HIGH SCHOOL EDITION
Journal of Student Research
Volume 13 Issue 3 (2024)

How Lung Cancer Severity Can Be Predicted Using Machine-Learning Based on Different Risk Factors

Lila Bag[1] and Paul Chong[#]

[1]Istanbul International Community School, Turkey
[#]Advisor

ABSTRACT

Lung cancer is among the top causes of death globally, so this study sought to create a medical diagnostic solution in surveying the relationships between features such as symptoms and risk factors for lung cancer severity. 1000 publicly-accessible, anonymized patient records, different machine learning models were utilized with classification accuracy ranging from 92.5 to 100%. These findings argue for a greater role of passive smoke exposure in lung cancer severity than previously established, though further research is encouraged.

Introduction

Lung cancer is among the top causes of death globally, resulting in around 1.8 million deaths in 2020 according to the World Health Organization (WHO).[1] Previous works have been performed with techniques including fuzzy logic and hybrid neuro-fuzzy systems that have failed due to difficulties in forming valid medical diagnostic systems with increasing sample sizes while remaining reliable.[2-4] A similar study on the association between the incidence rate of lung cancer and environmental risks using predictive machine learning models suggested that the average NO_2 concentration, employment percentage, and number of factories were also significant factors in incidence of lung cancer.[5] A survey on the progress, process, and challenges of lung cancer detection and classification talked about different machine learning models in the context of medical imaging, such CT scans.[6] The authors' study made use of machine learning (ML) analysis to develop a medical diagnostic solution that would overcome the limitations of previous systems. For that purpose, the relationship between the different features of people, including air pollution and alcohol use, with the severity of lung cancer was analyzed using ML.

Research question: How can the relationship between the number of people with lung cancer and the different risk factors be discovered using machine learning?

Materials & Methods

The dataset utilized for this paper contained publicly accessible, anonymized patient records which would be preprocessed with binarization for ML analysis using the Python platform and a range of classifiers, such as the Decision Tree Classifier, Random Forest Classifier, and the Multi-layer Perceptron (MLP) Classifier, using the Scikit-learn package.[5-7] First, a descriptive statistical analysis was conducted and a figure representing the average values of select features and the severity of lung cancer was made. Then, the set of classifiers had subsets to train on and a separate test dataset to generate predictions on; explainability techniques, including feature importance, aided in gaining insight into the ML models and their predictions.

See below for code of machine learning analysis using Decision Tree Classifier:

```
from sklearn import tree
import pandas as pd
```

```
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import confusion_matrix, accuracy_score, cohen_kappa_score
from matplotlib import pyplot as plt

dataset=pd.read_csv("/content/drive/MyDrive/Chong Research Group/lung_cancer.csv")
features = ["Age","Gender","Air Pollution","Alcohol use","Dust Allergy","OccuPational Hazards", "Genetic Risk", "chronic Lung Disease", "Balanced Diet", "Obesity", "Smoking", "Passive Smoker", "Chest Pain", "Coughing of Blood", "Fatigue", "Weight Loss", "Shortness of Breath", "Swal0ing Difficulty", "Clubbing of Finger Nails", "Frequent Cold", "Dry Cough", "Snoring"]
x=dataset[features].values
y=dataset["Level"].values
X_train, X_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=0)
clf = DecisionTreeClassifier(max_depth=3)
clf = clf.fit(X_train,y_train)
y_pred = clf.predict(X_test)
matrix = confusion_matrix(y_test, y_pred)
print(matrix)
feature_imp = pd.Series(clf.feature_importances_, index = features).sort_values(ascending = False)
print(feature_imp)
plt.figure(figsize=(16,7))
plt.barh(feature_imp.index, feature_imp)
plt.xlabel("Feature Importance")
plt.grid(axis="x")
plt.show()
fig = plt.figure(figsize=(25,20))
_ = tree.plot_tree(clf,
        feature_names=features,
        class_names=["0","1", "2"],
        filled=True)
```

See below for code of machine learning analysis using Random Forest Classifier:

```
from sklearn import tree
import pandas as pd
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import confusion_matrix, accuracy_score, cohen_kappa_score
from matplotlib import pyplot as plt

dataset=pd.read_csv("/content/drive/MyDrive/Chong Research Group/lung_cancer.csv")
features = ["Age","Gender","Air Pollution","Alcohol use","Dust Allergy","OccuPational Hazards", "Genetic Risk", "chronic Lung Disease", "Balanced Diet", "Obesity", "Smoking", "Passive Smoker", "Chest Pain", "Coughing of Blood", "Fatigue", "Weight Loss", "Shortness of Breath", "Swal0ing Difficulty", "Clubbing of Finger Nails", "Frequent Cold", "Dry Cough", "Snoring"]
x=dataset[features].values
y=dataset["Level"].values
X_train, X_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=0)
```

```
clf = RandomForestClassifier()
clf = clf.fit(X_train,y_train)
y_pred = clf.predict(X_test)
matrix = confusion_matrix(y_test, y_pred)
print(matrix)
feature_imp = pd.Series(clf.feature_importances_, index = features).sort_values(ascending = False)
print(feature_imp)
plt.figure(figsize=(16,7))
plt.barh(feature_imp.index, feature_imp)
plt.xlabel("Feature Importance")
plt.grid(axis="x")
plt.show()
```

See below for code of machine learning analysis using MLP Classifier:

```
from sklearn import tree
import pandas as pd
from sklearn.neural_network import MLPClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import confusion_matrix, accuracy_score, cohen_kappa_score
from matplotlib import pyplot as plt

dataset=pd.read_csv("/content/drive/MyDrive/Chong Research Group/lung_cancer.csv")
features = ["Age","Gender","Air Pollution","Alcohol use","Dust Allergy","OccuPational Hazards", "Genetic Risk", "chronic Lung Disease", "Balanced Diet", "Obesity", "Smoking", "Passive Smoker", "Chest Pain", "Coughing of Blood", "Fatigue", "Weight Loss", "Shortness of Breath", "Swal0ing Difficulty", "Clubbing of Finger Nails", "Frequent Cold", "Dry Cough", "Snoring"]
x=dataset[features].values
y=dataset["Level"].values
X_train, X_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=0)
clf = MLPClassifier(hidden_layer_sizes=(150,100,50), max_iter=300,activation = 'relu',solver='adam',random_state=1, verbose=1)
clf = clf.fit(X_train,y_train)
y_pred = clf.predict(X_test)
matrix = confusion_matrix(y_test, y_pred)
print(matrix)
feature_imp = pd.Series(clf.feature_importances_, index = features).sort_values(ascending = False)
print(feature_imp)
plt.figure(figsize=(16,7))
plt.barh(feature_imp.index, feature_imp)
plt.xlabel("Feature Importance")
plt.show()
```

Results and Discussion

Statistical Analysis

The patient dataset that was utilized in this study contained 1000 anonymized patient records, with an average age of around 34.2 years. The dataset included 407 males and 593 females. See Table 1 for a summary of select features, see Figure 1 for a visual representation of select features.

Table 1. Summary of Average Values of Select Features and Level of Lung Cancer. These were the results of the average values of the select features and the level of lung cancer. The findings showed that there's a trend in the average values and the severity of lung cancer, with the exception of the smoking feature, which did not exhibit the same pattern, (see Figure 1).

	Overall Data Set	Level 0	Level 1	Level 2
Passive Smoke Exposure	4.195	2.63366337	3.05421687	6.52876712
Shortness of Breath	4.24	2.49834983	4.63253012	5.32876712
Fatigue	3.856	2.17161716	3.48795181	5.5890411
Coughing of Blood	4.859	2.86138614	3.84638554	7.43835616
Obesity	4.465	2.40924092	3.90361446	6.68219178
Air Pollution	3.84	2.60066007	2.93373494	5.69315068
Smoking	3.948	3.02310231	2.45481928	6.0739726

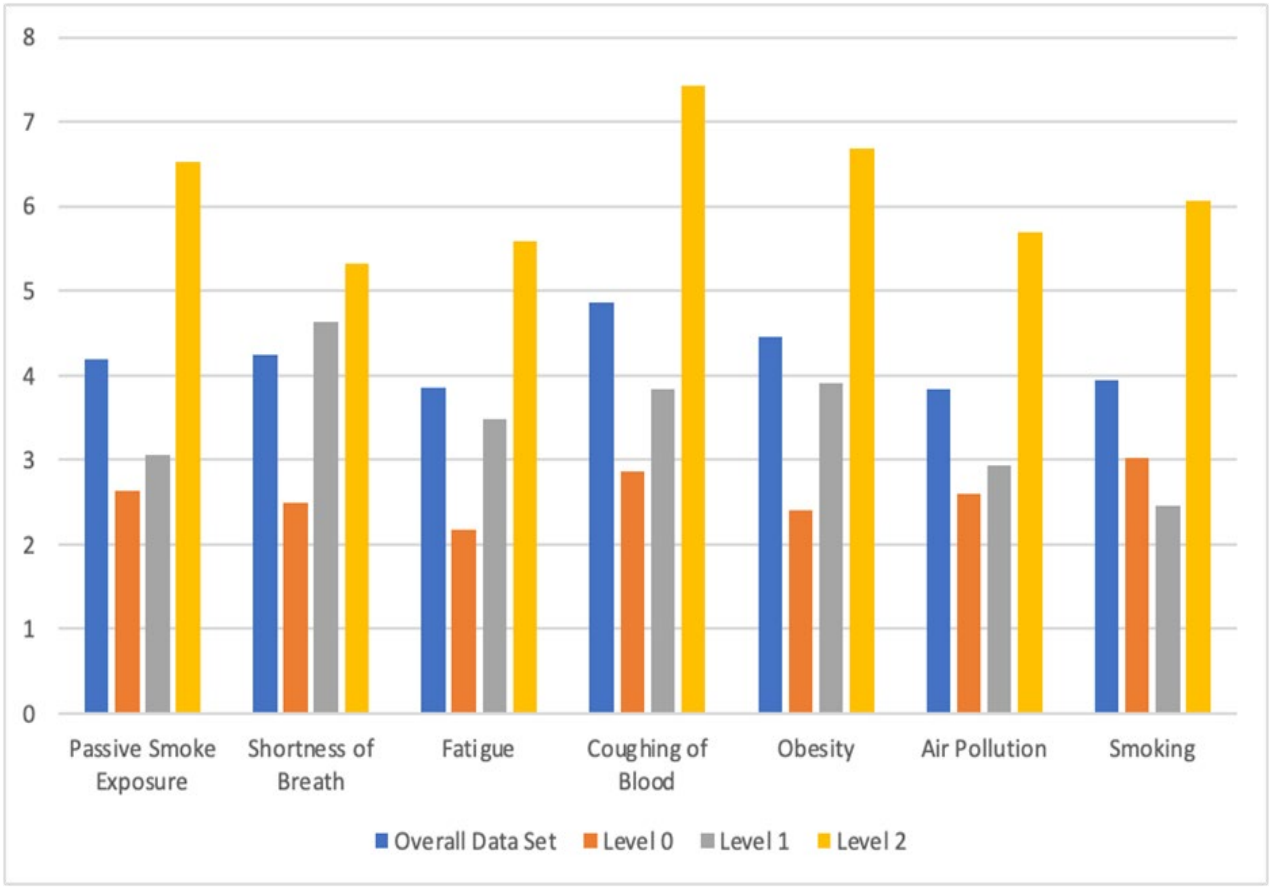

Figure 1. Figure Representing the Average Values of Select Features and Level of Lung Cancer. This is the visual representation of the values seen in Table 1.

Machine Learning Analysis

The Decision Tree Classifier's analysis for its prediction of the level of lung cancer based on the risk factors the test dataset provided showed an accuracy of 92.5%. The analysis of the feature importance suggests that the most important and impactful risk factor when determining the severity of lung cancer was the patient's passive smoking status, followed by signs of shortness of breath, fatigue, and coughing of blood (see Figure 2). See Figure 3 for decision tree visualization.

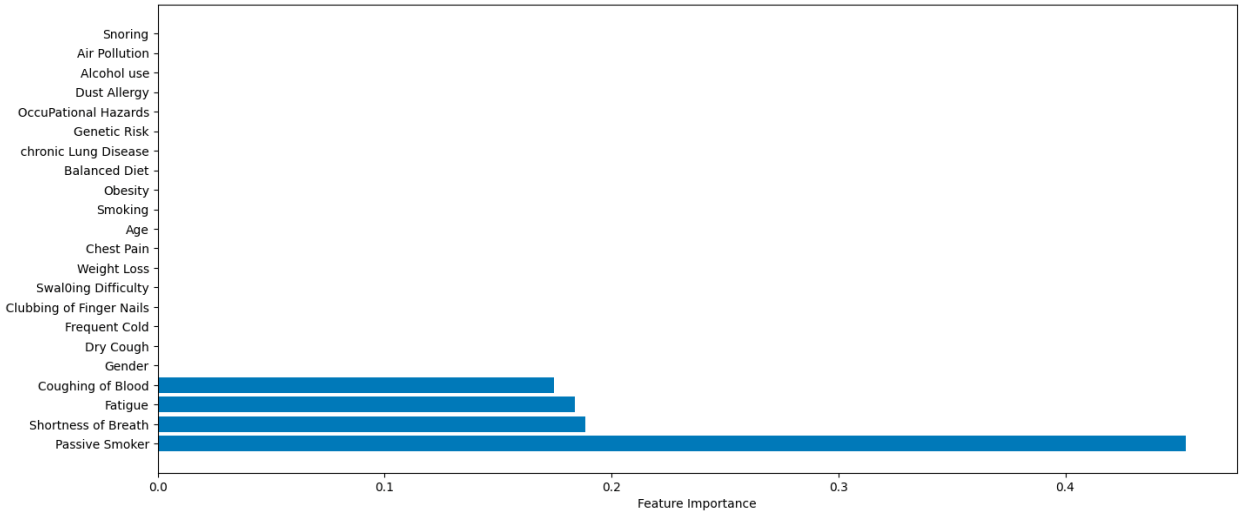

Figure 2. Feature importances for Decision Tree Classifier. These are the results of the feature importance analysis of the different risk factors of patients with lung cancer. The findings show that passive smoking is by far the most significant when compared to the other risk factors included.

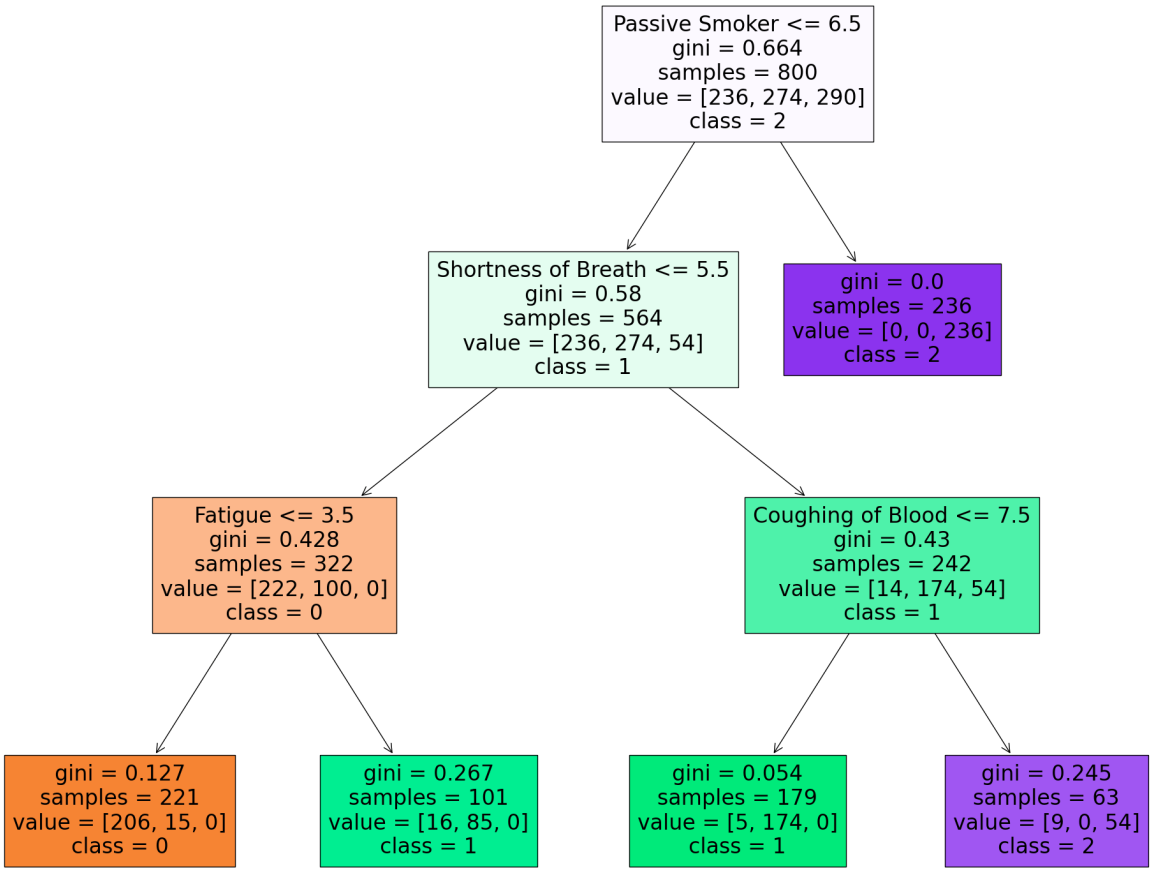

Figure 3. Visualization of Decision Tree Classifier. These are the results of the visualization of the Decision Tree Classifier. If the passive smoke exposure levels were above 6.5, then all patients, included in the sample dataset, were shown to have class 2 lung cancer.

The Random Forest Classifer's analysis in predicting the severity of lung cancer with the risk factors provided in the dataset for testing after training showed 100% accuracy. The feature importance analysis demonstrated again that the status of passive smoking was the most significant risk factor in deciding the level of lung cancer, with coughing of blood, obesity, and fatigue also being of significance (see Figure 4).

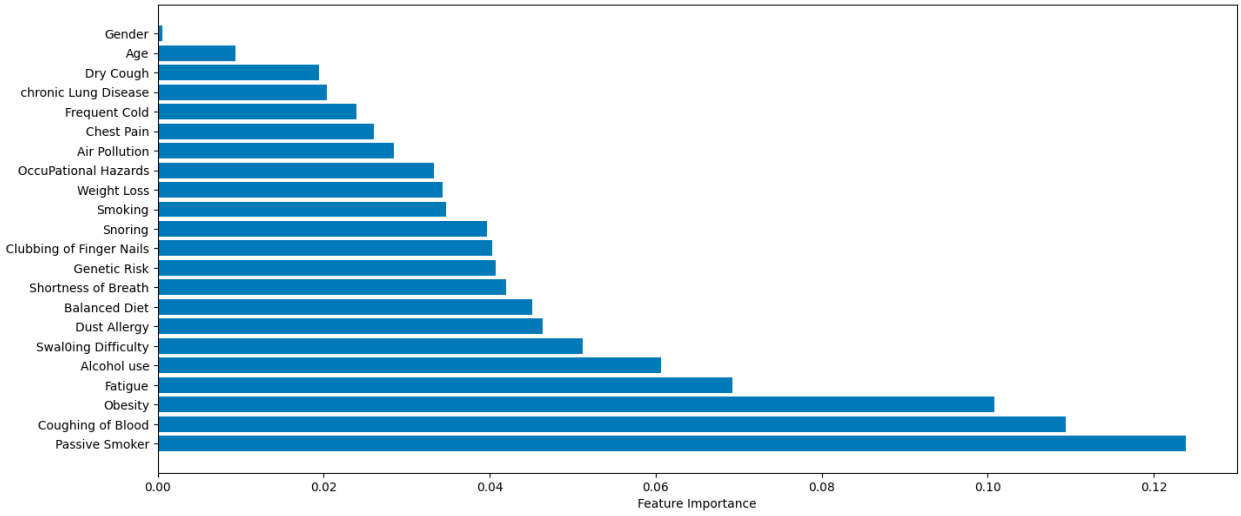

Figure 4. Feature importances of Random Forest Classifier. These results are the feature importance analysis of the different risk factors of patients with lung cancer. These findings show that passive smoking is the most significant risk factor, followed by coughing of blood and obesity.

The MLP Classifier displayed a 100% accuracy in its predictions of the level of lung cancer with the risk factors given in the dataset after training. The feature importance analysis could not be performed due to the type of the ML model. See Table 2 for the confusion matrices and accuracy of ML classifier models.

Table 2. Confusion matrices and accuracy of ML classifier models. These were the confusion matrices and the overall accuracy of the different ML classifier models used in this study. The Random Forest and MLP Classifiers both had perfect overall accuracy scores, however the Decision Tree Classifier had an overall accuracy of 92.5%.

ML classifier model	Confusion matrix	Overall accuracy
Decision Tree	[[57 9 1] [5 53 0 0 0 75]]	92.5%
Random Forest	[[67 0 0] [0 58 0 0 0 75]]	100%
MLP	[[67 0 0] [0 58 0 0 0 75]]	100%

As seen in Table 1, the relationship between the relative amount of passive smoke exposure and the severity of lung cancer is proportional. This is also mirrored in the feature importance analysis for the Decision Tree and Random Forest Classifiers. This relationship suggests that passive smoke exposure has an influential impact on the

level of lung cancer. This relationship also has a significant impact on determining the risk of the development of lung cancer in other symptoms and risk factors as seen in previous results. Previous literature has brought to the authors' attention that smoking and air pollution should have a greater effect on the level of lung cancer in the patients than what was observed, as the results showed that the relationship is inferior when compared to that of the aforementioned risk factors. [8-9]

The findings of this study showed that the Decision Tree Classifier, while predicting the lung cancer severity had fair accuracy while providing insight on what the individual impacts of the various risk factors and symptoms had shown. Although the Random Forest and MLP Classifiers achieved perfect accuracy in their predictions of the level of lung cancer in the patients, the individual impacts that the features and risk factors held were indeterminate beyond the feature importance analysis.

The conclusion of the authors' study tells that passive smoke exposure plays a greater role in the development and severity of lung cancer than previously thought. These results may indicate that exposure to passive smoking bears a greater risk than previously known, further research is encouraged. These findings imply that symptoms and risk factors such as levels of obesity and alcohol use mirror the severity of lung cancer.

The authors' results show that more balanced diets are associated with greater severity of lung cancer and that this connection is stronger than that between the level of lung cancer severity with smoking or higher levels of air pollution, though it is likely a coincidence. These conclusions imply that higher severities of obesity and alcohol use can be linked to an increased severity of lung cancer. The results do however challenge the mechanisms of the models; given the established evidence of the impact of smoking and air pollution on the risk of the development of lung cancer.

Limitations

There were a few notable limitations within this study, such as the lack of transparency of the Random Forest and MLP Classifiers. This is also seen in the aforementioned example of the balanced diet, smoking, and air pollution features.

Another example of a restraint in this study is its insufficiency in its representation of all populations since the dataset used was taken from China. The conclusion likely would have been different if the data had been gathered globally, resulting in some features, such as smoking and air pollution, having greater feature importance. A third limitation of this study would be the limited sample size of the dataset and the incapacity to show beyond doubt the statistical significance of these associations as seen by the ML models.

Conclusion

The relationship between lung cancer severity and various features was inspected using multiple ML models mentioned earlier with classification accuracy ranging from 92.5 to 100%. These findings suggest that passive smoke exposure plays a larger role in lung cancer severity and development than previously thought, though further research is encouraged.

Acknowledgments

I would like to thank my advisor for the valuable insight provided to me on this topic.

References

1. Lung cancer. www.who.int. Published June 26, 2023. https://www.who.int/news-room/fact-sheets/detail/lung-cancer#:~:text=GLOBOCAN%202020%20estimates%20of%20cancer

2. Ahmad AS, Mayya AM. A new tool to predict lung cancer based on risk factors. Heliyon. 2020;6(2):e03402. Published 2020 Feb 26. doi:10.1016/j.heliyon.2020.e03402

3. Tiwari SK, Walia N, Singh H, Sharma A. Effective Analysis of Lung Infection using Fuzzy Rules. International Journal of Bio-Science and Bio-Technology. 2015;7(6):85-96. doi:https://doi.org/10.14257/ijbsbt.2015.7.6.10

4. Billah M., Islam N. An early diagnosis system for predicting lung cancer risk using adaptive neuro fuzzy inference system and linear discriminant analysis. J. MPE Mol. Pathol. Epidoemiol. 2016;1(3):1–4.

5. Wang KM, Chen KH, Hernanda CA, Tseng SH, Wang KJ. How Is the Lung Cancer Incidence Rate Associated with Environmental Risks? Machine-Learning-Based Modeling and Benchmarking. *Int J Environ Res Public Health*. 2022;19(14):8445. Published 2022 Jul 11. doi:10.3390/ijerph19148445

6. Mridha MF, Prodeep AR, Hoque ASMM, et al. A Comprehensive Survey on the Progress, Process, and Challenges of Lung Cancer Detection and Classification. *J Healthc Eng*. 2022;2022:5905230. Published 2022 Dec 16. doi:10.1155/2022/5905230

7. Ahmad AS, Mayya AM. Lung Cancer Prediction. www.kaggle.com. Published September 19, 2017. https://www.kaggle.com/datasets/thedevastator/cancer-patients-and-air-pollution-a-new-link

8. Python Software Foundation. Python Language Reference, version 3.7.1. Available at http://www.python.org

9. Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

10. Klebe S, Leigh J, Henderson DW, Nurminen M. Asbestos, Smoking and Lung Cancer: An Update. Int J Environ Res Public Health. 2019;17(1):258. Published 2019 Dec 30. doi:10.3390/ijerph17010258

11. Cheng I, Yang J, Tseng C, et al. Traffic-related Air Pollution and Lung Cancer Incidence: The California Multiethnic Cohort Study. Am J Respir Crit Care Med. 2022;206(8):1008-1018. doi:10.1164/rccm.202107-1770OC