

# Non-Invasive Classification of EGFR Mutation from Thoracic CT Images Using Radiomics Features and LightGBM

**Reo Takahashi**

*Kyushu Institute of Technology, 1-1 Sensui, Tobata-ku, Kitakyushu, 804-8550, Japan*

**Tohru Kamiya**

*Kyushu Institute of Technology, 1-1 Sensui, Tobata-ku, Kitakyushu, 804-8550, Japan*

**Takashi Terasawa**

*University of Occupational and Environmental Health, 1-1 Iseigaoka, Yahatanishi-ku, Kitakyushu, 807-8555, Japan*

**Takatoshi Aoki**

*University of Occupational and Environmental Health, 1-1 Iseigaoka, Yahatanishi-ku, Kitakyushu, 807-8555, Japan*

*Email: takahashi.reo828@mail.kyutech.jp*

## Abstract

Cancer caused 9.7 million deaths in 2022, including 1.8 million from lung cancer the leading cause of cancer death. EGFR mutation testing is essential for lung cancer treatment planning, but it is invasive and visual identification from chest CT images is difficult. This paper proposes a computer-aided diagnosis system to identify EGFR mutation status. Lung tumor regions were automatically extracted and radiomics features were obtained. Dimensionality reduction was performed using null importance, variance inflation factor, and recursive feature elimination. The method was applied to 143 cases and achieved an accuracy of 59.1%, a true positive rate of 54.3% and a false positive rate of 36.1%. The results suggest that CAD (Computer-Aided Diagnosis) systems can improve the non-invasive detection of EGFR mutations in lung cancer.

*Keywords:* Computer Aided Diagnosis, Radiomics, U-Net, LightGBM

## 1. Introduction

Cancer is the world's deadliest disease, with an estimated 20 million new cases of cancer and 9.7 million deaths by 2022. According to estimates, one in five people will develop cancer in their lifetime, and one in nine men and one in twelve women die of the disease. In terms of lung cancer alone, 2.5million new cases will be diagnosed by 2022, accounting for one-eighth of all cancers worldwide. Lung cancer is also the leading cause of deaths by cancer site, accounting for about 1.8 million deaths [1]. This means that many people die from lung cancer each year, requiring early detection, early treatment, and effective therapy. In order to provide effective treatment, testing for the presence of driver gene mutations may be performed. Driver genes are a general term for genes involved in the development and progression of cancer. If a mutation in this gene is found, it allows the use of molecularly targeted drugs that can have a dramatic effect on cancer treatment [2]. This therapy is less stressful on the body and more effective than conventional anticancer drugs. However, testing for the presence of genetic mutations is usually done by biopsy, which is invasive for the patient [3]. Furthermore, it is difficult for physicians to confirm the presence or absence of genetic mutations from CT images. Therefore, to reduce the burden on physicians and patients, it is necessary to develop a computer-aided diagnosis [4] system that noninvasively classifies the presence or absence of EGFR gene mutations using CT images.

Although there is study [5] on this topic, require the physician to extract lung tumor regions. This task is very burdensome for physicians. In this paper, we propose an end-to-end method to automatically extract lung tumor regions and identify the presence or absence of genetic mutations in the obtained regions.

## 2. Methodology

The flow of the proposed method in this paper is shown in Fig. 1. Specifically, extraction is performed using a model based on U-Net with some modifications to optimally extract lung tumor regions. Next, radiomics features are extracted from those regions and dimensionality reduction is performed using a combination of null importance, Variance Inflation Factor and Recursive Feature Elimination. Then, LightGBM was used to classify the presence or absence of genetic mutations.

### 2.1. Extraction of Lung Tumor

In this paper, our proposed prior method, Improved U-Net [6], is used as the base model. This method introduces MultiRes Block [7] and CBAM (Convolutional Block Attention Module) [8] to U-Net [9]. These modifications allow us to extract features from multiple scales and to identify which features to focus on in those features. In this paper, we further applied ASPP (Atrous Spatial Pyramid Pooling) and ensemble learning

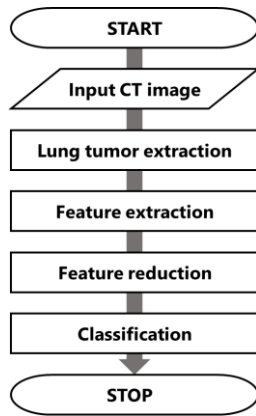


Fig.1. The flowchart of the method

to this model. ASPP enables the acquisition of diverse contextual information without increasing the number of parameters, allowing for more precise identification of the complex internal structure of tumors and their boundaries with surrounding lung tissue. Additionally, by ensembling two models with different layers, we aimed to improve the extraction accuracy of small lung tumors. Using this enhanced U-Net model, we performed automatic extraction lung tumors.

## 2.2. Feature extraction

In this paper, radiomics features were extracted from DICOM data for regions automatically segmented using an improved U-Net. Radiomics features enable the extraction of shape, intensity, and texture features from radiological images [10]. Compared to traditional biopsy-based analyses, this approach allows for the extraction of high-dimensional quantitative tumor characteristics with reduced burden on patients. Additionally, to analyze CT images across different frequency components, radiomics features were extracted from images processed with wavelet transformation.

## 2.3. Feature reduction

Feature reduction is the process of selecting only the important features among those obtained by feature extraction, thereby reducing the computational cost and preventing over-fitting to unnecessary noise data. In our previous study, feature reduction was performed by null importance [5]. However, most of the features selected were features obtained from wavelet transforms or texture features, which were not independent of each other. Therefore, in this paper, we devised a method to select features that are statistically independent from each other. Specifically, null importance was first applied separately to features obtained from the original images and those derived from wavelet-transformed images. This step retains broadly significant features and eliminates those that do not contribute to model learning. The separate application is necessary because applying null importance to the combined features would leave

few features from the original images. Next, VIF (Variance Inflation Factor) [11] was used to remove highly correlated features among those retained from the first stage. VIF quantifies the degree of multicollinearity, with higher values indicating stronger multicollinearity. Generally, VIF values above 10 suggest significant multicollinearity. In this paper, the VIF for each feature was calculated, and features were recursively removed until all remaining features had VIF values below 10. Finally, RFE (Recursive Feature Elimination) [12] was applied to the features obtained from the VIF step to further select the most important features. RFE recursively eliminates the least important features, starting with all features as input, thereby retaining only the features that most significantly impact model performance. By combining the filter-based, wrapper-based, and embedded methods for feature reduction, it is considered possible to select features with low redundancy and balanced representation.

## 2.4. Classification

For classification, we used LightGBM [13], a gradient boosting decision tree algorithm. It is a decision tree that grows Leaf-Wise instead of Level-Wise in the process of gradient boosting, which allows for quick and accurate learning. In this paper, we further implemented a two-stage learning approach for classification. Two-stage learning is a method that connects two models in series to compensate for each model's weaknesses. First, classification is performed using LightGBM as usual. When the predicted probability falls between 0.3 and 0.7, it is considered to have low confidence. Data that could not be accurately classified in this range are collected, and a second model is constructed to reclassify these instances. Finally, the results from the two models are combined to improve the overall accuracy.

## 3. Results and Discussion

### 3.1. Experimental and Evaluation Methods

The images used in this paper were obtained from the University of Occupational and Environmental Health Hospital, and the lung tumor regions were annotated under the guidance of physicians. A dataset consisting of 452 chest CT images from 143 cases was used. Leave-One-Out cross-validation was performed for model validation, and the evaluation metrics used were AUC (area under the curve), accuracy, TPR (true positive rate), FPR (false positive rate). As shown in Table 1, cases classified as having genetic mutations were considered positive cases, while those classified as not having genetic mutations were considered negative cases. Accuracy, TPR, and FPR were calculated using the following equations, where a, b, c, and d are defined as shown in Table 1.

$$\text{Accuracy} = \frac{a + d}{a + b + c + d} \times 100[\%] \quad (1)$$

$$\text{TPR} = \frac{a}{a + c} \times 100[\%] \quad (2)$$

$$\text{FPR} = \frac{b}{b + d} \times 100[\%] \quad (3)$$

### 3.2. Results

In this paper, as in previous research, classification was performed using LightGBM by integrating features obtained from chest CT images with clinical information, specifically gender. Table 2 compares the results of the single stage learning method without feature reduction, the proposed method, and the methods of the previous research [5]. The previous method used null importance for feature reduction and performed single-stage learning, while the proposed method combined null importance, VIF and RFE for feature reduction and performed both single-stage and two-stage learning. The proposed method obtained AUC=0.647, accuracy=59.1%, TPR=54.3%, FPR=36.1%. Compared to the previous method, the proposed method showed a 2.2% decrease in accuracy and a 3.1% decrease in TPR, indicating a decrease in discrimination accuracy. Comparing single-stage and two-stage learning, the two-stage learning reduced the accuracy by 1.4% and the TPR by 0.8%.

### 3.3. Discussion

In this paper, we adopted a multi-step feature selection method that combines Null Importance, VIF, and RFE. Table 3 shows the types of features selected by Null Importance and the proposed method. Table 3 shows that the number of features obtained from the original image increased with the proposed method, and the use of VIF reduced the correlation between features. On the other hand, many features were still selected from wavelet transforms and texture features, suggesting that features obtained from wavelet transforms are important for the model to discriminate the presence of genetic mutations. In addition, Table 2 shows that there is no significant difference in accuracy between the cases where feature selection was performed and those where no feature selection was performed. This may be because the extracted radiomics features did not contain many significant features for classification. Normally, the

Table 1. Valuation basis

	Predicted Positive	Predicted Negative
Test Positive	<i>a</i>	<i>c</i>
Test Negative	<i>b</i>	<i>d</i>

purpose of feature reduction is to reduce noise features that do not contribute to classification, but if the original data lacks important information, the reduced features are unlikely to help with classification. Therefore, it is assumed that this feature reduction did not lead to an improvement in classification accuracy. Furthermore, no improvement in accuracy was achieved with two-stage learning.

The purpose of two-stage learning is to focus on data that was difficult to classify in the first stage and improve classification accuracy. However, because the first stage was not properly trained, a large amount of data was transferred to the second stage, and as a result, the second stage may not have been properly trained as well. In particular, the second stage used the same features and hyperparameters as the first stage, which may have limited learning for difficult data. For two-stage learning to be effective, the first stage should be properly trained, and the features and parameters appropriate for the second stage should be reviewed. In a previous study, binary classification of the presence or absence of a genetic mutation based on manually extracted lung tumor regions had an accuracy of 92%. Although this method is very accurate, it has difficulties in practical application because it requires manual extraction of regions. On the other hand, this study proposed an end-to-end classification method and obtained an accuracy of 61.5%. This accuracy is still low and needs further improvement. In this study, we used 2D CT images to extract features, but the features obtained from 2D images were limited, which may have contributed to the lack of significant features for classification. In addition, the randomly selected slices included some cases with small tumor cross-sections, which probably made it more difficult to classify these cases. In the future, further improvement in accuracy is expected by automatically extracting lung tumor regions from 3D images and extracting more diverse features. In addition, the current dataset contains only 452 images, which is very limited. In such a situation, the model may not be able to learn enough diverse information, resulting in poor generalization performance. Therefore, expanding the dataset is also an important issue for the future.

### 4. Conclusion

In this paper, we developed a computer-aided diagnosis (CAD) system to identify the presence or absence of EGFR mutations from thoracic CT images, providing a less invasive method for EGFR mutation detection. By automating the extraction of lung tumor regions, this system not only reduces the burden on physicians, but also provides a non-invasive approach for patients. For feature selection, a achieved a classification performance of AUC = 0.647, accuracy = 59.1%, TPR = 54.3%, and FPR = 36.1%. To further improve the classification performance, we plan to improve the region extraction model, introduce new methods for optimal feature selection, and expand the dataset.

Table 2. Classification result (Acc. : Accuracy)

	TPR	FPR	Acc.	AUC
No feature reduction	51.4	30.0	61.5	0.661
Previous method	57.4	34.5	61.3	0.651
Single stage	55.1	33.9	60.5	0.654
Proposed method	54.3	36.1	59.1	0.647

Table 3. Selected Features Comparison

		shape	firstorder	texture
Previous method	Original	1	0	1
	Wavelet	-	5	11
Proposed method	Original	1	1	3
	Wavelet	-	3	8

5. References

1. F. Bray, M. Laversanne, H. Sung, J. Farlay, RL. Siegel, I. Soerjomataram and A. Jemal, Global Cancer Statistics 2022: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries, CA: A cancer Journal for Clinicians, Vol. 74(3); 2024, pp.229-263.
2. G. Q. Zhu, M. S. Zhang, X. X. Ding, B. He and Q. H. Zhang, Driver Genes in Non-small Cell Lung Cancer: Characteristics, Detection Methods, and Targeted Therapies, Oncotarget, Vol. 8(34); 2017, pp. 57680-57692.
3. J. Marrugo-Ramirez, M. Mir and J. Samitier, Blood-based Cancer Biomarkers in Liquid Biopsy: A Promising Non-invasive Alternative to Tissue Biopsy, International Journal of Molecular Sciences, Vol. 19(10); 2018, p. 2877.
4. K. Doi, Computer-Aided Diagnosis in Medical Imaging: Historical Review, Current Status and Feature Potential, Computerized Medical Imaging and Graphics, Vol. 31(4-5); 2007, pp. 198-211.
5. S. Watanabe, T. Kamiya, T. Terasawa and T. Aoki, Classification of Driver Gene Mutations from 3D-CT Images Based on Radiomics Features, In 2023 23<sup>rd</sup> International Conference on Control, Automation and Systems (ICCAS), 2023, pp. 1733-1736.
6. R. Takahashi, T. Kamiya, T. Terasawa and T. Aoki, Extraction of Lung Tumor Regions from Thoracic CT Images Using An Improved U-Net, In 2023 23<sup>rd</sup> International Conference on Control, Automation and Systems (ICCAS), 2023, pp. 1489-1493.
7. N. Ibtehaz and MS. Rahman, MultiResUNet: Rethinking the U-Net architecture for Multimodal Biomedical Image Segmentation, Neural Networks, Vol. 121; 2020, pp. 74-87.
8. S. Woo, J. Park, JY. Lee, IS. Kweon, Cbam: Convolutional Block Attention Module, Proceeding of the European Conference on Computer Vision (ECCV), 2018, pp. 3-19.
9. O. Ronneberger, P. Fischer and T. Brox, U-net: Convolutional Networks for Biomedical Image Segmentation, Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2015, pp. 234-241.
10. R. Thawani, M. Mclane, N. Beig, S. Ghose, P. Prasanna, V. Velcheti and A. Madabhushi, Radiomics and Radiogenomics in Lung Cancer: A review for The Clinician, Lung Cancer, Vol. 40(4); 2018, pp. 34-41.

11. M. O. Akinwande, H. G. Dikko and A. Samson, Variance Inflation Factor: As A Condition for the Inclusion of Suppressor Variable(s) in Regression Analysis, Open Journal of Statistics, Vol. 5(7); 2015, pp. 754-768.
12. I. Guyon, J. Weston, S. Barnhill and V. Vapnik, Gene Selection for Cancer Classification Using Support Vector Machines, Machine Learning, Vol. 46; 2002, pp. 389-422.
13. G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye and T. Y. Liu, LightGBM: A Highly Efficient Gradient Boosting Decision Tree, Proceedings of the 31st Advances in Neural Information Processing Systems (NeurIPS), Vol. 30, Curran Associates Inc., pp. 3149-3157.

Authors Introduction

Reo Takahashi



He is currently a master student in Kyushu Institute of Technology, Japan.

Tohru Kamiya, Prof., Ph.D.



He received his B.A. degree in Electrical Engineering from Kyushu Institute of Technology in 1994, the Master and Ph.D. degree from Kyushu Institute of Technology in 1996 and 2001, respectively. He is a professor in the Department of Mechanical and Control Engineering at Kyushu Institute of Technology. His research interests are focused on image processing and medical application of image analysis. He is currently working on computer aided diagnosis based on CT, MR imaging, fluorescence microscope imaging, and automatic classification of respiratory sound.

Takashi Terasawa, M.D., Ph.D.



He received his M.D. degree from University of Occupational and Environmental Health, Japan in 2011 and Ph.D. in medicine in 2020. He is a radiologist at the University of Occupational and Environmental Health, Japan.

Takatoshi Aoki, Prof. M.D., Ph.D.



He is a professor of Radiology at University of Occupational and Environmental Health, Japan, and the vice president of Japanese Society of Thoracic Radiology. His clinical and research interests include the imaging modalities of lung cancer, respiratory function.