 7: ROC curve for With Hyper parameter Tuning

3.2 Classification of types of Lung cancer

In the context of lung cancer classification, employing GLCM and SVM allows for the extraction of statistical parameters capturing texture information and subsequent classification based on learned patterns as tabulated in Table 6. Here, the outputs are categorised as Adenocarcinoma (1), Large cell carcinoma (2), Squamous cell carcinoma (3), and Normal (4). This approach enhances classification accuracy by incorporating local texture patterns into the model, facilitating precise differentiation between various cancer types.

Table 6: Statistical Parameter With Hyper-parameter Tuning

	FP R	TP R	Precisi on	Rec all	F1- Sco re	AU C	Accura cy
1	0.0	0.9	0.833	0.87	0.85	0.9	0.911
	73	27		5	3	85	

2	0.0 41	0.9 58	0.82	0.74 5	0.78	0.9 79	0.915
3	0.0 33	0.9 66	0.888	0.94 3	0.91	0.9 23	0.929
4	0	1	1	0.97 4	0.98 6	0.9 97	0.992

The integration of GLCM (Gray-Level Co-occurrence Matrix) analysis with SVM (Support Vector Machine) for lung cancer classification resulted in impressive accuracy rates for different types of lung cancer and normal cases. Specifically, the classification accuracies for each class were as follows: 91.1% for Adenocarcinoma, 91.5% for Large cell carcinoma, 92.9% for Squamous cell carcinoma, and 99.2% for Normal lung tissue. Additionally, the AUC (Area Under the Curve) values for all classes exceeded 90%, indicating excellent discriminative power and robustness of the model across different types of lung cancer and normal cases. This performance underscores the effectiveness of utilizing texture analysis techniques like GLCM in conjunction with machine learning algorithms like SVM for accurate and reliable lung cancer diagnosis and classification. The GLCM and SVM combination achieved a notable overall accuracy of 93.6%, demonstrating strong performance in accurately classifying lung cancer and normal cases. Similarly during the classification of types of lung cancer the proposed method achieved more than 90% for all the classes.

4. Conclusions

With advancements in technology and computer-aided diagnosis, many automated techniques have been developed to reduce false positives when evaluating cancer in CT images. To predict lung cancer, this study proposes an efficient CT classification system utilizing GLCM and SVM. Experiments were conducted to classify lung cancer into benign, malignant, and normal categories using a standard dataset. Further research extended to classification of types of lung cancer. The proposed system demonstrates that benign, malignant, and normal cases can be determined using hyper parameter tuning, achieving accuracies of 94.5%, 95.6%, and 97.5%, respectively. The overall accuracy of the system is 95.86% toward the detection of lung cancer. Similarly for the detection of types of lung cancer accuracy of 93.6% has been observed. For the same dataset, our model with hyper parameter tuning performs moderately better than the approach without hyper parameter tuning.

References

- [1] He, Jie, Ni Li, W. Q. Chen, Ning Wu, H. B. Shen, Yu Jiang, Jiang Li, Fei Wang, and J. H. Tian. "China guideline for the screening and early detection of lung cancer (2021, Beijing)." *Zhonghua Zhong Liu Za Zhi [chinese Journal of Oncology]* 43, no. 3 (2021): 243-268.
- [2] Jacobs, Colin, Arnaud AA Setio, Ernst T. Scholten, Paul K. Gerke, Haimasree Bhattacharya, Firdaus A. M. Hoesein, Monique Brink et al. "Deep learning for lung cancer detection on screening CT scans: results of a large-scale public competition and an observer study with 11 radiologists." *Radiology: Artificial Intelligence* 3, no. 6 (2021): e210027.
- [3] Lee, Jong Hyuk, Dongheon Lee, Michael T. Lu, Vineet K. Raghun, Chang Min Park, Jin Mo Goo, Seung Ho Choi, and Hyungjin Kim. "Deep learning to optimize candidate selection for lung cancer CT screening: advancing the 2021 USPSTF recommendations." *Radiology* 305, no. 1 (2022): 209-218.
- [4] Chen, Ke, Lei Liu, Bo Nie, Binchun Lu, Lidan Fu, Zichun He, Wang Li, Xitian Pi, and Hongying Liu. "Recognizing lung cancer and stages using a self-developed electronic nose system." *Computers in Biology and Medicine* 131 (2021): 104294.
- [5] Chabon, J.J., Hamilton, E.G., Kurtz, D.M., Esfahani, M.S., Moding, E.J., Stehr, H., Schroers-Martin, J., Nabat, B.Y., Chen, B., Chaudhuri, A.A. and Liu, C.L., 2020. Integrating genomic features for non-invasive early lung cancer detection. *Nature*, 580(7802), pp.245-251.
- [6] Asuntha, A., and Andy Srinivasan. "Deep learning for lung Cancer detection and classification." *Multimedia Tools and Applications* 79 (2020): 7731-7762.
- [7] Elnakib, Ahmed, Hanan M. Amer, and Fatma EZ Abou-Chadi. "Early lung cancer detection using deep learning optimization." (2020): 82-94.
- [8] Dr. Nagaraju C, Sangana Manasa Durga, Rachana N, "Ascertainment of Lung Cancer at an Early Stage", *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, ISSN: 2456-3307, Volume 2, Issue 4, 2017.
- [9] Ignatious S, Joseph R, "Computer Aided Lung Cancer Detection System." *IEEE, Proceedings of 2015 Global Conference on Communication Technologies (GCCT 2015)*, pp. 555–558, 2015.
- [10] Ueda, Daiju, Akira Yamamoto, Akitoshi Shimazaki, Shannon Leigh Walston, Toshimasa Matsumoto,

Nobuhiro Izumi, Takuma Tsukioka et al. "Artificial intelligence-supported lung cancer detection by multi-institutional readers with multi-vendor chest radiographs: a retrospective clinical validation study." *BMC cancer* 21 (2021): 1-8.

lung cancer detection on screening CT scans: results of a large-scale public competition and an observer study with 11 radiologists." *Radiology: Artificial Intelligence* 3, no. 6 (2021): e210027.

- [11] Acheampong, Felix, Trevor Ostlund, Mater Mahnashi, and Fathi Halaweish. "Estrone Analogs as Potential Inhibitors Targeting EGFR-MAPK Pathway in Non-Small Cell Lung Cancer." *Chemical Biology & Drug Design* (2023).
- [12] Bray, F., Ferlay, J., Soerjomataram, I., Rebecca, L.S., Torre L.A., Jemal, A. (2018). Global cancer statistics 2018: "GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for The Clinician*," 68(6): 394-424.
- [13] Hussain, M., Bird, J.J., Faria, D.R. (2019). "A study on CNN transfer learning for image classification." In *UK Workshop on Computational Intelligence*, Springer, Cham, pp. 191-202.
- [14] Thakur, Shailesh Kumar, Dharendra Pratap Singh, and Jaytrilok Choudhary. "Lung cancer identification: a review on detection and classification." *Cancer and Metastasis Reviews* 39 (2020): 989-998.
- [15] Schenk, Erin L., Tejas Patil, Jose Pacheco, and Paul A. Bunn Jr. "2020 Innovation-Based Optimism for Lung Cancer Outcomes." *The Oncologist* 26, no. 3 (2021): e454-e472.
- [16] He, Jie, Ni Li, W. Q. Chen, Ning Wu, H. B. Shen, Yu Jiang, Jiang Li, Fei Wang, and J. H. Tian. "China guideline for the screening and early detection of lung cancer (2021, Beijing)." *Zhonghua Zhong Liu Za Zhi [chinese Journal of Oncology]* 43, no. 3 (2021): 243-268.
- [17] Abdullah, Dakhaz Mustafa, and Nawzat Sadiq Ahmed. "A review of most recent lung cancer detection techniques using machine learning." *International Journal of Science and Business* 5, no. 3 (2021): 159-173.
- [18] Agarwal, Aman, Kritik Patni, and D. Rajeswari. "Lung cancer detection and classification based on alexnet CNN." In *2021 6th international conference on communication and electronics systems (ICCES)*, pp. 1390-1397. IEEE, 2021.
- [19] Ramaswamy, Anuradha. "Lung cancer screening: Review and 2021 update." *Current Pulmonology Reports* 11, no. 1 (2022): 15-28.
- [20] Jacobs, Colin, Arnaud AA Setio, Ernst T. Scholten, Paul K. Gerke, Haimasree Bhattacharya, Firdaus A. M. Hoessein, Monique Brink et al. "Deep learning for