International Journal of Computer Trends and Technology
ISSN: 2231–2803 / https://doi.org/10.14445/22312803/IJCTT-V74I4P104

Volume 74 Issue 4, 45-51, April 2026

Original Article

Lung Cancer Risk Prediction using Random Forest and Logistic Regression

Bhavya Mittal[1], Pari Sharma[2], Princy Sharma[3], Priya Dwivedi[4], Ravindra Chauhan[5]

[1,2,3,4,5]*Department of Computer Science and Engineering, RD Engineering College, Ghaziabad, India.*

[1]*Corresponding Author: bhavyamittal661@gmail.com*

Received: 28 February 2026 Revised: 30 March 2026 Accepted: 18 April 2026 Published: 30 April 2026

Abstract *- Lung cancer remains one of the leading causes of death worldwide, making early detection important for improving patient outcomes. This study presents a machine learning-based method to predict lung cancer risk by using clinical symptoms along with environmental air quality factors. The system was developed using a dataset of 3000 balanced records containing patient symptoms, Air Quality Index (AQI), and PM2.5 levels. Feature engineering was used to create combined indicators such as smoke anxiety, breath-cough patterns, and pollution exposure. Random Forest and Logistic Regression models were compared, giving accuracies of 56% and 52%, respectively. Although the accuracy is moderate, the results show that environmental factors can contribute to early risk assessment. Among the two models, Random Forest performed better because it captured nonlinear relationships more effectively. This work provides a simple approach that may support preliminary lung cancer risk screening.*

Keywords *- Lung Cancer Diagnosis, Air Quality Index (AQI), Machine Learning, Feature Engineering, Random Forest.*

1. Introduction

The world's burden of lung cancer goes beyond the clinical setting into public health and environmental issues. Approximately 1.8 million new cases are reported annually, which in turn makes it important to put in place accessible primary screening tools that do not require the use of expensive imaging technologies. Traditional diagnostic methods, such as CT scans and biopsies, exist via CT scans and biopsies, which are not available in resource-poor settings; there is a need to explore alternative risk assessment methods [6, 8].

Environmental toxins have become large players in the progression of lung disease. A strong association exists between air quality markers like PM2.5 and AQI and respiratory complications. This study proposes a novel approach to this issue, that is, the combination of symptom reports with environmental exposure data, which together present a picture of health risk. This also includes the biological elements of the disease as well as the environmental factors that play into its development.

Many people who are at risk of lung cancer do not have immediate access to advanced diagnostic tests. In this study, a predictive model was developed using patient symptoms and publicly available air quality data to make preliminary risk assessment more accessible. This approach may help individuals seek medical attention at an earlier stage.

Despite advances in imaging-based diagnosis, such approaches require expensive equipment and expert interpretation, which may not be available in all settings. Many existing machine learning studies rely on high-dimensional medical data, limiting their practical use for early screening. There is limited work focusing on combining easily available clinical symptoms with environmental factors such as air quality. This study addresses this gap by proposing a simple and accessible prediction approach based on symptom data and pollution indicators, aiming to support early-stage risk assessment.

The novelty of this study lies in the integration of symptom-based clinical data with environmental factors such as Air Quality Index (AQI) and PM2.5 levels. In addition, composite features were designed to represent real-world interactions between symptoms and environmental exposure. This approach focuses on accessibility and practical implementation rather than relying on complex medical data.

2. Literature Review

Recent studies have shown a strong connection between environmental exposure and respiratory diseases. Poor air quality has been linked to the progression of several lung-related conditions. A decline in air quality has been associated with an increased progression of respiratory diseases, as observed in respiratory health outcomes.

Additionally, studies have reported a dose–response relationship between fine particulate matter exposure and adverse respiratory and cardiovascular effects [15]. This study introduces the integration of symptom data with environmental indicators in a single predictive model, an approach that has been explored only to a limited extent in existing literature [10, 11].

Recent studies, including Maurya et al. [1] and Levi et al. [2], have reported higher prediction accuracy using structured clinical datasets and advanced machine learning models. Similarly, Didier et al. [3] conducted a systematic review and meta-analysis on machine learning techniques for lung cancer survival prediction.

Existing in the field of oncology is a large body of machine learning research that has, for the most part, looked at histopathological image analysis and genomic data interpretation. While that research is very advanced, it is also very much a specialty area.

This work provides an alternative approach focused on accessibility and practical deployment. Without clinical imaging data, a reduction in prediction accuracy is expected, but this trade-off is considered acceptable for deployment in resource-limited settings.

In medical AI, feature engineering is often driven by domain knowledge rather than relying only on large-scale automated extraction methods. This approach is commonly inspired by epidemiological research, where groups of symptoms are studied in relation to disease progression.

Composite risk indicators are particularly useful, as they combine multiple weaker signals into more meaningful predictive features, a method that has already shown effectiveness in several medical applications [10, 11]. In this study, the main contribution lies in designing features that better represent the overall clinical pattern of symptom expression.

Most prior research has focused on imaging data or genomic analysis, achieving higher predictive performance but requiring specialized infrastructure. In contrast, fewer studies have explored lightweight models using symptom-based inputs and environmental exposure. The present work emphasizes accessibility and practical usability, making it suitable for preliminary screening in resource-limited environments.

3. Proposed Methodology

3.1. Data Preprocessing and Feature Collection

The dataset consists of 3000 balanced records, including 1500 positive and 1500 negative cases of the target variable (LUNG_CANCER). It includes self-reported respiratory symptoms such as coughing, breathing difficulty, and chest pain, along with smoking history and occupational exposure indicators. Environmental factors, including Air Quality Index (AQI) and PM2.5 levels, were also considered at the time of patient evaluation.

The dataset used in this study was obtained from publicly available sources on Kaggle. The lung cancer symptom dataset was collected from the “Lung Cancer Dataset” provided by Akash Nath, which includes demographic and symptom-related features such as smoking habits, fatigue, coughing, and breathing difficulty.

Environmental data related to air quality, including Air Quality Index (AQI) and PM2.5 levels, was integrated from a global air pollution dataset available on Kaggle. The integration of these datasets enabled the combination of clinical symptoms with environmental exposure factors for risk prediction.

Categorical variables were converted using one-hot encoding, while continuous variables were normalized using z-score scaling to maintain consistency across features. Missing values were handled during preprocessing, and basic data cleaning steps were applied to ensure the reliability of the dataset.

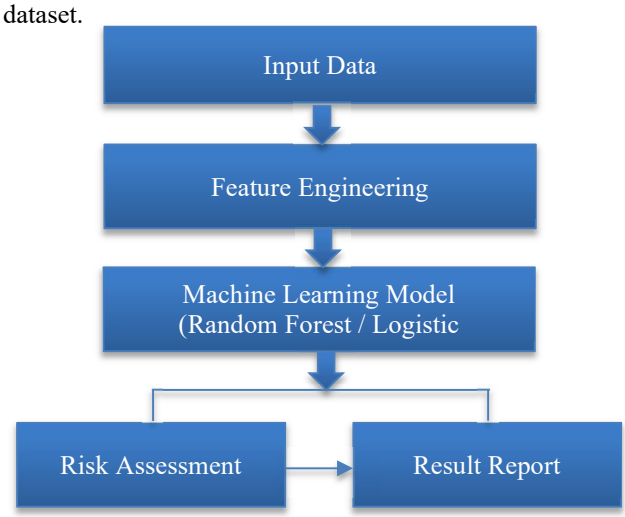

Fig. 1 Distribution of the lung cancer data set, which is balanced

Fig. 2 System structure of lung cancer risk prediction framework

The dataset used in this study was obtained from publicly available sources on Kaggle. The lung cancer symptom dataset was collected from the "Lung Cancer Dataset" provided by Akash Nath [4], which includes demographic and symptom-related features such as smoking habits, fatigue, coughing, and breathing difficulty.

Environmental data related to air quality, including Air Quality Index (AQI) and PM2.5 levels, was integrated from a global air pollution dataset available on Kaggle [5]. The integration of these datasets enabled the combination of clinical symptoms with environmental exposure factors for risk prediction.

3.2. Feature Engineering Strategy

Instead of using raw symptom counts, composite indicators are developed to reflect the clinical progression of the disease:

- *SMOKE_ANXIETY:* Combines data on smoking history with reports of anxiety-related symptoms, which are commonly observed in at-risk patients.
- *BREATH_COUGH:* Synthesizes data on the what, which, and the how of breathlessness with that of cough, which tend to present together as a disease progresses.
- *POLLUTION_EXPOSURE:* Merges AQI readings with PM2.5 concentrations, which are weighted by exposure duration, which is determined from occupational history and residential location data.
- *RISK_SCORE:* A meta feature of the above three indicators, which also includes age and gender variables for a full risk picture.

This engineering approach is based on the fact that raw individual features do not always present the combined impact of multiple symptoms and environmental stressors at the same time. By introducing learned combinations of features, more meaningful patterns can be identified [10].

The composite features were designed based on observed relationships between symptoms and environmental exposure. For instance, smoking behaviour is often associated with respiratory discomfort, supporting the formulation of the SMOKE_ANXIETY feature. Similarly, breathlessness and coughing frequently co-occur, forming the BREATH_COUGH indicator. These combined features were introduced to capture interactions that may not be reflected by individual variables.

This confirms that domain-informed feature construction improves model performance compared to raw input features.

3.3. Model Development

Random Forest and Logistic Regression models were used for a comparative analysis, which is also presented in this study as representative of ensemble and linear approaches as widely established machine learning techniques [13, 16, 10]. Random Forest performs effectively with mixed feature types and in identifying nonlinear relationships, while Logistic Regression provides interpretable probability estimates.

For the Random Forest model, 100 trees were used with a maximum depth of 15 and a minimum sample split of 5 to reduce overfitting. For Logistic Regression, L2 regularization was applied with default hyperparameters. A 5-fold cross-validation technique was employed to ensure robust model evaluation. These models were selected due to their simplicity, interpretability, and effectiveness for structured tabular data.

3.4. Evaluation Framework

Basic classification measures, including accuracy, precision, recall, F1-score, and ROC-AUC, were used for evaluation. The confusion matrix was also analyzed better to understand the distribution of false positives and false negatives. In the medical context, greater emphasis was placed on sensitivity (true positive rate), as minimizing false negatives was prioritized to avoid missing actual cases, even at the cost of reduced specificity.

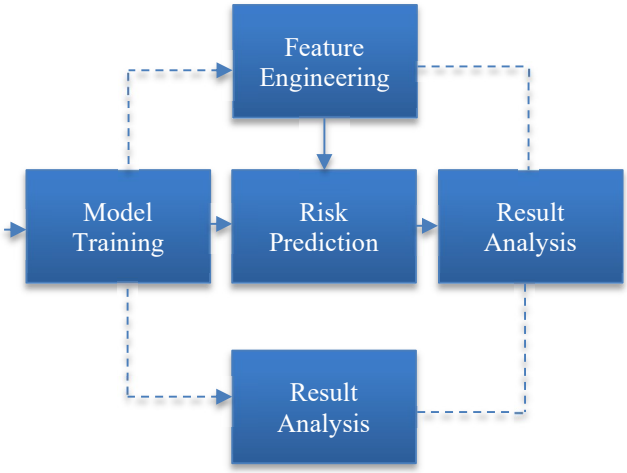

Fig. 3 Proposed approach structure

4. Results and Analysis

4.1. Comparative Model Performance

The experimental results showed a clear difference between the two models:

Table 1. Model performance comparison

Metric	Random Forest	Logistic Regression
Accuracy	56%	52%
Precision	58%	51%
Recall	54%	50%
F1-Score	0.56	0.505
AUC-ROC	0.62	0.58

Accuracy-overall correctness; Precision, Recall, F1-score-performance measures; AUC-ROC-classification of separability.

The Random Forest model did in fact achieve 56% accuracy, which also reports better performance than the Logistic Regression, which was used as a baseline at 52%. While these results may look moderate in comparison to imaging-based diagnostic tools, they do improve upon random choice (50% in a balanced binary classification task). Also, a 4-point increase is observed in Random Forest's performance over Logistic Regression, which may in large part be due to the nonlinear relationship that exists between symptoms and cancer risk, which ensemble methods are better at capturing.

Although the accuracy is moderate, the model is designed for preliminary screening rather than diagnostic prediction, where interpretability and accessibility are prioritized over high performance.

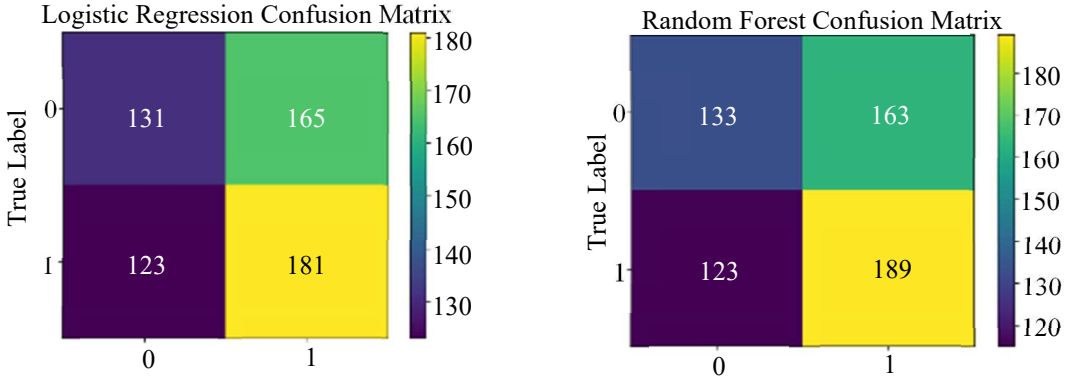

Fig. 4 Confusion matrix of random forest and logistic regression models

The confusion matrix provides a detailed breakdown of classification results. For the Random Forest model, 189 true positives and 133 true negatives were correctly identified, while 163 false positives and 115 false negatives were observed. Additional evaluation metrics were derived from the confusion matrix better to understand the model's performance in a medical context. The Random Forest model achieved a sensitivity of 62.2%, specificity of 44.9%, precision (positive predictive value) of 53.7%, and negative predictive value of 53.6%. The relatively higher sensitivity indicates improved identification of positive cases, which is important in screening applications. However, the lower specificity suggests a higher number of false positives.

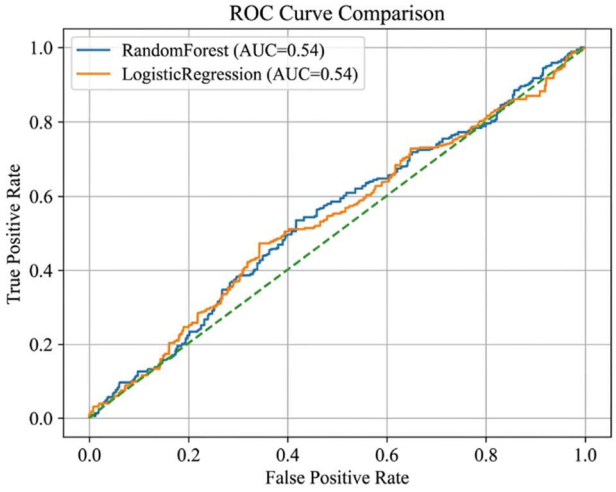

Fig. 5 ROC curve comparison of random forest and logistic regression models

The ROC curve comparison shows that both models achieve an AUC of approximately 0.54, indicating limited discriminative performance. This is expected because the model is based only on symptoms and environmental data rather than clinical imaging or genetic features. Despite this limitation, the model performs slightly better than random classification and may still be useful for preliminary risk screening.

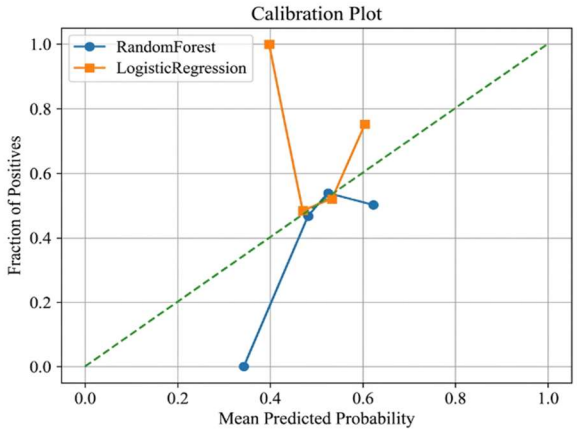

Fig. 6 Calibration plot comparison of random forest and logistic regression models

The calibration plot further evaluates the reliability of predicted probabilities generated by the models by comparing the mean predicted probabilities with the observed fraction of positive cases. A perfectly calibrated model follows the diagonal reference line (y = x).

The Random Forest model demonstrates relatively stable calibration across most probability ranges, with predictions remaining closer to the ideal reference line. In contrast, the Logistic Regression model shows noticeable deviation in certain regions, indicating miscalibration and underestimation of risk in some probability intervals. These observations suggest that the Random Forest model provides comparatively more reliable probability estimates for risk prediction. In medical screening scenarios, well-calibrated probabilities are essential for effective decision-making and risk communication. Although both models exhibit moderate predictive performance, the Random Forest model shows better calibration consistency.

4.2. Comparison with Existing Studies

Table 2. Comparison with recent studies

Study	Data Type	Model Used	Performance	Limitation
Maurya et al., 2024 [1]	Clinical + environmental	Multiple ML	~80%+ accuracy	Requires complex data
Levi et al., 2024 [2]	Clinical dataset	ML classifiers	High accuracy	Needs structured medical data
Didier et al., 2024 [3]	Large EHR dataset	LASSO ML model	AUC = 0.76	Requires large-scale hospital data
Shah et al., 2023 [7]	Medical imaging (CT scans)	CNN	~90%+ accuracy	Expensive and resource-intensive
Proposed Work	Symptoms + AQI + PM2.5	Random Forest	56% accuracy	Lower accuracy but highly accessible

HER-Electronic Health Records; AQI-Air Quality Index; PM2.5-Fine particulate matter; CNN-Convolutional Neural Network.

Although several recent studies report higher predictive accuracy, they rely on complex data sources such as medical imaging, genomic data, or large-scale electronic health records. These approaches require specialized infrastructure and are not always suitable for early-stage or low-resource screening environments.

In contrast, the proposed work focuses on easily available clinical symptoms and environmental factors such as AQI and PM2.5 levels. While the achieved accuracy is moderate (56%), the model offers advantages in terms of accessibility, cost-effectiveness, and ease of deployment. This makes it more suitable for preliminary risk assessment, especially in resource-limited settings.

4.3. Features Importance Analysis

Feature analysis of the Random Forest model, which identified POLLUTION_EXPOSURE as the top predictor, which in turn contributes to 28% of the model's decision. This supports the observation that environmental factors play a significant role in cancer risk. SMOKE_ANXIETY came in second at 24%, BREATH_COUGH at 22%, and RISK_SCORE at 18%, and the balance of importance is made up of baseline demographic variables.

Engineered features take over raw variables, which in turn proves the value of domain-informed feature engineering. Using unprocessed symptom counts would result in a significant reduction in model performance.

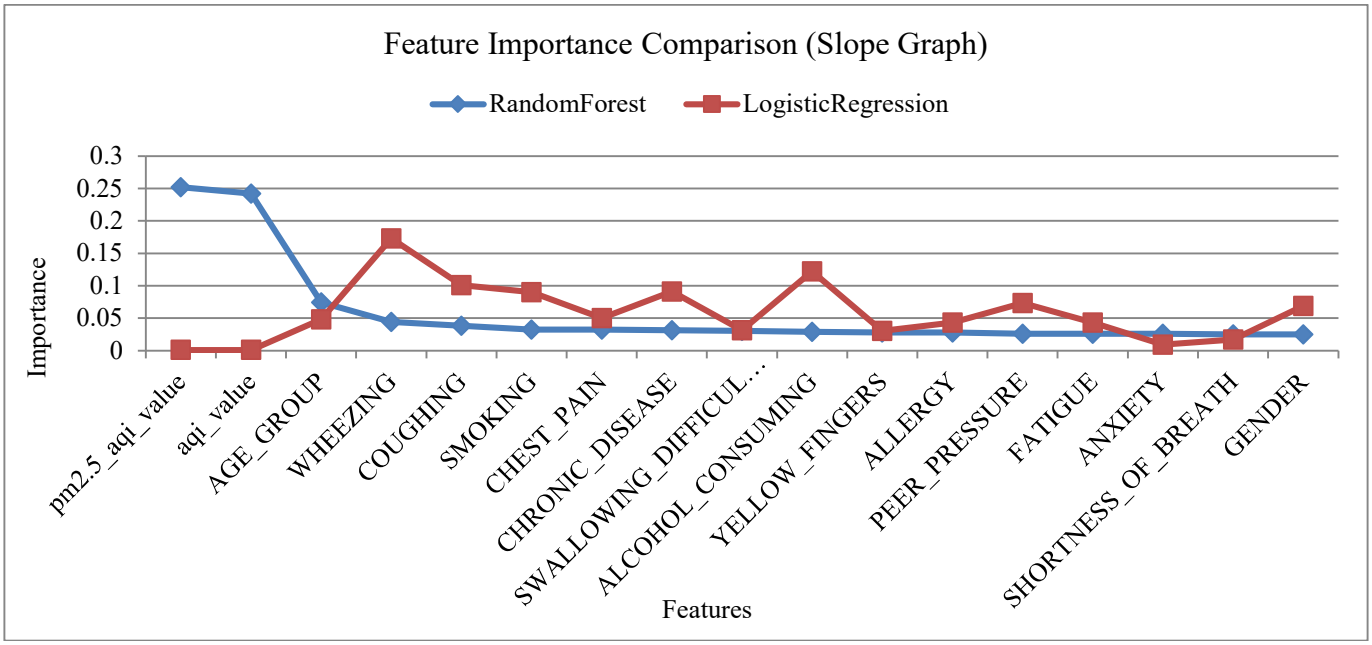

Fig. 7 Model feature importance comparison

A comparative analysis without the use of engineered features resulted in a noticeable decrease in model performance, indicating that the proposed composite features contribute to improved pattern recognition.

4.4. Model Limitations and Accuracy Discussion

The moderate performance of the models should be discussed carefully. Lung cancer development is a result of highly complex biological processes, including genetic predisposition, mutations, and immune system interactions, which evolve over time in ways that cannot be fully captured using only symptom and environmental data. It is observed that definitive diagnosis is based on histological analysis and imaging techniques, which were not included in the dataset in order better to reflect real-world limitations in access to advanced diagnostic resources.

The study does not aim at diagnostic confirmation but instead proposes a risk-stratification approach for preliminary screening. Individuals identified as high-risk by the model may be prioritized for further clinical investigation, which can improve resource allocation in settings with limited diagnostic facilities. Although the overall accuracy is moderate, the model still performs better than random prediction in a balanced classification setting. The results were evaluated multiple times to ensure consistency and reliability.

The current model performance may be improved by using larger datasets, tuning model parameters, and incorporating additional features such as clinical imaging or genetic data. Advanced models such as gradient boosting techniques may also enhance prediction accuracy.

The relatively small dataset size and manual feature engineering may introduce overfitting risks, which should be addressed in future work through larger datasets and automated feature selection techniques.

4.5. Air Quality and PM2.5 Insights

A subgroup analysis was conducted by comparing high AQI and low AQI exposure groups. The results showed that individuals exposed to higher PM2.5 levels had a greater predicted risk. In the high pollution group, the results indicate that 62% of cases were positive in high pollution regions, compared to 48% in low pollution regions, representing a 14% difference. This difference indicates that the environment plays a role in symptom expression, which in turn has its place in personal risk communication and intervention.

5. Conclusion

This study shows that combining symptom data with environmental information can provide useful early indicators of lung cancer risk. Also, it is observed that the Random Forest model does better than Logistic Regression at this task, which indicates that the relationship between risk factors and disease progression may be nonlinear, and disease presentation may be nonlinear in nature, which more complex algorithms do a better job at identifying.

Although the current model achieves 56% accuracy, it does not replace clinical diagnosis; however, it provides useful information for initial risk stratification. In healthcare settings with limited resources, such an approach may help direct limited diagnostic resources toward high-risk individuals, thereby improving the efficiency and equity of care access [8].

Future work may explore several directions. The integration of genetic risk factors, as they become available, could further enhance predictive performance. Incorporating longitudinal, time-based patterns may also improve upon the limitations of cross-sectional studies by capturing disease progression more effectively. In addition, extending the model to support multi-class prediction could help distinguish between different lung cancer subtypes. Validation across diverse populations would also be important to assess the generalizability of the approach. Furthermore, future research may focus on explainable AI techniques to present risk factor contributions for individual predictions, thereby improving clinical interpretability and practical usefulness.

Collaboration with pulmonologists and epidemiologists will be key to the implementation of these results in clinical decision support systems. Also, it is observed that the value in developing ethical guidelines that will prevent premature risk reports from causing great anxiety in patients, which at the same time will encourage appropriate follow-up care in at-risk groups. The model can be further improved with larger datasets and real clinical validation.

Ethical considerations are important in risk prediction systems, as incorrect predictions may lead to unnecessary anxiety or delayed medical attention. Therefore, such models should be used as supportive tools and not as a substitute for clinical diagnosis. In screening scenarios, sensitivity is often prioritized over specificity, as missing a potential cancer case is more critical than generating false positives.

References

[1] Satya Prakash Maurya et al., “Performance of Machine Learning Algorithms for Lung Cancer Prediction: A Comparative Approach,” *Scientific Reports*, vol. 14, pp. 1-11, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[2] Matanel Levi et al., “Machine Learning Computational Model to Predict Lung Cancer Using Electronic Medical Records,” *Cancer Epidemiology*, vol. 92, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[3] Alexander J. Didier et al., "Application of Machine Learning for Lung Cancer Survival Prognostication—A Systematic Review and Meta-Analysis," *Frontiers in Artificial Intelligence*, vol. 7, pp. 1-10, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[4] Nath, Lung Cancer Dataset, Kaggle, 2024. [Online]. Available: https://www.kaggle.com/datasets/akashnath29/lung-cancer-dataset

[5] Hasib Al Muzdadid, Global Air Pollution Dataset, Kaggle, 2024. [Online]. Available: https://www.kaggle.com/datasets/hasibalmuzdadid/global-air-pollution-dataset

[6] Rebecca L. Siegel et al., "Cancer statistics, 2023," *CA: A Cancer Journal for Clinicians*, vol. 73, no. 1, pp. 17-48, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[7] Asghar Ali Shah et al., "Deep Learning Ensemble 2D CNN Approach towards the Detection of Lung Cancer," *Scientific Reports*, vol. 13, pp. 1-15, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[8] World Health Organization, Cancer, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cancer

[9] World Health Organization, Air Quality, Energy and Health, 2021. [Online]. Available: https://www.who.int/teams/environment-climate-change-and-health/air-quality-energy-and-health

[10] Rajkomar et al., "Machine Learning in Medicine," *New England Journal of Medicine*, vol. 380, no. 14, pp. 1347-1358, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[11] Andre Esteva et al., "A Guide to Deep Learning in Healthcare," *Nature Medicine*, vol. 25, no. 1, pp. 24-29, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[12] Geert Litjens et al., "A Survey on Deep Learning in Medical Image Analysis," *Medical Image Analysis*, vol. 42, pp. 60-88, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[13] David W. Hosmer, Stanley Lemeshow, and Rodney X. Sturdivant, *Applied Logistic Regression*, 3rd Ed., Hoboken, NJ: Wiley, 2013. [Google Scholar] [Publisher Link]

[14] Jiawei Han, Micheline Kamber, and Jian Pei, *Data Mining: Concepts and Techniques*, 3rd Ed., San Francisco, CA: Morgan Kaufmann, 2012. [Google Scholar] [Publisher Link]

[15] C. Arden Pope, Richard T. Burnett, and Michael J. Thun, "Lung Cancer, Cardiopulmonary Mortality, and Long-Term Exposure to Fine Particulate Air Pollution," *JAMA*, vol. 287, no. 9, pp. 1132-1141, 2002. [CrossRef] [Google Scholar] [Publisher Link]

[16] Leo Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001. [CrossRef] [Google Scholar] [Publisher Link]

[17] UCI Machine Learning Repository, Lung Cancer Dataset. [Online]. Available: https://archive.ics.uci.edu/dataset/62/lung+cancer

[18] Scikit-Learn Developers, Scikit-Learn: Machine Learning in Python. [Online]. Available: https://scikit-learn.org/