International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
IJISAE
ISSN:2147-6799 www.ijisae.org Original Research Paper

Lung Cancer Detection, Prediction and Analysis of Lifestyle Parameters using ML and AI Techniques

Sarika Davare[1], Vishal Shirsath[2], Farook Sayyad[3]

Submitted:11/03/2024 **Revised:** 26/04/2024 **Accepted:** 03/05/2024

Abstract: Cancer poses a significant threat to human life, often diagnosed in later stages, highlighting the crucial need for early prediction. Literature extensively explores Machine Learning, Data Mining, and Artificial Intelligence techniques for the identification, classification, prediction, and detection of various cancers like lung, breast, prostate, skin, liver, and recurrence cancer. Predictive models rely on vast datasets for cancer prediction. Lung cancer's development is closely tied to lifestyle factors such as smoking, air pollution, and diet imbalance, emphasizing the potential of lifestyle indicators in early detection. A study focuses on constructing a model using lifestyle data to predict lung cancer and categorize its severity. Basic lifestyle parameters are initially examined, and if lung cancer potential is indicated, a second component of the model further analyzes each parameter to predict cancer level. Various Machine Learning techniques including Support Vector Machine, Logistic Regression, and Linear regression are applied to predict lung cancer risk and level with analysis of prediction using AI techniques. SVM emerges as the most effective classifier for predicting lung cancer risk based on lifestyle factors, while linear regression is optimal for total risk score prediction. Additionally, gender- and age-specific lifestyle parameters contributing to lung cancer are identified. The study's preliminary phase employs logistic regression and Support Vector Machine to predict lung cancer, achieving high accuracies of 94% and 90% respectively. The subsequent component utilizes SVM, Random Forest, KNN and Linear Regression to estimate cancer malignancy levels, with accuracies reaching 98% , 96 and 97% respectively. The study aims to predict lung cancer early using lifestyle data, offering insights into risk factors and preventive strategies.

Keywords: *Artificial Intelligence, Lifestyle, Lung Cancer, Machine Learning, Prediction*

1. Introduction

Uncontrolled cell proliferation is the root cause of cancer, a dangerous disease that can strike at any time and turn healthy cells into cancerous ones. Its roots are in the intricate interactions between age, lifestyle decisions, and genetic predispositions. A significant worldwide public health concern is lung cancer, marked by its high incidence rate and mortality. In 2020, as per the report of the World Health Organization (WHO) lung cancer is the cause for the highest number of cancer-related deaths globally, with approximately 1.8 million deaths among 2.2 million newly diagnosed cases in recent years [25]. Early identification becomes critical since it has the ability to change the course of this never-ending fight. Lifestyle choices are a crucial factor in determining cancer risk. In [2] reveal significant associations between family history, cancer screening, and lifestyle behaviours using logistic regression. Lifestyle factors and genetic scores calculated in determining CRC risk compared to family history [3].

The bulk of research efforts in this domain have been directed towards harnessing the wealth of information available within hospitals such as a wide array of medical records including CT scan images, Radiomic Data, Gnome data, and pathology reports reflecting a reliance on clinically relevant data sources for cancer prediction studies.

Aim of this research work is to predict lung cancer based on lifestyle parameters. Lifestyle choices significantly influence the developing risk of lung cancer, with tobacco use, primarily smoking, occupational hazard, air pollution serving as the leading cause of this devastating disease. The goals of this research work are to predict lung cancer disease further if the person is predicted to have lung cancer, then the risk level of developing lung cancer is predicted using machine learning and Artificial Intelligence techniques. This research work analyses the lifestyle factors which are cause for lung cancer using predictive models, this research seeks to assess individuals' susceptibility to the disease.

This research work will help to improve the health condition of a person and his/her family. Healthcare is something which affects the economy of the person. Prediction of life-threatening disease plays an important role in healthcare sector. Nowadays lifestyle is more immune to the various lives threatening disease. Based on the lifestyle parameter data we can predict the risk of occurrence of cancer at very early stage. Further if we can properly analyse the lifestyle parameters which are

[1] *Ajeenkya DY Patil School of Engineering. Pune, India*
ORCID ID: https://orcid.org/0009-0008-5794-085X
Email id: sarika.davare@adypu.edu.in
[2] *Ajeenkya DY Patil School of Engineering Pune, India*
ORCID ID: https://orcid.org/0000-0001-5251-1935
Email ID: vss.csit@gmail.com
[3] *Ajeenkya DY Patil School of Engineering Pune, India*
ORCID ID: https://orcid.org/0000-0003-0421-9344
Email id: farooksayyad@dypic.in

responsible for occurrence of cancer, we can change or improve that lifestyle parameter so that the probability of cancer can be reduced or nullified. If the analysis of the lifestyle parameter data is not taken seriously it will affect the health issues of a person and will also affect the economic condition of the family.

The rest of the contents are organized as literature review is covered in section 2. The overall Methodology with architecture is explained in section 3. The experimental results of the research are elaborated in section 4 and section 5 provides the conclusion.

2. Literature Review

In [1] authors introduce a streamlined solution for evaluating the likelihood of Non-Small Cell Lung Cancer (NSCLC) through the analysis of Lung Cancer microRNAs (LC-miRNAs) structures. The proposal advocates for the application of computational methods in miRNA and sequence analysis to enhance accuracy while reducing research expenses. In 2019 research [5] utilizes ensemble techniques and supervised learning algorithms to build a new model to predict breast cancer. Feature selection techniques are employed to enhance model performance, achieving higher accuracy using bagging, boosting, and stacking techniques. In [6], a lifestyle-based colorectal cancer (CRC) risk prediction model, integrates age, waist circumference, and dietary habits to accurately identify high-risk individuals in European populations. Demonstrating strong discrimination and calibration, it offers potential for personalized prevention strategies by motivating lifestyle changes. Its effectiveness is notable, particularly for individuals under 45 years old, showcasing its utility in early intervention. In 2021 authors [7] utilizes a data from a Chinese CRC screening trial, models incorporating lifestyle factors and genetic variants assess relative and 10-year absolute risks of colorectal neoplasm, informing personalized screening strategies. Individuals with unfavourable lifestyles and higher polygenic risk scores demonstrate significantly increased risks, emphasizing the potential of risk-adapted screening approaches for colorectal cancer prevention.

In [8] researchers investigated disease prediction for lifestyle diseases, employing periodical health checkup data, daily monitoring, and medical imaging for early detection. It introduces three approaches: fuzzy set-based analysis for uniform attribute manipulation and self-organizing maps to elucidate relationships among health examination data, with promising results for condition understanding and early detection of cerebral artery aneurysms using automated analysis methods.

In 2022 [9] developed a health assessment and disease prediction system using Decision Trees aims to predict individuals' likelihood of lifestyle-related diseases based on their health and lifestyle data. The system will offer personalized recommendations to improve lifestyle habits, ultimately reducing disease risk and promoting better health outcomes.

In 2023 research [10] addressed the urgency of lung cancer, proposing a risk prediction model, "Can Predict (lung)", in response to the UK's call for refined screening methods. various predictors including demographics, lifestyle factors, and medical history, exhibiting superior discrimination, calibration, and sensitivity compared to seven existing models. If implemented, it could aid in targeted screening, effectively identifying high-risk persons for lung cancer screening programs in primary care settings.

In [11] forecasted cancer incidence rates and onset ages globally from 2020 to 2040, using Bayesian models. Results indicate a majority of cancers will experience increasing incidence rates, with the proportion of cases occurring after age 60 expected to rise. Gender disparities are observed, with certain cancers showing a younger onset age in men by 2040, and despite demographic changes, a youth-oriented trend is noted in half of cancer morbidity globally, particularly in hormone-related and digestive tract cancers. In 2023 authors [12] assessed 11 lung cancer prediction models in Asian populations, finding existing models underestimate risk in certain subgroups. "Shanghai models" are developed, showing marginal overall improvement but outperforming Western models for low-intensity smokers and long-term quitters in Asia, addressing a gap in risk prediction for this population. In 2024 authors [13] introduced novel indicators, to evaluate the risks associated between lung cancer and Chronic Obstructive Pulmonary Disease (COPD), two metrics are used as Age at zero Mortality (AM0) and age at Average Mortality (AMa). It finds that while overall mortality risk for both diseases has decreased with environmental improvements, lung cancer exhibits lower AM0 values than COPD, indicating a potentially greater threat.

In [14] investigated lung cancer trends in China from 1990 to 2019, analysing incidence rates, death rates, and disability-adjusted life years (DALYs) using Global Burden of Disease data. Results reveal a notable increase in age-standardized incidence rates, particularly among men, with smoking, air pollution, and diet identified as primary risk factors. Predictive modelling suggests a potential rise in new cases and deaths over the next 15 years, emphasizing the need for targeted prevention strategies focusing on smoking cessation and reducing air pollution.

In study conducted in 2023 [16] to identify lung cancer severity levels and analyse risk factors for the disease using a machine learning algorithm. It used a decision tree-based ranking system to identify the most significant risk factors for lung cancer, and the most serious ones were found to be bloody coughing, air pollution, and obesity. It was suggested to use Extreme Gradient Boosting (XGBoost) in a machine learning model to identify patients' lung cancer severity. The model's efficacy was demonstrated by its excellent recall, accuracy, and precision rates.

In [17] authors created a precise machine learning model for early lung cancer prediction by analysing clinical and demographic variables. Utilizing various machine learning techniques, including Logistic Regression and Decision Trees, researchers aim to enhance healthcare outcomes. By customizing screening and prevention strategies based on patient records, the study aims to address the serious threat of lung cancer with the potential to save lives and improve patient care.

In [18] addresses the urgency of early lung cancer detection due to its poor survival rates and absence of universal screening. This highlights the potential of Machine Learning to enhance early detection and improve clinical outcomes in lung cancer management.

In [19] authors addressed Lung cancer, a major cause of mortality, results from genetic, environmental, and lifestyle factors, surpassing heart disease in developed nations. A data mining-based lung cancer prediction system (DMBCPS) is proposed, employing algorithms like Naive Bayes, SVM, and Random Forest to provide early warnings and performance analysis for improved prediction and management.

In 2023 [20] reveals gender-based disparities in healthy individuals and NSCLC, with ALB and TP identified as crucial nutrition indicators for accurate prediction. Their lung cancer prediction model achieves high accuracy, with an average AUC of 87.4% using five key nutrition indicators and 93.5% with a 15-index model, suggesting potential for screening and biopsy substitution. The composite index prediction method can be generalized to predict risks of various nutrition-related diseases, making it suitable for personalized health communication.

In study [21] machine learning algorithms offer potential for enhancing disease diagnostics, notably for lung cancer detection, despite challenges like high dimensionality and low accuracy. This study develops an ensemble classifier combining RF, SVM, NB, and KNN, outperforming individual techniques with 98.25% classification accuracy and the lowest error rate, particularly surpassing NB with 84.75% classification accuracy.

In [22] the authors developed a GUI application leveraging AI predicts lung cancer levels with 98% accuracy based on clinical features or histopathological images, while a comparative study identifies optimal ML and deep learning algorithms for classification.

In [23] researchers developed a lung cancer risk prediction model based on health examination data, showing good discrimination and calibration. Factors like age, sex, smoking intensity, BMI, COPD, pulmonary TB, and type 2 DM are considered, aiding individuals in decision-making for screening and lifestyle modifications, like smoking cessation, to reduce lung cancer mortality.

In [24] researchers endeavour the study to create a population-level lung cancer risk prediction tool using extensive data from southeast England, employing Machine Learning techniques like linear regression modelling. It identifies a final set of seven attributes for the Kent & Medway lung cancer risk prediction tool, outperforming existing screening programs by detecting more cases, showcasing the utility of Machine Learning in improving lung cancer risk assessment.

3. Methodologies

3.1 Lung cancer prediction Architecture

In the context of a Lung cancer prediction system, an architecture is shown in **Fig. 3.1**

The complete Architecture is divided in three layers as Presentation layer, Application Layer and Data Layer. Overall architecture is followed for data collection, prediction, and detailed data analysis. At Data Layer, data sources is utilized for thorough analysis. These datasets are then divided into training and testing sets for further processing. Moving into the application Layer, a range of machine learning (ML) and artificial intelligence (AI) techniques are applied to the dataset to predict lung cancer occurrence and assess its severity. Different machine learning techniques applied to predict lung cancer are Support Vector Machine (SVM) and Logistic regression. To predict lung cancer severity the machine learning techniques applied are SVM, Linear Regression and Random Forest. Artificial Intelligence techniques applied to further process lifestyle dataset are Generative AI and to analyse the machine learning techniques is Explainable AI. The application layer includes the in-depth analysis, involving statistical examination of lifestyle parameters. This critical analysis aims to extract meaningful insights and establish relevance to improve prediction accuracy. The presentation layer is for user interaction.

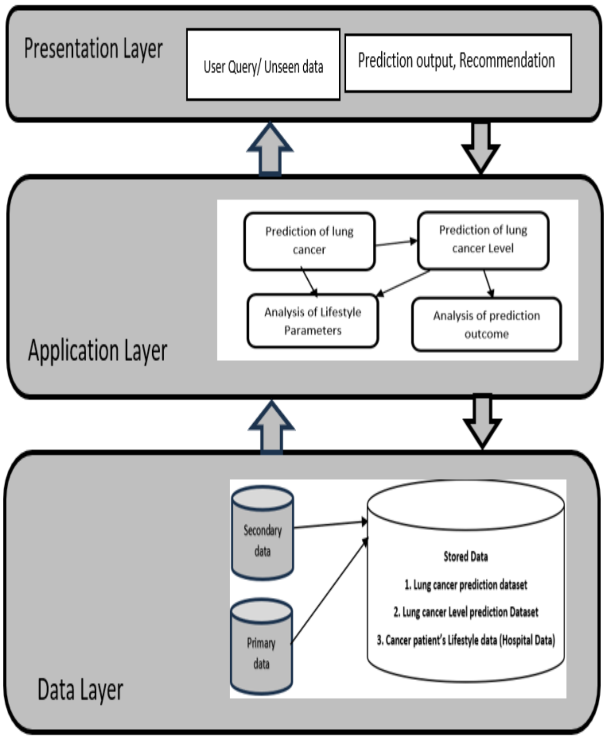

Presentation Layer
User Query/ Unseen data
Prediction output, Recommendation
Prediction of lung cancer
Prediction of lung cancer Level
Analysis of Lifestyle Parameters
Analysis of prediction outcome
Application Layer
Secondary data
Primary data
Stored Data
1. Lung cancer prediction dataset
2. Lung cancer Level prediction Dataset
3. Cancer patient's Lifestyle data (Hospital Data)
Data Layer

Fig. 3.1. Layered Architecture to Analyse Lung cancer data

3.1.1 Data Layer:

The data layer stores the data required for analysis. In this work three different types of datasets are used the first dataset is Lung cancer Prediction dataset which is used to predict the lung cancer as shown in the table 1.

Table 1. Attributes of Lung Cancer prediction dataset

Sr. No	Name of Attribute
1	Gender (M, F)
2	Age
3	Smoking
4	Yellow Fingers
5	Anxiety
6	Peer Pressure
7	Chronic Disease
8	Fatigue
9	Allergy
10	Wheezing
11	Alcohol
12	Coughing
13	Shortness of Breath
14	Swallowing Difficulty
15	Chest Pain
16	Lung Cancer (YES, NO)

The second dataset is used to predicts the severity level of lung cancer, categorized as Low, medium, or High as shown in table 2.

Table 2. Attributes of Lung cancer level prediction dataset

Sr. No	Name of Attribute
1	Patient id
2	Age (Range from 14 to 73)
3	Gender (Male and Female)
4	Air Pollution (Range from 1-8)
5	Alcohol use (Range from 1-8)
6	Dust Allergy (range from 1 to 8)
7	Occupational Hazards (range 1 to 8)
8	Genetic Risk (range from 1 to 7)
9	Chronic Lung Disease (range from 1 to 7)
10	Obesity (range from 1 to 7)
11	Smoking (range from 1 to 7)
12	Passive Smoker (range from 1 to 8)
13	Chest Pain (range from 1 to 9):
14	Coughing of Blood (range from 1 to 9)
15	Fatigue (range from 1 to 9)
16	Weight Loss (range from 1 to 8)
17	Shortness of Breath (range from 1 to 9)
18	Wheezing (range from 1 to 8)
19	Swallowing Difficulty (range from 1 to 8)
20	Frequent Cold (range from 1 to 7)
21	Dry Cough (range from 1 to 7)
22	Level(Low, Medium, High)

The third dataset is the primary dataset collected from the Cancer Hospital named as “RASAYU clinic, Pune” collection involved gathering primary data from individuals diagnosed with lung cancer at varying severity levels. This entailed obtaining information directly from patients currently experiencing different stages of lung cancer, ranging from low to high severity as shown in table 3.

Table 3. Attributes of Rasayu Hospital Data

Sr. No.	Name of Attribute
1	Gender (M, F)
2	Age
3	Alcohol (range from 1-8)
4	Smoking (range from 1-8)
5	Family history of cancer (range from 1-8)
6	Chest pain (range from 1-8)
7	Coughing of blood (range from 1-8)
8	Fatigue (range from 1-8)
9	Shortness of breath/ dyspnea (range from 1-8)
10	Swallowing difficulty/ dysphagia ((range from 1-8)
11	Wheezing (range from 1-8)
12	Weight loss (range from 1-8)
13	Cough (range from 1-8)
14	Level

3.1.2 Application Layer:

Application layer is divided in four main components as prediction of lung cancer, prediction of level of lung cancer, Analysis of lifestyle parameters and analysis of prediction outcome. For prediction of Lung cancer disease and the Level of lung cancer Machine Learning techniques are applied. For analysis of prediction algorithm Artificial Techniques are applied. For Analysis of Lifestyle parameters Random Forest method is applied.

3.1.3 ML techniques applied

To predict lung cancer, different machine learning techniques are applied to the Lung Cancer Prediction dataset. This dataset comprises lifestyle data from both lung cancer patients and non-patients. Two specific techniques, SVM and Logistic Regression, are utilized for analysis. Among these techniques, Logistic Regression emerges as the most suitable method for accurately predicting lung cancer, offering superior accuracy compared to SVM. To forecast the severity of lung cancer, diverse machine learning techniques are deployed on the Lung Cancer Level Prediction dataset. This dataset encompasses 23 lifestyle parameters from 1000 patients, categorizing them into Low, Medium, or High levels of lung cancer. Machine learning methods such as SVM, Random Forest, and Multiple Linear Regression are employed on this dataset. Among these approaches, SVM stands out as the most effective technique for predicting the level of lung cancer. Additionally, Linear Regression is utilized to predict both the level of lung cancer and the parameter scores of lung cancer patients. In the third dataset, known as the primary dataset, the SVM technique is employed for lung cancer prediction. This primary dataset is obtained directly from lung cancer patients. Through a conducted survey, data is collected from a total of 21 lung cancer patients. This data is treated as unseen data and is subsequently applied to the lung cancer prediction model for analysis.

3.1.4 AI Techniques applied

Generative AI and Explainable AI techniques applied on Lung cancer prediction dataset. For Generative AI technique this likely includes various health-related information such as medical history, lifestyle factors, genetic predispositions, and possibly recent diagnostic test results. After analyzing the patient's data, the system generates an overall health summary. This summary likely includes an assessment of the patient's current health status, identifying any existing health risks or conditions, and providing insights into areas of concern. The system identifies specific health risk parameters relevant to the patient's lifestyle. In this context, it appears to focus on parameters associated with the risk of developing lung cancer. These parameters might include smoking habits, exposure to environmental pollutants, dietary factors, and possibly genetic factors. Based on the identified health risk parameters, the system recommends preventive measures. These measures are designed to help the patient mitigate or avoid the identified health risks associated with lung cancer. They may include lifestyle changes such as quitting smoking, adopting a healthier diet, increasing physical activity, and avoiding exposure to known carcinogens.

Explainable AI presents the interpretation of a machine learning model's predictions for a specific instance. Each line corresponds to a feature and its associated contribution to the prediction as shown in figure.

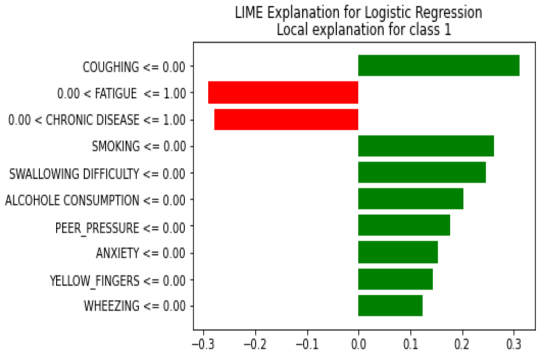


Fig 3.2: Lime Explanation for Logistic Regression

3.1.5 Analysis of Lifestyle parameters:

3.1.5.1 Analysis of lifestyle parameters based on dataset

Leveraging the Random Forest method facilitates a comprehensive examination of these features in two distinct datasets: the Lung Cancer Prediction Dataset and the Lung Cancer Level Prediction Dataset. Random Forest, a versatile ensemble learning technique, offers a robust framework for assessing feature importance by evaluating their contributions to the model's decision-making process.

In the context of lung cancer prediction, Fig. 3.3 illustrates the feature importance index for each lifestyle parameter within the dataset. Notably, Smoking and Fatigue emerge as the top two lifestyle parameters contributing to the prediction of lung cancer. This suggests that these factors play a significant role in the development or manifestation of the disease within the studied population.

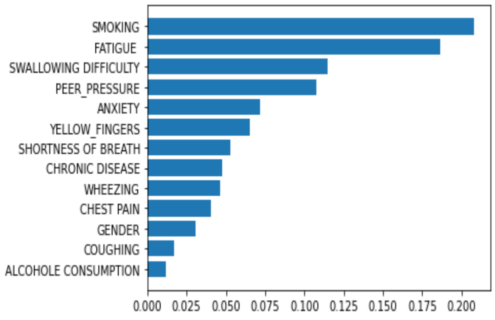


Fig. 3.3: Feature importance for Lung cancer prediction dataset

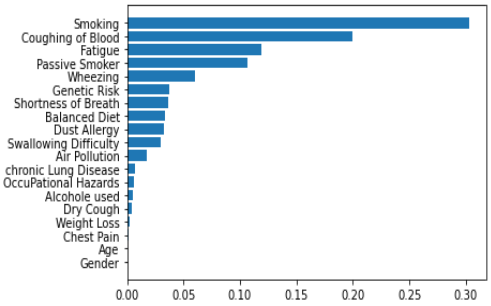


Fig. 3.4: Feature importance for Lung cancer level prediction

dataset

Fig. 3.4 shows the feature importance for each feature of Lung cancer level prediction dataset. As shown in the figure Smoking and Coughing of Blood are the parameters cause for the cancer.

3.1.5.2 Statistical analysis of parameters cause for Lung cancer

The top 4 parameters which are cause for cancer are considered and for each of these parameters following statistical analysis is performed.

a. Variance calculation is calculated by taking the average of the squared differences from the mean. A higher variance indicates that the data points are more spread out from the mean.

The formula for variance σ^2 of a sample is shown in equation 1:

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad (1)$$

Where x_i each individual data point, $\bar{x}$ is the mean of the data points, and n is the number of data points.

b. Standard deviation analysis for lifestyle parameters associated with lung cancer assesses the spread or dispersion of data points around the mean value. The formula for standard deviation σ is shown in equation 2:

$$\sigma = \sqrt{\sigma^2} \qquad (2)$$

Table 4. Variance and Standard Deviation for Lifestyle parameters

Sr. No.	Name of Lifestyle Parameters	Variance	Standard Deviation
1	Smoking	6.86	2.62
2	Coughing of Blood	5.89	2.42
3	Fatigue	5.03	2.24
4	Passive Smoker	5.34	2.31

As shown in table 4 for smoking parameter, a variance of 6.86 suggests that there is a moderate amount of variability in the smoking behaviour among the individuals surveyed. Some i b. Co-variance calculation individuals may smoke heavily, while others may smoke very little or not at all.

A standard deviation of 2.62 indicates that, on average, the smoking behaviour of individuals in the group deviates from the mean (average) smoking behaviour by about 2.62 units. This means that there is considerable variability in smoking habits among the individuals surveyed, with some smoking more or less than the average.

c. Covariance measures the degree to which two variables change together. It indicates the direction of the linear relationship between two variables. If the covariance is positive, it means that as one variable increases, the other variable tends to increase as well. If the covariance is negative, it means that as one variable increases, the other variable tends to decrease.

The formula for the covariance cov(X, Y) between two variables X and Y is specified in equation 3.

$$cov(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \qquad (3)$$

where x_i and y_i are individual data points of variables X , Y. $\bar{x}$ and $\bar{y}$ are the means of variables X and Y and n is the number of data points.

Table 5. Co-variance for top four parameters

	Smoking	Coughing of Blood	Fatigue	Passive Smoker
Smoking	6.86	4.24	1.39	3.58
Coughing of Blood	4.24	5.89	2.62	3.57
Fatigue	1.39	2.62	5.03	1.96
Passive Smoker	3.58	3.57	1.96	5.34

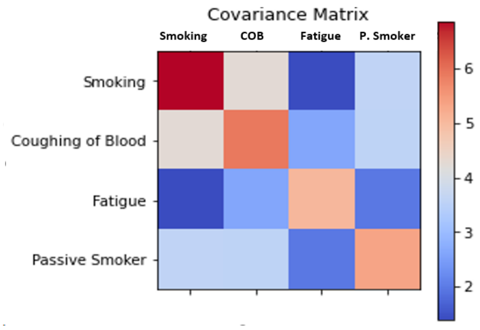


Fig. 3.5. Covariance Matrix for lifestyle parameters

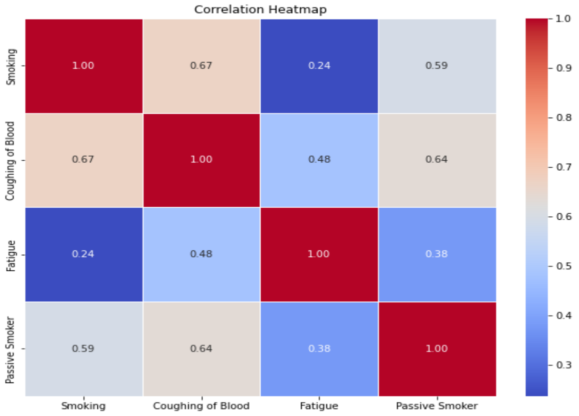
Table 5 presents a covariance table for top 4 lifestyle parameters and fig. 3.5 shows a Covariance for the top four lifestyle parameters for Lung cancer.

The positive covariance observed between lifestyle parameters associated with lung cancer suggests a tendency for these factors to vary together. A positive covariance indicated in fig.3.5 that as one lifestyle parameter increases or decreases, there is a corresponding tendency for the others to also increase or decrease, highlighting a potentially interconnected relationship among these factors in contributing to the risk of developing lung cancer.

d. Correlation is a standardized measure of the linear relationship between two variables. It tells us not only about the direction of the relationship (positive or negative) but also about the strength of the relationship. Correlation is always between -1 and 1, where 1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 indicates no linear relationship.

Fig. 3.6 presents a heatmap for the top four lifestyle parameters associated with cancer. According to the analysis, all four parameters exhibit positive correlations. This means that when one parameter shows a high value, there tends to be a corresponding increase in the values of the other parameters. In summary, the findings suggest a consistent positive correlation among all the lifestyle parameters studied.

Fig 3.6: Correlation heatmap for lifestyle parameters



3.1.5.3 Parameters Cause for Lung Cancer (Gender wise)

Lung cancer prediction dataset stores total 598 Female Lung cancer patients' data and 402 Male Lung cancer patients' data.

Gender wise analysis is performed and as per the results shown in fig. 3.7 it is observed that there is a strong association between smoking and the development of lung cancer specifically in male patients and Fig 3.8 shows that the Peer pressure, alcohol consumption and swallowing difficulty are the main causes of lung cancer in Female patients.

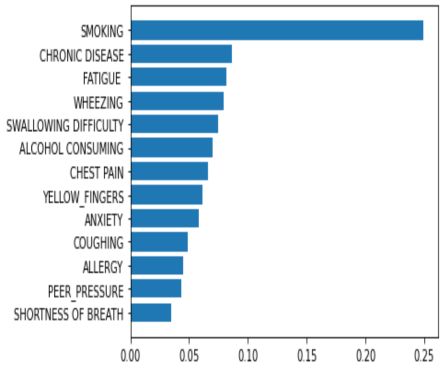


Fig 3.7. Parameters cause for lung cancer in male Patients

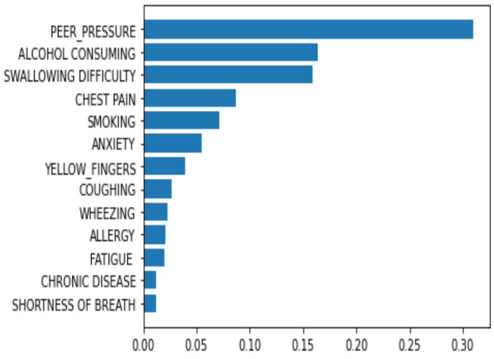


Fig 3.8. Parameters cause for lung cancer in female Patients

3.1.5.4 Parameters cause for lung cancer (Age wise)

The entire dataset is categorized into three age groups:

Young: Encompassing individuals aged 14 to 40 years.

Middle age: Including individuals aged 41 to 60 years.

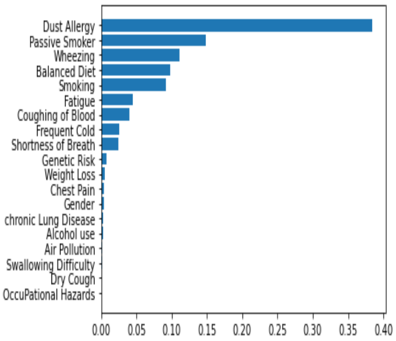


Fig 3.9: Parameters cause for lung cancer for young age

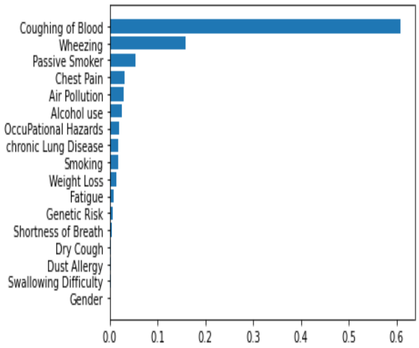


Fig 3.10: Parameters cause for lung cancer for middle age

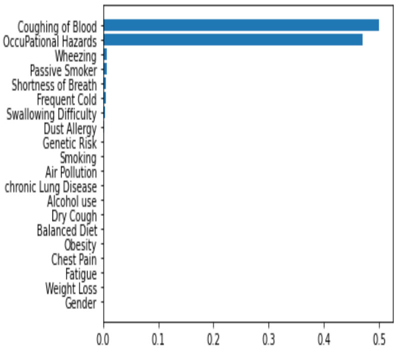


Fig 3.11: Parameters cause for lung cancer for old age

Old age: Encompassing individuals aged 61 to 80 years.

As shown in fig 3.9 among young individuals at risk of or diagnosed with lung cancer, dust allergy and passive smoking are

significant predictors or factors associated with the Lung cancer disease.

The findings, as presented in Figure 3.10, suggest that among middle-aged individuals diagnosed with lung cancer, coughing of blood and wheezing are notably prevalent symptoms or indicators of the disease. Among elderly individuals diagnosed with lung cancer, coughing of blood and wheezing are notably prevalent symptoms or indicators of the disease as shown in the fig 3.11.

4. Results and Discussion:

4.1 Results of Lung cancer prediction model:

Lifestyle factors play an essential role in the development of lung cancer. Utilizing machine learning techniques, it is possible to predict lung cancer well in advance by analyzing these lifestyle parameters. As depicted in table 6.

Table 6.: Performance Measures for lung cancer prediction model

Sr. No	Prediction techniques	Accuracy
1	Logistic Regression	94%
2	Support Vector Machine	90%

Two machine Learning techniques are applied on Lung cancer prediction dataset as logistic regression and support vector machine for lung cancer prediction. Notably, the Logistic regression technique demonstrated the highest accuracy in making these predictions.

4.2 Results of Lung cancer level prediction model:

Exploratory data analysis and preprocessing techniques were utilized to enhance the predictive accuracy of the study. By carefully examining and preparing the data, this study effectively predicted the risk level of lung cancer using twenty-two unique lifestyle factors. Six diverse classification models were employed, and their performance was thoroughly evaluated and compared. The results, depicted in Table 7 highlight the superior performance of all prediction techniques in accurately predicting the risk level of lung cancer.

As depicted in table 7, each machine learning model is evaluated based on several performance metrics such as accuracy, precision, recall, F1-score, ROC-AUC, and Matthew's correlation coefficient. SVM, Random Forest, and KNN exhibit similar performance with high accuracy, precision, recall, F1-score, and ROC-AUC values of 98%, while Linear Regression shows a perfect correlation coefficient of 0.98, suggesting strong predictive capability. The performance measures in the form of line chart is shown in fig.4.1.

Table 7: Performance Measures for Lung cancer level prediction Model

ML Model	Accuracy	Precision	Recall	Specificity	F1-Score	G-mean	MCC	ROC-AUC	R^2
SVM	98%	98%	98%	50%	97%	0.7	0.97	98%	
Random Forest	96%	96%	96%	50%	96%	0.69	0.94	98%	
KNN	97%	97%	97%	50%	97%	0.69	0.95	98%	
Linear Regression									0.98

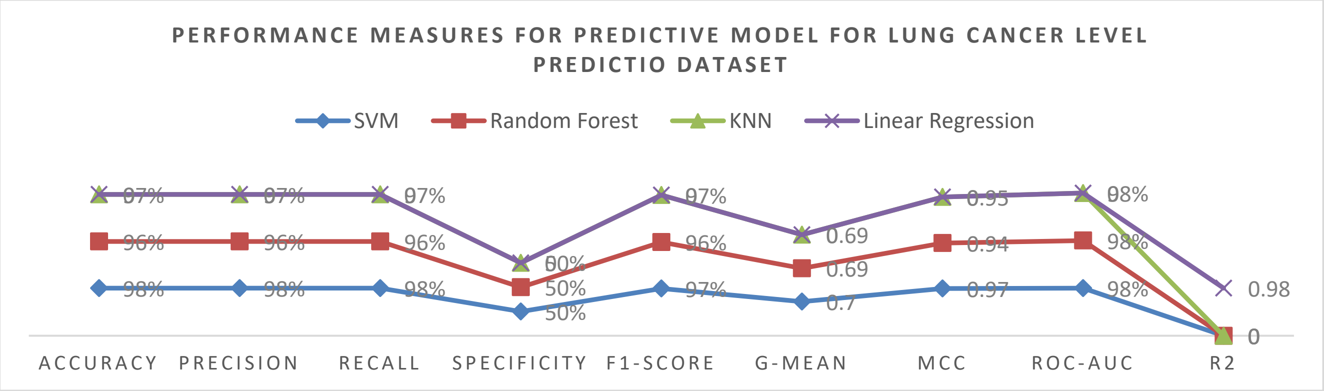


Fig. 4.1 Line chart for Lung cancer level prediction models

Calculation all measures is given below:

Accuracy: Accuracy is a simple and intuitive metric that measures the overall correctness of a model's predictions. Accuracy calculation is based on equation 4.

Accuracy= (TP + TN) / (TP + TN+ FP+FN) (4)

Precision: The proportion of true "Yes" instances out of all instances predicted as "Yes". Precision calculation is based on equation 5.

Precision: TP / (TP + FP) (5)

Recall: The proportion of true "Yes" instances correctly identified by the model out of all actual "Yes" instances. Recall calculation is based on equation 6.

Recall = TP / TP+FN (6)

F1-score:The F1 score is a harmonic mean of precision and recall, offering a balance between these two metrics. It is particularly useful when you want to consider both false positives and false negatives and want a single metric to evaluate model performance. F1-score calculation is based on equation 7.

F1-score = 2 * ((Precision * Recall) / (Precision + Recall)) (7)

G-Mean (Geometric Mean):

G-Mean calculates the geometric mean of sensitivity and specificity. It's useful in evaluating classification models, especially in imbalanced datasets, as it takes into account both the positive and negative classes.

The formula for G-Mean is shown in equation 8.

$$G - Mean = \sqrt{Sensitivity\ \times Specificity} \qquad (8)$$

Where:

- **Sensitivity** (also known as recall) is the true positive rate, i.e.,

  Sensitivity = TP / TP+FN (9)

- **Specificity** is the true negative rate, i.e.,

  Specificity= TN / TN+FP (10)

G-Mean ranges from 0 to 1, where a value closer to 1 indicates a better model performance in terms of both sensitivity and specificity.

The geometric mean of sensitivity (recall) and specificity. It provides an overall measure of a classifier's performance across both positive and negative classes.

Matthews Correlation Coefficient (MCC)

he Matthews Correlation Coefficient (MCC) is a unified metric that considers both true and false positives and negatives, making it a well-regarded measure suitable for scenarios with highly imbalanced class sizes. Its values span from -1, indicating complete discordance between predictions and actuals, to 1, signifying perfect alignment. A score of 0 denotes random prediction performance. The formula for MCC is shown in equation 11.

$$MCC = \frac{(TP\times TN)-(FP\times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN}} \qquad (11)$$

Where:

TP=TP_L+TP_M+TP_H

TN=TN_L+TN_M+TN_H

FP=FP_L+FP_M+FP_H

FN=FN_L+FN_M+FN_H

MCC is generally considered a good measure for binary classification problems, especially when the classes are imbalanced.

Value of MCC ranges from negative 1 to positive 1, where positive 1 indicates accurate prediction, 0 indicates prediction as random, and -1 indicates total discrepancy between prediction and observation.

Receiver Operating Characteristic (ROC)

In the realm of machine learning, the Receiver Operating Characteristic (ROC) curve and its associated metric, the Area Under the Curve (AUC), serve as crucial evaluation tools for binary classification models.

The ROC curve is a visual representation showcasing the balance between a model's true positive rate (sensitivity) and false positive rate (1 - specificity) across various threshold values. By plotting true positive rate against false positive rate at different thresholds, it offers insights into how well a model can distinguish between positive and negative instances. Models achieving higher true positive rates and lower false positive rates will exhibit ROC curves closer to the top-left corner, indicating superior performance.

The AUC summarizes the overall performance of a model by computing the area under its corresponding ROC curve. A perfect classifier yields an AUC score of 1, while a random classifier scores 0.5. Higher AUC values denote greater discriminatory ability in separating positive and negative instances. For instance, an AUC of 0.95 signifies outstanding predictive prowess, suggesting substantial distinction between positive and negative cases.

4.3 Results of predictive model for Primary dataset

Table 8 displays the SVM model performance for primary dataset. There are two different primary dataset results displayed in table. The first dataset is the collection of lifestyle data from cancer patients encompassing Lifestyle information of 21 patients. Additionally, a substantial dataset consisting of lifestyle data from 261 cancer patients was sourced from Rasayu Ayurvedic Cancer Hospital, Pune.

The SVM model is applied to a primary dataset consisting of data from 21 patients.

The overall prediction accuracy for this model is determined to be 93%. The SVM technique is applied to the augmented primary dataset (Rasayu Patients' dataset). The resulting model achieves an overall accuracy as 79%.

Table 8. Primary data performance Measures

Sr. No.	Dataset	Prediction Techniques	Performance
1	Hospital Dataset as unseen data for an improved model	Support Vector Machine	93%
2	Rasayu Hospital Patients' Dataset	Support Vector Machine	79%

4.4 Results for AI models

The Generative AI model analyzes diverse patient data, generating a comprehensive health summary that evaluates current health status, identifies risks, and suggests tailored preventive measures. Emphasizing lifestyle factors linked to lung cancer, such as smoking habits and genetic markers, it offers personalized recommendations like smoking cessation and dietary changes for risk reduction. The visual health summary highlights areas of concern and provides clear preventive action guidance, enhancing overall health through individualized data analysis.

Contrarily, the Explainable AI model's output, facilitated by LIME, clarifies machine learning predictions by breaking down each feature's contribution to the model's decision-making process. Positive coefficients indicate feature values that increase prediction likelihood, while negative coefficients signify the opposite. This transparency fosters trust and comprehension in AI-driven decision support systems, empowering stakeholders to understand the factors influencing predictions effectively within a concise and interpretable framework.

5. Conclusion

The review evaluates various Supervised Learning (Classification) and Unsupervised Learning (Clustering) techniques in predicting cancer, with Support Vector Machine (SVM) and Artificial Neural Network (ANN) showing superior accuracy. Supervised methods are favored due to their ability to use labeled data effectively. A study successfully predicted lung cancer risk using twenty-two lifestyle factors, favoring SVM, K-Nearest Neighbors (KNN), and Random Forest for their high accuracy. Machine learning's role in accurately predicting lung cancer risk is highlighted, aiding in identifying high-risk individuals and implementing preventive measures. Lifestyle factors are recognized as crucial in lung cancer development, with machine learning achieving over 90% accuracy in predicting based on lifestyle parameters. This facilitates early prediction and identification of lifestyle contributors, empowering individuals to make informed decisions and mitigate risks. The research underscores machine learning's potential in early detection and prevention of lung cancer, suggesting avenues for future research and clinical applications.

Acknowledgement

I would like to express my profound gratitude to my PhD Supervisor Dr. Vishal Shirsath and Co-Supervisor Dr. Farook Sayyad for their constant motivation and guidance. I would also like to thank to Dr. Yogesh Bendale and Dr. Poonam Birari of **Rasayu Cancer Care Hospital**, Pune for providing lung cancer patients' data and for providing me the necessary inputs required for the research work.

References:

[1] Nair B., Jeevakumar, A., & Anju, K. (n.d.). *Tobacco Smoking Induced Lung Cancer Prediction By LC-MicroRNAs Secondary Structure Prediction And Target Comparison. 2017 2nd International Conference for Convergence in Technology (I2CT).*

[2] Bostean G., Crespi, C. M., & McCarthy, W. J. (8 2013). Associations among family history of cancer, cancer screening and lifestyle behaviors: A population-based study. *Cancer Causes and Control*, *24*, 1491–1503. doi:10.1007/s10552-013-0226-9

[3] Jeon J., Du, M., Schoen, R. E., Hoffmeister, M., Newcomb, P. A., Berndt, S. I., … Hsu, L. (6 2018). Determining Risk of Colorectal Cancer and Starting Age of Screening Based on Lifestyle, Environmental, and Genetic Factors. *Gastroenterology*, *154*, 2152-2164.e19. doi:10.1053/j.gastro.2018.02.021

[4] Carr P. R., Weigl, K., Edelmann, D., Jansen, L., Chang-Claude, J., Brenner, H., & Hoffmeister, M. (7 2020). Estimation of Absolute Risk of Colorectal Cancer Based on Healthy Lifestyle, Genetic Risk, and Colonoscopy Status in a Population-Based Study. *Gastroenterology*, *159*, 129-138.e9. doi:10.1053/j.gastro.2020.03.016

[5] Pati J. (2019). Gene expression analysis for early lung cancer prediction using machine learning techniques: An eco-genomics approach. *IEEE Access*, *7*, 4232–4238. doi:10.1109/ACCESS.2018.2886604

[6] Aleksandrova K., Reichmann, R., Kaaks, R., Jenab, M., Bueno-de-Mesquita, H. B., Dahm, C. C., … Gunter, M. J. (12 2021). Development and validation of a lifestyle-based model for colorectal cancer risk prediction: the LiFeCRC score. *BMC Medicine*, *19*. doi:10.1186/s12916-020-01826-0

[7] Chen, H., Liu, L., Lu, M., Zhang, Y., Lu, B., Zhu, Y., … Dai, M. (7 2021). Implications of Lifestyle Factors and Polygenic Risk Score for Absolute Risk Prediction of Colorectal Neoplasm and Risk-Adapted Screening. *Frontiers in Molecular Biosciences*, *8*. doi:10.3389/fmolb.2021.685410

[8] Nii M., Momimoto, M., Kobashi, S., Kamiura, N., Hata, Y., & Sorachi, K. I. (3 2016). *Medical Checkup and Image Data Analysis for Preventing Life Style Diseases: A Research Survey of Japan Society for the Promotion of Science with Grant-in-Aid for Scientific Research (A) (Grant number 25240038)*. *2016-March*, 117–122. doi:10.1109/ICETET.2015.38

[9] Prachiti Gholap, V. P. A. P. (n.d.). DiseaseLens: A Lifestyle related Disease Predictor. *Published in: 2022 5th International Conference on Advances in Science and Technology (ICAST) Date of Conference: 02-03 December 2022 Date Added to IEEE Xplore: 13 February 2023*.

[10] Liao W., Coupland, C. A. C., Burchardt, J., & Baldwin, D. R. (n.d.). *Predicting the future risk of lung cancer: development, and internal and external validation of the CanPredict (lung) model in 19·67 million people and evaluation of model performance against seven other risk prediction models*.

[11] Xie P., Huang, X., Lin, D., Huang, X., Lin, S., Luo, S., … Weng, X. (n.d.). Long-term trend of future Cancer onset: A model-based prediction of Cancer incidence and onset age by region and gender. *Preventive Medicine Volume 177, December 2023, 107775*.

[12] Yang J., Wen, W., Zahed, H., & Zheng, W. (n.d.). Lung Cancer Risk Prediction Models for Asian Ever-Smokers. *Journal of Thoracic Oncology Volume 19, Issue 3, March 2024, Pages 451-464, Https://Doi. Org/10. 1016/j. Jtho. 2023. 11. 002*.

[13] Hui, L. (2024). Changes in threats from chronic obstructive pulmonary disorder and lung cancer with environmental improvements in China: Quantitative evaluation and prediction based on a model with age as a probe. *Heliyon Journal Homepage: Www. Cell. Com/Heliyon,April 01, 2024,DOI:Https://Doi. Org/10. 1016/j. Heliyon. 2024. E28977*.

[14] Zhao D., Lu, J., Zeng, W., Zhang, C., & You, Y. (n.d.). Changing trends in disease burden of lung cancer in China from 1990-2019 and following 15-year prediction. *Current Problems in Cancer Volume 48, February 2024, 101036, Https://Doi. Org/10. 1016/j. Currproblcancer. 2023. 101036*.

[15] A., M., Zulkifley, M. A., & Zainuri, M. A. A. M. (n.d.). A Review of Deep Learning Techniques for Lung Cancer Screening and Diagnosis Based on CT Images. *S. Diagnostics 2023, 13, 2617. Https://Doi. Org/10. 3390/ Diagnostics13162617*.

[16] Endalie D., & Abebe, W. T. (7 2023). Analysis of lung cancer risk factors from medical records in Ethiopia using machine learning. *PLOS Digital Health*, *2*, e0000308. doi:10.1371/journal.pdig.0000308

[17] Mohan K., & Thayyil, B. (9 2023). Machine Learning Techniques for Lung Cancer Risk Prediction using Text Dataset. *International Journal of Data Informatics and Intelligent Computing*, Vol. 2, pp. 47–56. doi:10.59461/ijdiic.v2i3.73

[18] Howell D., Analytica, Q., Buttery, R., Badrinath, P., George, A., Council, K. C., … Finnis, C. (2023). *Developing a risk prediction tool for Lung Cancer in Kent and Medway, England: Cohort Study using linked Data*. doi:10.21203/rs.3.rs-3100044/v1

[19] Shaoo P., Omrah, M., Somit, D., & Kumar, R. (10 2022). Lung Cancer Prediction Using Machine Learning and Big Data. *International Journal of Scientific Research in Computer Science Engineering and Information Technology*, *9*.

[20] Li, H., Cheng, Z. J., Liang, Z., Liu, M., Liu, L., Song, Z., … Sun, B. (1 2023). Novel nutritional indicator as predictors among subtypes of lung cancer in diagnosis. *Frontiers in Nutrition*, *10*. doi:10.3389/fnut.2023.1042047

[21] Abuya, T. K. (2023). Lung Cancer Prediction from Elvira Biomedical Dataset Using Ensemble Classifier with Principal Component Analysis. *Journal of Data Analysis and Information Processing*, *11*, 175–199. doi:10.4236/jdaip.2023.112010

[22] Idrissi S. E., Ben, I., Ouahab, A., Drider, Y., Bouhorma, M., & Ouaai, E. L. (2023). Prediction Of Lung Cancer Levels Based On Patient Lifestyle And Histopathological Images Using Artificial Intelligence. *Journal of Theoretical and Applied Information Technology*, Vol. 15. Retrieved from www.jatit.org

[23] Yeo Y., Shin, D. W., Han, K., Park, S. H., Jeon, K. H., Lee, J., Kim, J., & Shin, A. (2021). Individual 5-year lung cancer risk prediction model in korea using a

nationwide representative database. *Cancers*, *13*(14), Article 3496. https://doi.org/10.3390/cancers13143496

[24] Sim J., Kim, Y. A., Kim, J. H., Lee, J. M., Kim, M. S., Shim, Y. M., … Yun, Y. H. (12 2020). The major effects of health-related quality of life on 5-year survival prediction among lung cancer survivors: applications of machine learning. *Scientific Reports*, *10*. doi:10.1038/s41598-020-67604-3

[25] Anita Nath, Krishnan Sathishkumar, Priyanka Das, Kondalli Lakshminarayana Sudarshan, Prashant Mathur (2022). A clinic epidemiological profile of lung cancers in India – Results from the National Cancer Registry Program, Indian J Med Res 155, February 2022, pp 264-272, DOI: 10.4103/ijmr.ijmr_1364_21