International Journal of Advances in Engineering and Management (IJAEM)
Volume 7, Issue 05 May 2025, pp: 18-23 www.ijaem.net ISSN: 2395-5252

# A Review on Detection of Lung Cancer Using Machine Learning Techniques

Hetal Solanki[1]

[1] *Assistant Professor, Vidhyadeep Institute of Computer & Information Technology, Vidhyadeep University, Anita, Kim Gujrat - India*

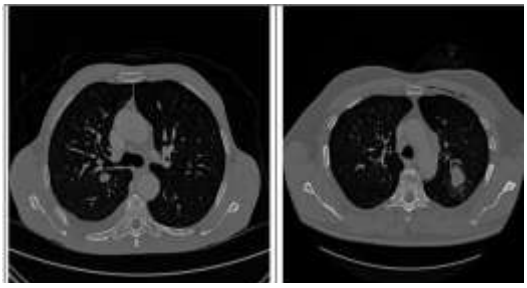
Date of Submission: 01-05-2025 Date of Acceptance: 10-05-2025

**ABSTRACT:**
One of the prime reasons for the cancer deaths worldwide is lung cancer and which is largely due to late diagnosis and limited access to firstly screening facilities. This research investigates that how machine learning techniques are applied to classify or diagnose lung cancer in its early stages using imaging and clinical data. Malignancy of lung nodule is forecasted using supervised learning methods like Convolutional Neural Networks (CNN), Random Forests, and Support Vector Machines (SVM). Quality, precision, recall, and F1-score are metrics using to train and evaluate the models based on the datasets such as, LIDC-IDRI. To assist healthcare doctor in early detection or the treatment planning also the research indicates that machine learning has significantly enhance the accuracy of diagnosis.

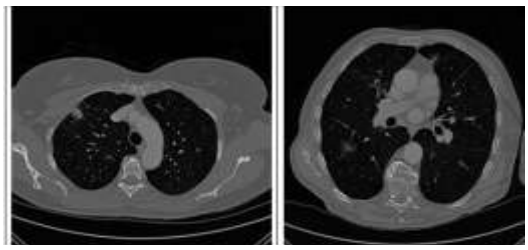
**Keyword:**Cancer Classification, Medical Imaging, Convolutional Neural Networks (CNN), Random Forests, Support Vector Machines (SVM), LIDC-IDRI Dataset, Supervised Learning, Computer-Aided Diagnosis (CAD.

## I. INTRODUCTION

Lung cancer is still one of the most dangerous cancers in the world, causing the largest number of cancer-related deaths each year. Although improvement in treatment approaches, the five-year survival rate for individuals with lung cancer remains terribly poor, primarily because of late-stage identification. Since lung cancer often shows no symptoms or hazy ones, it is particularly hard to detect in its early stages. Despite their efficiency, common examinations such as CT scans, chest X-rays, and biopsies are sometimes limited by their cost, accessibility, and need for specialized knowledge for accurate interpretation. These difficulties underline how urgently creative methods are needed to aid in the timely and precise detection of lung cancer, which could significantly enhance patient outcomes.



**(a)**



**(b)**

**Figure.1:**(a) Lung cancer and (b)non-cancer samples [1]

International Journal of Advances in Engineering and Management (IJAEM)
Volume 7, Issue 05 May 2025, pp: 18-23 www.ijaem.net ISSN: 2395-5252

The application of machine learning (ML) techniques in medical fields has accelerated dramatically in recent years. Computers may learn from data and make predictions or judgments without explicit task programming thanks to machine learning, a subset of artificial intelligence (AI). Machine learning algorithms have shown promise in detecting patterns and anomalies that may not be immediately noticeable to human observers, especially in medical imaging and clinical diagnostics. One intriguing way to get around the drawbacks of conventional diagnostic techniques is to incorporate machine learning (ML) into workflows for lung cancer detection. Machine learning models can interpret clinical data and medical imaging quickly, accurately, and consistently by utilizing vast datasets and complex algorithms.

The use of computer-aided diagnosis (CAD) systems has greatly advanced the use of medical imaging in the diagnosis of lung cancer. These techniques help radiologists detect cancerous tumours early on enhance diagnostic accuracy, and lower inter-observer variability. Starting with image capture and ending with lung nodule classification, the above graphic shows a systematic process for lung cancer detection and classification.

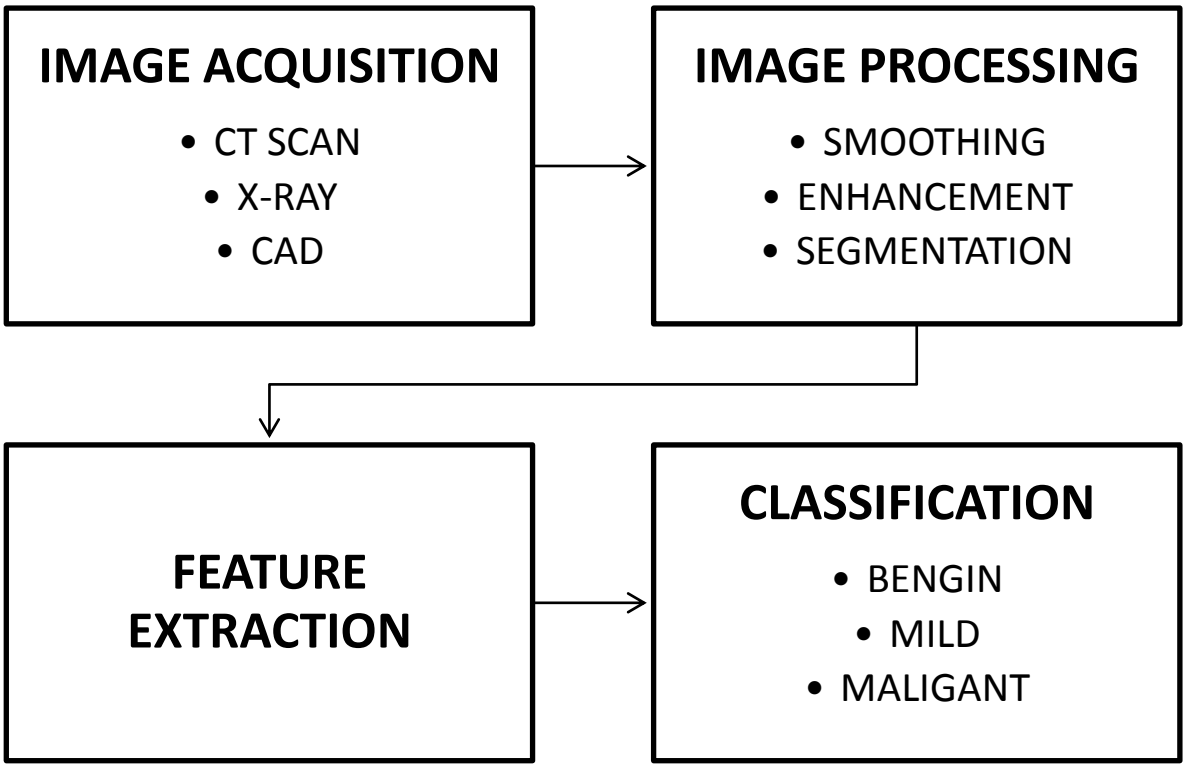



**Figure 2.** The steps of Image Processing Technique for Lung Disease Diagnosis (LDD)using Computed-Tomography (CT) and X-ray images.

**I) CT/X-Ray (Image Acquisition)**

The initial phase in this process is to obtain the lung images through chest X-rays or computed tomography (CT), which are frequently used in imaging techniques for identifying abnormalities in the lungs. When it comes to detecting small nodules, CT scans are more accurate than X-rays because they provide high-resolution cross-sectional pictures.

**II) Image Acquisition**

In this process, digital medical images are captured. Getting standardized inputs that can be processed and analysed further is important. The accuracy of downstream analysis is greatly affected by the quality and resolution of these images.

**III) Pre-processing (Smoothing, Enhancement, and Segmentation)**

This block includes three interconnected steps aimed at improving image quality and isolating relevant features:

- **Smoothing:** Ensures that the crucial information while reducing noise in the image. Methods such as, Gaussian filters are frequently employed.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 7, Issue 05 May 2025, pp: 18-23 www.ijaem.net ISSN: 2395-5252

- **Enhancement:** Making nodule or tumors easier to recognize by enhancing contrast and focusing important characteristics.
- **Segmentation:** Using this technique also thresholding, edge detection and deep learning segmentation networks, this picture are divided into regions such as interest (like suspicious nodules and lung areas).

**IV) Feature Extraction**

At this stage, meaningful features are extracted from the segmented images. These features may include:

- **Shape descriptors** (e.g., roundness, margin sharpness),
- **Texture features** (e.g., contrast, entropy), and
- **Intensity-based features**. This transforms raw image data **into structured data suitable for classification.**

**V) Classification**

The extracted features are then fed into machine learning classifiers, which categorize the lung nodules based on their malignancy risk. The classification is typically done using algorithms such as:

- **Support Vector Machines (SVM)**
- **Random Forests (RF)**
- **Convolutional Neural Networks (CNNs)**

The classifier compares the new input with the **database** of pre-labelled data and assigns it to one of the categories:

- **Benign** (non-cancerous)
- **Mild** (low risk or early-stage)
- **Malignant** (cancerous)

This study investigates the use of different machine learning methods for early lung cancer diagnosis and classification. Specifically, the efficacy of supervised learning techniques like Support Vector Machines (SVM), Random Forests (RF), and Convolutional Neural Networks (CNNs) in predicting the malignancy of lung nodules is examined. Because CNNs, a class of deep learning models, can automatically extract hierarchical features from raw images, they are particularly effective at processing and evaluating imaging data.Based on ensemble learning, Random Forests are renowned for their capacity to handle high-dimensional clinical datasets and their resilience to over fitting. In contrast, Support Vector Machines are very good at determining the best decision restrictions for classification tasks, especially when working with complicated, non-linear data distributions.

For the reason of training and proving these machine learning models, datasets like the LIDC-IDRI (Lung Image Database Consortium and Image Database Resource Initiative) offer a wealth of annotated medical images. The LIDC-IDRI dataset is a great resource for supervised learning techniques since it contains thoracic CT scans with marked-up annotated lesions for lung cancer screening and diagnosis purposes. Accuracy, precision, recall, and F1-score are the primary classification measures used in this study to assess model performance. These measures balance the trade-offs between false positives and false negatives, which are especially important in medical diagnosis, and offer a thorough grasp of the models' diagnostic capabilities.

The primary goal of using machine learning techniques in this situation is to support clinical decision-making procedures rather than to replace human expertise. By serving as a second set of eyes, machine learning models can help radiologists and oncologists spot worrisome nodules more quickly and precisely. Additionally, machine learning-assisted early detection can result in better patient survival rates, tailored treatment plans, and earlier intervention.

## II. LITERATURE REVIEW

S. S. M. Noor et al. [1]provide Traditional methods using classifiers like Support Vector Machines (SVM) and Random Forests (RF) were initially explored for lung cancer diagnosis, showing promising results in clinical settings.

Mohamed, Osman, and El-Nasr [2] explored various machine learning approaches for the early diagnosis of lung cancer, focusing on improving prediction accuracy using clinical and imaging data. Their study highlights the potential of advanced algorithms like SVM and ensemble methods in enhancing early detection rates.

Ahmed et al. [3] proposed an ensemble learning approach to classify lung cancer using CT imaging data, aiming to boost diagnostic performance. Their method combined multiple classifiers to reduce error rates and improve overall prediction accuracy in lung cancer detection. Researchers also applied ensemble methods to improve classification robustness [4].

Setio [5] developed a multi-view convolutional neural network (CNN) approach to detect pulmonary nodules in CT images, with a focus on reducing false positives. Their method significantly enhanced detection accuracy by analysing nodule candidates from multiple perspectives. The advent of deep learning, especially Convolutional Neural Networks (CNNs),

significantly improved performance on imaging datasets such as LIDC-IDRI. CNNs enabled automatic feature extraction, reducing the dependency on manual intervention [5], [6].

For medical image processing tasks, Tajbakhsh et al. [7]Looked at whether fully training or fine-tuning pre-trained convolutional neural networks (CNNs) produced better results. Their research offered valuable information about how to best apply deep learning techniques, particularly when working with sparse medical datasets.Later studies adopted transfer learning to leverage pre-trained models and boost small medical dataset's performance [7].In order to diagnose chest diseases, bar et al. Investigated the use of deep learning models that had previously been trained on non-medical datasets. Their research showed that even in the absence of substantial, specialized medical datasets, transfer learning could classify medical images with high accuracy [8].

Hybrid approaches combining imaging and clinical data for improved diagnosis gained attention. Fusion models integrating CT scans with patient metadata showed superior performance compared to models using imaging alone [9], [10]. A thorough analysis of computer-aided diagnosis (CADx) systems for examining mammogram masses and clustered micro calcifications was provided by Elter and Horsch [11]. Their study provided an overview of several machine learning methods designed to enhance mammography-based breast cancer diagnosis and classification. Techniques such as feature selection [11], [12] and dimensionality reduction [13] were employed to optimize model input space.

A thorough review of deep learning applications in medical imaging, with an emphasis on MRI technology, was given by Lundervold [14]. In addition to highlighting difficulties and potential paths forward in medical image processing, their work covered important deep learning architectures. Recent studies have focused on explain ability and interpretability of ML models in healthcare, ensuring that predictions are understandable to clinicians [15]. Capsule networks were presented by LaLonde and Bagci [16] as a unique method for medical imaging object segmentation. In comparison to conventional CNNs, their study showed that capsules could better capture spatial hierarchies in image data, increasing segmentation accuracy. Capsule Networks and attention mechanisms were also introduced to better capture spatial relationships in CT images.

In order to improve CNN performance in liver lesion classification, Frid-Adar et al. investigated the use of Generative Adversarial Networks (GANs) for synthetic medical picture augmentation. Their method showed that CNN accuracy could be greatly increased by using GAN-generated images, particularly in datasets with small sample sizes [19]. Emerging methods explored Generative Adversarial Networks (GANs) for augmenting medical datasets, thus tackling the small sample size problem [18], [19]. Researchers also investigated ensemble deep learning models that combine multiple CNN architectures [20], [21].

Using a radio genomics method, Liu et al. examined the connection between lung cancer gene alterations and imaging characteristics. Their research demonstrated how combining genetic and radiological data can enhance the precision of lung cancer diagnosis and tailored treatment plans. Beyond imaging, multimodal learning combining genomics, proteomics, and radiomics has also been explored for more accurate lung cancer prediction [22], [23].

In order to improve the performance of deep learning models by learning from unlabelled data, Chen et al. presented a straightforward framework for contrastive learning of visual representations. SimCLR, their approach, showed notable gains in image recognition tasks, making it an effective tool for applications involving medical imaging. Recent advancements involved self-supervised learning approaches [24], [25] that reduce the need for annotated data.

A thorough evaluation and meta-analysis of deep learning-based techniques for the automated identification of pulmonary nodules in chest CT scans was carried out by Nam et al. Their study reviewed the literature and assessed how well different deep learning models may increase the precision of lung cancer detection. Screening large populations through automated ML systems was explored, ensuring reduced radiologist workload while maintaining high sensitivity [26], [27]. Reinforcement learning approaches for sequential diagnosis decision-making are also under investigation [28].A reinforcement learning-based method for automatic lung cancer diagnosis was presented by Xu et al. [29], with an emphasis on sequential decision-making for diagnostic tasks. Their model optimized the decision-making process during cancer diagnosis, demonstrating the potential of reinforcement learning to increase diagnostic accuracy. Overall, machine learning techniques demonstrate vast potential in assisting lung cancer detection, yet challenges remain in dataset diversity, model generalization, and regulatory acceptance [30].

International Journal of Advances in Engineering and Management (IJAEM)
Volume 7, Issue 05 May 2025, pp: 18-23 www.ijaem.net ISSN: 2395-5252

## III. FUTURE SCOPE

Enhancing the generalization and scalability of models is crucial for the future of machine learning-based lung cancer diagnosis. The development of systems that can manage huge and varied datasets holds great promise for enabling these models to function well in a range of healthcare environments and demographics. To lessen bias, research might also concentrate on making algorithms more resilient so that models don't perform disproportionately better on particular data kinds or demographics.

The addition of multimodal data sources into unified diagnostic models, including imaging, clinical records, genomic data, and patient history, is another exciting field. Machine learning systems can provide more precise and individualized diagnostic insights by integrating these diverse information sources, which can result in customized treatment plans and better patient outcomes. Additionally, developments in transfer learning and self-supervised learning techniques might lessen the need for sizable labelled datasets, which would facilitate the use of AI models in contexts with a lack of annotated medical data.

Finally, a significant issue still lies in resolving these models' interpretability and explanation. It will be easier to build trust and hasten the adoption of AI-driven diagnostic tools in clinical practice if they are transparent and their judgments are intelligible to medical personnel. Future research in this field might concentrate on creating models that might assist clinicians in making decisions by offering thorough reasoning in addition to precise forecasts.

## IV. CONCLUSION

One interesting approach to overcoming the difficulties in early detection and categorization of lung cancer is the incorporation of machine learning (ML) tools into the diagnosis process. This study emphasizes how machine learning (ML) models, such as Support Vector Machines (SVM), Random Forests (RF), and Convolutional Neural Networks (CNNs), show significant promise in enhancing diagnostic precision, especially when used with complicated medical imaging data like CT scans. These models—CNNs in particular—have demonstrated remarkable capacity to automatically extract hierarchical characteristics, which lessens the need for human interpretation and improves detection efficiency.

Further enhancing model robustness and overall performance are ensemble learning techniques, such as integrating multiple classifiers or utilizing hybrid algorithms that incorporate imaging data with patient metadata. The difficulties presented by tiny and sparse medical datasets can be overcome by using transfer learning and data augmentation strategies, such as the creation of synthetic images using Generative Adversarial Networks (GANs).

In addition, multimodal techniques that integrate clinical, genetic, and radiological data have great potential to improve diagnostic accuracy and facilitate individualized treatment regimens. New avenues for minimizing reliance on labelled data and improving clinical decision-making processes are also made possible by the continuous study into self-supervised learning and reinforcement learning.

Although these developments, issues still exist, such as the requirement for a variety of datasets, the generalization of models, and regulatory approval. But with more study and advancement, machine learning methods can be a priceless addition to human knowledge, boosting early lung cancer diagnosis and, eventually, improving patient outcomes. Because of this, the future of lung cancer diagnosis depends on the cooperative efforts of intelligent machine learning models and human clinicians, which will enable earlier detection, more precise categorization, and individualized treatment plans.

## REFERENCES

[1]. S. S. M. Noor, H. M. Salleh, and A. A. Rahman, “Lung cancer classification using deep learning convolutional neural network,” Proc. IEEE Region 10 Conf., 2017, pp. 2302–2307.

[2]. A. R. Mohamed, H. S. Osman, and M. A. El-Nasr, “Machine learning approaches for early diagnosis of lung cancer,” IEEE Access, vol. 7, pp. 86964–86975, 2019.

[3]. H. G. Ahmed et al., “Ensemble learning for lung cancer classification using CT images,” IEEE Int. Conf. Imaging Syst. Tech., 2019.

[4]. S. Sharma and A. Aggarwal, “Random forests based classification model for early detection of lung cancer,” Proc. IEEE Int. Conf. Comput. Commun. Technol., 2018.

[5]. D. Setio et al., “Pulmonary nodule detection in CT images: False positive reduction using multi-view convolutional networks,” IEEE Trans. Med. Imaging, vol. 35, no. 5, pp. 1160–1169, 2016.

[6]. K. Ardila et al., “End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography,” Nat. Med., vol. 25, pp. 954–961, 2019.

[7]. H. Tajbakhsh et al., “Convolutional neural networks for medical image analysis: Full

training or fine tuning?," IEEE Trans. Med. Imaging, vol. 35, no. 5, pp. 1299–1312, 2016.

[8]. Y. Bar et al., "Deep learning with non-medical training used for chest pathology identification," Proc. SPIE Med. Imaging, 2015.

[9]. J. F. Gomes et al., "Combining image and clinical data for lung cancer diagnosis using deep learning approaches," IEEE Access, vol. 8, pp. 228883–228895, 2020.

[10]. S. M. Hosseini et al., "Early lung cancer detection using hybrid data fusion techniques," Proc. IEEE Symp. Comput. Intell. Healthc. Appl., 2018.

[11]. M. Elter and R. Horsch, "CADx of mammographic masses and clustered microcalcifications: A review," Med. Phys., vol. 36, no. 6, pp. 2052–2068, 2009.

[12]. R. Dua, M. Gaur, A. Ayyagari, and S. Vaidya, "Machine learning for early lung cancer detection: A systematic review," IEEE Rev. Biomed. Eng., vol. 16, pp. 1–15, 2022.

[13]. Z. Cheng et al., "Dimensionality reduction for lung cancer diagnosis using convolutional autoencoders," IEEE Access, vol. 7, pp. 150419–150429, 2019.

[14]. J. Lundervold and A. Lundervold, "An overview of deep learning in medical imaging focusing on MRI," Z. Med. Phys., vol. 29, no. 2, pp. 102–127, 2019.

[15]. S. Rajpurkar et al., "CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning," arXiv preprint arXiv:1711.05225, 2017.

[16]. R. LaLonde and U. Bagci, "Capsules for object segmentation," arXiv preprint arXiv:1804.04241, 2018.

[17]. A. Vaswani et al., "Attention is all you need," Adv. Neural Inf. Process. Syst., 2017.

[18]. X. Yi, E. Walia, and P. Babyn, "Generative adversarial network in medical imaging: A review," Med. Image Anal., vol. 58, 2020.

[19]. Y. Frid-Adar et al., "GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification," Neurocomputing, vol. 321, pp. 321–331, 2018.

[20]. X. Zhou et al., "Multi-level convolutional networks for lung nodule classification," IEEE Access, vol. 7, pp. 21267–21279, 2019.

[21]. L. Dou et al., "Multi-level contextual 3D CNNs for false positive reduction in pulmonary nodule detection," IEEE Trans. Biomed. Eng., vol. 66, no. 10, pp. 2853–2863, 2019.

[22]. C. Liu et al., "Radiogenomics of lung cancer: Imaging features associated with gene mutations," Eur. Radiol., vol. 29, no. 11, pp. 5452–5460, 2019.

[23]. D. Bi et al., "Multi-omics data integration for cancer outcome prediction," Brief. Bioinform., vol. 21, no. 4, pp. 1193–1205, 2020.

[24]. T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," Proc. 37th Int. Conf. Mach. Learn., 2020.

[25]. X. Zhuang et al., "Self-supervised learning for medical image analysis: A survey," arXiv preprint arXiv:2112.04333, 2021.

[26]. A. Nam et al., "Deep learning-based automatic detection of pulmonary nodules using chest CT images: A systematic review and meta-analysis," Eur. Radiol., vol. 32, pp. 1604–1616, 2022.

[27]. S. H. Kim et al., "Implementation of a CAD system in lung cancer screening: performance and impact on radiologists' detection," Radiology, vol. 283, no. 2, pp. 565–572, 2017.

[28]. H. Shao et al., "Reinforcement learning-based decision support for lung cancer diagnosis," Proc. IEEE Int. Conf. Bioinform. Biomed., 2019.

[29]. Z. Xu et al., "A reinforcement learning approach to automatic lung cancer diagnosis," Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2017.

[30]. A. Esteva et al., "A guide to deep learning in healthcare," Nat. Med., vol. 25, no. 1, pp. 24–29, 2019.

[31]. Pawar, Vikul J., Kailash D. Kharat, Suraj R. Pardeshi, and Prashant D. Pathak. "Lung cancer detection system using image processing and machine learning techniques." Cancer 3, no. 2020 (2020).