Early Detection of Lung Cancer among Indian Patients using Machine Learning Algorithms

Sarthak Gupta

Delhi Public School Sec-45, India

ABSTRACT

Lung cancer causes 1 in 5 cancer related deaths globally. Statistics show that less than 16% of lung cancer cases are detected at an initial stage, leading to a significant number of lung cancer deaths within a year of diagnosis. This emphasises a requirement of improved early diagnostic methodologies for lung cancer.For this primary research, we collected comprehensive data from a prominent Indian hospital for over 2500 Indian patients. Our objective was to employ machine learning(ML) models to identify individuals at a heightened risk of developing lung cancer and more specifically, to identify key symptoms that could guide early diagnosis within the Indian populace. Furthermore, we aimed to assess and compare the diagnostic effectiveness of various machine learning (ML) models using established metrics. By evaluating multiple ML models, our study sought to identify the most optimal approach for early lung cancer detection in the Indian population. Notably, among the evaluated models, the Random Forest model emerged as the most promising. It demonstrated remarkable performance across multiple metrics, ultimately achieving the highest diagnostic accuracy of 93.3%. The significant findings of this study provide a strong foundation for future advancements in lung cancer diagnosis in India.

Introduction

The lungs, vital respiratory organs in the human body, are located in the chest. Their primary purpose is to facilitate gaseous exchange between the body and the environment. Inhaled air enters the body through the nose and passes successively through the pharynx, larynx, trachea and subsequently, reaching the bronchi present in the lungs.

Once the air reaches the lungs, the process of respiration takes place. Respiration involves an exchange of gases, specifically oxygen and carbon dioxide, between blood and air in the lungs. In the lungs, millions of tiny sacs called alveoli provide the surface area for this reaction to occur. Deoxygenated blood from body organs leaves carbon dioxide, a metabolic product, and absorbs oxygen at the alveolar surface.

One of the leading causes of mortality across the world, cancer accounted for over 10 million deaths worldwide in 2020 [1]. It is characterised by the proliferation and uncontrolled growth of cells within the body [2]. In a healthy body, cells possess intricate regulatory mechanisms to maintain a proper rate of division. Normal cells receive signals instructing them to divide and are provided with cues to halt division. Conversely, cancerous cells are not reliant on customary signals for cellular division. This loss of dependence causes cancer cells to undergo perpetual division, contributing to the disease's progression or its potential to metastasize to distant sites in the body [3].

Cancer arises as a consequence of DNA alterations within the cells of the body, leading to changes in the behaviour and function of the affected cells. These alterations, known as 'genetic mutations', are the root cause of all cancers. In addition, environmental and lifestyle factors can influence the development of cancer. Consuming substances like tobacco and alcohol, or exposure to hazardous substances like asbestos, which contain known carcinogens, are potential causes of cancer [4].

Although no cure has been devised for the disease yet, there are several lines of treatment a diagnosed patient can pursue. Among these, notable treatment options include surgery, chemotherapy and radiation therapy [5]. Surgical intervention becomes necessary in cases of benign tumours. By surgically removing tumours from the affected area, the procedure can effectively eliminate localised masses of cells. Chemotherapy entails the use of drugs to eliminate cancer cells throughout the body. This form of treatment damages the DNA of the affected cells, therefore, interfering with their ability to divide and grow. Radiation therapy, on the other hand, utilises high-energy radiation to target and destroy cancer cells in a specific area of the body. The ultimate goal of a cure to cancer involves the complete eradication of cancer cells from the body.

Lung cancer poses a significant health challenge in India. In 2022 alone, over 70,000 patients in India were diagnosed with lung cancer. The situation is expected to worsen in coming years, with estimates of the number of cases to rise to over 80,000 by the year 2025 [6]. Lung cancer is the second most common cancer in Indian males, however rates of cases have steadily increased for both sexes across the past few years.

In the context of developing countries such as India, the 5-year survival rate for lung cancer is alarmingly low, at 5% [7]. This statistic represents a prevalent situation wherein the majority of cases are diagnosed at advanced stages, where the cancer may have spread to other parts of the body. Unfortunately, effective therapies for advanced-stage cancer are both limited and aggressive. Chances of long-term survival are diminished when lung cancer has progressed to a late stage. Therefore, the primary objective of early diagnosis is to significantly improve the 5-year survival rate for lung cancer patients, thereby enhancing the chances of long-term survival.

Artificial Intelligence (AI) has emerged as a promising tool in the field of cancer research. A study [8] showed the possibilities of applying AI in cancer prognosis and diagnosis, and revealed its potential to complement human diagnosticians and achieve high accuracy. Another study [9] exemplified the implementation of AI in identifying new drug targets for pancreatic cancer. This was achieved by analysing genomic data from over 1000 patients and using machine learning algorithms to identify potential drug targets.

Machine learning is a subset of AI that enables computers to learn and acquire skills, without being explicitly programmed to do so. It has come to be seen as a valuable tool in the field of medical diagnostics, including the detection of lung cancer. Machine learning algorithms can be trained to analyse patient data and assist in identifying lung cancer based on symptoms. Instead of relying on medical images, these algorithms learn patterns and features from a large dataset of labelled patient records, allowing them to make predictions on new cases

One approach in using machine learning for lung cancer detection is to employ classification algorithms. These algorithms learn from datasets that include information on patient demographics, medical history, and specific symptoms associated with lung cancer. By analysing this data, the algorithms can identify patterns and correlations that may indicate the presence of lung cancer.

To train a machine learning model for lung cancer detection based on Indian-specific symptoms, we have gathered a large dataset of patient records from Indian hospital.. The dataset includes both positive (lung cancer present) and negative (lung cancer absent) cases among Indian patients. The model is then trained to associate specific combinations of symptoms with the presence or absence of lung cancer in the Indian context.

Machine learning algorithms can offer several benefits in the Indian context for lung cancer detection based on symptoms. They can assist healthcare professionals by flagging Indian patients with a higher likelihood of having lung cancer, prompting further investigation and enabling earlier diagnosis, which is particularly crucial in a resource-constrained healthcare system of India.

By leveraging machine learning to prioritise cases, Indian healthcare professionals can work more efficiently and make timely diagnosis, potentially leading to improved patient outcomes. However, it is essential to emphasise that machine learning algorithms are not a replacement for clinical expertise. They act as supportive tools, assisting Indian healthcare professionals in making more accurate and efficient diagnoses.

Machine Learning Algorithms

Logistic Regression

Data points can be categorised into distinct groups according to their features using a technique called logistic regression. It uses the 'sigmoid curve' mathematical function to determine if a data point falls into a specific category. Each feature of the data is assigned a 'weight', indicating its significance in determining the category. For instance, in a dataset of animal and bird pictures, features like skin colour, fur or feather, and the presence of a beak would be given specific weights. Some features may carry more weight than others.

These feature weights, along with a constant term called 'bias', are combined using a mathematical formula. Each data point in the collection is given a numerical value as a result of this processing. After that, each value is processed through a "sigmoid function," which creates a probability that ranges from 0 to 1. The possibility of a data point falling into one category or the other is represented by this probability.

$$f(x) = \frac{1}{1 + e^{-x}}$$

Formula for a sigmoid function, here x is the combination of features and their weights in logistic regression. This sigmoid function takes this input x and transforms it into a value between 0 and 1.

Decision Trees

Decision trees are versatile tools used for solving both classification and regression problems. They are referred to as classification trees for classification tasks and regression trees for regression tasks. The fundamental idea behind a decision tree is to reach a single conclusion through a sequence of true/false or yes/no questions.

The construction of a decision tree, known as 'decision tree learning,' begins with a structured dataset. Each feature's 'Gini impurity' is independently calculated during this process (other methods like entropy gain may also be used). Gini impurity quantifies the probability of misclassifying a value within a chosen node. The feature with the lowest Gini impurity value is selected as the root node of the tree.

Once the root node is determined, the remaining features are tested for their Gini impurity, considering the two possible binary values of the root node (if it is a binary feature). The feature that results in the least Gini impurity is chosen as the first branch node. This process continues iteratively, selecting the feature that minimizes Gini impurity at each step, until a Gini impurity value of 0 is reached. At this point, the tree is complete, and the outputs associated with the terminal nodes are referred to as 'leaf nodes.'

Random Forest

Random Forest is a machine learning method that enhances predictive accuracy and reduces sensitivity to changes in training data by utilizing multiple decision trees.

The word "random" in Random Forest refers to the construction of many decision trees using random data features, which are then combined to give a single result.

In this approach, multiple decision trees are built, each trained on a different subset of the training data, chosen randomly. This is achieved through a technique called Bootstrap aggregating (or "bagging"). For each tree, a training subset is drawn at random from the original training dataset with replacement. This random sampling ensures that each tree observes the data in a slightly different way, reducing the risk of overfitting and enhancing the model's generalization capabilities. By combining the predictions of multiple trees, Random Forest can achieve higher accuracy and robustness in handling diverse datasets.

K-Nearest Neighbours

K-nearest neighbors (K-NN) is a supervised machine learning technique that makes predictions by measuring the similarity between instances in a dataset. This versatile algorithm finds applications in various use cases, including recommender systems, classification, regression, and clustering.

The fundamental idea behind K-NN is to determine the class or category of an unknown data point based on the classifications of its nearest neighbors. The value of "k" in K-NN represents the number of neighbors considered during the classification process. To classify a new data point, the algorithm calculates the distances between that point and every other data point in the dataset.

By identifying the k closest neighbors, the algorithm predicts the class of the new point using the class labels of those neighbors. This method operates on the assumption that data points belonging to the same class tend to cluster together in the classification space.

Data and Methods

Source of Data

The data for this study was sourced from Neera Hospital, a prominent healthcare facility situated in Lucknow, Uttar Pradesh, India. The dataset consists of information from 2512 Indian patients, making it a sizable sample for the conducted analysis. Details about the different attributes included in the dataset will be described in the following subsection.

Data Description

1. Gender: This attribute describes whether the sex of the patient is male or female.
2. Age: This attribute captures the age of the patient
3. Smoking: This particular attribute denotes the patient's smoking status, indicating whether they are currently a regular smoker or have been one at any point in their life.
4. Family History of Cancer: This attribute illustrates the genetic history of cancer in the patient’s family.
5. Dyspnea: This attribute shows if the patient suffers from dyspnea, or shortness of breath, and the degree of difficulty faced.
6. Chest Pain: This attribute captures whether or not the patient suffers from chest pain
7. Weight Loss: This attribute indicates if the patient has undergone recent weight loss
8. Coughing: This attribute signifies the presence of coughing symptoms in the patient
9. Previous Lung Disease: This attribute describes the patient’s past lung-related issues
10. Occupational Hazards: This attribute signifies whether or not the patient faces potential risk to lung cancer due to their occupation
11. Pollution Level in Residence City: This attribute reflects the degree of pollution in the city where the patient lives
12. Allergy: This attribute captures the patient’s susceptibility to allergic reactions caused by particular substances or environmental factors
13. Coughing Blood: This attribute indicates if the patient shows the symptom of coughing up blood from the respiratory tract
14. Immediate Family Smokers: This attribute signifies whether the patient has close family members that possess smoking habits
15. Fatigue: This attribute illustrates if the patient has persistent feelings of tiredness or lethargy
16. Hoarseness of Voice: This attribute describes vocal changes in the patient, characterised by a raspy, rough or a strained tone of voice.
17. Lung Cancer: This attribute shows whether or not lung cancer has been diagnosed in a patient

Data Preprocessing

One of the initial challenges we encountered was the presence of missing values in specific rows of the dataset. To address this issue, we decided to remove the rows containing missing values. This was done to ensure that the subsequent analysis is based on a complete dataset without any biases or erroneous values caused by other methods of preprocessing. Table 1 shows the number of missing values in each column of the dataset before removal.

Column	Missing Values
Family History of Cancer	74
Dyspnea	14
Chest Pain	121
Weight Loss	182
Previous Lung Disease	119
Occupational Hazards	135
Allergy	67
Immediate Family Smokers	102

Figure 1. Columns containing missing values

Subsequently, to enable effective utilisation of categorical features within the dataset within machine learning algorithms, appropriate methods of encoding were employed.

For (categorical) features that exhibited an order of hierarchy (Dyspnea, Chest Pain, Weight Loss, Occupational Hazards, Pollution Level in Residence City, Hoarseness of Voice), label encoding was employed. For this, unique numerical values were assigned to each category, based on their relative order. In the case where the categorical values lacked any inherent order, one-hot encoding was used (Gender, Smoking, Family History of Cancer, Coughing, Previous Lung Disease, Allergy, Coughing Blood, Immediate Family Smokers, Fatigue, Lung Cancer). This technique transforms each categorical value into a set of binary variables, creating separate columns for each category. The effect of this is seen in Table 3 (Section 3), where regression coefficients for columns that were encoded using label encoding do not yield a coefficient for a specific value, but rather for the entire column, while columns encoded using one-hot encoding provide a specific regression coefficient for each possible value of that column.

Results and Discussion

The dataset for this study was assembled from several patient sources and covers all patient characteristics. With the use of this information, our primary objective was to create a comparative assessment of numerous classifier models in identifying high-risk individuals to lung cancer. Additionally, we aimed to examine the potential of different machine learning approaches and algorithms to develop a predictive model for lung cancer detection. Given the unique nature of our dataset, which comprises comprehensive information on Indian patients, our study provides results that are specific to the Indian populace.

This section evaluates key insights from the dataset used in the study, as well as results obtained from performing exploratory data analysis on the provided data. By examining the results from the EDA, we aim to gain critical insights into the key constituents of the data, as well as the trends and relationship within the data. The factors analysed directly impact the prediction outcomes.

Throughout this section, we will present and discuss relevant insights from the dataset, and discuss the results derived from analysing the data through various machine learning techniques. Additionally, we will delve into the interpretation and significance of the obtained results. This will involve exploring the relationship between predictive features and lung cancer outcomes, and discussing notable patterns or trends discovered.

Evaluation

For this study, a variety of machine learning models, including Logistic Regression (LR), Decision Tree (DT), K-Nearest Neighbours (KNN), Random Forest (RF), XGBoost (XGB), Gradient Boosting Classifier (GBC), CatBoost, and AdaBoost. The effectiveness of these models was evaluated based on accuracy, precision and recall. In Table 1, the models' performance evaluations are provided.

Among all the tested models, the CatBoost model exhibited the best performance, with accuracy, precision and recall scores equal to 97.2%, 96.9% and 95.4% respectively. In addition, high accuracy scores were achieved by AdaBoost (95.5%) and GBC(96.1%).

One notable insight from our evaluation is that ensemble methods significantly outperformed non-ensemble methods. Ensemble methods such as XGB (93.3%), RF(93.89%), GBC (96.1%), CatBoost (97.2%) and AdaBoost (95.5%) demonstrated superior performance in their diagnosis.

	Accuracy	Precision	Recall
LR	0.9277	0.8955	0.909
DT	0.8111	0.7285	0.7727
KNN	0.8277	0.7777	0.7424
RF	0.9389	0.9365	0.8939
XGB	0.9333	0.9218	0.8939
GBC	0.9611	0.9836	0.909
CatBoost	0.9722	0.9692	0.9545
AdaBoost	0.9555	0.9393	0.9393

Figure 2. Performance Evaluation of Different Models

Symptoms Analysis

To gain insight into the relationships between features available in the utilised dataset and the target variable, we utilised regression coefficients and feature importances. Regression coefficients provide numeric values that indicate the direction and magnitude of the relationship between features and the target variable. Positive coefficients indicate that an increase in the feature value corresponds to a greater likelihood of a lung cancer diagnosis, while a negative coefficient suggests a decreased likelihood.

Among the examined features, several stood out as significant predictors of the diagnosis. Age showed a positive coefficient (0.1317), indicating that older individuals were more likely to be associated with lung cancer. This observation is consistent with studies conducted by de Groot et al. (2018) [10] and H.T. Bates et al. (2021) [11]. Dyspnea (0.6852)[12] , Coughing (*Yes*) (1.4379) [13], and Immediate Family Smokers (*Yes*) (1.70594) [14], also had positive coefficients, showing that these factors significantly contributed to the diagnosis. Table 2 shows the features with the highest regression coefficients.

Table 3. Symptoms that impact early diagnosis

Feature	***Regression Coefficient***
Age	0.131725
Dyspnea	0.685246
Chest Pain (None)	-1.12605
Coughing (Yes)	1.4379
Immediate Family Smokers (Yes)	1.705946
Hoarseness of Voice (None)	-0.63553
Hoarseness of Voice (Moderate)	0.83484
Family History of Cancer (Yes)	0.858354
Weight Loss (No)	-1.067209
Coughing Blood (Yes)	1.77654
Smoking (Never Smoker)	-1.31301

Lung cancer is strongly linked with smoking, with recent research confirming this association. According to a 2021 study, 80% of patients diagnosed with lung cancer were smokers[15]. The most common carcinogen consumed in India through smoking is tobacco. Tobacco products contain known carcinogens and elevated levels of reactive oxygen species (ROS) that impair cell function and promote inflammation. Additionally, another study [16] found a significant correlation between death from lung cancer and smoking. Our findings provide confirmation of previous knowledge regarding the association of smoking and lung cancer. The regression analysis for the 'Smoking (Never Smoker)' column was revealed to be approximately -1.313, which highlights the increased susceptibility to lung cancer caused by smoking and other associated factors.

Even though a patient may not smoke, they may still be at risk of developing lung cancer due to exposure to secondhand smoking (SHS), if their family members or close contacts, including family members, friends or colleagues who are smokers. Tobacco use kills 8 million people a year, of which 1.2 million deaths are from exposure to SHS. It has been noted in the past that individuals who have never smoked in their lifetime but are married to a smoker, face a 24% increased risk of developing lung cancer. Our analysis shows that those with immediate family members who smoke face a 1.7 times greater chance of developing lung cancer than patients who come from non-smoking families.

Another considerably impactful factor in the diagnosis of lung cancer is the history of cancer in the patient's family. Individuals with a family history of cancer among first relatives face a nearly 50% higher risk of developing lung cancer compared to those without such a family history. The analysis performed for our study indicates an 80% higher probability of lung cancer in patients with a family history of cancer. While it is understood that familial lung cancer is aggregated by specific shared susceptibility genes, inadequate research on genetic factors in the development of lung cancer has been conducted till date. Only a few cancer-specific genes have been identified to date, leaving a gap in our understanding of the genetic mechanisms contributing to familial lung cancer.

Discussion

For our study, a dataset with data on over 2500 Indian patients was compiled from a prominent Indian hospital. The data used consisted of features that captured risk factors to lung cancer (such as smoking, or dyspnea). The symptoms on which data was collected are commonly used to diagnose the presence of lung cancer in a patient. Moreover, the analysis conducted on the data revealed insights specific to the Indian subcontinent and the Indian populace. It affirmed the smoking-lung cancer association, as well as revealed highly impactful symptoms, or risk factors, that could potentially be useful markers in early diagnosis. We then trained several ML models using the obtained data, to be able to detect high-risk individuals using various risk factors and correctly predict their diagnosis. Even in case the model does not yield an approximately perfect score, it would certainly aid in the diagnostic process, where its predictive capabilities would complement the clinical expertise of healthcare professionals.

Machine learning has gained popularity as a tool for screening or prediction of several diseases. Previous studies have focused on the potential of ensemble models, specially boosting algorithms, in the context of a variety of cancers. These studies have reported promising results, indicating the usefulness of these models in identifying high-risk individuals. Providing insights about the biomedical markers about the disease, their findings have complemented healthcare analytics by developing predictive models for detection and prevention. However, conclusive work exploring this is yet to be done, especially in the field of lung cancer diagnosis. Our study revealed that ensemble methods outperformed non-ensemble methods in the prediction task. Moreover, our findings are specific to lung cancer in India, for which ensemble methods could be viable diagnostic tools. Future studies with larger and more diverse samples, nevertheless, should be conducted to validate and generalise our findings. Moreover, image data from CT scans of lungs can be integrated with categorical data to provide more robust predictive models for lung cancer risk assessment.

Our study was conducted on data collected from a single hospital, and it did not come from multiple healthcare centres or institutes, which could have provided a more generalised view of lung cancer characteristics in India. Nevertheless, the dataset we utilised provided results that validated previous findings, and provided insights into the various symptoms that can be used to diagnose the disease, and applicable predictive approaches.

Conclusion

Lung cancer is the leading cause of cancer-related deaths. For developing countries, this problem is amplified by a lack of adequate healthcare facilities, and research in early diagnosis and treatment. The disease in India has one of the lowest 5 year survival rates in the world. A majority of cases of lung cancer lead to fatality due to late diagnosis, where the cancer has spread to other parts of the body.

This study aimed to determine the most optimal approach to diagnosing lung cancer using machine learning. Various machine learning models, such as LR, KNN, DT, RF, GBC, CatBoost, and AdaBoost were chosen and evaluated in terms of accuracy, precision, and recall score, after being trained on a dataset collected from an Indian hospital, with data on over 2500 patients. From the results, we found that the CatBoost model outperformed all others with an accuracy of 97.2%. Furthermore, the AdaBoost and GBC models demonstrated high performance, with accuracies of 95.5% and 96.1% respectively.

An analysis was also performed to determine which factors, or symptoms, influenced the diagnosis to the greatest extent. The results verified already existing knowledge of lung cancer and its symptoms, as well as revealed insights into the potential markers of lung cancer in India. Smoking, Dyspnea, Immediate Family Smokers, and Family History of Cancer emerged as key indicators of lung cancer from the dataset.

In conclusion, our study aims to lay a foundation for more advanced predictive technologies for lung cancer. Using popular ML algorithms, we aim to create a comprehensive methodology for early diagnosis that has the potential to save thousands of lives and improve the healthcare landscape of India.

References

1. Beresford, M.J., Wilson, G.D. and Makris, A. (2006) ‘Measuring proliferation in breast cancer: Practicalities and applications’, *Breast Cancer Research*, 8(6). doi:10.1186/bcr1618.
2. Brennan, P. and Davey-Smith, G. (2021) ‘Identifying novel causes of cancers to enhance cancer prevention: New strategies are needed’, *JNCI: Journal of the National Cancer Institute*, 114(3), pp. 353–360. doi:10.1093/jnci/djab204.

3. *Cancer* (2022) *World Health Organization*. Available at: https://www.who.int/news-room/fact-sheets/detail/cancer (Accessed: 24 July 2023).
4. Falzone, L., Salomone, S. and Libra, M. (2018) 'Evolution of cancer pharmacological treatments at the turn of the Third Millennium', *Frontiers in Pharmacology*, 9. doi:10.3389/fphar.2018.01300.
5. Fares, J. *et al.* (2020) *Molecular principles of metastasis: A hallmark of cancer revisited*, *Nature News*. Available at: https://www.nature.com/articles/s41392-020-0134-x (Accessed: 18 September 2023).
6. Mathur, P. *et al.* (2022) 'A clinicoepidemiological profile of lung cancers in India – results from the National Cancer Registry Programme', *Indian Journal of Medical Research*, 155(2), p. 264. doi:10.4103/ijmr.ijmr_1364_21.
7. Huang, S., Yang, J., Fong, S., & Zhao, Q. (2020). Artificial Intelligence in cancer diagnosis and prognosis: Opportunities and challenges. *Cancer Letters, 471*, 61-71. doi:10.1016/j.canlet.2019.12.007
8. Iorio, F., Knijnenburg, T. A., Vis, D. J., Bignell, G. R., Menden, M. P., Schubert, M., . . . Garnett, M. J. (2016). A landscape of pharmacogenomic interactions in cancer. *Cell, 166*(3), 740-754. doi:10.1016/j.cell.2016.06.017
9. P.R., R., Nair, R.A.S. and G., V. (2019) 'A comparative study of lung cancer detection using machine learning algorithms', *2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)* [Preprint]. doi:10.1109/icecct.2019.8869001.
10. de Groot, P.M. *et al.* (2018) 'The epidemiology of Lung Cancer', *Translational Lung Cancer Research*, 7(3), pp. 220–233. doi:10.21037/tlcr.2018.05.06.
11. Bates JHT, Hamlington KL, Garrison G, Kinsey CM. Prediction of lung cancer risk based on age and smoking history. Comput Methods Programs Biomed. 2022 Apr;216:106660. doi: 10.1016/j.cmpb.2022.106660. Epub 2022 Jan 25. PMID: 35114461; PMCID: PMC8920760.
12. Damani A, Ghoshal A, Salins N, Muckaden MA, Deodhar J. High Prevalence of Dyspnea in Lung Cancer: An Observational Study. Indian J Palliat Care. 2019 Jul-Sep;25(3):403-406. doi: 10.4103/IJPC.IJPC_64_19. PMID: 31413456; PMCID: PMC6659529.
13. Harle, A.S.M. *et al.* (2019) 'Cough in patients with lung cancer', *Chest*, 155(1), pp. 103–113. doi:10.1016/j.chest.2018.10.003.
14. Gorlova, O.Y. *et al.* (2007) 'Aggregation of cancer among relatives of never-smoking lung cancer patients', *International Journal of Cancer*, 121(1), pp. 111–118. doi:10.1002/ijc.22615.
15. Singh, N. *et al.* (2021) 'Lung cancer in India', *Journal of Thoracic Oncology*, 16(8), pp. 1250–1266. doi:10.1016/j.jtho.2021.02.004.