Comparison of Prediction Models for Lung Cancer Data: Accuracy Analysis of Lasso Regression, SVM and XGBoost

Di Ren *

Department of Mathematics, Mt. San Antonio College, Walnut, CA 91789, United States

* Corresponding Author Email: dren1@student.mtsac.edu

Abstract. This study aimed to evaluate the performance of three predictive models on a lung cancer prediction dataset including Lasso Regression, Support Vector Machine (SVM), and eXtreme Gradient Boosting (XGBoost). To ensure data quality, preprocessing included converting binomial variables to numeric values and adjusting model weights to improve the detection of negative cases and reduce missed diagnoses. Model training was performed using a robust 5-fold cross-validation repeated three times to ensure accuracy and generalizability. For Lasso Regression, feature selection was optimized using both Minimum Squared Error (MSE) and 1-Standard Error (1SE) criteria. The SVM model's cost parameter was fine-tuned through cross-validation. For XGBoost, key hyperparameters include the number of boosting rounds (nrounds), maximum depth (max_depth), learning rate, column sampling by tree (colsample_bytree), and subsampling rate (subsample) were selected to maximize accuracy. Model performance was assessed primarily through accuracy. Among the three, XGBoost achieved the highest accuracy, indicating its superior capability in handling binary classification tasks such as lung cancer prediction. These findings suggest that XGBoost may be particularly effective for clinical decision-support systems. Future research could expand to larger datasets or explore threshold adjustments to reduce false positives and further enhance model performance. This study provides useful guidance for selecting effective predictive models in lung cancer diagnosis.

Keywords: Lung cancer, Lasso regression, SVM, XGBoost.

1. Introduction

Lung cancer is the second most prevalent cancer and the leading cause of cancer death in the United States. In 2020, approximately 228820 individuals were diagnosed with lung cancer, with about 135720 succumbing to the disease [1]. Given the critical role of the lungs in the respiratory system and overall survival, effective early detection and treatment are paramount.

Screening for lung cancer has shown promise in improving patient outcomes. A retrospective cohort study revealed that individuals who underwent screening had significantly better overall mortality rates compared to those who did not [2]. Despite these findings, the rate of missed diagnoses remains alarmingly high under the current screening guidelines, as evidenced by a 90.8% missed diagnosis rate in a health examination population in China [3, 4]. This highlights the urgent need for refined screening guidelines tailored to specific target populations.

In response to these challenges, there has been a significant shift towards integrating machine learning models to enhance the accuracy of lung cancer predictions. In 2022, only 9.4% of Americans eligible for lung cancer screening (LCS) discussed the test with their healthcare provider, and 19% had never heard of the test [5]. These statistics underscore the necessity of incorporating all relevant data, including induced correlations from the network of trial data, to make informed comparisons between treatments [6].

Various researchers, including Kumar, Nageswaran, and Onozato, have explored different machine learning approaches to predict lung cancer. Their work involved a comparative analysis of models such as Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbors (KNN), Logistic Regression, and Neural Networks [5-7]. These studies assessed predictive accuracy using metrics like accuracy, specificity, and sensitivity. One approach to enhancing the detection process from lung cancer datasets has involved using an SVM-based machine learning model. This model is

part of a broader set of techniques including Logistic Regression, KNN classifier, SVM classifier, Random Forest classifier, and ensemble algorithms such as XGBoost and Adaboost [8].

Additionally, image processing techniques have been applied to CT scans, utilizing methods like geometric mean filtering for image enhancement and k-means for image segmentation. Machine learning classifiers, including Artificial Neural Networks (ANN), KNN, and RF, have been used to analyze these images, demonstrating their potential in achieving high area under the curve (AUC) values in predictive accuracy [6, 7].

Given the high recurrence rates of lung cancer and the importance of post-operative prognosis, non-parametric methods like the Cox Proportional Hazards Regression model have also been employed to predict recurrence and survival, showcasing the diverse array of analytical tools available [9, 10].

Despite the favorable outcomes reported with the Random Forest model, there is a noted scarcity in direct comparisons between the predictive capabilities of Logistic Regression and RF. Most scholarly attention has been directed towards SVM and XGBoost models [11]. Therefore, this paper aims to specifically contrast the predictive accuracy of three machine learning methods including Lasso Regression, SVM, and XGBoost in the context of lung cancer. This comparative study will provide valuable insights into the application and effectiveness of these methods in improving lung cancer detection and treatment strategies.

2. Methods

2.1. Data Source

The initial data utilized in this study was sourced from Kaggle posted 2 years ago, with the dataset being owned by Muzdadid. It has been reviewed a total of 41809 times. In total, there are 309 participants included in this dataset.

2.2. Variables and Data Preprocessing

This dataset focuses on lung cancer status and comprises 309 observations with 16 key variables. This set includes 15 independent variables associated with the status of lung cancer. The remaining variable serves as the dependent variable, indicating the presence or absence of lung cancer. The dataset contains no missing data, and no variables have been omitted. Detailed information about these variables will be presented below Table 1.

Table 1. Description of Variables

Variable	Datatype	Information
Age	Numeric	range from 21 - 87
Allergy	Categorical	1[No], 2[Yes]
Anxiety	Categorical	1[No], 2[Yes]
Alcohol Consuming	Categorical	1[No], 2[Yes]
Chest Pain	Categorical	1[No], 2[Yes]
Coughing	Categorical	1[No], 2[Yes]
Chronic Disease	Categorical	1[No], 2[Yes]
Fatigue	Categorical	1[No], 2[Yes]
Gender	Categorical	M [Male], F [Female]
Peer Pressure	Categorical	1[No], 2[Yes]
Swallowing Difficulty	Categorical	1[No], 2[Yes]
Shortness of Breath	Categorical	1[No], 2[Yes]
Smoking	Categorical	1[No], 2[Yes]
Wheezing	Categorical	1[No], 2[Yes]
Yellow Fingers	Categorical	1[No], 2[Yes]
Lung Cancer	Categorical	Yes, No

2.3. Statistical Analysis

2.3.1. T-test

A two-sample t-test was conducted to compare the ages of individuals with and without lung cancer. As shown in Figure 1, the mean age was approximately 61 years for both groups. The resulting p-value was 0.18. This indicates that there is no significant evidence to suggest a difference in average age based on lung cancer status, but it does not rule out age as a contributing factor.

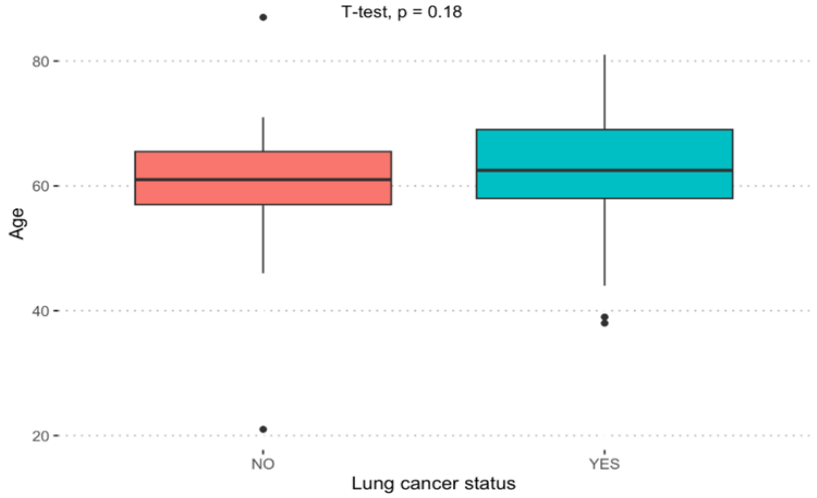

Figure 1. Comparison of Age Distribution by Lung Cancer Status

2.3.2. Chi-squared test

Figure 2 presents the results of Chi-squared tests, indicating that several symptoms and lifestyle factors differ significantly between the lung cancer and non-cancer groups. Notably, wheezing, difficulty swallowing, and coughing show particularly strong associations ($p < 0.001$). These variables could serve as strong predictors in lung cancer models. However, the lack of significance in expected variables like smoking highlights the need for further model-based multivariate analysis to better identify patterns and improve prediction accuracy for lung cancer status.

Figure 2. Proportional Comparison of Categorical Predictors by Lung Cancer Status Using Chi-Square Tests

2.3.3. Spearman correlation analysis

The Spearman correlation analysis illustrated in Figure 3 revealed several strong relationships between predictor variables. A high positive correlation was found between smoking and yellow fingers ($\rho \approx 0.85$). Similarly, wheezing and shortness of breath showed a strong correlation ($\rho \approx 0.72$). Coughing and chest pain were also significantly correlated ($\rho \approx 0.65$). Additionally, fatigue and chronic disease also showed a moderate correlation ($\rho \approx 0.5$), indicating possible underlying health links. These relationships reflect commonly co-occurring symptoms or behaviors. In contrast, some variables demonstrated unexpectedly weak correlations with lung cancer status. Smoking, age, and gender had very low correlation coefficients with lung cancer, suggesting that these variables, while clinically important, do not show strong monotonic relationships with lung cancer status in this dataset. These findings highlight the complexity of variable interactions and further support the use of multivariable modeling techniques that can account for interdependencies and nonlinear relationships.

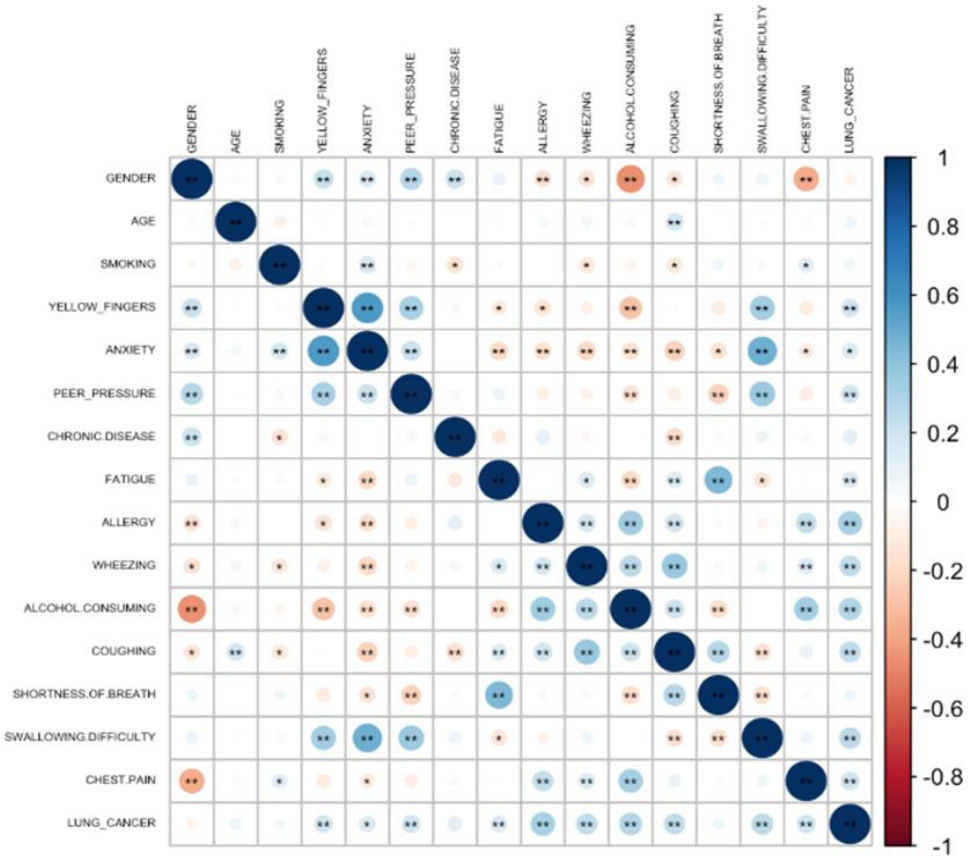

Figure 3. Spearman Correlation Matrix of Variables Associated with Lung Cancer

2.4. Machine Learning Models

In this paper, three different types of models are chosen, which are the Lasso Regression model (LR), Support Vector Machine (SVM), and eXtreme Gradient Boosting (XGB). All models in this study were trained and evaluated using a 70:30 split between the training and test sets. For the evaluation of the models, this paper chooses to use the accuracy calculated by the confusion matrix of $2 * 2$ for evaluation. The specific definition of the confusion matrix and the calculation formula are shown in Table 2 and Equation (1).

Table 2. Structure of a Binary Classification Confusion Matrix

		Prediction Condition	
		Positive (P)	Negative (N)
Actual Condition	Positive (P)	True Positive (TP)	False Negative (FN)
	Negative (N)	False Positive (FP)	True Negative (TN)

$$Accuracy = \frac{TN+TP}{TP+FN+FP+TN} \quad (1)$$

Feature selection was refined using both the Minimum Squared Error (MSE) and the 1-Standard Error (1SE) criteria to enhance model performance in Lasso regression model. The LR model is trained in reduced dimensionality data and evaluated on the test set to produce the results.

In the SVM model, the optimal parameters are selected through cross-validation to obtain the best model. The XGB model also used the dimensionality reduced data, cross-validation was performed to select the optimal parameters, and the final model was trained on the test set for prediction and evaluation. Finally, a conclusion is drawn by comparing the accuracy of LR model, SVM model and XGB model.

3. Results and Discussion

3.1. Lasso Regression Model

Initially applied with MSE shown in Figure 4, all 15 variables have been selected for cross validations. Subsequently, the 1SE rule was applied, refining the model by reducing the number of variables to 11 listed in Table 3, eliminating four that were less influential. Performance metrics, including accuracy, sensitivity, specificity was calculated to evaluate the model with the 11 selected variables.

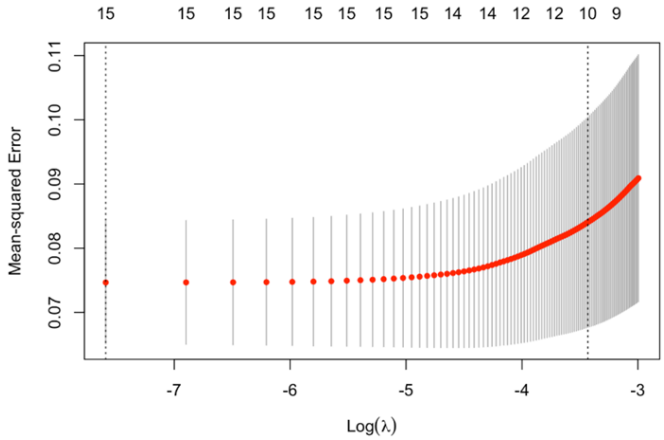

Figure 4. MSE Variable Selection on 15

Table 3. Model Coefficients after 1SE Rule

Variable	Information
Intercept	-4.8688
Gender	-
Age	0.015
Smoking	-
Yellow Finger	0.7715
Anxiety	0.6134
Peer Pressure	0.7645
Chronic Disease	0.4364
Fatigue	1.2522
Allergy	1.4454
Wheezing	-
Alcohol Consuming	0.8957
Coughing	1.1951
Shortness of Breath	-
Swallowing Difficulty	1.7766
Chest Pain	1.0616

From the Figure 5, the author can see the relationship between the regularization parameter and the cross-validation accuracy. When the regularization parameter is close to 0.028, the model reaches the highest accuracy of about 0.907. As the parameter increases, the accuracy gradually decreases. The final selected model parameters (alpha=1, lambda=0.028).

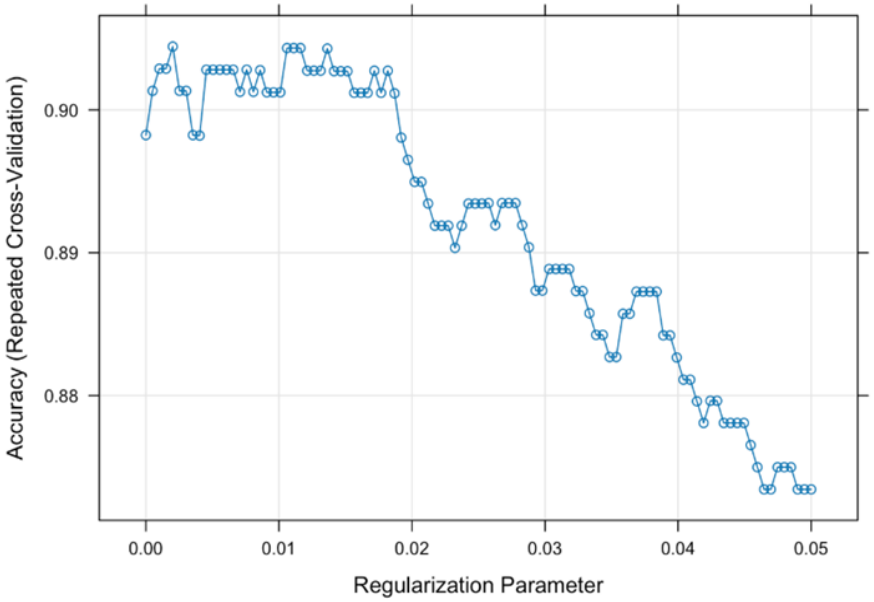

Figure 5. Repeated Cross Validation of Lasso Regression

According to the results of model testing demonstrated in Figure 6, the accuracy 0.860 is a high rate, suggesting the model is generally effective in classification tasks. The sensitivity 91.14% indicates strong performance in correctly identifying positive cases (72 true positives out of 79 actual Yes cases). However, the specificity is relatively low at 57.14%, suggesting weaker performance in detecting negative cases (only 6 true negatives out of 14 actual No cases). Overall, while the model is effective at minimizing false negatives, it tends to misclassify many No cases as Yes, as shown by the high number of false positives (8).

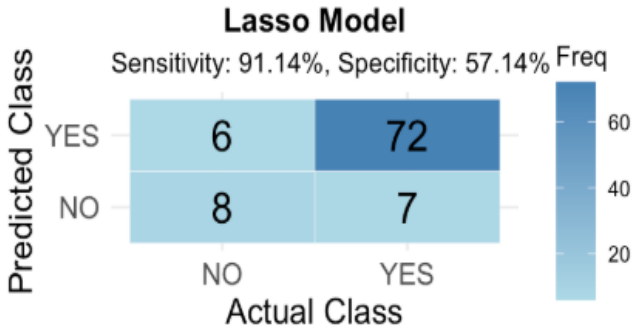

Figure 6. Confusion Matrix of Lasso Model on 11 selected variables

3.2. SVM Model

In realizing the support vector machine (SVM) model, this paper sets the random seed to "123" to ensure the replicability of the results based on all given variables. The results of parameter adjustment of the SVM model using the radial basis function kernel in lung cancer prediction show that under the condition of a fixed sigma value (0.04004), the model performance is best when the cost parameter C=4, reaches an accuracy of 91.98%. As can be seen from the figure 7, the accuracy increases first and then decreases with the increase of C value, reaching a peak at C=4 and then gradually decreasing, especially when C exceeds 16. The performance is significantly reduced, which shows that moderate regularization provides the best generalization ability for the model and balances the model complexity and data fitting degree.

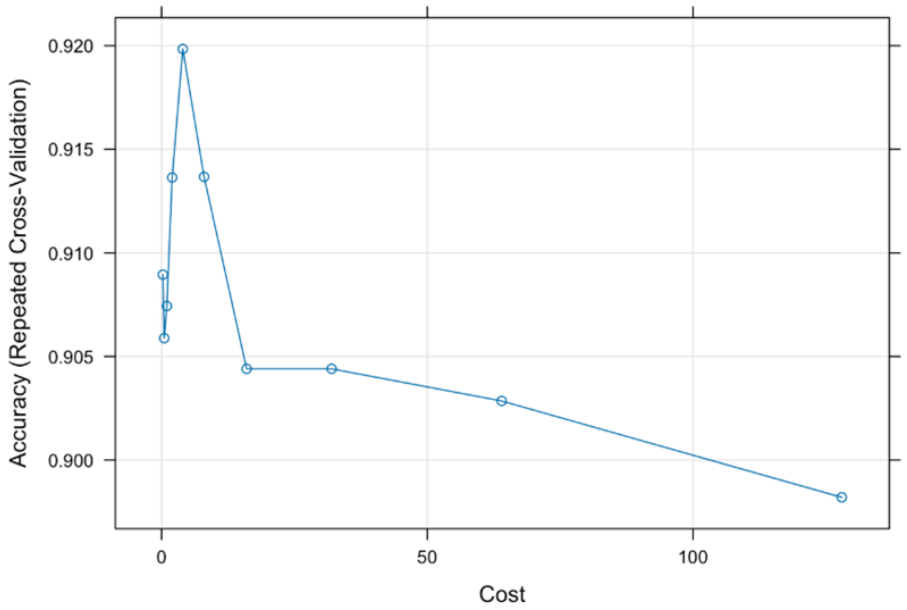

Figure 7. Repeated Cross Validation of SVM Model

The confusion matrix for the SVM model from Figure 8 shows that it achieves a very high sensitivity of 97.47%, correctly identifying 77 out of 79 actual Yes cases, indicating excellent performance in detecting positive instances. However, its specificity is low at 35.71%, meaning it correctly classifies only 5 out of 14 actual No cases. The model misclassifies many negative cases as positive (9 false positives), reflecting a tendency to overpredict the positive class.

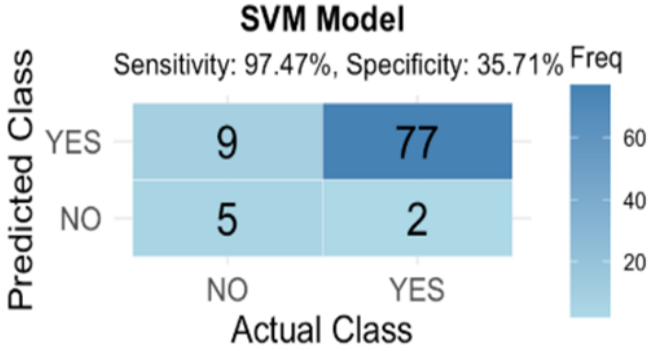

Figure 8. Confusion Matrix of SVM Model

This SVM model is highly sensitive, meaning it is excellent at detecting individuals with lung cancer and minimizing false negatives (which is critical in medical diagnosis). However, it has poor specificity, leading to many false positives which incorrectly classify healthy individuals as having lung cancer. This imbalance might cause unnecessary concern or follow-up testing in clinical settings.

3.3. XGBoost Model

The hyperparameter adjustment results of the XGBoost model in lung cancer prediction in Figure 9 show that under the conditions of fixed gamma value (0) and min_child_weight value (1), the best parameter combination of the model is: nrounds = 50, max_depth = 3, eta = 0.4, colsample_bytree = 0.8 and subsample = 0.5. This parameter combination is selected by the accuracy indicator, using the principle of taking the maximum value. The model setting reflects a moderate tree depth (max_depth=3) to prevent overfitting, a higher learning rate (eta=0.4) to accelerate convergence, and the model generalization ability is enhanced through column sampling (colsample_bytree=0.8) and row sampling (subsample=0.5) mechanisms. The combination of these parameters enables the model to capture the characteristics of lung cancer prediction while effectively controlling the model complexity and achieving good prediction performance.

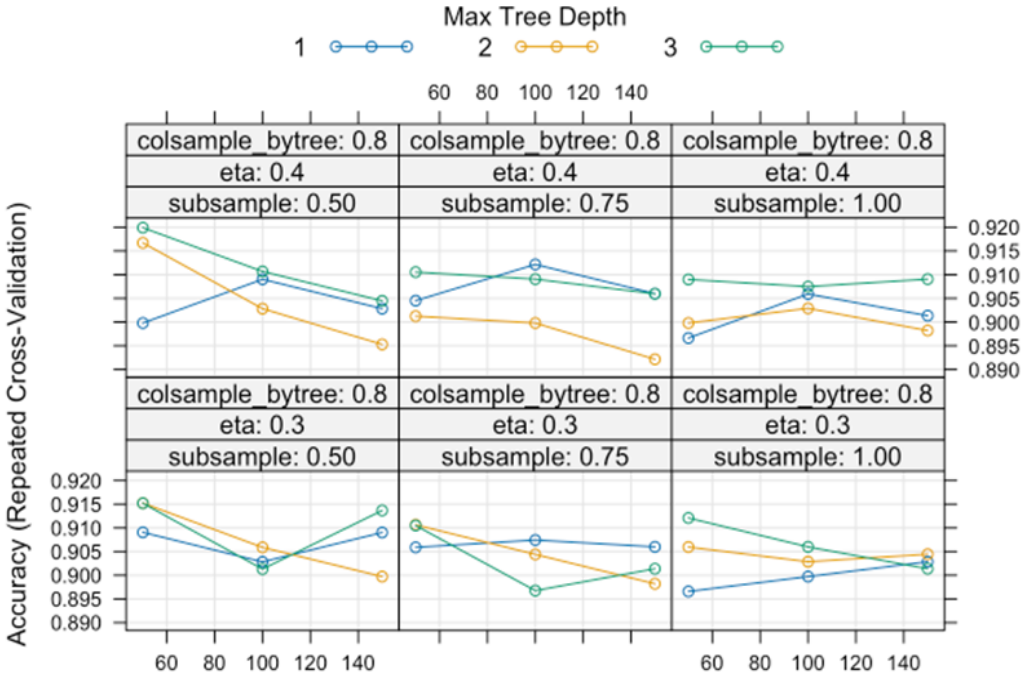

Figure 9. Repeated Cross Validation - Boosting Iteration

From Figure 10, the XGBoost model shows excellent sensitivity (92.41%) with only 6 false negatives, and improved specificity (71.43%) with 10 false positives, outperforming the previous Lasso model. The confusion matrix reveals better balanced classification (73 true negatives, 4 true positives), though class imbalance persists (83 "No" vs 10 "Yes"). This suggests stronger performance in identifying both positive and negative cases, with potential for further optimization through threshold adjustment or class rebalancing techniques.

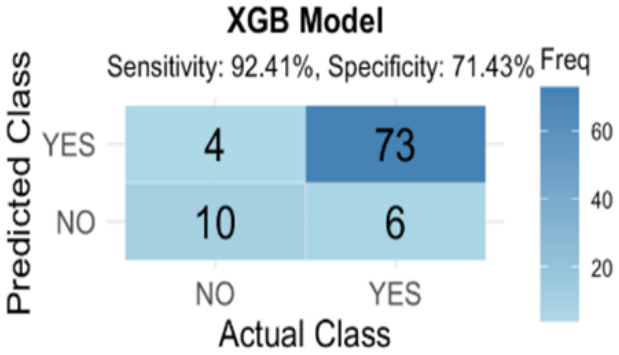

Figure 10. Confusion Matrix of XGB Model

The XGBoost model shows a strong balance between high sensitivity and moderate specificity, making it effective at detecting most lung cancer cases while also reducing false positives compared to the SVM model. This trade-off enhances the model's overall reliability for both identifying patients at risk and minimizing unnecessary alerts for non-cancer cases.

4. Conclusion

Overall, by comparing the performance of three models on lung cancer prediction, this study found that the XGB demonstrated the highest overall accuracy and offered a strong balance between sensitivity and specificity, making it particularly effective in identifying both positive and negative cases. The Lasso model, while effective at minimizing false negatives, suffered from lower specificity, and the SVM model, though highly sensitive, produced many false positives. These results highlight the importance of model selection based on the clinical context, where trade-offs between sensitivity and specificity must be carefully considered. Importantly, the performance of the model is affected

by multiple factors, including feature selection, data preprocessing, and choice of model parameters. Future work could focus on threshold tuning, rebalancing techniques, or applying these models to larger and more diverse datasets to enhance their generalizability and reduce diagnostic errors in lung cancer screening.

References

[1] Krist, Alex H, et al. Screening for Lung Cancer: US Preventive Services Task Force Recommendation Statement. JAMA: the journal of the American Medical Association, 2021, 325 (10): 962 - 970.

[2] Wu Fuzong, et al. Differences in Lung Cancer Characteristics and Mortality Rate between Screened and Non-Screened Cohorts. Scientific reports, 2019, 9: 19386 - 193867.

[3] Anil Kumar C, et al. Lung Cancer Prediction from Text Datasets Using Machine Learning. BioMed research international, 2022.

[4] Ji Guiyi, et al. Current Lung Cancer Screening Guidelines May Miss High-Risk Population: A Real-World Study. BMC cancer, 2021, 21: 50.

[5] Nour Mohammadi Liang, Patient-Provider Lung Cancer Screening Discussions: An Analysis of a National Survey. Clinical Lung Cancer, 2024, 25: 189 - 195.

[6] Hawkins Neil, et al. No Study Left Behind: A Network Meta-Analysis in Non–Small-Cell Lung Cancer Demonstrating the Importance of Considering All Relevant Data. Value in health, 2009, 12: 996 - 1003.

[7] Onozato Yuki, et al. Predicting Pathological Highly Invasive Lung Cancer from Preoperative [18F] FDG PET/CT with Multiple Machine Learning Models. European journal of nuclear medicine and molecular imaging, 2023, 50: 715 - 726.

[8] Lavanya C, et al. Novel Biomarker Prediction for Lung Cancer Using Random Forest Classifiers. Cancer informatics, 2023, 22.

[9] Pu Lucy, Rajeev Dhupar, Xin Meng. Predicting Postoperative Lung Cancer Recurrence and Survival Using Cox Proportional Hazards Regression and Machine Learning. Cancers, 2025, 17: 17 - 33.

[10] Kojima Kensuke, et al. Nonlinear Association between PD-L1 Expression Levels and the Risk of Postoperative Recurrence in Non-Small Cell Lung Cancer. Scientific reports, 2024, 14.

[11] Kumar Sandeep, et al. Optimized Predictive Models in Health Care Using Machine Learning. 1st ed. Newark: John Wiley & Sons, Incorporated, 2024.