2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Improved Genotype-Guided Deep Radiomics Signatures for Recurrence Prediction of Non-Small Cell Lung Cancer

Panyanat Aonpong, Yutaro Iwamoto, Xian-Hua Han, Lanfen Lin*, and Yen-Wei Chen*, *Member, IEEE*

Abstract— **Non-small cell lung cancer (NSCLC) is a type of lung cancer that has a high recurrence rate after surgery. Precise prediction of preoperative prognosis for NSCLC recurrence tends to contribute to the suitable preparation for treatment. Currently, many studied have been conducted to predict the recurrence of NSCLC based on Computed Tomography-images (CT images) or genetic data. The CT image is not expensive but inaccurate. The gene data is more expensive but has high accuracy. In this study, we proposed a genotype-guided radiomics method called GGR and GGR_Fusion to make a higher accuracy prediction model with requires only CT images. The GGR is a two-step method which is consists of two models: the gene estimation model using deep learning and the recurrence prediction model using estimated genes. We further propose an improved performance model based on the GGR model called GGR_Fusion to improve the accuracy. The GGR_Fusion uses the extracted features from the gene estimation model to enhance the recurrence prediction model. The experiments showed that the prediction performance can be improved significantly from 78.61% accuracy, AUC=0.66 (existing radiomics method), 79.09% accuracy, AUC=0.68 (deep learning method) to 83.28% accuracy, AUC=0.77 by the proposed GGR and 84.39% accuracy, AUC=0.79 by the proposed GGR_Fusion.**

Clinical Relevance—**This study improved the preoperative recurrence of NSCLC prediction accuracy from 78.61% by the conventional method to 84.39% by our proposed method using only the CT image.**

I. Introduction

Non-small cell lung cancer (NSCLC) is a type of epithelial lung cancer that is different from small cell lung cancer (SCLC) [1]. Of all lung cancer patients, more than 80 percent had NSCLC [1]. The doctors prefer to treat patients with surgery, despite there being several rounds of chemotherapy before the surgery. The patient may still be at high risk of tumor recurrence after the surgery [2]. The preoperative recurrence prediction can help the doctor to be prepared [2].

Recently, several machine learning methods have been proposed for the preoperative prediction of NSCLC recurrence based on genetic and radiomics information (computed tomography, CT image) [3-12]. Lambin et al. proposed the radiomics method that used the medical image to extract many quantitative image features and used them for the computer-aided diagnosis proposed [3]. The radiomics methods use only radiomics information (CT images) as their input. In [10], Wang et al. used the radiomics method to predict the recurrence of NSCLC using the principal component analysis (PCA) technique as the feature selection and classified by the various machine learning methods. In [11], Lee et al. used relief-F as feature selection and classified by the various machine learning methods. In [12], Christie et al. used LASSO as the feature selection and classification. However, the radiomics-based methods tend to have lower prediction accuracy. In [7], Aerts et al. reported the association of the handcrafted features and underlying gene expression association. Many recent studies attempted to use genetic-based information instead of using only the image for the diagnosis to increase the prediction accuracy [8][9]. For the genetic-based methods, they have their high costs as the limitation because the genetic examination is a high complexity test [8] and they are invasive diagnostic methods. In our previous work, we proposed a genotype-guided radiomics method (henceforth GGR) [13] to improve the prediction accuracy of the existing methods which use only the CT image [3-12]. The GGR consists of two models. The first model is gene estimation. This model will estimate the gene expression information from the CT images using deep learning-based. This model will work for each gene individually and do repeatedly for all related genes to avoid memory limitation. The second model is recurrence prediction. This model will use the estimated genes expression from the output of the first model to predict the recurrence. In the training phase, the GGR needs both CT image and gene expression data to create the mapping function. In the testing phase, the model needs only the CT image as input. The model uses the mapping function from the training phase to estimate the genes and use the estimated genes to predict the recurrence of NSCLC. In this paper, we further propose a deep genotype-guided radiomics fusion (GGR_Fusion) model, which is the improved version of the GGR by fusing the extracted radiomics features with the estimated genes data instead of using only estimated genes in the second model to improve the accuracy of the GGR.

• Panyanat Aonpong is with the College of Information Science and Engineering, Ritsumeikan University, Shiga, Japan. E-mail: gr0399rh@ritsumei.ac.jp

• Yutaro Iwamoto is with the College of Information Science and Engineering, Ritsumeikan University, Shiga, Japan. E-mail: yiwamoto@fc.ritsumei.ac.jp

• Xian-Hua Han is with Artificial Intelligence Research Center, Yamaguchi University, Japan. E-mail: hanxhua@yamaguchi-u.ac.jp

• Lanfen Lin is with the College of Computer Science and Technology, Zhejiang University, Hangzhou, China. Email: llf@zju.edu.cn

• Yen-Wei Chen is with the College of Information Science and Engineering, Ritsumeikan University, Shiga, Japan, Zhejiang Lab, Hangzhou, China and College of Computer Science and Technology, Zhejiang University, Hangzhou, China. E-mail: chen@is.ritsumei.ac.jp

• *Corresponding Authors: Yen-Wei Chen (chen@is.ritsumei.ac.jp), Lanfen Lin (llf@zju.edu.cn)

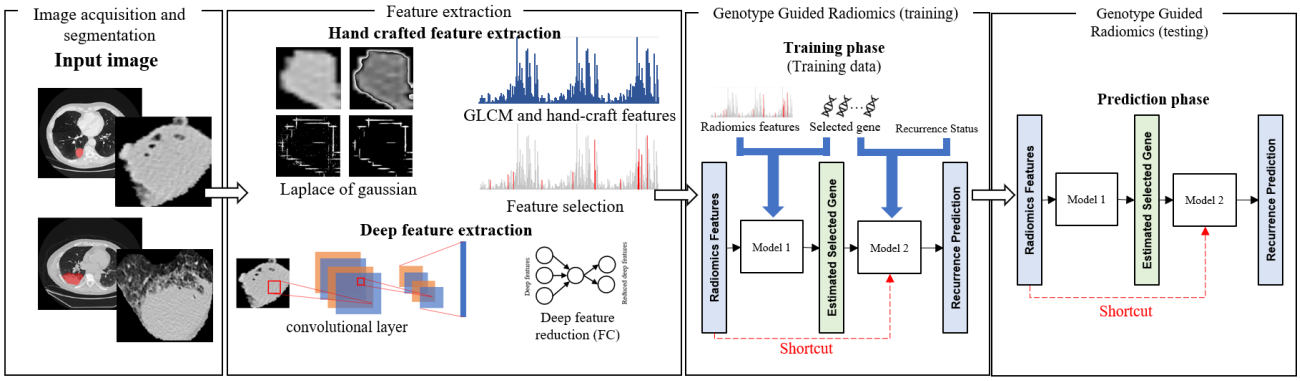

Figure 1. Overview of deep genotype-guided radiomics (GGR; without shortcut) and deep genotype-guided radiomics fusion (GGR_Fusion; with shortcut) for recurrence prediction of NSCLC.

II. Materials and Methods

The proposed GGR method consists of three parts: (1) image preprocessing, (2) feature extraction and selection (input features), and (3) prediction model. Unlike the existing traditional radiomics-based method, which uses a single model to predict the recurrence from the CT images, the GGR method consists of two models. The first model was created for gene estimation from CT images using handcrafted features and deep learning-based features, this will estimate a single gene for each time it runs, and repeatedly work for all related genes, and the second model is used to predict the recurrence using the estimated genes from the first model output [13].

The GGR_Fusion is developed from the GGR structure. It also consists of three parts: (1) image preprocessing, (2) feature extraction and selection (input features), and (3) prediction model. The main contribution of this technique is in the third part. We maximize the performance of the GGR by adding the shortcut of the data flow to allow the feature directly fed into the recurrence prediction model. In other words, we fuse the extracted features with the estimated genes from the gene estimation model. The workflow of GGR and GGR_Fusion is presented in Figure 1.

To make the experiment, all methods were trained using the NSCLC public radiogenomic dataset [14], which includes both CT images and corresponding gene data. The important idea of the proposed methods (GGR [13] and GGR_Fusion) is that we only require CT images to predict recurrence, while the models are training by pairs of gene data and CT images in the training phase to create the individual mapping function between each related gene and CT image using the deep learning model.

A. Dataset

In this research, we verify on a radiogenomics dataset of NSCLC [10], which now has open public access in The Cancer Imaging Archive (TCIA) [15]. This dataset was collected from 211 subjects. The data of every subject included CT image data and gene data [16][17].

B. Image Feature Extraction and Selection

Before the feature extraction procedure, we select the slice with the largest tumor mask area and select the adjacent above and below slices (three slices were selected in total). We cut the intensity information outside the range from -1000 HU to +400 HU (Hounsfield Unit), which covered the information that we want from the CT image [17]. We then normalized the value in the entire three slices to the value with a range from 0 to 255 using a linear transformation. In each CT image, the segmentation data attached to the CT image proceeds to multiply. We then cropped the area outside the bounding box around the masked image and resized the cropped image to 224 × 224 pixels. Next, the feature extraction must be performed.

In the handcrafted feature extraction, we extract the features based on gray level co-occurrence matrix (GLCM) in 4 directions, 0, 45, 90, and 135 degrees, and histogram-based statistics [18]. In the calculation of each part, Laplace of Gaussian (LoG) has been performed with five different σ values. Finally, 450 handcrafted features were extracted. Then select the related features using F-test [19], 12 handcrafted features are taken into account.

In the deep feature extraction, we applied the pre-trained ResNet50 structure [20] without a fully connected layer using the ImageNet dataset [21] to the feature extraction part. The NSCLC recurrence-related output features extracted from the ResNet50 will be selected using the F-test method as the feature selection. Finally, we obtain the related deep features which ready to be the inputs of both GGR and GGR_Fusion.

C. Gene Selection

We apply the feature selection methods to select the relevant genes. The methods used to select the corresponding genes in this study are including non-selected, LASSO [18], F-test (ANOVA) [19], CHI-2 [22], and the intersection of LASSO, F-test, and CHI-2. For F-test and CHI-2, we set the threshold at P-value < 0.02. We made a comparison of all 5 methods, including the nonselected method, and use the best efficient method as the gene selection. Finally, the intersection of the three methods has been chosen. By this method, we selected 74 related genes. The detailed results will be reported in Sec. 3.

D. Gene Estimation

In both GGR and GGR_Fusion, the gene estimation models are similar. They both use the deep neural network (DNN) regression model to estimate genes using the extracted deep feature as shown in Figure 2. A single DNN regression model will be declared to create a mapping function between a CT image of a patient and a gene. To estimate all the

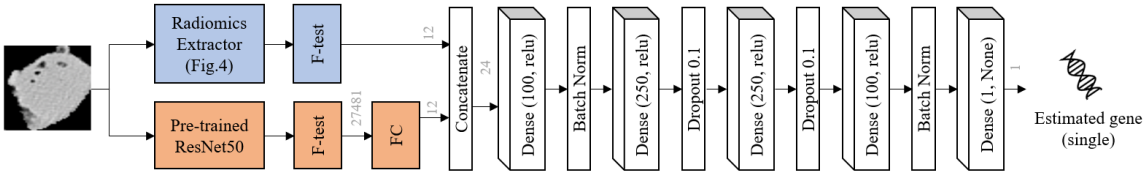

Figure 2. The DNN regression model's structure used to estimate a single gene in the GGR_Fusion. The grey numbers show the feature's number that passes the process.

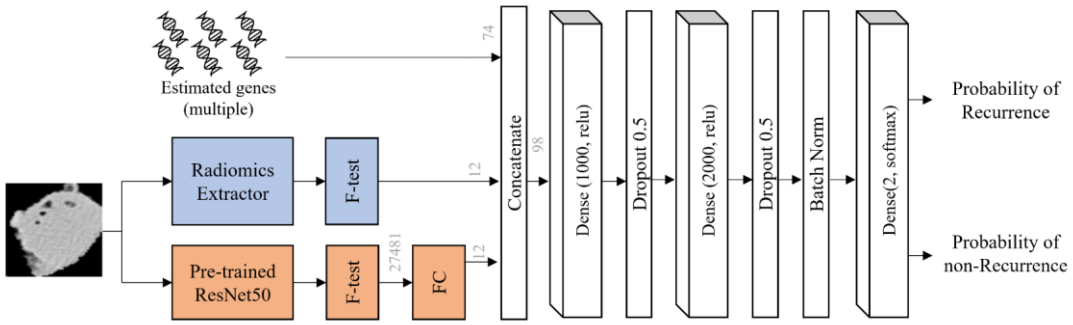

Figure 3. The model's structures used to predict the recurrence using estimated genes in GGR_Fusion model. The grey numbers show the feature's number that passes the process.

relevant genes, multiple DNN regression models must be created. The learning rate was set to 5e-6 to fit all gene estimation models. The mean square error loss was used as their loss function. As we described in Sec. 2.3, we select 74 genes to reduce the computation cost and enhance the prediction accuracy, the 74 same structure regression models are declared for all 74 genes with different weights.

E. The Recurrence Prediction

The recurrence prediction model is a two-class classification model (i.e., recurrence or non-recurrence). This proposed second model (model 2) is declared to predict the recurrence of NSCLC. The GGR_Fusion is trying to improve the performance of the GGR bypassing the features directly to model 2. We predict recurrence using a combination of the estimated gene data, handcrafted features, and the deep learning features as input. The learning rate was set to 0.05. The binary cross entropy loss was used as its loss function. The details of the GGR_Fusion's model 2 showed in Figure 3.

III. EXPERIMENTS

Every experiment in this study is performed based on 10-fold cross-validation to find the average accuracy. The area under the receiver operating characteristic (AUC) [23] is also provided to assess the efficiency. The Keras-GPU library version 2.2.4 on Python 3.6 was used to perform these experiment's actions. The related work models are compared with the GGR and the GGR_Fusion methods using the same control dataset.

A. Gene Selection Results

The results of gene selection performance are presented in Table 1. Since the method of intersection between the three methods and the F-test achieves the best and second performance ranking, respectively, we used 74 genes selected by the intersection of the three methods for our study.

TABLE I. PERFORMANCE OF THE RADIOMICS METHOD USING GENE DATA.

Feature selection method	Selected genes	Accuracy
Nonselected	5587	0.8141
LASSO	1123	0.8297
F-test	131	0.8689
CHI-2	2325	0.8339
Intersection of the three	**74**	**0.8806**

B. Recurrence Prediction Results

We show the summary results in terms of AUC in Table 2 and the area of the receiver operating characteristic curve (ROC) in Figure 4. The proposed methods are marked by bold texts. As shown in Table II, the proposed GGR and GGR_Fusion outperform the conventional radiomics-based method [3], and deep learning-based methods [6].

TABLE II. THE AREA UNDER THE RECEIVER OPERATING CHARACTERISTIC CURVE (AUC) OF THE GGR_FUSION COMPARED TO THE CONVENTIONAL METHODS.

Methods	ACC	AUC
CT-radiomics: PCA+DT [10]	61.36%	0.57
CT-radiomics: PCA+RF [10]	68.18%	0.62
CT-radiomics: PCA+SVM [10]	67.05%	0.58
CT-radiomics: Relief-F+SVM [11]	68.18%	0.56
CT-radiomics: Relief-F+RF [11]	67.04%	0.57
CT-radiomics: LASSO [12]	61.36%	0.68
CT-radiomics: F-test + ANN [5]	78.61%	0.66
ResNet50 [5]	79.09%	0.67
DenseNet121 [5]	77.36%	0.69
GGR [13]	83.28%	0.77
GGR_FUSION (proposed)	**84.39%**	**0.79**

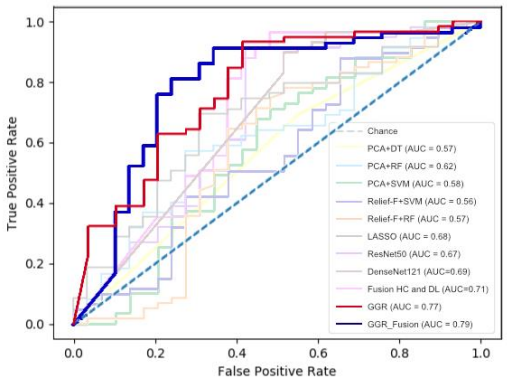

Figure 4. Average ROC of recurrence prediction using GGR_Fusion compared to other state-of-the-art methods.

IV. Discussion and Conclusion

In this study, we focused on NSCLC recurrence prediction using the CT image. As traditional radiological methods or newly proposed methods using only the CT image, the predictive performance is limited. From the experiments, we achieved AUC = 0.66 with conventional radiomics, AUC = 0.68 for deep learning-based method (ResNet50). In this study, the GGR is proposed to enhance the accuracy from the conventional radiomics, and the deep learning-based by using deep learning to make the genotype guidance instead of directly predict the recurrence. We further proposed the GGR_Fusion that uses the mapping function between CT image and genes to enhance the accuracy. These two proposed methods only use gene information in the training phase but are not required in the testing phase because they can estimate by the mapped function themselves. Different from the GGR, the GGR_Fusion directly passing the extracted features from the CT image to the recurrence prediction model to maximize the accuracy and AUC from the GGR. This proposed method can improve the prediction accuracy from 78.61% (AUC=0.66) by the traditional CT radiomics method and 79.09% (AUC = 0.68) to 83.28% (AUC = 0.77) by GGR and 84.39% (AUC = 0.79) by our proposed method. Although the GGR and GGR_Fusion are shown significantly better prediction performance compared to the conventional radiomics-based and the deep learning-based methods, our new proposed cannot reach the accuracy of the genomic-based method. We performed experiments with real gene expression for the recurrence prediction and achieved AUC = 0.92 for gene-expression analysis and AUC = 0.93 for the combination between gene expression and CT image, which are higher than the proposed methods.

In future efforts, we are going to improve the model performance and make them closer to the genomics-based methods but use only CT images. In the computational term, all GGR-based methods require large amounts of calculations because we must perform the individual estimation of 74 genes for a single patient.

Acknowledgment

This work was supported in part by the Grant-in Aid for Scientific Research from the Japanese Ministry for Education, Science, Culture and Sports (MEXT) under the Grant No. 20KK0234, No. 21H03470 and No. 20K21821.

References

[1] M. A. Bareschino, et al., "Treatment of advanced non-small cell lung cancer," Journal of thoracic disease, vol. 3, pp. 122, 2011.
[2] H. Uramoto, and T. Fumihiro, "Recurrence after surgery in patients with NSCLC," Translational lung cancer research, vol. 3, pp. 242, 2014.
[3] P. Lambin, et al., "Radiomics: extracting more information from medical images using advanced feature analysis," *European journal of cancer*, vol. 48, no. 4, pp. 441-446, 2012.
[4] M. Avanzo, et al., "Machine and deep learning methods for radiomics," Medical physics, vol. 47, pp. e.185-e202, 2020.
[5] P. Aonpong, et al., "Hand-Crafted and Deep Learning-Based Radiomics Models for Recurrence Prediction of Non-Small Cells Lung Cancers," in Proc. Innovation in Medicine and Healthcare. Springer, Singapore, 2020, pp. 135-144.
[6] P. Starkov, et al., "The use of texture-based radiomics CT analysis to predict outcomes in early-stage non-small cell lung cancer treated with stereotactic ablative radiotherapy," The British journal of radiology, vol. 92, 2019
[7] H. JWL. Aerts, et al., "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach," *Nature communications*, vol. 5, no. 1 pp. 1-9, 2014
[8] Y. Lu, et al., "MicroRNA profiling and prediction of recurrence/relapse-free survival in stage I lung cancer," Carcinogenesis, vol. 33, pp. 1046-1054, 2012.
[9] P. Marentakis, et al., "Lung cancer histology classification from CT images based on radiomics and deep learning models," *Medical & Biological Engineering & Computing*, vol. 59, pp. 215-226, 2021.
[10] X. Wang, D. Hui-hong, and N. Sheng-dong, "Prognostic recurrence analysis method for non-small cell lung cancer based on CT imaging," *2019 International Conference on Image and Video Processing, and Artificial Intelligence*, vol. 11321. International Society for Optics and Photonics, 2019.
[11] S. Lee, et al., "Radiomic feature-based prediction model of lung cancer recurrence in NSCLC patients," *International Workshop on Advanced Imaging Technology (IWAIT) 2020*, vol. 11515, International Society for Optics and Photonics, 2020.
[12] J. R. Christie, et al. "A multi-modality radiomics-based model for predicting recurrence in non-small cell lung cancer." Medical Imaging 2021: Biomedical Applications in Molecular, Structural, and Functional Imaging. Vol. 11600. International Society for Optics and Photonics, 2021.
[13] P. Aonpong, et. al., "Genotype-Guided Radiomics Signature for Recurrence Prediction of Non-Small Cell Lung Cancer," IEEE Access, Vol.9, pp.90244 - 90254, 2021
[14] S. Bakr, O. Gevaert, S. Echegaray, K. Ayers, M. Zhou, M. Shafiq, H. Zheng, W. Zhang, A. Leung, M. Kadoch, J. Shrager, A.Quon, D. Rubin, S. Plevritis, S. Napel, "Data for NSCLC Radiogenomics Collection," The Cancer Imaging Archive, 2017, http://doi.org/10.7937/K9/TCIA.2017.7hs46erv
[15] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, L. Tarbox, F. Prior, "The Cancer Imaging Archive (TCIA), Maintaining and Operating a Public Information Repository," Journal of Digital Imaging, vol. 26, pp. 1045-1057, December 2013.
[16] S. Bakr, et al., "A radiogenomic dataset of non-small cell lung cancer," Scientific data, vol. 5, pp. 1-9, 2018.
[17] B. A. Simon, et al., "Computed tomography studies of lung mechanics," In Proc. the American Thoracic Society, vol. 2, pp. 517-521, 2005.
[18] Y. Zhou, et al., "CT-based radiomics signature: a potential biomarker for preoperative prediction of early recurrence in hepatocellular carcinoma." Abdominal radiology, vol. 42, pp. 1695-1704, 2017.
[19] H. Scheffe, "The analysis of variance," John Wiley & Sons, vol. 72, 1999.
[20] K. He, et al., "Deep residual learning for image recognition," in Proc. the IEEE conference on computer vision and pattern recognition. 2016.
[21] R. Geirhos, et al., "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness," arXiv preprint arXiv:1811.12231, 2018.
[22] H. O. Lancaster, "The chi-squared distribution," 1969.
[23] T. Fawcett, "An introduction to ROC analysis," Pattern recognition letters, vol. 27, pp.861-874, 2006.