





Research Article

Machine Learning-Based Radiomics for Prediction of Epidermal Growth Factor Receptor Mutations in Lung Adenocarcinoma

Jiameng Lu ¹, Xiaoqing Ji,² Lixia Wang,³ Yunxiu Jiang,⁴ Xinyi Liu,⁴ Zhenshen Ma,⁵ Yafei Ning ¹, Jie Dong,⁵ Haiying Peng,⁴ Fei Sun,⁴ Zihan Guo,⁴ Yanbo Ji,² Jianping Xing ¹, Yue Lu,⁶ and Degan Lu ⁷

¹School of Microelectronics, Shandong University, Jinan 250100, China

²Department of Nursing, The First Affiliated Hospital of Shandong First Medical University & Shandong Provincial Qianfoshan Hospital, Jinan 250014, China

³Division of Disinfecting and Supply, Liaocheng People's Hospital, Liaocheng 252000, China

⁴Graduate School of Shandong First Medical University, Jinan 250000, China

⁵Department of Radiology, The First Affiliated Hospital of Shandong First Medical University & Shandong Provincial Qianfoshan Hospital, Shandong Medicine and Health Key Laboratory of Abdominal Medicine Imaging, Shandong Lung Cancer Institute, Shandong Institute of Neuroimmunology, Jinan 250000, China

⁶Department of Interventional Medicine, The Second Hospital, Cheeloo College of Medicine, Shandong University; Interventional Oncology, Institute of Shandong University, Jinan 250033, China

⁷Department of Respiratory, The First Affiliated Hospital of Shandong First Medical University & Shandong Provincial Qianfoshan Hospital, Shandong Institute of Respiratory Diseases, Shandong Institute of Anesthesia and Respiratory Critical Medicine, Jinan 250000, China

Correspondence should be addressed to Yafei Ning; ningyafei@sdu.edu.cn, Jianping Xing; xingjp@sdu.edu.cn, and Degan Lu; deganlu@126.com

Received 26 February 2022; Revised 13 April 2022; Accepted 23 April 2022; Published 7 May 2022

Academic Editor: Yan Yang

Copyright © 2022 Jiameng Lu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Identifying an epidermal growth factor receptor (EGFR) mutation is important because EGFR tyrosine kinase inhibitors are the first-line treatment of choice for patients with EGFR mutation-positive lung adenocarcinomas (LUAC). This study is aimed at developing and validating a radiomics-based machine learning (ML) approach to identify EGFR mutations in patients with LUAC. We retrospectively collected data from 201 patients with positive EGFR mutation LUAC (140 in the training cohort and 61 in the validation cohort). We extracted 1316 radiomics features from preprocessed CT images and selected 14 radiomics features and 1 clinical feature which were most relevant to mutations through filter method. Subsequently, we built models using 7 ML approaches and established the receiver operating characteristic (ROC) curve to assess the discriminating performance of these models. In terms of predicting EGFR mutation, the model derived from radiomics features and combined models (radiomics features and relevant clinical factors) had an AUC of 0.79 (95% confidence interval (CI): 0.77-0.82), 0.86 (0.87-0.88), respectively. Our study offers a radiomics-based ML model using filter methods to detect the EGFR mutation in patients with LUAC. This convenient and low-cost method may be of help to noninvasively identify patients before obtaining tumor sample for molecule testing.

1. Introduction

Lung cancer was the second most commonly diagnosed cancer and remained the leading cause of cancer-related death worldwide [1]. The most common histological subtype of lung cancer is lung adenocarcinoma (LUAC), accounting for approximately 40% of all cases [2]. Although tremendous progress has been made in the treatment of LUAC in the last decade, the prognosis of patients who are detected at advanced clinical stage remains unfavorable. Epidermal growth factor receptor (EGFR) is one of the most frequently mutated genes in LUAC [3], and EGFR tyrosine kinase inhibitors (TKI) have provided patients who harbor activating EGFR mutations with clinical benefit, such as high response rate and prolonged progression-free survival (PFS) [4]. Therefore, an EGFR-TKI has become the first-line treatment of choice for patients with positive EGFR mutation LUAC [5]. As a result, the detection of EGFR mutations is of great significance in determining treatment for patients with LUAC [6].

Detection of EGFR mutational profile is currently based on cytology and noncytology biopsy samples, and mutational sequencing has become the gold standard of EGFR mutation detection [7]. However, tissue sampling has some disadvantages. First, the tumor tissue is not easy to obtain in several cases. Second, the biopsied sample does not necessarily represent the tumor tissue due to intratumor heterogeneity [8]. Third, biopsy testing may potentially increase the risk of cancer metastasis, although the chance is small [9]. Finally, long turnaround time, unfeasibly repeated biopsy, and the relative high costs also account for the limited use of mutational sequencing [10]. Thus, it is a critical need to explore a noninvasive and convenient method to predict EGFR mutation status.

Radiomics is a rapidly evolving and important field because it can extract and analyze multiple features derived from digital medical images with the aim of enhancing clinical decision-making [11, 12]. Studies have revealed that somatic mutations, which ultimately lead to tumor phenotype, can be predicted by radiomics in different solid tumors, including lung cancer [10, 13]. Based on imaging information extracted from magnetic resonance imaging (MRI), computed tomography (CT), and positron-emission-tomography (PET), radiomics analysis can be performed to identify the presence of EGFR, anaplastic lymphoma kinase (ALK), Kirsten rat sarcoma viral oncogene (KRAS), and Erb-B2 receptor tyrosine kinase 2 (ERBB2) mutations in patients with non-small-cell lung cancer (NSCLC) [14–18]. With specific regard to EGFR mutation, previous studies have documented the potential for radiomics to predict EGFR 19Del and L858R based on the phenotypic appearance [14, 16, 19]. For example, Rossi et al. built a machine learning (ML) model to identify EGFR mutant and achieved an area under the receiver operating characteristic curve (AUC) of 0.89 [19]. By developing deep learning models, Zhang et al. reported that radiomics features from CT images can discriminate EGFR mutation with an AUC of 0.910 and 0.841 for the internal and external test cohorts, respectively [20]. Hong and colleagues [21] utilized features from enhanced CT imaging to recognize EGFR mutation status in advanced LUAC. They reported an AUC of 0.851 for predicting EGFR mutation with a model based on

radiomics features and clinical data [21]. Although previous studies have documented the association between radiological characteristics and EGFR mutation status, the role of CT-based radiomics ML in identifying EGFR mutation in LUAC remains to be further explored.

Selection of a subset of relevant predictor variables from highly dimensional data, which is termed as feature selection (FS), is a critical step in analysis of radiomics features [22]. FS is the core of classification which plays a fundamental role in ML and can reduce the learning complexity. As one of the FS methods, filter methods assess the goodness of features based on a simple weight score criterion [23]. In addition, filter methods select features independent of any specific classifiers and demand less computation [23]. As a result, filter models have been widely studied because of their efficiency and simplicity. However, few studies on prediction of EGFR mutation status were reported using filter approaches based on ML.

Therefore, the aim of this study is to develop a radiomics-based model to predict EGFR mutation status in patients with LUAC using filter methods. In the present study, CT-based radiomics features and ML methods were used to identify EGFR mutation status and the effect of this model on predicting EGFR mutation in LUAC was assessed. The outcome of this study may aid in distinguishing patients with EGFR mutations from those without and helping clinicians to make treatment decisions for patients.

2. Materials and Methods

2.1. Patients. The study population was retrospectively selected from patients diagnosed with LUAC from the First Affiliated Hospital of Shandong First Medical University (Jinan, China). The institutional review board approved this study with a waiver for the informed consent requirement. Patients who were (1) histologically diagnosed with primary LUAC, (2) classified as stage III-IV according to the Eighth Edition of the Lung Cancer Stage Classification, (3) having detected EGFR mutations based on PCR technology, (4) treatment-naïve subjects, and (5) receiving chest CT scan prior to biopsies or surgery met the inclusion criteria and were included. The exclusion criteria were given as follows: (1) lack of clinical data, such as age, gender, stage, and serum tumor marker, and (2) difficulty in drawing regions of interest (ROIs). In the end, 201 patients were included in this study. The flow chart of participant recruitment is shown in Figure 1. The enrolled patients were randomly classified into the training cohort and independent validation cohort with the ratio of 7 ($n = 140$):3 ($n = 61$). The workflow of the radiomics analysis is depicted in Figure 2.

2.2. Analysis of EGFR Mutation. Based on the tumor specimen, EGFR gene mutations in exons 18, 19, 20, and 21 were examined by an amplification refractory mutation system real-time technology using Human EGFR Gene Mutations Fluorescence Polymerase Chain Reaction (PCR) Diagnostic Kit (Amoy Diagnostics Co., Ltd, Xiamen, China). Wild-type EGFR in the present study referred to absence of mutation on those loci.

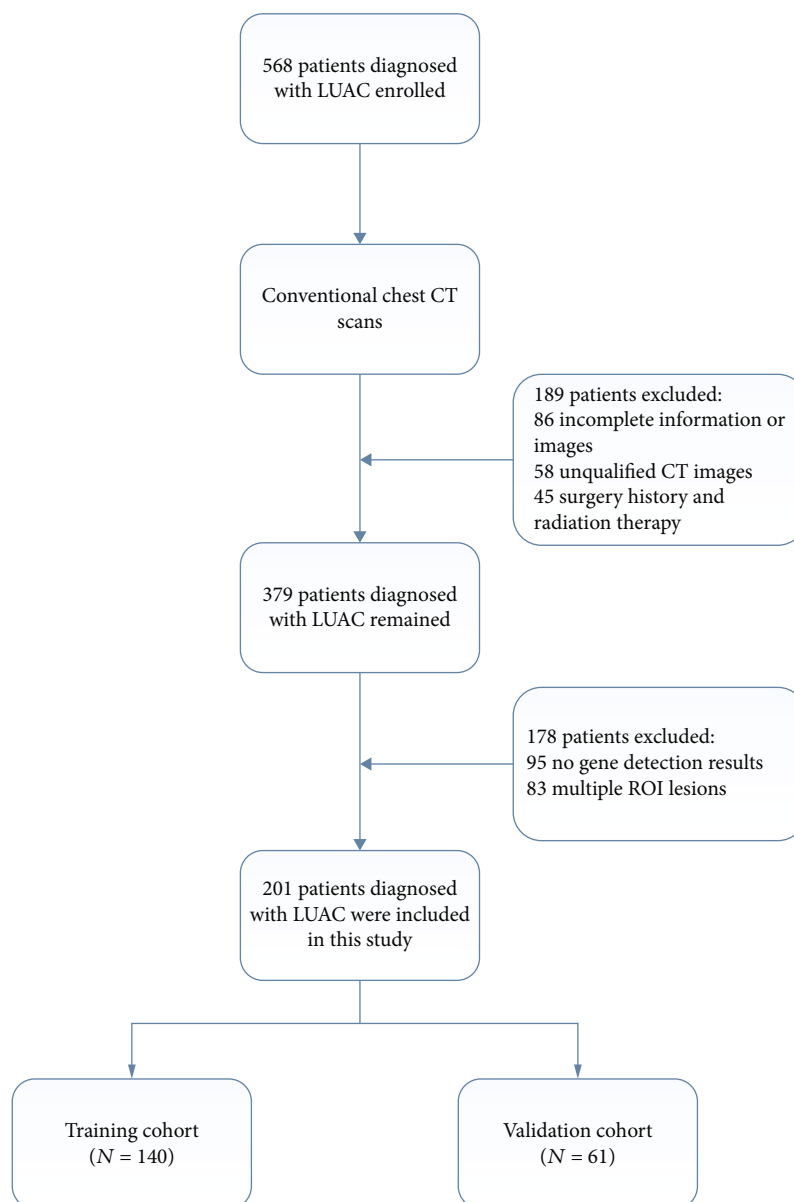


FIGURE 1: Patient recruitment workflow. In total, 201 of 568 patients were included in this study according to the selection criteria.

2.3. Image Acquisition. All patients included in this study underwent chest CT scans prior to any treatment using two CT scanners (GE Healthcare, Milwaukee, WI, USA; United Imaging, Shanghai, China). The scanning parameters were given as follows: the tube voltage, 120 kVp; tube current, 160–300 mA; detector collimation, 64 or 128 × 0.625 mm; field of view, 350 × 350 mm; the pitch, 0.992:1; and matrix of 512 × 512. All images were reconstructed with a section thickness of 2 mm and were stored in DICOM format in the Picture Archiving and Communication Systems (PACS) of our hospital.

2.4. Image Preprocessing. Because different CT scans were used in this study, image preprocessing prior to segmentation and feature extraction was undergone to make the radiomics fea-

tures more robust [24]. As previously reported by Hong et al. [21], a resampling method and Gaussian filter were used in this process.

2.5. Tumor Segmentation. Every lesion was independently evaluated and segmented manually slice by slice by two senior radiologists (both with more than 10-year experience of CT interpretation). The ROI was delineated in ITK-SNAP (version 3.6, <http://www.itksnap.org>) and confirmed by another chest radiologist with 15-year experience [25, 26]. If one patient has multiple lesions, the radiologist only delineates the tumor area where the biopsy was performed. All radiologists were blinded to the status of EGFR mutation.

To reduce the differences in manual segmentation between two radiologists, the intragroup correlation coefficient (ICC)

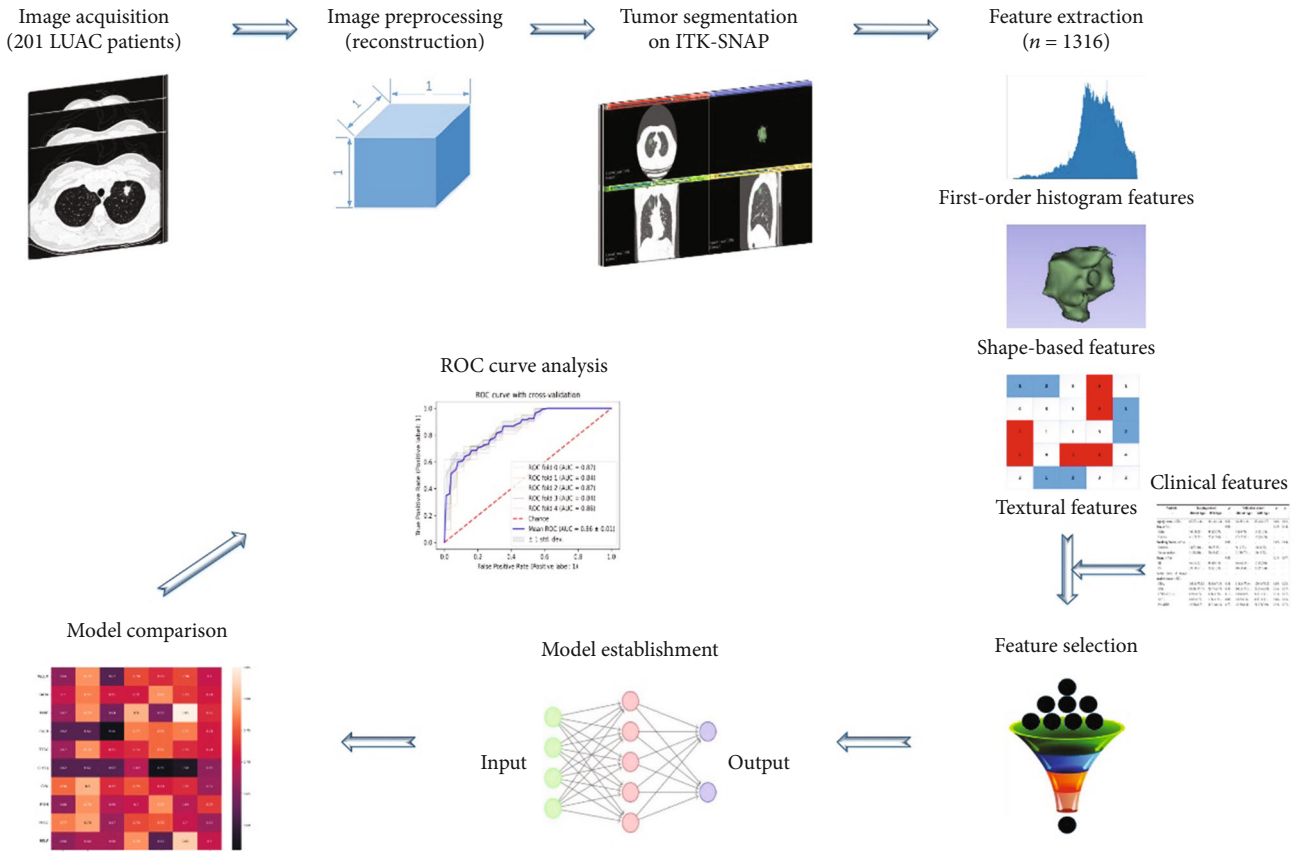


FIGURE 2: Workflow of the radiomics analysis.

for each feature was calculated [27, 28]. Only those with an ICC greater than 0.85 was considered highly stable and selected for the following analysis.

2.6. Feature Extraction. Based on the three-dimensional region of interest (3D ROI), radiomics features were extracted from each ROI using Pyradiomics package (<http://pyradiomics.readthedocs.io/en/latest/index.html>). A total of 1316 features were extracted, and these features can be divided into 3 categories: first-order statistics ($n = 18$ features), shape-based ($n = 14$ features), and textural feature [18]. The textural feature category includes Gray-Level Cooccurrence Matrix (GLCM) ($n = 24$ features), Gray-Level Run Length Matrix (GLRLM) ($n = 16$ features), Gray-Level Size Zone Matrix (GLSZM) ($n = 16$ features), Gray-Level Dependence Matrix (GLDM) ($n = 14$ features), and Neighboring Gray Tone Difference Matrix (NGTDM) ($n = 5$ features). In addition, two filters (including wavelet ($n = 744$ features) and Laplacian of Gaussian ($n = 465$ features) were also applied to the original CT images to obtain transformed images. By decomposing the image with wavelet transform, high- (H) or low- (L) pass filters in three dimensions were applied and 8 kinds of combinations were obtained: LHL, HHL, HLL, HHH, HLH, LHH, LLH, and LLL. To emphasize areas of gray-level change, the LoG filter was applied to the input image and yield a derived image for each sigma value specified [29]. In our study, five fil-

ters with different sigma values were applied ($\sigma = 1.0$ mm, 2.0 mm, 3.0 mm, 4.0 mm, and 5.0 mm). The specific number of features is listed in supplementary Table 1.

2.7. Feature Selection. At first, univariate analysis was performed for each feature and those with P values < 0.1 were considered to be associated with genetic mutations and selected [30]. Then, 10 FS techniques based on filter methods were used in the current analysis and they can be classified into two categories: univariate methods and multivariate methods [31]. The univariate methods included Fisher score (FSCR), Relief (RELF), t -test score (TTSC), chi-square (CHSQ), Wilcoxon rank sum (WLCX), Gini index (GINI), information gain (IFGN), F -ANOVA (FAOV), and Pearson correlation coefficient (PESC). The multivariate methods consisted of mutual information (MUIF). These approaches were chosen mainly due to their computational efficiency, simplicity in implementation, and applications in literature [32, 33]. Filter methods calculate a relevance score for each feature, and those which are lower than a given threshold will be removed [31].

FS methods, such as GINI, RELF, and IFGN, were performed using the “attrEval” function from the “CORElearn” package in R software package. FAOV, FSCR, TTSC, CHSQ, WLCX, PESC, and MUIF were implemented using the scikit-learn package in Python software (Python Software

Foundation: <http://www.python.org>). In order to describe various aspects of the EGFR mutation and avoid choosing features from a certain feature group, features were selected based on rankings in their own group rather than rankings among all features. With increased numbers of selected features, we found that the majority of classifiers showed the best predictive performance when the top 2 features are selected from each group. If no features passed the univariate test in a certain group, this group will be ignored.

2.8. Radiomics Model Establishment and Performance Evaluation. Seven ML algorithms were imported from the scikit-learn library in Python software to establish models [34]. These algorithms included decision tree (DT), AdaBoost classifier (AD), naïve Bayes (NB), random forest (RF), logistic regression (LR), support vector machines (SVM), extreme gradient boosting (XGBoost, XGB), and k nearest neighbors (KNN). In combination of 10 FS methods and 7 classifiers, we developed 70 ($10 \times 7 = 70$) models. The nomenclature of each model was established by two elements: the name of FS method and classifier. For example, NB-WLCX referred to a model combining naïve Bayes classifier with FS approach of Wilcoxon rank sum. The predictive ability of each algorithm was primarily assessed using AUC of receiver operating characteristic (ROC) curve analysis. Then, fivefold cross-validation was applied to examine all results and also evaluated by AUC. The model which gives the highest cross-validation accuracy was selected as the final model for further analysis.

2.9. Development and Validation of Models Combining Radiomics Features and Clinical Characteristics. To further increase the power of predicting EGFR mutation, some clinical characteristics were added to the aforementioned model. These clinical factors consisted of age, gender, smoking status, stage of disease, and serum level of tumor markers. The tumor markers included carcinoembryonic antigen (CEA), neuron-specific enolase (NSE), fragment of cytokeratin subunit 19 (CYFRA 21-1), squamous cell carcinoma antigen (SCC), and pro-gastrin-releasing peptide (Pro-GRP). The predictive performance of each algorithm was also evaluated based on the AUC of ROC curve analysis.

2.10. Statistical Analysis. Statistical analysis was performed using PRISM version 6 (GraphPad, La Jolla, CA, USA). Quantitative data were compared using Student's t -test, and categorical data were compared using the χ^2 test to identify baseline differences. The discrimination performance of models was evaluated by the ROC curve. All statistical tests were two-tailed, and $P < 0.05$ was considered statistically significant.

3. Result

3.1. Clinical Characteristics. The baseline clinical characteristics of the enrolled patients are listed in Table 1. No evident differences were found among the age, gender, stage of disease, and serum level of CEA, NSE, CYFRA 21-1, and Pro-GRP between the EGFR-mutated and EGFR wild-type group ($P > 0.05$). The smoking status was significantly different between the EGFR-mutated and EGFR wild-type group in

the training cohort ($P < 0.05$). The level of SCC in the serum was significantly different in the training and validation set ($P < 0.05$).

3.2. Selected Stable Features. In total, 1316 radiomics features were extracted. Subsequently, ICC for radiomics features in each group were calculated ($ICC = \text{mean} \pm \text{SD}$) and are depicted in supplementary Fig. 1: shape-based features ($ICC = 0.97 \pm 0.03$), first-order features ($ICC = 0.98 \pm 0.01$), GLCM features ($ICC = 0.98 \pm 0.02$), GLRLM features ($ICC = 0.99 \pm 0.01$), GLSZM features ($ICC = 0.98 \pm 0.02$), GLDM features ($ICC = 0.98 \pm 0.01$), NGTDM features ($ICC = 0.99 \pm 0.01$), wavelet transformed features ($ICC = 0.97 \pm 0.05$), and LoG-transformed features ($ICC = 0.95 \pm 0.06$). Overall, 1269 of the 1316 (96.4%) extracted radiomics features were identified as stability and were retained. These features consist of 14 shape-based features, 18 first-order features, 24 GLCM features, 16 GLRLM features, 16 GLSZM features, 14 GLDM features, 49 LoG features, 5 NGTDM features, 727 wavelet transformed features, and 435 LoG-transformed features. The histogram of the ICC values of the radiomics features is shown in supplementary Figure 1.

3.3. Model Performance Assessment. The mean AUC scores for each classifier across the different FS methods are presented in a heat map form (Figure 3). When analysis was based on radiomics features, the RF classifier performed better than the other classifiers and the median AUC of the 10 models using RF classifier was 0.74. With regard to FS approaches, MUIF provided the best predictive performance and the median AUC of the 7 models using MUIF FS method was 0.72. When various classifiers and FS methods are combined, RF-MUIF model provided the highest performance in the prediction of EGFR mutation and the AUC reached 0.79 (Figure 3(a)). Moreover, the RF-MUIF model achieved a sensitivity of 0.81, a specificity of 0.63, and an accuracy of 0.74 for predicting EGFR mutation status. Further, the XGBoost model outperformed other classifiers (median AUC 0.73) and MUIF generated better AUCs (median AUC 0.72) when the integrated model built with radiomics signature and clinical features was analyzed. The model of XGBoost-MUIF achieved the best predictive performance, and the AUC, sensitivity, specificity, and accuracy were 0.86, 0.95, 0.72, and 0.83, respectively (Figure 3(b)). The cross-validated AUC scores and AUC curve on the validation dataset are shown in (Figures 4(a)–4(d)).

3.4. Analysis of the Selected Radiomics and Clinical Features. Among the selected radiomics and clinical features, 10 features had lower values for EGFR mutant type than for EGFR wild type. These features included original_shape_Flatness (0.57 vs. 0.60, $P = 0.09$), original_firstorder_Kurtosis (5.47 vs. 8.12, $P = 0.004$), original_glrlm_GrayLevelNonUniformityNormalized (0.20 vs. 0.23, $P = 0.008$), original_ngtdm_Contrast (41.36 vs. 46.74, $P < 0.001$), wavelet-HLH_gldm_SmallDependence-HighGrayLevelEmphasis (0.38 vs. 0.40, $P < 0.001$) log-sigma-2-0-mm-3D_gldm_LargeDependenceEmphasis (247.74 vs.

TABLE 1: Characteristics of patients in training and validation cohorts.

Variable	Training cohort		P	Validation cohort		P	P
	Mutant type	Wild type		Mutant type	Wild type		
Age (y, mean ± SD)	65.27 ± 1.44	66.14 ± 1.24	0.66	64.35 ± 1.34	63.48 ± 1.57	0.68	0.83
Sex, n (%)			0.06			0.19	0.44
Male	34 (24.29)	40 (28.57)		9 (14.75)	19 (31.15)		
Female	41 (29.29)	25 (17.86)		17 (27.87)	16 (26.23)		
Smoking status, n (%)			0.01			0.05	0.64
Smoker	24 (32.00)	36 (55.38)		5 (19.23)	19 (54.29)		
Never smoker	51 (68.00)	29 (44.62)		21 (80.77)	16 (45.71)		
Stage, n (%)			0.43			0.15	0.07
III	46 (61.33)	44 (68.75)		16 (61.54)	15 (42.86)		
IV	29 (38.67)	20 (31.25)		10 (38.46)	20 (57.14)		
Serum level of tumor marker (mean ± SD)							
CEA	109.0 ± 75.82	30.86 ± 7.56	0.31	114.6 ± 77.44	129.3 ± 78.20	0.89	0.28
NSE	98.39 ± 75.75	28.75 ± 6.79	0.36	100.6 ± 77.31	29.05 ± 6.932	0.36	0.27
CYFRA 21-1	6.91 ± 0.79	9.36 ± 1.58	0.17	6.88 ± 0.82	9.63 ± 1.62	0.13	0.17
SCC	0.85 ± 0.77	1.26 ± 1.65	0.08	0.62 ± 0.38	0.97 ± 0.92	0.06	0.03
Pro-GRP	45.50 ± 8.23	49.31 ± 6.49	0.72	45.33 ± 8.40	51.17 ± 7.09	0.59	0.72

CEA: carcinoembryonic antigen; NSE: neuron-specific enolase; CYFRA 21-1: fragment of cytokeratin subunit 19; SCC: squamous cell carcinoma antigen; Pro-GRP: pro-gastrin-releasing peptide.

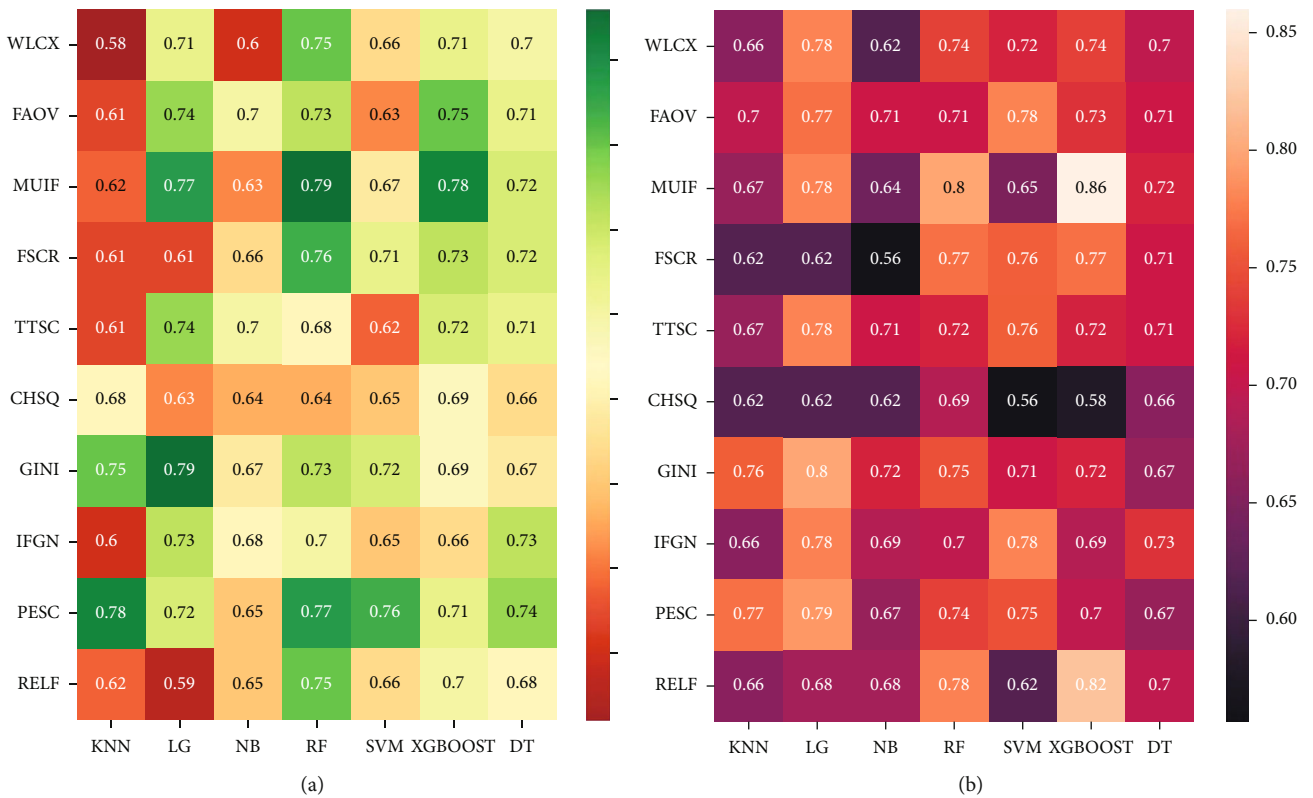


FIGURE 3: Heat maps with the AUC of different combinations of FS methods (rows) and classification algorithms (columns). (a) The average cross-validated AUC from 70 models based on radiomics features. (b) The average cross-validated AUC from 70 models based on radiomics features and clinical data.

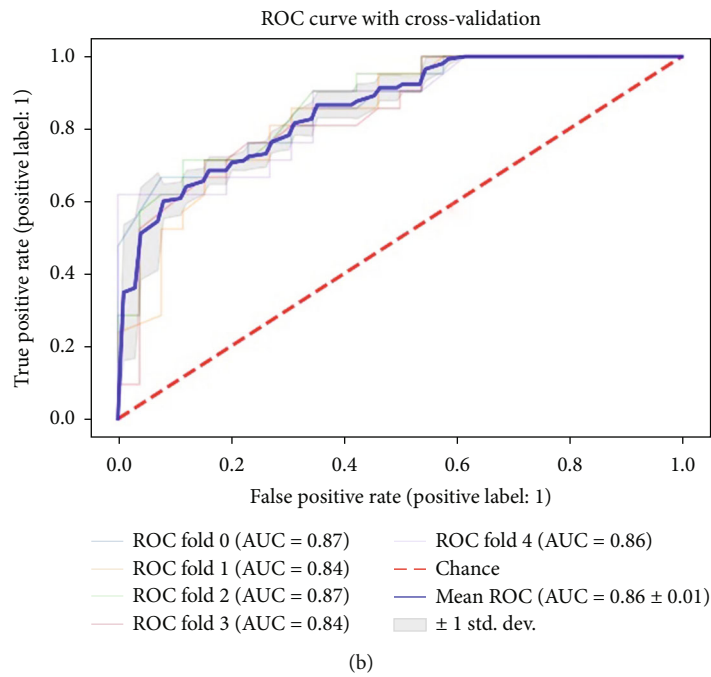
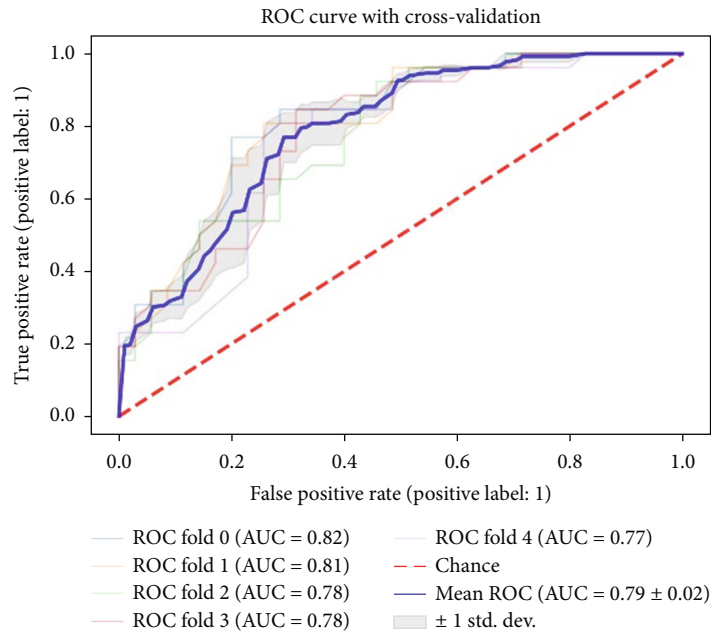


FIGURE 4: Continued.

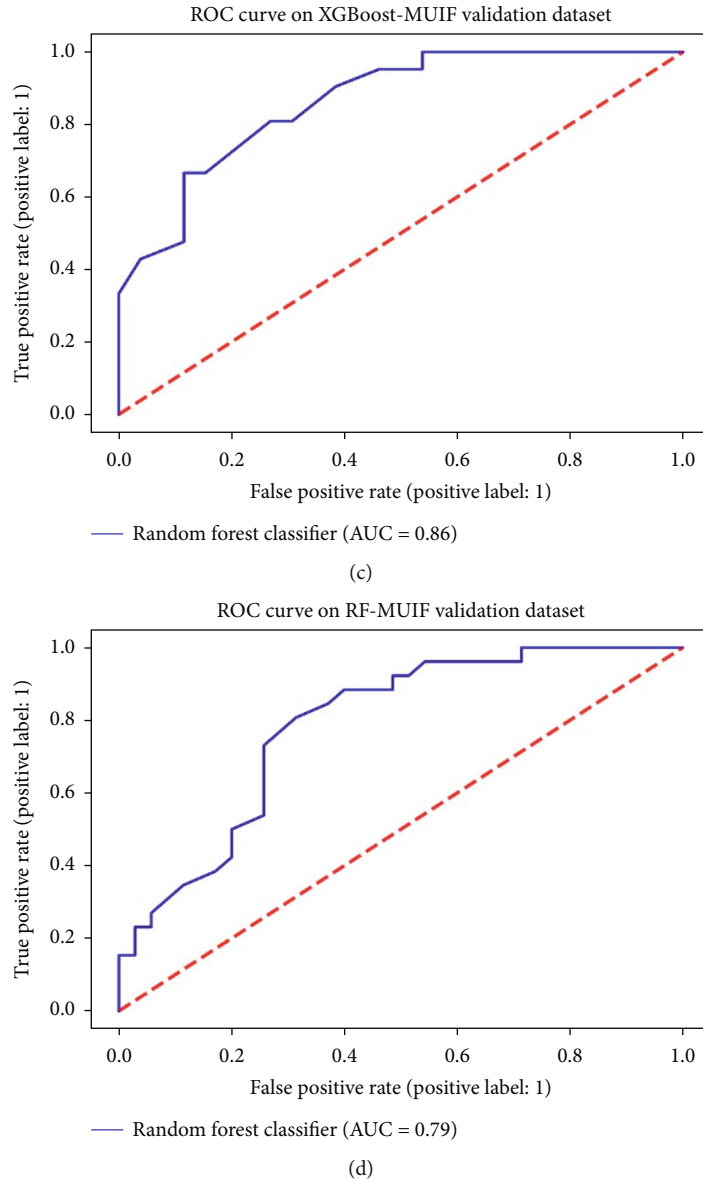


FIGURE 4: ROC curves of models on the training and validation sets. (a) The fivefold cross-validated ROC curve of model RF-MUIF. (b) The fivefold cross-validated ROC curve of model XGBoost-MUIF. (c) ROC curve of XGBoost-MUIF on the validation dataset. (d) ROC curve of RF-MUIF on the validation dataset.

338.90, $P < 0.001$), log-sigma-2-0-mm-3D_gldm_LargeDependenceLowGrayLevelEmphasis (0.91 vs 0.93, $P < 0.001$), original_gldm_Idmn (0.97 vs. 0.99, $P = 0.09$), original_glszm_SmallAreaHighGrayLevelEmphasis (20.63 vs. 23.87, $P = 0.06$), and SCC (3.91 vs. 6.03, $P = 0.15$). Five features showed higher value for EGFR mutant type compared with EGFR wild type. These features consisted of original_glrIm_ShortRunEmphasis (0.76 vs. 0.73, $P = 0.03$), original_glszm_LargeAreaLowGrayLevelEmphasis (2.36 vs. 2.20, $P = 0.07$), original_gldm_DependenceEntropy (8.40 vs. 5.84, $P < 0.001$), original_gldm_GrayLevelNonUniformity (9.11 vs. 8.44, $P < 0.001$), and wavelet-HLH_gldm_HighGrayLevelEmphasis (15.18 vs 13.56, $P < 0.001$) (Figure 5).

4. Discussion

In this retrospective study, we proposed a stable predictive model based on noninvasive CT images and clinical features in order to predict EGFR mutation status for patients with LUAC. The ML model was trained with 140 patients, and its performance was validated with 61 patients. This model showed favorable predictability in the validation set (AUC = 0.79). Similarly, the AUC of the integrated model built with radiomics features and clinical data was 0.86. This study demonstrated that the association was evident between CT image features and EGFR genotype and the ability of radiomics to identify the EGFR mutation status. Therefore, it

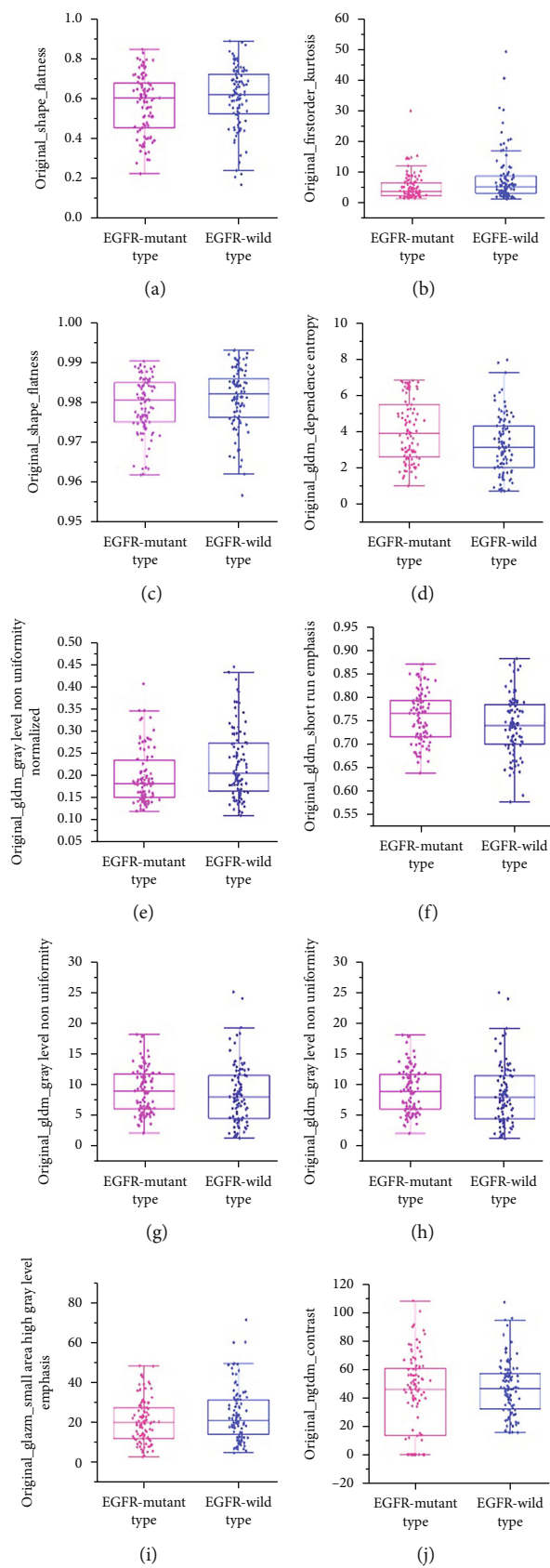


FIGURE 5: Continued.

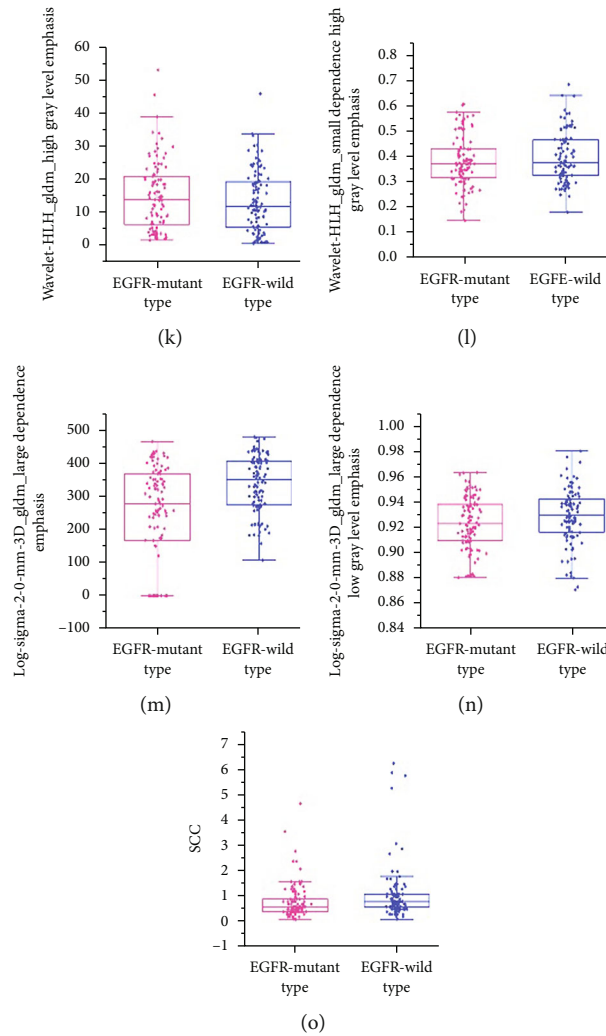


FIGURE 5: Boxplot illustrating the variation of the 15 features that finally incorporate into XGBoost-MUIF model between EGFR mutant type and EGFR wild type: (a) original_shape_Flatness; (b) original_firstorder_Kurtosis; (c) original_gldm_Idmn; (d) original_gldm_DependenceEntropy; (e) original_gldm_GrayLevelNonUniformityNormalized; (f) original_gldm_ShortRunEmphasis; (g) original_gldm_GrayLevelNonUniformity; (h) original_gldm_LargeAreaLowGrayLevelEmphasis; (i) original_gldm_SmallAreaHighGrayLevelEmphasis; (j) original_gldm_Contrast; (k) wavelet-HLH_gldm_HighGrayLevelEmphasis; (l) wavelet-HLH_gldm_SmallDependenceHighGrayLevelEmphasis; (m) log-sigma-2-0-mm-3D_gldm_LargeDependenceEmphasis; (n) log-sigma-2-0-mm-3D_gldm_LargeDependenceLowGrayLevelEmphasis; (o) SCC.

is possible to predict EGFR mutation before invasive biopsy and expensive molecular testing based on a noninvasive method. To the best of our knowledge, this is the only study which establishes ML models using filter methods to predict EGFR mutation status in patients of LUAC. The present study has made new contributions to the existing research in this field.

Radiomics is defined as the extraction of a myriad of radiographic image features and the further mining of these data with the intent of supporting adoption of precision medicine [35]. Radiomics analysis can be used to increase precision in establishing a diagnosis, assessing prognosis, and predicting therapy response in cancer patients. Some features have even been shown to identify genomic alterations in tumor tissue, which is termed as “radiogenomics” [36]. Radiogenomics examines the relationship between disease genomic characteristics and its radiomics features [37]. Although some limitations of the radiogenomics approach exist, radiogenomics will play an important role in cancer

research because it paves an avenue of obtaining important information from limited and incomplete data. This information might improve decision-making and, as a result, leads to better patient outcomes [38]. For example, recent studies have shown that radiogenomics can aid in treatment option and prognosis assessment in NSCLC patients [39, 40]. Additionally, radiogenomics can help in evaluating the efficacy of therapy and predicting outcomes of treatment [37, 39].

Previous studies have shown that EGFR mutation status can be predicted from image features in patients with NSCLC. For example, a study by Zhang et al. found that radiomics features are able to discriminate EGFR mutation in patients with NSCLC and the AUC was 0.862 and 0.873 for the training and validation cohort, respectively [41]. Mei et al. [42] analyzed the association between CT texture features and EGFR mutation statuses in patients with LUAC. They reported that AUC of combination with clinical and radiomics features to predict EGFR mutations was 0.664. Liu et al. [43] also predict EGFR

status with a model based on five radiomics features and obtain an AUC of 0.647 in surgically resected peripheral LUAC. When combined with clinical data, this model can reach an AUC of 0.709. In the study conducted by Gevaert et al. [44], the authors built a predictive model for the EGFR mutation and achieved an AUC of 0.89. Their work showed the potential of semantic image features to predict molecular properties. Recently, Wang et al. proposed a deep learning model to distinguish EGFR mutation status using CT images and clinical data. The AUC was 0.85 and 0.81 in the training and test cohorts, respectively [16]. Our results combined with previous studies clearly demonstrate that radiogenomics powered by ML can potentially aid in identifying patients who will benefit from targeted therapy.

FS is a process often used in ML, wherein a subset of predictor variables is selected from the input data for application of a learning algorithm [23]. FS is the core of classification which plays an essential role in image processing and ML [22]. The aims of FS include, but are not limited to, the following aspects: preventing overfitting of predictive and classifier models and achieving a good prediction performance, providing quicker and more optimizing computational solutions, and gaining a better insight into the underlying processes by which the data are generated [31, 32]. FS methods usually consist of three categories: wrapper, embedded, and filter. Most wrapper approaches are not computationally feasible for high-dimensional data sets [32]. Embedded methods search for the most optimal features during the training of the classifier, and they have better computational complexity than wrapper methods [45]. Filter methods calculate a score for each predictor variable and select those which exceed a defined threshold [31]. Unlike wrapper and embedded methods which are specific to a given learning algorithm, filter methods could be combined with any kind of predictive approaches [31]. Due to its independence of learning algorithms, filter approaches can prevent overfitting and demand less work in computation than wrapper and embedded methods [31]. As a result, although filter-based feature selection methods have some shortcomings, such as ignoring feature dependencies and providing feature subsets which perhaps contain redundant information, filter methods are increasingly used due to their efficiency, simplicity, and a good generalization capacity [46]. Zhang et al. built ML models based on CT radiomics features which were selected using filter methods to discriminate arteriovenous malformation-related intraparenchymal hematomas from those that were associated with other etiologies [47]. They obtained AUCs of 0.988 and 0.957 in the training and test cohorts, respectively. In the work presented by Parmar et al. [33], the authors showed that choosing WLCX, one of the filter methods, and/or RF classification method gets the highest performance in survival prediction based on 440 radiomics features extracted from 464 lung cancer patients. Our models achieved an AUC of 0.79 to identify EGFR mutation, which is comparable to the previous reports. It is worth noting that a deep learning approach has some shortcoming: requiring a huge amount of data for training, relying on more specialized hardware and computing power, and lack of interpretability [48, 49].

As a branch of artificial intelligence, ML is a method to identify patterns and relationships in data by building algorithmic models. ML has also been proven to be an interesting field in biomedical research and focuses on teaching computers to perform classification, prediction, or estimation and improve its own performance based on some experience (data) [50]. Supervised learning (training data are labeled) and unsupervised learning (training data are unlabeled) are two main common types of ML methods, and the former has been a dominant method in the data mining field [51]. Our retrospective study showed that it was feasible for 7 ML approaches to predict EGFR mutation status. When used in combination with the RF classifier, the majority of FS methods achieved the best predictive performance. This finding is in accordance with a recently reported study by Parmar et al. [33], who found that RF classification method yields the highest performance in the prediction of two-year patient survival in NSCLC patients. Gu et al. reported that RF-based radiomics classifier performed best (AUC = 0.776) in predicting the Ki-67 expression level in NSCLC [52]. Uddin et al. [51] compared different types of supervised ML algorithms to evaluate the potential for disease risk prediction. They found that the SVM algorithm is most frequently used whereas the RF algorithm gave superior accuracy comparatively. In addition, MUIF was found to have the highest predictive power with the majority of classifiers. MUIF can be used as relevant criterion for selecting predictive subsets of features [53]. Under some reasonable assumptions, features selected with MUIF are those whose mean squared error and mean absolute error are minimizing [54]. Our results combined with previous researches demonstrate that RF together with MUIF is a better ML approach for identifying EGFR mutations based on radiomics features.

The potential clinical utility of radiomics based model has also been assessed to predict EGFR mutation in this study. We identified SCC as the most important clinical predictor, which was consistent with previous reports [55, 56]. We found that age, gender, and s-CEA were not associated with the EGFR mutation status, which did not accord with previous studies [21, 57–59]. A meta-analysis of human epidemiologic data revealed that there are significantly increased odds of EGFR mutation in never smokers in comparison to ever smokers [60]. Hong et al. reported that female was more likely (OR = 3.124) to have EGFR mutations [21]. Wang et al. [57] demonstrated that high preoperative serum CEA levels (CEA > 20 ng/mL) were effective for predicting the EGFR mutation. With regard to the models integrating clinical characteristics and radiomics features, we found that the XGBoost-MUIF model performed better in predicting EGFR mutation status. These results are consistent with a previous study that reported that the genetic algorithm plus XGBoost classifier had the most favorable performance and reached an accuracy of 0.836 for detecting EGFR in patients with NSCLC [61].

The present study has some limitations. First, as the study was retrospective in nature, it was associated with flaws such as possible information and selection bias. Second, our sample size is relatively small. However, although larger data sets are associated with more power, radiomics analyses can be

performed with as few as 100 patients [62]. Further studies on large sample are required to assess the clinical applications as well as the stability of our models. Third, there were differences in the prevalence of EGFR mutations in LUAC and in subsequent treatments among different races [63], but all of subjects who were involved in this study were Chinese. Therefore, the results may lack universality and needs further verification within other racial and ethnic population. Finally, manual segmentation of ROI is time-consuming and its reproducibility should be evaluated by interobserver reproducibility analysis. Semiautomated or automated radiomics methods are expected in our future research to improve the robustness.

5. Conclusions

In conclusion, the present study showed that radiomics signature extracted from CT images in combination with clinical characters can identify EGFR mutation status in LUAC. Although these findings remain to be validated with a larger sample size, ML-based radiomics using filter methods provides a noninvasive and low-cost method to predict EGFR mutations, which may aid in screening patients before invasive sampling and developing personalized treatment design for optimizing the outcomes of patients with LUAC.

Data Availability

The original data supporting the conclusions of this paper will be provided unreservedly by the authors to any qualified researcher.

Ethical Approval

The studies involving human participants were reviewed and approved by the Institutional Review Committee of the First Affiliated Hospital of Shandong First Medical University (Jinan, China).

Consent

Written informed consent for participation in this study is not required in accordance with national legislative and institutional requirements.

Disclosure

The funders did not play a role in design of study, collection and analysis of data, or decision of preparing and publishing this manuscript.

Conflicts of Interest

The authors declare no conflict of interest.

Authors' Contributions

Jiameng Lu, Xiaoqing Ji, Degan Lu, Yafei Ning, and Jianping Xing were responsible for the conception and design. Jiameng Lu, Lixia Wang, Xiaoqing Ji, Yunxiu Jiang, Xinyi Liu, Haiying Peng, Fei Sun, Zihan Guo, Yue Lu, and Yanbo Ji

were responsible for collection and assembly of data. Jiameng Lu, Xiaoqing Ji, Lixia Wang, Zhenshen Ma, Jie Dong, Yunxiu Jiang, Xinyi Liu, Degan Lu were responsible for data analysis and interpretation.: All authors were responsible for manuscript writing and final approval of the manuscript.

Acknowledgments

This study was sponsored by the National Natural Science Foundation of China (No. 41904017), China Post-doctoral Science Foundation (No. 2021M691903), and the Collaborative Innovation Center for Intelligent Molecules with Multi-effects and Nanomedicine (No. 2019-01), Shandong Province, China.

Supplementary Materials

Table S1: the category and number of features. Figure S1: histogram of the ICC for radiomics features. ICC: intragroup correlation coefficient. (*Supplementary Materials*)

References

- [1] H. Sung, J. Ferlay, R. L. Siegel et al., "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a Cancer Journal for Clinicians*, vol. 71, no. 3, pp. 209–249, 2021.
- [2] B. D. Hutchinson, G. S. Shroff, M. T. Truong, and J. P. Ko, "Spectrum of lung adenocarcinoma," *Seminars in Ultrasound, CT, and MR*, vol. 40, no. 3, pp. 255–264, 2019.
- [3] M. D. Siegelin and A. C. Borczuk, "Epidermal growth factor receptor mutations in lung adenocarcinoma," *Laboratory Investigation*, vol. 94, no. 2, pp. 129–137, 2014.
- [4] D. H. Koo, K. P. Kim, C. M. Choi et al., "EGFR-TKI is effective regardless of treatment timing in pulmonary adenocarcinoma with EGFR mutation," *Cancer Chemotherapy and Pharmacology*, vol. 75, no. 1, pp. 197–206, 2015.
- [5] A. Passaro, T. Mok, S. Peters, S. Popat, M. J. Ahn, and F. de Marinis, "Recent advances on the role of EGFR tyrosine kinase inhibitors in the management of NSCLC with uncommon, non exon 20 insertions, EGFR mutations," *Journal of Thoracic Oncology*, vol. 16, no. 5, pp. 764–773, 2021.
- [6] T. Mitsudomi, T. Kosaka, and Y. Yatabe, "Biological and clinical implications of EGFR mutations in lung cancer," *International Journal of Clinical Oncology*, vol. 11, no. 3, pp. 190–198, 2006.
- [7] G. Ellison, G. Zhu, A. Moulis, S. Dearden, G. Speake, and R. McCormack, "EGFR mutation testing in lung cancer: a review of available methods and their use for analysis of tumour tissue and cytology samples," *Journal of Clinical Pathology*, vol. 66, no. 2, pp. 79–89, 2013.
- [8] J. P. O'Connor, C. J. Rose, J. C. Waterton, R. A. Carano, G. J. Parker, and A. Jackson, "Imaging intratumor heterogeneity: role in therapy response, resistance, and clinical outcome," *Clinical Cancer Research*, vol. 21, no. 2, pp. 249–257, 2015.
- [9] T. Uematsu and M. Kasami, "The use of positive core wash cytology to estimate potential risk of needle tract seeding of breast cancer: directional vacuum-assisted biopsy versus automated core needle biopsy," *Breast Cancer*, vol. 17, no. 1, pp. 61–67, 2010.

- [10] E. Rios Velazquez, C. Parmar, Y. Liu et al., "Somatic mutations drive distinct imaging phenotypes in lung cancer," *Cancer Research*, vol. 77, no. 14, pp. 3922–3930, 2017.
- [11] J. E. van Timmeren, D. Cester, S. Tanadini-Lang, H. Alkadhi, and B. Baessler, "Radiomics in medical imaging-"how-to" guide and critical reflection," *Imaging*, vol. 11, no. 1, p. 91, 2020.
- [12] M. E. Mayerhoefer, A. Materka, G. Langs et al., "Introduction to radiomics," *Journal of Nuclear Medicine*, vol. 61, no. 4, pp. 488–495, 2020.
- [13] S. S. Yip and H. J. Aerts, "Applications and limitations of radiomics," *Physics in Medicine and Biology*, vol. 61, no. 13, pp. R150–R166, 2016.
- [14] S. Li, T. Luo, C. Ding, Q. Huang, Z. Guan, and H. Zhang, "Detailed identification of epidermal growth factor receptor mutations in lung adenocarcinoma: combining radiomics with machine learning," *Medical Physics*, vol. 47, no. 8, pp. 3458–3466, 2020.
- [15] D. N. Ma, X. Y. Gao, Y. B. Dan et al., "Evaluating solid lung adenocarcinoma anaplastic lymphoma kinase gene rearrangement using noninvasive radiomics biomarkers," *Oncotargets and Therapy*, vol. Volume 13, pp. 6927–6935, 2020.
- [16] S. Wang, J. Shi, Z. Ye et al., "Predicting EGFR mutation status in lung adenocarcinoma on computed tomography image using deep learning," *The European Respiratory Journal*, vol. 53, no. 3, p. 1800986, 2019.
- [17] T. Zhang, Z. Xu, G. Liu et al., "Simultaneous identification of EGFR, KRAS, ERBB2, and TP53 mutations in patients with non-small cell lung cancer by machine learning-derived three-dimensional radiomics," *Cancers (Basel)*, vol. 13, no. 8, p. 1814, 2021.
- [18] M. Ren, H. Yang, Q. Lai et al., "MRI-based radiomics analysis for predicting the EGFR mutation based on thoracic spinal metastases in lung adenocarcinoma patients," *Medical Physics*, vol. 48, no. 9, pp. 5142–5151, 2021.
- [19] G. Rossi, E. Barabino, A. Fedeli et al., "Radiomic detection of EGFR mutations in NSCLC," *Cancer Research*, vol. 81, no. 3, pp. 724–731, 2021.
- [20] B. Zhang, S. Qi, X. Pan et al., "Deep CNN model using CT radiomics feature mapping recognizes EGFR gene mutation status of lung adenocarcinoma," *Frontiers in Oncology*, vol. 10, article 598721, 2021.
- [21] D. Hong, K. Xu, L. Zhang, X. Wan, and Y. Guo, "Radiomics signature as a predictive factor for EGFR mutations in advanced lung adenocarcinoma," *Frontiers in Oncology*, vol. 10, p. 28, 2020.
- [22] X. Li, Y. Wang, and R. Ruiz, "A survey on sparse learning models for feature selection," *IEEE Transactions on Cybernetics*, vol. 52, no. 3, pp. 1642–1660, 2020.
- [23] G. Zhao and S. Liu, "Estimation of discriminative feature subset using community modularity," *Scientific Reports*, vol. 6, no. 1, p. 25040, 2016.
- [24] M. Shafiq-Ul-Hassan, G. G. Zhang, K. Latifi et al., "Intrinsic dependencies of CT radiomic features on voxel size and number of gray levels," *Medical Physics*, vol. 44, no. 3, pp. 1050–1062, 2017.
- [25] P. A. Yushkevich and G. Gerig, "ITK-SNAP: an interactive medical image segmentation tool to meet the need for expert-guided segmentation of complex medical images," *IEEE Pulse*, vol. 8, no. 4, pp. 54–57, 2017.
- [26] P. A. Yushkevich, J. Piven, H. C. Hazlett et al., "User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability," *NeuroImage*, vol. 31, no. 3, pp. 1116–1128, 2006.
- [27] T. K. Koo and M. Y. Li, "A guideline of selecting and reporting intraclass correlation coefficients for reliability research," *Journal of Chiropractic Medicine*, vol. 15, no. 2, pp. 155–163, 2016.
- [28] R. T. Leijenaar, S. Carvalho, E. R. Velazquez et al., "Stability of FDG-PET Radiomics features: an integrated analysis of test-retest and inter-observer variability," *Acta Oncologica*, vol. 52, no. 7, pp. 1391–1397, 2013.
- [29] J. Morgado, T. Pereira, F. Silva et al., "Machine learning and feature selection methods for EGFR mutation status prediction in lung cancer," *Applied Sciences*, vol. 11, no. 7, p. 3273, 2021.
- [30] L. Yang, D. Dong, M. J. Fang et al., "Can CT-based radiomics signature predict KRAS/NRAS/BRAF mutations in colorectal cancer?," *European Radiology*, vol. 28, no. 5, pp. 2058–2067, 2018.
- [31] M. Piles, R. Bergsma, D. Gianola, H. Gilbert, and L. Tusell, "Feature selection stability and accuracy of prediction models for genomic prediction of residual feed intake in pigs using machine learning," *Frontiers in Genetics*, vol. 12, article 611506, 2021.
- [32] Y. Saeys, I. Inza, and P. Larranaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.
- [33] C. Parmar, P. Grossmann, J. Bussink, P. Lambin, and H. Aerts, "Machine learning methods for quantitative radiomic biomarkers," *Scientific Reports*, vol. 5, no. 1, p. 13087, 2015.
- [34] A. Abraham, F. Pedregosa, M. Eickenberg et al., "Machine learning for neuroimaging with scikit-learn," *Frontiers in Neuroinformatics*, vol. 8, p. 14, 2014.
- [35] P. Lambin, E. Rios-Velazquez, R. Leijenaar 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.
- [36] R. Thawani, M. McLane, N. Beig et al., "Radiomics and radiogenomics in lung cancer: a review for the clinician," *Lung Cancer*, vol. 115, pp. 34–41, 2018.
- [37] M. A. Mazurowski, "Radiogenomics: what it is and why it is important," *Journal of the American College of Radiology*, vol. 12, no. 8, pp. 862–866, 2015.
- [38] C. W. Wong and A. Chaudhry, "Radiogenomics of lung cancer," *Journal of Thoracic Disease*, vol. 12, no. 9, pp. 5104–5109, 2020.
- [39] L. Shi, Y. He, Z. Yuan et al., "Radiomics for response and outcome assessment for non-small cell lung cancer," *Technology in Cancer Research & Treatment*, vol. 17, p. 153303381878278, 2018.
- [40] M. Zhou, A. Leung, S. Echegaray et al., "Non-small cell lung cancer radiogenomics map identifies relationships between molecular and imaging phenotypes with prognostic implications," *Radiology*, vol. 286, no. 1, pp. 307–315, 2018.
- [41] L. Zhang, B. Chen, X. Liu et al., "Quantitative biomarkers for prediction of epidermal growth factor receptor mutation in non-small cell lung cancer," *Translational Oncology*, vol. 11, no. 1, pp. 94–101, 2018.
- [42] D. Mei, Y. Luo, Y. Wang, and J. Gong, "CT texture analysis of lung adenocarcinoma: can Radiomic features be surrogate biomarkers for EGFR mutation statuses," *Cancer Imaging*, vol. 18, no. 1, p. 52, 2018.

- [43] Y. Liu, J. Kim, Y. Balagurunathan et al., "Radiomic features are associated with EGFR mutation status in lung adenocarcinomas," *Clinical Lung Cancer*, vol. 17, no. 5, pp. 441–448.e6, 2016.
- [44] O. Gevaert, S. Echegaray, A. Khuong et al., "Predictive radiogenomics modeling of EGFR mutation status in lung cancer," *Scientific Reports*, vol. 7, no. 1, p. 41674, 2017.
- [45] B. Remeseiro and V. Bolon-Canedo, "A review of feature selection methods in medical applications," *Computers in Biology and Medicine*, vol. 112, article 103375, 2019.
- [46] R. Larracy, A. Phinyomark, and E. Scheme, "Machine learning model validation for early stage studies with small sample sizes," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, vol. 2021, pp. 2314–2319, Mexico, 2021.
- [47] Y. Zhang, B. Zhang, F. Liang et al., "Radiomics features on non-contrast-enhanced CT scan can precisely classify AVM-related hematomas from other spontaneous intraparenchymal hematoma types," *European Radiology*, vol. 29, no. 4, pp. 2157–2165, 2019.
- [48] G. Eraslan, Z. Avsec, J. Gagneur, and F. J. Theis, "Deep learning: new computational modelling techniques for genomics," *Nature Reviews. Genetics*, vol. 20, no. 7, pp. 389–403, 2019.
- [49] Y. Y. A. Teo, A. Danilevsky, and N. Shomron, "Overcoming interpretability in deep learning cancer classification," *Methods in Molecular Biology*, vol. 2243, pp. 297–309, 2021.
- [50] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Computational and Structural Biotechnology Journal*, vol. 13, pp. 8–17, 2015.
- [51] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction," *BMC Medical Informatics and Decision Making*, vol. 19, no. 1, p. 281, 2019.
- [52] Q. Gu, Z. Feng, Q. Liang et al., "Machine learning-based radiomics strategy for prediction of cell proliferation in non-small cell lung cancer," *European Journal of Radiology*, vol. 118, pp. 32–37, 2019.
- [53] C. O. Sakar, O. Kursun, H. Seker, and F. Gurgun, "Combining multiple clusterings for protein structure prediction," *International Journal of Data Mining and Bioinformatics*, vol. 10, no. 2, pp. 162–174, 2014.
- [54] B. Frenay, G. Doquire, and M. Verleysen, "Is mutual information adequate for feature selection in regression?," *Neural Networks*, vol. 48, pp. 1–7, 2013.
- [55] L. Wen, S. Wang, W. Xu et al., "Value of serum tumor markers for predicting EGFR mutations in non-small cell lung cancer patients," *Annals of Diagnostic Pathology*, vol. 49, 2020.
- [56] S. Wang, P. Ma, G. Ma et al., "Value of serum tumor markers for predicting EGFR mutations and positive ALK expression in 1089 Chinese non-small-cell lung cancer patients: a retrospective analysis," *European Journal of Cancer*, vol. 124, pp. 1–14, 2020.
- [57] Z. Wang, S. Yang, and H. Lu, "Preoperative serum carcinoembryonic antigen levels are associated with histologic subtype, EGFR mutations, and ALK fusion in patients with completely resected lung adenocarcinoma," *Oncotargets and Therapy*, vol. Volume 10, pp. 3345–3351, 2017.
- [58] J. Gu, S. Xu, L. Huang et al., "Value of combining serum carcinoembryonic antigen and PET/CT in predicting EGFR mutation in non-small cell lung cancer," *Journal of Thoracic Disease*, vol. 10, no. 2, pp. 723–731, 2018.
- [59] Z. Shi, X. Zheng, R. Shi et al., "Radiological and clinical features associated with epidermal growth factor receptor mutation status of exon 19 and 21 in lung adenocarcinoma," *Scientific Reports*, vol. 7, no. 1, p. 364, 2017.
- [60] A. M. Chapman, K. Y. Sun, P. Ruestow, D. M. Cowan, and A. K. Madl, "Lung cancer mutation profile of EGFR, ALK, and KRAS: meta-analysis and comparison of never and ever smokers," *Lung Cancer*, vol. 102, pp. 122–134, 2016.
- [61] N. Q. K. Le, Q. H. Kha, V. H. Nguyen, Y. C. Chen, S. J. Cheng, and C. Y. Chen, "Machine learning-based radiomics signatures for EGFR and KRAS mutations prediction in non-small-cell lung cancer," *International Journal of Molecular Sciences*, vol. 22, no. 17, p. 9254, 2021.
- [62] R. J. Gillies, P. E. Kinahan, and H. Hricak, "Radiomics: images are more than pictures, they are data," *Radiology*, vol. 278, no. 2, pp. 563–577, 2016.
- [63] J. Bauml, R. Mick, Y. Zhang et al., "Frequency of _EGFR_ and _KRAS_ mutations in patients with non small cell lung cancer by racial background: do disparities exist?," *Lung Cancer*, vol. 81, no. 3, pp. 347–353, 2013.