Lung Cancer Risk Assessment Through Machine Learning and Exploratory Data Visualization

Kishan Kumar[1], Nilesh Gupta[2]
[1]*M. Tech Student, Chouksey Group of Colleges, Bilaspur (C.G.), India*
[2]*Asst Professor Department of CSE, Chouksey Group of Colleges, Bilaspur (C.G.), India*

Abstract- Lung cancer is one of the most fatal cancers worldwide, largely because of advanced stage at diagnosis and lack of effective early detection methods. Data-driven methods for identifying high-risk individuals may have the potential of being instrumental in improving clinical outcomes. How the study will be conducted: This analysis employs an encompassing approach, unifying exploratory data visualization with machine learning-based predictive modeling to predict risk of lung cancer based on patient-level clinical, behavioral and demographic information. A visual analytics pipeline consisting of univariate distribution analysis, correlation heatmaps, density plots (violin) and feature interaction was used to discover hidden patterns, influential risk factors and feature consistency between patient groups. These visual clues led to the choice or the pre-processing of features for model building. A variety of machine learning models, such as Logistic Regression (LR), Random Forest, Support Vector Machines (SVM) and Gradient-Boosted Model were built and evaluated in a systematic way through the conventional performance indicators including accuracy, precision, recall F1-score and AUC-ROC. Our findings indicate that the integrated application of VA and PM generates more interpretable/clinically relevant results, while providing an insight into the dynamics associated with risk. The approach proposed here illustrates the value of machine learning–assisted analysis in early disease detection and the reinforcement of clinical decision-support systems for lung cancer risk assessments.

Keywords: Lung cancer, Machine learning, Risk assessment, Data visualization, Predictive modeling, Clinical analytics

1. INTRODUCTION

Lung cancer is one of the most common and fatal malignancies worldwide, making it responsible for a substantial fraction of the total global cancer-related mortality. Recent epidemiological studies indicate lack of early diagnose and absence of timely access to screening are the core concern for patients to survive this disease [1]. Consequently, there is an urgent requirement for early-stage risk assessment tools that can help in identifying populations of at-risk subjects who might benefit from preventive intervention and/or more extensive diagnostic evaluation by clinicians. Current diagnostic methods are based on imaging, clinical examination and a confirmatory biopsy; although these tools can be expensive and time-consuming, they may not provide adequate prediction of the likelihood of disease at an earlier stage [2]. Despite the abundant literature on this topic, the increasing availability of structured clinical data and progress in computational methods as well as statistical algorithms have paved the way to develop data-driven models that can reveal complex patterns related to lung cancer risk. Machine learning methods have been shown in their high potential to model nonlinear interactions between demographic, behavioral and clinical features related to the disease susceptibility [3]. Data exploration visualization extends this process by visualizing distributions of data, relationships between variables and the significance of features, making both interpretability and model selection [4] more informed. However, the integration of visual analytics and predictive modeling for a common analysis is still challenging. Most current researches pay attention respectively to model performance or visualization alone, which lead to an incomplete understanding of risk factors [5]. Motivated by these limitations, in this study we present a holistic framework, which integrates exploratory visualization and machine learning based classification for assessing the risk of lung cancer from patient-level data. The key contributions of this article are: (1) they provide an extensive exploration visualization of lung cancer risk indicators; (2)

developed several machine learning models and compared their performance for accurate prediction of the risk; (3) introduced a single workflow capable of enhancing interpretability via visualized feature analysis; and 4, findings that encourage early detection and clinical decision making. This model was developed to generate an interpretive connection between data interpretation and predictive performance, providing a strong foundation for the development of practical, data-based models using lung cancer risk assessment tools.

2. RELATED WORK

The use of machine learning (ML) and data visualization in lung cancer research has expanded tremendously in the past few years due to the urgency for early detection and better clinical management. Multiple studies have explored various ML algorithms for risk stratification of lung cancer with clinical, demographic, radiographic and biomarker data. The state of traditional methods like Logistic Regression, SVM and Random Forest have shown good precision of high-risk patients prediction with structured data [1]. For example, [2] used SVM and Decision Trees based on a patient survey data and reported substantial enhancement in predictive accuracy over traditional statistical tools. Deep learning techniques, in particular, Convolutional Neural Networks (CNNs), have also been applied extensively for radiological image analysis allowing automated detection of nodules and malignancy patterns from CT scans [3]. However, these methods may depend on big labeled sets and can be CPU expensive. Alongside ML methods, exploratory data visualization has been commonly used for increasing interpretability, exposing correlations among features, also revealing clinically relevant patterns. From feature distribution and interaction analysis, histogram, boxplot as well as correlation heatmap estimates and dimensionality reduction visualizations t-SNE or PCA have been applied on lung cancer datasets in [4]. Visualization-oriented researches claim that interpretability need to be highly desired in clinical application especially on high-dimensional domains where significances of features are not clear at beginning [5]. Although great progress has been made, current state-of-the-art work is frequently limited by their data size, lack of interpretability, and the gap between visualization and predictive modeling. Several researches show high accuracies but give no explanation regarding the synergic/rival effects between the risk factors and their effect on classification results [6]. Furthermore, there are only few works that integrate visualization-guided feature analysis with supervised learning pipelines, and so there remains significant space for tools to bridge interpretation and prediction.

To establish this work in the context of prior works, Table 1 and Table 2summarized important prior arts: one is oriented to ML models based on prediction; the other focuses on visualization-driven analysis.

Table 1. Summary of ML-Based Lung Cancer Prediction Studies

Study	Dataset Type	Methods Used	Key Findings
[1]	Clinical + demographic	Logistic Regression, SVM	SVM achieved highest accuracy for early risk classification.
[2]	Questionnaire survey	Decision Tree, Random Forest	RF improved interpretability and outperformed baseline models.
[3]	CT imaging	CNN, 3D-CNN	Deep models detected malignant nodules with high sensitivity.
[6]	Mixed clinical dataset	XGBoost	Achieved strong performance but lacked feature explanations.

Table 2. Summary of Visualization Approaches in Lung Cancer Research

Study	Visualization Method	Purpose	Limitation
[4]	Heatmaps, histograms	Show feature trends	Limited to numeric data
[5]	t-SNE, PCA	Explore high-dimensional patterns	Hard to interpret clinically

[7]	Box/violin plots	Compare group distributions	Only univariate insight
[8]	Survival curves	Examine prognosis	Needs longitudinal data
[9]	Cluster heatmaps	Group similar patients	Sensitive to scaling
[10]	ROC/PR curves	Evaluate model performance	No feature interpretation
[11]	SHAP plots	Explain model predictions	High compute cost
[12]	Bar/radar charts	Summarize feature importance	Oversimplifies relationships

3. DATASET & PREPROCESSING

The dataset used in this study is a blend of demographic, clinical and behavioral attributes representing lung cancer risk factors (age, gender, smoking frequency and history, COPD symptoms and early detection test results) as well lifestyle indicators like the number of hours you spend sitting or exercising. a qualitative classification ("low", "medium" or "high") of patients based on provided features. In heavier, the data was preprocessed to ensure analytic credibility including data cleaning (duplicate removal), missing values treatment using mean or median imputation for numerical attributes and mode substitution for categorical variables, and outlier detection under statistical distribution analysis. Non-numerical fields were transformed into machine-interpretable format via label encoding for the ordinal parameter, whereas the nominal attribute was transformed using one-hot encoding. For stable and fast convergence of machine learning models, numerical features with wide scale ranges were normalized into Z-score standardization or Min–Max scaling. This composite preprocessing process converted the raw data set into a well-structured, consistent and fully optimized format suitable for visualization, modeling and classification that can be used within the developed lung cancer risk assessment framework.

Exploratory Data Visualization (EDV) was used to investigate the distribution, inter-relationships and predictive importance of factors associated with risk of lung cancer. Distributions, histogram, density curve and violin plots showed significant diversity by demographic and behavioral characteristics with skewed distributions indicating different risk intensities among individuals; e.g. smoking frequency, age, environmental exposure. Correlation heatmaps were applied to explore linear relationships among numeric features, there were modest positive correlations between smoking-related-level features and reported symptom intensity, while weaker associations were observed for demographic such as gender. Pairwise feature plots yielded additional insights into feature interactions where high-risk classes cluster at higher values of smoking history, chronic cough, and fatigue. These visual patterns provided the initial highlights on the most relevant predictors and supported feature selection when creating machine-learning models.

Major observed trends in the exploratory visualizations are summarized in Table 3. It underscores differences in symptom severity, exposure to environmental risk factors, and lifestyle conditions among lung cancer risk groups. Regressors such as chest pain or coughing of blood, smoking, and air pollution show well-defined upward tendencies revealing their strong predisposition among high-risk categorization.

Table 3. Summary of Key Exploratory Visualization Insights

Visualization Method	Feature Observed	Key Insight	Interpretation
Histogram	Smoking Level	Strong right-skew pattern	Higher smoking intensity linked to elevated risk
Density Plot	Age Distribution	Uneven spread with older peak	Older individuals show higher symptom prevalence
Violin Plot	Symptom Severity	Wider spread in high-risk group	Severe symptoms consistently map to high-risk class

Box Plot	Exposure Level	Notable outliers in high exposure	Environmental exposure contributes to variability
Heatmap	Feature Correlation	Moderate links among smoking features	Behavioral habits show stronger risk associations
Scatter Plot	Exposure vs Symptoms	Clustering of high-risk instances	Combined exposure and symptom intensity predicts risk

4. RESULT ANALYSIS

4.1 Analysis of Feature Trends Across Lung Cancer Risk Levels

Pattern of features are evident in our feature trend analysis between lower-, mid-, and high-risk lung cancer group. Risk is related to smoking intensity, chronic lung disease, chest pain, blood-related symptoms and exposure. Weight loss among some features, has very little variation. Across all symptoms and exposures, at-risk individuals show consistently higher levels of these source categories than average risk individuals, reinforcing the importance of high-risk group membership for prediction of lung cancer risk levels

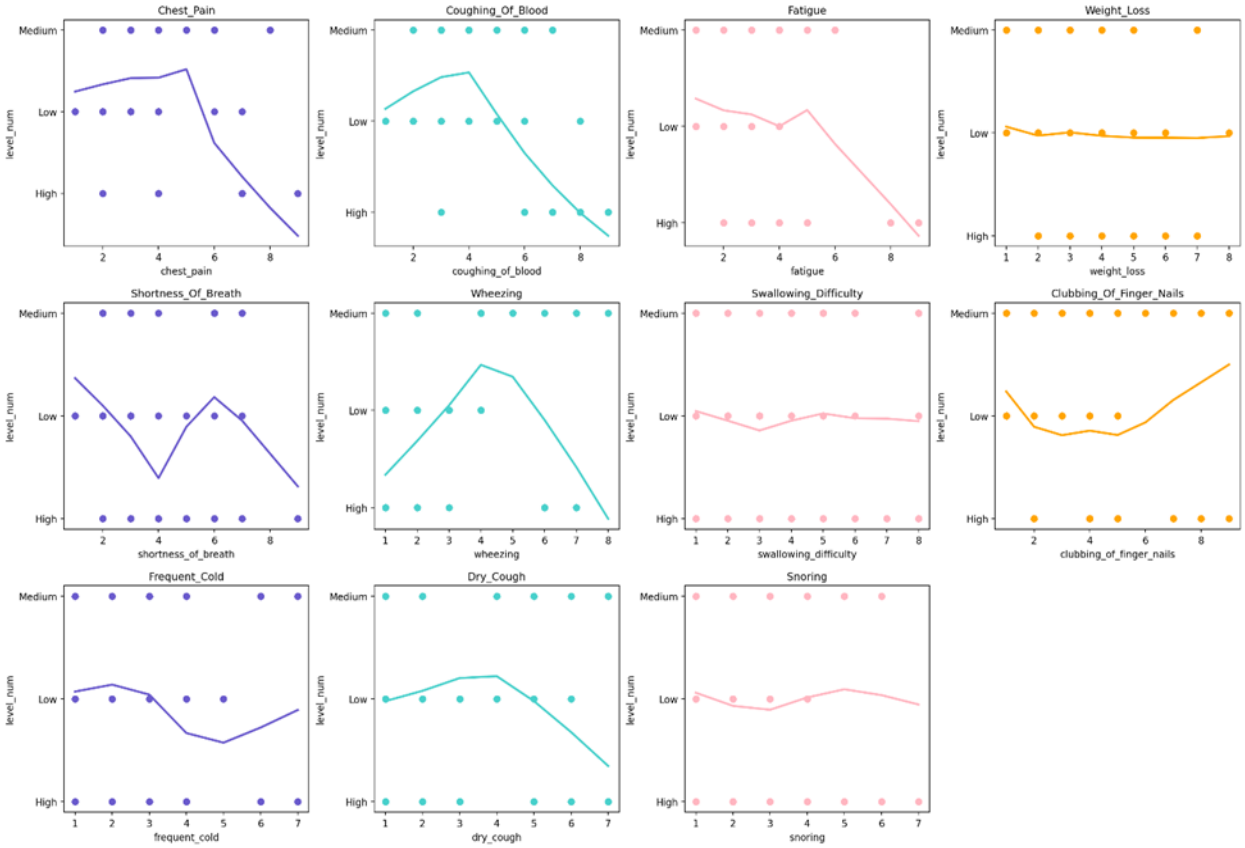

Figure 1: Level-Wise Trends of Clinical and Behavioral Features in Lung Cancer Risk Assessment

This Fig 1 shows differences in important clinical and behavioral characteristics across low, medium, and high lung cancer risk levels. The trends of chest pain, coughing of blood, fatigue and wheezing as symptoms were significantly decreasing-demonstrating a higher number of severe cases in the high-risk group. Nail clubbing shows a marked enlargement of the sinuses, indicating severe respiratory obstruction. Traits such as weight loss and snoring appear more constant, indicating less predictive power. In general, the trends indicate that increasing numbers of breathlessness and respiratory disease are highly associated with greater risk of lung cancer

Table 4: Analysis of Clinical & Behavioral Features by Lung Cancer Risk Level

Feature	Low (Mean)	Medium (Mean)	High (Mean)	Interpretation
Chest Pain	2.83	3.75	6.39	Strong increase → major high-risk symptom
Coughing of Blood	2.86	3.85	7.44	Sharp rise → severe indicator of lung disease
Fatigue	2.17	3.49	5.59	Consistent rise → fatigue intensifies with risk
Weight Loss	2.50	4.42	4.47	Moderate increase → mild predictor
Shortness of Breath	2.50	4.63	5.33	Clear upward trend → strong respiratory marker
Wheezing	2.57	4.76	3.88	Peaks at medium → variable but relevant
Swallowing Difficulty	2.76	4.16	4.19	Gradual increase → moderate association
Clubbing of Finger Nails	2.47	4.94	4.21	Strong medium/high-risk signal
Frequent Cold	2.37	3.67	4.38	Increases with risk → mild indicator
Dry Cough	2.91	3.70	4.78	Noticeable rise → symptom severity increases
Snoring	2.14	3.31	3.23	Mild rise → weak predictor

As shown in the table 4, symptomatic severities distribute significantly differently among lung cancer risk levels. High-risk patients have significantly greater mean scores for chest pain, coughing of blood, tiredness, shortness of breath and dry cough, thus these are clear clinical hallmarks. A moderate increase in clubbing of the fingernails, common cold and weight loss also add to their applicability for risk assessment. Wheezing and swallowing difficulty reach the maximum at around medium or high levels, demonstrating their variable effect. In general, the table shows that higher increasing intensity of symptoms is parallel with an at risk for lung cancer meaning and validate their role in a predictive model.

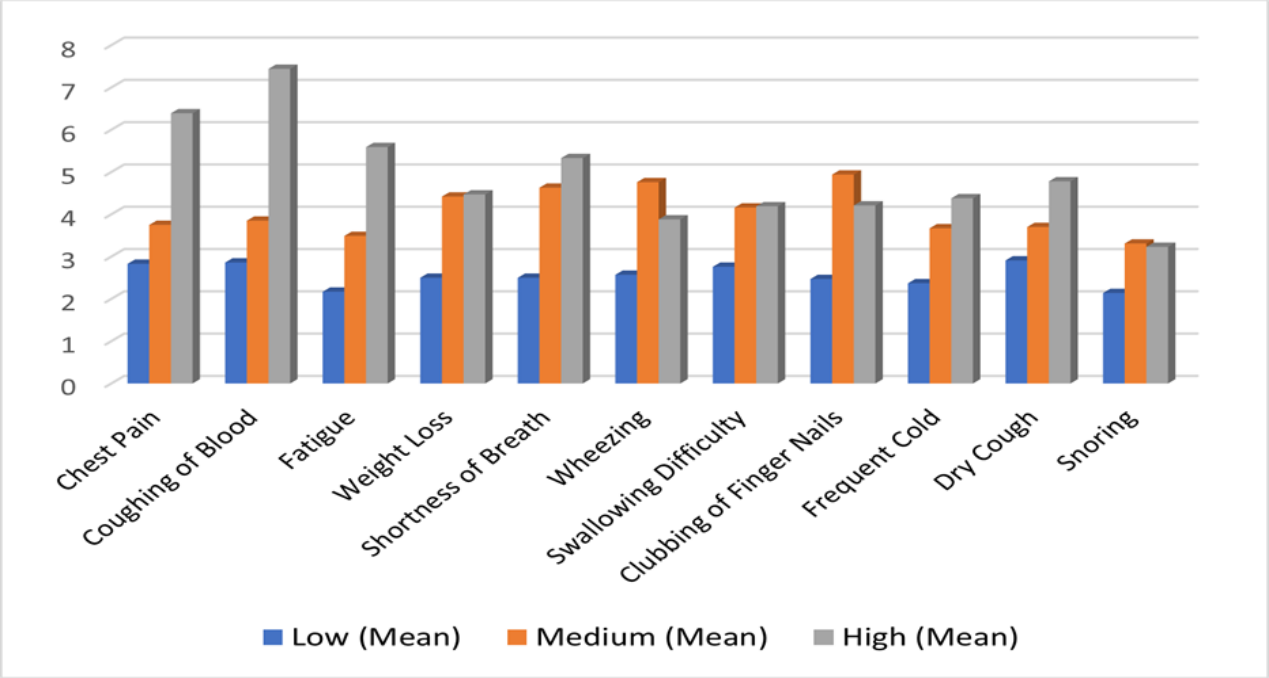

Figure 2: Mean Symptom Severity Across Low, Medium, and High Lung Cancer Risk Levels

The bar graph in Figure 2 presents the average values for selected clinical and behavioural characteristics by low, medium, and high lung cancer risk group. High-risk groups always have the highest symptom severity, especially in coughing of blood (7.44), chest pain (6.39), fatigue (5.59) and shortness of breath (5.33). These substantial number differences underscore their role as major predictors of lung cancer risk. Medium-risk scores are also generally substantially higher than low-risk values for all features, suggesting increasing symptom severity. Conversely, snoring and weight loss are less variable than overall severity level is, which implies that there is less direct explanation for the variability. In general, there is a clear positive relationship between the severity of symptoms and level of risk.

4.2 Distribution-Based Result Analysis

The distribution plots present important characteristics for the clinical, behavior and demographic facets of lung cancer dataset. Observed symptoms like chest pain, hemoptysis, weakness, dyspnea, wheezing and COPD demonstrate very high degree distributions at the higher range of the scale indicating that they are strongly related to a severe raised IF for their presentation. Behavioral variables including smoking, secondhand smoke exposure, air pollution exposure and occupational exposures have more outliers in the analyzed data that could be interpreted as a greater variability between people and a high impact of these factors on lung cancer risk. On the contrary, features such as weight loss, balanced diet and snoring reveal relatively similar patterns across groups meaning weak discriminatory power. The above Table 5 indicates that the predictors for higher level of lung cancer risk are dominated by worsening respiratory conditions and adverse exposures.

Table 5: Summary of Feature Distributions

Feature	Mean (μ)	Std. Dev (σ)	Interpretation
Age	37.2	12.0	Middle-aged group most represented; wide age spread
Gender	1.4	0.5	Majority of participants belong to one gender group
Air Pollution	3.8	2.0	Moderate pollution exposure with broad variation
Alcohol Use	4.6	2.6	High variability in alcohol consumption habits
Dust Allergy	5.2	2.0	Symptoms generally moderate to high
Occupational Hazards	4.8	2.1	Many individuals exposed to workplace risks
Genetic Risk	4.6	2.1	Genetic predisposition moderately high
Chronic Lung Disease	4.4	1.8	Presence of chronic conditions common in dataset
Balanced Diet	4.5	2.1	Variation in diet quality present
Obesity	4.5	2.1	Mid-range obesity levels with noticeable spread
Smoking	3.9	2.5	Wide variation; major risk-related behavior
Passive Smoker	4.2	2.3	Significant exposure to secondhand smoke
Chest Pain	4.4	2.3	Symptom frequently reported with high variance
Coughing of Blood	4.9	2.4	Severe symptom; distribution skewed to higher values
Fatigue	3.9	2.2	Moderate fatigue levels with notable variation
Weight Loss	3.9	2.2	Mild to moderate weight reduction patterns
Shortness of Breath	4.2	2.3	Strong respiratory symptom prevalence
Wheezing	3.8	2.0	Moderate respiratory difficulty across samples
Swallowing Difficulty	3.7	2.2	Symptom present but less variable
Clubbing of Nails	3.9	2.4	Potential indicator of chronic disease progression
Frequent Cold	3.5	1.8	Mild to moderate cold occurrences
Dry Cough	3.9	2.0	Persistently distributed respiratory symptom
Snoring	2.9	1.5	Lower score range; weak risk indicator

The table of distribution provides an overall picture on the central tendency and variety of each clinical, behavioral, and demographic characteristic in the lung cancer data. The higher average across symptoms such as chest pain, coughing up blood, shortness of breath and chronic lung disease suggests that these are prevalent and perhaps more influential in risk forecasting. Factors including smoking, passive smoking, and occupational exposure also vary widely between participants; this indicates major differences in behavior and environmental exposures of individuals. On the other hand, snoring and frequent cold have lower means and narrower ranges, indicating less clinical relevance. The table in general gives a measure of the importance in distribution and which features have more impact on disease severity.

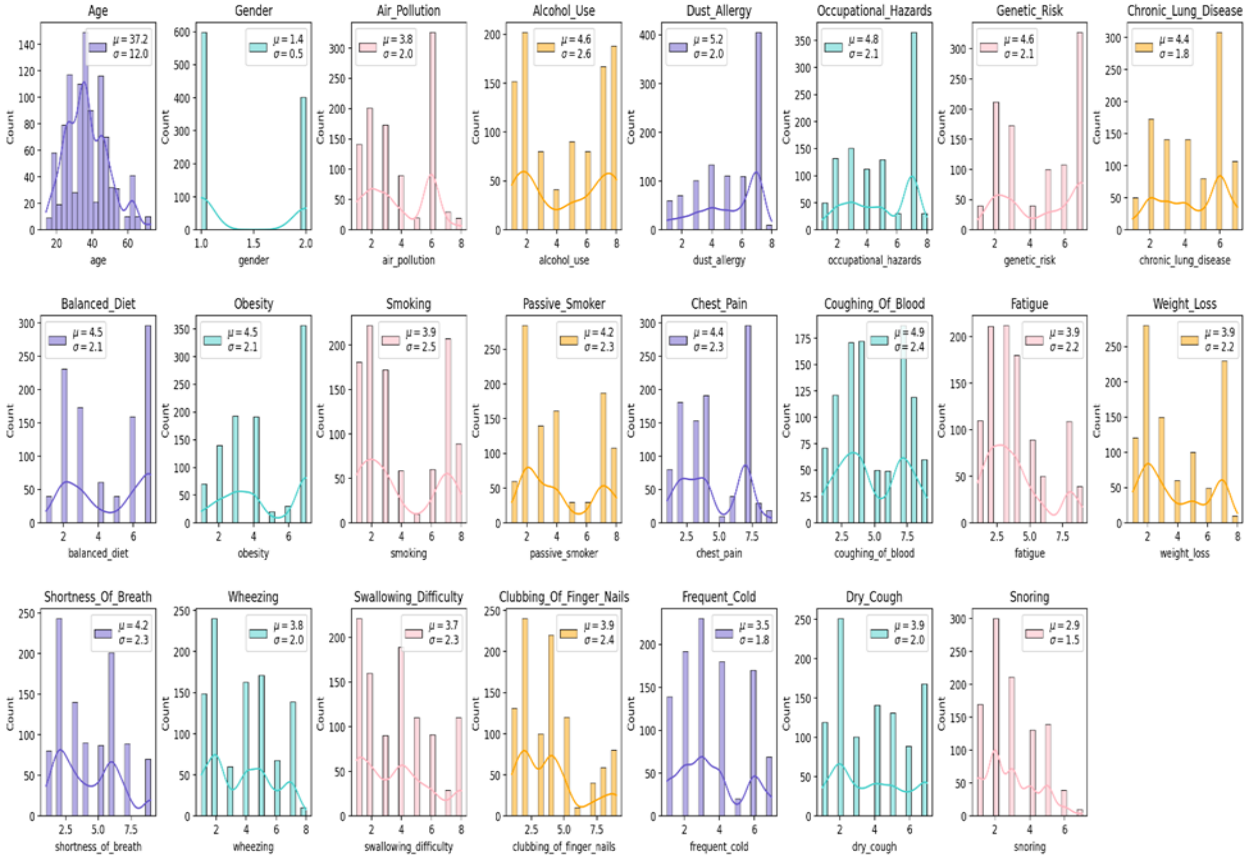

Figure 3: Distribution and Density Patterns of Clinical, Demographic, and Behavioral Features in the Lung Cancer Dataset

Figure 3 illustrates distribution and density patterns of clinical, demographic, and behavioral features, revealing variability, skewness, and symptom intensity differences that help distinguish lung cancer risk levels across the dataset.

4.3 Correlation-Based Result Analysis

The relationships between these clinical, behavioral and demographic variables on the prediction of lung cancer risk are best observed in the correlation matrix. There are strong positive correlations between air pollution, dust allergy, occupational exposure, genetic risk and chronic lung disease with chest pain and coughing blood suggesting that these factors tend to co-occur and contribute together toward disease severity. Likewise, behavioral indicators including smoking and passive smoking correlate moderately with multiple respiratory symptoms. Demographics such as age and gender, however, show relatively weak associations implying that there is little direct role on the risk. In general, the matrix illustrates groups of correlated factors which are important for prediction. Strongest and most relevant correlations of lung cancer–related features are summarized in Table 6. Positive associations that are markedly high, for instance between dust allergy and genetic risk as well as chronic lung disease, reflect these interactions between respiratory and hereditary features. Mild associations as between smoking and passive

smoking or frequent cold with dry cough are indicative of exposure overlap and symptom replication. In sum, these associations demonstrate two important interrelated predictors of lung cancer risk.

Table 6: Key Correlation Findings in Lung Cancer Risk Features

Feature Pair	Correlation Value	Relationship Strength	Interpretation
Air Pollution ↔ Dust Allergy	0.82	Strong	High pollution increases allergic respiratory reactions.
Dust Allergy ↔ Genetic Risk	0.88	Very Strong	Strong predisposition to developing allergies and lung issues.
Genetic Risk ↔ Chronic Lung Disease	0.83	Strong	Genetic factors heavily influence chronic lung conditions.
Occupational Hazards ↔ Chronic Lung Disease	0.78	Strong	Workplace exposure plays a major role in lung damage.
Chest Pain ↔ Coughing of Blood	0.81	Strong	Severe respiratory symptoms tend to co-occur.
Shortness of Breath ↔ Fatigue	0.47	Moderate	Breathing difficulty contributes to increased tiredness.
Smoking ↔ Passive Smoker	0.66	Moderate–Strong	Environments with smokers raise secondhand smoke exposure.
Balanced Diet ↔ Obesity	0.75	Strong	Poor diet quality strongly correlates with obesity levels.
Frequent Cold ↔ Dry Cough	0.52	Moderate	Recurrent colds often lead to ongoing coughing symptoms.
Chronic Lung Disease ↔ Chest Pain	0.62	Moderate–Strong	Lung conditions often manifest as persistent chest pain.

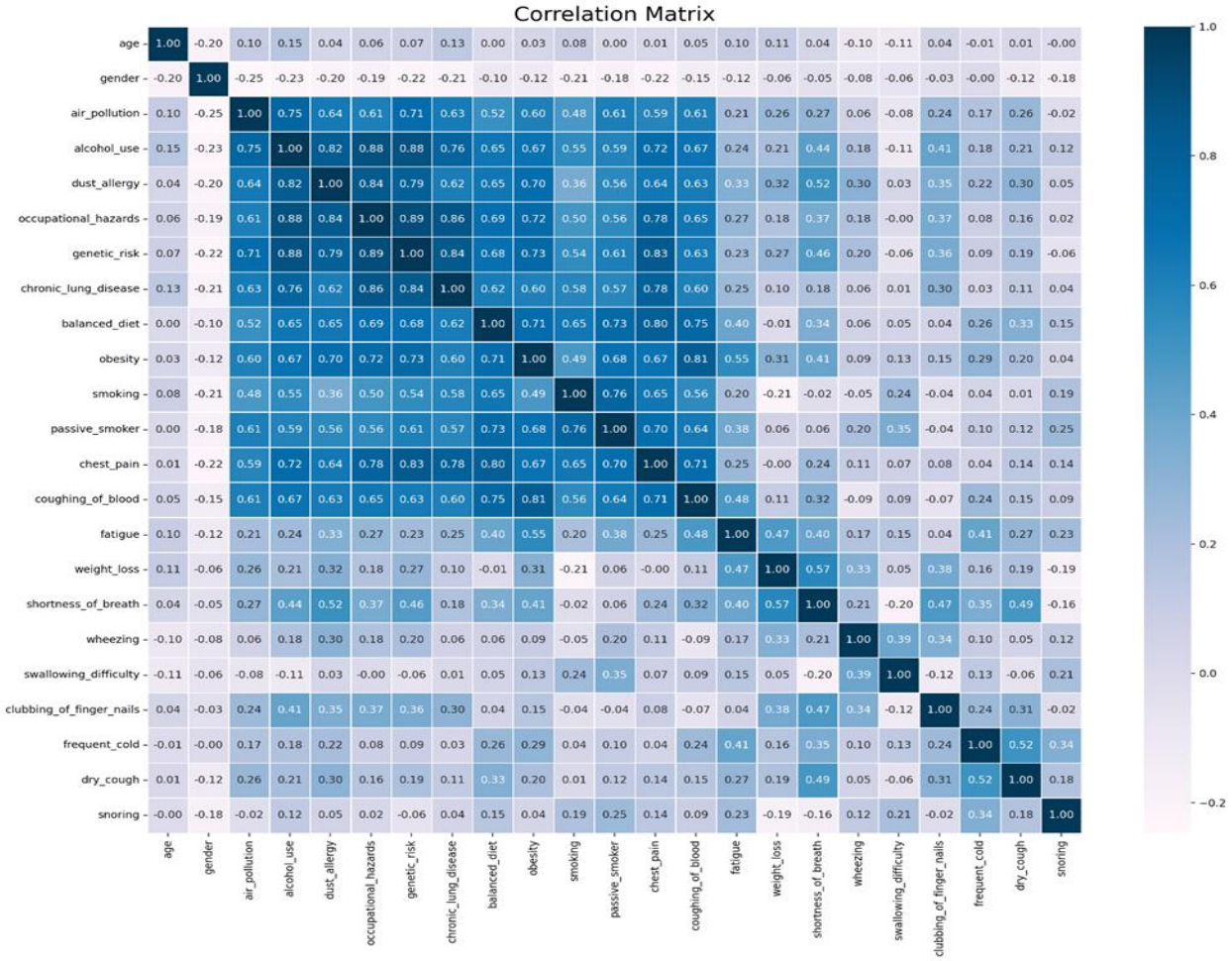

Figure 4: Correlation Matrix of Clinical, Behavioral, and Demographic Features

In Figure 4, the correlation matrix shows strong associations between various clinical and environmental characteristics. Air pollution, dust allergy, genetic risk factors, chronic lung disease and the symptom coughing of blood have relatively high positive correlations with each other It suggests that they will often occur simultaneously. Age and sex, in contrast, are only very weakly related or unrelated. In the end, the matrix makes transparent critical interrelated drivers of lung cancer risk.

5. CONCLUSION

This article illustrates the utility of combining exploratory data visualization, numerical analysis and correlation testing in order to develop an enhanced understanding of the clinical, behavioral, and environmental features associated with lung cancer risk. A consistent and highly significant influence of core respiratory symptoms such as chest pain, coughing of blood, shortness of breath and wheeze was seen even in high-risk individuals indicating their strong predictive properties. Behavioral and exposure-related factors such as smoking, passive exposure to smoke, air pollution and occupational hazards also have broad distribution ranges in the human population and evident elevating trends which implies that environmental and life style related factors make up a major part of risk elevation. The correlation analysis also shows clusters of associated features, including those between genetic predisposition and chronic lung disease, dust allergy, and respiration symptoms, suggesting that there are interconnections among them. In summary, the study supports that the risk of lung cancer results from interaction among symptom severity, clinical comorbidities and persistent environment exposure, which confirms the utility of data-driven approaches for preventive early screening and risk stratification.

5.1 Limitations

Limitations Though valuable insight was gained there are several limitations that must be noted. The sample size is small and could not fully represent the diversity in population. Clinical characteristics are self-reported and may be subject to reporting errors or recall bias. The dataset does not contain radiological or genomics information, constraining our efforts for a multimodal analysis. Second, the cross-sectional characteristic of the dataset does not allow to examine symptom development or temporal patterns. These constraints advise cautious interpretations of model-based predictions when extrapolating to additional populations.

5.2 Future Work

Further investigations should be carried out on larger, different populations and populations from diverse geographic areas to strengthen the results. Linking imaging information like CT images, genomic data and biomarker profiles should allow for more thorough multimodal prediction models. Sophisticated AI models, e.g., ensemble hybrid systems, deep neural networks or explainable AI frameworks can be used to increase prediction accuracy and interpretability of the model. Similarly, it would be interesting to have longitudinal studies following these patients over time- to see how their symptoms evolve and disease progresses. Lastly, creating more intuitive clinical dashboards or mobile health applications may facilitate real-time risk estimates and decisions at point of care, thereby potentially leading to earlier diagnosis and better outcomes for patients.

REFERENCES

[1] Cui, J. (2025). *A study on the risk factors of lung cancer using machine learning methods*. In Proceedings of the 2025 3rd International Conference on Image, Algorithms, and Artificial Intelligence (ICIAAI 2025) (pp. 522–529). Atlantis Press. https://doi.org/10.2991/978-94-6463-823-3_52

[2] Lee, P.-C., Lin, M.-W., Liao, H.-C., Lin, C.-Y., & Liao, P.-H. (2025). Applying machine learning to construct an association model for lung cancer and environmental hormone high-risk factors and nursing assessment reconstruction. *Journal of Nursing Scholarship, 57*(1), 140–151. https://doi.org/10.1111/jnu.12997

[3] Smyth, R., & Billatos, E. (2024). *Novel strategies for lung cancer interventional diagnostics*. Journal of Clinical Medicine, 13(7207). https://doi.org/10.3390/jcm13237207

[4] Firdaus, Q., Sigit, R., Harsono, T., & Anwar, A. (2020). *Lung cancer detection based on CT-scan images with detection features using gray level co-occurrence matrix (GLCM) and support vector machine (SVM) methods*. In 2020 International Electronics Symposium (IES) (pp. 643–650).

[5] He, K., Zhang, X., Ren, S., & Sun, J. (2016). *Deep residual learning for image recognition*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 770–778). https://doi.org/10.1109/CVPR.2016.90

[6] Beigi, P., McWilliams, A., Srinidhi, A., Lam, S., & MacAulay, C. E. (2015). smoking status effects on the early detection of early lung cancer in high-risk smokers using an electronic nose. *IEEE Transactions on Biomedical Engineering*. https://doi.org/10.1109/TBME.2015.2409092

[7] Aberle, D. R., Adams, A. M., Berg, C. D., Clapp, J. D., & Fagerstrom, R. M. (2011). Reduced lung-cancer mortality with low-dose computed tomographic screening. *New England Journal of Medicine, 365*(5), 395–409. https://doi.org/10.1056/NEJMoa1102873

[8] Alakwaa, W., Nassef, M., & Badr, A. (2017). Lung cancer detection and classification with 3D convolutional neural networks. *International Journal of Computer Applications, 157*(6), 1–5.

[9] Amma, G. K., & George, S. (2020). Predicting lung cancer using machine learning algorithms. *International Journal of Advanced Computer Science and Applications, 11*(9), 405–411.

[10] Armato, S. G., McLennan, G., Bidaut, L., McNitt-Gray, M. F., & Meyer, C. R. (2011). The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI). *Medical Physics, 38*(2), 915–931.

[11] Choi, H., Park, C. M., & Goo, J. M. (2018). Deep learning for pulmonary nodule detection in CT imaging. *European Radiology, 28*, 451–460. https://doi.org/10.1007/s00330-017-5005-5

[12] Dey, N., Ashour, A. S., & Balas, V. E. (2019). *Smart medical data sensing and IoT systems*. Springer.

[13] Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning, 63*(1), 3–42.

[14] Hawkins, S., Wang, H., Liu, Y., & Roth, H. (2016). Predicting malignant nodules from CT scans using deep learning. *Radiology, 280*(3), 772–781.

[15] Kim, H., Goo, J. M., & Lee, K. (2019). Deep learning-based risk classification for lung cancer screening. *Journal of Thoracic Imaging, 34*(6), 393–402.

[16] Kumar, D., Wong, A., & Taylor, G. (2015). Lung nodule classification using deep features in CT images. *Proceedings of the IEEE International Conference on Image Processing*, 162–166.

[17] Liao, F., Liang, M., Li, Z., & Hu, X. (2019). Evaluate deep learning models on LIDC-IDRI database. *IEEE Access, 7*, 32867–32875.

[18] Moghbel, M., Hamarneh, G., & Abugharbieh, R. (2017). 3D segmentation and classification of lung nodules using convolutional neural networks. *Medical Image Computing and Computer-Assisted Intervention*, 21, 665–673.

[19] Sarkar, M., & Madhavan, C. (2020). Machine learning models for lung cancer survival prediction. *BMC Medical Informatics and Decision Making, 20*(1), 1–13.

[20] Setio, A. A. A., Traverso, A., de Bel, T., & et al. (2017). Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in CT images. *IEEE Transactions on Medical Imaging, 36*(7), 143–155.