# Development of an Artificial Intelligence Model for Predicting Chemotherapy Response in Non-Small Cell Lung Cancer (NSCLC) Patients Using CT Imaging and Genomic Data

**Chanyanuch Oumyoo[1], Kornkamon Thonghom[1]**

**Rattaporn Klinbantom[1*]**

*[1]Princess Chulabhorn Science High School Lopburi, Huai Pong, Khok Samrong, Lopburi 15120, Thailand*

**E-mail: 04796@pccl.ac.th*

## Abstract

This study aimed to develop artificial intelligence models to predict chemotherapy response in patients with non-small cell lung cancer (NSCLC) using radiomic features from computed tomography (CT) images and genomic gene expression data. Radiomic features were quantitatively extracted from segmented tumor regions, while genomic data consisted of selected gene expression features associated with treatment response. Logistic Regression and Random Forest models were constructed and evaluated using the AUC, accuracy, and sensitivity. For radiomic-based prediction, Logistic Regression achieved superior performance with an AUC of approximately 0.83, an accuracy of about 0.82, and a sensitivity of around 0.89, compared with Random Forest (AUC ≈ 0.76, accuracy ≈ 0.68). These results suggest that radiomic and genomic data combined with machine learning techniques have potential value for predicting chemotherapy response in NSCLC patients and may serve as a foundation for future clinical decision support tools.

Keywords : NSCLC, Radiomics, Genomics, Machine Learning

## Introduction

Lung cancer remains one of the leading causes of cancer-related mortality worldwide, with Non-Small Cell Lung Cancer (NSCLC) representing the most prevalent subtype (Sung et al., 2021). Chemotherapy is one of the standard treatment options for patients with NSCLC. However, individual responses to chemotherapy vary considerably. While some patients achieve favorable therapeutic outcomes, others exhibit poor responses or experience severe adverse effects, resulting in reduced treatment efficacy and unnecessary toxicity (Thai et al., 2021).

In recent years, increasing attention has been directed toward the concept of Precision Medicine, which aims to provide personalized treatment strategies based on each patient's biological and clinical characteristics (Ashley, 2021). Among the most valuable sources of information for precision medicine are medical imaging and genomic data.

Computed tomography (CT) images can be quantitatively analyzed using Radiomics, a technique that extracts high-dimensional imaging features, including tumor shape, intensity, and texture. These quantitative features may reflect the underlying biological characteristics of tumors in a non-invasive manner, eliminating the need for additional surgical procedures or tissue biopsies (Lambin et al., 2021). At the same time, genomic data, particularly gene expression profiles, provide valuable insights into the molecular activities of cancer cells, including mechanisms associated with tumor growth, metastasis, and responses to chemotherapeutic agents (Vogelstein et al., 2021). However, genomic datasets typically contain thousands of genes, making feature selection an essential step for identifying the most informative biomarkers and improving predictive performance (Liu et al., 2023).

Recent advances in Artificial Intelligence (AI) and Machine Learning (ML) have enabled the integration and analysis of large-scale multimodal datasets, facilitating the development of predictive models with improved accuracy and clinical applicability (Topol, 2024). Therefore, this project aims to develop an AI-based predictive model for chemotherapy response in patients with Non-Small Cell Lung Cancer (NSCLC) by integrating two complementary data modalities: radiomic features extracted from CT images and gene expression data. Furthermore, the study compares the predictive performance of these two data types to demonstrate the potential of combining medical imaging and biological information for supporting clinical decision-making and advancing precision medicine.

## Methodology

### 1) Data Sources

1. Radiomics Data : Computed tomography (CT) images of patients diagnosed with Non-Small Cell Lung Cancer (NSCLC) were used to extract quantitative radiomic features. These features characterize the tumor in terms of its morphology, intensity, and texture, providing non-invasive information about tumor heterogeneity.

2. Genomic Data : Gene expression profiles were obtained for each patient to represent the expression levels of thousands of genes within cancer cells. These data provide

molecular information related to tumor biology and treatment response.

3. Clinical Response Labels : Clinical outcomes following chemotherapy were used as the ground-truth labels for supervised machine learning. Each patient was classified into one of two categories, Response = 1: Responsive to chemotherapy and Response = 0: Non-responsive to chemotherapy

**2) Data Preprocessing**

Radiomic features were extracted from CT images using radiomics analysis techniques. Quantitative imaging features describing tumor shape, intensity, and texture were generated for each patient.

The gene expression dataset contains a large number of genes, many of which may not be informative for predicting chemotherapy response. Therefore, feature selection was performed to identify genes significantly associated with treatment outcomes.

The Pearson correlation coefficient was calculated between the expression level of each gene and the chemotherapy response label. The Pearson correlation coefficient is defined as:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

The selected radiomic features and gene expression features were integrated with the corresponding chemotherapy response labels using the Patient ID as the unique identifier. The resulting dataset served as the input for machine learning model development.

**3) Model Development**

1. Logistic Regression, a linear classification model used to predict whether a patient would respond to chemotherapy.

2. Random Forest, an ensemble learning algorithm that combines multiple decision trees to improve prediction accuracy and reduce overfitting.

Separate predictive models were developed using: Radiomics features only and Gene expression features only. The performance of these models was subsequently compared.

**4) Model Evaluation**

The integrated dataset was divided into training and testing subsets. Model performance was evaluated using several classification metrics, including Accuracy, Receiver Operating Characteristic (ROC) Curve and Area Under the ROC Curve (AUC).

The predictive performance of models developed from radiomics and genomics data was compared to determine which data modality provided superior predictive capability for chemotherapy response.

**5) Results Visualization**

The experimental results were presented using both tables and graphical visualizations.

ROC curves were generated to illustrate the diagnostic performance of each predictive model. In addition, comparative performance graphs were created to facilitate direct comparison between different machine learning models and data modalities. All evaluation results and figures were saved for inclusion in the final research report.

**6) Research Tools**

Python was used as the primary programming language for data preprocessing, statistical analysis, machine learning model development, and result visualization due to its extensive ecosystem of scientific computing libraries. The following software packages were utilized are Pandas and NumPy, Scikit-learn, Matplotlib and 3D Slicer (Version 5.10.0)

## Result and Discussion

**1) Radiomics**

This analysis included radiomic features extracted from CT images and clinical information from 40 patients diagnosed with NSCLC.

Logistic Regression achieved an AUC of 0.83, demonstrating good predictive performance despite limited accuracy in identifying surviving patients due to class imbalance.

Random Forest achieved an AUC of 0.76, showing slightly lower overall performance.

Overall, Logistic Regression outperformed Random Forest, suggesting that it is more suitable for small and imbalanced radiomics datasets.

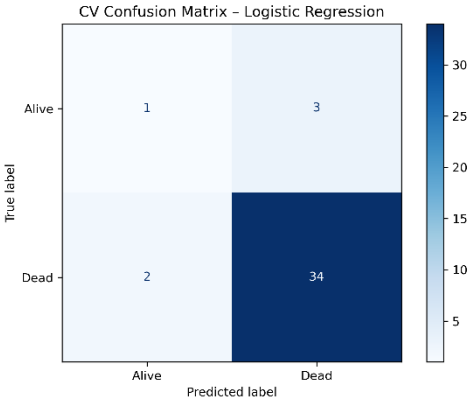


**Fig.1** Prediction performance of the Logistic Regression model.

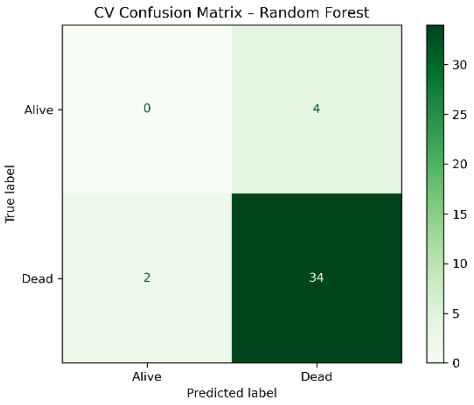


**Fig.2** Prediction performance of the Random Forest model.

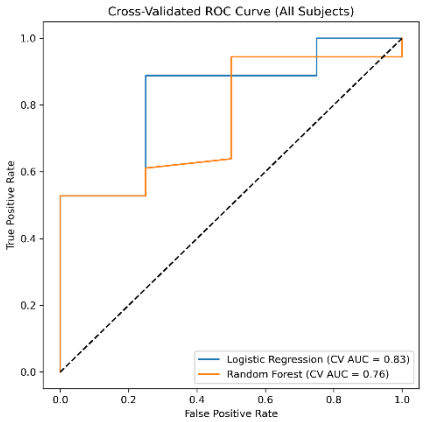


**Fig.3** ROC curve comparison between Logistic Regression and Random Forest using radiomic features.

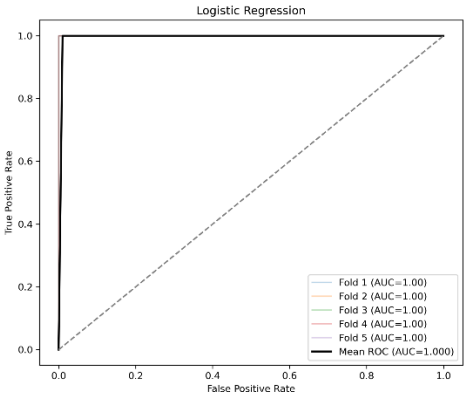


**Fig.4** ROC curves of Logistic Regression using genomic features.

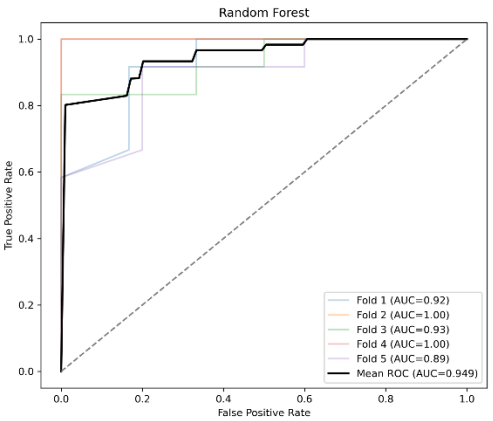


**Fig.5** ROC curves of Random Forest using genomic features.

### 2) Genomics

The genomic dataset included 89 NSCLC patients with more than 60,000 gene expression features. After Pearson correlation analysis, the 500 most relevant genes were selected for model training.

Logistic Regression achieved a mean AUC of 1.000, indicating excellent predictive performance, although the possibility of overfitting should be considered.

Random Forest achieved a mean AUC of 0.949, also demonstrating excellent classification performance.

Overall, genomic features provided higher predictive accuracy than radiomic features, highlighting their potential for predicting chemotherapy response and supporting Precision Medicine in NSCLC patients.

### Conclusions

Analysis of the radiomics dataset consisting of 40 NSCLC patients demonstrated that both Logistic Regression and Random Forest were capable of predicting patient survival outcomes. Logistic Regression achieved an AUC of 0.83, outperforming Random Forest (AUC = 0.76), suggesting that linear models may perform better

when datasets are relatively small and class imbalance is present. For the genomic dataset comprising 89 patients, more than 60,000 gene expression features were reduced to the 500 most informative genes using correlation-based feature selection. Using these selected features, Logistic Regression achieved a mean AUC of 1.000, whereas Random Forest achieved a mean AUC of 0.949. These findings demonstrate the strong predictive capability of genomic biomarkers for chemotherapy response in NSCLC patients.

## Recommendations for Future Research

Future studies should include larger and more balanced patient cohorts to improve model robustness and predictive accuracy. External validation using independent datasets collected from multiple institutions should also be conducted to assess the generalizability of the proposed models. Furthermore, integrating radiomic and genomic data within a Radiogenomics framework may further enhance predictive performance and contribute to the advancement of Precision Medicine for patients with Non-Small Cell Lung Cancer (NSCLC).

## Acknowledgments

Special thanks to The Cancer Imaging Archive Website for information and public data, and Teacher Mrs. Rattaporn Klinbantom for advice and data collection. We also thank Princess Chulabhorn Science High School, Lopburi, for support in carrying out this study.

## References

Bera, K., Braman, N., Gupta, A., Velcheti, V., & Madabhushi, A. (2022). Predicting cancer outcomes with radiomics and artificial intelligence in radiology. Nature Reviews Clinical Oncology, 19(2), 132–146. https://doi.org/10.1038/s41571-021-00560-7

Cancer Imaging Archive. (2024). NSCLC Radiogenomics Collection. The Cancer Imaging Archive (TCIA). https://www.cancerimagingarchive.net/collection/nsclc-radiogenomics/

Chen, Q., Wang, S., Shi, J., Wu, S., & Wang, Y. (2021). Current status and quality of radiomic studies for predicting immunotherapy response and outcome in patients with non-small cell lung cancer: A systematic review and meta-analysis. European Journal of Radiology, 145, 110041. https://doi.org/10.1016/j.ejrad.2021.110041

Imaging Data Commons. (2024). National Cancer Institute Imaging Data Commons. https://portal.imaging.datacommons.cancer.gov/