Proceedings of the 6th International Conference on Data Intelligence and Cognitive Informatics (ICDICI-2025)
DVD Part Number: CFP25VL6-DVD; ISBN: 979-8-3315-0312-3

A COMPREHENSIVE REVIEW ON LUNG CANCER PREDICTION USING ML AND DL TECHNIQUES

Leema Raina F
Research scholar
Department of Computer Science and Information Technology
VISTAS, Pallavaram, Chennai
rainaleema@gmail.com

Dr.C.Anbarasi
Assistant professor
Department of Computer Science and Information Technology
VISTAS, Pallavaram, Chennai
canbarasi.scs@vistas.ac.in

***Abstract* - Lung Cancer is the leading cause of cancer-related death worldwide. Because it speeds up and improves the process that follows clinical board, early detection, prediction, and diagnosis of lung cancer has therefore become crucial. Lung cancer is a dangerous type of cancer that is hard to find. Because it typically results in demise for both women and men, it is more important for caretakers to promptly and accurately check nodules. As a result, a number of methods have been used to discoverthe lung cancer earlier. This paper presents a review analysis of various Deep Learning (DL) and Machine Learning (ML)based models for the detection of lung cancer. It can now be diagnosed using many different techniques, most of which use CT scan images and some that use x-ray images. In addition, different segmentation algorithms are combined with multiple classifier techniques for the detection of lung cancer nodules through image data. In general the CT scan images are used majorly for the identification of cancerous cells. Additionally, results from DL based strategies were more accurate than those from strategies that were applied using typical ML techniques.**

***Keywords:* Lung cancer detection, Machine learning, Deep learning, Classification.**

I. INTRODUCTION

Lung cancer causes a deadly illness that kills a great number of people worldwide. Reducing the patient mortality rate requires a primal encounter with lung cancer. Therefore, identifying and diagnosing lung cancer is a significant challenge for medical professionals and researchers. Medical images like computed tomography, chest X-rays, MRI scans, etc., can be used to detect lung cancer. Machine learning techniques identify the key features of intricate lung cancer datasets. Early in the 1980s, a Computer-Aided Diagnosis (CAD) was created to increase the effectiveness and survival rate of medical image interpretation. ML models such as Random Forest (RF), Decision trees, Support Vector Machine (SVM), Naïve Bayes (NB), Linear Regression (LR) significantly influence the healthcare industry. In addition, this study explores DL techniques and algorithms which is useful in diagnosing, detecting, and predicting different types of cancer. Using DL model it can provide a clear and concise review of recent research efforts across various types of cancer, with a particular emphasis on lung cancer prediction. Nowadays lung cancer becomes more commonin people with emphysema or chest conditions [1]. The lung cancer can be categorized into two primary types. The Small Cell Lung Cancer (SCLC) is mostly caused due to smoking and increases more quickly. The Non Small Cell Lung Cancer (NSCLC) is more prevalent and spreads more slowly, and this process is known as variedsmall/Large cell cancer [2]. It was predicted to see about 234,030 new cases of lung cancer in 2018, with about 85% of those cases being NSCLC, according to recent statistics [3].

The development of lung cancer without symptoms is the main cause of this disease's high death rate. Almost 25% of participants had no symptoms of cancer. Since lung cancer can spread rapidly, early diagnosis is crucial. According to advancements in imaging technology, lung cancer can now be identified early [4]. The location and size of the tumor determine the classification of symptoms. Since it may not cause any pain or symptoms in the early stages, it is challenging to analyze and detect.

Machine learning is the ability of computers to learn and make decisions based on algorithms, allowing them to generate results without explicit programming. It is widely recognized as a subfield of Artificial Intelligence (AI) [5]. ML is widely used for generating decisions and classifying difficult data. It is closely related to mathematical optimization, theories, and implementation frameworks for computational tasks [6].

Figure 1 shows the system model for lung cancer detection and classification phases. Normal (Benign) nodules illustrate that the patient is not cancerous, while malignant nodules signify that the patient has cancer.

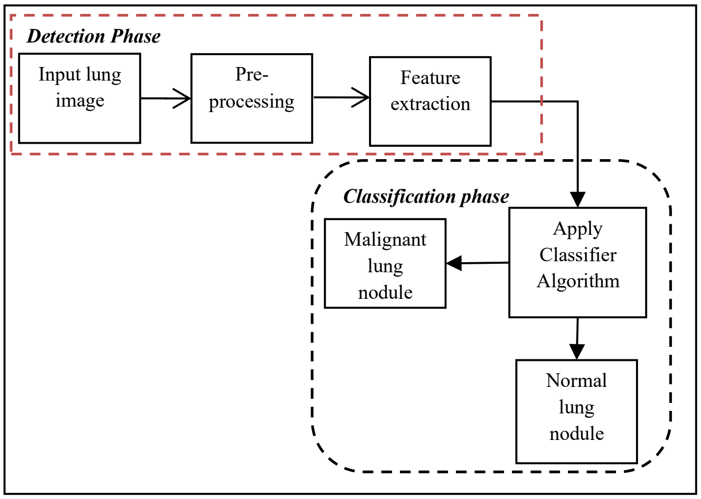

Fig 1: Basic System Model for Lung cancer detection and Classification

II. RELATED WORKS

Over the past decade, extensive research has been analyzed using machine learning [7] and deep learning models to enhance the early detection of lung cancer. These advanced techniques have been widely explored due to their ability to analyze medical imaging and clinical data with high accuracy. The ML and DL models predict the disease by undergoing certain phases such as preprocessing, segmentation, feature extraction, and classification.

A. Pre-Processing Stage

Pre-processing processes are the first step in almost any computer-aided diagnostics. Data are typically taken in the form of CT scans, MRIs, and X-rays. Therefore, pre-processing these inputs is necessary to achieve better outcomes that are desired. Every additional feature was primarily based on this method. Pre-processing is essentially the catalyst for initiating the execution since the collected inputs should not be directly applicable to the system. The original image needs be modified up in order to produce more exact and accurate data. There are various pre-processing methods. When it comes to medical imaging, image quality is crucial for identifying illnesses. To improve inputs, various quality enhancement approaches are available. Resolution, signal-to-noise ratio, and variance all affect image quality. Noises and other distortions must be eliminated as the initial step in lung CT images. CT and MRI images cannot be seen even though they have been pre-processed.

By taking neighbor pixel characteristics in consideration, contrast can be adjusted using Adaptive Neighborhood Contrast Enhancement (ANCE) [8]. Histogram equalization must be used as a pre-processing approach if the image's foreground and background are identical. The erosion technique, the second step of pre-processing after noise reduction, is typically utilized to remove unwanted areas from lung CT scans, including the sternum portion. Every object's structure is a crucial factor in the identification of cancer.

B. Segmentation Phase

To give the most accurate information on damaged lung cells, lung nodule segmentation must be done accurately. To divide all required regions before our target, object detection must be completed once the image quality has been improved. Specific characteristics must be extracted in order to identify the gland for segmentation. Texture-based features are often derived from SURF or Local Binary Patterns (LBP) descriptors and use LBP as ROI. Segmentation can be carried out using threshold values and global, local, or dynamic morphological systems, where each pixel determines the value. It may also be based on prior knowledge, a location, or a contour. Morphological operations, three-stage processes, k-means clustering with the use of different color spaces, region-based approaches, adaptive thresholding, mean shift segmentation are some of the segmentation techniques utilized in cytology images. Allotting the membership levels and then utilizing them to allocate data to one or more clusters is known as fuzzy clustering [15].

A fully automated pipeline was presented for detecting and segmenting NSCLC in 1328 thoracic CT scans [16]. The method demonstrates faster and more reproducible performance with detailed quantitative analysis based on image thickness, size of the tumor, interpretation complexity and location of the tumor.

Lung segmentation can be achieved by combining morphological procedures with an algorithm based on a stochastic technique called EM (expectation-maximization) [17]. In addition to differentiating objects, segmentation separates the area of emphasize from the surrounding environment and identifies anomalies. Adaptive template matching is the best method for identifying oddities, whereas genetic algorithms combined with thinning algorithms are ideal for detecting lung edges.

Based on volume, there is another kind of lung nodule segmentation, which integrates volumetric nodule segmentation such as region growth, thresholding, morphology, deformable model in an iterative process which is proposed based on the morphology model. Nodule extraction can be accomplished using morphological opening operations such as directional gradient concentration, closure and opening whereas segmentation can be accomplished using Feed-Forward Neural Networks [18]. Greyscale thresholding is a segmentation model can be effectively utilized to delineate the pleural wall through a level set based propagation model. This approach enables the contour to evolve adaptively, guided by shape priors and localized image gradients, ensuring accurate boundary detection even in regions with low contrast. The contrast-based differentiation enhances the model's ability to isolate nodules from adjacent tissues, improving the reliability of detection in medical imaging applications.

C. Feature Extraction Phase

The reduction of dimension is the objective of feature extraction. Nodule size, form, energy, entropy, and contrast are among the characteristics that were discovered. A feature extraction method called Local Energy-based Shape Histogram (LESH) was recently developed with the goal of diagnosing lung cancer [9]. The sequential forward method is used to choose features by considering statistical features. The two main categories of shape extraction strategies are region-based and contour-based approaches. However semantic or high order spatial relationships in lung nodules are not captured deeply.

Generally, there are three primary classifications of texture extraction algorithms: harmonic, numerical, and morphological. Textures are viewed by structural strategies as repetitions of fundamental primitive patterns with specific placement rules. Image characteristics show that intensity is a first-order statistic that is dependent on the vasslues of individual pixels. For categorization, neural networks, fuzzy c-means clustering, and the fuzzy min-max algorithm [10] are employed. A supervised neural network classifier that creates hyper box sets of fuzzy information for both training and classification is called fuzzy min-max classification with neural networks. The procedure of learning involves the hyper regions expanding and contracting. The accuracy value for detecting lung cancer using feature extraction techniques is shown in Table 1.

TABLE I. METHODS FOR FEATURE EXTRACTION

S.NO	Methods for Feature Extraction	Accuracy (%)
1.	Lung Image Database Consortium (LIDC)	95.06
2.	Local energy based shape histogram[12]	99
3.	GLCM with MLP	85
4.	LIDS –Entropy of features	91.66
5.	Correlation based Feature selection (CFS) [13]	92.46
6.	Texture and Morphology [14]	96.3
7.	Compactness, Circularity, Solidity, Convexity	85

The watershed segmentation technique was employed to identify potential nodules by effectively separating overlapping or adjacent structures based on intensity gradients and topological features. This method allowed for the detection of candidate regions that might correspond to pulmonary nodules. To characterize these candidate regions, the Histogram of Oriented Gradients (HOG) technique was utilized for feature extraction, capturing the edge orientation and local shape information critical for distinguishing nodules from other anatomical structures. To further minimize False Positives (FP), a two-stage classification process was implemented, combining a rule-based classifier with SVM. The performance of this approach was validated using 10-fold cross-validation, yielding a high sensitivity of 93.9%, indicating the method's effectiveness in correctly identifying true nodules [11].

D. Classification

In medical imaging, classification methods are crucial, particularly for tumor identification and categorization [19]. Since classification is the last stage of the CAD system, it is essential to achieving the best results possible given the features at present.

Nodule detection is accomplished by classification and clustering. Fuzzy rule-based categorization systems utilizing knowledge-based interpretations can be utilized to accomplish it. A rule-based method is a common technique used for classification, where a system uses "if-then" rules to make decisions based on the specific problem. A set of rules was created to keep as many cancerous cell areas as possible. These rules help reduce the chances of mistakenly identifying healthy cells as cancerous. During the feature extraction step, radiomic features are taken from each lung image separately. Only the significant features are selected using an improved version of the whale optimization algorithm based on graph clustering. After selecting the key features, different classification methods are used to identify cancerous cells. These include ensemble classifiers such as SVM, KNN and RF.

An Artificial Neural Network (ANN) was created for classification through the integration of geometric information with multi-dimensional intensity scales. Using pre-computed sections with the same texture, an interactive method for lung segmentation is provided, which trains the haziness nodules by improving the current KNN classifier. For classification the KNN classifier is applied using the fisher score ranking algorithm among 22 different features. When compared to other ML models, ANN generates the best results, with an 82% success rate. For candidate detection, a multi-scale 2D filtering technique is employed, and FP are decreased through logical processing.

For the accurate identification of the severe lung cancer condition, an optimal classifier based on computational intelligence approaches had been developed.

2D & 3D CNN Classifier:

CNN is a feed-forward, multi-layered technique that works particularly well for detection. There are fewer training parameters and a straightforward network layout. CNN's weight-sharing network topology makes it more similar to biological brain networks. The network's layers include convolution, activation, pooling, and a fully linked layer. Deep CNN provides considerably more processing power but more explicit findings when compared to traditional approaches.

The 2D CNN algorithm was proposed to categorize the development of malignant and non-malignant (benign) cells in the lungs. This automated technique, which uses CT images to identify lung cancer nodules, saves a great deal of efforts, time, and error [20]. Using the LIDC dataset, the Particle Swarm Optimization (PSO) algorithm, which is based on 2D CNN, was presented to eliminate the need for physical network hyper-parameter searching. It reduces FP by combining an evolutionary method with a deep learning methodology. For nodule feature classification, a combination of XGBoost and RF is used in conjunction with 2D U-Net. The convolution technique is the primary distinction between 3D CNN and 2D CNN [21]. The input layer in this system is a 4D tensor, which includes four parts: depth, height, width, and channels. The added "depth" dimension helps in processing 3D medical images. When a filter scans the input, it moves across three directions such as height, width, and depth.

3D CNN can be used for detecting lung nodules, and reinforcement learning can be applied to improve this detection. To enhance accuracy, several deep learning models have been used such as 3D U-Net, DenseNet, and Region Proposal Network (RPN). In this setup, U-Net combined with RPN helps to find possible nodule candidates, and DenseNet reduces incorrect classifications. A 3D version of the Faster R-CNN model [22], which works similarly to U-Net, has also been used to locate nodules. A deep CNN is then used to confirm the presence of these nodules.

Two models such as a Faster R-CNN and a regular CNN were trained to detect lung nodules from the features. The improved version of the Faster R-CNN helped boost the recall rate (how many real nodules were correctly found) from 40.1% to 56.8%, and it also increased detection precision (how many detected nodules were actually correct) from 76.4% to 90.7%.To estimate the chance of cancer, the XGBoost model was used to train a logistic regression based on feature values. Important features like nodule diameter, speculation (sharp edges), and lobulation (bumpy shapes) were used in the logistic regression model to help reduce false positives. Separating the left and right lungs in the images helped simplify the detection process. Higher accuracy for identifying lesions of any size is also offered by four different sized 3D CNNs. The primary goal of 3D CNN is for identifying lung cancer lesions with high sensitivity while lowering the false positive rate.

Therefore for structured or small datasets machine learning algorithms such as SVM, RF, KNN, etc are used, for image based tasks like segmentation and classification CNNs can be used due to their strong spatial feature learning, in order to handle sequential data RNNs or hybrid RNN-CNN models can be used.

Table 2 provides a comparison summary table for lung cancer prediction and detection techniques.

TABLE II. COMPARISON OF EXISTING MODELS

Stages	**Models**		
	Technique	***Advantages***	***Limitations***
Preprocessing	ANCE	Improves local contrast using neighboring pixel information	Noise amplification in low SNR regions
	Histogram Equalization	Foreground and background intensities are similar	It has uneven illuminations
Segmentation	Morphological Operations	Useful for feature extraction	Parameter dependent and less robust
	Fuzzy Clustering	It allows overlapping between clusters	Requires fine tuning and high computational cost
Feature Extraction	Entropy Contrast	Interpretable	Global spatial patterns are not extracted efficiently
	Radiomic features	Focuses on relevant features using search-based selection	Non-linear or abstract patterns are not captured efficiently
Classification	ANN	Able to learn complex nonlinear relationships	Requires large labeled dataset and fine tuning
	SVM and RF	Performs well in small to mid-sized datasets	SVM is less interpretable while RF prone to overfitting with noisy features
	CNN	Performs both feature extraction and classification with high accuracy	Requires GPU hardware

From the analysis of this review of lung cancer there are some limitations found like pre-processing methods ANCE, histogram equalization, etc. are more data specific and the generalization of CAD system capability is reduced. The fully automated process for lung cancer detection includes data pre-processing, automated segmentation, automated radiomic and deep feature extractors for selecting most relevant features, classifying the benign and malignant nodules by employing trained ML or DL classifiers.

To improve the performance of the lung cancer detection using medical image analysis the high-quality and annotated datasets are used. Also advanced deep learning models like CNN, RNN with attention or ensemble techniques are used. Enhancing pre-processing, effective handling of class imbalance, selecting relevant features and applying high level classifier models can significantly improves the accuracy which also improves the detection performance.

By using multimodal fusion, ensemble models (CNN+SVM) and advanced deep learning architectures such as deeper CNNs helps in learning more discriminative features, 3D-CNN, etc. are the techniques used for improving the precision for lung cancer detection and classification. This process significantly reduces False Positives (FP) and provides more reliable outcomes.

IV CONCLUSION

Machine learning and deep learning algorithms are very useful because they can work with different types of data and can be used in many programming languages. These methods are now being used in more and more fields. In recent years, deep learning has become especially popular in medicine, particularly because of its success in working with image and video data. In this area, CNN's are often used, especially for the classification of image tasks. Studies show that CNN models perform very well in identifying and classifying lung nodules as either benign or malignant in CT scan images. In this review, a detailed overview of the latest medical image processing methods used for detecting and classifying lung cancer with ML and DL is provided. However the medical imaging datasets are limited and suffers from class imbalance. In addition deep models require high performance hardware (GPU) for both inference and training. The future work can be carried out by integrating the optimization models with deep learning algorithms so that the accuracy can be improved which increases the performance.

REFERENCE

1. Radhika, P. R., Nair, R. A., &Veena, G. (2019, February). A comparative study of lung cancer detection using machine learning

algorithms.In 2019 IEEE international conference on electrical, computer and communication technologies (ICECCT) (pp. 1-4).IEEE.

2. Hussain, L., Rathore, S., Abbasi, A. A., &Saeed, S. (2019, March). Automated lung cancer detection based on multimodal features extracting strategy using machine learning techniques. In Medical imaging 2019: physics of medical imaging (Vol. 10948, pp. 919-925).SPIE.

3. Siegel, R. L., Miller, K. D., &Jemal, A. (2018). Cancer statistics, 2018. CA: a cancer journal for clinicians, 68(1), 7-30.

4. Günaydin, Ö.,Günay, M., &Şengel, Ö. (2019, April). Comparison of lung cancer detection algorithms.In *2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT)* (pp. 1-4).IEEE.

5. Chiu, H. Y., Chao, H. S., & Chen, Y. M. (2022). Application of artificial intelligence in lung cancer.Cancers, 14(6), 1370.

6. Mahesh, B. (2020). Machine learning algorithms -a review. International Journal of Science and Research (IJSR) [Internet], 9(1), 381-386.

7. Radzi, S. F. M., Karim, M. K. A., Saripan, M. I., AbdRahman, M. A., Osman, N. H., Dalah, E. Z., & Noor, N. M. (2020). Impact of image contrast enhancement on stability of radiomics feature quantification on a 2D mammogram radiograph. IEEE Access, 8, 127720-127731.

8. Chaki, J. (2023). Cancer Data Pre-Processing Techniques.In Current Applications of Deep Learning in Cancer Diagnostics (pp. 19-33).CRC Press.

9. Wajid, S. K., Hussain, A., Huang, K., &Boulila, W. (2016, August). Lung cancer detection using Local Energy-based Shape Histogram (LESH) feature extraction and cognitive machine learning techniques. In 2016 IEEE 15th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC) (pp. 359-366). IEEE.

10. Deshmukh, S., &Shinde, S. (2016, September). Diagnosis of lung cancer using pruned fuzzy min-max neural network. In 2016 International Conference on automatic control and dynamic optimization techniques (ICACDOT) (pp. 398-402).IEEE.

11. Firmino, M., Angelo, G., Morais, H., Dantas, M. R., &Valentim, R. (2016). Computer-aided detection (CADe) and diagnosis (CADx) system for lung cancer with likelihood of malignancy.Biomedical engineering online, 15, 1-17.

12. Wajid, S. K., & Hussain, A. (2015). Local energy-based shape histogram feature extraction technique for breast cancer diagnosis. Expert Systems with Applications, 42(20), 6990-6999.

13. Abdullah, D. M., Abdulazeez, A. M., & Sallow, A. B. (2021). Lung cancer prediction and classification based on correlation selection method using machine learning techniques. Qubahan Academic Journal, 1(2), 141-149.

14. Jena, S. R., George, T., &Ponraj, N. (2019, February). Texture analysis based feature extraction and classification of lung cancer. In 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT) (pp. 1-5).IEEE.

15. Ajai, A. K., &Anitha, A. (2022). Clustering based lung lobe segmentation and optimization based lung cancer classification using CT images. Biomedical Signal Processing and Control, 78, 103986.

16. Primakov, S. P., Ibrahim, A., van Timmeren, J. E., Wu, G., Keek, S. A., Beuque, M., &Lambin, P. (2022). Automated detection and segmentation of non-small cell lung cancer computed tomography images. Nature communications, 13(1), 3423.

17. Sharafeldeen, A., Elsharkawy, M., Alghamdi, N. S., Soliman, A., & El-Baz, A. (2021). Precise segmentation of COVID-19 infected lung from CT images based on adaptive first-order appearance model with morphological/anatomical constraints. Sensors, 21(16), 5482.

18. Nanglia, P., Mahajan, A. N., Rathee, D. S., & Kumar, S. (2020). Lung cancer classification using feed forward back propagation neural network for CT images. International Journal of Medical Engineering and Informatics, 12(5), 447-456.

19. Hussain, S., Mubeen, I., Ullah, N., Shah, S. S. U. D., Khan, B. A., Zahoor, M., ... & Sultan, M. A. (2022). Modern diagnostic imaging technique applications and risk factors in the medical field: a review. BioMed research international, 2022(1), 5164970.

20. Biradar, V. G., &Pareek, P. K. (2022, October). Lung cancer detection and classification using 2D convolutional neural network. In 2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon) (pp. 1-5). IEEE.

21. Yu, J., Yang, B., Wang, J., Leader, J., Wilson, D., & Pu, J. (2020). 2D CNN versus 3D CNN for false-positive reduction in lung cancer screening. Journal of Medical Imaging, 7(5), 051202-051202.

22. Xu, J., Ren, H., Cai, S., & Zhang, X. (2023). An improved faster R-CNN algorithm for assisted detection of lung nodules. Computers In Biology And Medicine, 153, 106470.