Diagn Interv Radiol 2025; DOI: 10.4274/dir.2024.242972

CHEST IMAGING

ORIGINAL ARTICLE

Automatic machine learning accurately predicts the efficacy of immunotherapy for patients with inoperable advanced non-small cell lung cancer using a computed tomography-based radiomics model

Siyun Lin[1,2*]
Zhuangxuan Ma[3*]
Yuanshan Yao[4]
Hou Huang[2]
Wufei Chen[3]
Dongfang Tang[1]
Wen Gao[1]

[1]Huadong Hospital, Fudan University, Department of Thoracic Surgery, Shanghai, China

[2]Shanghai Key Laboratory of Clinical Geriatric Medicine, Shanghai, China

[3]Huadong Hospital, Fudan University, Department of Radiology, Shanghai, China

[4]Shanghai Chest Hospital, Shanghai JiaoTong University School of Medicine, Department of Thoracic Surgery, Shanghai, China

***These authors are joint first authors.**

Corresponding authors: Wufei Chen, Dongfang Tang, Wen Gao

E-mails: chenwufei_2008@163.com, tangdongfangchest@163.com, gaowenchest@163.com

Received 08 August 2024; revision requested 13 September 2024; last revision received 12 October 2024; accepted 18 November 2024.

PURPOSE

Patients with advanced non-small cell lung cancer (NSCLC) have varying responses to immunotherapy, but there are no reliable, accepted biomarkers to accurately predict its therapeutic efficacy. The present study aimed to construct individualized models through automatic machine learning (autoML) to predict the efficacy of immunotherapy in patients with inoperable advanced NSCLC.

METHODS

A total of 63 eligible participants were included and randomized into training and validation groups. Radiomics features were extracted from the volumes of interest of the tumor circled in the preprocessed computed tomography (CT) images. Golden feature, clinical, radiomics, and fusion models were generated using a combination of various algorithms through autoML. The models were evaluated using a multi-class receiver operating characteristic curve.

RESULTS

In total, 1,219 radiomics features were extracted from regions of interest. The ensemble algorithm demonstrated superior performance in model construction. In the training cohort, the fusion model exhibited the highest accuracy at 0.84, with an area under the curve (AUC) of 0.89–0.98. In the validation cohort, the radiomics model had the highest accuracy at 0.89, with an AUC of 0.98–1.00; its prediction performance in the partial response subgroup outperformed that in both the clinical and radiomics models. Patients with low rad scores achieved improved progression-free survival (PFS); (median PFS 16.2 vs. 13.4, $P = 0.009$).

CONCLUSION

autoML accurately and robustly predicted the short-term outcomes of patients with inoperable NSCLC treated with immune checkpoint inhibitor immunotherapy by constructing CT-based radiomics models, confirming it as a powerful tool to assist in the individualized management of patients with advanced NSCLC.

CLINICAL SIGNIFICANCE

This article highlights that autoML promotes the accuracy and efficiency of feature selection and model construction. The radiomics model generated by autoML predicted the efficacy of immunotherapy in patients with advanced NSCLC effectively. This may provide a rapid and non-invasive method for making personalized clinical decisions.

KEYWORDS

Advanced non-small cell lung cancer, immunotherapy, radiomics, automatic machine learning, models

Epub: 16.01.2025

Publication date:

DOI: 10.4274/dir.2024.242972

You may cite this article as: Lin S, Ma Z, Yao Y, et al. Automatic machine learning accurately predicts the efficacy of immunotherapy for patients with inoperable advanced non-small cell lung cancer using a computed tomography-based radiomics model. *Diagn Interv Radiol.* 16 January 2025 DOI: 10.4274/dir.2024.242972 [Epub Ahead of Print].

Non-small cell lung cancer (NSCLC) is a prevalent and malignant tumor with high incidence and mortality rates globally.[1] Over 30% of new NSCLC cases are diagnosed at locally advanced stages [tumor–node–metastasis (TNM) stage III]. The absence of notable early symptoms often leads to diagnoses at advanced stages or after local metastasis has occurred, which frequently delays surgical treatment.

The current standard treatment for patients with advanced NSCLC involves concurrent chemoradiotherapy followed by immunotherapy.[2] Definitive efficacy and improved prognoses have been achieved in all stages of NSCLC with the use of immune checkpoint inhibitors (ICIs), either alone or in combination with chemotherapy.[3,4] In the CHECKMATE-816 clinical trial, nivolumab combined with chemotherapy extended event-free survival (EFS) by 10.8 months and decreased the risk by 37% compared with the control group [hazard ratio (HR) 0.63, confidence interval (CI): 0.43–0.91, $P = 0.0052$].[5] Furthermore, the recent NEOTORCH trial reported a similar extension in EFS and a significantly higher pathological complete response (CR) rate (24.8% vs. 1.0%, $P < 0.0001$) in the group receiving combined immune-chemotherapy.[6] However, in the Pacific trial (NCT02125461), only one-third of patients who received adjuvant therapy with durvalumab remained disease-free after 5 years,[7,8] indicating that immunotherapy may not be suitable for all patients due to factors such as the specific tumor immune microenvironment, residual toxicity, and societal expense. Effective immunotherapy is often positively correlated with high programmed death-ligand 1 (PD-L1) expression and the tumor mutation burden (TMB), but these require tissue from biopsies for detection. The challenge of not being able to perform repeated biopsies after developing chemo-resistance complicates treatment options for patients at an advanced stage. Therefore, there is an urgent need to develop non-invasive methods to accurately predict the efficacy of immunotherapy, which could benefit a broader group of patients.

Main points

- Radiomics modeling based on computed tomography images predicted the efficacy of immunotherapy in patients with advanced non-small cell lung cancer effectively.
- Automatic machine learning can integrate multiple algorithms to obtain improved predictive capabilities.
- The diagnostic performance of the radiomics model outperformed that of the clinical model.
- Patients with lower rad scores achieved superior progression-free survival.

In recent years, thin-slice computed tomography (CT) scans have become integral in diagnosing and staging NSCLC.[9,10] With advancements in medical imaging, there has been a transition from traditional qualitative diagnosis to the extraction of multimodal image data for quantitative analysis. Radiomics, a promising tool in image analysis, allows for the extraction of high-throughput features from imaging data. These features, combined with specific modeling techniques, can enhance the accuracy of disease diagnosis, differentiation, and prognosis evaluation.[11] Previously, we developed and implemented delta radiomics diagnostic features to refine and personalize the diagnosis of invasive adenocarcinoma in lung partial solid nodules.[12]

Automatic machine learning (autoML) algorithms have facilitated the analysis of complex, large-sample data into predictive models and automated classifications. By integrating substantial amounts of data from radiology, pathology, genomics, and proteomics, autoML has enhanced clinical decision-making.[13] In the present study, we aimed to identify effective radiomics features in CT images using autoML and integrate them with clinical features to develop a fusion model for individualized efficacy prediction and progression assessment in patients with advanced NSCLC receiving immunotherapy.

Methods

Study design and population

In this retrospective observational single-center study, we reviewed patients with NSCLC who underwent ICI treatment at Huadong Hospital between January 2020 and December 2022. The inclusion criteria were as follows: (1) >18 years; (2) receiving ICI treatment (anti-PD-1/PD-L1) at Huadong Hospital for the first time; (3) a clinically confirmed diagnosis of unresectable locally advanced stage NSCLC [stage III–IV, Union for International Cancer Control/American Joint Committee on Cancer (8th edition)]; and (4) available thin-slice CT images (1–1.25 mm), with lesions delineated and evaluated. The exclusion criteria were as follows: (1) a pathologically confirmed diagnosis of small cell lung cancer; (2) a history of malignancies other than NSCLC; (3) poor CT image quality with artifacts; and (4) failure to extract radiomics features due to other reasons.

Finally, a total of 63 eligible cases were enrolled (Figure 1). The clinical features before receiving ICIs were collected from medical records, including age, gender, smoking history, the time of diagnosis, pathological type, tumor location, the maximum diameter of the primary tumor site, clinical tumor stage, metastatic location, driver gene mutation, the start time and type of ICI treatment, treatment regimen, and disease progression and survival information. The efficacy evaluation

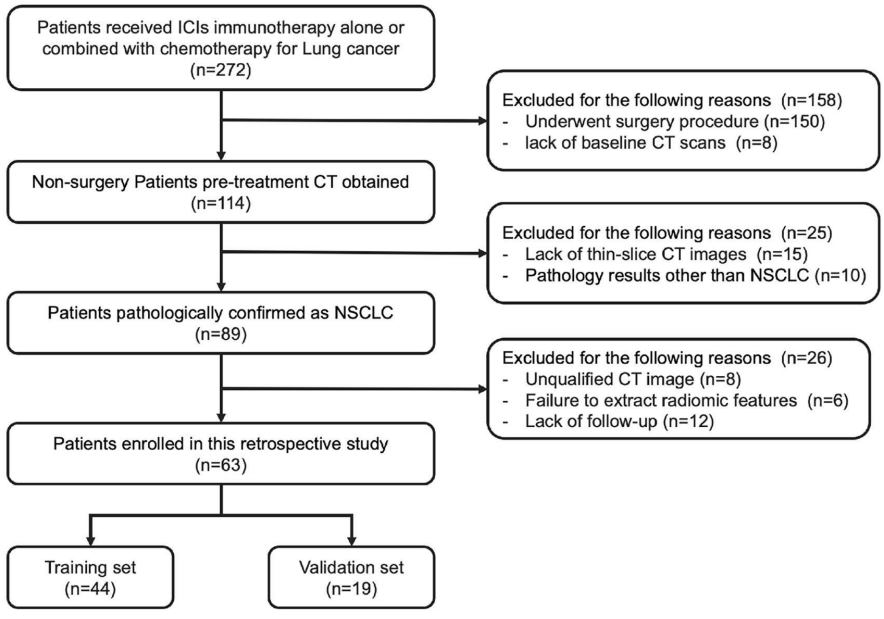


Figure 1. Study flowchart. ICIs, immune checkpoint inhibitors; CT, computed tomography; NSCLC, non-small cell lung cancer.

was based on the immune-related response evaluation criteria in solid tumors,[14] which classifies outcomes as CR, partial response (PR), stable disease (SD), and PD. The disease control rate (DCR) refers to the sum of all patients who were CR, PR, and SD. All the enrolled cases were further separated into a training and a validation cohort randomly after adjusting for potential confounders. The study was approved by the Ethics Committee of Huadong Hospital, and the requirement for informed consort was waived (approval no.: 2022K033, date:XXX).

Computed tomography image acquisition

The patients in this study were all subjected to non-contrast-enhanced CT performed on two scanners: a Somatom Definition Flash scanner (Siemens Medical Solutions, Erlangen, Germany) and a GE Discovery CT750 HD scanner (GE Healthcare, MO, USA) at 120 kV. The detailed scanning parameters are shown in Supplementary Table 1. The overall scanning range was from the lung apex to the bilateral adrenal gland. During the examination, the patients were instructed to lie in a supine position and inhale deeply with both arms raised.

Target segmentation and radiomics features extraction

According to the target lesions on the axial slices of the initial CT scans, the volumes of interest (VOIs) were manually marked by two experienced radiologists, each with 5 years' expertise in diagnosing chest CT images, to achieve three-dimensional (3D) segmentation using the open-source 3D Slicer software (version 4.13.0; National Institutes of Health).

The extraction of radiomic features from these tumor VOIs was automatically performed using pyRadiomics (version: 3.0.1).[15] To assess the inter-rater reliability between the radiologists, the intraclass correlation coefficient (ICC) was employed, with ICC >0.75 indicating a high level of agreement. The types of radiomic features extracted included grayscale, shape, texture, and wavelet transform features.

Feature selection and model construction

Due to the broad variability in the initial dataset, the data underwent normalization to control the radiomics features within a standardized intensity range. Feature selection was performed within the training cohort. The MLJAR platform, an open-source software based on Python, was employed for predictive feature selection and modeling.[16] This platform is designed to automatically address missing data by implementing strategies such as mean or median imputation to maintain data integrity. It also manages categorical variables by automatically performing encoding transformations, such as one-hot encoding or label encoding, enabling machine learning algorithms to effectively interpret these features. Subsequently, a feature engineering step was undertaken to create "golden features" that possess enhanced predictive power, derived from the original dataset features through operations such as addition, subtraction, multiplication, and division. Throughout the training phase, MLJAR assessed the significance of each feature using techniques such as permutation importance or SHapley Additive exPlanations, providing a quantitative measure of each

Table 1. Basic characteristics of the enrolled patients in the training cohort and validation cohorts

	Training cohort (n = 44)					Validation cohort (n = 19)				
	Total	PR	SD	PD	*P*	Total	PR	SD	PD	*P*
Age										
<60	11	3	3	5	0.484	7	4	0	3	0.731
≥60	33	12	4	17		12	6	1	5	
Gender										
Male	34	14	4	16	0.13	10	6	1	3	0.396
Female	10	1	3	6		9	4	0	5	
Pathological type										
LSCC	13	8	2	3	0.034*	13	7	1	5	0.739
LUAD	31	7	5	19		6	3	0	3	
Tumor location										
Right	30	10	5	15	0.975	13	7	0	6	0.311
Left	14	5	2	7		6	3	1	2	
cT stage										
T1–T2	22	5	3	14	0.179	9	4	1	4	0.509
T3–T4	22	10	4	8		10	6	0	4	
cN stage										
N0	9	2	2	5	0.663	1	0	0	1	0.484
N+	35	13	5	17		18	10	1	7	
cM stage										
M0–M1a	30	11	6	13	0.365	13	8	1	4	0.311
M1b–M1c	14	4	1	9		6	2	0	4	
cTNM stage										
III	11	5	2	4	0.563	4	3	1	0	0.041*
IV	33	10	5	18		15	7	0	8	
Driver gene mutation										
Negative	35	13	6	16	0.533	15	7	1	7	0.577
Positive	9	2	1	6		4	3	0	1	
Smoking status										
Never	22	7	5	10	0.464	12	7	0	5	0.383
Ex- or current	22	8	2	12		7	3	1	3	
Treatment										
Without CHT	9	1	3	5	0.137	5	2	1	2	0.221
With CHT	35	14	4	17		14	8	0	6	
PD-L1 expression										
<50%	33	11	6	16	0.573	11	6	1	4	1.000
≥50%	11	6	1	4		8	5	0	3	

*Means statistical significance existed (Fisher exact probability test, $P < 0.05$). PR, partial response; SD, stable disease; PD, progressive disease; LSCC, lung squamous cell carcinoma; LUAD, lung adenocarcinoma; CHT, chemotherapy; cT stage, clinical tumor stage; cN stage, clinical node stage; cM stage, clinical metastasis stage; cTNM stage, clinical tumor-node-metastasis stage; PD-L1, programmed death-ligand 1.

feature's impact on the model's predictive accuracy and offering insight into the underlying decision-making processes of the model.

Afterward, in the "competition" mode of MLJAR, the software sought the most effective algorithms from a range, including linear regression, light gradient-boosting machine (LightGBM), eXtreme gradient boosting, neural networks (NN), and random forest (RF). Additionally, it considered assembling multiple algorithms to finalize the modeling process. The rad score was obtained by multiplying the coefficients of each feature by its value and then summing the results to get the final value.

The predictive model, which included clinical, radiomics, and fusion models, was developed using the aforementioned autoML algorithms. The efficacy of each model was assessed through receiver operator characteristic (ROC) curves for both the training and validation cohorts. Subsequently, the area under the curve (AUC) was calculated to determine the predictive accuracy of each constructed model.

Statistical analysis

The feature extraction and statistical analysis procedures were conducted using R software (version 3.6.2; http://www.Rproject.org and SPSS 22 (IBM, IL, USA). Categorical variables were analyzed using Fisher's exact test. To evaluate the multi-class ROC curves, both the macro-AUC and micro-AUC were calculated. The macro-AUC averaged the AUC values from each category, whereas the micro-AUC computed the weighted average after evaluating each category independently. Furthermore, model performance was assessed using statistical metrics such as accuracy, precision, recall, and F1-score.

Model performance was evaluated by ROC analysis, and the significance level of curves was compared using the DeLong test. A COX regression analysis was utilized to investigate factors associated with disease progression and survival. Survival rates were analyzed using the Kaplan–Meier method, and survival data comparisons were conducted with the log-rank test. A two-sided *P* value less than 0.05 was considered statistically significant for all tests.

Results

Basic characteristics of patients

The basic characteristics of the patients are listed in Table 1. In total, 63 patients with advanced NSCLC who had received ICIs in our hospital were randomly divided into the training cohort (n = 44, PR: 15, SD: 7, and PD: 22) and the validation cohort (n = 19, PR: 10, SD: 1, and PD: 8) based on the efficacy evaluation (Supplementary Table 2).

In the training cohort, differences were observed in the tumor pathological types of patients with various curative effects [lung squamous cell cancer vs. lung adenocarcinoma (LUAD), 13 vs. 31, $P = 0.034$]. In the validation cohort, a difference in the clinical TNM (cTNM) stage was observed (cTNM III vs. cTNM IV, 4 vs. 15, $P = 0.041$). No differences were observed in age, gender, tumor location, driver gene mutations, smoking history, PD-L1 expression, or combination therapy among the patients (all $P > 0.05$).

Selection of radiomics and clinical golden features

The radiomics feature selection workflow is shown in Figure 2. The VOIs were automatically extracted, yielding a total of 1,219 features. Within the training cohort, the golden features, regarded as the most predictive features, were selected for the subsequent model construction by autoML. Among the radiomics features, based on the superior performance of the LightGBM algorithm, log-sigma-4-0mm_ Glrlm_ Lowgraylevelrunemphasis had the highest mean of feature importance; the top 25 golden features are listed in Supplementary Figure 1. The rad scores for patients undergoing ICI treatment were significantly lower in the DCR group than in the PD group in both the training (0.105 ± 0.284 vs. 0.502 ± 0.318, $P < 0.001$) and the validation cohorts (0.119 ± 0.224 vs. 0.528 ± 0.262, $P = 0.002$) (Supplementary Figure 2).

Among the clinical features, ten golden features were identified and selected for model building using autoML. Among these, the feature representing the combination with chemotherapy (feature 11) was identified as the most critical (Supplementary Figure 3).

Model construction and performance comparison

Based on the input of golden features with the highest importance, different learning algorithms were selected for establishing each model (Supplementary Figure 4). The ensemble algorithm demonstrated the low-

Table 2. Performance evaluation of the clinical, radiomics, and fusion models

	Training cohort			Validation cohort		
	Clinical model	Radiomics model	Fusion model	Clinical model	Radiomics model	Fusion model
Micro-AUC	0.93	0.94	0.93	0.90	0.98	0.97
Macro-AUC	0.92	0.94	0.95	0.92	0.99	0.98
Accuracy	0.80	0.77	0.84	0.74	0.89	0.84
AUC (95% CI)						
PR	0.92 0.646 to 0.953	0.92 0.614 to 0.928	0.89 0.602 to 0.913	0.88 0.474 to 0.895	0.99 0.737 to 1.000	0.96 0.698 to 0.968
SD	0.87 0.549 to 0.925	0.96 0.638 to 1.000	0.98 0.676 to 0.996	1.00 /	1.00 /	1.00 /
PD	0.96 0.664 to 0.965	0.92 0.696 to 0.928	0.91 0.688 to 0.921	0.83 0.605 to 0.934	0.98 0.605 to 1.000	0.94 0.653 to 0.956
***P* value**	0.004*	0.005*	<0.001*	0.015*	0.060	0.010*
Precision						
PR	0.87	0.80	0.80	0.80	0.90	0.90
SD	0.71	0.86	0.86	1.00	1.00	1.00
PD	0.77	0.73	0.86	0.62	0.88	0.75
Recall						
PR	0.76	0.75	0.86	0.80	1.00	0.82
SD	0.56	0.67	0.75	0.50	0.50	1.00
PD	0.94	0.84	0.86	0.71	0.88	0.86
F1-score						
PR	0.81	0.77	0.83	0.80	0.95	0.86
SD	0.85	0.75	0.80	0.67	0.67	1.00
PD	0.63	0.78	0.86	0.67	0.88	0.80

*Means statistical significance existed between the AUC values among the models (DeLong test, $P < 0.05$). AUC, area under the curve; 95% CI, 95% confidence interval; PR, partial response; SD, stable disease; PD, progressive disease.

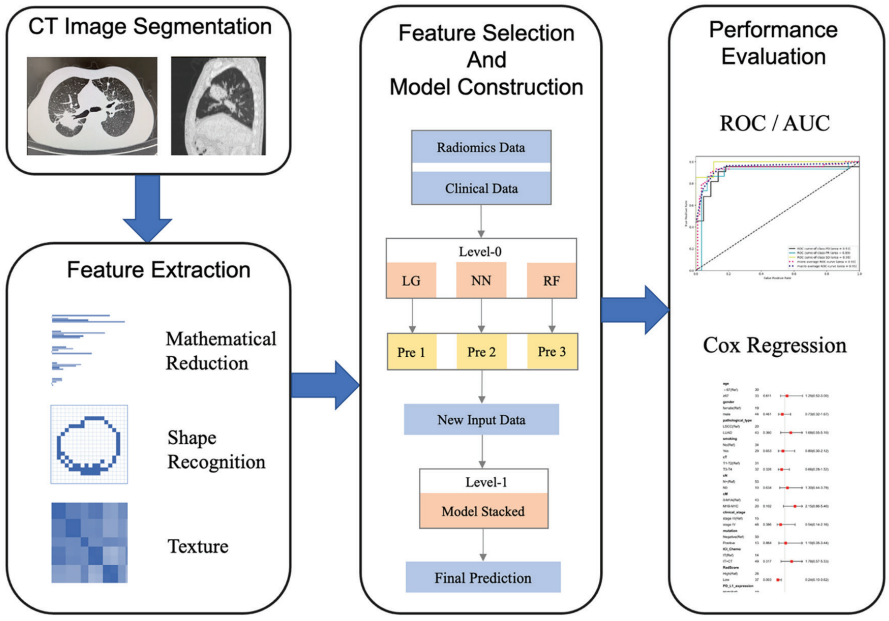


Figure 2. Workflow for the radiomics analysis. ROC, receiver operator characteristic; AUC, area under the curve.

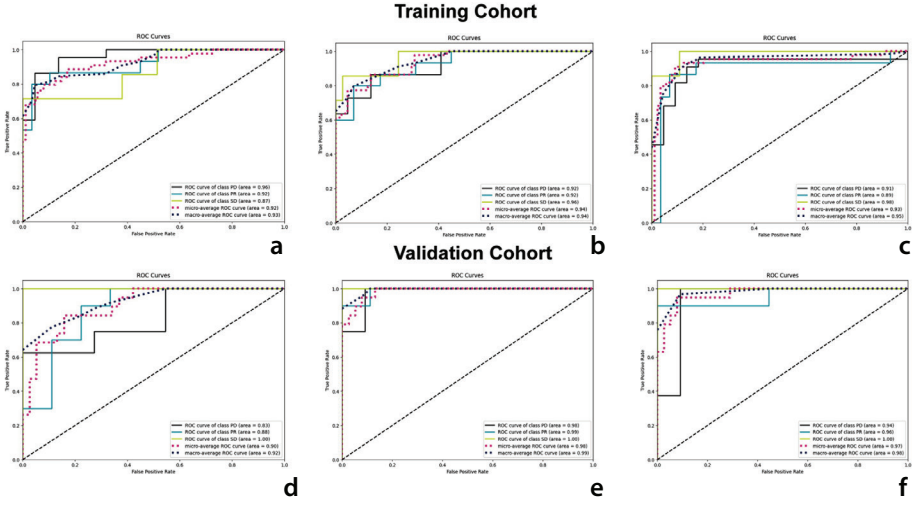


Figure 3. Evaluation of the performance of the different models. **(a)** Clinical model in the training cohort; **(b)** radiomics model in the training cohort; **(c)** fusion model in the training cohort; **(d)** clinical model in the validation cohort; **(e)** radiomics model in the validation cohort; **(f)** fusion model in the validation cohort. ROC, receiver operator characteristic.

est log-loss value in both the clinical and fusion models, indicating greater accuracy and a superior alignment between the predicted results and actual outcomes. In the radiomics model, the performance matched that of LightGBM, also suggesting improved accuracy and consistency.

Our study has shown that in both the radiomics and fusion models, the micro-AUC and macro-AUC were higher than those in the clinical model across the training and validation cohorts. In terms of accuracy, the fusion model scored the highest in the training cohort with 0.84, whereas the radiomics model outperformed the other models in the validation cohort with 0.89. In the training cohort, the radiomics and fusion models both exhibited optimal performance in SD, with an AUC of 0.96 (95% CI, 0.638–1.000) and 0.98 (95% CI, 0.676–0.996), respectively. In the validation cohort, the AUC of the radiomics model in three subgroups (PR, PD, and SD) were all higher than in the clinical and fusion models. Additionally, in the validation cohort, the PR subgroup exhibited better recall values and F1-scores than the SD and PD subgroups in both the clinical and radiomics models, suggesting enhanced predictive performance for this subgroup (Table 2, Figure 3).

Model prediction of progression-free and overall survival

All the enrolled patients were followed up for progression-free survival (PFS) and overall survival (OS), including 30 disease-progressed cases and 8 deaths, with a median follow-up time of 20 months (range: 3–47 months). Based on a nomogram derived from the multivariate COX regression analysis, patients undergoing ICI treatment were divided into high and low rad-score groups, with a threshold of 0.3 (Figure 4a). Regression analysis confirmed that the rad score was a more accurate predictor of progression risk than clinical factors (HR: 0.25, 95% CI: 0.10–0.63, P = 0.004) (Figure 4b). Although there was no significant difference in OS between the high and low rad-score groups (20.2 vs. 21.8 months, P = 0.056), the median PFS was notably longer in the low-score group, at 16.2 months, compared with 13.4 months in the high-score group (P = 0.009) (Supplementary Figure 5). The above data suggest that patients with low rad scores, as determined by the radiomics model, tend to experience less progression following immunotherapy.

Discussion

In the present study, we developed and validated a radiomics-based model using autoML algorithms to non-invasively assess the efficacy of immunotherapy in patients with inoperable advanced NSCLC. The findings revealed that the model, which incorporates features from CT images, displayed robust capabilities for diagnostics as well as for predicting therapeutic efficacy and disease progression.

In addition to PD-L1 expression, recent studies have shown that ICIs are highly effective in patients with high microsatellite instability or deficient mismatch repair (dMMR). Tumor cells with dMMR characteristics tend to have a higher TMB, which leads to the production of a considerable number of neoantigens. These neoantigens facilitate the recruitment of lymphocytes that become tumor-infiltrating lymphocytes, inhibiting tumor growth and enhancing the efficacy of immunotherapy.[17,18] However, these markers are typically identified through pathological immunohistochemistry or next-generation sequencing analysis, which require invasive tissue sampling and are costly. Therefore, there is a need for non-invasive, cost-effective, and accurate predictive methods using radiomics.

Progress in computerized imaging technology has led to the production of high-

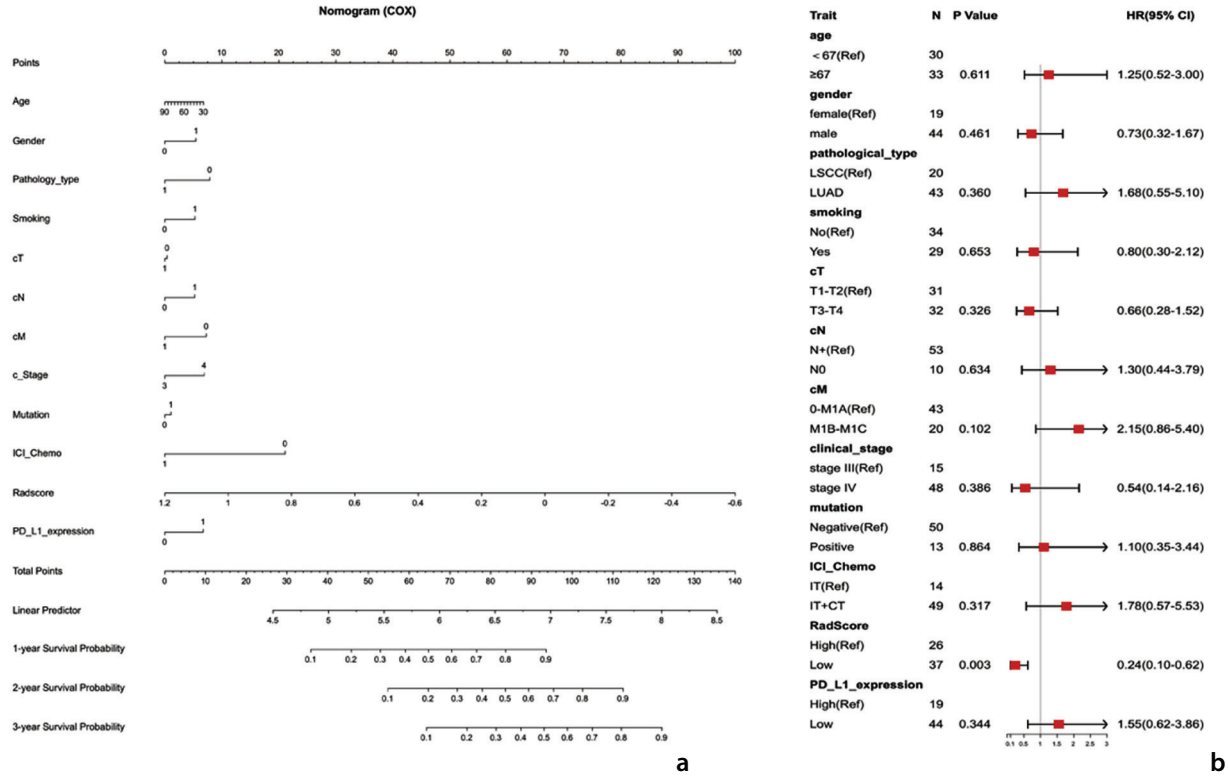


Figure 4. Rad score reflecting the risk of progression using COX regression analysis. **(a)** Nomogram of the rad scores and clinical risk factors; **(b)** results of the COX regression analysis.

er-definition images, enhancing radiomics' ability to extract more intricate features than traditional imaging methods. This advancement supports the performance of high-dimensional quantitative analysis, providing additional insights for clinical decision-making.[19] At present, numerous researchers have developed models with refined features that demonstrate high evaluation efficacy in various NSCLC application scenarios. These models have proven effective in predicting lesion benignity and malignancy, lymph node metastasis, driver mutations, and the severity of adverse effects.[20-23] For example, Yoon et al.[24] discovered that CT imaging features could non-invasively predict PD-L1 expression, identifying that validated radiomics models had greater discriminatory power than those generated from clinical features alone in an advanced LUAD cohort. Similarly, Trebeschi et al.[25] identified a non-invasive machine learning biomarker capable of differentiating between responders and non-responders to immunotherapy, and this model achieved an AUC value of 0.83 in lung cancer studies.[24]

In all our models, the predictive performance for the PR subgroup exceeded that for the PD subgroup. These results suggest that our model aided in identifying patients who are likely to benefit from immunotherapy. However, the diagnostic consistency for the SD subgroup in the validation cohort remained uncertain due to the limited sample size. Previous studies typically focused on binary outcomes, such as categorizing responses as effective or ineffective or progressive and non-progressive, which often excluded patients in the SD subgroup. The antitumor effect in the SD subgroup is considered ambiguous, leading to no significant differences in OS compared with the PR or PD subgroups. Although fusion models are generally regarded as having superior predictive capabilities, in this study, they only excelled in the SD subgroup compared with both clinical and radiomics models alone. This occurred because the features extracted from the images, when processed by autoML, might yield diagnoses that contradict clinical features, thereby reducing the predictive accuracy of the fusion model.

In the survival analysis, variations in PFS were observed among patients with differing rad scores ($P = 0.056$), though there was no statistically significant difference in OS. This lack of significance in OS could be due to all patients being in the advanced stages of the disease (cIII–cIV) and exhibiting either lymph node or distant metastasis, both of which are associated with higher risks. In studies with smaller sample sizes and shorter follow-up times, PFS may be a more suitable endpoint than OS, although OS remains the gold standard for measuring clinical benefit.

Furthermore, a positive result in PFS does not always translate to a benefit in OS. This discrepancy can arise because the toxic side effects of a treatment might cause a statistical bias in the PFS assessment, with drugs that have higher side effects potentially showing a "false" PFS advantage during shorter follow-up periods. In this study, the high rad-score group accounted for more than half of the recurrences (median PFS: 13.8 months), whereas the low rad-score group did not reach the median PFS. Median OS was not achieved in either group. The median follow-up time was 20 months, exceeding the median PFS by 6.2 months, which may also indicate robust results.

With the progression in central processing unit and graphics processing unit technology, deep learning and autoML methods have gained popularity.[26] In the present study, various algorithms were sequentially employed to develop clinical, radiomics, and fusion models via autoML. Among these, the ensemble models that integrated multiple classifiers demonstrated superior performance. However, the radiomics model, developed using LightGBM, achieved prediction levels in the training cohort comparable to those of the ensemble model. LightGBM is a framework that implements the gradient-boosting decision tree algorithm. This algorithm is well-regarded in machine learning for its

ability to iteratively train weak classifiers to derive an optimal model, notable for its efficient parallel training, improved accuracy, and capability to prevent overfitting.[27,28] In response to the characteristics of the dataset, different machine learning algorithms have demonstrated their respective performance advantages. For instance, Wiesweg et al.[29] applied support vector machine modeling to analyze RNA expression from biopsy samples in patients with advanced NSCLC, identifying seven genes predictive of immunotherapy response. Similarly, using a cytokine-based ICI response index, Wei et al.[30] employed RF modeling to predict responses to ICIs in patients with NSCLC. In the present study, we harnessed autoML to amalgamate multiple algorithms, developing models that exhibited enhanced predictive efficacy. This approach could significantly aid in predicting the effectiveness and survival outcomes of ICI treatment in patients with advanced NSCLC.

The current study has several limitations. First, being a single-center retrospective study with a small sample size in the training cohort, there is a potential impact on the specificity of the models, necessitating the collection of multicenter clinical data to confirm the models' robustness. Second, CT images were obtained from two scanning devices, which might have an adverse effect on radiomics feature extraction caused by uniformity. The MLJAR platform offers capabilities for model interpretation. As the complexity of the autoML models increases, their interpretability decreases, making it difficult for clinicians to understand and trust the model outputs, which could affect the reliability of model outcomes and the quality of decision-making. Moreover, the assessment of PD-L1 expression was limited by the amount of tissue available for fine-needle biopsy, resulting in some patients not being accurately assessed. It is also crucial in practice to select the most suitable combination of autoML algorithms, tailored to the specific characteristics of the data.

Furthermore, although the primary goal of this study was to provide surgical and oncology specialists with a predictive tool for treatment efficacy in patients with advanced NSCLC, challenges have arisen in accurately identifying lesions on CT images. To address this, Jiang et al.[31] developed a multi-scale convolutional NN method that integrates features from different resolutions to segment lung tumors accurately, facilitating the precise and automated tracking of tumor volumes. Integrating similar diagnostic models could enhance the utility of autoML in clinical settings. Moreover, although autoML allows for the training of numerous deep learning models with minimal coding or data input, the performance of these models can vary, and there remains room to improve both efficiency and prediction accuracy. Models that are designed and refined by experts may prove more reliable, and further clarification is needed on their clinical relevance and guidelines for practical diagnosis and treatment.

In conclusion, autoML has the ability to accurately predict the efficacy of immunotherapy and the short-term prognosis of patients with inoperable advanced NSCLC by constructing CT-base radiomics models, aiding the clinical evaluation and screening of a broader population and the development of personalized treatment strategies.

Conflict of interest disclosure

The authors declared no conflicts of interest.

Funding

The work was supported by the National Natural Science Foundation of China (62203117), the National Key R&D Program of China (2022YFF1203301), Research Project Plan of Shanghai Municipal Health Commission (20214Y0309) and Huadong Hospital Clinical Trial Project (HDLC2022012).

References

1. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2022. *CA Cancer J Clin*. 2022;72(1):7-33. [CrossRef]
2. Isaka T, Ito H, Nakayama H, Yokose T, Yamada K, Masuda M. Effect of epidermal growth factor receptor mutation on early-stage non-small cell lung cancer according to the 8th TNM classification. *Lung Cancer*. 2020;145:111-118. [CrossRef]
3. Lahiri A, Maji A, Potdar PD, et al. Lung cancer immunotherapy: progress, pitfalls, and promises. *Mol Cancer*. 2023;22(1):40. [CrossRef]
4. Borghaei H, Gettinger S, Vokes EE, et al. five-year outcomes from the randomized, phase III Trials CheckMate 017 and 057: nivolumab versus docetaxel in previously treated non-small-cell lung cancer. *J Clin Oncol*. 2021;39(7):723-733. [CrossRef]
5. Forde PM, Spicer J, Lu S, et al. Neoadjuvant nivolumab plus chemotherapy in resectable lung cancer. *N Engl J Med*. 2022;386(21):1973-1985. [CrossRef]
6. Lu S, Zhang W, Wu L, et al. Perioperative toripalimab plus chemotherapy for patients with resectable non-small cell lung cancer: the neotorch randomized clinical trial. *JAMA*. 2024;331(3):201-211. [CrossRef]
7. Spigel DR, Faivre-Finn C, Gray JE, et al. Five-year survival outcomes from the PACIFIC trial: durvalumab after chemoradiotherapy in stage III non-small-cell lung cancer. *J Clin Oncol*. 2022;40(12):1301-1311. [CrossRef]
8. Antonia SJ, Villegas A, Daniel D, et al. Durvalumab after chemoradiotherapy in stage III non-small-cell lung cancer. *N Engl J Med*. 2017;377(20):1919-1929. [CrossRef]
9. Saad MB, Hong L, Aminu M, et al. Predicting benefit from immune checkpoint inhibitors in patients with non-small-cell lung cancer by CT-based ensemble deep learning: a retrospective study. *Lancet Digit Health*. 2023;5(7):e404-e420. [CrossRef]
10. Thomas A, Pattanayak P, Szabo E, Pinsky P. Characteristics and outcomes of small cell lung cancer detected by CT screening. *Chest*. 2018;154(6):1284-1290. [CrossRef]
11. Ligero M, Garcia-Ruiz A, Viaplana C, et al. A CT-based radiomics signature is associated with response to immune checkpoint inhibitors in advanced solid tumors. *Radiology*. 2021;299(1):109-119. [CrossRef]
12. Chen W, Wang R, Ma Z, et al. A delta-radiomics model for preoperative prediction of invasive lung adenocarcinomas manifesting as radiological part-solid nodules. *Front Oncol*. 2022;12:927974. [CrossRef]
13. Kaul V, Enslin S, Gross SA. History of artificial intelligence in medicine. *Gastrointest Endosc*. 2020;92(4):807-812. [CrossRef]
14. Seymour L, Bogaerts J, Perrone A, et al. iRECIST: guidelines for response criteria for use in trials testing immunotherapeutics. *Lancet Oncol*. 2017;18(3):e143-e152. [CrossRef]
15. van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational radiomics system to decode the radiographic phenotype. *Cancer Res*. 2017;77(21):e104-e107. [CrossRef]
16. A. Plonska PP. MLJAR. 2021. [CrossRef]
17. Taieb J, Svrcek M, Cohen R, Basile D, Tougeron D, Phelip JM. Deficient mismatch repair/microsatellite unstable colorectal cancer: diagnosis, prognosis and treatment. *Eur J Cancer*. 2022;175:136-157. [CrossRef]
18. Li N, Wan Z, Lu D, Chen R, Ye X. Long-term benefit of immunotherapy in a patient with squamous lung cancer exhibiting mismatch repair deficient/high microsatellite instability/high tumor mutational burden: a case report and literature review. *Front Immunol*. 2023;13:1088683. [CrossRef]
19. Mu W, Jiang L, Zhang J, et al. Non-invasive decision support for NSCLC treatment using PET/CT radiomics. *Nat Commun*. 2020;11(1):5228. [CrossRef]
20. Le NQK, Kha QH, Nguyen VH, Chen YC, Cheng SJ, Chen CY. Machine learning-based radiomics signatures for EGFR and KRAS

mutations prediction in non-small-cell lung cancer. *Int J Mol Sci*. 2021;22(17):9254. [CrossRef]

21. Liu G, Xu Z, Ge Y, et al. 3D radiomics predicts EGFR mutation, exon-19 deletion and exon-21 L858R mutation in lung adenocarcinoma. *Transl Lung Cancer Res*. 2020;9(4):1212-1224. [CrossRef]
22. He L, Huang Y, Yan L, Zheng J, Liang C, Liu Z. Radiomics-based predictive risk score: a scoring system for preoperatively predicting risk of lymph node metastasis in patients with resectable non-small cell lung cancer. *Chin J Cancer Res*. 2019;31(4):641-652. [CrossRef]
23. Shu Y, Xu W, Su R, et al. Clinical applications of radiomics in non-small cell lung cancer patients with immune checkpoint inhibitor-related pneumonitis. *Front Immunol*. 2023;14:1251645. [CrossRef]
24. Yoon J, Suh YJ, Han K, et al. Utility of CT radiomics for prediction of PD-L1 expression in advanced lung adenocarcinomas. *Thorac Cancer*. 2020;11(4):993-1004. [CrossRef]
25. Trebeschi S, Drago SG, Birkbak NJ, et al. Predicting response to cancer immunotherapy using noninvasive radiomic biomarkers. *Ann Oncol*. 2019;30(6):998-1004. [CrossRef]
26. Faes L, Wagner SK, Fu DJ, et al. Automated deep learning design for medical image classification by health-care professionals with no coding experience: a feasibility study. *Lancet Digit Health*. 2019;1(5):e232-e242. [CrossRef]
27. Fu XY, Mao XL, Wu HW, et al. Development and validation of LightGBM algorithm for optimizing of Helicobacter pylori antibody during the minimum living guarantee crowd based gastric cancer screening program in Taizhou, China. *Prev Med*. 2023;174:107605. [CrossRef]
28. Liu X, Zhu B, Dai XW, et al. GBDT_KgluSite: an improved computational prediction model for lysine glutarylation sites based on feature fusion and GBDT classifier. *BMC Genomics*. 2023;24(1):765. [CrossRef]
29. Wiesweg M, Mairinger F, Reis H, et al. Machine learning reveals a PD-L1-independent prediction of response to immunotherapy of non-small cell lung cancer by gene expression context. *Eur J Cancer*. 2020;140:76-85. [CrossRef]
30. Wei F, Azuma K, Nakahara Y, et al. Machine learning for prediction of immunotherapeutic outcome in non-small-cell lung cancer based on circulating cytokine signatures. *J Immunother Cancer*. 2023;11(7):e006788. [CrossRef]
31. Jiang J, Hu YC, Liu CJ, et al. Multiple resolution residually connected feature streams for automatic lung tumor segmentation from CT images. *IEEE Trans Med Imaging*. 2019;38(1):134-144. [CrossRef]

Supplementary Table 1. Scanning parameters of two scanners

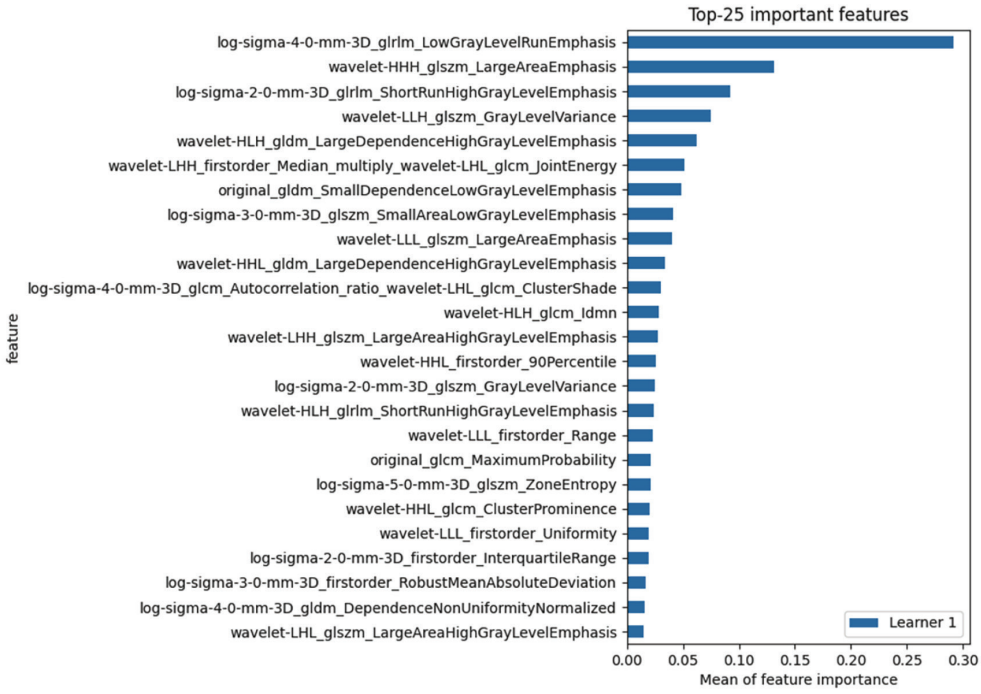
Parameters	GE Discovery CT750 HD	Somatom definition flash
Tube voltage (kVp)	120	120
Tube current (mAs)	200	110
Pitch	0.984:1	1.0
Collimation (mm)	0.625*64	0.6*64
Rotation time (s/rot)	0.5	0.33
SFOV (cm)	50	50
Slice thickness of reconstruction (mm)	1.25	1
Slice interval of reconstruction (mm)	1.25	1
Reconstruction algorithm	STND	Medium sharp

kVp, kilovoltage peak; mAs, milliampere-seconds; SFOV, scan field of view; STND, standard reconstruction algorithm.

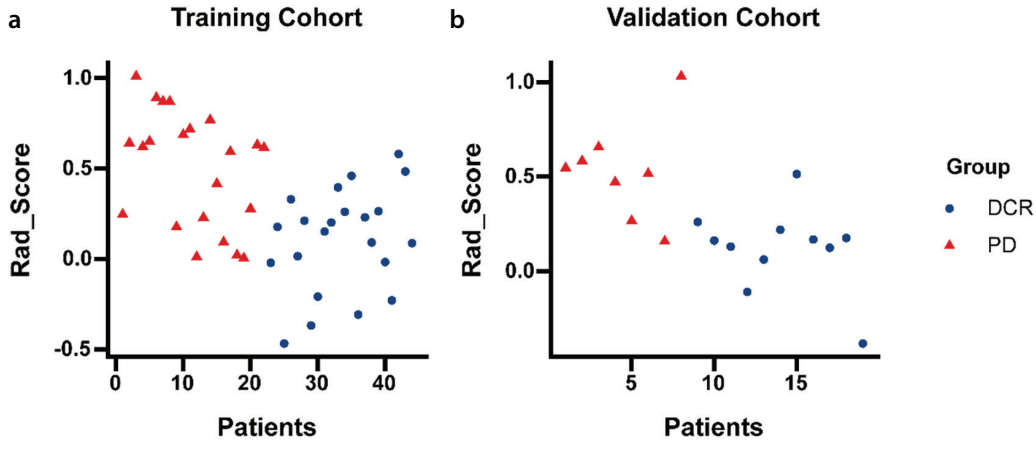
Supplementary Table 2. Evaluation of tumor response to immunotherapy

Tumor response	All patients (n = 63)
CR	0
PR	25
SD	8
PD	30 (47.6%)
DCR (CR + PR + SD)	33 (52.4%)

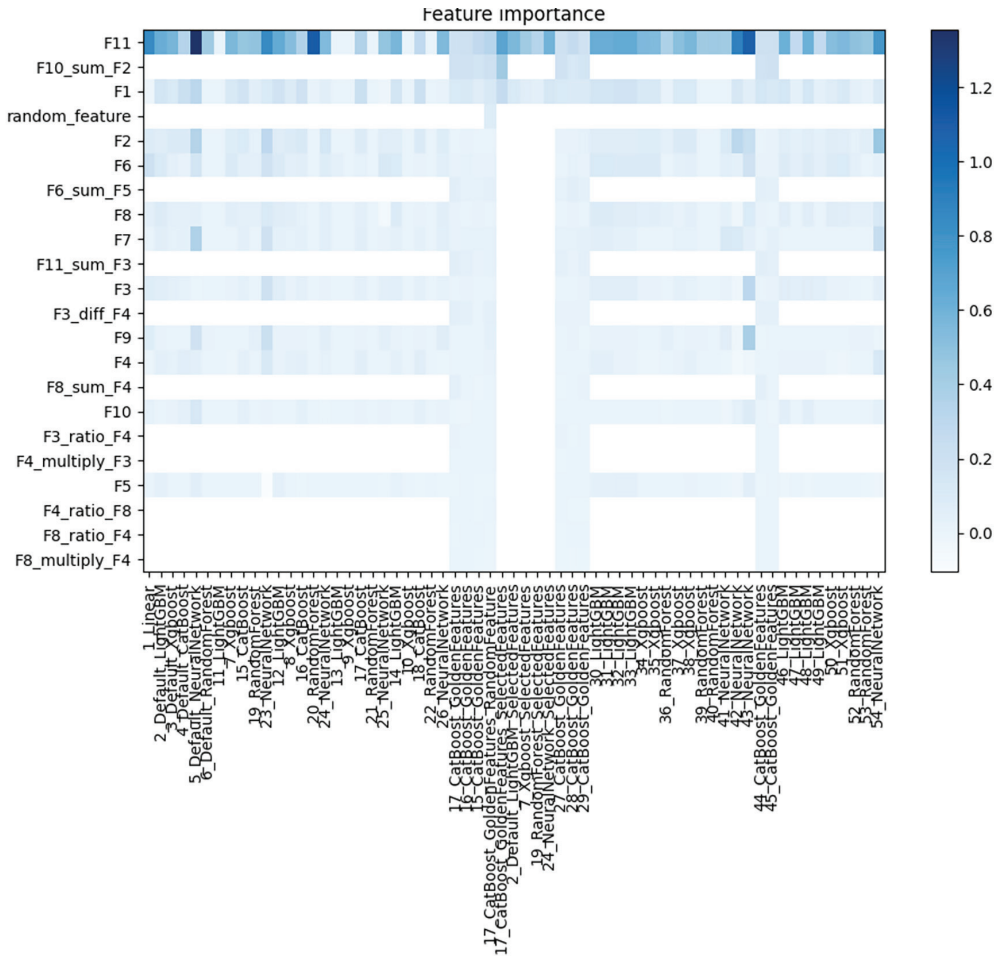
CR, complete response; PR, partial response; SD, stable disease; PD, progressive disease; DCR, disease control rate.



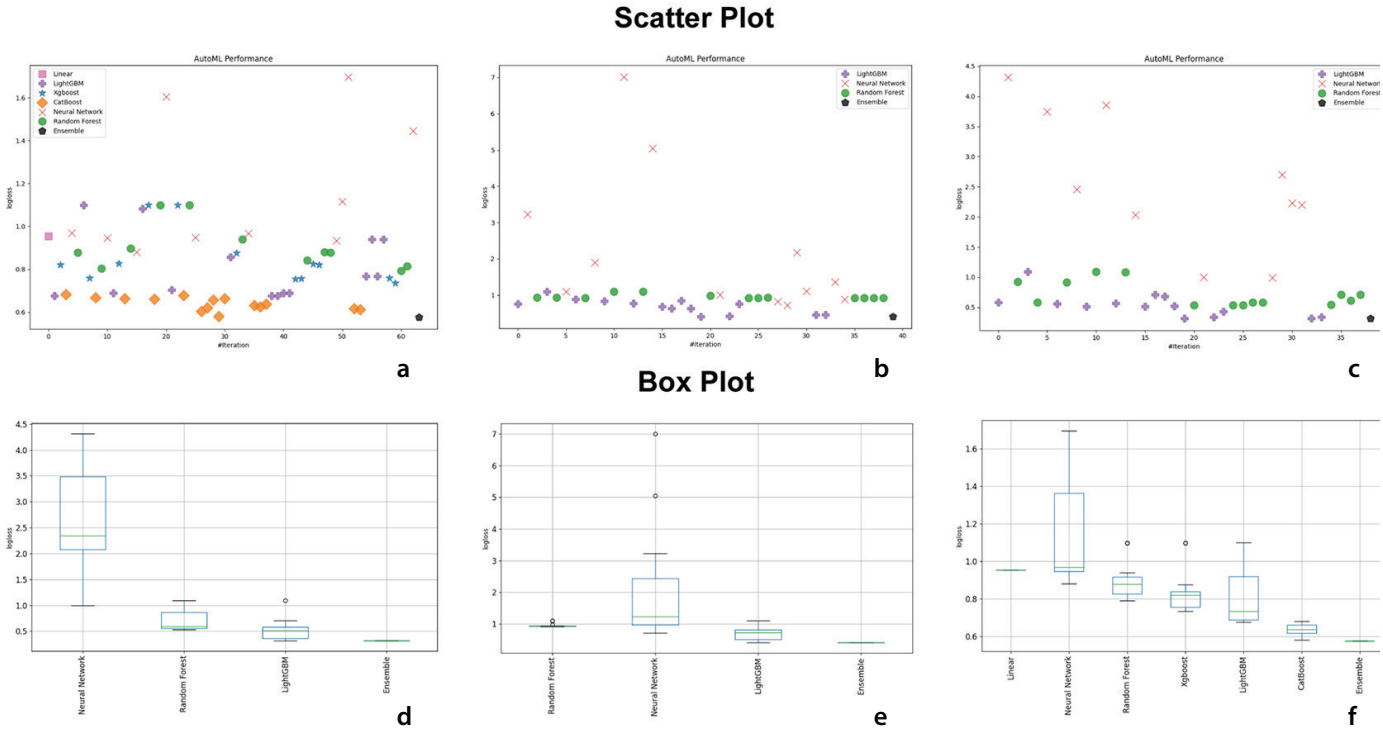
Supplementary Figure 1. Top 25 important radiomics features selected by LightGBM algorithm. LightGBM, light gradient-boosting machine.



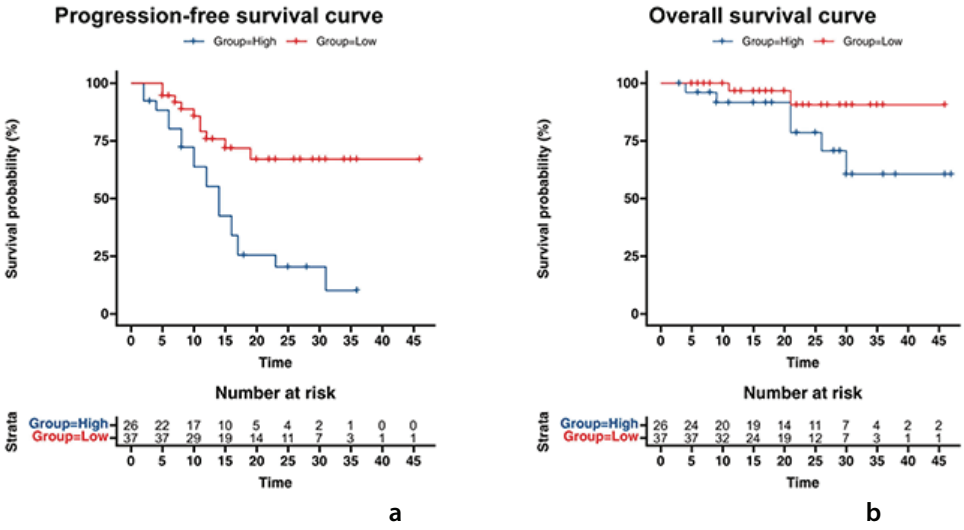
Supplementary Figure 2. The rad scores of patients in DCR and PD subgroups. **(a)** The training cohort; **(b)** the validation cohort. DCR, disease control rate; PD, progressive disease.



Supplementary Figure 3. Predictive clinical features generated by ensemble algorithm.



Supplementary Figure 4. The performance of the detection models in the training cohort. **(a, d)** clinical model; **(b, e)** radiomics model; **(c, f)** fusion model.



Supplementary Figure 5. Survival analyses in different groups of disease progression risk classified by the radiomics model. **(a)** Progression-free survival in different groups of Rad scores ($P < 0.01$); **(b)** Overall survival in different groups of Rad scores ($P = 0.056$).