International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(11s) 2026 pp. 942-958
https://doi.org/10.70917/ijcisim-2026-3825
Article

Early Lung Cancer Analysis and Prediction Using Machine Learning Classification Techniques

A. Prasannakumari[1], G. Kannan[2]

[1]Department of Computer and Information Science, Annamalai University,
Annamalai Nagar, Chidambaram, Tamil Nadu, India.
Email: abinithivelan@gmail.com
[2]Department of Computer Science, Government Arts and Science College, Manalmedu, (Deputed from Department of CIS, Annamalai University)
Email: gmkanna@gmail.com

Abstract: Lung cancer is one of the leading causes of death worldwide and early detection is important for improving patient survival. Lung cancer develops due to the abnormal growth of cells in the lungs. This study applies machine learning classification algorithms to detect lung cancer accurately and efficiently. Three algorithms, namely K-Nearest Neighbors (KNN), Random Forest (RF), and Support Vector Machine (SVM), are used to classify lung cancer cases. The performance of these models is evaluated using accuracy, sensitivity, specificity, and Area Under the Curve (AUC). The experimental results show that machine learning techniques can support early lung cancer detection and assist healthcare professionals in making reliable diagnostic decisions.
Keywords: Lung Cancer, Machine Learning, K-Nearest Neighbors (KNN), Random Forest (RF), Support Vector Machine (SVM), Classification.

1. INTRODUCTION

Cancer is the second-leading cause of death globally behind heart diseases. In men, lung cancer, prostate cancer, colorectal cancer, stomach cancer, and liver cancer are the most common cancers. The most common cancers among females are bone, colorectal, lung, cervical, and thyroid. The burden of cancer is steadily increasing world over. This has a huge physical, emotional and financial toll on individuals, families, communities and health systems. Health systems in low- and middle-income nations have struggled to cope with this rising burden with many patients facing delays and poor-quality diagnosis and treatment.

Pinpointing cancer helps to diagnose it properly and also ensures that it does not spread to another organ. Detecting cancer genes is essential in preventing cancer. The most popular methods for detecting cancer include:

1. Tissue sample.
2. Examination of body tissues.
3. Radiography methods.
4. CT scan
5. MR Imaging
6. Molecular biology methods.
7. Lung cancer statistics in India.

One of the most common types of cancer diagnosed in India is lung cancer, which has steadily increased in incidence. The country saw over 70000 new instances of lung cancer in 2020. Another major public health problem, which accounted for more than 60,000 deaths in 2020, is also one of the leading causes of cancer deaths.

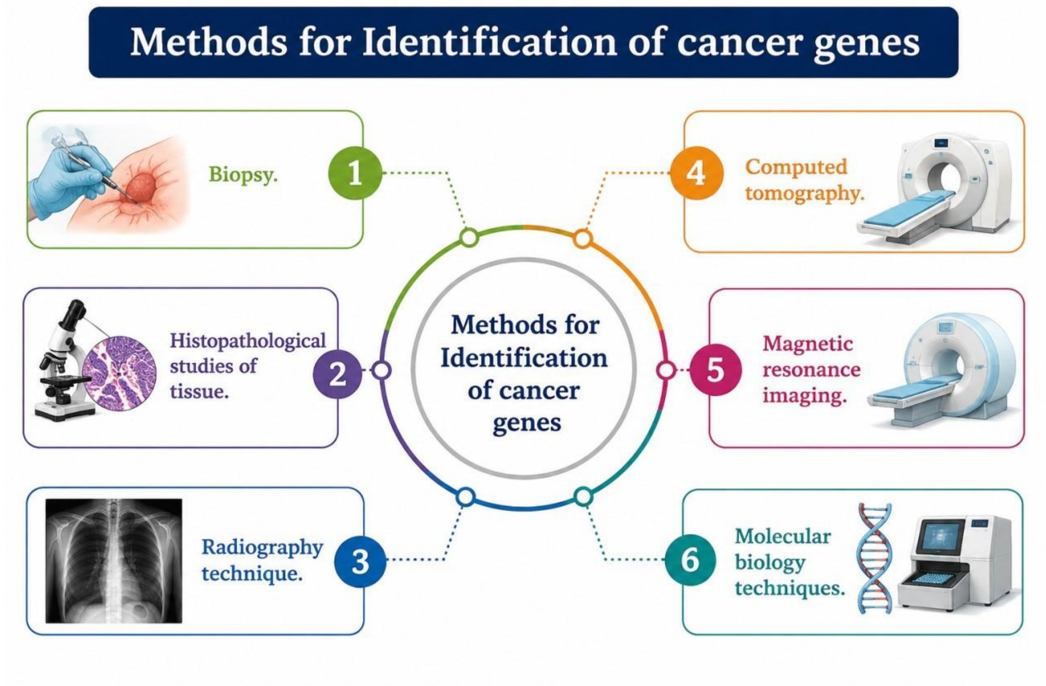

*Figure 1.*Various factors associated with the development of lung cancer.

Lung cancer is one of the most serious health challenges in the world and India, with a high incidence and death rate. Tobacco control, public health campaigns and improving air quality are some of the key strategies to reduce it. Enhancing awareness and early detection can enhance the health of more at-risk people. The aforementioned elements may be represented by Figure 1. Depending on the risk factors and previous exposure of each individual, these can differ. Reliable medical sources and research studies should be consulted for complete information.

Lung cancer leads the ongoing battle as it dramatically increases every year. It is the result of the uncontrolled and rapid multiplication of cells. It is one of the most common and fatal cancers globally, contributing to a large number of deaths from cancer.

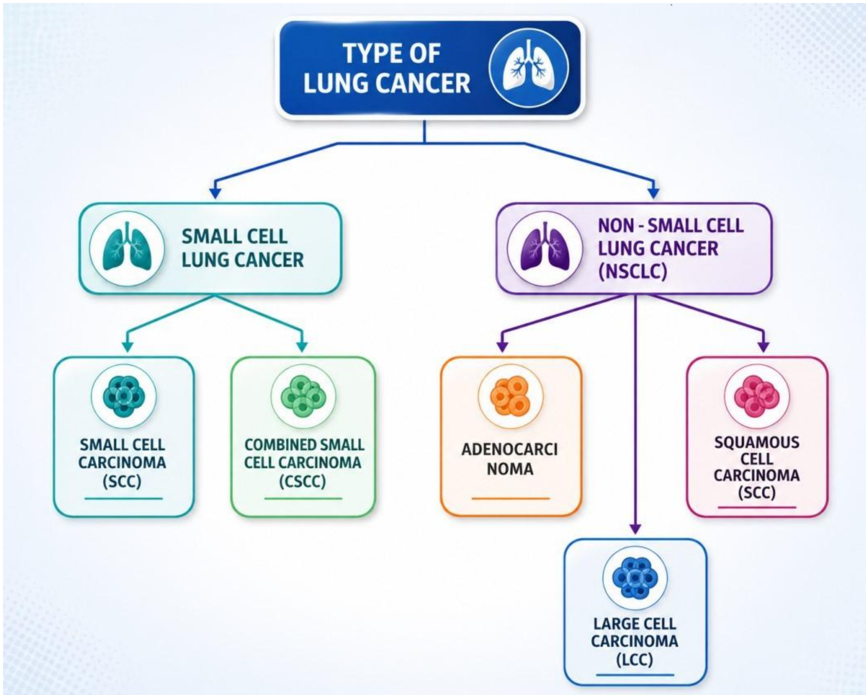

Figure 2: Causes of Lung cancer

This complete introduction will provide you with the full details regarding lung cancer, types, and stages. Lung cancers can be divided into two main groups depending on their origins, growth rate and treatments.

- Small Cell Lung Cancer (SCLC)
- **Non-Small Cell Lung Cancer (NSCLC)**

Table: 1 Differences between SCLC and NSCLC

Features	SCLC	NSCLC
Incidence	Decreasing	Increasing
Association with Smoking	Universal	Highly Variable
Growth Kinetics	Rapid	Variable
Biologic Diversity (Historic & Molecular)	Homogeneous	Distinct Sub types
Early Metastases	Universal	Variable
Sensitivity to DNA-damaging Chemotherapy	High	Variable
Sensitivity to Radiotherapy	High	Variable

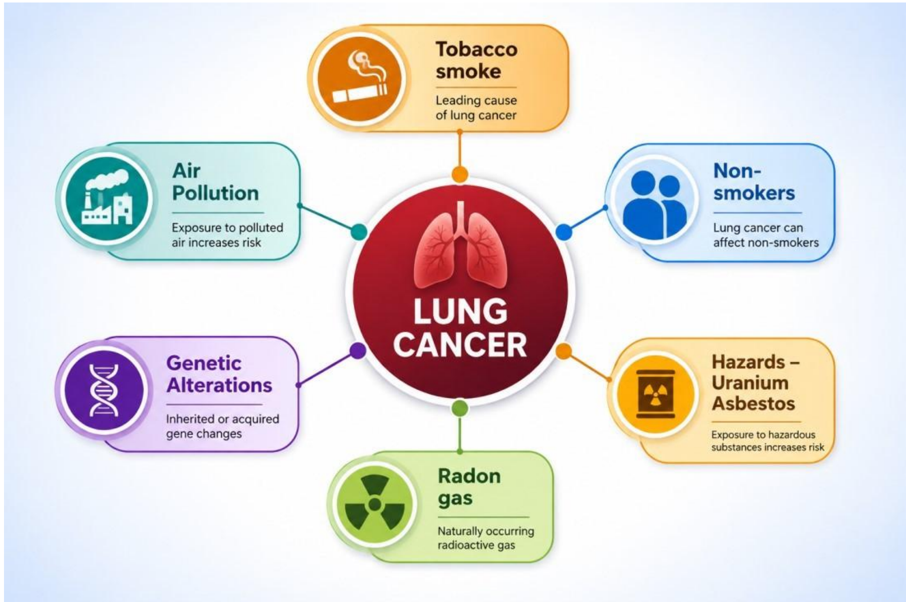

*Figure 3 P*rimary lung cancer types: NSCLC and SCLC.

Lung cancer cases NSCLC are responsible for and they have three major types.

- Adenocarcinoma forms in the outer parts of your, it is the most common type of.
- Lungs.
- Squamous cell carcinoma tends to arise in bronchial tube linings and is highly associated with smoking.
- Large cell carcinoma is rare, but aggressive and spreads to other tissues at earlier stages.

Every type behaves differently and responds differently to treatment. Thus, accurate classification is important for management.

Lung cancer is one of the most common cancers in the world, thus having an early diagnosis is crucial for lung cancer survival. The development of computer-aided diagnosis (CAD) systems for predicting and classifying lung cancer has transformed dramatically in recent years thanks to machine learning (ML) and deep learning (DL), radiomics, and explainable artificial intelligence (XAI).

Research carried out between 2024 - 2026 focuses on the hybrid multimodal clinically interpretable AI modelling rather than traditional ML algorithms. Notable investigations comprise.

Maurya et al. (2024): Maurya et al. compared lung cancer prediction using clinical symptoms, behavioral risk factors, and twelve ML algorithms. K-Nearest Neighbor (KNN) and Bernoulli Naïve Bayes were found to provide superior performance and combination of the demographic feature with the symptom-based feature enhances accuracy.

Cai et al. (2024):) reviewed the AI development in pulmonary nodule detection and lung cancer diagnosis. According to them, CNNs outperform standard ML techniques by automatically learning from the data in the analyses of CT images. The increasing use of GANs, ensemble learning and transfer learning to facilitate early detection was also highlighted in the assessment.

Patel et al. (2024): focused on deep learning with a conclusion stating that CNN-based architectures are the key focus of recent research as it overcomes hand-crafted feature extraction and provide a high accuracy on benchmark datasets. According to them, training time and computational resources are reduced with transfer learning, making it clinically feasible.

Research in 2025 will contribute causes with more robust, explainable, and applicable to clinical AI models for lung cancer detection.

Meeradevi et al. (2025) came up with an image-based classification framework that utilized machine learning and deep learning for chest X-ray analysis. The model classified benign and malignant lung anomalies efficiently and accurately classified pulmonary diseases such as atelectasis, infiltration, nodules and pneumonia. As observed in this study, automated image analysis significantly reduces the workload of radiologists while maintaining high accuracy.

Shatnawi et al. (2025) proposed a robust deep learning framework by developing enhanced CT images in conjunction with state-of-the-art CNN architectures to diagnose lung cancer at an early stage. Their method includes image preprocessing, feature enhancement via ConvNeXt classification, which improved prediction accuracy and robustness to image noise. According to the authors, the performance of the model is dependent on the image enhancement techniques.

The progress achieved here represents a significant advance towards achieving clinically accurate and reliable AI platforms for lung cancer diagnosis.

A number of significant studies in 2025 enhanced the comprehension and application of machine learning (ML) techniques for lung cancer diagnosis.

Farshchiha et al. (2025) study a wide comparative study between ML algorithms such as Logistic Regression, Decision Tree, Random Forest, Catboost, Lightgbm, Xgboost, Adaboost and Deep Neural Network. The results showed that XGBoost, LightGBM and Logistic Regression gave the highest performances in overall with low time-cost and minimal overfitting. Despite the rapid advancement of deep learning techniques, the older ensemble ML algorithms are still quite competitive.

Shahriyar et al. (2025) proposes combinations of feature selection methods and classifiers. Chi-square feature selection in conjunction with Random Forest and Support Vector Machine (SVM) improves the classification accuracy as well as minimises the computational complexity. It demonstrates the importance of selecting informative clinical variables before training.

Lamba et al. (2025) applied synthetic minority over-sampling technique (smote) to the class imbalance which was identified code optimization predictive framework. In their experimental study, they found that balancing the datasets prior to classification substantially improves sensitivity and reduces false negative predictions which are a key medical diagnosis.

A comprehensive review by **Arshad et al. (2025)** on latest developments of AI in diagnosing lung cancer concluded that their clinical use in the future depends on multicentre validation, standardised imaging protocols, explainable AI, and integration into clinical workflows. According to them, combining imaging biomarkers, electronic health records and genomic data can help develop personalized diagnostic system.

These studies show how the field of AI methods for the diagnosis of lung cancer is evolving towards robustness, interpretability, and clinical relevance. Research on multimodal intelligence and clinical translation focusses more on the AI system for lung cancer diagnosis by 2026.

Abumohsen et al. (2026) Abumohsen et al. (2026) liquidated an review machine learning and deep learning model, in which discusses several barriers for clinical use, such as limited external validation, limited explainability, heterogeneous training set, regulatory issues, and approval and adaptation. Develop standardized and explainable AI that is multicenter validated for routine clinical use.

Zahid et al. (2026) presented a survey was performed comparing the ML, DL and hybrid models for lung cancer prediction. The findings showed that hybrid approaches using radiomics, genomics, pathology, and clinical data outperform single-modality approaches. According to them, explainable AI and multimodal learning will shape the next generation of intelligent clinical decision-support systems.

Shweikeh et al. (2026) presents a dual-branch deep learning model integrating a convolutional neural network and a deep neural network for lung cancer classification.

- The steps that are involved in the lung cancer detection and classification (LCDC) methodology.
- Lung cancer diagnosis and classification system using machine learning.
- Lung CT image data set
- Image input and reporting output of Lung Cancer classification results.
- Image pre-processing is step two.

- If the image is noise free then goes to step 3 else go to step 2.2
- Proceed to the next step to improve the data in preparing the data of this stage.
- Step Three: Extraction of Features.
- In step 3.1, we extract the features such as Area, Perimeter, Centroid, Gray Level Co-occurrence Matrix (GLCM), Local Binary Patterns (LBP).
- Step 3.2: Extract statistical features such as Mean, Standard Deviation and Smoothness.

In the fourth stage, begin the process of developing a classification algorithm to train and forecast the Tumour to be benign or malignant.

The classifications algorithm can be applied for training and prediction of all classes of lung cancer.

In the last step, the performance of the model is evaluated using the evaluation parameters.

In step 7 repeat steps 1 to 6 for all images in the dataset.

No more algorithm

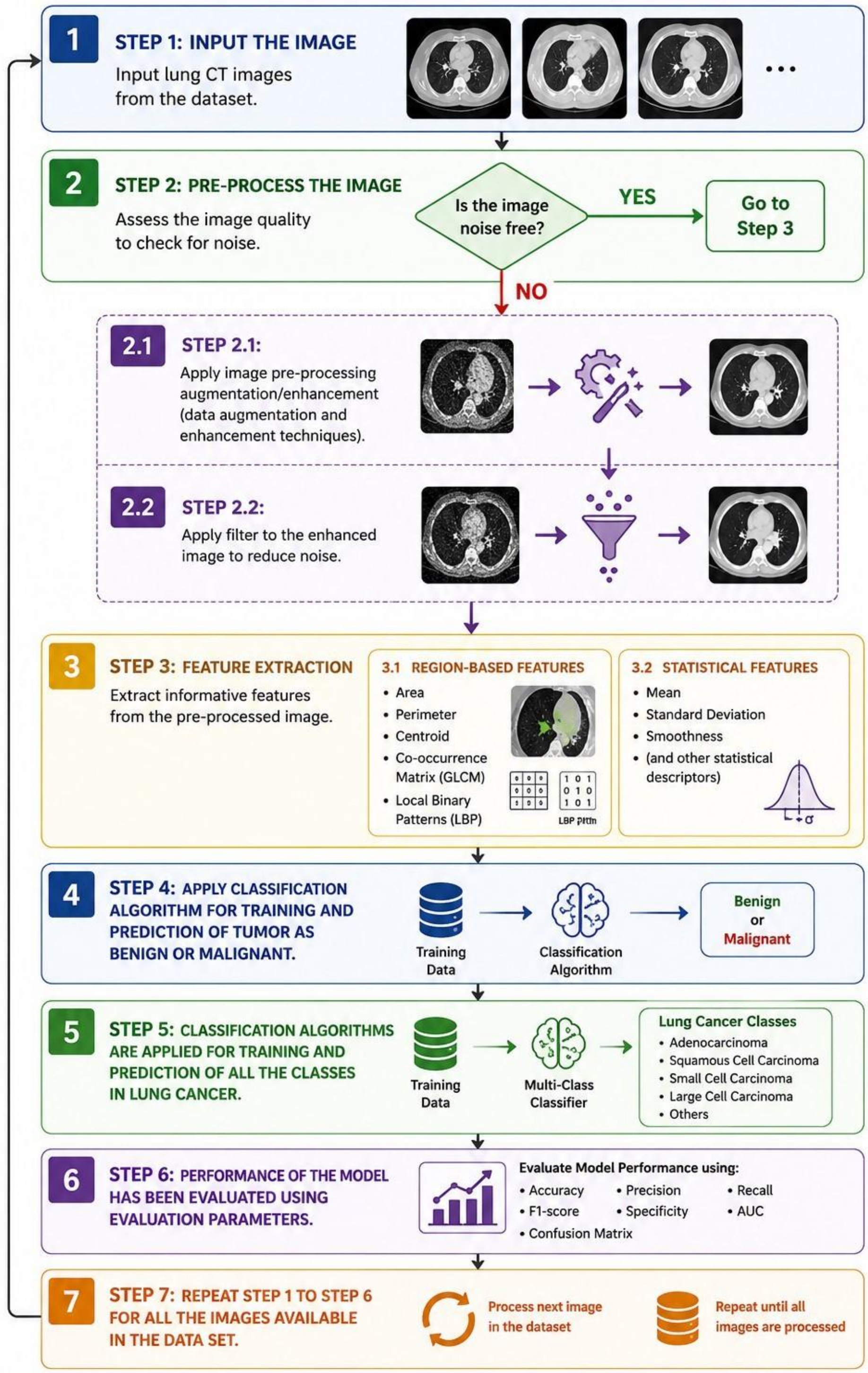

Figure 4 Steps for Lung Cancer Detection and Classification

Data Collection

- To enhance the performance of forecasting models, we curated a dataset with raw data sourced from various public places in Asia.

• The performance of ML needs datasets sourced from trusted and credible sources. Ideally, the datasets must come outfitted with a comparatively high number of lung CT images.

• Public and custom datasets.

Many datasets are available in the public domain to detect Lung diseases.

Some researchers also create custom datasets for more specificity and accuracy [86].

• Illustration of the Sample.

As shown by Figure 3.2, representative lung CT scan images demonstrate the type and quality of data used in the experimental analysis [87].

• Used Dataset:

The proposed model was trained and tested using the LIDC-IDRI lung CT scan dataset.

• Selection of Samples

A particular set of samples was chosen from the LIDC-IDRI dataset to approve the findings of the study.

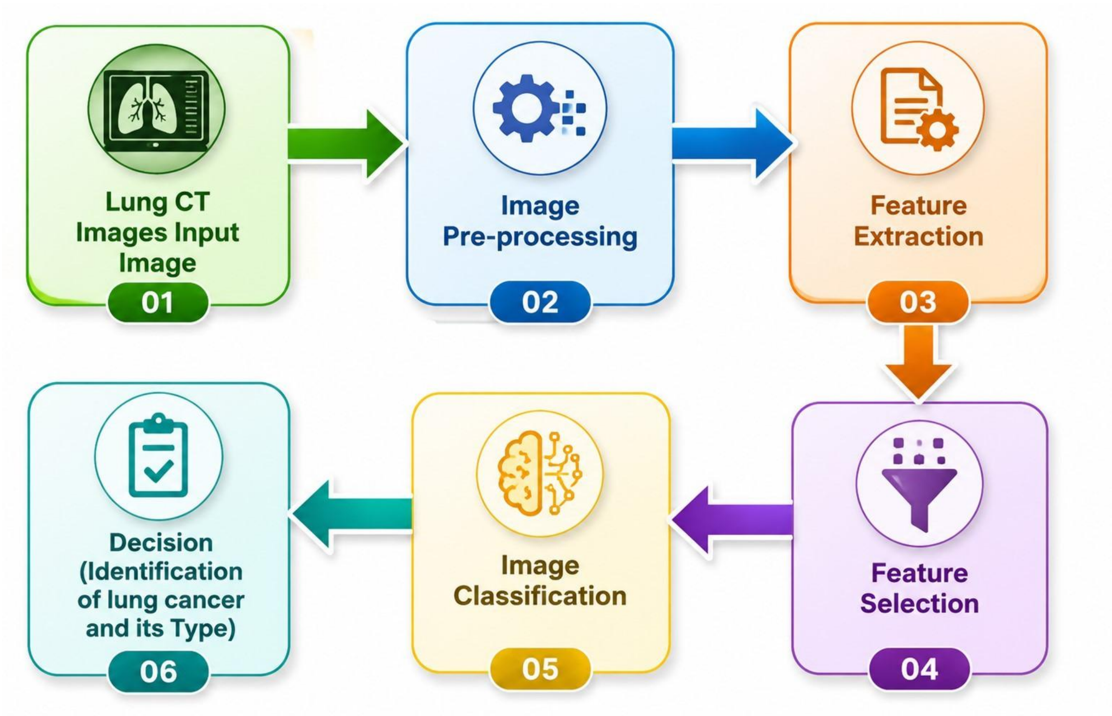

Figure 5 Steps for Lung Cancer Identification

Lung Cancer Patient Dataset: Features Overview

This data set provides information that will help to understand different risk factors that lead to lung cancer disease. Every patient record has the following attributes.

• Code assigned to each patient for identification

• Patient's numeric age (5 words)

• Patient's Gender (Categorical).

• The level of exposure to air pollution (categorical).

• Frequency or amount of alcohol consumption (categorical) you did

• Whether a person has a dust allergy or not.

- Occupational Hazards: exposure to workplace environmental risks (categorical).
- Genetic predisposition; family history (categorical).
- Longstanding lung illnesses: pre-existing chronic lung diseases (categorical).
- Nutrition level depending on the legal status to offer the curated distinct balanced diet.
- The degree of overweight status.
- Current smoking behaviour (categorical).
- Passive Smoker status: Exposure to passive smoke (categorical).
- The degree and intensity of chest discomfort or pain.
- Coughing of Blood: Categorical hemoptysis incidence.
- The category of fatigue one experiences.
- Degree of unintentional loss of weight (categorical)
- Presence and severity of shortness of breath.
- Frequency of wheezing episodes (categorical).
- Trouble swallowing (categorical).
- The nails on the fingers were clubbed (device).
- Often Falling Sick: Has a history of frequent (categorical) getting sick.
- Dry cough occurred with what frequency?
- Presence And Severity Of Snoring

Purpose:

- Reducing lung cancer mortality could be tremendous, and we will begin by assessing the main risk factor.
- Data cleaning.
- Integrity of Data.
- Identify and amend outliers, discrepancies, and missing data.
- Use high level input data. Do thorough R.D cleaning.
- Scaling Dimensions
- Implement normalization or standardization to make the scales of features consistent
- Preparation of Images:
- Ensure uniformity in image size, format and quality.
- Make use of noise reduction and image enhancement techniques, such as.
- A method of adjusting brightness of the image.
- Wiener noise adjustment.
- Adjustable Gaussian Filtering
- Gaussian and Gabor filtering.
- It is essential to carry out the image normalization, size/format correction, and noise reduction for the dataset to be used in machine learning.
- Key sub-processes
- Smoothing of Images.

- Apply a median filter to remove salt and pepper noise while retaining edges 信息[88]
- improving image quality
- Utilize contrast modification to enhance image quality by saturating approximately 1% of information at low & high intensity level.

Feature Extraction

This proposal seeks to convert images into numerical features for downstream applications and classification.

•Method for Feature Extraction from Segmented CT Scan Regions of Interest (ROIs)

•Key Attribute Features

Properties Related to Geometry

Region.

Pinnacle.

Weirdness, Abnormality.

These characteristics assist in evaluating the size and shape of tumor [89].

Distribution

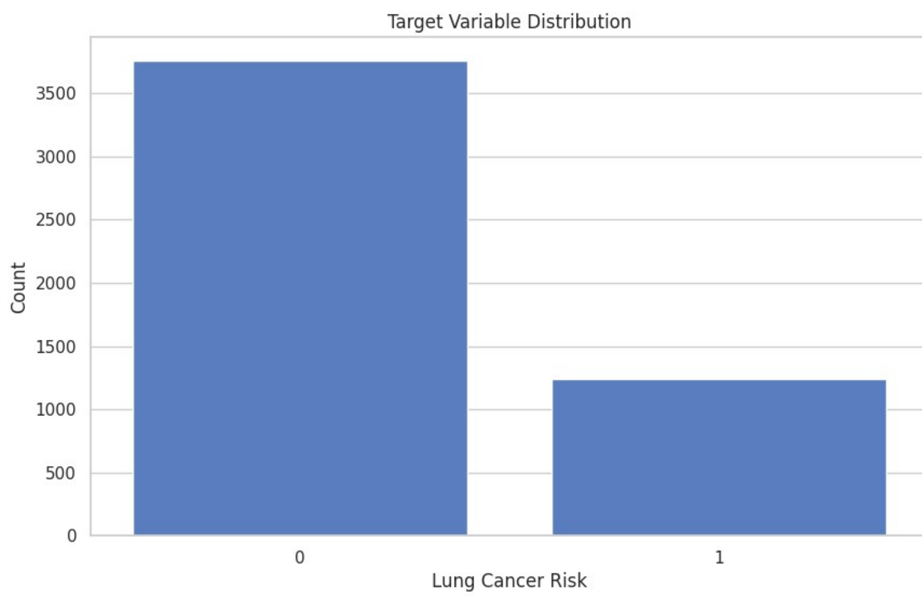

Figure 6 Distributions

Feature Distribution

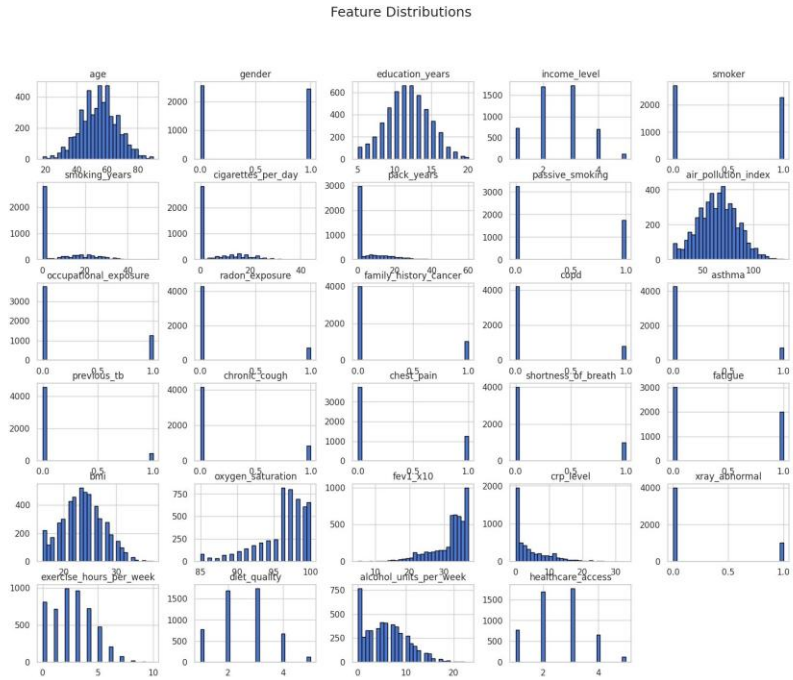

Figure .7 Feature Distribution

Heatmap

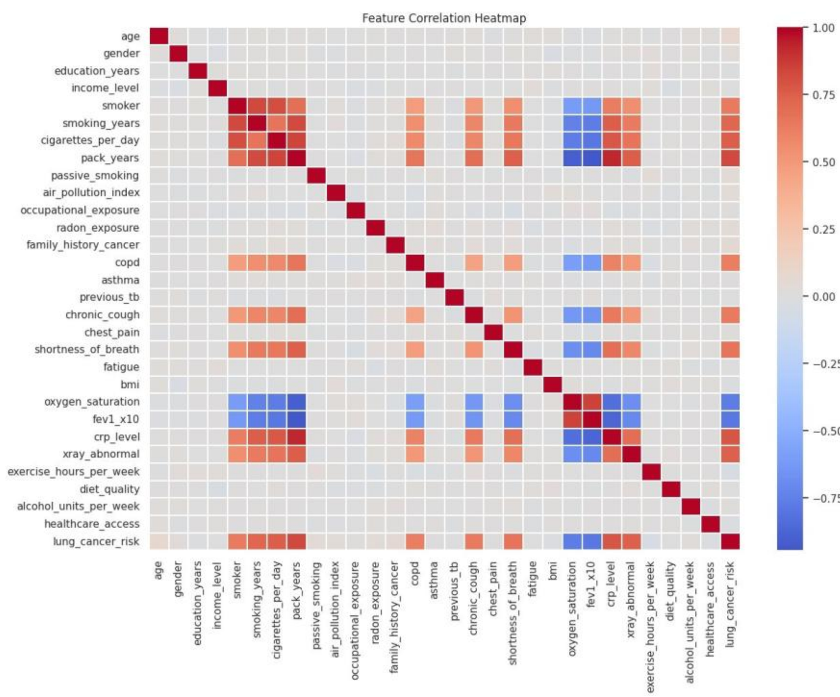

Figure .8 Heat maps

Feature Importance

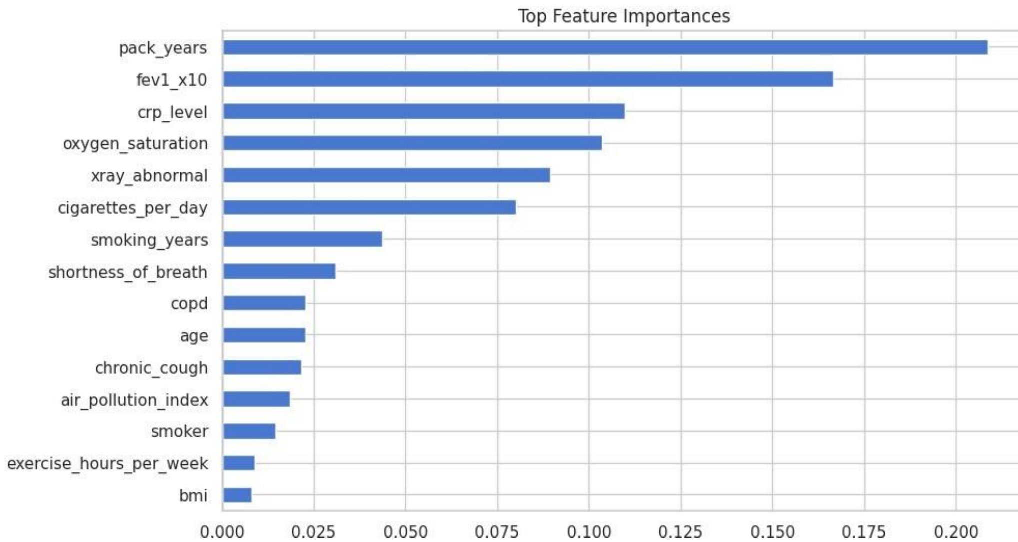

Figure 9 Feature Importance

The study focuses on the potential use of machine learning algorithms for the detection of lung cancer. The accuracy as well as efficiency of lung cancer detection processes has been greatly improved with the help of K-Nearest Neighbors (KNN), Random Forest (RF), Support Vector Machines (SVM) and others. After investigation, it was concluded that the algorithms have performed quite well with respect to sensitivity, specificity, and other overall performance measure so as to classify the tissue as cancerous or non-cancerous.

Table 2: Performance analysis of Lung Cancer using SVM classifier

Fold\Class	**TPR**	**FPR**	**Precision**	**Recall**	**F1-Score**	**Accuracy**
1	0.99	0.39	0.99	0.99	0.99	98.9
2	0.99	0.43	0.91	0.99	0.99	91.2
3	0.99	0.05	0.98	0.98	0.99	97.3
4	0.97	0.11	0.95	0.96	0.95	93.7
5	0.99	0.30	0.93	0.99	0.97	94.1
Average	0.99	0.26	0.95	0.99	0.98	95.1

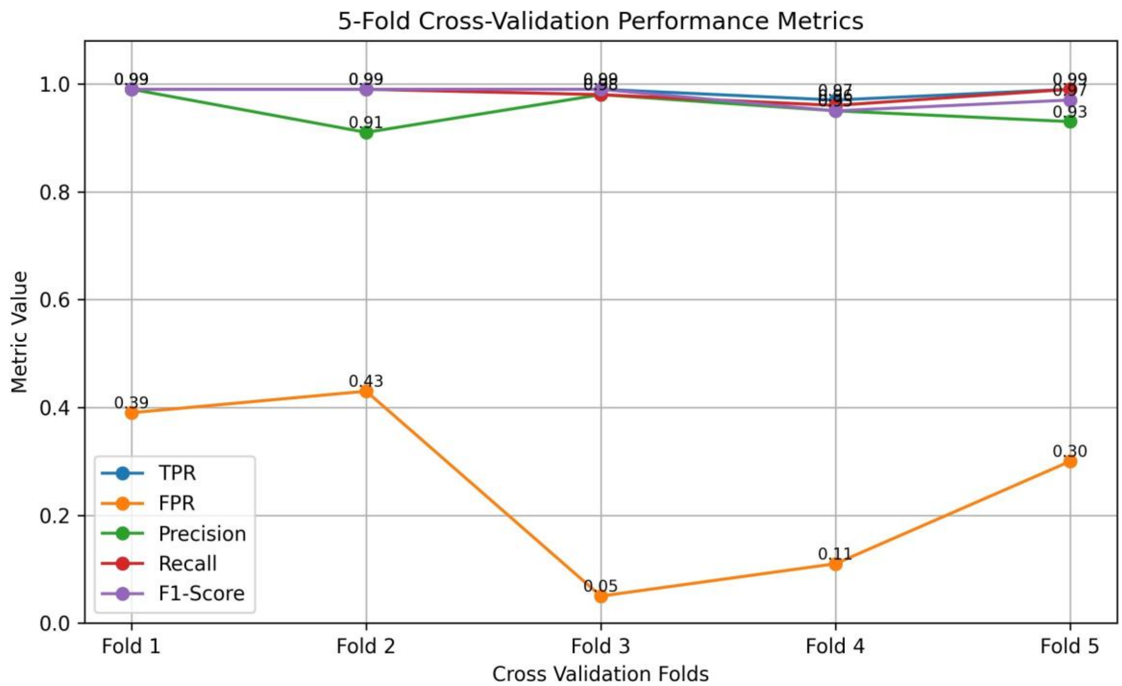

Figure 10 Accuracy

Table 6.1 presents the findings of the lung cancer prediction performance assessment utilizing an SVM classifier. A five-fold cross will give a realistic estimate of the model.

Process for Validation was used. The system attains 95.06% accuracy for predicting an overall a.

Specific type of lung cancer. This study also examined other important relevant indicators.

Overall functioning. Metrics entail precision, recall, AUC (Area Under Curve) and F1.

rating. The same 5-fold was utilized to calculate and tabulate each metrics per individual classes.

Merged results were analysed and compared with cross-validation technique.

Table 3: Performance analysis of lung cancer using KNN classifier

Fold\Class	**TPR**	**FPR**	**Precision**	**Recall**	**F1-Score**	**Accuracy**
1	0.99	0.43	0.89	0.99	0.94	89.9
2	0.99	0.33	0.82	0.99	0.89	86.0
3	0.99	0.30	0.82	0.99	0.90	86.6
4	0.96	0.35	0.81	0.96	0.88	83.8
5	0.99	0.49	0.87	0.99	0.93	88.1
Average	0.98	0.38	0.84	0.98	0.91	86.89

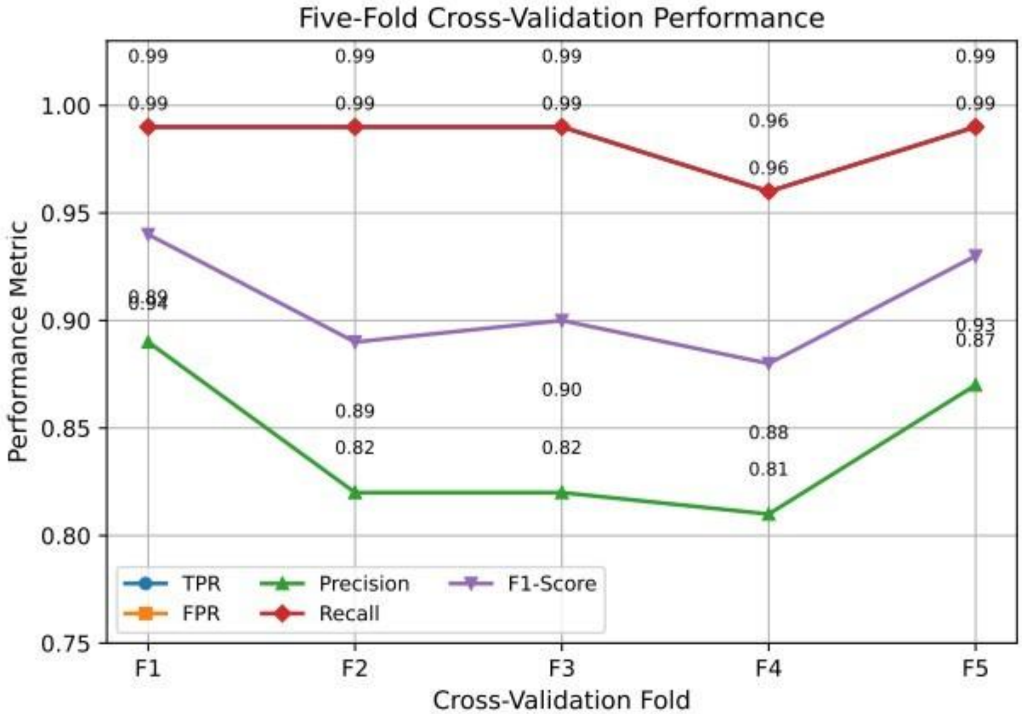

Figure 11 Accuracy

The accuracy of KNN classifier for lung cancer prediction has been found to be 86.89% as shown in table 6.2 which is the performance evaluation of KNN classifier. The performance analysis of Lung Cancer using Random Forest classifier is shown in table 6.3 which has a 81.96% accuracy. The classifiers also classify the various sub-types of lung cancer

In the classification of lung cancer, SVM demonstrates superior performance compared to KNN and Random Forest. This is attributed to SVM's ability to cope well with the data being studied using more features owing to its high dimensionality. This is further exemplified in figure 6.2. The confusion matrix that is shown indicates how with lung cancer data, SVM manages to identify complex patterns well. Thus, SVM performs with more accuracy and reliability in classifying data than the KNN or Random Forest algorithms.

Table 4: Performance analysis of lung Cancer using Random Forest classifier

Fold\Class	TPR	FPR	Precision	Recall	F1-Score	Accuracy
1	0.99	0.57	0.76	0.99	0.86	78.8
2	0.99	0.50	0.82	0.99	0.89	83.9
3	0.98	0.37	0.77	0.98	0.86	82.6
4	0.96	0.40	0.77	0.96	0.85	80.8
5	1.00	0.58	0.82	1.00	0.90	83.8
Average	0.98	0.48	0.78	0.98	0.87	82.0

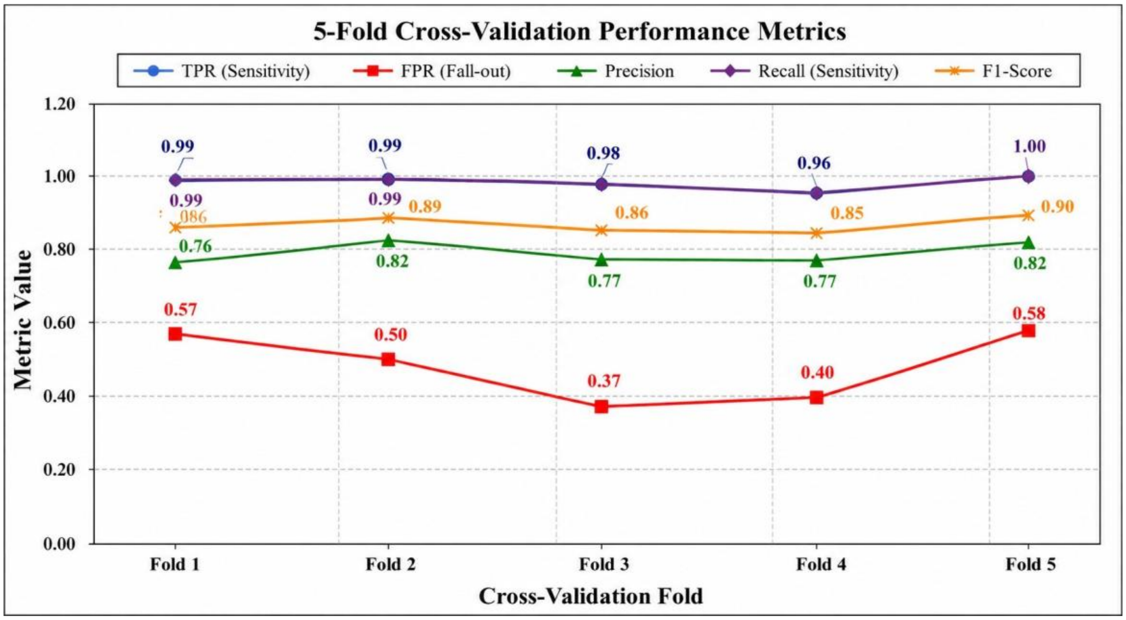

Figure 12 Accuracy

Results indicated that SVM achieved the highest accuracy, reaching an impressive 95.05%. Proposed work has been carried out for different combinations as in table 6.4 below

Table 6.4 : Comparison of proposed ML algorithm with state of art methods

Table: 5 Differences between SCLC and NSCLC

Sl. No	Authors	Accuracy in %
1	Patra	81.25
2	Gultepe	83
3	Maurya	91.07
4	Faisal	90
5	Ansari	94.45
6	Shanbhag	88
7	Yakar	70
8	Proposed LCDC	95.05

These accuracy rates demonstrate the impact of diverse combinations of machine learning algorithms on lung cancer classification. In the realm of lung cancer detection as well as classification capability machine learning algorithms, Patra et al. obtained an accuracy of 81.25%, Gultepe et al. achieved 83% and Maurya et al. reported 91.07% accuracy. Faisal et al. achieved an accuracy rate of 90%. Meanwhile, Ansari et.al achieved an accuracy of 94.45%. Afterwards, Shanbhag et.al obtained an additional accuracy. Yarkar et.al achieved an accuracy of 70% while obtained an accuracy of 88%. Conversely, the proposed model attained an astonishing 95.05 percent accuracy.

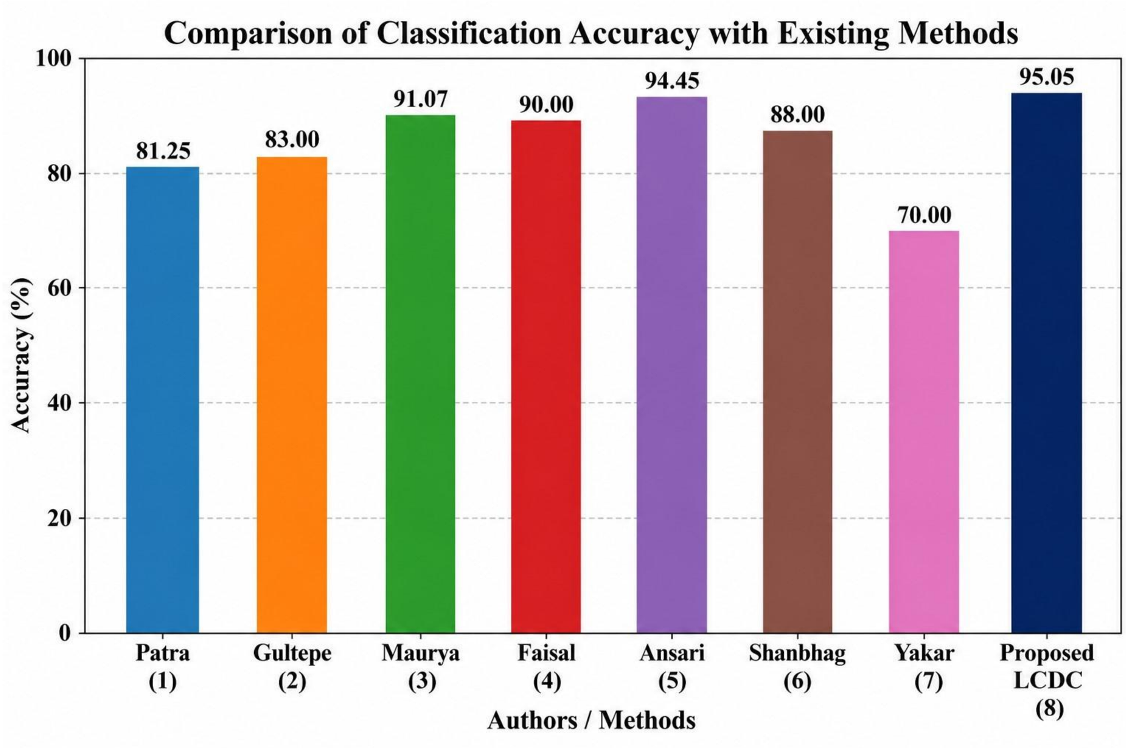

Figure 13 Accuracy

This comparison underscores the effectiveness and superiority of our proposed approach in accurately detecting and classifying lung cancer compared to previous studies.

2. CONCLUSION

Along with this different validation for cancer detection and classification ensures each of the aspect performance of the model is validated and optimized separately for reliable lung cancer diagnosis. These results suggest that advanced computer techniques could be an effective accompaniment to existing diagnostic tools, supporting a more accurate detection of lung cancer in a clinical setting.

3. FUTURE RESEARCH

Future research can focus on improving the accuracy and reliability of lung cancer detection by incorporating advanced machine learning and deep learning algorithms with larger and more diverse datasets. Additional validation techniques, including cross-validation, external validation, and real-world clinical testing, can be employed to ensure that every aspect of the prediction model is thoroughly evaluated and optimized for accurate lung cancer diagnosis. Furthermore, future studies can integrate multimodal medical data such as CT images, histopathological images, patient clinical records, and genomic information to improve diagnostic performance. The proposed system can also be developed into a real-time clinical decision-support tool and extended to detect different stages and types of lung cancer, enabling healthcare professionals to make faster, more accurate, and reliable diagnostic decisions.

References

1. Ansari, M.M.; Kumar, S.; Heyat, B.B.; Ullah, H.; Hayat, M.A.B.; Parveen, S.S.; Ali, A.; Zhang, T. SVMVGGNet-16: A Novel Machine and Deep Learning Based Approaches for Lung Cancer Detection using Combined SVM and VGGNet-16. Curr. Med. Imaging 2025, 21, 1–18.
2. Daoud, Maisa, and Michael Mayo. "A survey of neural network-based cancer prediction models from microarray data." Artificial intelligence in medicine (2019).
3. Faisal, M.I.; Bashir, S.; Khan, Z.S.; Khan, F.H. An evaluation of machine learning classifiers and ensembles for early stage prediction of lung cancer. In Proceedings of the 2018 3rd International Conference on Emerging Trends in Engineering, Sciences and Technology (ICEEST), Thrissur, Kerala, India, 18–20 January 2018; pp. 1–4.

4. Fenwa, Olusayo D., Funmilola A. Ajala, and A. Adigun. "Classification of cancer of the lungs using SVM and ANN." Int. J. Comput. Technol. 15.1 (2016): 6418-6426.
5. Gültepe, Y. Performance of Lung Cancer Prediction Methods Using Different Classification Algorithms. Comput. Mater. Contin. 2020, 67, 2015–2028.
6. Jin, Xin-Yu, Yu-Chen Zhang, and Qi-Liang Jin. "Pulmonary nodule detection based on CT images using convolution neural network." 2016 9th International symposium on computational intelligence and design (ISCID). Vol. 1. IEEE, 2016.
7. Lynch, Chip M., et al. "Prediction of lung cancer patient survival via supervised machine learning classification techniques." International journal of medical informatics 108 (2017): 1-8.
8. Maurya, S.P.; Sisodia, P.S.; Mishra, R.; Singh, D.P. Performance of machine learning algorithms for lung cancer prediction: A comparative approach. Sci. Rep. 2024, 14, 18562.
9. Öztürk, Şaban, and Bayram Akdemir. "Application of feature extraction and classification methods for histopathological image using GLCM, LBP, LBGLCM, GLRLM and SFTA." Procedia computer science 132 (2018): 40-46.
10. Palani, D., and K. Venkatalakshmi. "An IoT based predictive modelling for predicting lung cancer using fuzzy cluster based segmentation and classification." Journal of medical systems
11. 43.2 (2019): 21.
12. Patra, R. Prediction of lung cancer using machine learning classifier. In Proceedings of the International Conference on Computing Science, Communication and Security, Gujarat, India, 26–27 March 2020; Springer: Berlin/Heidelberg, Germany; pp. 132–142.
13. Shanbhag, G.A.; Prabhu, K.A.; Reddy, N.V.S.; Rao, B.A. Prediction of Lung Cancer using Ensemble Classifiers. J. Phys. Conf. Ser. 2022, 2161, 012007.
14. Sumathipala, Yohan, et al. "Machine learning to predict lung nodule biopsy method using CT image features: A pilot study." Computerized Medical Imaging and Graphics 71 (2019): 1-8.
15. Yakar, M.; Etiz, D.; Yilmaz, S.; Celik, O.; Guntulu, A.K.; Metintas, M. Prediction of survival and progression-free survival using machine learning in stage III lung cancer: A pilot study. Turk. Oncol. Derg. 2021, 36, 446–458.