Cureus Journal of Computer Science
https://doi.org/10.7759/s44389-026-00044-8

Review Article

A Journeying on Machine Learning and Deep Learning Strategies for Lung Carcinoma Forecasting

Pragnya Das [1, 2], Satya N. Tripathy [2], Kali P. Rath [3]

1. Department of Computer Science and Engineering, Centurion University of Technology and Management, Parlakhemundi, IND
2. Department of Computer Science and Engineering, Berhampur University, Berhampur, IND
3. Department of Mathematics, NIST University, Berhampur, IND

Received: February 18, 2026 | Review began: February 24, 2026 | Review ended: March 17, 2026 | Published: March 23, 2026

Abstract

Lung cancer, which ranks second globally in terms of prevalence, is a significant factor linked to a lower likelihood of survival. In the context of cancer-related illnesses, it significantly affects survival disparities between sexes. Preventive lung cancer diagnosis is one of the most important ways to improve the prognosis for those who have the disease. A variety of machine learning and deep learning techniques have significantly improved outcomes for lung cancer. To enable early detection of lung cancer, researchers have developed precise prediction models based on machine learning and deep learning techniques. By placing greater emphasis on the dataset augmentation component, that is, expanding the sample size and incorporating ensemble approaches that could aid in defining the various stages of cancer progression, the current research endeavor can be further improved. The article provides a thorough overview of the numerous studies conducted by various researchers that use various technical techniques to explain how lung cancer develops as a predictive strategy.

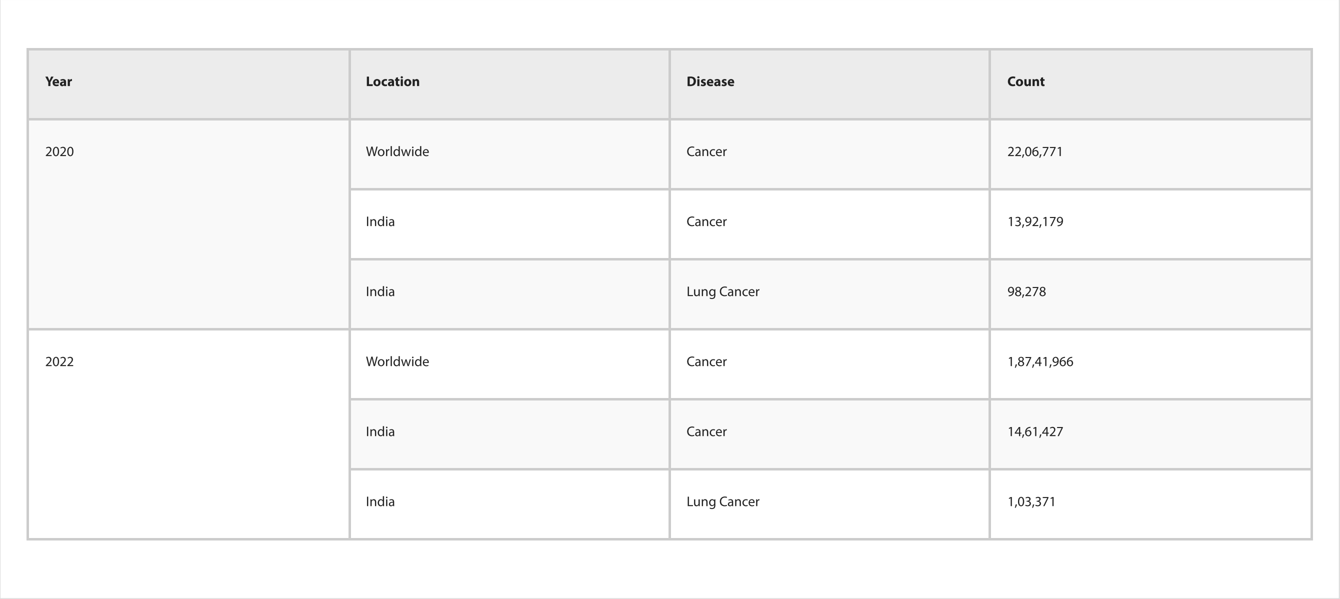
Categories: AI/ML-based decision support systems, Computer Vision, Deep Learning

Keywords: lung cancer, internet of things, machine learning, deep learning, disease forecast

Introduction And Background

Cancer is a neoplastic disease, which is the uncontrolled growth of cells and appears in different anatomical sites of the human body. One of its subtypes, which is the most common, is lung carcinoma and is the next most common malignancy globally. In the United States, the mortality rate caused by lung carcinoma in the year 2022 was 130,180 fatalities (68,820 men and 61,360 women), with 236,740 adults having a lung carcinoma diagnosis (117,910 men and 118,830 women) in the United States (Cancer.org [1]). In 2020 alone, it is estimated that 2,206,771 individuals were diagnosed with lung carcinoma and roughly 1,796,144 of them died of the condition. In 2022, the projections of cancer in India showed that the expected rate of cancer incidence was 1,461,426 with an estimated 12.8% projected increase by 2025 compared to 2020 [2]. Table *1* represents the 2020 and 2022 statistics of cancer and lung cancer.

How to cite this article:
Das P, Tripathy S N, Rath K P (March 23, 2026) A Journeying on Machine Learning and Deep Learning Strategies for Lung Carcinoma Forecasting. Cureus J Comput Sci 3 : es44389-026-00044-8. DOI https://doi.org/10.7759/s44389-026-00044-8



Year	Location	Disease	Count
2020	Worldwide	Cancer	22,06,771
	India	Cancer	13,92,179
	India	Lung Cancer	98,278
2022	Worldwide	Cancer	1,87,41,966
	India	Cancer	14,61,427
	India	Lung Cancer	1,03,371

TABLE 1: Cancer and lung cancer statistics

Early diagnosis of lung cancer can increase survival rates. Different classification methods may be used to differentiate between malignant (cancerous) and normal cells in lung tissue. Modern efforts are being made to make health surveillance systems smarter in order to improve the efficiency of healthcare delivery. Smart health monitoring systems help systematically organize and store patients' health records, making a valuable contribution to healthcare providers and patients, and enabling health examinations regardless of geographical location.

Internet of Things

According to IBM's definition, the Internet of Things (IoT) is a large approach that comprises interconnections of devices and people and enables obtaining and sharing information in a way that suits the environmental situational understanding [3]. IoT plays a central role in the healthcare ecosystem, in terms of affording the healthcare practitioners with the ability to conduct round-the-clock patient care, and at the same time, the ability to provide patients with the power to manage their own medical records. In addition, IoT architecture grants greater data security concerning the protection of patient records. Notably, IoT technology proves instrumental in the identification of malignant lung cells, wherein computed tomography (CT) images sourced from diverse healthcare infrastructures undergo comprehensive analysis to discern the malignancy status of lung cells.

Machine learning

In this new age of artificial intelligence (AI), computer systems can learn through machine learning (ML) from previous information. The three variants of ML paradigms that are used to address various domains of problems and contribute towards facilitating this knowledge elicitation are supervised learning, unsupervised learning, and reinforcement learning.

In the case of supervised ML techniques, annotated datasets act as the basis for training algorithms. In order to perform predictions or classifications, these algorithms extract feature patterns and relationships from the data. Regression and classification are two basic categories that comprise supervised learning. Both simple linear regression and multiple linear regression are two different regression techniques best suited for continuous data inputs. However, classification techniques including random forest (RF), decision tree (DT), K-nearest neighbor (KNN), naïve Bayes, support vector machine (SVM), and logistic regression (LR) have been used to handle categorical inputs and are shown in Figure *1*.

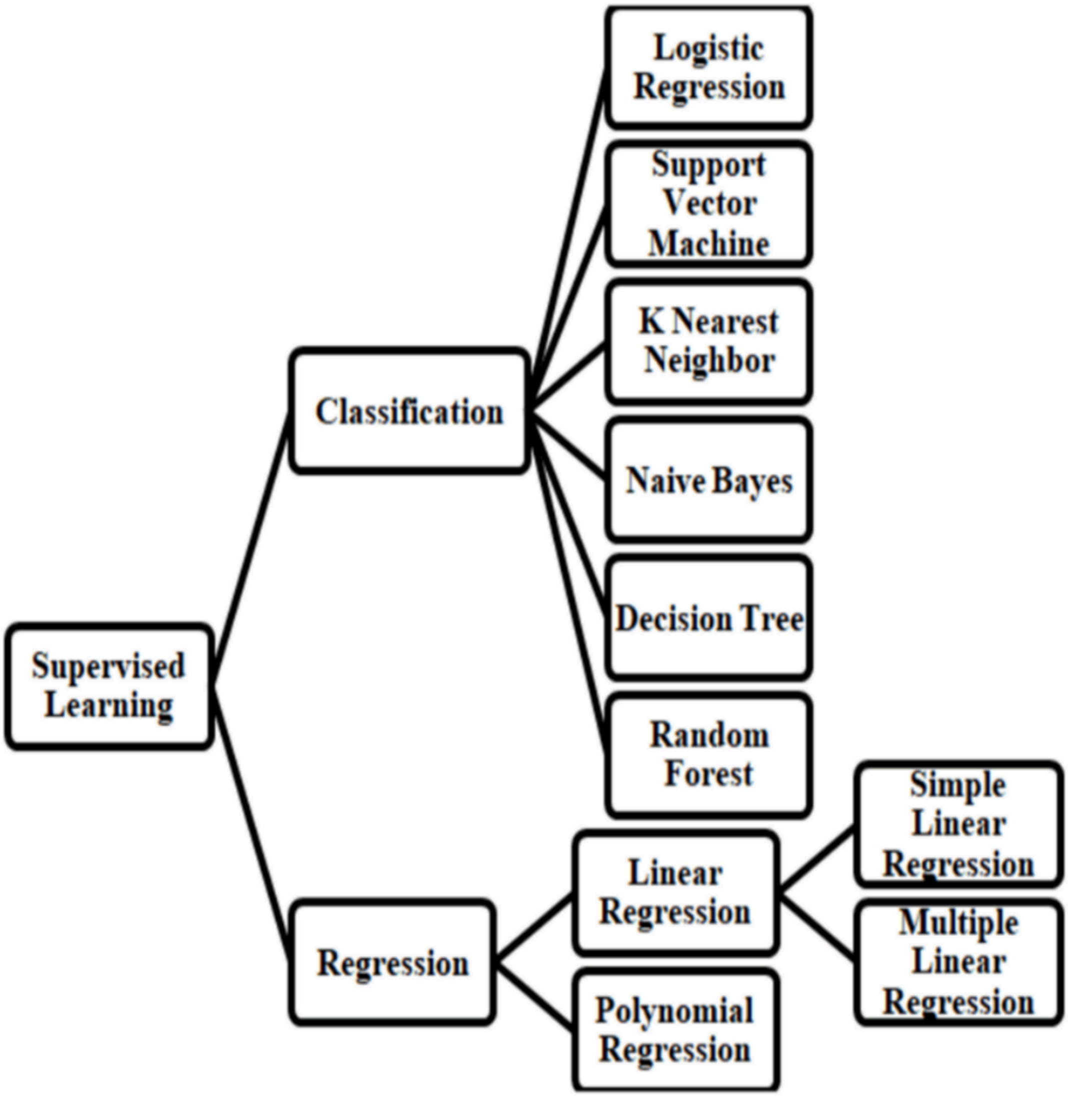


FIGURE 1: Supervised learning type

As a part of ML techniques, unsupervised ML algorithms use unannotated data for training any computational model. By utilizing feature similarity analysis, this model automatically divides the data into labeled groups. Association and clustering are two different subcategories of unsupervised ML. Again, fuzzy clustering and hierarchical clustering are two different clustering algorithms, while the fuzzy pattern growth algorithm and the a priori algorithm are two different association techniques. Unsupervised algorithms are shown in Figure *2*.

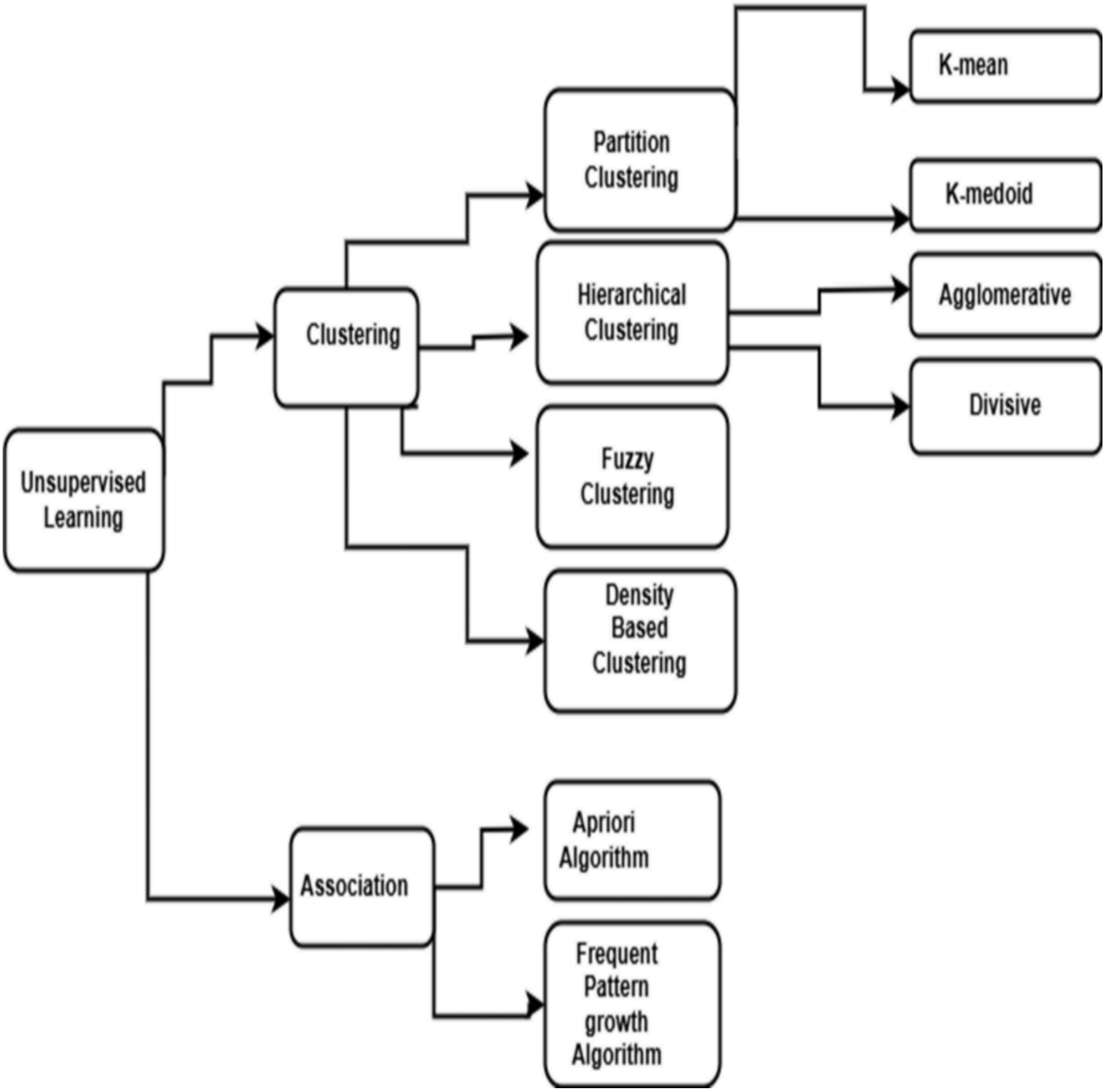


FIGURE 2: Unsupervised learning type

Reinforcement learning is a feedback-based ML method in which an agent learns by trying different actions and observing what works.

Deep learning

Deep learning (DL), as a subset of ML within the broader domain of AI, exhibits the capability to address complex problem-solving tasks. This capability arises from artificial neural networks as its foundation, enabling it to learn intricate patterns and relationships in data. Deep neural networks exist in various forms, including convolutional neural networks (CNNs), multilayer perceptrons (MLPs), and recurrent neural networks (RNNs).

As a notable deep neural network, a CNN is used to categorize images by extracting feature patterns from the image. In the CNN architecture, key characteristics include the ability to share parameters and the consideration of spatial relationships across network layers, despite the absence of recurrent connections.

Conversely, MLPs comprise a distinct type of DL methodologies mostly suited for managing tabular data. These networks do not include parameter sharing, spatial relationships, or recurrent connections, focusing solely on processing structured data.

On the other side, as part of the deep neural network family, RNNs are tailored to sequences of data such as textual information, time series, and audio signals. Recurrent connections and parameter sharing facilitate the modeling of temporal dependencies in RNNs, while spatial relationships are not a central consideration within this technique. Different algorithmic techniques of DL are shown in Figure *3*.

How to cite this article:
Das P, Tripathy S N, Rath K P (March 23, 2026) A Journeying on Machine Learning and Deep Learning Strategies for Lung Carcinoma Forecasting. Cureus J Comput Sci 3 : es44389-026-00044-8. DOI https://doi.org/10.7759/s44389-026-00044-8

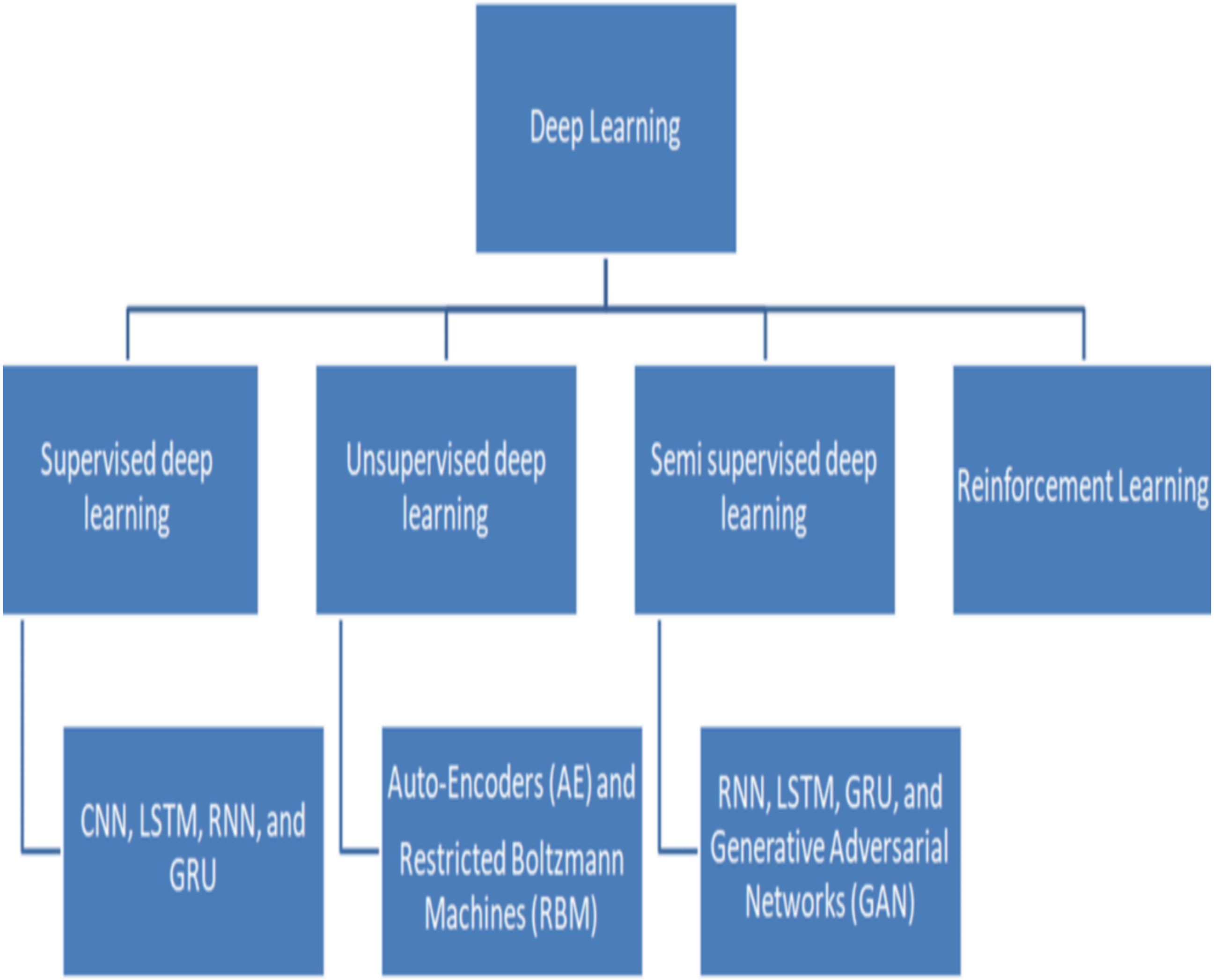
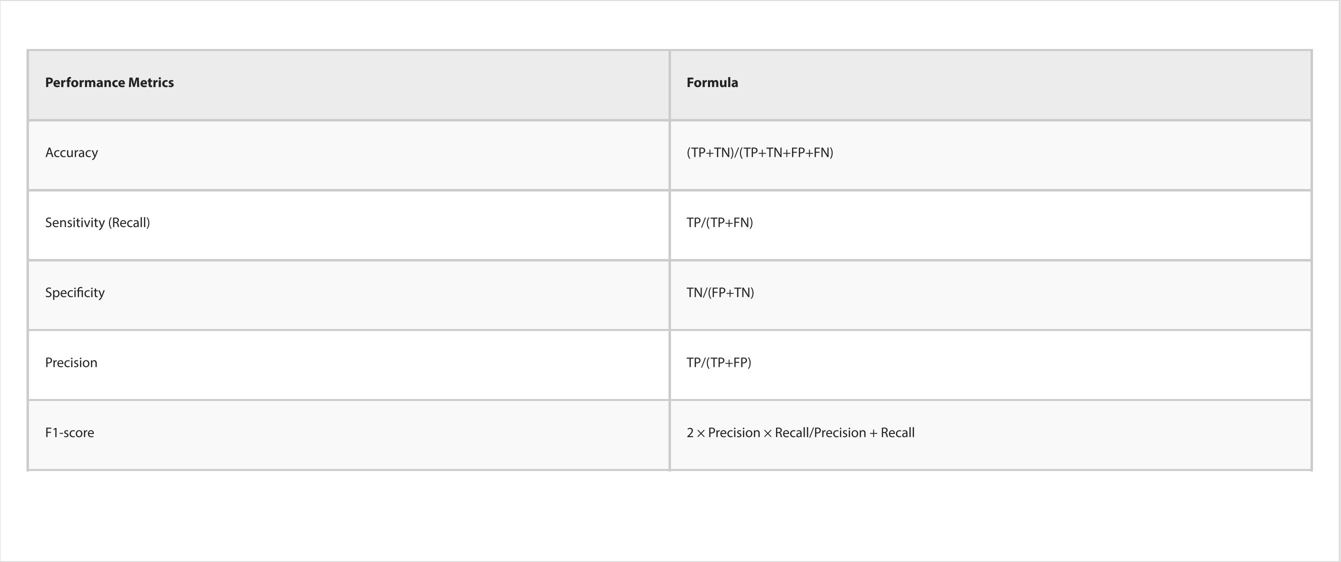


FIGURE 3: Deep learning type

CNN, Convolutional Neural Network; GRU, Gated Recurrent Unit; LSTM, Long Short-Term Memory; RNN, Recurrent Neural Network

Evaluation metrics

The most frequently reported evaluation metrics in the studies of lung cancer prediction were extracted and compared in this review. The performance metrics used to assess accuracy, sensitivity, specificity, precision, F1-score, and the area under the ROC curve (AUC) were recorded in the majority of studies. They were recorded in order to be able to compare the diagnostic accuracy in a similar way, especially in terms of evaluating the capacity of a model to detect malignant cases correctly (sensitivity) and with no false positives (specificity). The records of such standardized measurements can offer an objective and comparable analysis of the work of both ML and DL models under different datasets and methodological circumstances. Table *2* represents the performance metrics used with their formulas.



Performance Metrics	Formula
Accuracy	(TP+TN)/(TP+TN+FP+FN)
Sensitivity (Recall)	TP/(TP+FN)
Specificity	TN/(FP+TN)
Precision	TP/(TP+FP)
F1-score	2 × Precision × Recall/Precision + Recall

TABLE 2: Performance metrics with formulas

TP, True Positive; TN, True Negative; FP, False Positive; FN, False Negative

The main research question that will be discussed in this review is how ML and DL methods can enhance the accuracy, reliability, and timeliness of early diagnosis and prognosis of lung cancer. As medical imaging data is increasingly becoming available, sophisticated computational models are being actively investigated in automated detection and classification. It conducts a systematic review of publications of the last five years (2016 and 2025) and includes articles involving AI-based models used on CT or positron emission tomography (PET)/CT data. This paper will discuss the role of ML and DL solutions in detecting malignant patterns and assisting clinicians with their decisions. Special attention is paid to the contribution of data augmentation, heterogeneous datasets, and ensemble learning techniques. All these methods are meant to improve the predictive performance and to more accurately describe the intricate dynamics of lung cancer.

Review

This review adopted a systematic methodology to identify and evaluate research on lung cancer classification using ML and DL approaches. Using this approach, the analysis and usefulness of both ML and DL algorithms in lung cancer prediction are examined. The review was performed using diverse sources such as conference papers, journal articles, book chapters, and scholarly articles. The flowchart of the proposed review methodology is shown in Figure *4*.

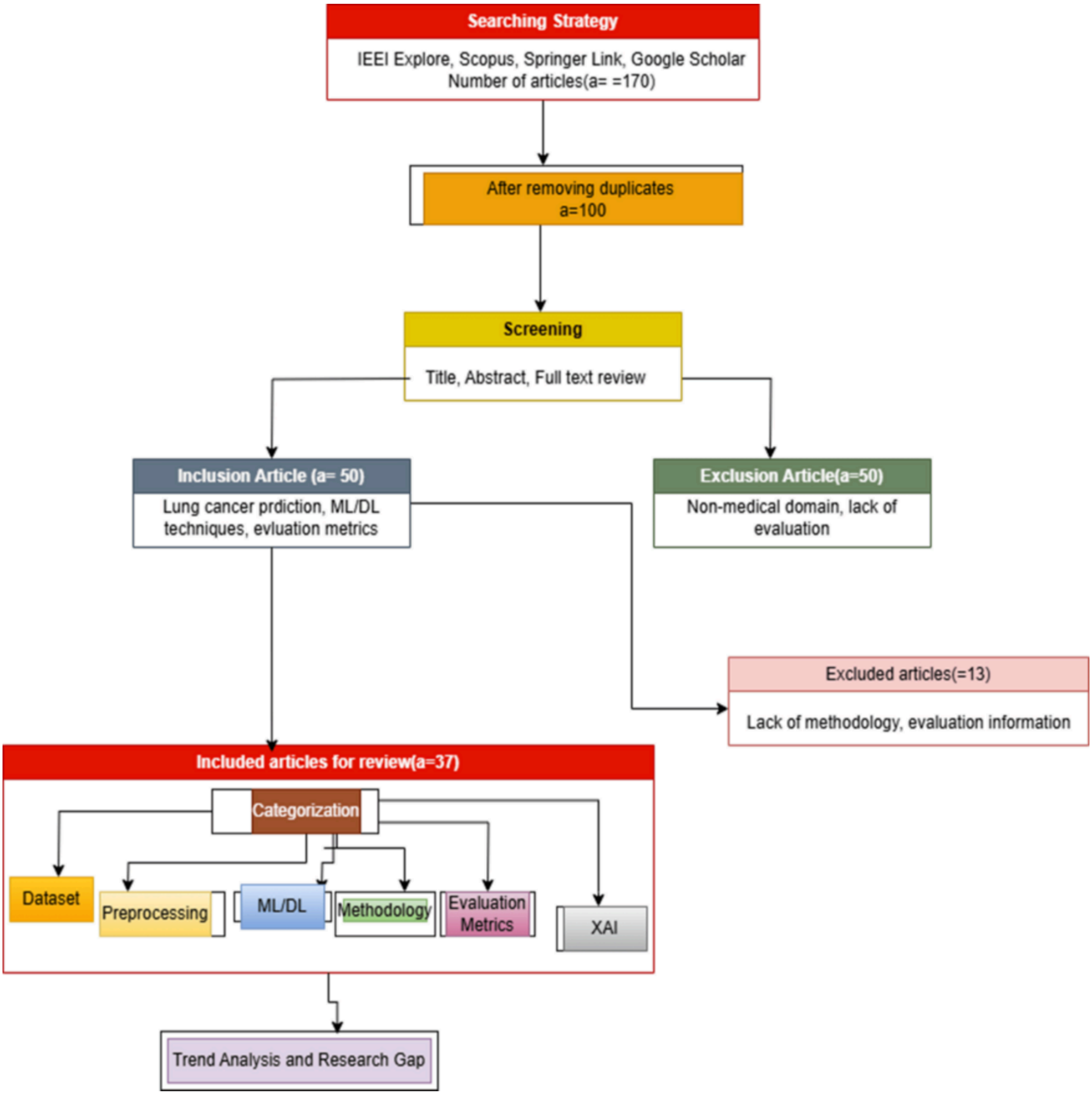


FIGURE 4: Review methodology flow chart

DL, Deep Learning; ML, Machine Learning; XAI, Explainable Artificial Intelligence

Searching strategy

This review was conducted according to the PRISMA guidelines to ensure methodological transparency and reproducibility. A comprehensive literature search was performed in PubMed, Scopus, IEEE Xplore, Web of Science, and Google Scholar to identify studies published between 2015 and 2025 on ML- and DL-based lung cancer detection, classification, or prognosis using medical images as input. The primary keywords used in various combinations included lung cancer, ML, DL, CT imaging, medical image classification, nodule detection, predictive modeling, and AI-based diagnosis. After removing duplicate records, studies were screened based on titles and abstracts, followed by full-text screening.

Dataset characteristics

In this review, we assessed research based on the nature and quality of datasets used for lung cancer prediction, following PRISMA guidelines. Only articles involving medical imaging datasets (CT, PET/CT, or chest X-ray) or studies reporting the use of clinical or biometric datasets were considered. Important features, such as source repository, sample size, imaging modality, and annotation method, were recorded to evaluate the comparability of studies. This systematic review ensures that the selected studies represent a variety of datasets that are prolifically applied in the study of diagnosing lung cancer by using ML/DL.

Screening

The screening of all relevant articles occurs based on the title, abstract, and full-text review. Title screening is done to remove irrelevant works, and abstract screening is used to determine whether the research framework applied ML/DL algorithms on lung cancer data. Finally, full-text review is conducted to provide complete methodological reporting as well as to derive performance measures.

Inclusion and exclusion policy

Inclusion criteria included (i) the studies that involved the use of ML or DL models being applied to lung cancer image datasets (CT, PET, X-ray), (ii) articles that reported quantitative measures of performance as either accuracy, sensitivity, specificity, F1-score, or AUC, and (iii) articles that were published in English.

The exclusion criteria were non-medical imaging studies, non-ML/DL methods, reviews, editorials, incomplete datasets, or studies that were not adequately described. Information was extracted from each study, including dataset features, dataset preprocessing methods, model design, validation strategy, and major performance metrics. The final group of studies included in the review is considered the most relevant and methodologically appropriate work on AI-based lung cancer prediction.

Watching and tabulation of data

In a systematic review of inclusion criteria, articles with ambiguity in their methodology and evaluation metric values were removed. The remaining included articles were reviewed carefully to derive important information and classify them in a tabular format, as represented in Table *3*.

In the research work, Qadir and HamaKarim [4] used two different classification algorithms, namely RF and neural network, for classifying lung carcinoma data. They used a lung cancer dataset of 310 records in which 15 features were dependent variables, while one feature was an independent type. By considering the kappa statistic and accuracy as important factors, the accuracy of both models was calculated. The neural network algorithm achieved 0.75 accuracy with 0.25 mean squared error (MSE), whereas the RF classification achieved 0.89 accuracy with 0.21 MSE. Through this approach, they aimed to assess a supervised predictive ML algorithm in terms of precision, accuracy, and minimum error.

The research by Ahmad and Mayya [5] was dedicated to predicting the risk of lung cancer based on a structured dataset of demographic, behavioral, and clinical risk factors instead of imaging information. Age, smoking history, genetic predisposition, level of exposure, and comorbidities are the variables contained in the dataset. Missing value handling, label encoding, and normalization procedures were used in preprocessing. The methodology made use of various ML algorithms such as DTs and RF. Accuracy, sensitivity, and specificity were used as evaluation metrics, and the RF classifier had the highest predictive performance (approximately 92%). No explainable AI (XAI) methods were used, so the analysis was limited in terms of interpretability [5].

Yang et al. [6] employed a clinical follow-up dataset with treatment background, recurrence data, biomarkers retrieved through imaging, and survival data to construct customized prognostic models of lung cancer. Removal of irregular clinical records as well as standardization of continuous parameters was done in preprocessing. The ML methods used by the authors include DT, neural network, SVM, Cox proportional hazards-based survival models, etc. The approach aims to incorporate multi-modal clinical variables and train models to predict recurrence risk and survivability interval. The concordance index (C-index) is used to report performance, along with accuracy and AUC, and the best models are those with high discrimination for recurrence prediction (C-index > 0.75). No XAI methodology is used, which limits its application in clinical practice. Limitations include uneven data distribution, possible biases in follow-up duration, and non-integration of imaging-based DL [6].

In the study by Mishra et al. [7], a predictive model for lung cancer was developed by integrating the Internet of Health Things (IoHT) and computational intelligence techniques. Preprocessing involved noise reduction of sensor signals, continuous feature normalization, and timestamp synchronization of multi-sensor estimates. The experiment used ML algorithms such as Heuristic Greedy Best First Search, RF, SVM, and ANN in order to identify a patient's risk condition in real time. The approach incorporates cloud-based processing, a smart decision-making model, and nonstop physiological surveillance. The proposed model demonstrated optimal accuracy, specificity, and sensitivity of 98.8%, 97.5%, and 97.8%, respectively. The system lacks XAI, which limits clinical transparency in automatic alerts. Weaknesses include the lack of medical imaging information, use of non-clinical sensor features, and testing in actual patient populations [7].

The CT imaging dataset used by Venkatesh and Bojja [8] contains lung nodules and non-nodules for cancer detection. Noise reduction using Gaussian filtering, segmentation using adaptive segmentation, and enhancement using morphological operators were the steps in the preprocessing pipeline. An optimization algorithm (e.g., Particle Swarm Optimization or a Genetic Algorithm variation) based on bio-inspired techniques was applied to improve the quality of feature selection and the efficiency of classification. The methodology used SVM classification to enhance detection accuracy. The performance measures used to evaluate the effectiveness of the swarm intelligence framework compared to conventional algorithms included accuracy, specificity, sensitivity, and training time. The model achieved an outstanding accuracy, sensitivity, and specificity of 93.53, 92.96, and 98.52, respectively. It does not use any XAI technique, which limits clinical interpretation [8].

The structured clinical data used by Abdullah et al. [9] include demographic, lifestyle, and symptom-based characteristics that are used to classify the lung cancer status. Preprocessing is the step of correlation-based feature selection. Some classification algorithms considered are SVM, CNN, and KNN. It focuses on the best features selection by applying correlation-based feature selection to improve the performance of classifiers and minimize processing time. According to performance evaluation, the accuracy of KNN, SVM, and CNN is 88.40%, 95.56%, and 92.11%, respectively. There is no XAI framework, which restricts the interpretability of the selected features. The limitations are a small sample size, absence of imaging data, and no advanced ensemble or DL comparisons [9].

Hussain et al. [10] employed CT imaging data, with textures extracted based on the Gray Level Co-occurrence Matrix, to predict lung cancer by enhancing the robustness of classification. Histogram equalization, image enhancement, and lung region of interest isolation using segmentation were applied in the preprocessing step. On the extracted Gray Level Co-occurrence Matrix textures, ML classifiers such as SVM, RF, and Gradient Boosting were used. Accuracy, recall, precision, specificity, and F1-score performance metrics are reported, and the highest accuracy (96-100%) is obtained with SVM. It does not apply any XAI method because the method is based on handcrafted features rather than saliency-map-based DL techniques. Weaknesses include reliance on handcrafted features, lack of comparison with DL methods, and limited scalability to heterogeneous CT collections [10].

Nanglia et al. [11] came up with a hybrid ML classifier (SVM and neural network) to classify lung cancer using clinical and image-based features. The preprocessing involved feature scaling and feature reduction on the basis of correlation. They used neural networks to first generate discriminative features followed by SVM to carry out final classification on the extracted features, performing the experiment with 98.08% accuracy. No XAI technique was used in the study, and the findings were based only on traditional ML performance measures [11].

Gupta et al. [12] introduced an IoT-based Smart Healthcare Kit that allows real-time remote control of essential physiological parameters such as temperature, pulse rate, and respiratory indicators using built-in low-cost sensors. The system has a cloud-based architecture for active data delivery and remote accessibility, providing a viable early warning system for respiratory abnormalities. In contrast to state-of-the-art ML-based diagnostic models, the proposed kit is targeted at lightweight embedded processing to allow the model to operate quickly and efficiently using limited resources in home or under-infrastructure environment. It contributes by offering a low-cost, compact, and scalable health-monitoring system that builds the foundation for future AI-based lung disease prediction systems [12].

According to Kalaivani et al. [13], the DL-based lung cancer detection model was trained using CT scan datasets, with preprocessing involving image resizing, lung region extraction, and contrast enhancement using the histogram equalization method. In the study, the authors applied a CNN architecture to automatically extract and classify features and reported an accuracy of 90.85% in tumor classification. The strategy consisted of a series of steps including segmentation, feature learning, and classification using softmax. Although the model showed good predictive performance, no XAI methods were used, which restricts interpretability [13].

Pawar et al. [14] propose a new extension of CNN that also uses blockchain technology to achieve secure transfer of lung cancer diagnostic results. Preprocessing of CT images through normalization and noise filtering was performed, and the images were then fed into the extended CNN for classification. The hybrid model achieved an accuracy of more than 96%, with blockchain contributing traceability and secure model deployment. XAI techniques were not used because the emphasis was on model security and performance rather than interpretability [14].

Makaju et al. [15] employed CT scan images from publicly available datasets, and steps such as filtering (median filter and Gaussian filter) and lung segmentation based on watershed segmentation were applied to the preprocessing. A CNN-based classifier was developed that is capable of detecting malignant nodules with a promising accuracy (92%) and sensitivity (100%), even using a rather limited dataset. The approach was based on automated deep feature extraction. There was no inclusion of XAI components, which restricted clinical explainability [15].

Palani and Venkatalakshmi [16] propose an IoT-based predictive modeling system to detect lung cancer that involves the use of fuzzy clustering to augment prediction and a hybrid classification system. The technique is based on Otsu thresholding and Fuzzy C-means clustering to isolate transitional areas and appropriately segment lung images, assisted by edge detection, morphological thinning, and cleaning, as well as refining tumor edges using region-filling operations. To improve predictive accuracy, the authors present an incremental classification algorithm that integrates Association Rule Mining, DT with time-dependent features, and CNN. The classification employed fuzzy models based on ML methods, providing moderate accuracy for use in remote monitoring. The lack of XAI and diverse datasets were the major shortcomings [16].

Saleh et al. [17] suggested a hybrid CNN-SVM architecture to classify lung cancer based on a CT-scan dataset from publicly available repositories, which mainly included benign and malignant nodule images. Noise reduction using a Gaussian filter, lung region extraction, and enhancement were the preprocessing steps. They employed DCNN in their approach to automatically extract hierarchical image features, which were then passed to an SVM classifier to obtain final malignancy predictions, providing better decision boundaries than those obtained with softmax-based classification. The hybrid model showed good performance, with accuracy values above 97%, along with high sensitivity and specificity, and showed some distinct benefits compared to standalone CNN models and classical ML models. Nevertheless, no XAI methods were used in the research [17].

Shanbhag et al. [18] present a multi-stage carcinoma identification process aimed at determining benign and malignant CT lung images. To enhance predictive accuracy, the authors developed an ensemble classifier of five ML models: SVM, LR, MLP, DT, and KNN, thus using the complementary advantages of different algorithms. The ensemble method achieved an accuracy of 85%, as well as precision and recall values that indicate moderate reliability in the detection of malignant nodules. Although the methodology demonstrates the usefulness of having a combination of various classifiers, the lack of XAI schemes and the dependence on restricted representations of features limit clinical interpretability, and this can prevent generalization across different conditions of CT imaging [18].

The research suggests that a combination of two cascaded modules to complement one another can be utilized to introduce a comprehensive system of lung cancer prognosis. The first module is concerned with nodule detection, which uses a combination of heterogeneous DL models, i.e., a 3D-CNN-type detector and a recently trained RNN-type detector, the complementary feature extraction processes of which significantly enhance the recall by reducing missed nodules.

The second module is cancer risk assessment, which is based on the use of 3D-CNN models to assess morphological and structural features of identified nodules, formulating medically interpretable intermediate features. A total of 13 high-risk features of nodules are obtained and inputted into a regression model to predict the morbidity grade consistent with oncologist-labeled features. The system has good prognostic results with the log-loss of 0.408 and revealed the important morphological characteristics that are most likely to correlate with the cancer severity [19].

Shakeel et al. [20] present a developed automated lung cancer detection system called RDCNN-TriHorn-Net-WHOA-ALCD, which is aimed at reducing the difficulties in CT lung cancer diagnosis. Images of Chest CT-Scan Images Dataset and a Formatted and Augmented variation are processed via Sub-Aperture Keystone Transform Matched Filtering (SAKTMF) to remove noise before being segmented by Automatically Weighted Binary Multi-View Clustering (AW-BMVC). The Second-Order Synchroextracting Wavelet Transform (SOSWT) is used to extract features, and Robust Deformed Convolutional Neural Network (RDCNN) models are used to classify the features under three different strategies. The RDCNN-TriHorn-Net with the Wader Hunt Optimization Algorithm (WHOA) proved to be the most powerful of them, achieving better results than other models, including RDCNN-ResNeXt-50 (AEALOA) and RDCNN-CoAtNet (DTOA). The proposed system was able to achieve successful multi-class classification of large cell lung Carcinoma, adenocarcinoma, squamous cell carcinoma, and normal cases with significant improvements of 13.67, 27.55, and 14.67 in Dice similarity, as well as 22.23, 24.11, and 25.56 in log-loss. This makes RDCNN-TriHorn-Net-WHOA-ALCD a robust pipeline for improving the lung cancer detection in CT imaging [20].

Ali et al. [21] suggest an end-to-end CNN that would be used to identify dysfunctional nodules in the lung and non-dysfunctional nodules with minimal false-positive rates. In order to deal with massive class imbalance, the authors use oversampling, whereby images of minority classes are modified so that they train the CNN much stronger. The model has a considerably low false positive of 1.5, which implies a high degree of reliability in the prevention of over-detection, but its false negative of 31.76 is relatively high, which indicates the difficulty in detecting the small nodules. Altogether, the system has a sensitivity of 68.66 and a specificity of 98.42, which is good, and there is room to enhance the results on the accurate identification of non-nodules, but not of the actual cancerous nodules. Nevertheless, in spite of such limitation, the article draws attention to the efficiency of specific CNN frameworks along with data augmentation techniques to classify nodules [21].

Lanjewar et al. [22] introduce a variant of a DL framework, DenseNet201, with more layers to detect lung cancer by increasing features in the original framework and improving features through the introduction of new layers. Attributes generated out of the improved DenseNet201 are again narrowed down through two feature-selection techniques and handed to various ML classifiers to achieve final forecasts. The performance of the system is strictly measured using confusion matrices, ROC analysis, Matthews correlation coefficient, kappa score, p-value calculation, and 5-fold cross-validation. The proposed approach gives superior outcomes with a perfect 100% accuracy in some of the settings, average accuracy of 95%, and a p-value of less than 0.001 having a statistical significance [22].

To make a prediction regarding the disease of lung cancer, Al-Tawalbeh et al. [23] conducted the usage of two different ML algorithms: SVM, KNN, naive Bayes, and narrow neural network. They have analyzed the symptoms of the patient. As per their results, KNN is least accurate (85.87), and SVM is most accurate (92.6%).

Doppalapudi et al. [24] utilized DL models in their research methodology and discovered that they are more effective compared to the traditional ML models that use categorization and regression. As compared to the traditional ML, where the classification strategy and RMSE strategy had 61.12% and 14.87% accuracy, respectively, DL models had 71.18% accuracy in the classification strategy, 13.5% RMSE, and 0.5 R^2, respectively. Table *3* demonstrates the literature search of lung cancer based on DL methods.

Nassif et al. [25] used CNN with F-test feature selection algorithm in their study design to distinguish between the severity of lung cancer using the data on gene expression. They have used this model on two datasets, which are coded LUAD and LUSC, and have achieved an almost 93.94% and an 88.42% accuracy, respectively. The possibility of overfitting existed due to the high levels of features of high-dimensional genes. In future studies, they will be more concerned with larger datasets, improved sample size, and balanced qualities to overcome this problem.

A study by Dutta [26] offers a comparative analysis of ML and DL algorithms in lung cancer prediction using a Kaggle dataset that contained symptoms and lifestyle-related factors of 1,000 patients. To enhance the model robustness, the author used a wide range of preprocessing, including feature selection via Pearson correlation, elimination of outliers, and normalization of the data. Several ML classifiers were tested in Weka and neural networks models with 1-3 hidden layers coded in Python. The accuracy on performance evaluation is 92.86%, with the absence of XAI methods [26].

This paper suggested a hybrid of the fuzzy and neuro structure of classification, which is applied to the medical images in order to detect lung cancer. Reddy et al. [27] apply the dataset of CT scans that contains both benign lung nodules and malignant ones and perform the required preprocessing methods, including the elimination of the noise with the assistance of the median and the Gaussian filters, localization of the lung region with the assistance of the thresholding and the morphological operations, and contrast enhancement to make the nodules more noticeable. The basic algorithm is a mixture of fuzzy neural network and fuzzy logic addressing the ambiguity in pixel variation along with the neural network part, which learns higher-level patterns, with the help of handcrafted texture and statistical features of segmented nodules. The model has attained an accuracy of 96.67 and is more competitive than conventional standalone classifiers. This approach did not have any XAI techniques [27].

A paper [28] describes a transfer learning system for the classification of lung carcinoma based on a DL model using publicly available CT image datasets of various types of lung cancer and normal samples. The preprocessing pipeline includes data augmentation and transformation methods. The classification of lung nodules was performed using transfer learning models such as VGG16, VGG19, and Xception. VGG16, VGG19, and Xception achieved accuracies of 98.83%, 97.4%, and 98.05%, respectively, during performance evaluation. The research does not use any XAI framework, including Gradient-weighted Class Activation Mapping (Grad-CAM), Local Interpretable Model-Agnostic Explanations (LIME), or Shapley Additive Explanations (SHAP), which restricts interpretation of the fine-tuned convolutional layers and impedes clinical validation [28].

Shatnawi et al. [29] built a CNN-based diagnostic system trained on a dataset of chest CT-scan images containing malignant and benign lung cases obtained from publicly available clinical repositories. The preprocessing pipeline involved noise reduction using a median filter, contrast enhancement using histogram equalization and contrast-limited adaptive histogram equalization (CLAHE), followed by morphological operations and data augmentation. The authors used an improved CNN and several pre-trained models, including ConvNeXt, VGG16, ResNet50, InceptionV3, and EfficientNetB0. The accuracies of ConvNeXt, VGG16, ResNet50, InceptionV3, and EfficientNetB0 were 87%, 99%, 94.5%, 76.9 %, and 97.9%, respectively. Although the study primarily focused on predictive performance, no specific XAI technique was applied [29].

Prasad et al. [30] employed a hybrid metaheuristic algorithm-based deep neural network to classify and detect lung cancer in biomedical lung images gathered from the LIDC/IDRI dataset and chest X-ray images. They applied bicubic interpolation for noise removal, along with a hybrid Spotted Hyena Optimization and Seagull Algorithm for optimal feature selection and generative modeling in the preprocessing pipeline. A hybrid CNN-Long Short-Term Memory model was used for classification. The accuracy, sensitivity, specificity, and precision achieved were 99.6%, 99.8%, 99.34%, and 99.14% on the LIDC/IDRI dataset, and 99.70%, 99.62%, 97.80%, and 97.50% on the chest X-ray dataset, respectively. No XAI methods were used in this approach [30].

Mohandass et al. [31] used a curated CT scan dataset of benign, malignant, and normal lung CT images from publicly available registries, supplemented with hospital scans to balance the classes. The preprocessing pipeline included Hounsfield unit normalization, contrast enhancement using CLAHE, lung region segmentation using a threshold-based method, and noise reduction using Gaussian and median filters to enhance the visibility of the nodules. The authors developed an optimized attention-based CNN model and restrained DenseNet-201 transfer learning model, which allowed rich hierarchical feature extraction, while attention modules increased focus on clinically significant nodule regions. Their approach consisted of modifying DenseNet-201 and adding channel-wise and spatial attention blocks, followed by training with data augmentation, dropout regularization, and Adam optimization. The optimized architecture recorded high diagnostic results, with an accuracy of 98.92%, sensitivity of 98.40%, specificity of 99.10%, and an F1-score of more than 98%, outperforming baseline CNN and traditional DenseNet variants. The study lacked any XAI framework, even though the model proved to have good predictive ability [31].

The designed Lung-AttNet framework presents a lightweight CNN architecture supplemented with a Lightweight Global Attention Module (LGAM) that enhances discrimination among lung cancer classes. The model is trained on Kaggle CT scan data of adenocarcinoma, large cell carcinoma, squamous cell carcinoma, and normal images, extracting both low- and high-level features using convolutional blocks, while LGAM captures relationships in large-scale spatial and channel dependencies. It includes extensive experimentation, ablation studies, and 5-fold cross-validation, and XAI tools such as Grad-CAM and LIME are used to check the relevance of the features and the transparency of the models. In its classification process, Lung-AttNet achieves an average accuracy of 91.5%. To address limitations in medical data sharing, the model is incorporated into a federated learning framework, which allows model training using aggregated local weights rather than patient data around the world. Lung-AttNet achieves up to 92% accuracy in federated learning settings with two and three clients, demonstrating its robustness, privacy protection, and applicability to real-world clinical applications [32].

Kanavati et al. [33] introduced a loosely monitored DL architecture to classify lung carcinoma based on whole-slide histopathology images, a task dealing with the issue of few pixel-scale annotations. They used a large dataset of digitized lung tissue whole-slide images (WSIs), collected across several clinical centers in Japan, including cases of adenocarcinoma, squamous cell carcinoma, and benign samples, where only slide-level labels were used. Preprocessing included color normalization to minimize staining variability, patch identification of gigapixel WSIs, artifact removal, and tiling strategies to ensure balanced sampling of diagnostically relevant tissue regions. A CNN trained under a weakly supervised multiple-instance learning paradigm was used, enabling the model to learn discriminative patch-level features even in the absence of structured annotations, and an attention-based multiple-instance learning pooling scheme was added to emphasize the most relevant regions for final decision-making. The system achieved high diagnostic performance, with accuracy over 96% and AUC greater than 0.95 for major carcinoma subtypes. For interpretability, the authors used attention heatmaps as an intrinsic XAI mechanism to visualize histological regions contributing most to the classification, but did not use external XAI methods such as Grad-CAM, LIME, or SHAP [33].

Naveenraj and Vijayakumar [34] developed an enhanced lung cancer classification framework that combines an optimized hybrid DL method to improve diagnosis using CT images. The study used a publicly available CT scan dataset comprising malignant, benign, and healthy cases, which was augmented to address class imbalance. Preprocessing involved segmentation of lung regions using thresholding and morphological operations, removal of noise using Gaussian and median filtering, and contrast enhancement using CLAHE to enhance the visibility of nodules. The approach involved the combination of a dual deep CNN (DCNN) and Long Short-Term Memory network, which complement one another in terms of the overall features extraction and learning over time. The proposed model achieved a classification accuracy of more than 98%, along with high sensitivity, specificity, precision, and F1-score, and outperformed traditional CNN and single-model architectures. Nevertheless, no XAI techniques were incorporated by the authors [34].

The computer-aided decision support system proposed here also has a strong 3D-DCNN framework that can assist radiologists in detecting and diagnosing lung nodules more accurately. Through a combination of median intensity projection and a multi-Region Proposal Network, the system successfully extracts rich 3D contextual information from CT scans and automatically detects high-quality candidate regions, improving diagnostic reliability. The model, tested on benchmark datasets such as LUNA16, ANODE09, and LIDC-IDRI, demonstrates high diagnostic capability, with 98.4% sensitivity, 92% specificity, 96% AUROC, and 98.51% accuracy, along with only 2.1 false positives per scan. Scalability is further enhanced through cloud-based computation, which allows training and validation on large clinical datasets from Shanghai Sixth People’s Hospital as well as public repositories. Altogether, the cloud-based 3D-DCNN outperforms several state-of-the-art systems, achieving an impressive 98.7% sensitivity with 1.97 false positives per scan, highlighting its potential as a reliable second-opinion tool for lung cancer screening and clinical decision-making [35].

The framework introduced by Zhang and Kong [36] is a multi-scene DL system used to automatically identify lung nodules through CT scan datasets such as LIDC-IDRI, which contains a wide variety of nodule types and cases. The preprocessing pipeline includes lung segmentation, Hounsfield unit normalization, noise removal using Gaussian filtering, and multi-scale patch extraction to capture nodules of varying sizes. A multi-branch CNN model is integrated and trained to learn complementary features from different scene representations, including whole-lung views, region of interest patches, and contextual slices, which increases the capability to detect the nodules that are either subtle or small. The system had high diagnostic performance, as the reported accuracy and sensitivity are greater than 98%, as compared to the conventional single-view CNN methods. The study, however, does not indicate discussable AI techniques [36].

DFD-Net study presents a DL-based framework to detect lung cancer based on CT scan data such as LIDC-IDRI, with an emphasis on high-quality nodule analysis using advanced denoising techniques. A dual filtering method using Gaussian and median filters is applied in the preprocessing pipeline to reduce noise and enhance contrast, followed by feeding the processed images into the model. The proposed DFD-Net architecture combines an auto-denoising network with a feature-extracting CNN, which allows effective learning of robust representations from low-dose or noisy CT scans. The model achieved an accuracy of more than 96%, with high sensitivity and specificity, which proves that the model is reliable in detecting cancer at an early stage. Nevertheless, the research fails to implement XAI methods, including Grad-CAM or LIME, restricting its clinical applicability of the identified nodules [37].

Wahab Sait [38] introduces a DL architecture to detect lung cancer using PET/CT scans, including sophisticated preprocessing and augmentation techniques that help reduce noise and artifacts. The author used DenseNet-121 for robust feature extraction and deep autoencoders to efficiently reduce the dimensionality of features. A MobileNet V3-Small classifier was then used for lightweight and accurate identification of light subjects of lung cancer subtypes. Quantization-aware training, early stopping, and Adam optimization were used to improve performance, which guaranteed high accuracy at low computational cost. The model was tested on the Lung-PET-CT-Dx dataset and achieved 98.6% accuracy and a Cohen's kappa of 95.8, indicating a strong model with minimal parameters. The proposed strategy has potential for real-time clinical use, and future research considers liquid neural networks and ensemble strategies to enhance diagnostic capabilities of the approach [38].

A study [39] proposed two CNN-based DL architectures to diagnose lung cancer in CT images, but they included a straight 3D-CNN with softmax and a hybrid 3D-CNN with an Radial Basis Function-SVM classifier. Benchmarks were applied using modified versions of 3D-AlexNet and 3D-GoogleNet. The experimental results indicate that both proposed models perform better than the baseline architectures, and the hybrid 3D-CNN + SVM shows high diagnostic potential. It achieved an accuracy of 91.81%, a sensitivity of 88.53%, and a precision of 91.91%, which is better than the straight 3D-CNN, and demonstrates the usefulness of deep feature extraction combined with an SVM decision layer for better lung cancer classification [39].

A study [40] used a multi-output CNN model to classify and stage PET/CT images of lung cancer. Based on the TNM staging system and histologic subtype prediction, the authors used a VGG16-based feature extractor and a three-branch classifier to independently predict tumor size (T), nodal involvement (N), and histologic categories. The experimental results using the Lung-PET-CT-Dx dataset show strong performance, with tumor size classification accuracy of 0.94 and AUC of 0.97, and N staging accuracy of 0.98. The model is also more effective than existing methods in histologic subtype classification, indicating the effectiveness of multi-task learning in PET/CT-based lung cancer evaluation [40].

Author	Year	Contribution	Dataset	Pre-processing	ML/DL/Methodology	Evaluation Metrics	XAI	Limitation
Qadir et al. [4]	2024	Lung cancer prediction using machine learning	NN, RF	Not clear	ML: RF, NN	Accuracy of NN = 0.75, MSE of NN = 0.25, Accuracy of RF = 0.89, MSE of RF = 0.21	No	Small non-imaging dataset
Ahmed et al. [5]	2020	Lung cancer prediction using novel approaches based on risk factors	Risk factor dataset with clinical data	Missing value handling, label encoding, and normalization	ML: RF, WEKA tool	Accuracy = 93.33%, Sensitivity = 100%, Specificity = 90.47%	No	Small non-imaging dataset with limited generalizability
Yang et al. [6]	2022	Application of machine learning for personalized prediction of lung cancer recurrence and survivability	NSCLC with two major subtypes: LUAD (511) and LUSC (487)	Removal of irregular clinical records, standardization of continuous parameters	ML: DT, NN, SVM	C-index > 0.75	No	Non-uniform data distribution
Mishra et al. [7]	2021	An IoHT-driven intelligent healthcare monitoring system for long-term lung cancer risk	IOHT sensor data	Noise filtering and time series normalization	ML: Gradient Boosting Feature Selection with RF	Accuracy = 98.8%, Specificity = 97.5%, Sensitivity = 97.8%	No	Dependence on IoT stability; absence of imaging data

		assessment						
Venkatesh et al. [8]	2020	A bio-inspired algorithm for CT scan analysis and IoT cloud-based secure data transmission for lung cancer detection	CT images from private and public database	Filtering (median filter), segmentation (Otsu thresholding), Particle Swarm Optimization, and feature extraction (LBP)	ML: SVM	Accuracy = 93.53%, Sensitivity = 92.96%, Specificity = 98.52%	No	Lack of clear dataset description
Abdullah et al. [9]	2021	Forecasting and classification of lung cancer using correlation-based feature selection in machine learning	Lung cancer dataset from the UCI Machine Learning Repository	Correlation-based feature selection	ML: KNN, SVM; DL: CNN	Accuracy of SVM = 95.56%, Accuracy of KNN = 88.40%, Accuracy of CNN = 92.11%	No	Small structured dataset
Hussain et al. [10]	2022	Prediction of lung cancer based on strong machine learning and image enhancement technique on the extracted graylevel co-occurrenc	Digital Imaging and Communications in Medicine images from public datasets	Image enhancement technique, feature extraction, thresholding technique, gamma correction	ML: SVM, Decision Tree, Naïve Bayes	Without image enhancement, accuracy of SVM, RBF, and polynomial = 99.89%. With image enhancement, accuracy of SVM, RBF, and	No	Classical ML technique used

		e matrix features				polynomial = 100%		
Nanglia et al. [11]	2021	Classification of lung cancer using hybrid SVM and neural network	Low-dose CT scan image datasets	Feature scaling and correlation-based feature reduction	ML: SVM, NN	Accuracy = 98.08	No	Small dataset used
Gupta et al. [12]	2016	IOT-based smart healthcare kit	IoT sensor signals	Signal filtering, calibration	IoT health kit, simple ML	Reliability (represents a low cost, compact, and scalable health monitoring system that provides quick and efficient results)	No	No medical image, less diagnostic depth
Kalaivani et al. [13]	2020	Recognition and categorization of lung carcinoma using deep learning	CT scan images	Image resizing, lung region extraction, and contrast enhancement using histogram equalization	DL: CNN-based architecture	Accuracy = 90.85%	No	No comparison with other DL techniques
Pawar et al. [14]	2022	Prediction of lung cancer using blockchain technology integrated with an	CT scan images	Normalization and contrast adjustment	DL: Extended CNN, blockchain	Accuracy = 96.88%	No	High computational cost

		extended CNN						
Makaju et al. [15]	2018	Identification of lung carcinoma from CT scan images	CT scan images	Filtering (median and Gaussian filters) and segmentation using the watershed method	ML: SVM	Accuracy = 92%, Sensitivity = 100%	No	Limited architecture depth
Palani and Venkatalakshmi [16]	2019	Lung cancer prediction using fuzzy clustering-based segmentation and classification in an IoT-based model	Standard images and health data collected via IoT	Fuzzy C-means clustering, Otsu thresholding method, morphological operations	ML: Association Rule Mining, Decision Tree, CNN	Accuracy = 85%, Sensitivity = 87%, Specificity = 85.9%	No	Limited dataset
Saleh et al. [17]	2021	Classification of lung carcinoma using a combined CNN-SVM model	CT images	Data augmentation	ML: SVM, DL: CNN	Accuracy = 97.91%, Sensitivity = 97.9%, Specificity = 99.32%, Precision = 97.96%, AUC = 1.00	No	No clear preprocessing
Shanbhag et al. [18]	2022	Lung cancer prediction using ensemble classifiers	CT scan images	Noise filter using Gaussian filter, segmentation using Otsu thresholding, feature extraction	Ensemble classifier of SVM, LR, MLP, Decision Tree, KNN	Accuracy = 85%	No	limited feature representations

Wang and Chakraborty [19]	2021	Automatic prognosis of lung cancer using diverse deep learning models for nodule detection and morphological feature analysis	CT scan images	Noise reduction, feature extraction	3DCNN, RNN	Log loss = 0.408	No	Lack of clinical interpretability
Shakeel et al. [20]	2022	Automatic lung cancer detection from CT image using improved deep neural network and ensemble classifier	Chest CT scan image	Noise reduction using Sub-Aperture Keystone Transform Matched Filtering (SAKTMF), segmentation using Automatically Weighted Binary Multi-View Clustering (AW-BMVC), feature extraction using Second-Order Synchroextracting Wavelet Transform (SOSWT)	Robust Deformed CNN	Improvement in dice similarity by 13.67%, 27.55%, and 14.67 and 22.23%, 24.11%, and 25.56% reductions in log-loss	No	Lack of clinical interpretability
Ali et al. [21]	2023	Lung nodule recognitio	CT images	Noise reduction, oversampl	2D CNN model	Accuracy = 83%, Specificity	No	Lack of clinical

		n and anomaly detection using a two-dimensional CNN architecture		ing to avoid class imbalance		= 98.42%, Sensitivity = 68.66%		interpretability
Lanjewar et al. [22]	2023	Recognition of lung carcinoma using a modified DenseNet with feature selection and machine learning classifiers on CT images	CT scan image	feature extraction technique	DenseNet 201	Accuracy = 95%	No	Lack of clinical interpretability
Al-Tawalbeh et al. [23]	2022	Lung carcinoma classification using machine learning algorithms	Textual data related to patient symptoms	Data missing handling, label encoder, normalization	SVM, KNN, and Naive Bayes, narrow neural network	Accuracy of SVM = 92.6%, Accuracy of KNN = 85.87%	No	Lack of clinical interpretability
Doppalapudi et al. [24]	2021	Prediction and explication of lung carcinoma survival using deep learning	CT images	Noise reduction, feature extraction	ANN, CNN, RNN	Deep learning model classification approach = 71.18%, Regression approach RMSE = 13.5%, R2 = 0.5	No	Lack of clinical interpretability
Nassif et al. [25]	2025	Gene expression data-	LUAD, LUSC	F-test feature	CNN	With LUAD dataset, accuracy =	No	Lack of clinical

		based deep learning in classifying the severity of lung cancer		selection technique		93.94%. With LUSC dataset, accuracy = 88.42%		interpretability
Dutta [26]	2025	Comparative analysis of ML and DL models for lung cancer prediction based on symptoms and lifestyle factors	Patient symptoms	Data missing handling, label encoder, normalization	Decision Trees, KNN, RF, Naïve Bayes, AdaBoost, LR, and SVM	Accuracy = 92.86%	No	Lack of clinical interpretability
Reddy et al. [27]	2019	Lung cancer prediction using hybrid fuzzy neural network	CT scan images	Noise reduction using median and Gaussian filter; lung region extraction using thresholding and contrast enhancement	Fuzzy Neural Network	Accuracy = 96.67%	No	Lack of clinical interpretability
Humayun et al. [28]	2019	Classified lung nodules using transfer learning approach	CT scan images	Augmentation, transformation	VGG16, VGG19, Xception	Accuracy of VGG16 = 98.83%, Accuracy of VGG19 = 98.05%, Accuracy of Xception = 97.4%	No	Lack of clinical data

Shatnawi et al. [29]	2025	Diagnosed lung cancer using deep learning approach	Chest CT scan image dataset	Noise reduction using median filter; contrast enhancement using HE, CLAHE, morphological operation, data augmentation	Enhanced CNN, ConvNext, VGG16, ResNet50, Inception V3, EfficientNetB0	Accuracy of enhanced CNN = 100%, Accuracy of ConvNext = 87%, Accuracy of VGG16 = 99%, Accuracy of ResNet50 = 94.5%, Accuracy of Inception V3 = 76.9%, Accuracy of EfficientNetB0 = 97.9%	No	Lack of clinical data
Prasad et al. [30]	2023	Used hybrid metaheuristic algorithm-based deep neural network for the classification and detection of lung cancer	LIDC/IDRI, chest X-ray dataset	bicubic interpolation for noise elimination, hybrid spotted Hyena optimization with Seagull algorithm for optimum feature selection and generative modelling technique for data augmentation	Hubrid CNN-LSTM	For LIDC/IDRI, Accuracy = 99.6%, Sensitivity = 99.8%, Specificity = 99.3%, Precision = 99.14%. For chest X-ray, Accuracy = 99.7%, Sensitivity = 99.62%, Specificity = 97.8%, Precision = 97.5%	No	Lack of clinical interpretability

Mohandas s et al. [31]	2024	Used optimized attention-based CNN with DenseNet 201 for the classification of lung cancer	CT scan images from LIDC/IDRI	Hounsfield unit normalization, contrast enhancement using CLAHE, lung region extraction through threshold segmentation, and noise reduction with Gaussian and median filters	Used optimized attention-based CNN with DenseNet 201	Accuracy = 98.2%, Sensitivity = 98.4%, Specificity = 99.1%, Precision = 98%	No	Lack of clinical interpretability
Saha et al. [32]	2025	Lung cancer detection using federated learning with an attention-based CNN architecture	CT scan images	Image resize, image enhancement using CLAHE	Lung-AttNet(convolution block+ LGAM)	Accuracy = 92%	Grad-CAM, LIME	Lack of clinical data
Kanavati et al. [33]	2020	Weakly supervised deep learning framework for lung carcinoma classification using whole-slide histopathology images	Whole-slide histopathology images	Color normalization to reduce staining variability, patch extraction from gigapixel WSIs, artifact removal, and tiling	CNN	Accuracy = 96%	No	Lack of clinical interpretability

				strategies that ensured balanced sampling of diagnostically relevant tissue regions				
Naveenraj and Vijayakumar [34]	2025	Developed optimized hybrid DL framework for lung cancer classification	CT scan images	Lung region segmentation using thresholding and morphological operations, followed by noise removal through Gaussian and median filtering and contrast enhancement using CLAHE	DCNN + LSTM	Accuracy = 98%	No	Lack of clinical interpretability
Schwyzer et al. [35]	2020	Detection and diagnosis of lung cancer in chest CT using a cloud-based automated clinical decision support system	LUNA16, ANODE09, and LIDC-IDRI	3D contextual information extraction By leveraging median intensity projection and a multi-Region Proposal Network	Cloud-based 3DCNN	Sensitivity = 98.7%, Specificity = 92%, AUROC = 96%, Accuracy = 98.51%	No	Lack of clinical interpretability

Zhang and Kong [36]	2020	Lung cancer detection using a deep learning-based framework	LIDC/IDRI	Segmentation, Hounsfield unit normalization, noise reduction using Gaussian filtering, and multi-scale patch extraction	Multi-branch CNN architecture	Accuracy = 98.7%	No	Lack of clinical interpretability
Sori et al. [37]	2021	Detection of lung cancer using deep learning-based framework	LIDC/IDRI	Dual filtering (Gaussian +median) for noise reduction, contrast enhancement	CNN	Accuracy = 96%	No	Lack of clinical interpretability
Wahab Sait [38]	2023	Lung cancer prediction using deep learning techniques	Lung-PET-CT-Dx dataset	Noise reduction, data augmentation	DenseNet-121–based CNN, deep autoencoders, MobileNet V3-Small	Accuracy = 98.6%	No	Lack of clinical interpretability
Polat et al. [39]	2019	Lung cancer classification using a hybrid 3D CNN + SVM model	CT scan images	Noise reduction, augmentation, enhancement	3DCNN+SVM	Accuracy = 91.81%, Sensitivity = 88.53%, Precision = 91.91%	No	Lack of clinical interpretability
El Hamdi et al. [40]	2024	Lung cancer classification of PET/CT images using CNN	PET/CT	Noise reduction, augmentation, enhancement	Multi-output CNN, VGG	Accuracy = 98%	No	Lack of clinical interpretability

TABLE 3: Categorical representation of the reviewed articles

ANN, Artificial Neural Network; AUROC, Area Under the Receiver Operating Characteristic Curve; CLAHE, Contrast Limited Adaptive Histogram Equalization; CNN, Convolutional Neural Network; CT, Computed Tomography; DCNN, Deep Convolutional Neural Network; DL, Deep Learning; DT, Decision Tree; HE, Histogram Equalization; IoHT, Internet of Health Things; IoT, Internet of Things; KNN, k-Nearest Neighbors; LIDC/IDRI, Lung Image Database Consortium,Image Database Resource Initiative; LR, Logistic Regression; LSTM, Long Short-Term Memory; LUAD, Lung Adenocarcinoma; LUSC, Lung Squamous Cell Carcinoma; ML, Machine Learning; MLP, Multilayer Perceptron; MSE, Mean Squared Error; NN, Neural Network; NSCLC, Non-Small Cell Lung Cancer; PET, Positron Emission Tomography; RBF, Radial Basis Function; RF, Random Forest; RNN, Recurrent Neural Network; SVM, Support Vector Machine; UCI, University of California, Irvine (Machine Learning Repository); XAI, Explainable Artificial Intelligence

Trend analysis and research gap

Recent studies on lung cancer show that there have been significant changes toward DL-focused, multifaceted, and clinically incorporated diagnostic models. The prevailing trend is now toward CNNs, 3D-CNNs, transfer learning, attention mechanisms, and feature selection (using a combination of classical ML methods, e.g., SVM, KNN, DT, and correlation-based methods) in CT, PET/CT, and multi-omics data (e.g., [12,16,18,20,28,31,33,39,40]). Another trend is the combination of IoT- and IoHT-based monitoring, blockchain-enhanced security, cloud models, and federated learning to achieve privacy-preserving clinical implementation [6,7,11,13,15,31]. Multi-task learning is also a common trend in many studies, extending beyond predicting the presence of cancer to include subtypes, severity, recurrence, survivability, TNM staging, and morphological characteristics [5,14,18,23,24,35,40]. Furthermore, image repair, denoising, and multi-view/multi-scene feature retrieval have become significant to enhance the performance in the real clinical environment [9,21,36,37]. The collective trend suggests a shift toward lightweight, interpretable, federated, and deployable AI systems, with a focus on clinical realism, multi-center datasets, and high generalizability. Although there has been significant advancement, there are still a number of research limitations in the studies. To begin with, most models are based on smaller, homogeneous datasets or single-centered scans (e.g. LIDC-IDRI, LUNA16, Kaggle lung CT sets) that limit generalizability to the real world; few studies would consider multi-institutional or large-scale clinical PET/CT data [20,28,34,40]. Second, despite superior results of attention modules, 3D-CNNs, and hybrid ML-DL, small nodules and lesions in their early stages have a high false negative rate, and the literature has not considered how to correct the class imbalance by use of clinically safe augmentation methods. Third, the use of XAI is not consistent, as most of the studies have not demonstrated high interpretability beyond Grad-CAM/LIME, limiting the level of clinical trust and regulatory compliance. Fourth, there are systems and federated learning frameworks that are enabled by IoT/IoHT, but little validation is done in actual hospital settings; almost no studies show end-to-end integration with PACS, RIS, or oncologist workflows. Fifth, the majority of works are still concerned with detection/classification, but not with longitudinal prognosis, survival prediction, recurrence risk, and response-to-therapy estimation, which are essential to precision oncology. Selection of favorable dataset, preprocessing techniques, and classification algorithms for achieving better performance is highly necessary [41]. Lastly, little comparative modalities (CT vs. PET/CT vs. clinical + genetic data) benchmarking exists, and there is no common framework of robustness testing in the presence of noise, low-dose imaging, demographic variability, or adversarial perturbations.

Critical analysis of included studies

A closer analysis of the studied works [4-40] shows that there are several methodological limitations that go beyond the generally reported high accuracy rates. The strong dependency on small or single-source datasets is a key issue, as seen in studies such as [4,9,23], and it restricts the applicability of the models to a wider population. Most DL-based methods [13,20,31] have shown good performance; however, such performance is usually achieved under controlled experimental conditions with minimal external validation and cross-dataset testing, which increases the risk of overfitting and reduces reproducibility. Moreover, some research [10,22,30] relies on manual feature extraction or hybrid systems, making it challenging to compare results with end-to-end DL models due to inconsistent preprocessing pipelines. Class imbalance is another issue that has not been directly addressed in many studies, such as [6,18,26], and may result in biased models where the majority class is privileged. Moreover, the validation strategies differ greatly; some works use basic train-test splits [11,15] rather than more rigorous methods, such as k-fold cross-validation, which also impacts the validity of reported findings. Ensemble and hybrid models [17,18,20] are also shown to be more effective, but these models are not interpreted and are not tested in clinically realistic condition. Additionally, even with the progress in the field of DL, there is a deficit of XAI methods integration in most studies that limits the possibility of clinical trust and use. On the whole, despite the promising changes that are evident in the literature on the use of ML and DL to predict lung cancer, the existence of limitations in datasets, lack of overall methodological consistency, and lack of validation can prove the necessity of more standardized, transparent, and clinically focused research frameworks.

Discussion and analysis

Dataset, Dataset Limitations, and Generalizability

Lung cancer datasets include both imaging data and clinical data collected from different sources. The most frequently used lung cancer dataset information is shown in Table *4*.

Dataset	Types	Size
LIDC-IDRI (Lung Image Database Comsortium)	CT scan images	1,010 patients
TCGA-LUAD (Lung Adenocarcinoma)	CT, pathology, clinical, genomic	69
Lung Cancer Segmentation (Lung-RADs)	CT images+segmentation masks	708 training images+264 test images
NLST (National Lung Screening Trial)	Clinical data+imaging	2,100 records
LUNA16	3D chest CT images	888 CT scans, 1,186 annotated nodules
CPTAC-LSCC (Lung Squamous Cell Carcinoma)	Radiology images (CT, PET) + Histopathology (WSI) + Clinical/Proteomic/Genomic data	Radiology (CT/PT): ~52,019 images from 36 subjects; Histopathology (SVS WSI): ~1,081 images from 212 subjects; Total ~444.6 GB
Lung-PET-CT-Dx	PET/CT image	1,294 PET/CT images
NSCLC Radiogenomics Dataset	Textual data	211 NSCLC cases
Pulmonary Chest X-Ray Dataset	X-ray image	800–1,500 X-rays

TABLE 4: Lung cancer dataset details

CT/PET, Computed Tomography/Positron Emission Tomography; NSCLC, Non-Small Cell Lung Cancer; SVS, ScanVision Slide format; WSI, Whole Slide Image

A major drawback that can be noticed in most of the reviewed papers [4-40] is that they use small-scale or single-source datasets, which strongly affects the applicability and strength of the proposed models. Several studies, such as [4,9,23,26], used a limited or institution-specific datasets, making them effective in controlled settings but less generalizable to a wide range of clinical populations. In imaging-based research [13-20,31-39], data are often obtained from a single repository, or the data are not variable enough in terms of demographics, imaging procedures, and disease progression. This lack of diversity increases the risk of overfitting and reduces the reliability of models in real-world clinical settings. In addition, few studies use multi-center or cross-dataset validation, which is critical for evaluating the strength of the models. Although recent approaches such as federated learning [32] strive to eliminate the drawbacks of data sharing, they are not widely used. Therefore, future research must focus on large-scale, multi-institutional, and heterogeneous datasets, along with standardized benchmarking systems, to develop more valid and clinically useful lung cancer prediction models.

Preprocessing

This phase represents the preliminary stage wherein image data undergoes preprocessing operations encompassing denoising, normalization, binarization, and thresholding. Subsequently, a segmentation technique is employed to partition the image into distinctive regions, distinguishing between those exhibiting similarity and dissimilarity. Till now, no specific preprocessing pipeline has been finalized.

Augmentation

As large amounts of data are necessary to train a model in DL, and it is not always possible to have a huge amount of data, data augmentation techniques are applied to the existing dataset. Different augmentation operations such as flipping, rotation, shearing, and cropping have been used to enhance the dataset size.

Classification

In different model architectures suggested by various researchers, various ML classification algorithms such as LR, SVM, KNN, DT, Naive Bayes, neural networks, RF, and ensemble classifiers, as well as DL models including CNN, 3D-CNN, and transfer learning models (VGG19, ResNet50, DenseNet201, InceptionV3, EfficientNetB0), have been used in many studies. It is important to note that the highest accuracy results were always obtained in studies that implement DL methods for lung cancer prognosis. Overall, RF and the ensemble technique XGBoost have achieved high accuracy, ranging from 97% to 99.5%, compared to other ML techniques. Among DL models, CNN models achieved the highest accuracy, which ranges from 93% to 98%. The accuracy of a model mainly depends on the quality of the dataset used, the preprocessing procedures used, and the cross-validation strategies used.

The frequency of different datasets, preprocessing techniques, ML algorithms, DL algorithms, and evaluation metrics used by different researchers in the review section is shown in Figures *5*-*9*, respectively.

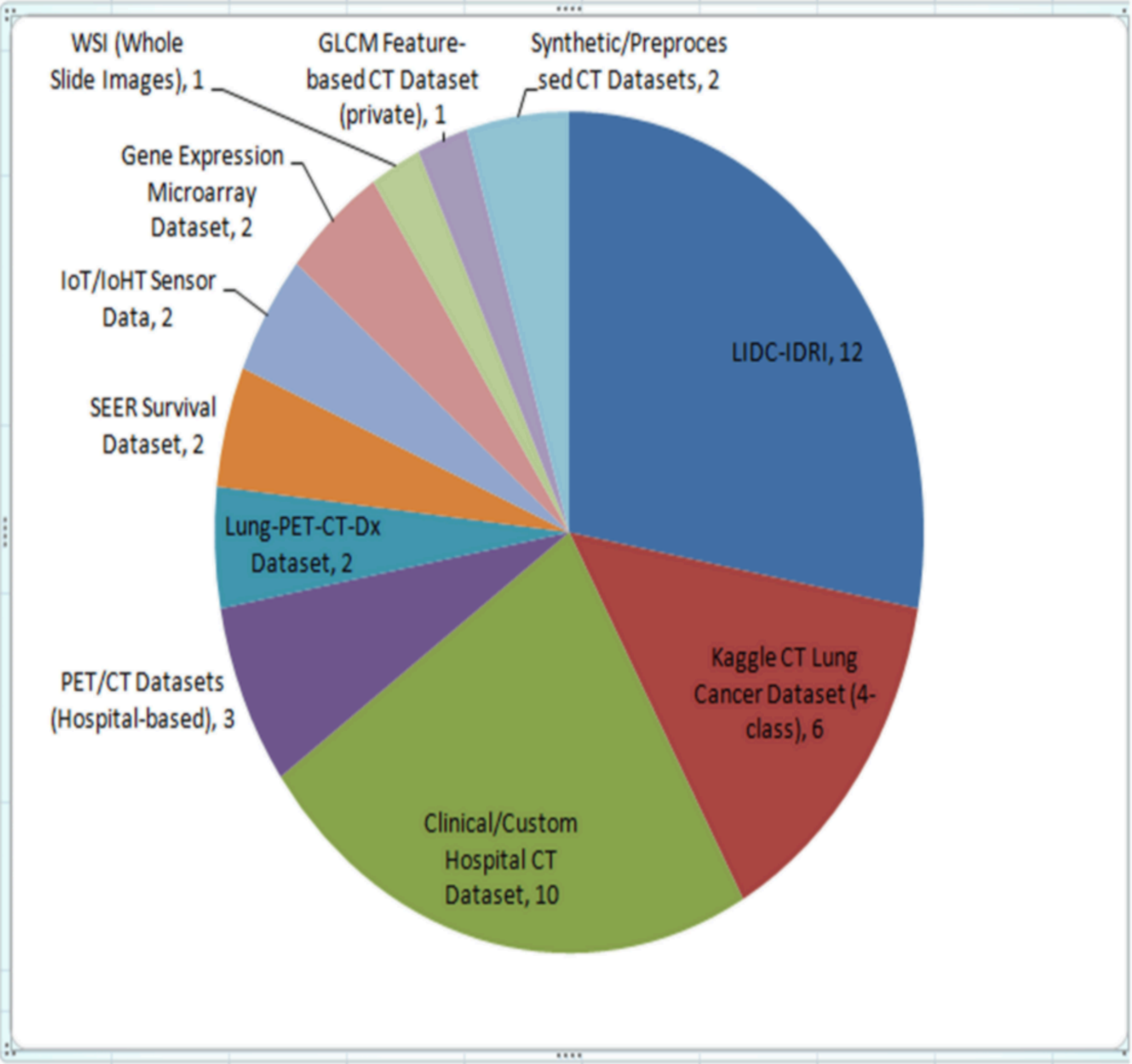


FIGURE 5: Frequency of datasets used in the review section

CT, Computed Tomography; IoHT, Internet of Health Things, IoT, Internet of Things; PET, Positron Emission Tomography

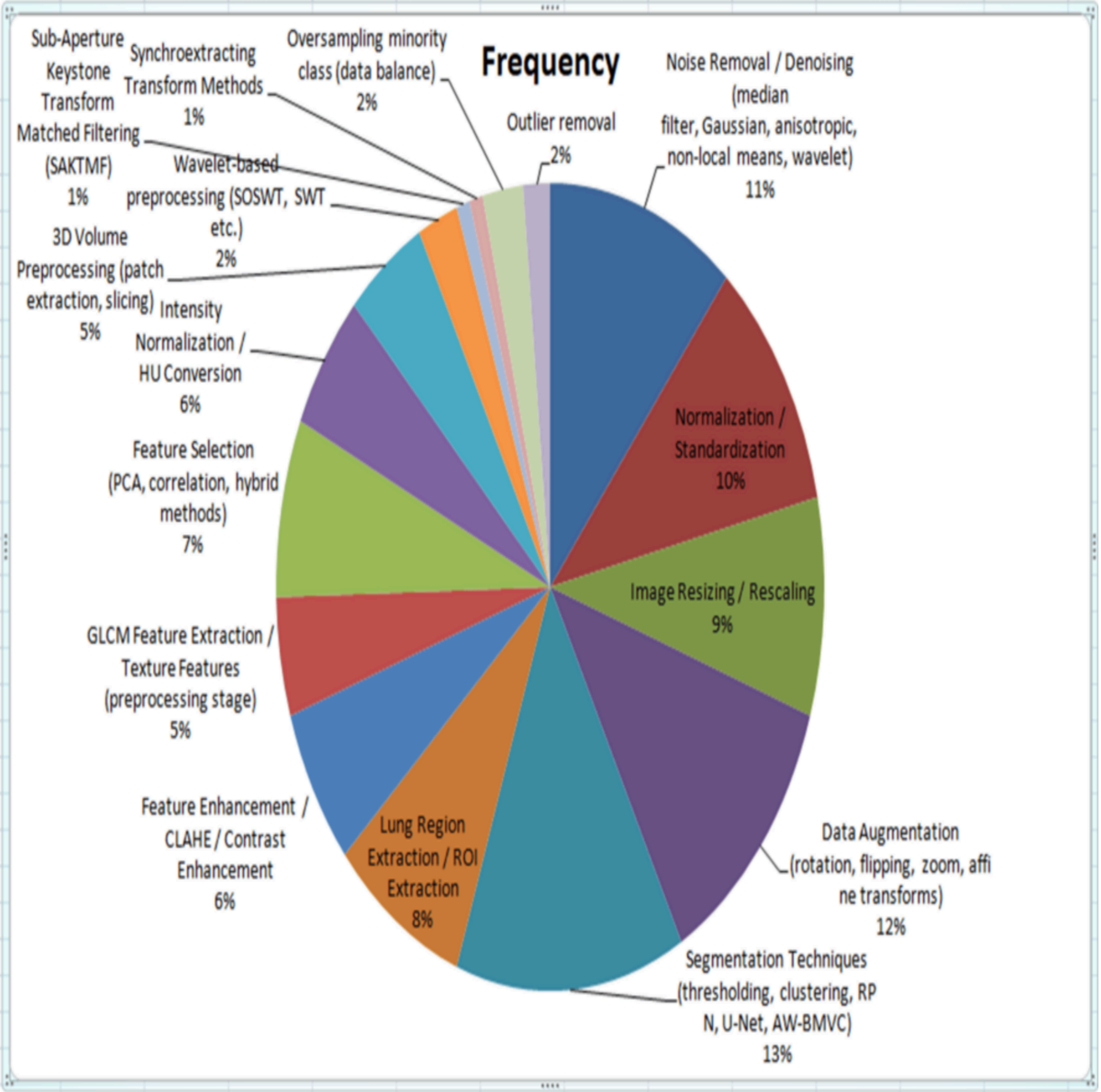


FIGURE 6: Frequency of preprocessing techniques used in the review section

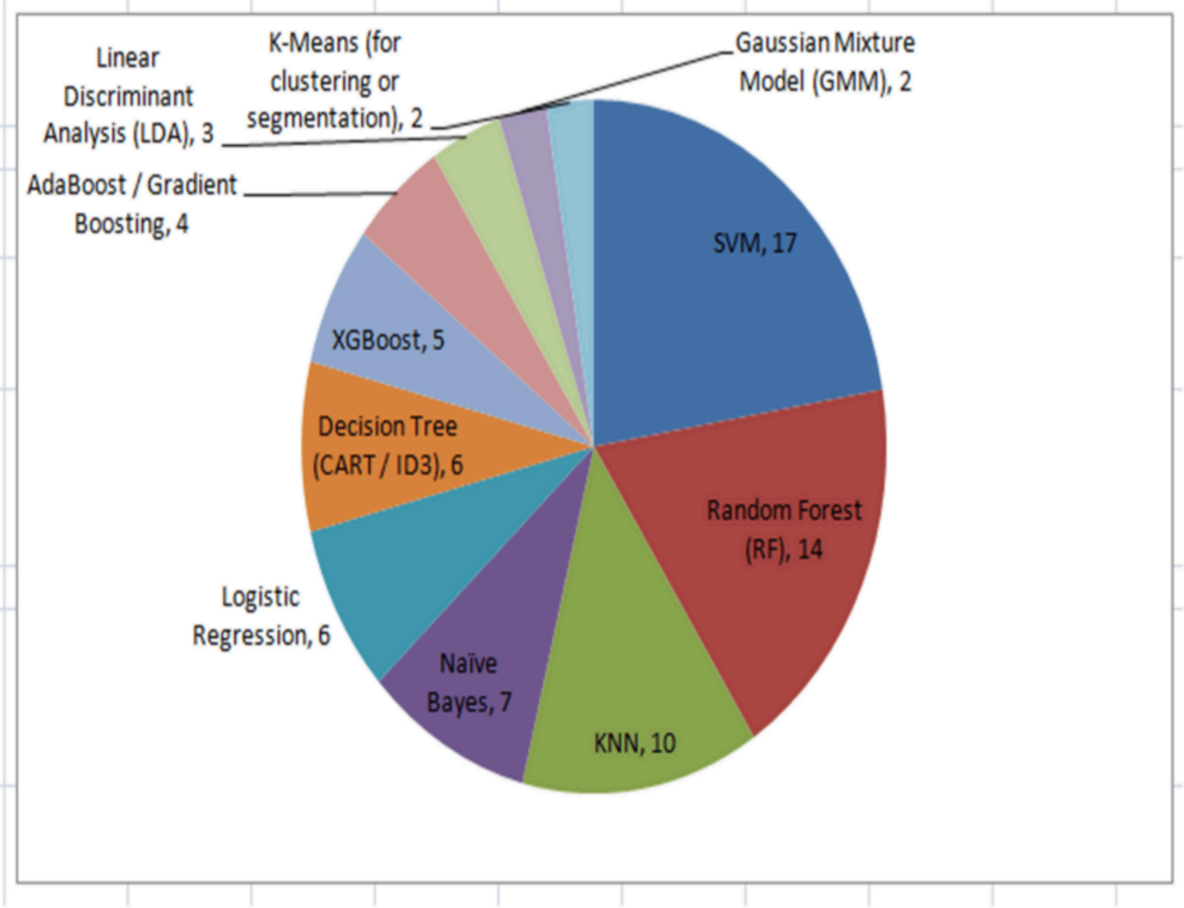


FIGURE 7: Frequency of machine learning techniques used in the review section

KNN, k-Nearest Neighbors; SVM, Support Vector Machine

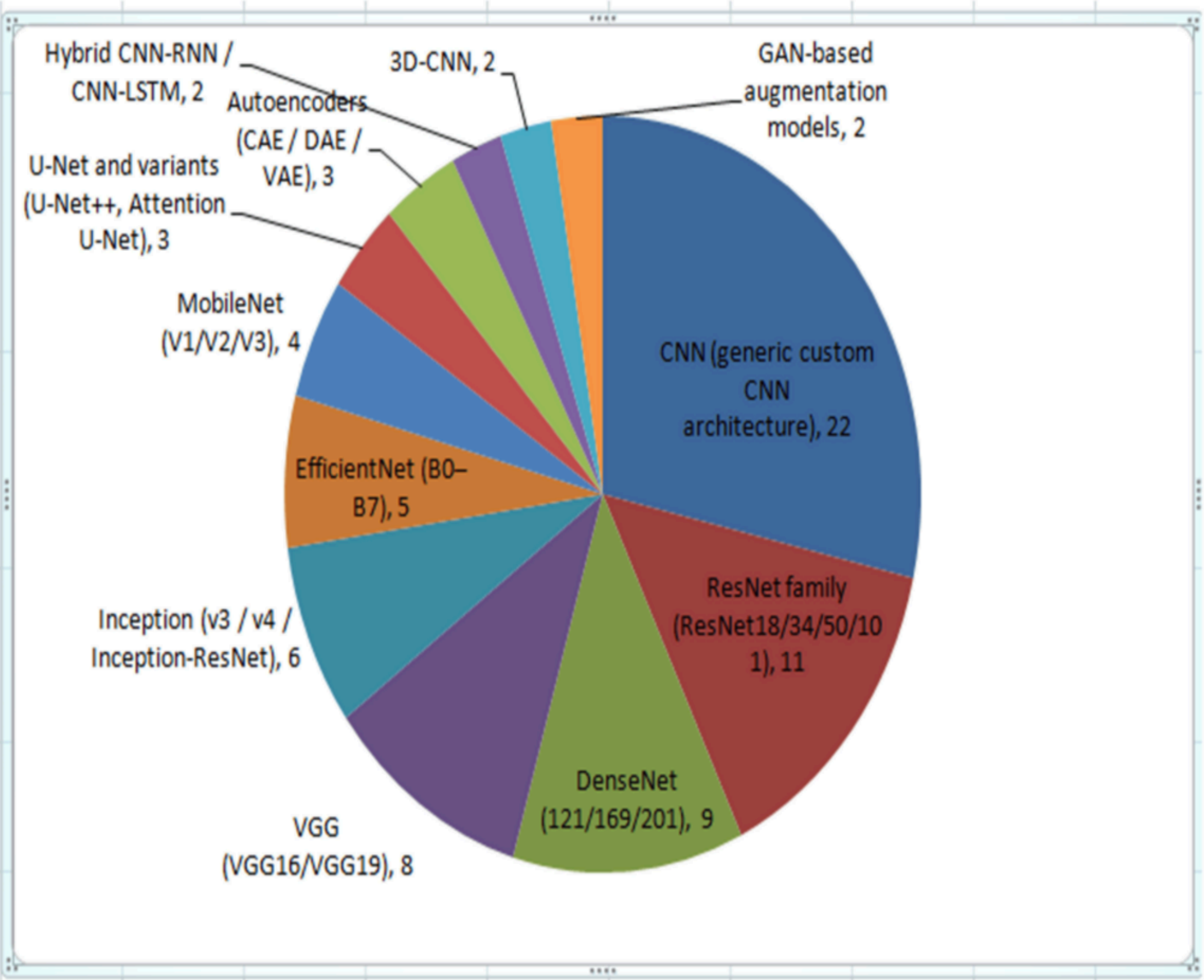


FIGURE 8: Frequency of deep learning techniques used in the review section

CAE, Convolutional Autoencoder; CNN, Convolutional Neural Network; GAN, Generative Adversarial Network; DAE, Denoising Autoencoder; LSTM, Long Short-Term Memory; VAE, Variational Autoencoder

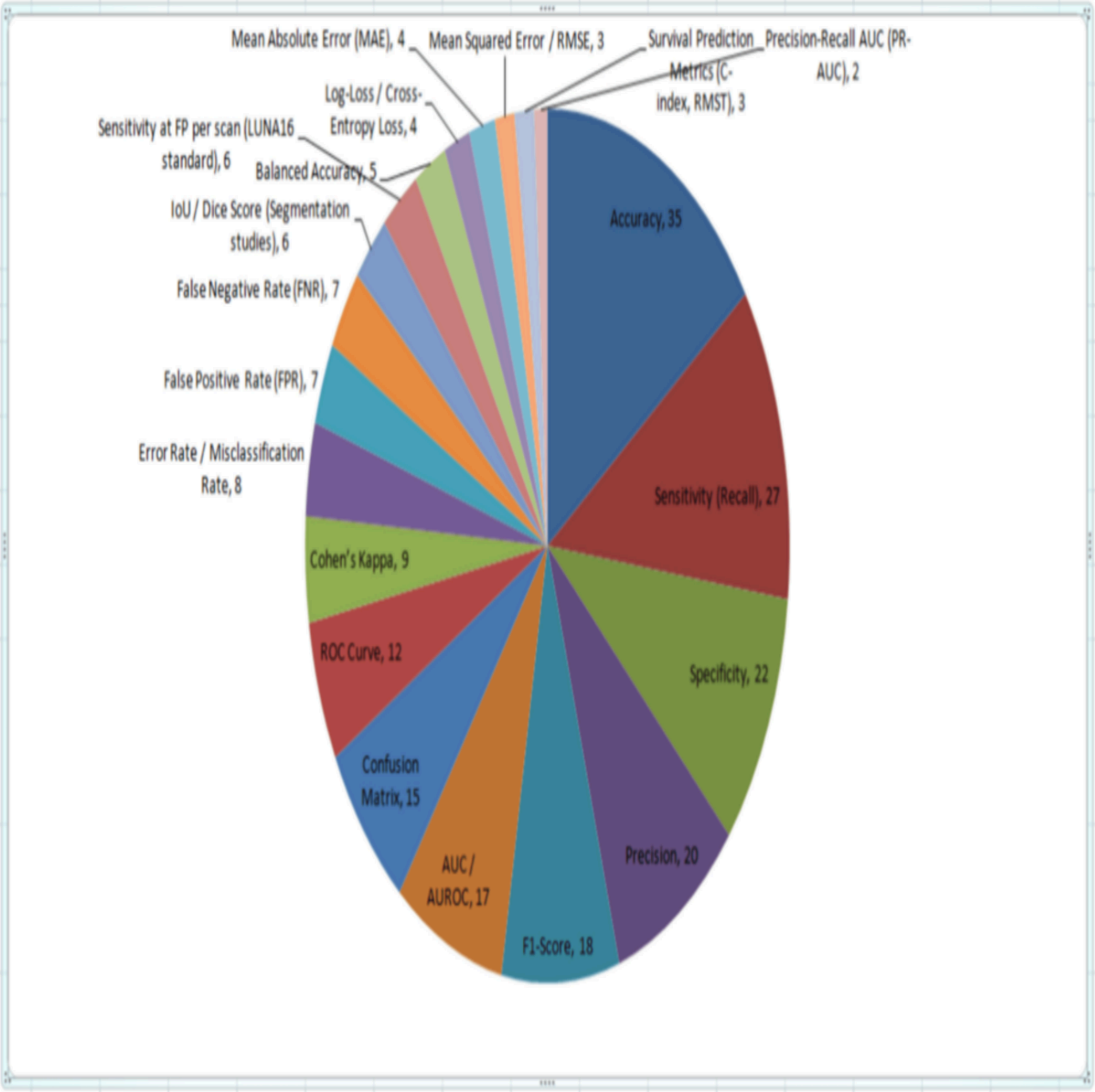


FIGURE 9: Frequency of different evaluation metrics used in the review section

AUC, Area Under the Curve; AUROC, Area Under the Receiver Operating Characteristic Curve; PR, Precision-Recall; RMSE, Root Mean Squared Error; RMST, Restricted Mean Survival Time

XAI and Clinical Interpretability

Although ML and DL models have proven in the literature to be highly effective in lung cancer detection and classification activities, they are not yet applicable in clinical practice because of the absence of interpretability. Most of the existing models are black boxes, and they do not provide a clear explanation of how decisions are made. This non-transparency decreases clinician trust and limits their application in real-world healthcare contexts where interpretability is essential in making informed decisions. Recent initiatives in XAI such as Grad-CAM, LIME, and SHAP techniques present possible solutions since they offer visualization and interpretation of model predictions. The techniques can be used to detect significant areas in medical images and measure the contribution of features and, thus, improve the transparency and reliability of the models. Nevertheless, the use of XAI methods in the existing lung cancer predictive research is still constrained. In future studies, consideration should be made to integrate interpretable models to guarantee sufficient predictive accuracy as well as clinical trust, accountability, and usability.

Beyond Classification: Toward Clinically Meaningful Outcomes

Even though a significant portion of the available literature is dedicated to classification accuracy as a way of distinguishing malignant and benign lung nodules, this represents only one aspect of clinically relevant decision-making. In real-world applications, patient survival prediction, disease progression, risk stratification, and treatment response assessment are equally or even more important outcomes. Nevertheless, most ML and DL studies are limited by the scarcity of longitudinal and well-labeled clinical data, restricting them to classification-based problems. Consequently, new models tend to fail to describe the complexity of individual patient disease courses. Future studies should extend clinical applicability beyond simple classification by incorporating multimodal data streams such as imaging, clinical records, genomic data, and follow-up information. Integrative methods like these may allow creating more holistic predictive models that help in planning personalized treatment and enhancing patient outcomes.

Conclusions

Lung cancer is an incapacitating non-communicable neoplastic disease that has great physiological and socioeconomic consequences for individuals and society. It is essential to carry out timely prognostication to prevent the negative effects of this disease. This paper provides an extensive review of systemic studies on a continuum of ML and DL approaches that can support predictive operations at their initial stages to identify lung cancer using medical images, and they can also classify the medical images as malignant or benign. As a solution to enable early detection of cancer, a new ensemble DL model, enhanced by strict cross-validation processes, will be presented in future work. Furthermore, the different phases of cancer development are also intended to be properly outlined, which increases the accuracy of treatment distribution to patients affected by the disease. In future work, hybrid ensemble DL methods can be supported with XAI. XAI methods can be applied to analyze and explain the significance of features playing important roles in prediction. There is a high need to validate an ML or DL approach before applying it in regular clinical practice. These findings suggest that more sophisticated models, together with different types of data, data augmentation procedures, and the use of ensemble learning, can provide more precise and secure predictions. These methods are especially useful in capturing various phases of the disease and facilitating early diagnosis. Nevertheless, these approaches require additional studies to demonstrate their effectiveness on large-scale and clinically heterogeneous datasets. Altogether, it is possible to note that the combination of these methods is a promising trend toward better AI-assisted lung cancer predictors.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Pragnya Das, Satya N. Tripathy, Kali P. Rath

Acquisition, analysis, or interpretation of data: Pragnya Das, Satya N. Tripathy, Kali P. Rath

Drafting of the manuscript: Pragnya Das, Satya N. Tripathy, Kali P. Rath

Critical review of the manuscript for important intellectual content: Pragnya Das, Satya N. Tripathy, Kali P. Rath

Supervision: Pragnya Das, Satya N. Tripathy, Kali P. Rath

Disclosures

Conflicts of interest: In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Data Availability Statements

Data sharing is not applicable. This work is theoretical; no datasets were generated or analyzed.

References

1. Key Statistics for Lung Cancer. (2025). Accessed: January 20, 2025: https://www.cancer.org/cancer/types/lung-cancer/about/key-statistics.html.
2. Sathishkumar K, Chaturvedi M, Das P, Stephen S, Mathur P: Cancer incidence estimates for 2022 & projection for 2025: Result from National Cancer Registry Programme, India. Indian Journal of Medical Research. 2022, 156:598-607. 10.4103/ijmr.ijmr_1821_22
3. What is the Internet of Things (IoT)?. (2023). Accessed: December 18, 2024: https://www.ibm.com/think/topics/internet-of-things.
4. Qadir RH, HamaKarim KM: Prediction of lung cancer disease using machine learning techniques. UHD Journal of Science and Technology. 2024, 8:75-83. 10.21928/uhdjst.v8n2y2024.pp75-83
5. Ahmad S, Mayya AM: A new tool to predict lung cancer based on risk factors. Heliyon. 2020, 6:e03402. 10.1016/j.heliyon.2020.e03402
6. Yang Y, Xu L, Sun L, Zhang P, Farid SS: Machine learning application in personalised lung cancer recurrence and survivability prediction. Computational and Structural Biotechnology Journal. 2022, 20:1811-1820. 10.1016/j.csbj.2022.03.035
7. Mishra S, Thakkar HK, Mallick PK, Tiwari P, Alamri A: A sustainable IoHT based computationally intelligent healthcare monitoring system for lung cancer risk detection. Sustainable Cities and Society. 2021, 72:103079. 10.1016/j.scs.2021.103079
8. Venkatesh C, Bojja P: Lung cancer detection using bio-inspired algorithm in CT scans and secure data transmission through IoT cloud. International Journal of Advanced Computer Science and Applications. 2020, 11:1-7. 10.14569/ijacsa.2020.0111148
9. Abdullah DM, Abdulazeez AM, Sallow AB: Lung cancer prediction and classification based on correlation selection method using machine learning techniques. Qubahan Academic Journal. 2021, 1:141-149. 10.48161/qaj.v1n2a58
10. Hussain L, Alsolai H, Hassine SBH, et al.: Lung cancer prediction using robust machine learning and image enhancement methods on extracted gray-level co-occurrence matrix features. Applied Sciences. 2022, 12:6517. 10.3390/app12136517
11. Nanglia P, Kumar S, Mahajan AN, Singh P, Rathee D: A hybrid algorithm for lung cancer classification using SVM and neural networks. ICT Express. 2021, 7:335-341. 10.1016/j.icte.2020.06.007
12. Gupta P, Agrawal D, Chhabra J, Dhir PK: IoT based smart healthcare kit. 2016 International Conference on Computational Techniques in Information and Communication Technologies (ICCTICT), New Delhi, India. 2016, 237-242. 10.1109/ICCTICT.2016.7514585
13. Kalaivani N, Manimaran N, Sophia S, Devi DD: Deep learning based lung cancer detection and classification. IOP Conference Series: Materials Science and Engineering. 2020, 994:012026. 10.1088/1757-899x/994/1/012026
14. Pawar AB, Jawale MA, William P, Chhabra GS, Rakshe DS, Korde SK, Marriwala N: Implementation of blockchain technology using extended CNN for lung cancer prediction. Measurement: Sensors. 2022, 24:100530. 10.1016/j.measen.2022.100530
15. Makaju S, Prasad PWC, Alsadoon A, Singh AK, Elchouemi A: Lung cancer detection using CT scan images. Procedia Computer Science. 2018, 125:107-114. 10.1016/j.procs.2017.12.016
16. Palani D, Venkatalakshmi K: An IoT based predictive modelling for predicting lung cancer using fuzzy cluster based segmentation and classification. Journal of Medical Systems. 2019, 43:21. 10.1007/s10916-018-1139-7

17. Saleh AY, Chin CK, Penshie V, Al-Absi HRH: Lung cancer medical images classification using hybrid CNN-SVM. International Journal of Advances in Intelligent Informatics. 2021, 7:151-162. 10.26555/ijain.v7i2.317
18. Shanbhag GA, Prabhu KA, Reddy NVS, Rao BA: Prediction of lung cancer using ensemble classifiers. Journal of Physics: Conference Series. 2022, 2161:012007. 10.1088/1742-6596/2161/1/012007
19. Wang W, Chakraborty G: Automatic prognosis of lung cancer using heterogeneous deep learning models for nodule detection and eliciting its morphological features. Applied Intelligence. 2021, 51:2471-2484. 10.1007/s10489-020-01990-z
20. Shakeel PM, Burhanuddin MA, Desa MI: Automatic lung cancer detection from CT image using improved deep neural network and ensemble classifier. Neural Computing and Applications. 2022, 34:9579-9592. 10.1007/s00521-020-04842-6
21. Ali S, Asad S, Asghar Z, Ali A, Kim D: End-to-end 2D convolutional neural network architecture for lung nodule identification and abnormal detection in cloud. Computers, Materials & Continua. 2023, 75:461-475. 10.32604/cmc.2023.035672
22. Lanjewar MG, Panchbhai KG, Charanarur P: Lung cancer detection from CT scans using modified DenseNet with feature selection methods and ML classifiers. Expert Systems with Applications. 2023, 224:119961. 10.1016/j.eswa.2023.119961
23. Al-Tawalbeh J, Alshargawi B, Alquran H, Al-Azzawi W, Mustafa WA, Alkhayyat A: Classification of lung cancer by using machine learning algorithms. 2022 5th International Conference on Engineering Technology and its Applications (IICETA), Al-Najaf, Iraq. 2022, 528-531. 10.1109/IICETA54559.2022.9888332
24. Doppalapudi S, Qiu RG, Badr Y: Lung cancer survival period prediction and understanding: Deep learning approaches. International Journal of Medical Informatics. 2021, 148:104371. 10.1016/j.ijmedinf.2020.104371
25. Nassif AB, Abujabal NA, Omar AA: Classification of lung cancer severity using gene expression data based on deep learning. BMC Medical Informatics and Decision Making. 2025, 25:184. 10.1186/s12911-025-03011-w
26. Dutta B: Comparative analysis of machine learning and deep learning models for lung cancer prediction based on symptomatic and lifestyle features. Applied Sciences. 2025, 15:4507. 10.3390/app15084507
27. Reddy UJ, Reddy BRVR, Reddy BE: Recognition of lung cancer using machine learning mechanisms with fuzzy neural networks. Traitement du Signal. 2019, 36:87-91. 10.18280/ts.360111
28. Humayun M, Sujatha R, Almuayqil SN, Jhanjhi NZ: A transfer learning approach with a convolutional neural network for the classification of lung carcinoma. Healthcare. 2022, 10:1058. 10.3390/healthcare10061058
29. Shatnawi MQ, Abuein Q, Al-Quraan R: Deep learning-based approach to diagnose lung cancer using CT-scan images. Intelligence-Based Medicine. 2025, 11:100188. 10.1016/j.ibmed.2024.100188
30. Prasad U, Chakravarty S, Mahto G: Lung cancer detection and classification using deep neural network based on hybrid metaheuristic algorithm. Soft Computing. 2024, 28:8579-8602. 10.1007/s00500-023-08845-y
31. Mohandass G, Krishnan GH, Selvaraj D, Sridhathan C: Lung cancer classification using optimized attention-based convolutional neural network with DenseNet-201 transfer learning model on CT image. Biomedical Signal Processing and Control. 2024, 95:106330. 10.1016/j.bspc.2024.106330
32. Saha C, Saha S, Rahman MA, et al.: Lung-AttNet: an attention mechanism-based CNN architecture for lung cancer detection with federated learning. IEEE Access. 2025, 13:57369-57386. 10.1109/ACCESS.2025.3554744
33. Kanavati F, Toyokawa G, Momosaki S, et al.: Weakly-supervised learning for lung carcinoma classification using deep learning. Scientific Reports. 2020, 10:9297. 10.1038/s41598-020-66333-x
34. Naveenraj M, Vijayakumar P: Enriched lung cancer classification approach using an optimized hybrid deep learning approach. Scientific Reports. 2025, 15:37573. 10.1038/s41598-025-07322-w
35. Schwyzer M, Ferraro DA, Muehlematter UJ, et al.: Automated detection of lung cancer at ultralow dose PET/CT by deep neural networks - Initial results. Lung Cancer. 2018, 126:170-173. 10.1016/j.lungcan.2018.11.001
36. Zhang Q, Kong X: Design of automatic lung nodule detection system based on multi-scene deep learning framework. IEEE Access. 2020, 8:90380-90389. 10.1109/ACCESS.2020.2993872
37. Sori WJ, Feng J, Godana AW, Liu S, Gelmecha DJ: DFD-Net: Lung cancer detection from denoised CT scan images using deep learning. Frontiers of Computer Science. 2021, 15:152701. 10.1007/s11704-020-9050-z
38. Wahab Sait AR: Lung cancer detection model using deep learning technique. Applied Sciences. 2023, 13:12510. 10.3390/app132212510

39. Polat H, Danaei MH: Classification of pulmonary CT images by using hybrid 3D-deep convolutional neural network architecture. Applied Sciences. 2019, 9:940. 10.3390/app9050940
40. El Hamdi D, Elouedi I, Slim I: Computer-aided classification of cell lung cancer via PET/CT images using convolutional neural network. International Journal of Image and Graphics. 2024, 24:2450040. 10.1142/s0219467824500402
41. Aggarwal R, Kaur K: Machine learning approaches in software fault prediction: a review. Cureus Journal of Computer Science. 2026, 3:es44389-025-00021-7. 10.7759/s44389-026-00021-7

How to cite this article:
Das P, Tripathy S N, Rath K P (March 23, 2026) A Journeying on Machine Learning and Deep Learning Strategies for Lung Carcinoma Forecasting. Cureus J Comput Sci 3 : es44389-026-00044-8. DOI https://doi.org/10.7759/s44389-026-00044-8