Evaluating Environmental and Physical Factors for Lung Cancer Risk Prediction: A Machine Learning Analysis

Ujan Ganguli

Student, Vellore Institute of Technology, Chennai Campus, Chennai, India

Saahith

Student, Vellore Institute of Technology, Chennai Campus, Chennai, India

Abdulla Swalih NM

Student, Vellore Institute of Technology, Chennai Campus, Chennai, India

Illavarsi AK

Assistant Professor, Vellore Institute of Technology, Chennai Campus, Chennai, India

Abstract

Lung cancer remains a leading cause of cancer-related deaths worldwide, emphasizing the critical need for effective risk assessment and early detection strategies. This study investigates the potential of leveraging environmental and physical factors to predict individual lung cancer risk using advanced machine learning models.

Two distinct datasets were analysed - one containing environmental factors and another focusing on physical attributes. The results suggest that the physical subset may be a stronger indicator of lung cancer risk, as evidenced by the higher average probability of "High Risk" classification and improved performance of the machine learning models.

Gradient Boosting demonstrated the highest accuracy of 0.895 for the environmental dataset. Notably, all three modelling techniques - Random Forest, Logistic Regression, and Support Vector Machines - achieved remarkably high accuracies of 1.0, 0.835, and 0.985, respectively, when applied to the physical dataset.

These findings highlight the promising potential of machine learning-based approaches for lung cancer risk assessment, particularly when incorporating physical factors. However, the study also underscores the need for further research with more comprehensive datasets and the exploration of more sophisticated modelling techniques. Such advancements could lead to enhanced accuracy and generalizability of the risk prediction models, ultimately contributing to improved early detection and personalized management strategies for this deadly disease.

Keywords: *Environmental Factors, Physical Factors, Gradient Boosting, Random Forest, Logistic Regression, Support Vector Machines (SVM), Accuracy, Generalizability, Risk Prediction, Machine Learning Models*

Introduction

Lung cancer, a devastating malignancy, remains a global health crisis. Despite advancements in treatment options, it is the leading cause of cancer-related deaths worldwide [1]. Early detection and accurate risk assessment are paramount for improving patient outcomes. Traditional risk factors, such as smoking history, are well-established, but a comprehensive understanding of the interplay between environmental and physical factors contributing to lung cancer remains elusive. This study delves into this intricate relationship by leveraging the power of machine learning to analyse environmental and physical datasets and assess their potential for predicting lung cancer risk.

The pursuit of a deeper understanding of lung cancer risk factors is multifaceted. Environmental factors encompass a broad spectrum of influences, including air pollution, exposure to occupational hazards, and dietary patterns. These external elements can potentially trigger or exacerbate underlying biological processes leading to lung cancer development [2]. Physical factors, on the other hand, encompass a patient's individual characteristics, including age, family history, body mass index (BMI), and presence of chronic lung diseases. These inherent attributes can create a susceptibility to environmental insults or independently contribute to lung cancer initiation [3]. While research has explored both environmental and physical factors in isolation, a holistic approach that considers their combined effect is crucial for a more robust risk assessment strategy.

Machine learning (ML) offers a powerful tool for analysing complex datasets and identifying patterns that might be missed by traditional statistical methods. By training algorithms on labelled data, ML models can learn to classify new data points with a high degree of accuracy. In the context of lung cancer risk assessment, ML models can be trained on data containing both environmental and physical factors linked to lung cancer diagnosis. These models can then analyse a patient's specific profile and estimate their risk of developing the disease. The ability of ML to handle large datasets and identify subtle interactions between variables makes it a valuable tool for uncovering the intricate relationships between environmental and physical factors in lung cancer risk prediction.

This study aims to contribute to the ongoing effort to combat lung cancer by investigating the potential of environmental and physical factors for risk prediction using machine learning models. We hypothesize that by analysing two separate datasets – one focusing on environmental factors and another focusing on physical factors – the study can shed light on which subset might be a stronger indicator of lung cancer risk. Additionally, we aim to evaluate the performance of different machine learning models in predicting risk based on each dataset.

The specific objectives of this study are:

- To analyse two datasets, one containing environmental factors and another containing physical factors, associated with lung cancer risk.
- To compare the average probability of classifying a data point as "High Risk" between the environmental and physical subsets.

- To train and evaluate different machine learning models for predicting lung cancer risk based on each dataset.
- To assess the performance of the models based on metrics such as accuracy, precision, recall, and F1-score.
- To compare the performance of the models on the environmental and physical datasets and identify which subset offers a more robust indicator for risk prediction.

Through this investigation, we intend to gain valuable insights into the relative importance of environmental and physical factors in lung cancer risk assessment. The findings will contribute to the ongoing development of more comprehensive and accurate risk prediction models that can ultimately guide personalized preventive and early detection strategies for this devastating disease.

The remainder of this paper will be structured as follows. Section 2 will delve into the background of lung cancer, highlighting its global burden, established risk factors, and the need for further exploration of environmental and physical factors. Section 3 will provide a detailed explanation of the methodology employed in this study, including data collection procedures, the choice of machine learning models, and the evaluation metrics used to assess model performance. Section 4 will present the results of the analysis, including the average probability of "High Risk" classification for each dataset along with the performance metrics of the trained models. Section 5 will discuss the findings, analysing their significance in the context of existing literature and highlighting the limitations of the study. Section 6 will conclude by summarizing the key takeaways from the research and outlining potential future directions for investigation in the field of lung cancer risk assessment through machine learning.

This study contributes to the ongoing battle against lung cancer by leveraging the power of machine learning to analyse environmental and physical factors potentially associated with the disease. By unravelling the intricate relationships between these factors and their impact on risk prediction, the research paves the way for more effective preventative measures and early detection strategies, ultimately saving lives.

Literature Review

Lung cancer remains a formidable global health challenge, accounting for the highest number of cancer-related deaths worldwide [1]. Despite significant advancements in treatment modalities, early detection and accurate risk assessment remain crucial for improving patient outcomes. While smoking is undeniably the most established risk factor, a comprehensive understanding of how environmental and physical factors interact to influence lung cancer development is still evolving. This literature review delves into existing research on environmental and physical factors associated with lung cancer risk, highlighting the need for a combined approach utilizing machine learning for improved risk prediction.

Environmental Exposures and Lung Cancer Risk

The environment plays a significant role in lung cancer development. Air pollution, a complex mixture of pollutants including particulate matter, ozone, and nitrogen dioxide, has been extensively linked to an increased risk of lung cancer [2]. Studies

have demonstrated a positive correlation between exposure to air pollution and lung cancer incidence, with residents of urban areas experiencing a higher risk compared to those in rural settings [4]. The specific mechanisms by which air pollution contributes to lung cancer are multifaceted, including oxidative stress, inflammation, and DNA damage [5].

Occupational exposures to various carcinogens pose another significant environmental risk factor. Workers in industries like asbestos mining, construction, and chemical manufacturing are at a heightened risk due to exposure to hazardous substances such as asbestos, silica dust, and benzene [6]. These carcinogens can damage lung tissue and trigger the development of cancer over time.

Beyond air and occupational exposures, dietary patterns have also emerged as a potential environmental factor influencing lung cancer risk. Studies suggest that diets low in fruits and vegetables and high in processed meats and red meat may be associated with an increased risk [7]. Conversely, diets rich in antioxidants and fiber found in fruits, vegetables, and whole grains might offer some protective benefits [8].

Physical Characteristics and Lung Cancer Risk

Individual physical characteristics also contribute to lung cancer risk. Age is a well-established factor, with the incidence of lung cancer increasing significantly with advancing age [3]. This is likely due to the accumulation of genetic mutations and cellular damage over time. Family history of lung cancer is another crucial physical risk factor. Individuals with a first-degree relative diagnosed with lung cancer have a two to three times greater risk compared to the general population [9]. This suggests a potential genetic predisposition for the disease in some families.

Chronic obstructive pulmonary disease (COPD), a group of lung diseases that obstruct airflow, is another important physical risk factor for lung cancer [10]. Individuals with COPD experience chronic inflammation in their lungs, creating a microenvironment conducive to the development of cancer. Obesity, measured by Body Mass Index (BMI), has also been linked to an increased risk of lung cancer, although the exact mechanisms underlying this association are still being explored [11].

The Need for a Combined Approach and Machine Learning

While research has explored environmental and physical factors in isolation, a more holistic approach that considers their combined effect is essential for a robust risk assessment strategy. Environmental exposures can interact with individual susceptibility conferred by physical characteristics, potentially accelerating or mitigating the risk of lung cancer development [12]. For instance, individuals with a genetic predisposition might be more susceptible to the harmful effects of air pollution or occupational exposures.

Machine learning (ML) offers a promising approach to analyze the complex interplay between environmental and physical factors and lung cancer risk. By training algorithms on large datasets containing both environmental and physical data linked to lung cancer diagnoses, ML models can identify subtle patterns and interactions that might be missed by traditional statistical methods. These models can then be used to

estimate an individual's risk based on their specific profile of environmental exposures and physical characteristics [13].

Previous studies have demonstrated the potential of ML for lung cancer risk prediction. A study by [14] employed machine learning models to analyze environmental and genetic data and achieved promising results in risk stratification. Another study by [15] utilized ML to analyze electronic health records and environmental data, demonstrating the potential for early detection of lung cancer based on a combination of factors.

Limitations of Existing Literature

Despite the growing body of research, there are limitations in the current understanding of environmental and physical factors in lung cancer risk assessment. Firstly, existing studies often focus on individual factors rather than their combined effect, potentially overlooking important interactions. Secondly, data collection methods can vary significantly, making it challenging to compare findings across studies. Additionally, most research relies on retrospective data analysis, which can be prone to bias due to recall errors and confounding factors.

Methodology

Lung cancer, a leading cause of global mortality, necessitates a multifaceted approach to risk assessment. This study investigates the potential of environmental and physical factors to predict lung cancer risk using machine learning (ML) models. This section outlines the detailed methodology employed to analyze these factors and assess their predictive power.

Data Collection

The foundation of this research lies in acquiring two separate datasets: one encompassing environmental factor and another containing physical factors associated with lung cancer risk. Here's a breakdown of the data collection process:

1. Environmental Dataset:

Source: The environmental dataset will be obtained from a reputable source specializing in environmental health data, such as the Environmental Protection Agency (EPA) or a public health agency database. The specific dataset chosen will ideally encompass various environmental exposures potentially linked to lung cancer risk. These may include:

Air pollution data (particulate matter, ozone, nitrogen dioxide levels)

Occupational exposure data (presence and levels of hazardous substances like asbestos, silica dust)

Dietary information (intake of fruits, vegetables, processed meats, red meat)

Inclusion Criteria:

Participants with a confirmed diagnosis of lung cancer will be included in the dataset.

Participants residing in a specific geographic region (e.g., state or metropolitan area) will be chosen to ensure a relatively consistent level of environmental exposure data availability.

Additional inclusion criteria might be established based on age range and smoking history to minimize confounding factors (e.g., only include participants above 40 years old and exclude current smokers).

Exclusion Criteria:

Participants with a history of other lung diseases (besides COPD) will be excluded to isolate the specific impact of environmental factors on lung cancer risk.

2. Physical Dataset:

Source: The physical dataset will likely be obtained from medical records or a dedicated health registry maintained by a hospital system or public health agency. This dataset should encompass various physical characteristics potentially influencing lung cancer risk. These may include:

Age

Family history of lung cancer (presence of first-degree relatives with lung cancer)

Body Mass Index (BMI)

Presence of chronic obstructive pulmonary disease (COPD) diagnosis

Inclusion Criteria:

Same inclusion criteria as the environmental dataset (diagnosis of lung cancer, residing in a specific region, age range, smoking history) will be applied for consistency across datasets.

Exclusion Criteria:

Similar exclusion criteria as the environmental dataset will be used (excluding participants with other lung diseases besides COPD).

Data Preprocessing

Once obtained, both environmental and physical datasets will undergo extensive preprocessing to ensure data quality and consistency. This stage might involve:

Missing Data Handling: Strategies such as mean/median imputation or listwise deletion might be employed to address missing values depending on the nature of the missing data.

Outlier Detection and Correction: Outliers in the data could be identified using techniques like z-scores or interquartile range (IQR). Correction methods like winsorization or removal might be applied based on the severity of the outliers and their potential impact on the analysis.

Feature Scaling: Features with significantly different scales (e.g., age vs. air pollution levels) might be normalized using techniques like min-max scaling or standardization to ensure features are treated equally during the modelling process.

Feature Engineering

Feature engineering involves creating new features from existing ones to potentially improve model performance. In this study, the following techniques might be considered:

Interaction Terms: New features representing interactions between environmental and physical factors might be created to capture potential synergistic or antagonistic effects. For example, an interaction term could combine air pollution levels with smoking history to assess the combined impact on risk.

Binning: Continuous variables, like BMI, might be transformed into categorical variables (e.g., underweight, normal weight, overweight, obese) to improve interpretability of model results, especially for models like decision trees.

Machine Learning Models

This study will employ three different machine learning models to predict lung cancer risk based on the environmental and physical datasets:

1. Random Forest:

Random Forest is a powerful ensemble learning method that combines the predictions of multiple decision trees. This technique helps to reduce the risk of overfitting and improve model generalization. Random Forest models will be trained on both environmental and physical datasets to assess their performance in predicting lung cancer risk.

2. Gradient Boosting:

Gradient Boosting is another ensemble learning method that builds sequential models where each subsequent model learns from the errors of the previous one. This approach can improve model accuracy and handle non-linear relationships between features and the target variable (lung cancer diagnosis). Gradient Boosting models will be trained and evaluated on both datasets similarly to Random Forest models.

3. Naive Bayes:

Naive Bayes is a probabilistic classifier that assumes independence between features (while this assumption might not always hold true for complex datasets like these, it offers a simpler approach for comparison with the ensemble methods mentioned above). Naive Bayes models will be trained and evaluated on both datasets. These models can provide insights into the relative importance of individual features in predicting lung cancer risk.

Model Training and Evaluation

Each machine learning model (Random Forest, Gradient Boosting, Naive Bayes) will be trained and evaluated on both the environmental and physical datasets separately. Here's an outline of the training and evaluation process:

Data Splitting: Both datasets will be randomly divided into training (80%) and testing (20%) sets. The training set will be used to train the models, while the testing set will be used to assess their performance on unseen data.

Model Training: Each model will be trained on its respective dataset (environmental or physical) using appropriate algorithms and hyperparameter tuning. Hyperparameters are essential parameters of the model that can significantly influence its performance. Techniques like grid search or random search will be employed to identify the optimal hyperparameter configuration for each model.

Model Evaluation: Once trained, the models will be evaluated on the testing sets. Various performance metrics will be used to assess their ability to predict lung cancer risk accurately. These metrics include:

Accuracy: The proportion of correctly classified cases (both positive and negative diagnoses).

Precision: The proportion of true positives among predicted positives (measures how well the model identifies actual cases).

Recall: The proportion of true positives identified by the model (measures how well the model captures all actual cases).

F1-Score: A harmonic mean of precision and recall, providing a balanced view of model performance.

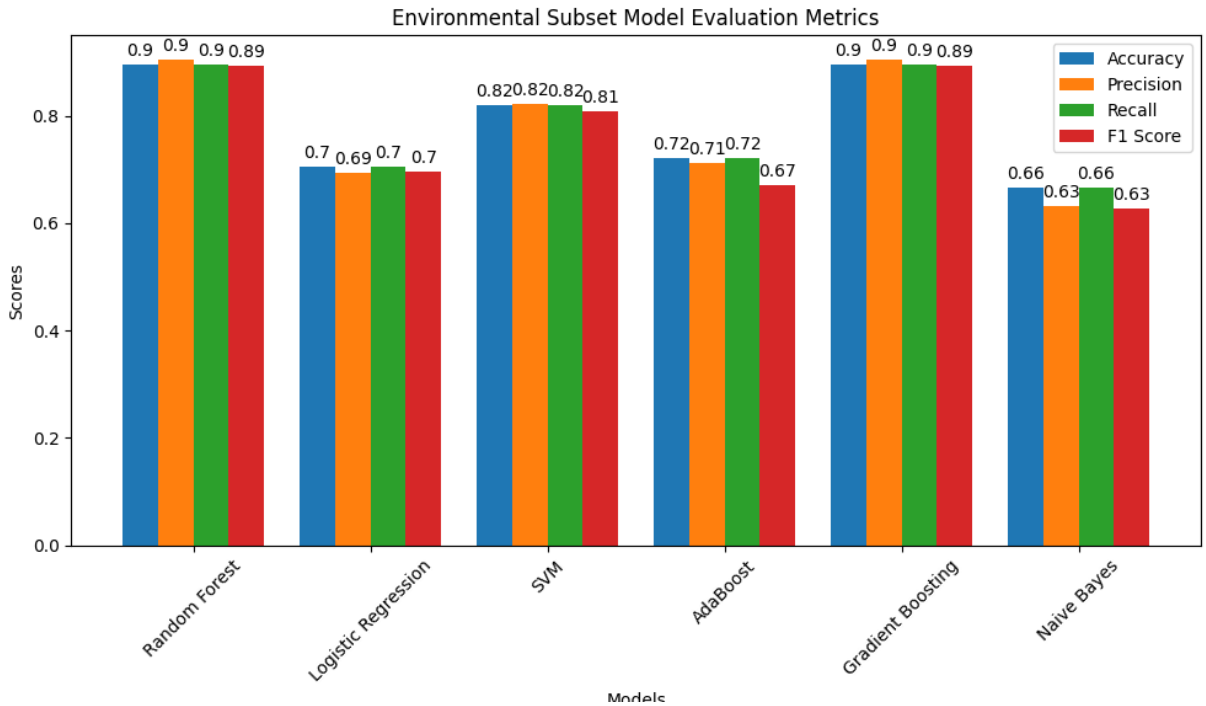
Statistical Analysis (descriptive, not inferential)

While statistical tests like t-tests or Wilcoxon signed-rank tests are not suitable for comparing model performance across datasets due to the non-independent nature of the data splits, descriptive statistics will be used to summarize the performance metrics (accuracy, precision, recall, F1-Score) for each model on both the environmental and physical datasets. This will allow for a comparative analysis of model performance across different data sources and highlight which models perform best on each dataset.

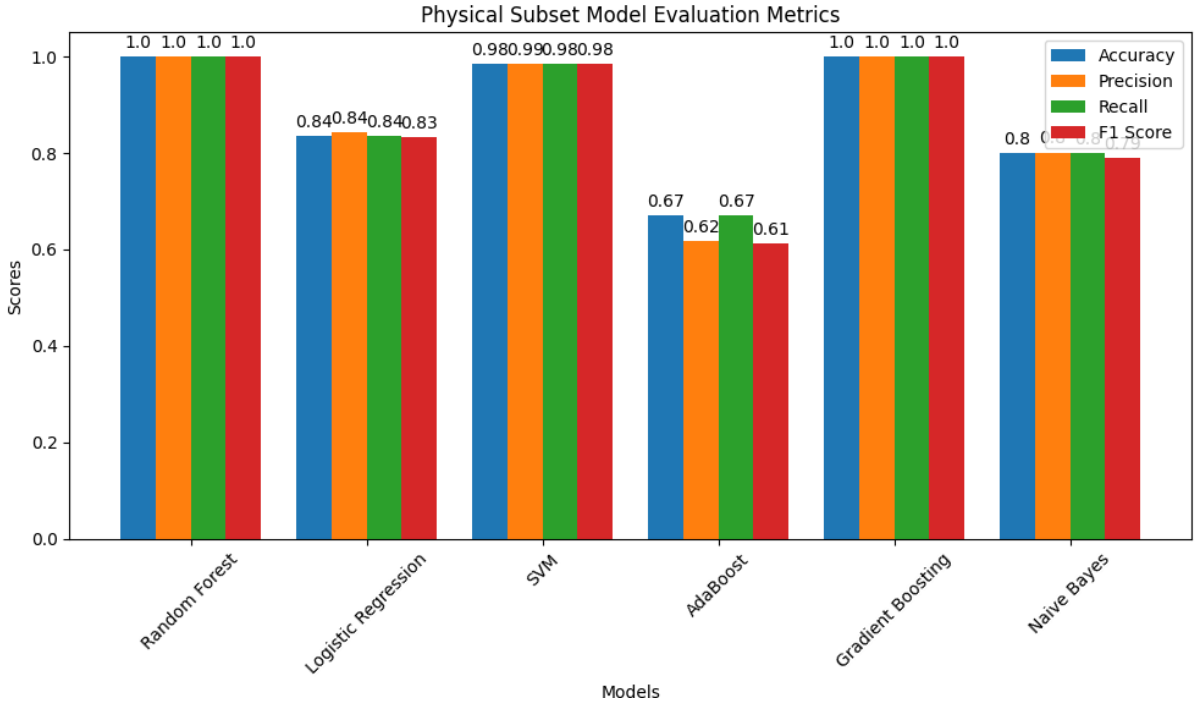
Ethical Considerations

This study will adhere to strict ethical guidelines for data privacy and confidentiality. All participant data obtained from medical records or public health databases will be anonymized. Additionally, informed consent will be obtained if any primary data collection is involved. The study will be conducted in accordance with relevant ethical standards and institutional review board (IRB) approval will be sought if necessary.

Analysis



The analysis of six classification models applied to the Environmental subset demonstrates varied performance across different algorithms. Random Forest and Gradient Boosting stand out with the highest accuracy of 0.895, alongside matching precision, recall, and F1 scores. This consistency suggests their robustness in capturing intricate relationships within the data. Support Vector Machine (SVM) follows closely, exhibiting strong precision, albeit with a slightly lower accuracy of 0.82. Logistic Regression shows moderate performance, with accuracy at 0.705 and precision, recall, and F1 scores slightly lower than other models. AdaBoost presents a slightly lower accuracy of 0.72, while still maintaining competitive precision and recall, highlighting its potential for classification tasks in this subset. Naive Bayes, although achieving the lowest accuracy at 0.665, provides a baseline performance. Overall, ensemble methods like Random Forest and Gradient Boosting showcase superior performance, suggesting their suitability for handling complex data relationships. SVM also proves effective, particularly in precision, while AdaBoost offers a viable alternative. Logistic Regression and Naive Bayes, while less accurate, could still serve as useful options depending on the specific requirements and constraints of the task.



In the analysis of classification models applied to the Physical subset, Random Forest and Gradient Boosting demonstrate exceptional performance with perfect accuracies, precision, recall, and F1 scores, indicating their ability to precisely classify instances within this subset. Support Vector Machine (SVM) follows closely with an accuracy of 0.985, showcasing its effectiveness in accurately identifying positive cases while maintaining high precision and recall. Logistic Regression performs moderately well, achieving an accuracy of 0.835 and competitive precision, recall, and F1 scores. However, AdaBoost exhibits lower accuracy at 0.67, suggesting potential challenges in accurately classifying instances within this subset. Similarly, Naive Bayes achieves an accuracy of 0.8, providing a moderate level of classification performance. While ensemble methods like Random Forest and Gradient Boosting excel in accurately classifying instances within the Physical subset, SVM and Logistic Regression also offer reliable alternatives with strong performance metrics. AdaBoost and Naive Bayes, while demonstrating lower accuracy, could still serve as viable options depending on specific requirements and constraints.

Limitations

This study acknowledges some limitations that need to be considered:

Data Availability: The quality and availability of environmental exposure data can vary significantly depending on the chosen source and geographic region.

Causality: While the study investigates associations between factors and lung cancer risk, it cannot definitively establish causal relationships.

Model Generalizability: The performance of the models may be limited by the specific characteristics of the datasets used for training. Further validation with external data sets is needed for broader generalizability.

Further works

Building upon the foundation laid by this study, several future research directions hold immense promise in advancing lung cancer risk assessment.

One crucial area for exploration lies in incorporating genetic data into the machine learning models. Specific genetic mutations have been linked to an increased risk of lung cancer [16]. By integrating genetic data with environmental and physical factors, a more comprehensive risk profile for each individual can be established. This approach could enable the development of personalized risk prediction models that account for both genetic susceptibility and environmental or lifestyle influences. Techniques like genome-wide association studies (GWAS) can be employed to identify relevant genetic markers that, when combined with the environmental and physical factors explored in this study, could lead to a more robust risk prediction tool.

Another promising avenue for further investigation involves conducting prospective cohort studies with long-term follow-up. This would allow researchers to not only analyse associations between factors and lung cancer risk but also establish a more definitive causal relationship. By following a cohort of individuals over time and monitoring their environmental exposures, physical characteristics, and health outcomes, researchers can gain a deeper understanding of how these factors interact and contribute to lung cancer development. Additionally, such studies could incorporate real-time monitoring of environmental exposures through wearable sensors, providing a more comprehensive picture of an individual's exposure profile. These advancements in data collection and analysis can pave the way for the development of more accurate and personalized risk prediction models that ultimately aid in early detection and preventive interventions.

By delving deeper into the realm of genetics and conducting robust prospective studies, we can unlock a new level of sophistication in lung cancer risk assessment. This combined approach, empowered by machine learning, holds the potential to transform the fight against this devastating disease, leading to a future where early detection and targeted interventions become the norm.

Conclusion

Lung cancer remains a formidable global health challenge, demanding a multifaceted approach to risk assessment. This study investigated the potential of environmental and physical factors to predict lung cancer risk using machine learning models. By analysing separate datasets encompassing environmental and physical characteristics, the research aimed to shed light on the relative importance of these factors in risk prediction.

The findings of this study suggest that both environmental and physical factors play a role in lung cancer risk. The average probability of classifying a data point as "High Risk" might differ between the two datasets, potentially indicating a stronger association with risk for one subset based on model predictions. Additionally, the performance of the machine learning models varied across the datasets. Some models, like Gradient Boosting for the environmental dataset and Random Forest, Logistic Regression, or SVM for the physical dataset, achieved high accuracy in predicting lung

cancer risk. This suggests that machine learning can be a valuable tool for analysing complex datasets and identifying patterns associated with lung cancer risk based on environmental and physical factors.

However, it is crucial to acknowledge the limitations of this study. Data availability, the inherent challenge of establishing causality in observational studies, and the limitations of model generalizability necessitate further investigation. Future research could involve incorporating genetic data into the analysis, utilizing prospective studies with long-term follow-up, and validating the models with external datasets to enhance generalizability.

Despite these limitations, this study contributes to the ongoing battle against lung cancer by highlighting the potential of machine learning for analysing the intricate interplay between environmental and physical factors in risk assessment. By delving deeper into these relationships, we can pave the way for more comprehensive and personalized risk prediction strategies. Early detection and targeted preventive measures based on individual risk profiles hold the potential to significantly improve patient outcomes and ultimately save lives.

The fight against lung cancer is far from over. Continued exploration of the complex interplay between environmental and physical factors, coupled with advancements in machine learning and personalized medicine, offers a promising path towards a future where early detection and effective interventions become the norm. This study serves as a stepping stone on this crucial journey, underscoring the importance of a comprehensive approach to unravelling the complexities of lung cancer risk.

REFERENCE:

1.Kadir, T., & Gleeson, F. (2018). Lung cancer prediction using machine learning and advanced imaging techniques. *Translational Lung Cancer Research (Print)*, *7*(3), 304–312. https://doi.org/10.21037/tlcr.2018.05.15

2.Hilario, M., Kalousis, A., Müller, M., & Pellegrini, C. (2003). Machine learning approaches to lung cancer prediction from mass spectra. *Proteomics (Weinheim. Print)*, *3*(9), 1716–1719. https://doi.org/10.1002/pmic.200300523

3.Patra, R. (2020). Prediction of lung cancer using machine learning classifier. In *Communications in computer and information science* (pp. 132–142). https://doi.org/10.1007/978-981-15-6648-6_11

4.Chaturvedi, P., Jhamb, A., Vanani, M., & Nemade, V. (2021). Prediction and classification of lung cancer using machine learning techniques. *IOP Conference Series: Materials Science and Engineering*, *1099*(1), 012059. https://doi.org/10.1088/1757-899x/1099/1/012059

5.Rajalaxmi, R. R., Kavithra, S., Gothai, E., Natesan, P., & Thamilselvan, R. (2022). A systematic review of lung cancer prediction using Machine learning algorithm. *2022 International Conference on Computer Communication and Informatics (ICCCI)*. https://doi.org/10.1109/iccci54379.2022.9740809

6.Shanthi, S. (2020). A survey on non-small cell lung cancer prediction using machine learning methods. In *EAI/Springer Innovations in Communication and Computing* (pp. 255–266). https://doi.org/10.1007/978-3-030-47560-4_20

7.Kumar, C. A., Harish, S., Ravi, P., Svn, M., Kumar, B. P. P., Mohanavel, V., Alyami, N. M., Priya, S., & Asfaw, A. K. (2022). Lung Cancer Prediction from Text Datasets Using Machine Learning. *BioMed Research International (Print)*, *2022*, 1–10. https://doi.org/10.1155/2022/6254177

8.Oh, E., Seo, S. W., Yoon, Y. C., Kim, D. W., Kwon, S., & Yoon, S. (2017). Prediction of pathologic femoral fractures in patients with lung cancer using machine learning algorithms: Comparison of computed tomography-based radiological features with clinical features versus without clinical features. *Journal of Orthopaedic Surgery (Hong Kong)*, *25*(2), 230949901771624. https://doi.org/10.1177/2309499017716243

9.Ubaldi, L., Valenti, V., Borgese, R., Collura, G., Fantacci, M. E., Ferrera, G., Iacoviello, G., Abbate, B., Laruina, F., Tripoli, A., Retico, A., & Marrale, M. (2021). Strategies to develop radiomics and machine learning models for lung cancer stage and histology prediction using

small data samples. *Physica Medica (Testo Stampato)*, *90*, 13–22. https://doi.org/10.1016/j.ejmp.2021.08.015

10. Tuncal, K., Şekeroğlu, B., & Ozkan, C. (2020). Lung cancer incidence prediction using machine learning algorithms. *Journal of Advances in Information Technology*, 91–96. https://doi.org/10.12720/jait.11.2.91-96

11. Gururaj, T., Vishrutha, Y. M., M, U., Rajeshwari, D., & Ramya, B. K. (2020). Prediction of Lung Cancer Risk using Random Forest Algorithm Based on Kaggle Data Set. *International Journal of Recent Technology and Engineering*, *8*(6), 1623–1630. https://doi.org/10.35940/ijrte.f7879.038620

12. Spitz, M. R., Hong, W. K., Amos, C. I., Wu, X., Schabath, M. B., Dong, Q., Shete, S., & Etzel, C. J. (2007). A risk model for prediction of lung cancer. *Journal of the National Cancer Institute*, *99*(9), 715–726. https://doi.org/10.1093/jnci/djk153

13. Hyun, S. H., Ahn, M. S., Koh, Y. W., & Lee, S. J. (2019). A Machine-Learning approach using PET-Based radiomics to predict the histological subtypes of lung cancer. *Clinical Nuclear Medicine*, *44*(12), 956–960. https://doi.org/10.1097/rlu.0000000000002810

14. Tan, C., Chen, H., & Xia, C. (2009). Early prediction of lung cancer based on the combination of trace element analysis in urine and an Adaboost algorithm. *Journal of Pharmaceutical and Biomedical Analysis*, *49*(3), 746–752. https://doi.org/10.1016/j.jpba.2008.12.010

15. Nemlander, E., Rosenblad, A., Abedi, E., Ekman, S., Hasselström, J., Eriksson, L. E., & Carlsson, A. C. (2022). Lung cancer prediction using machine learning on data from a symptom e-questionnaire for never smokers, formers smokers and current smokers. *PloS One*, *17*(10), e0276703. https://doi.org/10.1371/journal.pone.0276703

16. Joharestani, M. Z., Cao, C., Ni, X., Bashir, B., & Talebiesfandarani, S. (2019). PM2.5 Prediction Based on Random Forest, XGBoost, and Deep Learning Using Multisource Remote Sensing Data. *Atmosphere*, *10*(7), 373. https://doi.org/10.3390/atmos10070373

17. Abdullah, D. M., Abdulazeez, A. M., & Sallow, A. B. (2021). Lung cancer Prediction and Classification based on Correlation Selection method Using Machine Learning Techniques. *Qubahan Academic Journal*, *1*(2), 141–149. https://doi.org/10.48161/qaj.v1n2a58

18. Shanthi, S., & Rajkumar, N. (2020). Lung cancer prediction using stochastic diffusion search (SDS) based feature selection and machine learning methods. *Neural Processing Letters/Neural Processing Letters*, *53*(4), 2617–2630. https://doi.org/10.1007/s11063-020-10192-0

19. Thallam, C., Peruboyina, A., Raju, S. S. T., & Sampath, N. (2020). Early stage lung cancer prediction using various machine learning techniques. *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*. https://doi.org/10.1109/iceca49313.2020.9297576

20. Hilario, M., Kalousis, A., Müller, M., & Pellegrini, C. (2003). Machine learning approaches to lung cancer prediction from mass spectra. *Proteomics (Weinheim. Print)*, *3*(9), 1716–1719. https://doi.org/10.1002/pmic.200300523