Using Machine Learning Methods to Predict Lung Cancer Risk

Brandon Wood, Isaac Mower, Blake Simpson

Introduction

Pulmonary disease, including lung cancer, remains one of the leading causes of morbidity and mortality worldwide. According to the U.S. Centers for Disease Control and Prevention (CDC), lung cancer is the third most common cancer in the United States and the leading cause of cancer death among both men and women. Early identification of individuals at high risk is essential for timely intervention, improved health outcomes, and reduced mortality. However, diagnosing pulmonary disease presents numerous challenges due to its complex etiology, which includes a wide range of risk factors such as tobacco use, genetic predisposition, environmental exposure (e.g., air pollution, radon), and variability in symptom presentation (CDC, 2025; Corewell Health; NHS, 2022).

In this project, we frame the prediction of lung cancer as a supervised classification task, using labeled data to train models that can distinguish between individuals with and without the disease. Lung cancer diagnosis is a supervised learning problem because our study uses a labeled dataset of participants with features such as age, smoking habits, and symptoms with already known diagnoses from medical experts. This is similar to studies done by Elais Dristas and Radhanath Patra. Both Elias Dristas and Radhanath Patra used these machine learning methods to help in early diagnosis of lung cancer, which is crucial in successfully treating lung cancer (Dristas 2025 and Radhanath 2020). The broader goal is to develop a reliable and interpretable model that not only classifies risk accurately but also provides insight into which risk factors are most predictive of pulmonary disease.

The motivation behind this project is twofold: (1) to create an effective predictive model for identifying individuals at risk of pulmonary disease, and (2) to gain a better understanding of the underlying risk factors that contribute most significantly to disease onset. Machine learning is ideal to address this problem as it can help us understand the data and give us models to use in future predictions. To address these goals, we implemented and compared three supervised learning models: Linear Discriminant Analysis (LDA), Random Forests, and LASSO. LDA is used as a simplistic baseline model to be compared against; Random Forests offer a flexible, non-parametric approach capable of modeling nonlinear interactions and variable importance; and LASSO introduces regularization to help with feature selection in high-dimensional settings by shrinking less relevant coefficients to zero.

By comparing these models, we aim to strike a balance between predictive accuracy, interpretability, and the ability to identify influential variables. This approach not only allows for effective disease classification but also enhances our understanding of which features—such as smoking history, exposure to environmental pollutants, or symptoms like chronic cough and chest discomfort—play a pivotal role in predicting pulmonary disease.

Data

This is a supervised learning problem. The data we are using can be found at Lung Cancer Prediction Dataset | Kaggle. It was posted on Kaggle by Shantanu Garg, the csv has 5,000 observations and 18 factors related to lung cancer risk factors and prediction, such as lifestyle habits, medical history, and symptoms associated with pulmonary disease. We performed a minor preprocessing step by recoding the PULMONARY_DISEASE variable from categorical values ("YES"/"NO") to a binary numeric format, where 1 indicates the presence of pulmonary disease and 0 indicates its absence. To evaluate model performance and

generalizability, we began by randomly splitting the dataset into three subsets: training (60%), validation (20%), and testing (20%). A fixed seed (set.seed(123)) was used to ensure reproducibility of results. This approach allows us to train the model on a representative sample while preserving a held-out test set to assess out-of-sample performance. The training set was used for both model training, the validation set was used for tuning, while the test set was strictly reserved for final evaluation metrics such as accuracy, precision, recall, specificity, MSE, F1 score, and AUC. Because the cost of a false negative is particularly high in medical diagnosis, we place a strong emphasis on recall as a performance metric. In this context, failing to identify someone who truly has pulmonary disease could delay critical intervention, so we prefer models that prioritize sensitivity even at the expense of a small drop in overall accuracy.

Linear Discriminant Analysis (LDA)

For this analysis, we selected Linear Discriminant Analysis (LDA) as one of the classification methods due to its simplicity, efficiency, and interpretability. LDA is a linear model that performs well when the predictors are approximately normally distributed and when class separation is linear. Given that our dataset includes binary predictors and has a sufficiently large sample size, LDA should still perform effectively despite the binary nature of some predictors. In contrast to more complex models like Random Forest, LDA is less computationally intensive and provides clear insights into how predictors contribute to class separation. We also explored LASSO logistic regression, which is useful in high-dimensional settings, but can struggle when predictors are highly correlated—an area where LDA can sometimes perform better by considering the covariance structure. We implemented LDA using the caret package in R, with 10-fold cross-validation for model evaluation. LDA does not involve many hyperparameters, so no extensive tuning was needed, and the focus was on ensuring stable performance through

validation. Overall, LDA offered a good balance of performance and interpretability compared to the other models tested.

LASSO Logistic Regression

We chose LASSO logistic regression for our data set because the response variable is binary, and we wanted to know which predictor variables were important in diagnosing lung cancer. To create our LASSO logistic regression, we used the R software and used the following packages: glmnet (used for creating and tuning the model), verification (used for calculation of metrics), and pROC (used for building and finding ROC/AUC). LASSO introduces a penalty term that shrinks less important feature coefficients toward zero, effectively performing both variable selection and regularization simultaneously. After splitting the available data into training, validation, and testing subsets, a logistic regression model with L1 regularization was fitted to the training data. The strength of the regularization was controlled through the lambda parameter, which influences how many variables are retained in the final model. A range of lambda values was evaluated, and their corresponding prediction errors were measured on a validation set to identify the best balance between model complexity and predictive performance. LASSO logistic regression was selected over traditional logistic regression because it automatically performs variable selection, reducing model complexity and improving generalization to unseen data.

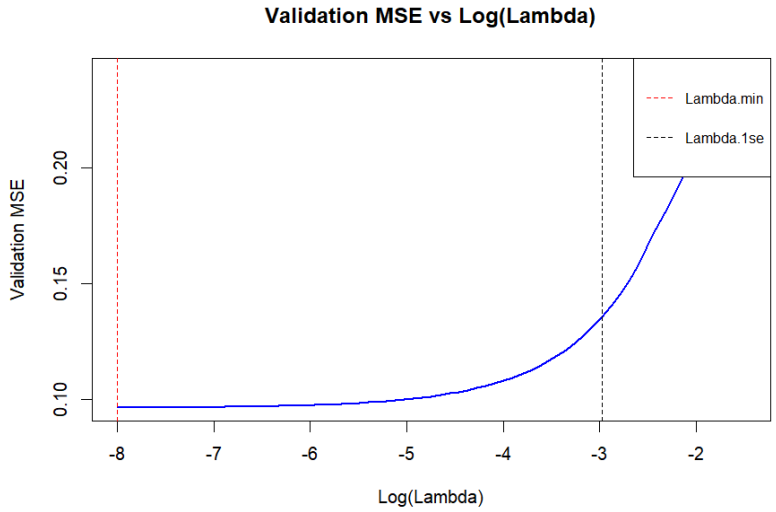

Figure 1: Binomial Deviance vs the natural log of the regularization parameter.

In *Figure 1,* two vertical dashed lines highlight important model selection points: the red line identifies the minimum lambda, the lambda value that achieves the minimum deviance (optimal predictive performance on the validation data), and the black line marks lambda 1 s.e., the largest lambda within one standard error of the minimum lambda value. Selecting the minimum lambda would yield the most accurate model, while choosing lambda 1 s.e. would result in a simpler model that maintains strong predictive performance with less chance for overfitting and more interpretability.

Once the optimal lambda values were selected based on validation performance, the final model was evaluated on a separate testing set to assess its predictive ability on new data. The model produced probability estimates for lung cancer diagnosis, which were converted into binary classifications using both the normal probability threshold equal to 0.5 and a lowered probability threshold equal to 0.4 in order to prioritize recall. This adjustment was made to ensure that the model is more sensitive to identifying individuals at risk for lung cancer, even if it results in a higher rate of false positives.

Random Forests

We included Random Forests due to their strong performance on classification tasks and their ability to quantify variable importance to be compared against the results from LDA and LASSO. We used the randomForest, verification, caret, and pROC libraries in R for this analysis. Given the strong baseline performance of Random Forests and the scope of our project, we chose to use default hyperparameters without additional tuning. As such we trained the Random Forest on the entire training dataset, then used 10-fold cross-validation to evaluate the initial model. We then fit that model to the test data to obtain our predictions.

Results

We begin reporting our results by presenting the results of the Linear Discriminant Analysis (LDA) model, which serves as a baseline for evaluating the performance of other classification methods in this study. As LDA involves minimal hyperparameter tuning, the model was applied directly to the training data without additional optimization.

	Reference	
Prediction	0	1
0	532	41
1	59	368

Table 1: Confusion Matrix for LDA

Table 1 presents the confusion matrix for the LDA model, which achieved an overall accuracy of 90%, with both recall and specificity also at 90%. These results are favorable for our application, as a high recall is particularly important; failing to identify individuals with cancer could lead to serious and potentially life-threatening consequences.

	LD1
AGE	-0.0007405599
GENDER	0.0070203331
SMOKING	1.6372996787
FINGER_DISCOLORATION	0.0529135384
MENTAL_STRESS	-0.0574776140
EXPOSURE_TO_POLLUTION	0.3732909700
LONG_TERM_ILLNESS	-0.0520880783
ENERGY_LEVEL	0.0453753536
IMMUNE_WEAKNESS	-0.0170860284
BREATHING_ISSUE	1.3842174781
ALCOHOL_CONSUMPTION	-0.0018881048
THROAT_DISCOMFORT	1.2584765497
OXYGEN_SATURATION	0.0083083073
CHEST_TIGHTNESS	0.0634505575
FAMILY_HISTORY	-0.1674161033
SMOKING_FAMILY_HISTORY	0.9621970364
STRESS_IMMUNE	0.9205543882

Figure 2: Coefficients of the Linear Discriminant Function

Figure 2 displays the scaled coefficients from the LDA model, reflecting the relative weight assigned to each predictor variable. Variables with higher absolute coefficient values are considered more influential in class separation. Notably, Smoking, Throat_Discomfort, and Breathing_Issue exhibit the largest coefficients, indicating their strong predictive importance in identifying cases of lung cancer.

Model	Accuracy	Recall	Specificity	Precision	F1 Score	MSE	AUC
Lambda_min (Thresh 0.5)	0.909	0.890	0.922	0.888	0.889	0.091	0.930
Lambda_1se (Thresh 0.5)	0.875	0.831	0.905	0.859	0.845	0.133	0.916
Lambda_min (Thresh 0.4)	0.887	0.914	0.868	0.827	0.869	0.091	0.930
Lambda_1se (Thresh 0.4)	0.843	0.895	0.807	0.763	0.823	0.133	0.916

Table 2: Accuracy metrics for different lambda values and probability thresholds.

To evaluate model performance for LASSO, we compared key classification metrics for the models selected by minimum lambda and lambda 1 s.e., these metrics are shown in *Table 2*. *Table 2* shows that decreasing our probabilistic threshold from 0.5 to 0.4 increases our recall without tanking our accuracy, specificity, or precision. After looking at the metrics, the optimal model is the model with a minimized regularization parameter and a probabilistic threshold of

0.4. This model has the second-highest total accuracy and the highest recall. The 2% decrease in total accuracy from the most accurate model for a 2.5% increase in recall is an acceptable tradeoff for catching more people who are at risk for lung cancer, as a false positive is better than a false negative in this case.

The final model did not reduce any variable coefficients to zero, but did reduce some variable coefficients to near zero. The least significant variables were age (coefficient = -0.000336), alcohol consumption (0.00919), and gender (-0.0124). The most significant variables were smoking (coefficient = 3.302), presence of breathing issues (3.031), and throat discomfort (2.592). LASSO showing that smoking is the best predictor is good because it aligns with conventional medical knowledge. Our other two most significant predictors are common symptoms of lung cancer. This shows that our model catches the same details that doctors do. Our insignificant variables confirm the same thing, except for age. Age is a common predictor of lung cancer, as people over the age of 65 are at a higher risk of lung cancer than those who are younger. The age variable in our data set has a mean value of about 57.2 years old. The high average age in our dataset is likely the reason that age was an insignificant predictor in our model.

The Random Forest was able to obtain the highest accuracy of the models, reaching 91.9% accuracy. It also performed well in the other metrics, specifically, we got a specificity of 89.16% and a precision of 93.85%.

	Reference	
Prediction	0	1
0	549	45
1	36	370

Table 3: Random Forest Confusion Matrix

Looking at the confusion matrix, the Random Forest model achieved strong performance, correctly identifying 370 true positives and 549 true negatives. While it maintained high precision and specificity, its 45 false negatives highlight a modest risk of missing true cases of pulmonary disease, which is critical in a medical context.

We also used Random Forests built-in importance metrics to help us determine pulmonary disease risk factors.

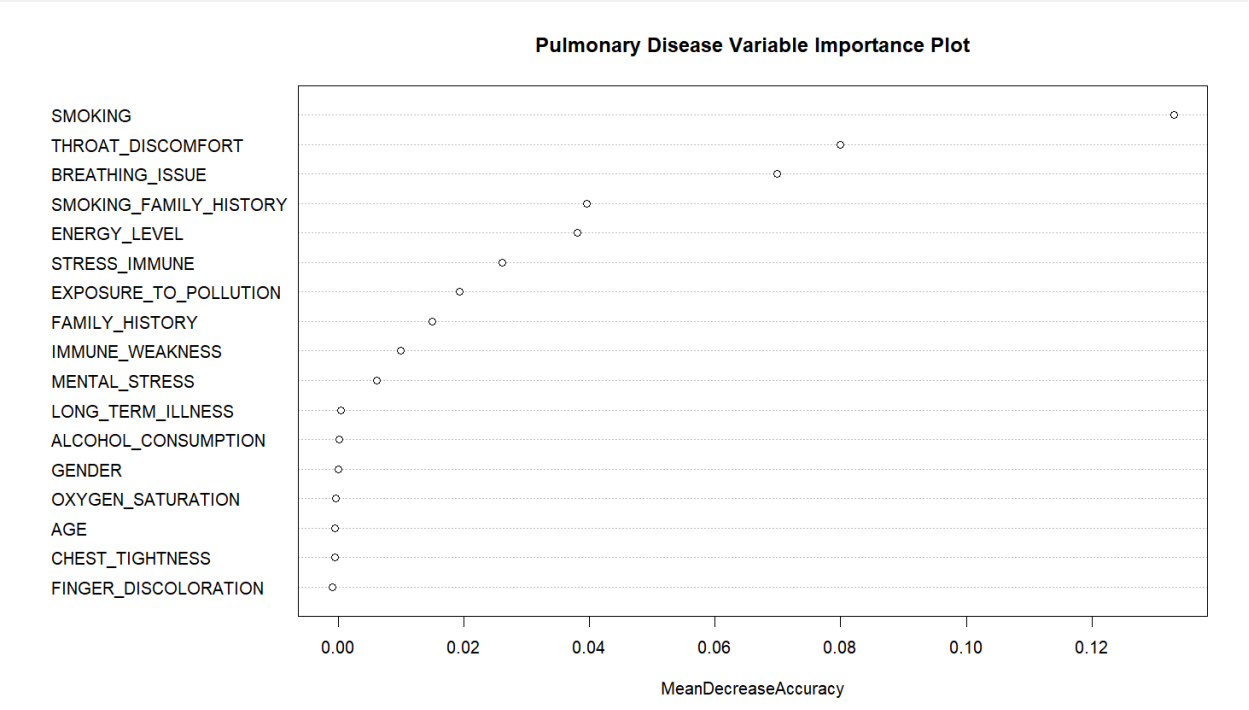

Figure 3: Random Forest Importance Plot

As we can see in *Figure 3,* the most important variables for prediction are the same ones we found using LDA and LASSO: Smoking, Throat_Discomfort, and Breathing_Issue. There is then a pretty big drop off after these variables, with the next important ones going to Smoking_Family_History, and Energy_Level as other helpful variables. Interestingly, the variables with approximately 0 mean decrease accuracy are Age, Chest_Tightness, and Finger_Discoloration.

Looking at Kaggle, there are quite a few other people who have been exploring this dataset with similar goals to ours. Specifically, the most impressive performance we found is from Lee Seungmin and can be found here. In their analysis, they used Xgboost with a grid search for parameter tuning. They were able to obtain an accuracy score of 92%, with a precision of 93% and a recall of 92%.

Overall, our results demonstrate consistent and robust model performance, with Smoking, Throat_Discomfort, and Breathing_Issue consistently being the most influential predictors of lung cancer risk. The LDA model provided a reliable baseline, while the LASSO regression model allowed for tuning that prioritized recall with minimal loss in overall accuracy, an important consideration given the high cost of false negatives in the diagnosis of lung cancer. The Random Forest model achieved the highest overall accuracy, specificity, precision, F-1 score, and AUC. Finally, a comparison with external studies, such as the XGBoost model developed by Seungmin Lee, suggests that our modeling approach is both competitive and consistent with broader findings on Kaggle.

Conclusion

Model	Accuracy	Recall	Specificity	Precision	F1 Score	MSE	AUC
LDA	0.900	0.900	0.900	0.862	0.88	0.100	0.93
LASSO	0.887	0.914	0.868	0.827	0.869	0.091	0.930
Random Forest	0.919	0.8916	0.9385	0.9113	0.9013	0.081	0.933

Table 4: Metrics of our models.

Table 4 shows that Random Forests outperformed both LDA and LASSO in every metric but recall. If the goal of the implementer of our models was just a high accuracy test, random forests would be the best option. However, diagnosing someone who has lung cancer as free from cancer is much worse than diagnosing someone free from cancer with cancer. So an implementer might want the model with the highest Recall. LDA performs at the middle or

bottom of the pack in every metric and likely would not be selected when implementing one of our models. One challenge of creating our models was maximizing recall because a false negative is very problematic when diagnosing lung cancer. We tried a variety of different probability thresholds and eventually settled on 0.4 since the recall plateaued when the threshold was lower, and the accuracy/sensitivity plummeted after that point. One surprising result we found was the insignificance of age in predicting lung cancer, which could be explained by the high average age of our sample. If we had more time, we would train and tune more time-consuming and accurate models like neural networks or support vector machines. These models typically outperform the models we used, though they are more computationally expensive, and we have no reason to believe they would not increase accuracy here. One major limitation of our analysis is that the model trainers are not experts in lung cancer and did not have access to experts in lung cancer. This access would have allowed us to make more informed decisions throughout the model training process, as well as have a better evaluation of the models' results. Despite these limitations, our results demonstrate the value of even relatively simple models in addressing complex medical challenges. Ultimately, the integration of machine learning techniques into the diagnostic process holds significant promise for enhancing early lung cancer detection, potentially saving countless lives through timely and data-driven diagnosis.

References

Corewell Health. *Lung Cancer Risk Factors.* Corewell Health. https://www.beaumont.org/conditions/lung-cancer-risk-factors

Dritsas, Elias, and Maria Trigka. "Lung cancer risk prediction with machine learning models." Big Data and Cognitive Computing 6.4 (2022): 139.

Lung Cancer Risk Factors. (2025, February 13). U.S. Centers for Disease Control and Prevention (CDC). https://www.cdc.gov/lung-cancer/risk-factors/index.html

Lung Cancer - Symptoms. (2022, November 4). English National Health Service. https://www.nhs.uk/conditions/lung-cancer/symptoms/

Patra, Radhanath. "Prediction of lung cancer using machine learning classifier." Computing Science, Communication and Security: First International Conference, COMS2 2020, Gujarat, India, March 26–27, 2020, Revised Selected Papers 1. Springer Singapore, 2020.