International Journal of Engineering and Information Systems (IJEAIS)
ISSN: 2643-640X
Vol. 8 Issue 6 June - 2024, Pages: 59-63

# A Machine Learning Approach for Lung Cancer Detection

**Yousuf Sk[1], Joydeep Mukherjee[2],**

**1 M.Tech Scholar, School of Education Technology, Jadavpur University, Kolkata, India**
**2Assistant Professor, School of Education Technology, Jadavpur University**

***Abstract:*** *Early detection of lung cancer plays a pivotal role in improving patient outcomes and reducing mortality rates. This paper presents a novel approach utilizing machine learning techniques for the early detection of lung cancer based on symptom-based numerical attributes. Leveraging a dataset comprising numerical attributes extracted from symptoms such as Yellow fingers, Anxiety, Chronic Disease, Fatigue, Allergy, Wheezing, Coughing, Shortness of Breath, Swallowing Difficulty, and Chest pain, our proposed model aims to accurately classify individuals at risk of lung cancer. Through rigorous experimentation and validation, our study demonstrates the efficacy of machine learning algorithms in accurately identifying potential cases of lung cancer at an early stage. The findings underscore the potential of integrating machine learning with symptom-based datasets for enhanced lung cancer detection, thereby contributing to improve prognosis and patient care. Here I have used four type of machine learning algorithm - Logistic Regression, MLPC, ANN, SVM and made a comparison between them to select the best model for detection of lung cancer.*

**Keywords:** Lung Cancer Detection, Machine Learning, EDA, Feature Extraction, Logistic Regression, MLPC, ANN, SVM, Feature Importance, Data Oversampling , Z-score Method

## 1.Introduction:

Lung cancer remains one of the most prevalent and deadly forms of cancer worldwide, posing a significant public health challenge. Lung cancer is among the most commonly diagnosed cancers globally. According to the World Health Organization (WHO), there were an estimated 2.2 million new cases of lung cancer globally in 2020. Lung cancer is the leading cause of cancer-related deaths worldwide. In 2020, it was estimated that lung cancer accounted for nearly 1.8 million deaths globally, making up approximately 18% of all cancer deaths. Early detection plays a crucial role in improving patient outcomes and reducing mortality rates associated with the disease. Conventional methods of lung cancer diagnosis often rely on invasive procedures and may not detect the disease until it has reached an advanced stage, limited treatment options and diminishing survival rates. In recent years, there has been a growing interest in leveraging machine learning techniques for the early detection of lung cancer. Machine learning offers the potential to analyze large volumes of data, including symptom-based attributes, to identify individuals at risk of the disease. By utilizing Numerical attributes extracted from symptoms such as Yellow fingers, Anxiety, Chronic Disease, Fatigue, Allergy, Wheezing, Coughing, Shortness of Breath, Swallowing Difficulty, and Chest pain, machine learning models can learn patterns and relationships within the data to accurately classify individuals at risk of lung cancer. This paper presents a novel approach to lung cancer detection using machine learning algorithms. Leveraging a dataset comprising symptom-based numerical attributes, we aim to develop a robust and accurate model capable of identifying potential cases of lung cancer at an early stage. By comparing the performance of four different machine learning algorithms - Logistic Regression, Multi-Layer Perceptron Classifier (MLPC), Artificial Neural Network (ANN), and Support Vector Machine (SVM) - we seek to determine the most effective model for lung cancer detection. Through rigorous experimentation and validation, our study aims to demonstrate the efficacy of machine learning in enhancing lung cancer detection and contributing to improved prognosis and patient care. By harnessing the power of machine learning and symptom-based datasets, we hope to pave the way for more effective screening early intervention strategies for lung cancer, ultimately saving lives and improving the quality of patient outcome.

## 2.Literature Review

In this paper [1] Ibrahim M. Nasser, Samy S. Abu-Naser investigates artificial neural networks (ANNs) for cancer detection, employing training and validation stages with an 80-20 data split. It likely covers ANN architecture, training, and performance evaluation.

This study [2] by Baidaa Al-Bander, Yousra Ahmed Fadil, and Hussain Mahdi develops a lung cancer prediction model using a multi-criteria decision aid system. It combines web survey data, AHP for feature selection, and MLP for symptom classification, improving early detection and treatment strategies.

In this paper [3] Radhanath Patra explores machine learning classifiers to distinguish between benign and malignant lung cancer using data from the UCI machine learning repository.

In this paper [4] V.Krishnaiah explores the application of classification-based data mining techniques like Rule-based, Decision tree, Naïve Bayes, and Artificial Neural Network to healthcare data for early detection and diagnosis of lung cancer.

In this research [5] Trailokya Raj Ojha examines various machine learning algorithms, including Support Vector Machine, Adaptive Boosting, k-Nearest Neighbour, Logistic Regression, J48, and Naïve Bayes, to predict lung cancer based on medical history and physical activities. The study finds that while all algorithms show high predictive accuracy, Logistic Regression stands out with an accuracy and f-measure of 94.7%, making it the most effective tool for lung cancer prediction and classification.

**3.Problem Statement:**

Lung cancer's high mortality is often due to delayed diagnosis and limited treatments. Traditional methods frequently miss early detection, leading to poor outcomes. This research leverages machine learning to improve early detection by analysing symptom-based numerical attributes. It also introduces a method to calculate individual lung cancer probability by examining patient-specific symptoms and risk factor.

**4.Methodology:**

This process of building a machine learning model to predict lung cancer and calculate the probability of lung cancer any individual patient consists of several stages, they are:

1. The dataset consists of symptoms were used to diagnose the lung cancer, these symptoms such as Yellow fingers, Anxiety, Chronic Disease, Fatigue, Allergy, Wheezing, Coughing, Shortness of Breath, Swallowing Difficulty and Chest pain. The target column is ‘Lung cancer’ which classifies if a patient has lung cancer or not.

2. We import the dataset to Jupyter Notebook.

3. Then the dataset is checked for any null or missing value. My dataset doesn’t contain any null value. Then, the dataset is checked for any outliers using z-score method. And after removing the outliers we then get a fully cleaned data.

4. Now, random oversampling technique is applied on the dataset to increase the size of the dataset as well as making it a balanced dataset. Then the process of random shuffling begins. Then the process of feature extraction takes place.

5. After separating the target attribute from the independent features, the dataset is split into training and testing by 8:2 ratio respectively. Then the following machine learning models are built and tested with the finally encoded and scaled train and test dataset accordingly.

6. Some sort of command is implemented using each corresponding model to calculate the probability of lung cancer in each patient.

Some major aspects of the proposed framework are:

**i. EDA:**

EDA, or Exploratory Data Analysis, is an approach to analyzing data sets to summarize the main characteristics, often with visual methods. It involves various techniques they are :

1. Understanding the Data's Structure: Identifying data types, distributions, and relationships between variables

1. Understanding the Data's Structure: Identifying data types, distributions, and relationships between variables

2. Detecting Outliers: Finding anomalies that might affect the analysis.
3. Uncovering Patterns: Revealing trends, clusters, and hidden structures.
4. Testing Hypotheses: Formulating and testing assumptions about the data.
5. Data Cleaning: Identifying and handling missing or inconsistent data.

**ii. Z-Score Method for Outlier Detection:**

1. Compute the mean (`62.705298` for `AGE`) and standard deviation of the data.
2. Loop Through Data: Iterate over each data point in the `AGE` column.
3. Calculate Z-Score: For each data point, calculate the z-score using $Z=(X-\mu)/\sigma$ where $Z$ is the z-score.
   - $X$ is the individual data point.
   - $\mu$ (mu) is the mean of the data set.
   - $\sigma$ (sigma) is the standard deviation of the data set.
4. Compare with Threshold: Identify outliers where the absolute z-score is greater than `3`.

5. Append Outliers: Add data points exceeding the threshold to the `outliers` list.
6. Return Outliers: Return the list of outliers. For `AGE`, the outlier identified is `21`.

**iii. Data Oversampling:**

To balance my imbalanced dataset, I used random oversampling, enhancing model generalization and feature learning. This technique randomly duplicates samples from the minority class ('NO') until both classes ('YES' and 'NO') reach the desired size of 10000 instances each. The process involves identifying minority and majority classes, determining the desired size, calculating sampling sizes, and randomly selecting samples with replacement. After oversampling both classes, the dataset is concatenated and shuffled to prevent bias during model training. While effective in mitigating class imbalance and improving model performance, careful implementation is essential to avoid overfitting.

**iv. Lung Cancer Prediction of Individual Patient:**

After building and training the machine learning models certain commands will be used with the model to calculate the probability of lung cancer by evaluating the clinical attributes or symptoms on the test dataset. Basically, this command will show the probability in the range of 0 to 1 where value closer to 0 denotes absence of lung cancer and value closer to 1 denotes presence of lung cancer. These results are same as the previously given results by the models.

## 5. Dataset Description:

| Feature | Levels |
|---|---|
| GENDER | Male(M), Female(F) |
| AGE | Age of the patient |
| SMOKING | YES=2, NO=1 |
| YELLOW_FINGERS | YES=2, NO=1 |
| ANXIETY | YES=2, NO=1 |
| PEER_PRESSURE | YES=2, NO=1 |
| CHRONIC DISEASE | YES=2, NO=1 |
| FATIGUE | YES=2, NO=1 |
| ALLERGY | YES=2, NO=1 |
| WHEEZING | YES=2, NO=1 |
| ALCOHOL CONSUMING | YES=2, NO=1 |
| COUGHING | YES=2, NO=1 |
| SHORTNESS OF BREATH | YES=2, NO=1 |
| SWALLOWING DIFFICULTY | YES=2, NO=1 |
| CHEST PAIN | YES=2, NO=1 |
| LUNG_CANCER | YES=2, NO=1 |

## 6. Results and Discussion:

I have proposed four different machine learning models: Logistic Regression, MLPC, ANN, SVM. After being trained with the trained dataset, i.e,80% of the oversampled shuffled dataset each model is then tested with the rest of 20% of the oversampled shuffled dataset. Then we run certain command using each model to calculate the probability of lung cancer of any individual patient by evaluating its clinical attributes. Also, we have shown the feature importance for each model. And we have got these results:

**a) Logistic Regression:**

Logistic regression is a statistical method used for binary classification tasks, where the goal is to predict the probability that a given input belongs to one of two classes. It is a type of regression analysis but is specifically designed for classification rather than prediction of a continuous outcome. The accuracy of logistic regression on test dataset is 92%. The model is also tested to see if it can accurately measure the

probability of lung cancer in each patient present in test dataset. And the model accurately does the same.

**b) MLPC:**

MLPC stands for Multi-Layer Perceptron Classifier. It is a type of artificial neural network used for classification tasks. The Multi-Layer Perceptron (MLP) is a feedforward artificial neural network model that maps sets of input data onto a set of appropriate outputs. The accuracy of this model is 100%.

**c) ANN:**

An Artificial Neural Network (ANN) is a computational model influenced by the manner in which biological neural networks in the human brain process information. ANNs are used in a variety of applications for pattern recognition, classification, regression, and many other tasks. They are a fundamental part of machine learning and deep learning. The accuracy of this model is 97.6%.

**d) SVM:**

Support Vector Machine (SVM) is a supervised machine learning algorithm commonly used for classification, regression, and outlier detection tasks. SVM is effective in high-dimensional spaces and is particularly useful for cases where the number of dimensions exceeds the number of samples. It is known for its robust theoretical foundation and its ability to create non-linear decision boundaries using kernel methods. The accuracy of this model is 92.4%.

**Conclusion:**

This study demonstrates the potential of machine learning techniques to improve early detection of lung cancer. By analysing symptom-based numerical attributes, we developed and compared models using Logistic Regression, Multi-Layer Perceptron Classifier (MLPC), Artificial Neural Network (ANN), and Support Vector Machine (SVM). Our findings highlight that machine learning models can significantly enhance early lung cancer diagnosis, enabling timely intervention and better patient outcomes. Future work will focus on refining these models and integrating them into clinical practice.

**References:**

[1] M. Nasser and S. S. Abu-Naser, "Lung Cancer Detection Using Artificial Neural Network," International Journal of Engineering and Information Systems (IJEAIS), vol. 3, no. 3, pp. 17-23, Mar. 2019.

[2] T. Divya and J. V. Gripsy, "An Integrated Deep Learning Based Enhanced Grey Wolf Optimization for Lung Cancer Prediction," Journal of Theoretical and Applied Information Technology, vol. 102, no. 6, pp. [include page numbers if available], Mar. 2024. ISSN: 1992-8645, E-ISSN: 1817-3195.

[3] N. Devihosur and R. K. M. G., "Enhancing Precision in Lung Cancer Diagnosis Through Machine Learning Algorithms," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 14, no. 8, 2023.

[4] M. TIA, O. N. Oyelade, and A. E. Ezugwu, "Automatic detection and classification of lung cancer CT scans based on deep learning and ebola optimization search algorithm," PLoS One, vol. 18, no. 8, p. e0285796, Aug. 2023, doi: 10.1371/journal.pone.0285796. PMID: 37590282; PMCID: PMC10434933.

[5] B. R. Pandit, A. Alsadoon, P. W. C. Prasad, S. Al Aloussi, T. A. Rashid, O. H. Alsadoon, and O. D. Jerew, "Deep Learning Neural Network for Lung Cancer Classification: Enhanced Optimization Function," Multimedia Tools and Applications, 2022, doi: 10.1007/s11042-022-13566-9.

[6] M. S. Kumar and K. V. Rao, “Prediction of Lung Cancer Using Machine Learning Technique: A Survey,” in 2021 IEEE International Conference on Computer Communication and Informatics (ICCCI), Jan. 2021, pp. 1–5.

[7] B. R. Manju, et al., "Efficient multi-level lung cancer prediction model using support vector machine classifier," IOP Conference Series: Materials Science and Engineering*, vol. 1012, p. 012034, 2021

[8] D. M. Abdullah, A. M. Abdulazeez, and A. B. Sallow, "Lung cancer Prediction and Classification based on Correlation Selection method Using Machine Learning Techniques," Qubahan Academic Journal, vol. 1, no. 2, pp. 141-149, 2021, doi: 10.48161.

[9] S. Mukherjee and S. U. Bohra, "Lung Cancer Disease Diagnosis Using Machine Learning Approach," in Proceedings of the Third International Conference on Intelligent Sustainable
Systems (ICISS 2020), Amravati, India, 2020. IEEE Xplore, Part Number: CFP20M19-ART, ISBN: 978-1-7281-7089-3.

[10] R. Patra, "Prediction of Lung Cancer Using Machine Learning Classifier," in Proceedings of COMS2 2020, CCIS 1235, N. Chaubey et al. (Eds.), Springer Nature Singapore Pte Ltd., 2020, pp. 132–142.

[11] G. A. P. Singh and P. K. Gupta, "Performance analysis of various machine learning-based approaches for detection and classification of lung cancer in humans," Neural Computing and Applications, vol. 31, no. 10, pp. 6863–6877, Oct. 2019.

[12] A. Axwe, E. Ayo, A. Uso, J. A. I. Joshua, and C. O. V. Chinedu, "Lung Cancer: A Chronic Disease Epidemiology; Prevalence Study," Asian Journal of Advanced Research and Reports, pp. 1–7, 2019.

[13] Y. Xie, "Knowledge-based Collaborative Deep Learning for Benign Malignant Lung Nodule Classification on Chest CT," in Proceedings of the IEEE Conference, 2018.

[14] M. I. Faisal, S. Bashir, Z. S. Khan, and F. H. Khan, "An evaluation of machine learning classifiers and ensembles for early-stage prediction of lung cancer," in 2018 IEEE 3rd International Conference on Emerging Trends in Engineering, Sciences and Technology (ICEEST), 2018, pp. 1–4.

[15] A. Khosravi and A. Khatami, "Lung cancer classification using deep learned features on low population dataset," in 2017 Canadian Conference on Electrical and Computer Engineering (CCECE), 2017, pp. [include page numbers if available]. IEEE, ISBN: 978-1-5090-5538-8.

[16] R. Sammouda, "Segmentation and analysis of CT chest images for early lung cancer detection," in Proceedings of the Global Summit on Computer & Information Technology (GSCIT), 2017, pp. [include page numbers if available]. IEEE, ISBN: 978-1-5090-2659-3.

[17] P. R. Hachesu, N. Moftian, M. Dehghani, and T. S. Soltani, "Analyzing a lung cancer patient dataset with the focus on predicting survival rate one year after thoracic surgery," Asian Pacific Journal of Cancer Prevention (APJCP), vol. 18, no. 6, p. 1531, 2017.

[18] V. Krishnaiah, G. Narsimha, and N. Subhash Chandra, "Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques," International Journal of Computer Science and Information Technologies (IJCSIT), vol. 4, no. 1, pp. 39–45, 2013.