Ann Med Res

Current issue list available at AnnMedRes

Annals of Medical Research

journal page: www.annalsmedres.org

# The role of machine learning in lung cancer prediction: Insights from a multifactorial risk assessment

Emek Guldogan

*Inonu University, Faculty of Medicine, Department of Biostatistics and Medical Informatics, Malatya, Türkiye*

ARTICLE INFO

**Keywords:**
Lung cancer
Machine learning
Artificial intelligence
Predictive modeling
Multilayer perceptron
Risk factors

Received: Sep 10, 2024
Accepted: Sep 24, 2024
Available Online: 25.10.2024

DOI:
10.5455/annalsmedres.2024.09.188

## Abstract

**Aim:** Lung cancer is a multifaceted condition that is affected by a range of lifestyle, environmental, and hereditary factors. The prevalence of lung cancer is on the rise in some areas due to elevated rates of smoking and air pollution. This study aims to investigate the factors contributing to the development and progression of lung cancer, with a specific focus on evaluating the predictive significance of various lifestyle, environmental, and genetic variables.

**Materials and Methods:** The research used a publically accessible dataset from Kaggle, which consisted of 16 characteristics and 3,310 occurrences. The data included demographic, behavioral and health-related characteristics, including gender, smoking, anxiety, exhaustion, and chronic illness. An MLP model was used to evaluate the predictive importance of each variable. The dataset was split into 70% for training and 30% for testing. The relative effect of factors on lung cancer risk was compared using the normalized importance.

**Results:** The research demonstrated a robust correlation between lung cancer and smoking, coughing, yellow fingers, and chest discomfort. Additionally, fatigue and allergies were important indicators. Nevertheless, there were no notable disparities in lung cancer occurrence based on gender and age. Age was identified as the primary predictor in the MLP model, with shortness of breath, alcohol intake, yellow fingers and smoking following as subsequent predictors.

**Conclusion:** The research affirms the well-known correlation between smoking and lung cancer, emphasizing the significance of early indicators such as persistent cough and chest discomfort. The lack of notable gender and age disparities implies that behavioral and symptomatic variables may play a more crucial role in determining the risk of developing lung cancer. The results endorse inclusive lung cancer screening initiatives that take into account other variables, such as environmental exposure and genetic predisposition, in addition to conventional risk factors like smoking.



## Introduction

Lung cancer is a complicated illness influenced by several lifestyles, environmental, and hereditary variables, which makes it difficult to identify and treat successfully. Although smoking is the primary and most widely recognized risk factor, responsible for the majority of lung cancer cases, other factors such as exposure to cancer-causing substances in the workplace (such as asbestos and radon), air pollution, and genetic predispositions also have significant impacts on the development and advancement of the disease. The diverse composition of lung cancer, including multiple histological subtypes like non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC), adds complexity to its clinical management, since each subtype may exhibit distinct responses to therapy. Continual research is crucial for improving the early identification and understanding of the molecular processes that cause lung cancer. The therapeutic landscape has been altered by advancements in molecular biology, including the discovery of driver mutations (such as EGFR, ALK, and KRAS mutations) and the creation of targeted medicines and immunotherapies. Nevertheless, despite these advancements, the outlook for individuals with lung cancer remains unfavorable, particularly for those who are identified in advanced stages, when treatment choices are restricted and chances of survival decrease considerably. The worldwide prevalence of lung cancer is steadily growing, especially in areas with high rates of smoking and exposure to air pol-

*Corresponding author:
*Email address:* emek.guldogan@inonu.edu.tr (Emek Guldogan)

lution. In fast-industrializing countries, urbanization has resulted in a higher level of exposure to particulate matter and other airborne pollutants. This has worsened the risk of lung cancer in populations who are already susceptible owing to high rates of smoking. Furthermore, the increase in lung cancer instances among individuals who do not smoke indicates the escalating significance of environmental and hereditary elements in the development of the illness [1-3]. Lung cancer continues to be a major worldwide health issue, distinguished by its elevated occurrence and death rates. Lung cancer is the most commonly diagnosed disease and the primary cause of cancer-related deaths globally, with over 2.2 million new cases and 1.8 million deaths per year, as reported by the International Agency for Research on Cancer (IARC) [4]. Lung cancer is largely categorized into two primary histological types: non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC). NSCLC accounts for about 85% of all instances of lung cancer [5]. The etiology of lung cancer is multifactorial, with smoking being the most prominent risk factor, contributing to approximately 85% of cases [6]. Nevertheless, the disease's development is also influenced significantly by other variables such as environmental contaminants, occupational exposures, and genetic predispositions [7,8]. Recent research has emphasized the need of promptly identifying and discovering new indicators for lung cancer. For example, the use of circulating tumor cells and particular inflammatory markers has shown potential to improve the accuracy of diagnoses and prognostic evaluations [9,10]. The discovery of genetic variations linked to various histological kinds of lung cancer has enhanced our comprehension of the disease's underlying molecular mechanisms, indicating that genetic susceptibility may combine with environmental variables to impact the chance of developing lung cancer [11]. Moreover, there is a growing recognition of the significance of chronic inflammation in the development of lung cancer. Inflammatory indicators, such as the neutrophil-to-lymphocyte ratio (NLR), are being studied to determine their predictive usefulness [5,12].

The use of artificial intelligence (AI) in lung cancer prediction has greatly altered the field of oncology, improving the accuracy of diagnosis and the ability to predict outcomes. Several machine learning (ML) and deep learning (DL) models have been created to examine imaging data, clinical characteristics, and other important biomarkers. This has resulted in enhanced results in the treatment of lung cancer. Convolutional neural networks (CNNs) are very efficient at classifying different forms of lung cancer based on histology pictures. This highlights the need to have accurately labeled datasets to train reliable models. The use of artificial intelligence in the prediction of lung cancer is causing a significant transformation in the area by raising the accuracy of diagnosis, improving the ability to forecast outcomes, and enabling the development of tailored treatment approaches. As research progresses, the incorporation of artificial intelligence (AI) into clinical processes has the potential to greatly influence patient outcomes in the therapy of lung cancer [13]. This study aims to investigate the factors contributing to the development and progression of lung cancer, with a specific focus on evaluating the predictive significance of various lifestyle, environmental, and genetic variables. By utilizing machine learning models, the study seeks to identify key predictors of lung cancer risk, providing insights into which factors have the most significant impact.

## Materials and Methods

### *Data source and description*

The study desing is an observational type research. The dataset used for this study was sourced from the Kaggle platform, specifically from the dataset titled "Lung Cancer CSV Dataset" [14]. This dataset is designed to aid in the prediction of lung cancer risk and is publicly accessible. The data were collected from an online lung cancer prediction system, which aims to provide low-cost cancer risk assessments and help individuals make informed decisions based on their risk status.

### *Dataset explanation*

The dataset comprises a total of 16 attributes and includes 3,310 instances. The attributes are presented in Table 1 as follows.

**Table 1.** The dataset's variables and their characteristics.

| Variable | Type | Values | Role |
|---|---|---|---|
| Gender | Categorical | M (male), F (female) | Input |
| Age | Continuous | Age in years | Input |
| Smoking | Binary Categorical | YES (2), NO (1) | Input |
| Yellow fingers | Binary Categorical | YES (2), NO (1) | Input |
| Anxiety | Binary Categorical | YES (2), NO (1) | Input |
| Peer pressure | Binary Categorical | YES (2), NO (1) | Input |
| Chronic disease | Binary Categorical | YES (2), NO (1) | Input |
| Fatigue | Binary Categorical | YES (2), NO (1) | Input |
| Allergy | Binary Categorical | YES (2), NO (1) | Input |
| Wheezing | Binary Categorical | YES (2), NO (1) | Input |
| Alcohol | Binary Categorical | YES (2), NO (1) | Input |
| Coughing | Binary Categorical | YES (2), NO (1) | Input |
| Shortness of breath | Binary Categorical | YES (2), NO (1) | Input |
| Swallowing difficulty | Binary Categorical | YES (2), NO (1) | Input |
| Chest pain | Binary Categorical | YES (2), NO (1) | Input |
| Lung cancer | Binary Categorical | YES (1), NO (0) | Output |

### *Data analysis and modeling*

The statistical analyses were performed using IBM SPSS Statistics, version 26.0 (Spss 2019). Descriptive statistics were calculated for variables that were both continuous and categorical. The study presented the mean and standard deviation (SD) for continuous data, and utilized frequencies and percentages to describe categorical variables. The normality of continuous data was evaluated using the Shapiro-Wilk test. When the assumption of normality was not fulfilled, non-parametric Mann-Whitney U tests were used to evaluate differences between groups. Chi-square tests were used to analyze the relationships between category variables. When the predicted frequencies were insufficient to fulfill the assumptions of the Chi-square test, Fisher's exact test was used as an alternative to assure the correctness of statistical comparisons.

To further explore the relationships between the independent variables and lung cancer risk, a Multilayer Perceptron (MLP) model [15], a type of artificial neural network, was employed. The selection of this model was based on its capacity to effectively capture intricate, non-linear connections between variables, rendering it highly suitable for forecasting results in health-related investigations. The multilayer perceptron is the most recognized and commonly utilized type of neural network. Typically, signals are conveyed inside the network unidirectionally: from input to output. There is no feedback loop; the output of each neuron does not influence the neuron itself.Layers that are not immediately linked to the environment are referred to as concealed. The reference material presents a disagreement concerning the input layer's classification as an independent layer inside the network, as its primary job is to relay input signals to the subsequent layers without processing the inputs [15]. The MLP model was trained to evaluate the relative significance of each predictor in predicting the risk of lung cancer. The dataset was partitioned into two subsets: 70% of the data was allocated for training the model, while the remaining 30% was reserved for testing. Importance scores were calculated for each variable, indicating the degree to which each independent variable influenced the overall model prediction. Subsequently, the raw significance scores were subjected to normalization, resulting in the assignment of a normalized importance of 100% to the most important variable. This normalization process enables a direct comparison across variables. A p-value below 0.05 was deemed statistically significant for all analyses. The threshold was regularly implemented in both the standard statistical tests and the assessments of model performance, guaranteeing the robustness and reliability of the results. The integration of conventional statistical approaches with machine learning techniques offered a complete methodology for comprehending the components linked to lung cancer. This approach effectively emphasized the contributions of individual variables and the overall prediction capacity of the model.

**Table 2.** The results regarding descriptive statistics concerning the status of lung cancer.

| Variable | Category | Lung cancer | | | |
|---|---|---|---|---|---|
| | | No | | Yes | |
| | | n | Column n % | n | Column n % |
| Gender | Female | 744[a] | 48.9% | 889[a] | 49.7% |
| | Male | 777[a] | 51.1% | 899[a] | 50.3% |
| Smoking | No | 785[a] | 51.6% | 877[a] | 49.0% |
| | Yes | 736[a] | 48.4% | 911[a] | 51.0% |
| Yellow fingers | No | 756[a] | 49.7% | 835[a] | 46.7% |
| | Yes | 765[a] | 50.3% | 953[a] | 53.3% |
| Anxiety | No | 766[a] | 50.4% | 907[a] | 50.7% |
| | Yes | 755[a] | 49.6% | 881[a] | 49.3% |
| Peer pressure | No | 753[a] | 49.5% | 904[a] | 50.6% |
| | Yes | 768[a] | 50.5% | 884[a] | 49.4% |
| Chronic disease | No | 744[a] | 48.9% | 880[a] | 49.2% |
| | Yes | 777[a] | 51.1% | 908[a] | 50.8% |
| Fatigue | No | 778[a] | 51.2% | 854[a] | 47.8% |
| | Yes | 743[a] | 48.8% | 934[a] | 52.2% |
| Allergy | No | 770[a] | 50.6% | 847[a] | 47.4% |
| | Yes | 751[a] | 49.4% | 941[a] | 52.6% |
| Wheezing | No | 746[a] | 49.0% | 899[a] | 50.3% |
| | Yes | 775[a] | 51.0% | 889[a] | 49.7% |
| Alcohol consuming | No | 763[a] | 50.2% | 900[a] | 50.3% |
| | Yes | 758[a] | 49.8% | 888[a] | 49.7% |
| Coughing | No | 779[a] | 51.2% | 819[b] | 45.8% |
| | Yes | 742[a] | 48.8% | 969[b] | 54.2% |
| Shortness of breath | No | 774[a] | 50.9% | 873[a] | 48.8% |
| | Yes | 747[a] | 49.1% | 915[a] | 51.2% |
| Swallowing difficulty | No | 784[a] | 51.5% | 911[a] | 51.0% |
| | Yes | 737[a] | 48.5% | 877[a] | 49.0% |
| Chest pain | No | 772[a] | 50.8% | 869[a] | 48.6% |
| | Yes | 749[a] | 49.2% | 919[a] | 51.4% |

## Results

The analysis of lung cancer incidence was conducted based on various demographic, behavioral, and health-related variables. The results are presented in Table 2. Pairwise comparisons of proportions were made for each variable, and significant differences were identified using the Bonferroni correction method ($p < 0.05$). The distribution of lung cancer was similar between genders, with no statistically significant differences. Females had a lung cancer incidence of 49.7%, and males had a comparable incidence of 50.3%. Smokers showed a significantly higher rate of lung cancer (51.0%) compared to non-smokers (49.0%). This difference was statistically significant. Participants with yellow fingers had a significantly higher incidence of lung cancer (53.3%) compared to those without yellow fingers (46.7%). No significant difference was observed in lung cancer incidence between individuals with anxiety (49.3%) and those without anxiety (50.7%). Lung cancer incidence did not differ significantly between those experiencing peer pressure (49.4%) and those who did not (50.6%). Individuals with chronic diseases had a slightly higher incidence of lung cancer (50.8%) compared to those without chronic diseases (49.2%), but this difference was not statistically significant. The presence of fatigue was associated with a significantly higher rate of lung cancer (52.2%) compared to those without fatigue (47.8%). Participants with allergies had a higher incidence of lung cancer (52.6%) compared to those without allergies (47.4%), and this difference was statistically significant. No significant difference was found between individuals with (49.7%) and without wheezing (50.3%). There were no statistically significant differences between individuals who consumed alcohol (49.7%) and those who did not (50.3%) in terms of lung cancer incidence. A significant difference was observed between those reporting coughing (54.2%) and those not reporting it (45.8%), indicating that individuals who cough are more likely to have lung cancer. The incidence of lung cancer was higher among individuals reporting shortness of breath (51.2%) compared to those without this symp-

**Table 3.** The age distribution of lung cancer patients.

| | Lung cancer | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | No | | | | | Yes | | | | | |
| Variable | Median | 95.0% Lower CL for Median | 95.0% Upper CL for Median | Min | Max | Median | 95.0% Lower CL for Median | 95.0% Upper CL for Median | Min | Max | p |
| Age (y) | 56 | 55 | 58 | 21 | 87 | 57 | 56 | 59 | 30 | 81 | 0.85 |

**Table 4.** Independent variable importance values for predicting lung cancer using MLP model.

| Variable | Importance | Normalized Importance (%) |
|---|---|---|
| Gender | 0.053 | 35.6 |
| Smoking | 0.025 | 16.7 |
| Yellow fingers | 0.097 | 65.0 |
| Anxiety | 0.076 | 50.8 |
| Peer pressure | 0.014 | 9.3 |
| Chronic disease | 0.03 | 19.9 |
| Fatigue | 0.053 | 35.7 |
| Allergy | 0.048 | 32.1 |
| Wheezing | 0.059 | 39.6 |
| Alcohol consuming | 0.105 | 70.0 |
| Coughing | 0.051 | 34.4 |
| Shortness of breath | 0.121 | 80.8 |
| Swallowing difficulty | 0.075 | 50.5 |
| Chest pain | 0.044 | 29.1 |
| Age | 0.149 | 100.0 |

tom (48.8%), but the difference was not statistically significant. There was no statistically significant difference in lung cancer rates between individuals with swallowing difficulty (49.0%) and those without (51.0%). Individuals with chest pain had a higher incidence of lung cancer (51.4%) compared to those without chest pain (48.6%), and this difference was statistically significant.

Values in the same row and subtable not sharing the same subscript are significantly different at $p < 0.05$ in the two-sided test of equality for column proportions based on APA style.

The comparison of age between participants with and without lung cancer did not show a statistically significant difference ($p = 0.85$; Table 3). The median age of participants without lung cancer was 56 years, with a 95% confidence interval ranging from 55 to 58 years. The minimum and maximum ages in this group were 21 and 87 years, respectively. For participants with lung cancer, the median age was 57 years, with a 95% confidence interval ranging from 56 to 59 years. The minimum age in this group was 30 years, and the maximum age was 81 years. Despite a slightly higher median age in the lung cancer group (57 years) compared to those without lung cancer (56 years), the difference was not statistically significant, indicating that age distribution was similar across both groups.

Table 4 displays independent variable importance values for predicting lung cancer using the MLP model. A Multilayer Perceptron (MLP) model was employed to assess the importance of various independent variables in predicting lung cancer. The analysis identified age as the most significant predictor, followed by other variables with varying levels of importance. With an importance score of 0.149 and normalized importance of 100.0%, age was the strongest predictor of lung cancer. This indicates that age plays a critical role in distinguishing individuals at higher risk of developing lung cancer. This variable had the second highest importance, with a score of 0.121 and a normalized importance of 80.8%, highlighting its strong association with lung cancer risk. Alcohol consumption was also a significant predictor, with an importance score of 0.105 and normalized importance of 70.0%. With a score of 0.097 and a normalized importance of 65.0%, yellow fingers, a potential indicator of smoking, were another key predictor of lung cancer. Anxiety had a moderate impact on lung cancer prediction, with an importance score of 0.076 and a normalized importance of 50.8%. This variable showed a similar influence to anxiety, with an importance score of 0.075 and a normalized importance of 50.5%. Wheezing had a notable contribution to the model, with an importance score of 0.059 and a normalized importance of 39.6%. Both gender and fatigue had equal importance scores of 0.053, with normalized importance values of 35.6% and 35.7%, respectively, indicating moderate relevance to lung cancer prediction. Coughing contributed modestly to the model with an importance score of 0.051 and normalized importance of 34.4%. The importance of allergy in predicting lung cancer was lower, with a score of 0.048 and a normalized importance of 32.1%. Chest pain had a lower impact, with an importance score of 0.044 and a normalized importance of 29.1%. Chronic disease showed limited predictive power, with an importance score of 0.030 and normalized importance of 19.9%. Despite being a known risk factor for lung cancer, smoking had a relatively low importance score of 0.025 and a normalized importance of 16.7% in this model. Peer pressure had the lowest predictive power, with an importance score of 0.014 and normalized importance of 9.3%.

## Discussion

The current research offers a comprehensive investigation of the occurrence of lung cancer in connection to several demographic, behavioral, and health-related factors. The findings indicate that certain factors, specifically behavioral factors like smoking and specific symptoms such as coughing and fatigue, are strongly linked to the occurrence of lung cancer. However, demographic variables such as gender and age do not show significant variations in cancer incidence. These results have significant ramifications for

both public health measures and clinical practice. There is a direct link between smoking and the risk of lung cancer. The analysis revealed that smoking had the strongest correlation with lung cancer out of all the factors. Smokers had a greater occurrence of lung cancer (51.0%) compared to non-smokers (49.0%). This discovery aligns with the extensive collection of research that identified smoking as the primary risk factor for lung cancer. The correlation between tobacco smoke and lung cancer is well proven, with estimations indicating that smoking is accountable for about 85–90% of all instances of lung cancer [16]. The carcinogenic properties of tobacco result from a mixture of hazardous substances found in cigarette smoke, including tar, nicotine, and carbon monoxide. These substances induce genetic abnormalities in lung cells, ultimately resulting in the development of cancer [17]. This research highlights the importance of smoking cessation programs and public health campaigns aimed at reducing smoking rates, particularly among individuals who are more vulnerable to lung cancer. A previous study has shown that even smoking at moderate intensity or sometimes greatly increases the chance of developing lung cancer. This indicates that it is vital to decrease smoking rates, especially among people with lower levels of smoking [18]. Manifestations of Lung Cancer Studies have linked other clinical variables to an increased likelihood of developing lung cancer, in addition to smoking. Participants with persistent coughing had a notably greater occurrence of lung cancer (54.2%) comparing to those who did not (45.8%). A chronic cough is sometimes seen as the first clinical indication of lung cancer, especially in those who smoke or are exposed to environmental contaminants. According to the research, a persistent cough might be a sign of tumors blocking the airways or irritation of lung tissue, which typically occurs before the diagnosis of lung cancer [19].

This research found a strong correlation between chest discomfort, weariness, and lung cancer. Participants who reported chest pain had a greater prevalence of lung cancer (51.4%) compared to those who did not report chest pain (48.6%). These results support what doctors have seen in the clinic: chest pain may be the first sign of lung cancer, especially when it's caused by tumors pressing on nearby tissues like the pleura [20]. This research demonstrated a similar link between tiredness, a non-specific symptom, and lung cancer. 52.2% of the participants with fatigue received a lung cancer diagnosis, while 47.8% of those without fatigue did not. Studies have extensively recorded the correlation between tiredness and cancer, suggesting that the body's inflammatory response to cancer or cancer-related metabolic alterations may cause fatigue [21]. Furthermore, it is fascinating to note that people with allergies had a notably higher incidence of lung cancer (52.6% vs. 47.4%). Although the relationship between allergies and cancer remains unclear, numerous studies suggest that allergic responses to persistent inflammation could contribute to the development of cancer cells. Nevertheless, this continues to be a disputed field in cancer research, necessitating more examination [22].

Even though gender and age are often considered essential demographic determinants in cancer epidemiology, this research surprisingly found no significant association between lung cancer incidence and these parameters. The incidence rates of lung cancer in men (50.3%) and females (49.7%) were almost equal, indicating that gender did not have a significant impact on susceptibility to lung cancer. A recent study supports the idea that while males used to have higher rates of lung cancer because they smoked more, the difference between genders has been narrowing in recent years. This is because the number of women who smoke has increased in many regions of the globe [23]. The absence of a significant age disparity between people diagnosed with lung cancer and those without it (with a median age of 57 years and 56 years, respectively) indicates that the age distribution is comparable in both groups. Other variables, such as smoking or genetic predisposition, may have a more significant impact on lung cancer incidence in older populations. While age is a contributing factor, prior research indicates that the gradual accumulation of carcinogens over time, rather than age alone, primarily influences the occurrence of lung cancer [24]. In addition, this research found no significant association between characteristics such as anxiety and peer pressure, which might indirectly affect smoking habits, and the risk of lung cancer. Anxiety and peer pressure may exacerbate maladaptive coping strategies, such as heightened smoking prevalence, but their direct impact on cancer development is contingent upon the presence of other risk variables [25]. The results of this study have important implications for programs aimed at preventing and detecting lung cancer at an early stage. Given the strong correlation between smoking and lung cancer, it is imperative that public health initiatives persistently prioritize smoking cessation programs, especially among populations at higher risk. Early detection programs should also closely follow patients who display early symptoms such as persistent cough, chest discomfort, or exhaustion since these symptoms might be signs of undiscovered lung cancer. Furthermore, the absence of notable gender and age disparities emphasizes the need for widespread lung cancer screening across all demographic groups rather than limiting it to conventional risk factors such as age or gender. A future study should focus on investigating the intricate relationships between behavioral, psychological, and symptomatic aspects in the development of lung cancer.

The results of the Multilayer Perceptron (MLP) model emphasize the significance of age as the primary and most influential predictor of lung cancer, with a normalized value of 100% and the highest overall predictive power. Considering that age is a significant risk factor for lung cancer due to the gradual accumulation of cancer-causing substances over time, this result aligns with the extensive study undertaken on the topic [26]. Furthermore, the model highlights the importance of incorporating additional variables, such as the prevalence of shortness of breath (80.8% of the individuals) and alcohol consumption (70.0% of the individuals), which have been acknowledged in the scientific literature as factors that contribute to the likelihood of developing lung cancer [27]. It is intriguing that smoking, which is often considered a major risk factor for lung cancer [28], was shown to have a relatively low significance in our model (16.7%). Consequently, this suggests that other elements, such as behavioral indicators

(e.g., yellow fingers at 65.0%) and psychological aspects (e.g., anxiety at 50.8%), may exhibit a more influential prediction ability in certain datasets. This is corroborated by research that illustrates the correlation between stress and the advancement of cancer [29]. The relatively lower significance of chronic diseases (19.9%) and peer pressure (9.3%) compared to other variables suggests that although they are useful, these factors may not be primary indicators of lung cancer risk in this particular context. The findings of this research endorse the adoption of a comprehensive approach to lung cancer screening. This approach would include considering demographic, behavioral, and symptomatic factors to enhance the accuracy of prediction and effectiveness of intervention strategies. Future research should be focused on Bayesian Model Averaging, which has strengths against model uncertainties, may be an alternative to the approaches utulized in the current study [30].

## Conclusion

In conclusion, the results of this study not only reinforce the well-established link between smoking and lung cancer, but also emphasize the importance of symptoms such as chronic coughing, chest tightness, and excessive fatigue as significant early indicators of the disease. The findings indicate that demographic factors, such as gender and age, may have a less substantial impact on the development of lung cancer than previously thought. Viewed from this perspective, the study underscores the importance of focusing on behavioral risk factors and early symptoms when formulating methods for the prevention, identification, and treatment of lung cancer. These findings might potentially provide valuable guidance for the formulation of future public health policies and the creation of lung cancer awareness campaigns. This will guarantee that a heightened emphasis is placed on these crucial elements.

### *Ethical approval*

This study does not require ethical approval and informed consent because the open-source data set is used.

## References


1. Yuan Z, Wang L, Hong S, Shi C, Yuan B. Diagnostic value of HSP90$\alpha$ and related markers in lung cancer. Journal of Clinical Laboratory Analysis. 2022;36(6):e24462.
2. Malhotra J, Malvezzi M, Negri E, La Vecchia C, Boffetta P. Risk factors for lung cancer worldwide. European Respiratory Journal. 2016;48(3):889-902.
3. Barta JA, Powell CA, Wisnivesky JP. Global epidemiology of lung cancer. Annals of global health. 2019;85(1).
4. Kai W, Junhui W, Huiling Y, Gongxian M, Meng X. Effect of Petroleum Ether Extract from Rhizoma Amorphophalli (SLG) on Proliferation and Apoptosis of Non-small Cell Lung Cancer A549 Cells. MEDS Clinical Medicine. 2023;4(7):30-5.
5. Chen J-l, Wu J-n, Lv X-d, Yang Q-c, Chen J-r, Zhang D-m. The value of red blood cell distribution width, neutrophil-to-lymphocyte ratio, and hemoglobin-to-red blood cell distribution width ratio in the progression of non-small cell lung cancer. PloS one. 2020;15(8):e0237947.
6. Wu H, Yang J, Wang H, Li L. Mendelian randomization to explore the direct or mediating associations between socioeconomic status and lung cancer. Frontiers in Oncology. 2023;13:1143059.
7. Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A. Global cancer statistics, 2012. CA: a cancer journal for clinicians. 2015;65(2):87-108.
8. Carr SR, Akerley W, Hashibe M, Cannon-Albright LA. Evidence for a genetical contribution to non-smoking-related lung cancer. Thorax. 2015;70(11):1033-9.
9. Qiu X, Zhang H, Zhao Y, Zhao J, Wan Y, Li D, et al. Application of circulating genetically abnormal cells in the diagnosis of early-stage lung cancer. Journal of Cancer Research and Clinical Oncology. 2022;148(3):685-95.
10. Zhu X, Chen Y, Cui Y. Absolute neutrophil count and mean platelet volume in the blood as biomarkers to detect lung cancer. Disease Markers. 2020;2020(1):1371964.
11. Jiang L, Sun Y-Q, Brumpton BM, Langhammer A, Chen Y, Mai X-M. Body mass index and incidence of lung cancer in the HUNT study: using observational and Mendelian randomization approaches. BMC cancer. 2022;22(1):1152.
12. Song Z, Yang F, Du H, Li X, Liu J, Dong M, et al. Role of artemin in non-small cell lung cancer. Thoracic cancer. 2018;9(5):555-62.
13. Ramanaiah PK. Classification and Diagnosis of Lung Cancer Based Using CNN with VGG-19. 2024.
14. Kaggle. Lung Cancer CSV dataset 2024 [September 6, 2024]. Available from: https://www.kaggle.com/datasets/auu23egcse045/lung-cancer-csv-dataset/data.
15. Popescu M-C, Balas VE, Perescu-Popescu L, Mastorakis N. Multilayer perceptron and neural networks. WSEAS Transactions on Circuits and Systems. 2009;8(7):579-88.
16. Lushniak BD. A historic moment: The 50th anniversary of the first Surgeon General's Report on Smoking and Health. Public Health Reports. 2014;129(1):5-6.
17. Hecht SS. Lung carcinogenesis by tobacco smoke. International journal of cancer. 2012;131(12):2724-32.
18. Inoue-Choi M, Hartge P, Liao LM, Caporaso N, Freedman ND. Association between long-term low-intensity cigarette smoking and incidence of smoking-related cancer in the national institutes of health-AARP cohort. International journal of cancer. 2018;142(2):271-80.
19. Silvestri GA, Gonzalez AV, Jantz MA, Margolis ML, Gould MK, Tanoue LT, et al. Methods for staging non-small cell lung cancer: Diagnosis and management of lung cancer: American College of Chest Physicians evidence-based clinical practice guidelines. Chest. 2013;143(5):e211S-e50S.
20. Gadgeel SM, Kalemkerian GP. Lung Cancer: Overview. Lung Cancer Metastasis: Novel Biological Mechanisms and Impact on Clinical Practice. 2010:1-27.
21. Lozano R, Naghavi M, Foreman K, Lim S, Shibuya K, Aboyans V, et al. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. The lancet. 2012;380(9859):2095-128.
22. Turner MC, Krewski D, Diver WR, Pope III CA, Burnett RT, Jerrett M, et al. Ambient air pollution and cancer mortality in the cancer prevention study II. Environmental health perspectives. 2017;125(8):087013.
23. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2018. CA: a cancer journal for clinicians. 2018;68(1):7-30.
24. Tammemagi CM, Pinsky PF, Caporaso NE, Kvale PA, Hocking WG, Church TR, et al. Lung cancer risk prediction: prostate, lung, colorectal and ovarian cancer screening trial models and validation. Journal of the national cancer institute. 2011;103(13):1058-68.
25. Basu M, Das P, Mitra S, Ghosh S, Pal R, Bagchi S. Role of family and peers in the initiation and continuation of smoking behavior of future physicians. Journal of Pharmacy and Bioallied Sciences. 2011;3(3):407-11.
26. Levine ME, Hosgood HD, Chen B, Absher D, Assimes T, Horvath S. DNA methylation age of blood predicts future onset of lung cancer in the women's health initiative. Aging (Albany NY). 2015;7(9):690.
27. Bagnardi V, Blangiardo M, La Vecchia C, Corrao G. Alcohol consumption and the risk of cancer: a meta-analysis. Alcohol Research & Health. 2001;25(4):263.
28. Hecht SS. Tobacco smoke carcinogens and lung cancer. Journal of the national cancer institute. 1999;91(14):1194-210.
29. Wang L, Omrani H, Zhao Z, Francomano D, Li K, Pijanowski B. Analysis on urban densification dynamics and future modes in southeastern Wisconsin, USA. PloS one. 2019;14(3):e0211964.
30. Bahrami S, Hajian-Tilaki K, Bayani M, Chehrazi M, Mohamadi-Pirouz Z, Amoozadeh A. Bayesian model averaging for predicting factors associated with length of COVID-19 hospitalization. BMC Medical Research Methodology. 2023;23(1):163.