Detection of Lung Nodules from CT Image Based on Ensemble Learning

Natsuho Baba[1], Tohru Kamiya[1], Takashi Terasawa[2], Takatoshi Aoki[2]
[1] *Kyushu Institute of Technology, 1-1 Sensui, Tobata, Kitakyushu, Fukuoka, 804-8550, Japan*
[2] *University of Occupational and Environmental Health, 1-1, Iseigaoka, Yahatanishi, Kitakyushu, Fukuoka, 807-8555, Japan*
Email: baba.natsuho781@mail.kyutech.jp, kamiya@cntl.kyutech.ac.jp

Abstract

Lung cancer is the most frequently diagnosed cancer worldwide and the leading cause of cancer-related deaths, making early detection and treatment crucial. Temporal subtraction system, one of the CAD, emphasize the differences between the current and previous images. In this study, radiomics features are extracted as explanatory variables from the temporal subtraction images. Feature selection is performed using Elastic Net, followed by the application of machine learning methods. Finally, ensemble learning is applied to classify unknown data into two categories: positive and negative lung nodules.

Keywords: Computer Aided Diagnosis, Machine Learning, Temporal Subtraction Technique, Radiomics, Ensemble Learning.

1. Introduction

According to GLOBOCAN 2022, there were 2.2 million new lung cancer cases (11.4%) and an estimated 1.8 million deaths from lung cancer in 2020 [1]. Lung cancers at early stages often appear as nodules include ground-glass opacities (GGO). Because these lesions are low density or small size, there is concern that they may go undetected by the interpreting physician.

One of the computer-aided diagnosis (CAD) systems that addresses this challenge is the temporal subtraction imaging technique. In this method, a temporal subtraction image is created by performing a subtraction operation between current and previous images of the same patient. The temporal subtraction image technique erases normal tissues, such as blood vessels and ribs, that do not change over time. It also highlights newly developed lesions or changes in existing abnormal tissues by comparing them to previous images [2]. The studies have reported that the use of temporal subtraction images can improve the accuracy of physician interpretation [3], [4].

In this paper, we propose a method for extracting and classifying lung cancer lesions from chest computed tomography (CT) images with the aim of reducing the workload of radiologists and improving detection accuracy. First, temporal subtraction images are generated, and a bounding box is created for each candidate region, which may contain both lung cancer lesions and artifacts. Next, radiomics features are extracted from these images as explanatory variables, and feature selection is performed using Elastic Net. Machine learning methods are then applied to classify the data. Finally, ensemble learning is used to combine the results of the machine learning models for the final classification.

2. Methodology

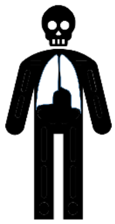
In this paper, the proposed method is applied to a dataset of 21 cases, each consisting of images acquired with a multidetector row CT scanner. A set of current and previous images of the same patient is considered one case.

2.1. *The temporal subtraction technique*

One of the CAD techniques anticipated for use in comparative reading in chest CT imaging is the temporal subtraction technique. This technique highlights temporal changes by subtracting two medical images that have been acquired at different times [5]. In addition, normal structures such as ribs and blood vessels are effectively removed, facilitating detection of newly emerging lung lesions [6].

Figure 1 shows an image of the temporal subtraction technique. In the figure, (a) is the original image, (b) is the previous state image, and (c) is the result image where image (b) has been subtracted from image (a), leaving only the lesion candidate area. Image (c) shows that a newly appeared lesion is enhanced and unchanged structures such as blood vessels, bones, and muscles are removed.

(a) The current state image



(b) The previous state image

(c) The temporal subtraction

Fig.1 The image of the temporal subtraction technique

2.2. *Extraction of candidate lesion areas from the temporal subtraction images*

The temporal subtraction images contain numerous residual artifacts. Therefore, in this study, regions with a diameter of 5 mm or greater are extracted from the generated temporal subtraction images as candidate lesion regions. Additionally, the extracted candidate lesion area is enlarged vertically and horizontally by 2 mm, and a Bounding Box is defined as the final region of interest (ROI). In this study, two methods are used: one is to use lesion candidate regions with a diameter of 5 mm or more, and the other is to use features extracted from a Bounding Box with an enlarged region.

2.3. *Radiomics features*

Radiomics features are quantifiable data derived from medical images using mathematical formulas and computer-based algorithms. These features capture and provide detailed information about the shape, brightness, and texture of lesions in medical images [7]. In this study, shape and texture features were extracted using the open-source library Pyradiomics [8].

2.4. *Feature selection*

Feature selection is a process that identifies only significant features from a dataset to reduce computational cost and minimize overfitting due to irrelevant noise. In this paper, we use a method called Elastic Net [9], [10] to identify the most relevant features for classification from the set of 1026 radiomics features.

Elastic Net is a penalized linear regression model that applies a linear combination of regularization penalties to the loss function during training. It is an analytical method that combines the characteristics of Ridge regression and Lasso (Least absolute shrinkage and selection operator), allowing it to perform both automatic feature selection, a characteristic of Lasso, and continuous shrinkage, a characteristic of Ridge regression, simultaneously.

2.5. *Machine learning classification*

In this paper, the features selected by Elastic Net are used as explanatory variables and machine learning methods. Three machine learning methods are used: XGBoost, LightGBM, and CatBoost. While machine learning is powerful due to its ability to handle numerous features, it involves many hyperparameters that significantly affect model performance. Optimizing these hyperparameters is crucial for improving the accuracy of machine learning models. In this study, hyperparameter tuning is performed using Optuna [11], a black-box optimization framework. Finally, ensemble learning is used to combine the results of the three machine learning methods to generate the final output.

(i) XGBoost (Extreme Gradient Boosting)
XGBoost [12], [13] is a gradient-boosted decision tree (GBDT) method for based classification and regression introduced in 2014. It is characterized by fast processing speed, high accuracy, and flexibility in data requirements. Its strengths include strong model generalization ability, high scalability, and fast computation.

(ii) LightGBM (Light Gradient Boosting Machine)
LightGBM [14], [15], developed in 2017, is a classification and regression method based on GBDT, similar to XGBoost. LightGBM is known for faster speed and lower memory consumption compared to XGBoost, achieving high efficiency with large datasets while maintaining high accuracy with smaller datasets.

(iii) CatBoost (Categorical Boosting)
CatBoost [16], [17], developed in 2018, is another classification and regression method based on GBDT. CatBoost excels in handling categorical variables and includes a built-in cross-validation function to mitigate overfitting. In addition, it performs well when handing outliers and missing values in the data.

(iv) Ensemble Learning
Ensemble learning [18], [19] is a technique that combines predictions from multiple models to improve reliability and generalizability. This approach reduces the risk of overfitting while maintaining strong predictive performance by averaging predictions from different models. Ensemble learning methods include bagging and stacking, which use boosting algorithms. These models are trained in parallel on different random subsets of data using alternative sampling, and the predictions from all models are aggregated.

3. Experimental Results

In this study, temporal subtraction images were generated from 21 lung cancer cases. Multiple slice images with lesions are selected and their performance is evaluated by 4-fold cross-validation using 231 candidate lesions. In the temporal subtraction images, newly appearing nodules over time were classified as positive lesions, while artifacts such as blood vessels and ribs were classified as negative lesions. The distribution of positive and negative lesions was 111 positive lesions and 120 negative lesions. The evaluation metrics were accuracy, true positive rate (TPR), false positive rate (FPR), and area under the curve (AUC). The results of the classification are shown in the table, with the input data in Table 1 as candidate lesion areas obtained from the temporal subtraction images and the input data in Table 2 as ROIs. Both Table 1 and Table 2 shows the results for methods that performed classification without feature reduction, methods that used Elastic Net for feature selection in previous study [20], and methods that incorporated ensemble learning.

Figure 2 shows an example image of the experimental results with the input data as bounding boxes. In Figure 2, (a) is the current image, (b) is the previous image, (c) is the temporal subtraction image, (d) is the mask image generated using the temporal subtraction image, (e) is a magnified view focusing on the Bounding Box of the current image, and (f) is a magnified view focusing on the Bounding Box of the temporal subtraction image. Arrows indicate lung cancer areas. The rectangular areas at (e) and (f) are the ROIs.

4. Discussions

Table 1 shows that ensemble learning achieves the best overall results when the lesions obtained from the temporal subtraction images are used as input data. In contrast, Table 2 shows that when the ROI, Bounding Box, is used as input data, the accuracy of ensemble learning is slightly less than that of LightGBM. This result is likely due to the greater variation in the accuracy of the three machine learning models in Table 2 compared to those in Table 1.

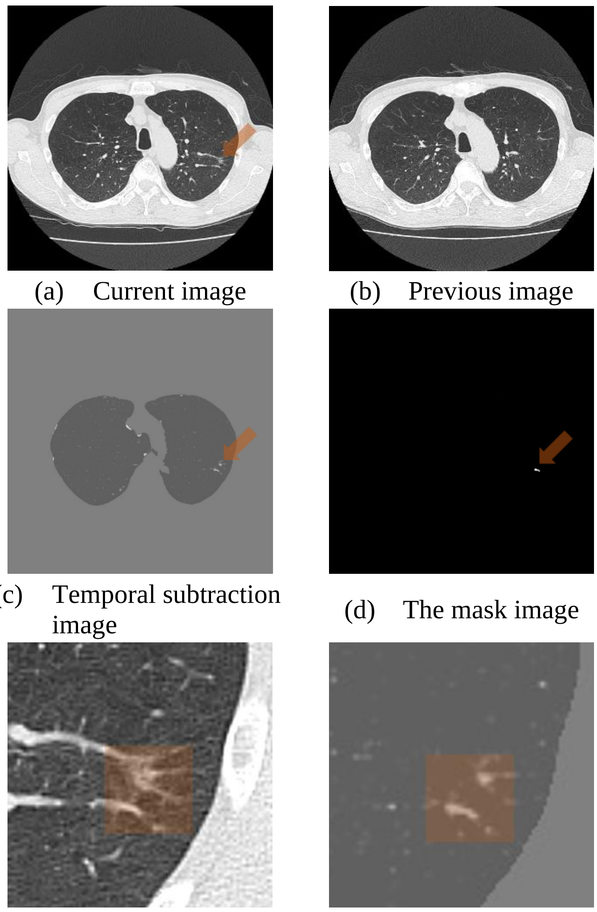
From Tables 1 and 2, the results in Table 2, where the lesion is changed to a ROI, show an overall improvement. Figure 2 also shows an example of improved discrimination by changing the input data from lesion candidate areas to ROIs. In images (e) and (f), the lesion candidate regions obtained from the temporal subtraction images were used as input data, so they were misclassified as blood vessels by feature extraction based on the shape of the region. However, by changing the input data to the bounding box, we were able to extract the features of regions containing light whitish areas, which is a characteristic of GGO, and we believe that we were able to correctly classify the data.

Table 1. Identification result from regions of interest (Acc.: Accuracy)

		TPR	FPR	Acc.	AUC
XGBoost	No feature reduction	67.29	30.6	67.11	0.721
	Elastic Net	83.60	21.04	80.96	0.836
LightGBM	No feature reduction	77.77	20.33	77.96	0.799
	Elastic Net	81.69	21.41	79.24	0.835
CatBoost	No feature reduction	68.59	26.97	70.15	0.782
	Elastic Net	80.82	28.30	75.35	0.812
Ensemble Learning		84.68	26.67	78.79	0.843

Table 2. Identification result from bounding box (Acc.: Accuracy)

		TPR	FPR	Acc.	AUC
XGBoost	No feature reduction	85.32	15.98	84.39	0.904
	Elastic Net	80.95	19.41	80.07	0.897
LightGBM	No feature reduction	87.65	10.78	87.87	0.917
	Elastic Net	90.92	11.92	89.17	0.949
CatBoost	No feature reduction	85.99	16.08	84.83	0.913
	Elastic Net	87.38	13.99	86.15	0.947
Ensemble Learning		90.09	15.00	87.45	0.939



(a) Current image
(b) Previous image
(c) Temporal subtraction image
(d) The mask image
(e) Enlarged image of the current image
(f) Enlarged image of the temporal subtraction image

Fig.2 Experimental results with bounding box

5. Conclusion

In this paper, we proposed a method for extracting and classifying lesions from chest CT images with the aim of reducing radiologist workload and improving lesion detection accuracy. The proposed method was applied to ROIs obtained from temporal subtraction images of 21 cases and achieved AUC 0.939, accuracy 87.45%, TPR 90.09%, and FPR 15.00% by ensemble learning. Future work includes further improving ensemble learning techniques, expanding the dataset, refining the feature selection methods, and optimizing the ensemble classification using XGBoost and LightGBM.

References

1. F. Bray et al., "Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries", A Cancer Journal for Clinicians, Vol. 74, No. 3, pp. 229-263, 2024.
2. T. Aoki et al., "CT temporal subtraction: techniques and clinical applications", Quant Imaging Med Surg, Vol. 11, No. 6, pp. 2214-2223, 2021.
3. M. C. Difazio et al., "Digital chest radiography: Effect of temporal subtraction images on detection accuracy", Radiology, Vol. 202, No. 2, pp. 447-452, 1997.
4. N. Miyake et al., "Optimizing early cancer diagnosis and detection using a temporal subtraction technique", Technological Forecasting and Social Change, Vol. 167, 2021.
5. T. Aoki., "Clinical usefulness of temporal subtraction system for lung nodule detection", Medical Imaging and Information Science, Vol. 30, No. 2, pp. 36-38, 2013.
6. Y. Tanabe and T. Ishida., "Development of a novel detection method for changes in lung conditions during radiotherapy using a temporal subtraction technique", physical and Engineering Sciences in Medicine, Vol. 44, pp. 1341-1350, 2021.
7. A. A. Ardakani et al., "Interpretation of radiomics features – A pictorial review", Computer Methods and programs in Biomedicine, Vol. 215, 2022.
8. J. J. M. van Griethuysen et al., "Computational radiomics system to decode the radiographic phenotype", Cancer Research, Vol. 77, pp. 104-107, 2017.
9. H. Zou and T. Hastie, "Regularization and variable selection via the elastic net" Statistical Methodology, Vol. 67, pp. 301-320、2005.
10. H. Zou and H. H. Zhang, "On the adaptive elastic-net with a diverging number of parameters", The Annals of Statistics, Vol. 37, No. 4, pp1733-1751, 2009.
11. T. Akiba et al., "Optuna: A next-generation hyperparameter optimization framework", ACM SIGKDD International Conference, pp. 2623-2631, 2019.
12. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system", Machine Learning, Vol. 3, 2016, pp. 785-794.
13. X. Guan et al., "Construction of the XGBoost model for early lung cancer prediction based on metabolic indices", BMC Medical Informatics and Decision Making, Vol. 23, 2023.
14. G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree", Advances in Neural Information Processing Systems, Vol. 30, pp. 3149-3157, 2017.
15. E. A-R. Hamed et al., "An efficient combination of convolutional neural network and LightGBM algorithm for lung cancer histopathology classification", Diagnostics, Vol. 13, No. 15, 2023.
16. L. Prokhorenkova et al., "CatBoost: unbiased boosting with categorical features", Neural Information Processing Systems, pp. 6639-6649, 2018.
17. J. T. Hancock and T. M. Khoshgoftaar, "CatBoost for big data: an interdisciplinary review", Journal of Big Data, Vol. 7, No. 94, 2020.
18. J. M. Ahn et al., "Ensemble machine learning of gradient boosting (XGBoost, LightGBM, CatBoost) and attention-based CNN-LSTM for harmful algal blooms forecasting", Toxins, Vol 15, No. 10, pp.608-623, 2023.
19. Y. Zhang et al., "A review of ensemble learning algorithms used in remote sensing applications", Appled Sciences, Vol. 12, No. 17, pp. 8654-8674, 2022.
20. N. Baba et al., "A detection method for nodular shadows from temporal subtraction images using a machine learning technique incorporated radiomics features", International Conference on Control, Automation and Systems, 2023, pp. 871-881.

Authors Introduction

Ms. Natsuho Baba

She received her Bachelor's degree in Engineering in 2023 from the Faculty of Engineering, Kyushu Institute of technology in Japan. She 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, medical application of image analysis, etc.

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.