Early Analysis and Prediction of Lung Cancer using Machine Learning Classifiers

A.Prasannakumari
Research Scholar
Department of Computer and Information Science,
Annamalai University , Annamalai Nagar, Chidambaram,
Tamil Nadu – INDIA

Dr.G.Kannan
Assistant Professor
Department of Computer Science
Government Arts and Science College
Department of CIS , Annamalai University, Chidambaram
Tamil Nadu - INDIA

Abstract

Machine learning (ML) models have been employed to analyze clinical data, medical imaging, and patient demographics, demonstrating promising accuracy rates. ML classification models is a rapidly evolving field that aims to improve diagnostic accuracy and patient outcomes. This proposed research presents the novel ML Models for early prediction and analysis of lung cancer.

Keywords: Machine Learning, data,medical imaging, patient ,prediction

I. INTRODUCTION

A malignant tumor in one or both lungs' tissue is known as lung cancer. Either the spongy lung tissue or the bronchi may contain a tumor. Primary lung cancer is a tumor that originates in the lung. Cancer that has spread through the blood from another part of the body, like the breast, colon, or prostate, can also cause lung tumors; these cancers are referred to as lung "secondary" or "metastases." The information that follows relates to primary lung cancer.

Similar to other types of cancer, lung cancer arises from the unchecked proliferation and multiplication of lung cells. Breathing becomes difficult, resulting in pain and symptoms associated with the loss of normal lung function, as this aberrant cell proliferation gradually grows into an ever larger mass that begins to invade functional areas of the lung. This aberrant cell cluster is referred described as a "tumor" by doctors. Unchecked growth and division of these aberrant cells can lead to their eventual spread throughout the body if treatment is not received.

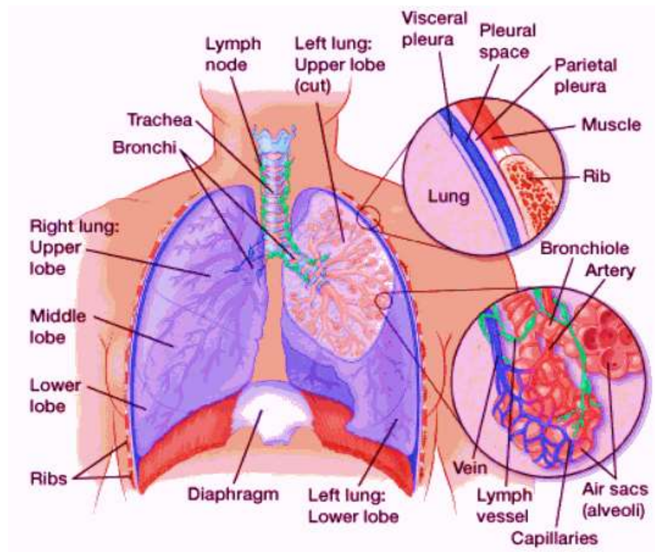

Fig 1. Lung Cancer

Non-small cell lung cancer (NSCLC)

The most prevalent kind of lung cancer is called non-small cell lung cancer (NSCLC). More than 80 percent of instances of lung cancer are caused by it. Squamous cell cancer and adenocarcinoma are frequent varieties. Two less frequent forms of NSCLC are sarcomatoid carcinoma and adenosquamous carcinoma.

Small cell lung cancer (SCLC)

NSCLC is easier to cure than small cell lung cancer (SCLC), which grows more quickly. A relatively modest lung tumor that has already migrated to other parts of your body is how it is typically discovered. Combination small cell carcinoma and small cell carcinoma, commonly known as oat cell carcinoma, are two distinct forms of SCLC.

II. LITERATURE REVIEW

Key performance metrics like accuracy, precision, and recall are essential for assessing model effectiveness(Basha et al., 2024). Hyperparameter tuning has been shown to improve model performance, exemplified by an increase in Logistic Regression accuracy from 84% to 85%(Ruqiya et al., 2024). **Accuracy, Precision, and Recall**: These metrics are crucial for assessing model effectiveness. For instance, SVC and RF models excelled in these areas, indicating their reliability in clinical settings(Zapata-Paulini & Cabanillas-Carbonell, 2024)(Basha et al., 2024).

Hyperparameter Tuning: This process significantly enhances model performance, as seen with LR(Ruqiya et al., 2024).

While machine learning classifiers show great promise in lung cancer prediction, challenges remain in dataset diversity and model generalization. Future research should focus on integrating more comprehensive datasets and exploring advanced algorithms to further enhance predictive accuracy.

Types of ML Models for Lung Cancer Prediction

Supervised Learning Techniques

Several supervised learning techniques have been applied to classify lung cancer patients, focusing on survival prediction. Techniques such as Gradient Boosting Machines (GBM), Support Vector Machines (SVM), and custom ensembles have been used, with GBM often showing superior performance in terms of predictive accuracy1. Decision Trees, however, may be less applicable due to limited discrete outputs1.

Deep Learning Models

Deep learning models, particularly Convolutional Neural Networks (CNNs), have been effectively used for classifying lung cancer from histopathology images and CT scans. These models can distinguish between subtypes like adenocarcinoma and squamous cell carcinoma with high accuracy, comparable to that of pathologists2 5. Deep Neural Networks (DNNs) and ensemble methods have also shown robust performance in classifying lung cancer levels3.

Data Types and Feature Selection

Imaging Data

CT scans and histopathology images are commonly used in ML models for lung cancer classification. Image processing techniques, such as feature extraction and segmentation, are crucial for enhancing the quality and interpretability of these images before classification4 9. Models like ML-xResNet and EOSA-CNN have been developed to improve classification accuracy by optimizing feature extraction and model parameters8 9.

Genomic and Multi-Omics Data

Integrating multi-omics data, including mRNA, miRNA, and DNA methylation, with deep learning models has shown promise in predicting lung cancer stages and subtypes. This approach allows for a comprehensive analysis of genetic alterations associated with lung cancer, leading to improved predictive performance6.

Key ML Models for Lung Cancer Prediction

1. **Ensemble Learning Models:**
2. **Random Forest (RF):** Frequently identified as a top performer for lung cancer prediction tasks, including EGFR mutation prediction and venous thromboembolism (VTE) risk, with high AUC values indicating strong predictive power2 3 8.
3. **Gradient Boosting Machines (GBM):** Noted for its accuracy in survival prediction, often used within custom ensembles to enhance performance1.
4. **Support Vector Machines (SVM):**
5. While SVMs have been used, they often underperform compared to ensemble methods like GBM and RF, though they can provide distinctive outputs in certain contexts1.
6. **Logistic Regression (LR):**
7. Demonstrated superior performance in predicting post-chemotherapy lung infections, with high accuracy and AUC values, indicating its utility in specific clinical scenarios4.
8. **Graph Convolutional Networks (GCN):**

   Used for survival analysis in early-stage lung cancer, outperforming traditional models by leveraging imaging data, which provides a robust prediction of patient outcomes9.
9. **Artificial Neural Networks (ANN):**

   Effective in combining conventional indicators with tumor markers for early lung cancer diagnosis, showing high AUC values and clinical significance7.

Comparative Insights

- **Performance Metrics:** Ensemble models like RF and GBM generally outperform other models in terms of AUC and accuracy across various prediction tasks1 2 3 8.
- **Fairness and Bias Mitigation:** Some studies focus on ensuring fairness across racial and demographic groups, with models like LungFlag demonstrating equitable performance across different subpopulations5 6.

- **Inter-Institutional Generalizability:** ML models have shown high generalizability across different institutions, making them suitable for widespread clinical application10.

III. PROPOSED METHDOLOGY

Data Preprocessing Techniques

- Effective data preprocessing methods, including normalization and feature selection, significantly impact prediction accuracy(Nada & Dutta, 2025).
- The use of diverse datasets, such as those from the UCI machine learning repository, aids in training robust models for distinguishing between benign and malignant tumors(Nada & Dutta, 2025).

The present work relied on a public dataset [39]. The number of participants is 309, and all the attributes (15 as input to the ML models and 1 for the target class) are described as follows:

[1] **Gender** : This characteristic indicates whether the individual is male or female.
[2] **Age** (years) : This attribute indicates how old the individual is.
[3] **Smoking** : This characteristic shows whether or not the subject smokes.
[4] **Yellow fingers** : This characteristic indicates whether or not the subject has yellow fingertips.
[5] **Anxiety** : This attribute indicates whether or not the person is experiencing anxiety
[6] **Peer pressure** : This attribute determines whether or not the participant experiences peer pressure.
[7] **Chronic disease** : This characteristic indicates whether or not the subject has a chronic illness.
[8] **Fatigue** : This characteristic appears whether or not the subject experiences fatigue.
[9] **Allergy** : This feature indicates if the individual is allergic or not.
[10] **Wheezing** : This trait indicates whether or not a subject has wheezing.
[11] **Alcohol** : This feature indicates whether or not the user drinks alcohol.
[12] .**Coughing**: This characteristic indicates whether or not the subject coughs.
[13] **Shortness of breath**: This characteristic indicates whether or not the subject experiences dyspnea.
[14] **Swallowing difficulty**: This feature shows whether or not the participant has trouble swallowing.
[15] **Chest pain** : This characteristic indicates whether or not the subject experiences chest pain.
[16] **Lung Cancer**: This feature displays whether or not the individual has received a lung cancer diagnosis.

All the features are nominal except for age, which is numerical.

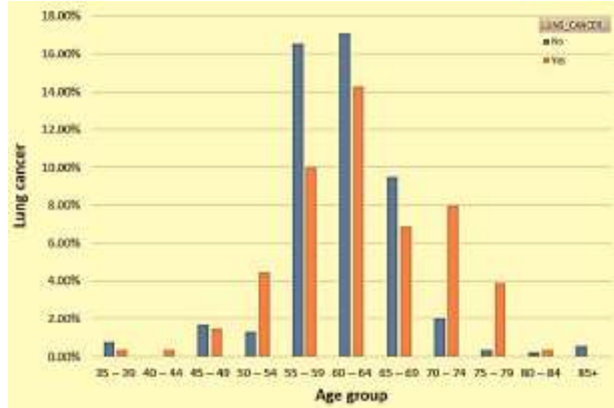

Fig 3. Age group Vs Lung Cancer

Table 1. Statistics of Lung Cancer Dataset

	count	mean	std	min	25%	50%	75%	max
AGE	309.0	62.67	8.21	21.0	57.0	62.0	69.0	87.0
SMOKING	309.0	1.56	0.50	1.0	1.0	2.0	2.0	2.0
YELLOW_FINGERS	309.0	1.57	0.50	1.0	1.0	2.0	2.0	2.0
ANXIETY	309.0	1.50	0.50	1.0	1.0	1.0	2.0	2.0
PEER_PRESSURE	309.0	1.50	0.50	1.0	1.0	2.0	2.0	2.0
CHRONIC DISEASE	309.0	1.50	0.50	1.0	1.0	2.0	2.0	2.0
FATIGUE	309.0	1.67	0.47	1.0	1.0	2.0	2.0	2.0
ALLERGY	309.0	1.56	0.50	1.0	1.0	2.0	2.0	2.0
WHEEZING	309.0	1.56	0.50	1.0	1.0	2.0	2.0	2.0
ALCOHOL CONSUMING	309.0	1.56	0.50	1.0	1.0	2.0	2.0	2.0
COUGHING	309.0	1.58	0.49	1.0	1.0	2.0	2.0	2.0
SHORTNESS OF BREATH	309.0	1.64	0.48	1.0	1.0	2.0	2.0	2.0
SWALLOWING DIFFICULTY	309.0	1.47	0.50	1.0	1.0	1.0	2.0	2.0
CHEST PAIN	309.0	1.56	0.50	1.0	1.0	2.0	2.0	2.0

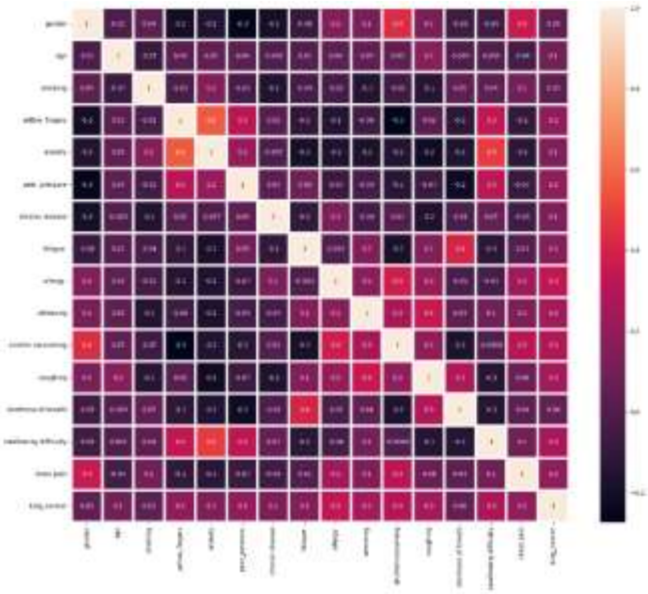

Fig 4. Heatmap Lung Cancer

Table 2. Feature Values and Class Label

Feature	Lung Cancer		Feature	Lung Cancer	
Gender	No	Yes	Allergy	No	Yes
Female	20.11%	23.18%	No	49.07%	18.07%
Male	[illegible]	[illegible]	Yes	[illegible]	[illegible]
Smoking	No	Yes	Wheezing	No	Yes
No	30.20%	21.30%	No	47.41%	18.81%
Yes	[illegible]	[illegible]	Yes	[illegible]	[illegible]
Yellow Fingers	No	Yes	Alcohol	No	Yes
No	[illegible]	19.81%	No	48.70%	19.44%
Yes	20.19%	30.19%	Yes	1.30%	30.60%
Anxiety	No	Yes	Coughing	No	Yes
No	[illegible]	[illegible]	No	[illegible]	18.70%
Yes	[illegible]	[illegible]	Yes	[illegible]	[illegible]
Peer Pressure	No	Yes	Shortness of Breath	No	Yes
No	48.13%	33.13%	No	11.67%	17.41%
Yes	1.85%	26.88%	Yes	[illegible]	[illegible]
Chronic Disease	No	Yes	Swallow Difficulty	No	Yes
No	41.85%	23.70%	No	49.07%	34.07%
Yes	8.15%	[illegible]	Yes	[illegible]	[illegible]
Fatigue	No	Yes	Chest Pain	No	Yes
No	15.93%	13.00%	No	32.59%	20.37%
Yes	34.07%	[illegible]	Yes	17.41%	[illegible]

IV. Evaluation Metrics

Table 3. Models' comparison

Model	Accuracy
SVM	95
ANN	94.6
NB	95
DT	93.7
KNN	95.2

V. CONCLUSION

While ML models have demonstrated potential in lung cancer prediction, challenges remain in optimizing model performance and ensuring generalizability across diverse datasets. Future research should focus on refining model parameters, exploring new data types, and integrating temporal treatment information to enhance prediction accuracy and clinical applicability10. Additionally, explainable ML models are crucial for understanding the influence of various features on prediction outcomes, thereby aiding in clinical decision-making8. Despite the advancements, challenges remain in the widespread adoption of machine learning models in clinical practice. The "black-box" nature of some models, particularly deep learning, poses ethical and regulatory challenges8. There is a need for more interpretable models, such as those combining deep learning with decision trees, to enhance clinical acceptance8. Future research should focus on integrating multiomic data to improve predictive accuracy and develop comprehensive models that can guide personalized treatment strategies10

REFERENCES

[1] Barroso, A.T.; Martín, E.M.; Romero, L.M.R.; Ruiz, F.O. Factors affecting lung function: A review of the literature. *Arch. De Bronconeumol.* **2018**, *54*, 327–332. [Google Scholar] [CrossRef]

[2] Barta, J.A.; Powell, C.A.; Wisnivesky, J.P. Global epidemiology of lung cancer. *Ann. Glob. Health* **2019**, *85*, 8. [Google Scholar] [CrossRef] [Green Version]

[3] Bell, S.C.; Mall, M.A.; Gutierrez, H.; Macek, M.; Madge, S.; Davies, J.C.; Burgel, P.R.; Tullis, E.; Castaños, C.; Castellani, C.; et al. The future of cystic fibrosis care: A global perspective. *Lancet Respir. Med.* **2020**, *8*, 65–124. [Google Scholar] [CrossRef] [Green Version]

[4] Bradley, S.H.; Kennedy, M.; Neal, R.D. Recognising lung cancer in primary care. *Adv. Ther.* **2019**, *36*, 19–30. [Google Scholar] [CrossRef] [Green Version]

[5] Dotan, Y.; So, J.Y.; Kim, V. Chronic bronchitis: Where are we now? *Chronic Obstr. Pulm. Dis. J. COPD Found.* **2019**, *6*, 178. [Google Scholar] [CrossRef]

[6] Hervier, B.; Russick, J.; Cremer, I.; Vieillard, V. NK cells in the human lungs. *Front. Immunol.* **2019**, *10*, 1263. [Google Scholar] [CrossRef] [PubMed] [Green Version]

[7] Mandell, L.A.; Niederman, M.S. Aspiration pneumonia. *N. Engl. J. Med.* **2019**, *380*, 651–663. [Google Scholar] [CrossRef]

[8] Mirza, S.; Clay, R.D.; Koslow, M.A.; Scanlon, P.D. COPD guidelines: A review of the 2018 GOLD report. In *Mayo Clinic Proceedings*; Elsevier: Amsterdam, The Netherlands, 2018; Volume 93, pp. 1488–1502. [Google Scholar]

[9] Schiller, H.B.; Montoro, D.T.; Simon, L.M.; Rawlins, E.L.; Meyer, K.B.; Strunz, M.; Vieira Braga, F.A.; Timens, W.; Koppelman, G.H.; Budinger, G.S.; et al. The human lung cell atlas: A high-resolution reference map of the human lung in health and disease. *Am. J. Respir. Cell Mol. Biol.* **2019**, *61*, 31–41. [Google Scholar] [CrossRef] [PubMed]

[10] Stern, J.; Pier, J.; Litonjua, A.A. Asthma epidemiology and risk factors. In *Seminars in Immunopathology*; Springer: Berlin/Heidelberg, Germany, 2020; Volume 42, pp. 5–15. [Google Scholar]