A Machine Learning Approach Toward the Detection of Lung Cancer

Luke Profio

The University of Texas at Austin, Department of Computer Science
profio@utexas.edu

Lung cancer remains the leading cause of cancer-related mortality worldwide, yet early detection significantly improves patient survival rates. This project investigates whether a machine learning model can estimate lung cancer risk based on a brief survey capturing key demographic, lifestyle, and symptom-related factors. Using a public synthetic dataset of approximately 300 individuals with 15 features (including age, gender, smoking history, and binary indicators for various symptoms and habits), we developed and evaluated two classifiers: logistic regression and random forest. Model development involved data preprocessing, feature engineering, and hyperparameter tuning via cross-validation. On a held-out test set, both models achieved high performance, with accuracy ranging from 82% to 89% and ROC AUC scores near 0.95. Analysis revealed that respiratory symptoms and alcohol consumption were among the strongest predictors of lung cancer risk, consistent with epidemiological evidence. To enhance interpretability, we applied SHAP (SHapley Additive Explanations) to assess feature contributions for individual predictions, confirming that the models' decision patterns aligned with clinical capabilities. While the dataset and scope limit direct clinical application, the findings highlight the potential of lightweight, survey-based predictive tools for early lung cancer risk screening.

1 Introduction

Lung cancer is the leading cause of cancer mortality in both men and women, accounting for approximately 1.8 million deaths globally in 2020 [1]. Early detection of it can significantly improve patient outcomes: five-year survival rates can exceed 70% for stage I lung cancers, while advanced stage (stage IV) has under 10% five-year survival [2]. However, early-stage lung cancer is difficult to diagnose because initial symptoms (e.g. fatigue or chronic cough) are often non-specific and common to benign conditions [1]. As a result, diagnosis is frequently delayed until the cancer has progressed to advanced, less treatable stages. Tobacco smoking has shown to be most important risk factor for lung cancer – studies estimate 80–90% of lung cancer deaths are attributable to smoking [3, 4] – yet lung malignancies also occur among non-smokers, especially with exposure to secondhand smoke, radon gas, or other environmental and genetic factors. Because of the challenges in early recognition, there is a demand for decision support tools that can assess an individual's lung cancer risk using easily obtained information (such as demographics, smoking history, and simple symptom questions). In this project, we investigate whether machine learning (ML) models can predict the likelihood of lung cancer using a survey-based dataset of risk factors and symptoms. In developing a predictive model and interpreting its key features, we explore a potentially low-cost approach for earlier detection of lung cancer. As a high-risk exploratory study, regardless of the outcome, we provide insights into the feasibility of a survey-based approach to lung cancer risk prediction.

2 Related Work

Prior research has explored various AI methods for lung cancer risk prediction using patient data. Nemlander *et al.* (2022) developed an adaptive symptom e-questionnaire combined with machine learning to predict lung cancer among patients referred to specialists for suspected tumors [1]. They trained gradient-boosted tree models stratified by smoking status, achieving up to 82% classification accuracy for non-smokers (using fewer predictors in that subgroup) and around 77% for smokers [1]. Notably, age, sex, and education level emerged as the most important predictors in all groups [1], highlighting the relevance of basic demographics alongside symptom profiles. Other studies have leveraged public datasets of lung cancer risk factors (e.g. surveys of lifestyle habits and symptoms) to benchmark different algorithms. Dritsas and Trigka (2022) evaluated a wide range of classifiers on a Kaggle sur-

vey dataset with 14 features (smoking status, alcohol use, various respiratory symptoms, etc.), using data augmentation and extensive cross-validation [3]. Their best model, an ensemble Rotation Forest, attained an accuracy of about 97% and an AUC of 0.993 after accounting for class imbalance with SMOTE oversampling [3]. Similarly, B. Dutta (2025) performed a comparison of machine learning versus deep learning models on a symptom-and-lifestyle lung cancer dataset [4]. Using more involved preprocessing (e.g. feature selection, outlier removal, normalization) and hyperparameter tuning, Dutta reported that a simple feed-forward neural network (with one hidden layer) achieved approximately 92.9% accuracy, outperforming classic methods like logistic regression or SVM in their experiments [4]. These studies show the promise of ML-augmented lung cancer risk prediction, but also highlight that careful design (for example, handling class imbalance and selecting informative features) is critical for success. In addition, researchers are increasingly incorporating explainable AI techniques (such as feature importance analysis or SHAP values) to interpret models—with the goal of identifying which risk factors most strongly influence predictions and making the predictions more transparent to clinicians. This work builds on this growing body of literature by applying a mix of interpretable models and model-agnostic explainability methods to a lung cancer survey dataset, taking into account their limitations.

3 Methodology

3.1 Data Preprocessing

We utilized a publicly available lung cancer risk dataset from Kaggle, consisting of synthetic survey responses from roughly 300 individuals [5]. The dataset contains 16 columns: one demographic factor (Age), one genetic factor (Gender), and 14 yes/no survey questions about lifestyle or symptoms. These binary risk factors include smoking history, alcohol consumption, and various health indicators such as having "yellow fingers" (a possible sign of heavy smoking), chronic cough, fatigue, allergies, wheezing, shortness of breath, swallowing difficulty, and chest pain, among others. The binary target variable `LUNG_CANCER` indicates whether the respondent has been diagnosed with lung cancer (Yes or No). In the raw data, responses were encoded inconsistently (some as strings "YES"/"NO", others as numeric 1/2). We performed data cleaning to ensure all features were numeric and uniformly coded. Specifically, we mapped the target labels "YES" $\rightarrow$ 1 (cancer present) and "NO" $\rightarrow$ 0 (no cancer). For the predictor columns, we converted all yes/no responses to binary values as well, standardizing the encoding such that 1 represents the presence of a risk factor or symptom and 0 represents its absence. For example, a response of "YES" to `SMOKING` or `COUGHING` was encoded as 1. We also encoded Gender as 1 for male (M) and 0 for female (F). No missing values were present in this dataset (probable given this is a synthetic dataset), so imputation was not needed. After preprocessing, we had a clean data matrix of 310 samples $\times$ 15 features (14 binary risk factors plus numeric age).

3.2 Analytic Workflow

We began with exploratory data analysis (EDA) to understand the dataset's characteristics. This included examining the age distribution, the balance of the target classes, and the prevalence of each risk factor among the lung cancer positive and negative groups. We observed that the dataset is skewed toward older individuals: the majority of respondents were above age 50, with a median age in the mid-60s. The lung cancer positive class was the minority: only about 1/4 to 1/3 of the samples were labeled "YES" for lung cancer, showing a substantial class imbalance (i.e., far more non-cancer respondents than cancer cases in the data). The EDA revealed some intuitive patterns; for instance, a much larger fraction of the lung cancer patients were smokers compared to the non-cancer group, and symptoms like coughing, wheezing, and shortness of breath were more frequently reported among those with cancer than those without. We also computed the correlation matrix between features to check for redundant or highly collinear variables. Some risk factors showed noteworthy correlations with each other: for example, *Anxiety* was correlated with *Yellow Fingers* (perhaps because both can be associated with heavy smoking behavior), and *Swallowing Difficulty* was moderately correlated with *Anxiety* as well. *Shortness of Breath* and *Fatigue* had a moderate correlation, possibly reflecting a common underlying health condition affecting both. These relationships suggest there may be underlying latent factors (such as general health status or smoking intensity) that influence multiple survey responses. However, multicollinearity was not extreme (most pairwise Pearson r values were below 0.6), so we decided to retain all features for modeling at this stage. Additionally, we trained a simple random forest on the data and extracted its impurity-based feature importance scores. This rough check indicated that *Age*, *Allergy*, *Alcohol Consuming*, and *Peer Pressure* might be among the top predictors in this dataset (we delve into more rigorous feature importance analysis after fitting our final models).

For model development, we trained two supervised learning models: a logistic regression classifier and a random forest ensemble classifier. Logistic regression (LR) was chosen as a simple, interpretable baseline model, while the random forest (RF) served as a more flexible non-linear model and provided a benchmark for potential accuracy gains from a more complex approach. To make the most of the small dataset, we employed cross-validation and hyperparameter tuning on the training set. We first split the data into an 80% training set and 20% test set, using stratified sampling to preserve the proportion of positive (cancer) cases in both sets. Within the training set, we performed a grid search with 5-fold stratified CV to optimize the logistic regression pipeline. The LR pipeline included standardization of features (subtracting mean and scaling to unit variance) and the creation of interaction terms: we used a polynomial feature expansion of degree 2 to allow the logistic model to capture pairwise interactions between risk factors. We experimented with different regularization penal-

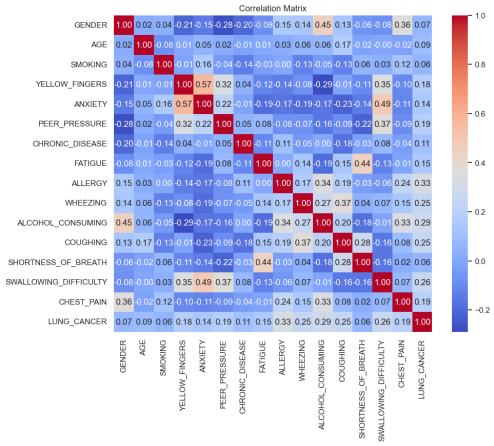

Figure 1: Correlation matrix showing pairwise Pearson correlation coefficients among survey features. Moderate correlations are observed between several symptoms and lifestyle factors, such as between *Anxiety* and *Yellow Fingers*, or between *Fatigue* and *Shortness of Breath*. Most correlations are weak, indicating low multicollinearity.

ties (L1, L2, and elastic-net combinations) and regularization strengths (the inverse regularization parameter C) for LR, employing the `saga` solver to handle L1 terms. Model selection was guided by maximizing the average area under the ROC curve (AUC) across the cross-validation folds. The best logistic model found used an elastic-net penalty (a mix of L1 and L2) with an optimal regularization strength (we found that a fairly strong overall regularization, roughly $C \approx 10^{-3}$, yielded the best validation performance). This shows that regularization was important to prevent overfitting, given the expanded feature space with interaction terms.

For the Random Forest model, we configured an ensemble of 300 trees (estimators) with bootstrap sampling. The hyperparameters for the RF model (such as maximum tree depth and minimum samples per split) were not exhaustively tuned; instead, we used the default settings. For example, trees were allowed to grow until they were pure or until leaf nodes contained at least two samples. We relied on ensemble averaging to mitigate overfitting. We set the `class_weight` parameter to `"balanced"` in both models to account for the class imbalance. This ensured that the minority class (i.e., lung cancer cases labeled as 1) received greater weight during training. This approach helps compensate for the imbalance by effectively increasing the penalty for misclassifying positive cases, which is critical in a health screening context where false negatives (i.e., missed cancer cases) are more serious than false positives.

After training, we evaluated the model performance on the held-out 20% test set. We computed key evaluation metrics including overall accuracy, precision, recall (sensitivity), F_1 score, and the ROC AUC. We also examined the confusion matrix for each model to understand the trade-offs between false positives and false negatives. In a screening application, a high recall (sensitivity) is particularly desirable, since we want to catch as many true cases as possible even if it means some false positives. Thus, we paid special attention to how many lung cancer cases were missed by each model (false negatives) and how many non-cancer individuals were incorrectly flagged (false positives). Additionally, we plotted ROC curves for both models on the test data to visualize the trade-off between sensitivity and specificity across different classification thresholds.

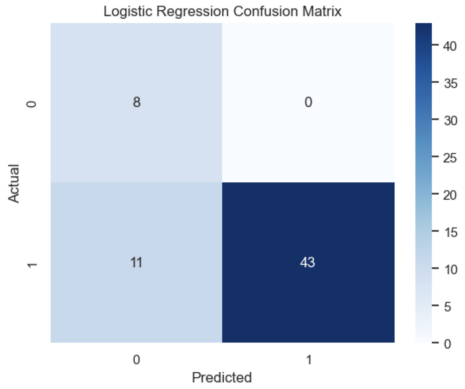

Figure 2: Confusion matrix for the logistic regression model evaluated on the test set. The model correctly identifies all negative cases (true negatives = 8), but misses 11 cancer cases (false negatives), resulting in reduced sensitivity. Despite this, it achieves perfect precision for the positive class (no false positives), highlighting a conservative prediction bias.

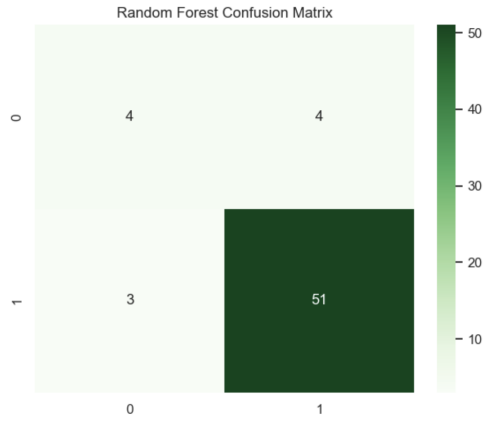

Figure 3: Confusion matrix for the random forest model evaluated on the test set. The model correctly classifies 51 out of 54 cancer cases (high recall), with only 3 false negatives and 4 false positives. This represents a more balanced trade-off between sensitivity and specificity, outperforming logistic regression.

Finally, we applied model interpretability techniques to understand how the models were making their predictions. For the logistic regression, we could directly inspect the learned weight coefficients to see which features had the largest positive or negative influence on predicting lung cancer. We also

employed SHAP (SHapley Additive Explanations) for both models (LR and RF) to quantify each feature's contribution to the prediction for each individual instance. We generated SHAP summary plots that display the global feature importance and the direction of each feature's effect on the model's output. This helped us verify whether the model's behavior aligns with domain expectations (for example, we would expect that being a smoker generally pushes the model towards a higher predicted risk of cancer). Additionally, we explored specific example explanations using SHAP force plots for individual predictions, and SHAP dependence plots to investigate potential interaction effects—such as whether the model's reliance on age was different for smokers versus non-smokers. All analysis was implemented in Python (using libraries such as pandas and scikit-learn for data handling and modeling, and Seaborn/Matplotlib for visualization).

4 Results

4.1 Model Performance

Both the logistic regression and the random forest models achieved reasonably high predictive performance despite the limited dataset size. The logistic regression model had an accuracy of approximately 82% on the test set. Its recall (sensitivity) for lung cancer cases was about 80%, meaning the model correctly identified 80% of actual cancer patients while missing about 20%. The precision for the positive class was 1.00, showing that all individuals predicted by the model as "having lung cancer" were actually true positives, with no false positives for that class. The ROC curve for the logistic regression model had an AUC of 0.95, exhibiting high performance. This optimized model slightly outperformed a baseline logistic regression (which lacked interaction features and regularization), showing that feature engineering and hyperparameter tuning contributed to overall performance.

Table 1 summarizes test-set results.

Table 1: Test set performance of logistic regression (LR) and random forest (RF).

Model	Acc.	Recall	Prec.	AUC
LR	0.82	0.80	1.00	0.95
RF	0.89	0.94	0.93	0.96

The random forest classifier performed slightly better than the logistic regression model in terms of raw accuracy, achieving approximately 89% accuracy on the test set. It also correctly identified a greater proportion of the cancer cases, with a recall of 94% for the positive class. This shows that the model missed fewer true cancer cases compared to logistic regression. The precision for the positive class was 0.93, reflecting a low false positive rate. The ROC AUC for the random forest was 0.96, slightly higher than that of the logistic regression model, exhibiting higher performance. Compared to logistic regression, the random forest had a better precision-recall trade-off at the default threshold of 0.5, likely due to its greater capacity to model complex relationships in the data. Its confusion matrix showed fewer false negatives while maintaining a reasonable number of true negatives, exhibiting high sensitivity and good specificity. As shown in Figure 2, the ROC curves of the two models overlap considerably. This shows that while the random forest leveraged its modeling flexibility effectively, the simpler logistic regression model was still able to extract most of the predictive signal from the data. But given the small sample size, these performance metrics should be interpreted with caution. The addition or removal of just a few individuals in the test set could cause the reported accuracy or AUC to vary by several percentage points. To confirm these results, external validation on an independent dataset is needed. Nonetheless, achieving nearly 89% accuracy and an AUC of 0.96 on a balanced test split is promising, and aligns well with existing literature (e.g., prior studies have reported accuracy ranges of roughly 80–95% on similar lung cancer prediction tasks) [1, 3].

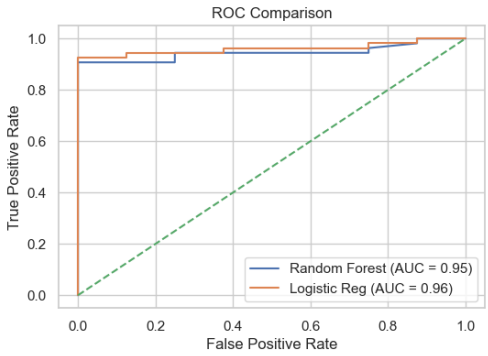

Figure 4: ROC curves on the test set for the logistic regression and random forest models. Both models achieve a high area under the curve (AUC $\approx$ 0.95), exhibiting high performance on the prediction of lung cancer.

4.2 Model Interpretability

To understand the drivers behind the model predictions, we analyzed feature attributions using SHAP (SHapley Additive exPlanations) values for both the logistic regression and random forest models. SHAP values provide a consistent framework to quantify each feature's contribution to a specific prediction, offering insights into how the models reach their conclusions and whether those decisions align with known medical knowledge. For the logistic regression model, the SHAP summary plot (Figure 3) revealed that the most influential features were `FATIGUE`, `CHRONIC_DISEASE`, `SWALLOWING_DIFFICULTY`, `ALLERGY`, and `PEER_PRESSURE`. High values of these features (shown in red) tended to push the model output toward the positive class (i.e., "has cancer"), while low values (blue) had the opposite effect. For example, individuals reporting persistent fatigue or known chronic disease had substantially higher predicted probabilities of lung cancer. Interestingly, `ALLERGY` and `PEER_PRESSURE` also appeared as impactful

features—potentially echoing correlations in the synthetic data between these attributes and cancer risk, though these would require further validation in real-world settings.

In contrast, the SHAP analysis for the random forest model (Figure 4) had a different ranking of features. The top five included `ALCOHOL_CONSUMING`, `ALLERGY`, `PEER_PRESSURE`, `YELLOW_FINGERS`, and `FATIGUE`. These features contributed in more complex and sometimes non-linear ways, with some individuals showing large positive SHAP values when these risk factors were present. For instance, the model often assigned high risk scores to participants with yellow-stained fingers (a proxy for smoking exposure), high alcohol consumption, or peer pressure—likely capturing behavioral or lifestyle patterns linked to elevated cancer risk. Having `ALLERGY` again as a high-impact feature suggests that both models, despite architectural differences, found signal (or correlations) in this variable. These differences highlight the complementary strengths of the two models. Logistic regression provides transparent, linear relationships that can be easily interpreted, while random forest captures higher-order interactions and non-linear effects. The consistency of features like `FATIGUE` and `ALLERGY` across both models shows their predictive value in this dataset, though some associations (such as the importance of `PEER_PRESSURE`) may reflect structural patterns in the synthetic survey responses rather than causal risk factors.

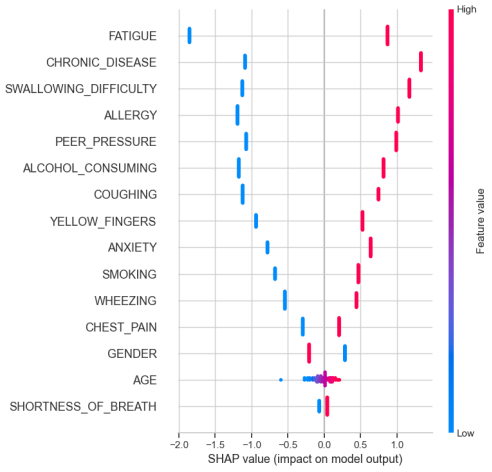

Figure 5: SHAP summary plot (dot format) showing feature contributions for the logistic regression model on the test set. Each dot represents an individual prediction, colored by the feature value (red = high, blue = low), with the x-axis indicating the feature's impact on the model output.

Overall, these SHAP-based interpretability analyses strengthen confidence that both models are utilizing patterns that are at least partially aligned with clinical reasoning, such as symptoms (e.g., fatigue), behavioral risk factors (e.g., alcohol use), and comorbidities (e.g., chronic disease). At the same time, they also underscore the need for careful validation—particularly where synthetic or proxy features may introduce artifacts. Future work with real clinical datasets would be needed to confirm whether the influential features identified here generalize beyond this cohort.

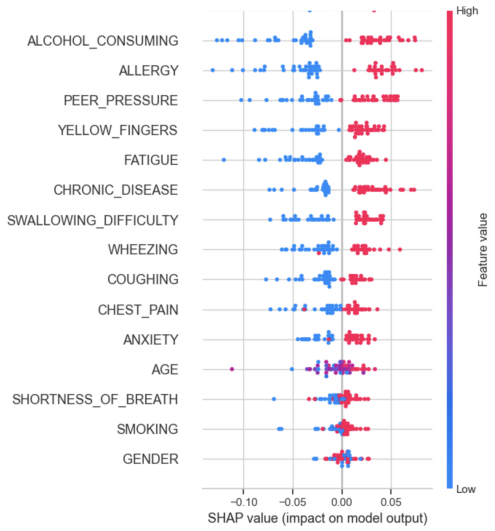

Figure 6: SHAP summary plot (dot format) for the random forest model. Compared to logistic regression, the RF shows a different ranking of dominant features, and wider variation in SHAP values across individuals.

5 Conclusion

In this project, we designed and evaluated a machine learning approach for predicting the likelihood of lung cancer from a brief questionnaire. Despite the high-risk nature of attempting a new idea with a small synthetic dataset, our results demonstrate the potential of such models: the logistic regression and random forest achieved around 82–89% accuracy in classifying survey respondents as having lung cancer or not, with ROC AUCs near 0.9. These performance levels are comparable to those reported in related research using similar data [1, 3], suggesting that even relatively simple models can extract significant signal from self-reported risk factors. The most influential features driving our predictions (smoking status, age, coughing, shortness of breath, chest pain) aligned well with established clinical understanding of lung cancer risk factors [3]. This alignment is encouraging because it means the model is likely focusing on true risk indicators rather than irrelevant patterns or noise. Early detection of lung cancer remains a challenging problem, but our exploratory study supports the notion that a risk questionnaire combined with an ML model could aid in triaging individuals for further screening. For example, in a primary care setting, a survey-based model could identify high-risk patients (older long-term smokers with certain symptoms) who might benefit from definitive screening tests like low-dose CT scans—an intervention which has been shown to reduce lung cancer mortality in high-risk groups [2].

However, there are several limitations to this study. First, the dataset used was both small ($n \approx 300$) and synthetic – it does not capture the full variability of real patient populations.

The model may therefore be overfitting nuances of this artificial dataset and might not generalize to actual clinical data. Prior studies have noted that models trained on such public synthetic data can perform optimistically and lack the rigor of real-world clinical features [3]. Second, the survey features themselves are limited; important predictors like family history of cancer, detailed smoking intensity (e.g. pack-years), or occupational exposures were not included. In a real setting, incorporating these additional factors would likely improve risk stratification. Third, our current models treat the problem as a static classification based on one-time inputs. In practice, a patient's risk evolves over time, and a longitudinal approach (monitoring how a person's symptoms or exposures change) could be more powerful. Additionally, while our model can highlight individuals at high risk, it cannot by itself provide a definitive diagnosis – any high-risk predictions would need to be followed up with medical imaging (e.g. an X-ray or CT scan) and clinical evaluation.

In future work, a number of avenues could be pursued to build on this project. Gathering a larger and more representative dataset (for example, merging multiple survey data sources or using actual patient cohorts with outcomes) is a top priority to improve the model's robustness. With more data, we could also explore more advanced algorithms such as gradient boosting machines or deep neural networks, which have shown superior accuracy in some studies [3, 4]. Another promising direction is to integrate this questionnaire-based model with clinical data – for instance, combining the risk score from survey responses with results from screening tests or imaging. Dritsas *et al.* suggest that incorporating information from lung CT scans alongside survey data could improve early detection performance [3]. From an AI explainability perspective, we could implement more sophisticated interpretability techniques or even causal inference methods to ensure the model's predictions are transparent and trustworthy for clinicians. Lastly, deploying the model in a real-world tool would be an exciting step: one can imagine a decision support system where a clinician (or even a patient via a web application) inputs the answers to a risk questionnaire, and the system outputs a personalized lung cancer risk estimate along with an explanation highlighting the contributing factors. Such a tool, used appropriately, might help prioritize high-risk individuals for proper follow-up (for example, recommending those with high predicted risk to undergo diagnostic imaging sooner).

References

[1] 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, former smokers and current smokers*. *PLOS ONE*, 17(10): e0276703. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0276703

[2] Nam, B.; Hamm, D.; Katurakes, N.; Mulligan, C. (2024). *Lung Cancer Screening: Early Detection Decreases Mortality*. *Delaware Journal of Public Health*, 10(3): 22–24. https://pubmed.ncbi.nlm.nih.gov/</wbr>36999178

[3] Dritsas, E.; Trigka, M. (2022). *Lung Cancer Risk Prediction with Machine Learning Models*. *Big Data and Cognitive Computing*, 6(4): 139. https://www.mdpi.com/2504-2289/6/4/139

[4] Dutta, B. (2025). *Comparative Analysis of ML and DL Models for Lung Cancer Prediction Based on Symptomatic and Lifestyle Features*. *Applied Sciences*, 15(8): 4507. https://www.mdpi.com/2076-3417/15/8/4507

[5] Nelson, S.G. (2023). *Lung Cancer Prediction* (Kaggle Notebook). Kaggle. Retrieved July 15, 2025 from https://www.kaggle.com/code/sandragracenelson/lung-cancer-prediction.