European Radiology (2023) 33:3918–3930
https://doi.org/10.1007/s00330-022-09337-7

CHEST

A pre-treatment CT-based weighted radiomic approach combined with clinical characteristics to predict durable clinical benefits of immunotherapy in advanced lung cancer

Zhenchen Zhu[1,2] · Minjiang Chen[3] · Ge Hu[1] · Zhengsong Pan[1,2] · Wei Han[4] · Weixiong Tan[5] · Zhen Zhou[5] · Mengzhao Wang[3] · Li Mao[5] · Xiuli Li[5] · Xin Sui[1] · Lan Song[1] · Yan Xu[3] · Wei Song[1] · Yizhou Yu[6] · Zhengyu Jin[1]

Received: 17 June 2022 / Revised: 8 September 2022 / Accepted: 29 November 2022 / Published online: 14 December 2022
© The Author(s), under exclusive licence to European Society of Radiology 2022

Abstract

Objectives To develop a pre-treatment CT-based predictive model to anticipate inoperable lung cancer patients' progression-free survival (PFS) to immunotherapy.

Methods This single-center retrospective study developed and cross-validated a radiomic model in 185 patients and tested it in 48 patients. The binary endpoint is the durable clinical benefit (DCB, PFS ≥ 6 months) and non-DCB (NDCB, PFS < 6 months). Radiomic features were extracted from multiple intrapulmonary lesions and weighted by an attention-based multiple-instance learning model. Aggregated features were then selected through L2-regularized ridge regression. Five machine-learning classifiers were conducted to build predictive models using radiomic and clinical features alone and then together. Lastly, the predictive value of the model with the best performance was validated by Kaplan-Meier survival analysis.

Results The predictive models based on the weighted radiomic approach showed superior performance across all classifiers (AUCs: 0.75–0.82) compared with the largest lesion approach (AUCs: 0.70–0.78) and the average sum approach (AUCs: 0.64–0.80). Among them, the logistic regression model yielded the most balanced performance (AUC = 0.87 [95%CI 0.84–0.89], 0.75 [0.68–0.82], 0.80 [0.68-0.92] in the training, validation, and test cohort respectively). The addition of five clinical characteristics significantly enhanced the performance of radiomic-only model (train: AUC 0.91 [0.89–0.93], p = .042; validation: AUC 0.86 [0.80–0.91], p = .011; test: AUC 0.86 [0.76–0.96], p = .026). Kaplan-Meier analysis of the radiomic-based predictive models showed a clear stratification between classifier-predicted DCB versus NDCB for PFS (HR = 2.40–2.95, p < 0.05).

Conclusions The adoption of weighted radiomic features from multiple intrapulmonary lesions has the potential to predict long-term PFS benefits for patients who are candidates for PD-1/PD-L1 immunotherapies.

Key Points

- *Weighted radiomic-based model derived from multiple intrapulmonary lesions on pre-treatment CT images has the potential to predict durable clinical benefits of immunotherapy in lung cancer.*
- *Early line immunotherapy is associated with longer progression-free survival in advanced lung cancer.*

Zhenchen Zhu and Minjiang Chen contributed equally to this work.

✉ Lan Song
songl@pumch.cn

✉ Yan Xu
maraxu@163.com

✉ Wei Song
cjr.songwei@vip.163.com

[1] Department of Radiology, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100730, China

[2] 4+4 Medical Doctor Program, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100730, China

[3] Department of Respiratory and Critical Care Medicine, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100730, China

[4] Department of Epidemiology and Biostatistics, Institute of Basic Medicine Sciences, Chinese Academy of Medical Sciences and School of Basic Medicine, Peking Union Medical College, Beijing 100005, China

[5] Deepwise Artificial Intelligence (AI) Lab, Beijing Deepwise & League of PhD technology Co. Ltd, Beijing 100080, China

[6] Department of Computer Science, The University of Hong Kong, Hong Kong, China

Keywords Immunotherapy · Immune checkpoint inhibitors · Lung neoplasms · Machine learning · Progression-free survival

Abbreviations

a-MIL	Attention-based multiple-instance learning
AUC	Area under the curve
CT	Computed tomography
DCB	Durable clinical benefit
DICOM	Digital Imaging and Communications in Medicine
EGFR	Epidermal growth factor receptor
HR	Hazard ratio.
ICC	Intraclass correlation coefficient
ICI	Immune checkpoint inhibitor
K-M	Kaplan–Meier
KRAS	Kirsten rat sarcoma 2 viral oncogene homologue
LD	Linear discriminant
LL	Largest lesion
ML	Average-multiple-lesion
MLP	Multilayer perception
NCCN	National Comprehensive Cancer Network
NDCB	Nondurable clinical benefit
NSCLC	Non-small-cell lung cancer
PD-1	Programmed cell death protein 1
PD-L1	Programmed cell death protein ligand 1
PFS	Progression-free survival
RECIST	Response evaluation criteria in solid tumors
SVM	Support vector machines
VOI	Volume of interest
WL	Weighted-multiple-lesion
Xgboost	Extreme gradient boosting

Introduction

The rapid development of immune checkpoint inhibitor (ICI) agents targeting programmed cell death protein 1 (PD-1) or programmed cell death protein ligand 1 (PD-L1) has granted immunotherapy a key role in the treatment of advanced lung cancer in the past ten years [1]. The National Comprehensive Cancer Network (NCCN) has recommended PD-1/PD-L1 blockade therapy for locally advanced and metastatic non-small-cell lung cancer (NSCLC) without targetable genetic mutations [2]. Nonetheless, the beneficial outcome only exists in 15–40% of patients based on previous literature [3, 4]. The need for identifying more efficient predictive biomarkers of immunotherapy responses is therefore crucial.

PD-L1 expression status has been clinically used to select candidates for PD-1/PD-L1 ICIs, but its efficacy as a predictive biomarker is controversial, which is partially due to the quantification nonuniformity and intratumoral heterogeneity [5–9]. In addition, it is an invasive procedure that is not suitable for all patients. Therefore, a noninvasive biomarker is still needed for the precise stratification of patients receiving immunotherapy.

Poor prognosis was associated with several clinical features, such as late lines of immunotherapy and the presence of distant metastases before the treatment, but those findings were based on relatively small cohorts [10, 11]. Computed tomography (CT) image-based radiomics have shown promising results in evaluating tumor responses to immunotherapy, but challenges remain in the aggregation of predictions made at a lesion level to predict a patient-level outcome [12]. Assigning the same patient response to all lesions is a typical lesion-wise radiomic analysis yet it fails to account for effects induced by the unique immune-related response, specifically, dissociated responses [13]. Other patient-level approaches include selecting the largest lesion or averaging by the mean, but the outcomes are far from satisfactory [11, 14]. Chang et.al used a tumor volume-based weighted radiomic features to establish the patient-level outcome in brain metastases with preliminary success [15]. However, this weighting strategy does not consider factors other than tumor size. Recent studies show that an attention-based multiple instance learning (a-MIL) technique may help build predictive models by attributing more weights to the most relevant subregions associated with a specific classification task [16–19]. Therefore, we aimed to explore whether using the adaptively weighted sum of radiomic features from multiple intrapulmonary lesions on the pre-treatment CT scans can enhance the performance of radiomic models for predicting long-term progression-free survival (PFS $\geq$ 6 months) benefit of immunotherapy. Furthermore, we wanted to examine the complementary ability of clinical variables to the radiomic model.

Materials and methods

Patient population and clinical data collection

This single-center study retrospectively reviewed a total of 309 patients at our hospital from June 2015 to November 2020 with pathologically confirmed advanced lung cancer treated with at least one cycle of either PD-1 or PD-L1 ICI therapies. The inclusion and exclusion criteria are detailed in Fig. 1. The final 233 patients were randomly split into a training (n = 185) and an internal test cohort (n = 48). Patient clinical data were collected through electronic medical records and are displayed in Table 1. The endpoint of our study was defined as the durable clinical benefit (DCB: PFS $\geq$ 6 months) or the non-DCB (NDCB: PFS < 6 months) group. PFS was defined as

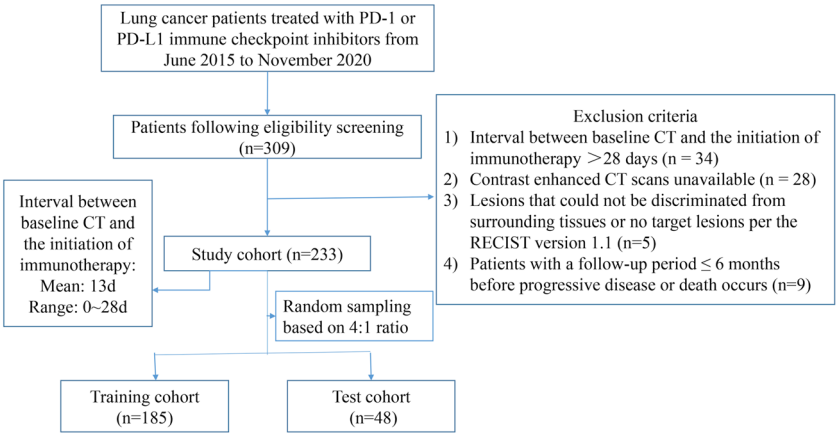


Fig. 1 Flow chart shows inclusion/exclusion and data split for the single-center cohort. Abbreviations: PD-1, programmed cell death protein-1; PD-L1, programmed cell death protein ligand-1; RECIST, response evaluation criteria in solid tumors

the time between the initiation of ICI to the progression of lung cancer or the death of the patient. The progression status was retrieved from the medical records and approved by a senior pulmonologist (M.C.) and a senior radiologist (X.S.) according to the response evaluation criteria in solid tumors (RECIST) version 1.1. Detailed methods for quantifying PD-L1 expression are illustrated in Supplementary Note 1.

Table 1 The definitions and scoring rules for clinical characteristics

Clinical feature	**Definition**	**Scoring**
Age	Age of patients	0, ≤ 60 years old; 1, > 60 years old
Sex	Gender of patients	0, male; 1, female
Smoking history	The history of smoking cigarettes	0, non-smokers; 1-smokers (ever smoker or current smoker)
Current smoker	Have a smoking history within the past 6 months.	0, no; 1, yes
Clinical stage	The clinical TNM stage was determined according to the 8th edition of the American Cancer Society guideline for lung cancer staging.	0, stage III; 1, stage IV
ICI treatment	If multiple PD-1/PD-L1 ICI treatments are involved during the disease course, the first one is used. They are divided into three groups: pembrolizumab, nivolumab, and other PD-1/PD-L1 inhibitors.	Pembrolizumab: 0, no; 1, yes Nivolumab: 0, no;1, yes
Line of therapy	A line of therapy consists of ≥ 1 complete cycle of a single agent, a regimen consisting of a combination of several drugs, or a planned sequential therapy of various regimens.	1, First line; 2, Second line; 3, Third line or more
Chemotherapy	Whether or not the ICI regimen involves the use of chemotherapy.	0, no; 1, yes
Pathology	Pathology subtypes are divided into three groups: adenocarcinoma, squamous cell carcinoma, and other subtypes of lung cancer (small-cell lung cancer, large-cell lung cancer, and adenosquamous carcinoma).	Adenocarcinoma: 0, no;1, yes Squamous cell carcinoma:0, no;1, yes
Driver gene mutation	EGFR gene mutation	0, no; 1, yes
	KRAS gene mutation	0, no; 1, yes
	ROS1 gene mutation	0, no; 1, yes
Metastasis location	Pulmonary	0, no; 1, yes
	Pleural	0, no; 1, yes
	Brain	0, no; 1, yes
	Bone	0, no; 1, yes
	Adrenal gland	0, no; 1, yes
	Liver	0, no; 1, yes
	Others (thyroid gland, prostate, distant lymph node, and etc.)	0, no; 1, yes

Note: Other PD-1/PD-L1 ICI agents include atezolizumab, durvalumab, tislelizumab, penpulimab, and sintilimab. Abbreviations: *PD-1*, programmed cell death protein 1; *PD-L1*, programmed cell death protein ligand 1; *ICI*, immune checkpoint inhibitor; *EGFR*, epidermal growth factor receptor; *KRAS*, Kirsten rat sarcoma 2 viral oncogene homologue; *ROS1*, ROS proto-oncogene 1

Image acquisition and lesion delineation

Contrast-enhanced chest CT scans were carried out from the lung apex to the lung base using various sets of CT scanners as shown in Table 2. The contrast-enhanced scans were acquired at 35 s after the injection of 80–100 mL of nonionic contrast material (Ultravist 300, Bayer Schering Pharma AG; or Ioversol 320, Hengrui Pharmaceutical Co., Ltd) intravenously at a rate of 2.5 mL/s.

The anonymized original Digital Imaging and Communications in Medicine (DICOM) images were normalized and standardized before being imported into the Dr. Wise research platform, on which lesions were automatically delineated using algorithms described in the previous literature [20]. The results were confirmed and modified on axial images slice by slice with mediastinal (width: 450 HU, level: 50 HU) and lung (width: 1200 HU, level: −600 HU) window settings by a senior thoracic radiologist (L.S.) without knowledge of response status. The volume of interest (VOI) was drawn based on the criteria described in Supplementary Note 2 and Supplementary Fig. 1. A maximum of five lesions were retained for each patient based on their longest diameters. For the purpose of reliability analysis, a randomly selected 30 cases were drawn by a second senior radiologist (W.S.).

Feature extraction and aggregation methods

A total of 1454 radiomic features were extracted for each VOI using the PyRadiomics (version 3.0.1) package in Python (version 3.8). More details can be found in Supplementary Note 3.

All features were normalized before being aggregated to form the patient-level radiomic features in the following three ways:

a. Largest lesion approach (LL): radiomic feature of the largest (3D diameter) lesion.
b. Average-multiple-lesion approach (ML): average of summed radiomic features of up to five lesions.
c. Weighted-multiple-lesion approach (WL): sum of adaptively weighted radiomic feature of up to five lesions. The weighted coefficients were determined by an a-MIL model that was developed in the training cohort. The weight coefficients for lesions in the test cohort were generated by the attention-based algorithm embedded in the a-MIL model. A detailed illustration of this technique is shown in Supplementary Note 4 and the codes in Python were available on GitHub at https://github.com/zhjtwx/immunity_WL for reproducibility purposes.

Feature selection and model development

The feature selection and model development were conducted in the training cohort. The patient-level radiomic features underwent three preliminary procedures in a sequential order as shown in Fig. 2d. Firstly, the features that were robust to changes in variations in contour delineation (intraclass correlation coefficient (ICC) of interobserver variability > 0.8) were retained. Secondly, features that were significantly related to the clinical outcome (p value < 0.05 in the Mann–Whitney U test

Table 2 Parameters of CT scanners

	Parameters	Peking Union Medical College Hospital				
CT system information	CT scanner system	Dual Source CT (Siemens Healthcare)		Spectral CT (Discovery CT750 HD scanner, GE Medical Systems)	Spectral CT (IQon CT, PHILIPS)	64-channel CT (Aquilion 64 CT, Toshiba)
		Somatom Definition Flash	Somatom Force			
	Number of patients	101	94	22	8	8
CT scan parameters	Tube voltage	120 kVp		120 kVp	120 kVp	120 kVp
	Tube current	Variable tube current with automatic tube-current modulation activated		Variable tube current with automatic tube-current modulation activated	Variable tube current with automatic tube-current modulation activated	Variable tube current with automatic tube-current modulation activated
	Rotation time	0.5 s		0.6 s	0.5 s	0.5 s
	Detector collimation	64 × 0.6 mm		64 × 0.625 mm	64 × 0.625 mm	64 × 0.625 mm
	Pitch	1.2		0.984	1.2	0.984
	Arterial phase	35 s after injection		35 s after injection	35 s after injection	35 s after injection
	Image matrix	512 × 512		512 × 512	512 × 512	512 × 512
	Field of view	350 × 350 mm		350 × 50 mm	350 × 350 mm	350 × 350 mm
	Reconstruction slice thickness/slice increment	1 mm/1 mm		0.625 mm/0.625 mm	1 mm/1 mm	5 mm/5 mm
	Reconstruction algorithm	standard resolution		standard resolution	standard resolution	standard resolution

analysis) were remained. Lastly, Pearson's correlation coefficient (PCC) of each of the two features was calculated and for the correlated pair (PCC ≥ 0.85), the one with the lower p value was remained. Both the radiomic and clinical features were then fed to the L2-regularized ridge-embedded logistic regression (ridge regression) to select the representative features (coefficient > 0.001) that were associated with the clinical outcome. Unsupervised hierarchical clustering was performed on these representative radiomic features using heatmaps to comprehend their structure (Fig. 2e).

The radiomic model was then built based on the selected radiomic features using five machine learning classifiers: logistic regression (LR), support vector machines (SVM), extreme gradient boosting (Xgboost), multilayer perception (MLP), and linear discriminant (LD). The fivefold cross-validation technique was applied and the average performance in the 4 sub-datasets and 1 sub-dataset from five iterations was reported as the training and validation performance respectively. Similar approaches were used to build the integrated model with the additional selected clinical features. All models were validated in the test cohort. The entire workflow is depicted in Fig. 2.

Statistical analysis

Differences in all variables between the DCB and the NDCB were assessed using the Mann–Whitney U test for continuous variables and the chi-square test or Fisher's exact test for categorical variables as appropriate. The reliability of segmentation was analyzed using the Dice similarity coefficient, and for radiomic features, the ICC and Bland-Altman plots were used. The Gradient weighted Class Activation Mapping was used to visualize the representative radiomic features.

The diagnostic performance was evaluated by classification sensitivity, specificity, accuracy, F1 score, positive predictive value, negative predictive value, and area under the curve (AUC). A two-sided 95% confidence interval for AUC was constructed following the approach of Hanley and McNeil [21]. Performance among different models was compared with the Delong test. Calibration curves and decision curve analysis were performed to evaluate the predictive accuracy and clinical utility of the models.

The Kaplan–Meier (K-M) survival curve method and Cox proportional hazards model were used to analyze PFS. To generate a binary classification, the cutoff thresholds for the prediction probabilities generated by all models were

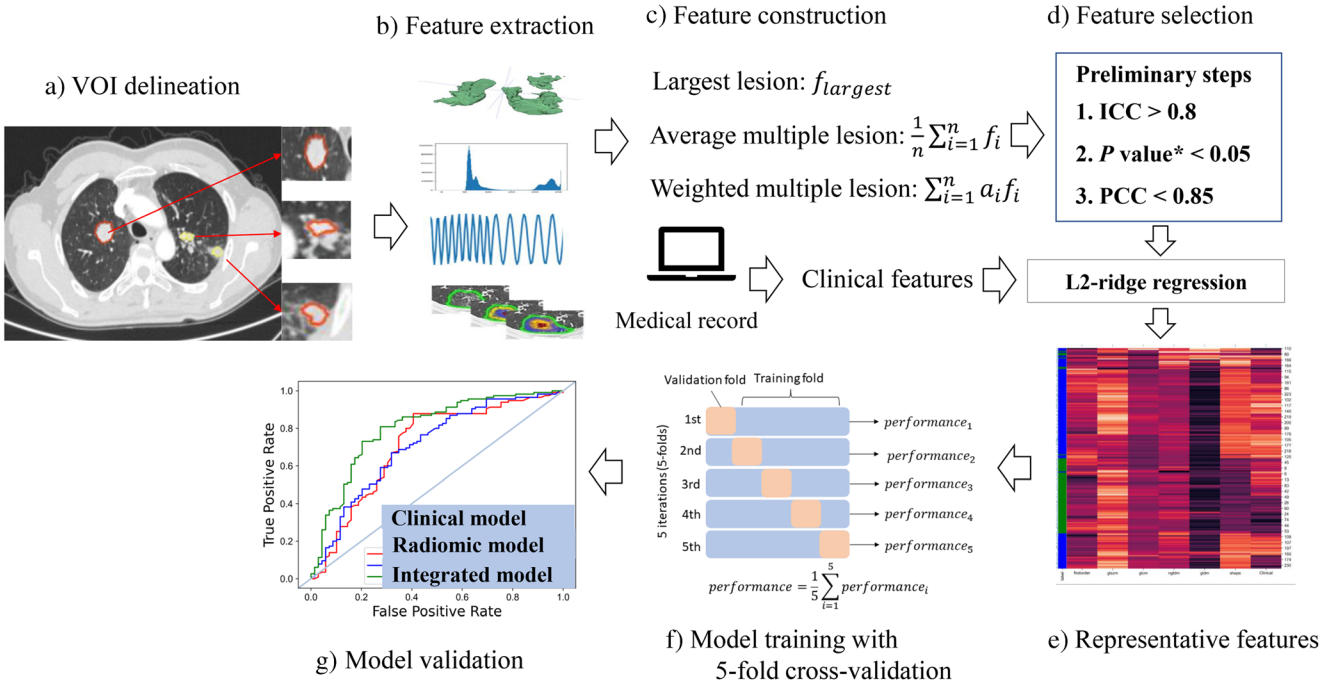


Fig. 2 Workflow of model construction. **a** Multiple lesions were delineated on the Deep-wise labeling system online. **b** Radiomic features were extracted from multiple lesions. **c** The three patient-level analyses were performed with (1) radiomic feature from the largest lesion (LL), (2) average radiomic features from multiple lesions per patient (ML), and (3) weighted sum of radiomic features from multiple lesions per patient (WL). See the "Materials and methods" section for details. **d** The robust and uncorrelated radiomic features were then selected through three preliminary steps. The L2-ridge regression was the last feature selection step to reduce redundancy. **e** Representative features were displayed using unsupervised hierarchical clustering. **f** The model was cross-validated in the training cohort and then validated in the independent test cohort. **g** Receiver operating characteristic (ROC) curves were drawn. *: denotes the p value of radiomic features between the durable clinical benefit (DCB) and the non-DCB group in the training cohort examined by the Mann–Whitney U test. Abbreviations: ICC, intraclass correlation coefficient; PCC, Pearson's correlation coefficient

established using the maximum Youden index in the training cohort. Different curves were compared using the log-rank test. A subgroup analysis was performed to evaluate the model's stratification ability in patients using pembrolizumab. A two-sided p value < 0.05 was used to indicate statistical significance throughout the study. All statistical analyses were performed with the R statistical package.

Results

Patient cohort

Patients with DCB account for 62.9% and 64.9% of the patients in the training and test cohort respectively. The median PFS in the entire cohort was 7.7 months (training: 7.6 months; test: 8.4 months). The data for the PD-L1 expression level were available in 89 (38% of the entire cohort) patients. The clinical and demographic characteristics of the patients in our analysis are summarized in Tables 3 and 4. There were no significant differences in the demographic and clinical characteristics between the two cohorts ($p > 0.05$). Early line therapy, KRAS genetic mutation, and the combination of chemotherapy were significantly associated with DCB, while the presence of bone metastasis before immunotherapy was associated with NDCB. For chemotherapy agents, compared with single-drug, the dual-drug regimen was significantly associated with better clinical outcomes in the test cohort (Supplementary Table 1). Nevertheless, no single chemotherapy regimen showed remarkable superiority over

Table 3 Demographic characteristics and the therapy regimen of patients in the analysis

Characteristics		Training cohort n = 185		p value	Test cohort n = 48		p value
		NDCB	DCB		NDCB	DCB	
Age, median (range)		64 (34–78)	64 (40–79)	.252	60 (36–72)	63 (48–77)	.237
Sex	Male	49 (71)	90 (78)	.317	12 (71)	24 (77)	.731
	Female	20 (29)	26 (22)		5 (29)	7 (23)	
Smoking history	Non-smokers	31 (45)	38 (33)	.098	8 (47)	12 (39)	.760
	Smokers	38 (55)	78 (67)		9 (53)	19 (61)	
Current smoker	No	48 (70)	69 (60)	.169	13 (77)	19 (61)	.350
	Yes	21 (30)	47 (40)		4 (23)	12 (39)	
Clinical stage	III	10 (15)	29 (25)	.090	0 (0)	10 (32)	**.009***
	IV	59 (85)	87 (75)		17 (100)	21 (68)	
ICI treatment	Pembrolizumab	33 (48)	66 (57)	.062	7 (41)	21 (68)	.203
	Nivolumab	20 (29)	17 (15)		6 (35)	6 (19)	
	Others#	16 (23)	33 (28)		4 (24)	4 (13)	
Line of therapy	First	22 (32)	75 (65)	**≤ .001***	6 (35)	22 (71)	**.044***
	Second	31 (31)	33 (28)		7 (41)	7 (23)	
	Third+	16 (23)	8 (7)		4 (24)	2 (6)	
Chemotherapy	No	44 (64)	46 (40)	**.003***	9 (53)	12 (39)	.518
	Yes	25 (36)	70 (60)		8 (47)	19 (61)	
Chemo. agent	AC	9 (36)	31 (44)	.732	1 (13)	5 (26)	**.017***
	AP	1 (4)	2 (3)		0 (0)	2 (11)	
	CE	2 (8)	9 (13)		0 (0)	0 (0)	
	DOC	1 (4)	0 (0)		2 (25)	0 (0)	
	EP	0 (0)	1 (1)		0 (0)	0 (0)	
	GC	0 (0)	1 (1)		1 (13)	0 (0)	
	GP	2 (8)	4 (6)		2 (25)	1 (5)	
	T	1 (4)	1 (1)		1 (13)	0 (0)	
	TC	9 (36)	21 (30)		1 (13)	11 (58)	

Note: Values are expressed as number (%), if not defined otherwise. *: p value < 0.05. #: Other ICI treatments include atezolizumab, durvalumab, tislelizumab, penpulimab, and sintilimab

Abbreviations: *NDCB*, non-durable clinical benefit; *DCB*, durable clinical benefit; *ICI*, immune checkpoint inhibitor; *Chemo.*, chemotherapy; *AC*¸ pemetrexed and carboplatin; *AP*¸ pemetrexed and cisplatin; *CE*¸ carboplatin and etoposide; *DOC*¸ docetaxel; *EP*¸ etoposide and cisplatin; *GC*¸ gemcitabine and carboplatin; *GP*¸ gemcitabine and cisplatin; *T*¸ paclitaxel; *TC*¸ paclitaxel and carboplatin

another. The elevated expression of PD-L1 was associated with epidermal growth factor receptor (EGFR)-wild type and Kirsten rat sarcoma 2 viral oncogene homologue (KRAS) mutation (Supplementary Table 2).

Representative features

There was high agreement between the segmentations drawn by two radiologists (Dice coefficient of 0.89 [95%CI 0.87–0.91] for the largest lesion and 0.90 [95%CI 0.87–0.93] for multiple lesions). Five representative clinical features were identified: age (≤ 60 or > 60), clinical stage (III or IV), bone metastasis, line of therapy (first, second, or third+), and the use of pembrolizumab. Nineteen, twenty-one, and twenty-five radiomic features were selected individually using the LL, ML, and WL approaches. The number of features that remained at every selection step is shown in Supplementary Fig. 2. ICCs and the Bland-Altman plots showed excellent robustness of the selected features (Supplementary Table 3 and Supplementary Fig. 3). The unsupervised clustering analysis of all representative features resulted in three clusters. Features showed differential expression between the DCB and NDCB cases in both cohorts (Supplementary Fig. 4). A complete list of the representative features and their coefficients in the integrated models with different feature construction methods is shown in Supplementary Table 4. In the WL-based integrated model, the most contributable clinical feature

Table 4 Immunopathologic features and metastasis statuses of tumor before the initiation of immunotherapy

Characteristics		Training cohort $n = 185$		p value	Test cohort $n = 48$		p value
		NDCB	DCB		NDCB	DCB	
Pathology	ADC[&]	29 (42)	48 (41)	.759	10 (59)	15(52)	.556
	SCC	33 (48)	52 (45)		7 (41)	16(50)	
	Others^	7 (10)	16 (14)		0 (0)	0(0)	
EGFR mutation	No	58 (84)	106 (91)	.129	11 (65)	28(90)	.051
	Yes	11 (16)	10 (9)		6 (35)	3(10)	
KRAS mutation	No	67 (97)	100 (86)	**.019***	15 (88)	30(97)	.283
	Yes	2 (3)	16 (14)		2 (12)	1(3)	
ROS1 mutation	No	63 (91)	112 (97)	.178	16 (94)	29(94)	1.000
	Yes	6 (9)	4 (3)		1 (6)	2(6)	
PD-L1 expression	TPS < 1%	10 (14)	13 (11)	.113[#]	2 (12)	5(16)	.571[#]
	1% ≤ TPS < 50%	11 (16)	18 (16)		2 (12)	5(16)	
	TPS ≥ 50%	3 (4)	17 (15)		0 (0)	3(10)	
	Unknown	45 (65)	68 (59)		13 (76)	18(58)	
Pulmonary metastasis	No	40 (58)	80 (69)	.130	10 (59)	19(61)	1.000
	Yes	29 (42)	36 (31)		7 (41)	12(39)	
Pleural metastasis	No	46 (67)	79 (68)	.840	11 (65)	19(61)	1.000
	Yes	23 (33)	37 (32)		6 (35)	12(39)	
Brain metastasis	No	59 (85)	108 (93)	.092	14 (82)	27(87)	.686
	Yes	10 (15)	8 (6.9)		3 (18)	4(13)	
Bone metastasis	No	45 (65)	92 (79)	**.034***	11 (65)	28(90)	.051
	Yes	24 (35)	24 (21)		6 (35)	3(10)	
Adrenal gland metastasis	No	59 (85)	97 (84)	.733	14 (82)	25(81)	1.000
	Yes	10 (15)	19 (16)		3 (18)	6(19)	
Liver metastasis	No	60 (87)	109 (94)	.112	13 (77)	27(87)	.428
	Yes	9 (13)	7 (6)		4 (3)	4(13)	
Other metastases	No	58 (84)	102 (88)	.456	15 (88)	27(87)	1.000
	Yes	11 (16)	14 (12)		2 (12)	4(13)	

Note: Values are expressed as number (%), if not defined otherwise. *: p value < 0.05. #: p values are calculated using the cases with known PD-L1 statuses. &: The ADC subtype includes 1 invasive mucinous adenocarcinoma. ^: Other pathology types include small-cell lung cancer ($n = 12$), large-cell lung cancer ($n = 7$), and adenosquamous carcinomas ($n = 4$)

Abbreviations: *NDCB*, non-durable clinical benefit; *DCB*, durable clinical benefit; *ADC*, adenocarcinoma; *SCC*, squamous cell carcinoma; *TPS*, tumor proportion score; *EGFR*, epidermal growth factor receptor; *ROS1*, ROS proto-oncogene 1; *KRAS*, Kirsten rat sarcoma 2 viral oncogene homolog

was the line of therapy, and for radiomic features, GLCM_Correlation and GLDM_SDHGLE showed the largest coefficient in the negative and positive directions respectively (Supplementary Fig. 5).

Comparison of model performance

Among the three aggregation methods, the AUCs of the WL-based models were superior to those of the other two approaches in all classifiers but MLP, in which the performance of the WL-based radiomic model was not significantly better than that of the LL-based radiomic model in the test cohort (Fig. 3).

The WL-based radiomic model with logistic regression classifier yielded the most balanced performance to discern DCB from NDCB with AUCs of 0.87 [0.84–0.89], 0.75 [0.68–0.82], 0.80 [0.68–0.92] in the training, validation, and test cohort respectively (Supplementary Table 5). With the addition of five clinical characteristics, the WL-based integrated model reached a significantly better AUC than the radiomic model and the clinical model, as presented in Table 5 and Fig. 4. The calibration and decision curve analysis curves for the above models are shown in Supplementary Fig. 6 and 7. The performance of the a-MIL model for differentiating DCB from NDCB was given in Supplementary Note 5.

Figure 5 illustrates the discriminability of the log GLDM (SDHGLE) feature and the wavelet GLCM (correlation) feature for representative DCB and NDCB patients before ICI therapy. We observed a higher textural heterogeneity pattern on lesions of the DCB patient compared with the NDCB patient. In addition, heavier weights were attributed to the smaller lesion in most circumstances.

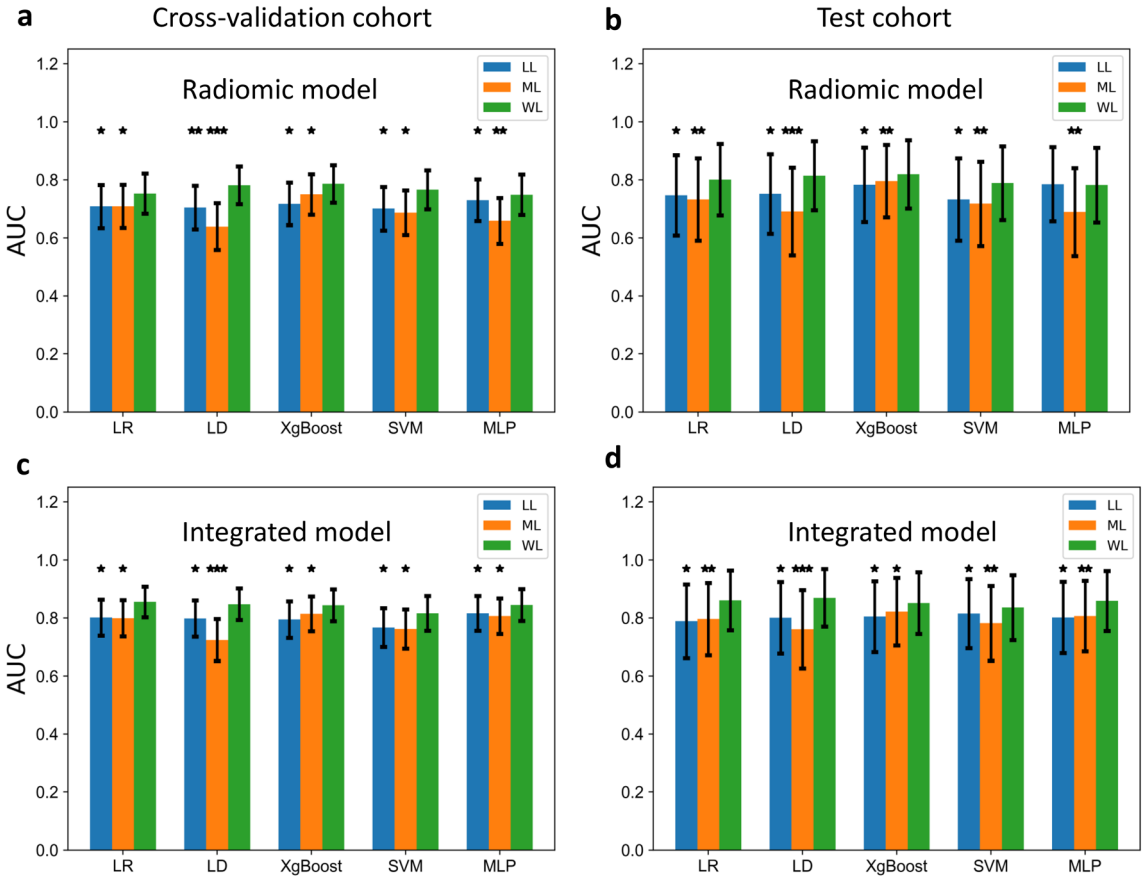


Fig. 3 Comparison of the area under the curves (AUCs) of different predicative models based on three patient-level analyses. **a**, **b** The AUCs of the radiomic models for discerning DCB (PFS ≥ 6 months) from NDCB (PFS < 6 months) in the cross-validation (**a**) and the test cohort (**b**). **c**, **d** The AUCs of the integrated models discerning DCB from NDCB in the cross-validation (**c**) and the test cohort (**d**). The Y axis represents AUCs and the X axis represents different classifiers. The bar in green denotes the model's performance based on WL-based radiomic features. *P* values were obtained by comparing the AUC of the integrated model with the AUCs of the other two models [LL-(blue) and ML-(orange) based radiomic models] using the Delong test. Note: * denotes *p* value < 0.05, ** denotes *p* < 0.01, *** denotes *p* value < 0.001. Abbreviations: LL, largest-lesion approach; ML, average-multiple-lesion approach; WL, weighted-multiple-lesion approach; DCB, durable clinical benefit; NDCB, non-durable clinical benefit; LR, logistic regression; SVM, support vector machines; Xgboost, extreme gradient boosting; MLP, multilayer perception; LD, linear discriminant

Table 5 The performance of the logistic regression-based integrated model was compared with the clinical and the radiomic model in each of three feature construction approaches

DCB vs. NDCB		Training cohort (validation fold)							Test cohort							
Model type	AUC [95%CI]	ACC	F1	SPE	SEN	PPV	NPV	*p* value	AUC [95%CI]	ACC	F1	SPE	SEN	PPV	NPV	*p* value
Clinical	0.71 [0.64, 0.78]	0.77	0.84	0.59	0.88	0.78	0.75	0.047*	0.80 [0.67, 0.92]	0.79	0.83	0.71	0.84	0.84	0.71	0.59
LL_Radiomic	0.71 [0.63, 0.78]	0.67	0.72	0.68	0.67	0.78	0.55	0.016*	0.75 [0.61, 0.89]	0.75	0.80	0.71	0.77	0.83	0.63	0.039*
LL_Integrated	0.80 [0.74, 0.86]	0.80	0.76	0.79	0.80	0.73	0.86	Ref.	0.79 [0.66, 0.92]	0.79	0.79	0.83	0.76	0.81	0.86	Ref.
Clinical	0.71 [0.64, 0.78]	0.77	0.84	0.59	0.88	0.78	0.75	0.051	0.80 [0.67, 0.92]	0.79	0.83	0.71	0.84	0.84	0.71	0.084
ML_Radiomic	0.71 [0.63, 0.78]	0.71	0.68	0.71	0.78	0.63	0.83	0.003**	0.73 [0.59, 0.87]	0.71	0.72	0.94	0.58	0.95	0.55	0.002**
ML_Integrated	0.80 [0.74, 0.86]	0.80	0.75	0.80	0.70	0.78	0.81	Ref.	0.80 [0.67, 0.92]	0.80	0.77	0.80	0.88	0.71	0.92	Ref.
Clinical	0.71 [0.64, 0.78]	0.77	0.84	0.59	0.88	0.78	0.75	0.008**	0.80 [0.67, 0.92]	0.79	0.83	0.71	0.84	0.84	0.71	0.008**
WL_Radiomic	0.75 [0.68, 0.82]	0.78	0.82	0.77	0.79	0.85	0.69	0.011*	0.80 [0.68, 0.92]	0.81	0.86	0.71	0.87	0.84	0.75	0.026*
WL_Integrated	0.86 [0.80, 0.91]	0.83	0.86	0.78	0.85	0.87	0.76	Ref.	0.86 [0.76, 0.96]	0.85	0.89	0.82	0.87	0.90	0.78	Ref.

Note: *p* value*:* the area under the curve (AUC) of the clinical model and the radiomic model were compared to the integrated models using the DeLong test. *: *p* value < 0.05. **: *p* value < 0.01. Abbreviations: *NDCB*, non-durable clinical benefit; *DCB*, durable clinical benefit; *CI*, confidence interval; *ACC*, accuracy; *SPE,* specificity; *SEN,* sensitivity; *PPV*, positive predictive value; *NPV*, negative predictive value; *LL*, largest lesion approach; *ML*, average-multiple-lesion approach; *WL*, weighted-multiple-lesion approach, *Ref.*, reference

Stratified pretreatment PD-L1 expression as a predictor of durable PFS

As illustrated in Supplementary Table 6, the positivity rate for PD-L1 expression was 66% (59 out of 89) if the cut-off was 1%, with an accuracy of 61.8% (55 of 89) and an AUC of 0.57 (95% CI: 0.44–0.70) in differentiating DCB from NDCB. If the cut-off was set at 50%, the positivity rate reached 26% (23 out of 89), with an accuracy of 50.6% (45 of 89) and an AUC of 0.61 (95% CI: 0.49–0.73) in differentiating DCB from NDCB. More than 46% (41 of 89) of patients with low expression of PD-L1 (tumor proportion score < 50%) experienced DCB.

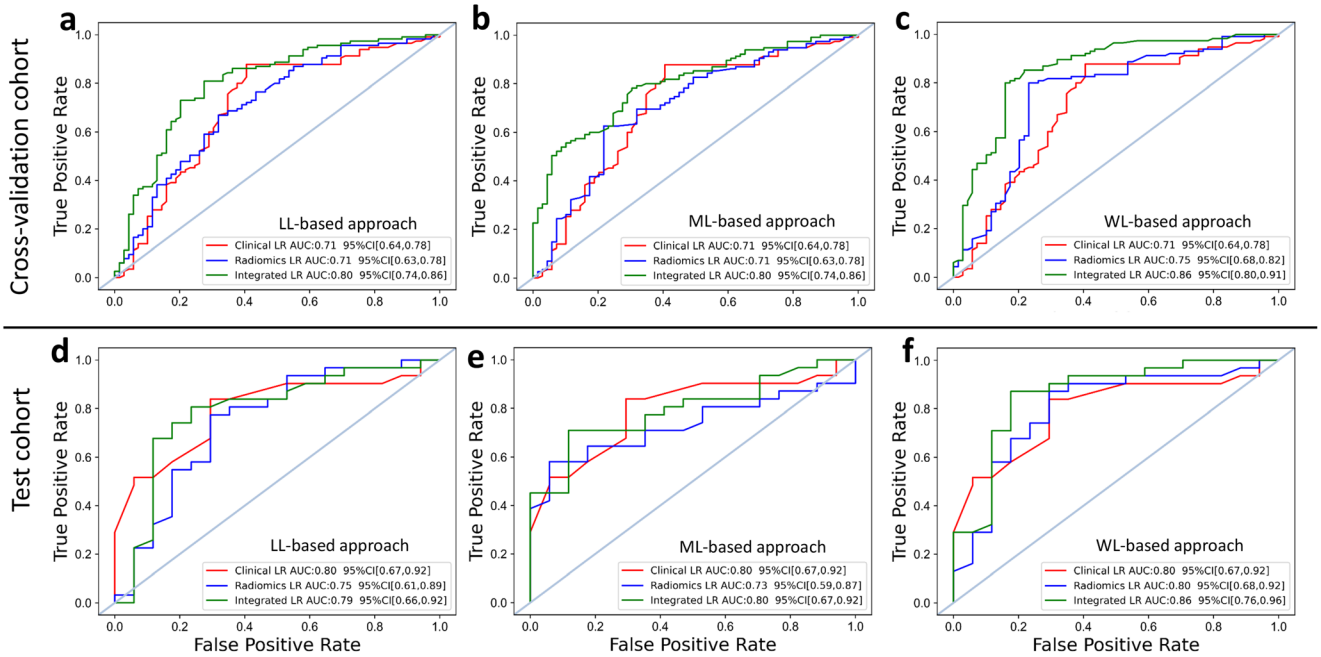
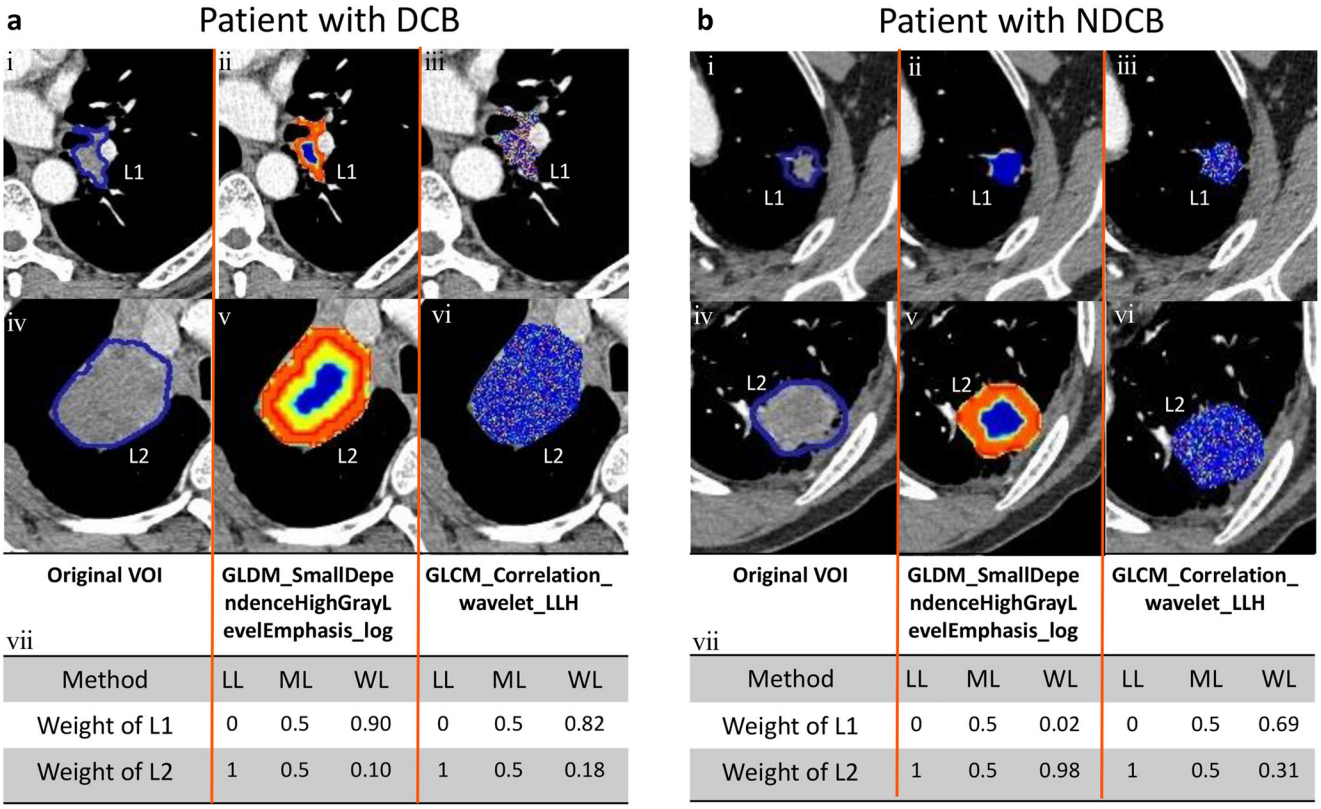


Fig. 4 Displays of the receiver operating characteristic (ROC) curves of clinical, radiomic, and integrated models. **a–c** ROC curves of clinical (red), radiomic (blue) and integrated (green) logistic regression model built with the LL-based radiomic features (**a**), ML-based radiomic features (**b**) and WL-based radiomic features (**c**) for differentiating DCB from NDCB in the cross-validation cohort. **d–f** ROC curves of clinical (red), radiomic (blue) and integrated (green) logistic regression model built with the LL-based radiomic features (**d**), ML-based radiomic features (**e**), and WL-based radiomic features (**f**) for differentiating DCB from NDCB in the test cohort



Method	LL	ML	WL	LL	ML	WL
Weight of L1	0	0.5	0.90	0	0.5	0.82
Weight of L2	1	0.5	0.10	1	0.5	0.18

Method	LL	ML	WL	LL	ML	WL
Weight of L1	0	0.5	0.02	0	0.5	0.69
Weight of L2	1	0.5	0.98	1	0.5	0.31

Fig. 5 Baseline chest contrast-enhanced CT (CECT) images and visualizing heatmaps of class activation in an image of two lung cancer patients presented with durable clinical benefit (DCB) and non-DCB (NDCB). **a** This DCB case used penpulimab (PFS = 19.1 months). **b** This NDCB case used sintilimab (PFS = 5 months). L1 and L2 each represents an individual tumor lesion identified on CECT scans. Both lesions shrank significantly at follow-up 6 weeks after the first course of therapy. (i, iv) represent two original tumor lesions in the mediastinal window of CECT. (ii, v) represent GLDM_SmallDependenceHighGrayLevelEmphasis (SDHGLE) feature heatmaps with corresponding tumor lesions of CECT. (iii, vi) represent GLCM_correlation feature heatmaps with corresponding tumor lesions of CECT. (vii) shows the attributed weight coefficients of corresponding radiomic features of L1 and L2 in the LL (largest lesion), ML (average-multiple-lesion) and WL (weighted-multiple-lesion) approaches, respectively. Note: GLDM-SDHGLE measures the joint distribution of small dependence with higher gray-level values, and a greater value indicates a smaller dependence of higher gray-level values and less homogeneous textures; GLCM-Correlation measures the linear dependency of gray-level of neighbouring pixels, and a higher value indicates a less smooth gradient of the pattern in the image

Predictive ability for PFS of different predictive models

As illustrated in Fig. 6 and Supplementary Table 7, the integrated model showed better performance for predicting PFS than the other two models (HR = 2.90 [95% CI: 2.15–3.84], p = 0.014 in the test cohort). In the pembrolizumab subgroup analysis, a higher score stratified by the radiomic and integrated models was significantly associated with a longer PFS (Supplementary Fig. 8). The results of Cox regression and K-M analysis for the pembrolizumab subgroup are displayed in Supplementary Table 8.

Discussion

In this study, we collected a relatively large cohort of advanced lung cancer patients and constructed models to identify patients who were more likely to obtain durable clinical benefits using PD-1/PD-L1 targeted therapies. In the meantime, we explored the method of weighting the sum of radiomic features from multiple intrapulmonary lesions to construct the predictive models and found that it exhibited superior performance to discriminate DCB from NDCB compared with the conventional approaches. Furthermore, an integrated predictive model was constructed using the WL-based radiomic features and five clinical features, reaching AUCs of 0.86 in both the cross-validation dataset and test cohorts.

Considering the presence of immunotherapy-specific unconventional response patterns, the patient-level radiomic analysis that incorporates features from multiple lesions is gaining more attention [12, 14, 22]. MIL is a useful tool to aggregate features from multiple imaging patches (instances) that represent one bag-level characteristic [23]. Back in 2020, Zhang et al adopted an MIL-based supporter vector machine

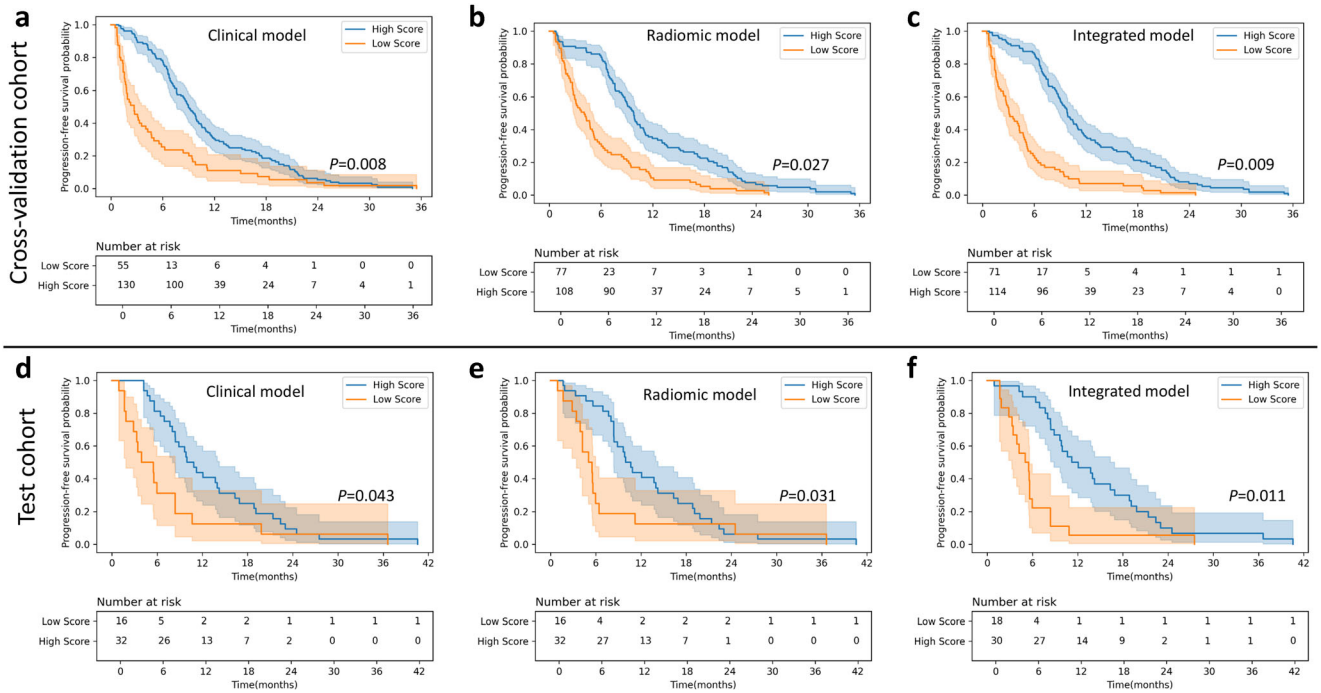


Fig. 6 Kaplan-Meier (KM) progression-free survival (PFS) curve analyses. **a**–**c** KM curves on the cross-validation cohort for model scores generated by (**a**) logistic regression (LR)-based clinical model, (**b**) LR-based radiomic model and (**c**) LR-based integrated model. **d**–**f** KM curves on the test cohort for scores generated by (**d**) LR-based clinical model, (**e**) LR-based radiomic model, and (**f**) LR-based integrated model. All radiomic and the integrated models displayed here were built with the weighted-multiple-lesion (WL)-based patient-level radiomic features. The cutoff threshold of the clinical, radiomic, and integrated model for the PFS risk stratification is 0.51, 0.43, and 0.67 respectively

to identify the survival-related high-risk subregions in magnetic resonance imaging (MRI) scans for glioblastoma [16]. More recently, Li et.al. proposed an attention-based MIL framework to compute weights for each segmented patch in an abundance of chest CT images and identified regions that were most correlated with the assessment of COVID-19 severity [17]. Another study in histopathology by Lu et al aggregated patch-level features into slide-level representations and assigned scores to each patch to represent the significance to the collective slide-level representations for a specific classification (e.g., clear cell renal carcinoma) [18]. Here, we treated each delineated lesion as an instance and adaptively weighed them to represent the patient-level clinical outcome. As illustrated in Fig. 5, the difference in radiomic features between the DCB and NDCB group is distinctive in the smaller lesion, to which a higher weight was attributed. It demonstrates the strength and validity of our method.

The later line of immunotherapy was associated with poor prognosis. A similar finding was reported in Tunali et al's study, in which they argued that multiple systematic treatments induced an "immune-desert" microenvironment that compromised the efficacy of immunotherapy [10, 24]. Currently, using pembrolizumab or atezolizumab as first-line therapy for metastatic lung cancers with high expression of PD-L1 has gained increasing acknowledgment [2, 25]. However, our study recognized that a significant proportion of patients who had low expression of PD-L1 (< 50%) could still reach DCB from PD-1/PD-L1 targeted immunotherapies. It highlights the need to identify a more precise predictive biomarker. Given our predictive model only requires pretreatment CT images and basic clinical information, it can serve as an alternative and noninvasive biomarker to direct personalized therapeutic immunotherapy regimen, especially for those with unknown PD-L1 statuses.

GLDM_SDHGLE and GLCM_Correlation are the two textural features from our integrated model that had the largest coefficients in the positive and negative directions respectively. By visualizing them on the heatmaps, we identified a ring structure surrounding the tumor margin. Previous literature suggested that peritumoral texture features are associated with tumor infiltrating lymphocytes that can predict tumor response to immunotherapy [26–28]. The identified marginal characteristics in our study, though not being biologically validated, may be correlated with the recruitment of active lymphocytes.

The positive association between KRAS-mutated status and immunotherapy's efficacy was identified in our study and in previous literature as well. Chen et al argued that KRAS-mutation induced an inflammatory tumor microenvironment that may result in the elevation of tumor burden [29, 30]. Other studies indicate that this environment also triggers the

elevated expression of PD-L1 although no agreement has been made yet [31, 32]. Our finding suggests the prognostic value of KRAS mutation in PD-1/PD-L1 targeted therapies in lung cancer. Nonetheless, further studies in a larger KRAS-mutated cohort are needed to further warrant this statement.

We acknowledge the limitations of our study. The first is the retrospective nature of this single-center study. Although an internal independent test cohort was adopted, further external validation in a prospective cohort is warranted. Second, we used a relatively small sample size to train a deep learning–based model, although the precise annotations and radiomic features were used to reduce the network complexity. Third, the peritumoral region was not included in our radiomic analysis, which may result in the loss of useful information related to the distribution of tumor-infiltrated lymphocytes around the tumor. Fourth, PD-L1 expression data were unavailable for most patients in our cohort. Combing it with our radiomic signature may enhance the predictive performance of the models. Fifth, we aggregated the radiomic features to give more weight to the most relevant lesion but did not capture inter-lesion heterogeneity and differential patterns of response in patient. Lastly, there is a deficiency of biological validation due to the retrospective nature of our data. Further studies are warranted to help explain the biological significance of the radiomic biomarker.

Conclusion

Our noninvasive predictive model based on the weighted sum of radiomic features from multiple intrapulmonary lesions holds considerable promise as a new approach to bring substantial survival benefits to lung cancer patients who are candidates for immunotherapy.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00330-022-09337-7.

Funding This study has received funding from the CAMS Innovation Fund for Medical Sciences (CIFMS, 2021-I2M-C&T-A-007), the AI + Health Collaborative Innovation Cultivation Project of Beijing Municipal Commission of Science and Technology (No. Z201100005620008), the Scientific and Technological Innovation 2030- New Generation Artificial Intelligence Project of the National Key Research and Development Program of China (No. 2020AAA0109503), the National Natural Science Foundation of China (NSFC No. 82171934), and the 2021 SKY Imaging Research Fund of Chinese International Medical Exchange Foundation (No. Z-2014-07-2101). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Declarations

Guarantor The scientific guarantor of this publication is Zhengyu Jin from Peking Union Medical College Hospital, Department of Radiology.

Conflict of interest The authors of this manuscript declare no relationships with any companies whose products or services may be related to the subject matter of the article.

Statistics Dr. Wei Han (one of the authors) kindly provided statistical advice for this manuscript.

Informed consent Written informed consent was waived by the Institutional Review Board of our institution due to the retrospective nature of the study.

Ethical approval This study was approved by the institutional review board and the ethics committee of Peking Union Medical College Hospital (Beijing, China) (No.S-K196Z), and the requirement for informed patient consent was waived due to the retrospective nature of the study.

Methodology
- retrospective
- diagnostic or prognostic study
- performed at one institution

References

1. Finck A, Gill SI, June CH (2020) Cancer immunotherapy comes of age and looks for maturity. Nat Commun 11(1):3325 https://doi.org/10.1038/s41467-020-17140-5
2. National Comprehensive Cancer Network (2022) NCCN clinical practice guidelines in oncology. Non-small Cell Lung Cancer Version 4.2022. Available via https://www.nccn.org/. Accessed 4 September 2022
3. Sui H, Ma N, Wang Y et al (2018) Anti-PD-1/PD-L1 therapy for non-small-cell lung cancer: toward personalized medicine and combination strategies. J Immunol Res 2018:6984948 https://doi.org/10.1155/2018/6984948
4. Rizvi NA, Hellmann MD, Snyder A et al (2015) Cancer immunology. mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. Science 348:124–128
5. Chen R, Tao Y, Xu X et al (2018) The efficacy and safety of nivolumab, pembrolizumab, and atezolizumab in treatment of advanced non-small cell lung cancer. Discov Med 26:155–166
6. Chen J, Jiang CC, Jin L, Zhang XD (2016) Regulation of PD-L1: a novel role of pro-survival signalling in cancer. Ann Oncol 27:409–416
7. Langer CJ, Gadgeel SM, Borghaei H et al (2016) Carboplatin and pemetrexed with or without pembrolizumab for advanced, non-squamous non-small-cell lung cancer: a randomised, phase 2 cohort of the open-label KEYNOTE-021 study. Lancet Oncol 17:1497–1508
8. Hellmann MD, Paz-Ares L, Bernabe Caro R et al (2019) Nivolumab plus ipilimumab in advanced non-small-cell lung cancer. N Engl J Med 381:2020–2031
9. Yi M, Jiao D, Xu H et al (2018) Biomarkers for predicting efficacy of PD-1/PD-L1 inhibitors. Mol Cancer 17(1):129. https://doi.org/10.1186/s12943-018-0864-3
10. Tunali I, Gray JE, Qi J et al (2019) Novel clinical and radiomic predictors of rapid disease progression phenotypes among lung cancer patients treated with immunotherapy: An early report. Lung Cancer 129:75–79

11. Liu Y, Wu M, Zhang Y et al (2021) Imaging biomarkers to predict and evaluate the effectiveness of immunotherapy in advanced non-small-cell lung cancer. Front Oncol 11:657615. https://doi.org/10.3389/fonc.2021.657615
12. Sun R, Henry T, Laville A et al (2022) Imaging approaches and radiomics: toward a new era of ultraprecision radioimmunotherapy? J Immunother Cancer 10(7):e004848. https://doi.org/10.1136/jitc-2022-004848
13. Ligero M, Garcia-Ruiz A, Viaplana C et al (2021) A CT-based radiomics signature is associated with response to immune checkpoint inhibitors in advanced solid tumors. Radiology 299:109–119
14. Trebeschi S, Drago SG, Birkbak NJ et al (2019) Predicting response to cancer immunotherapy using noninvasive radiomic biomarkers. Ann Oncol 30:998–1004
15. Chang E, Joel MZ, Chang HY et al (2021) Comparison of radiomic feature aggregation methods for patients with multiple tumors. Sci Rep 11(1):9758. https://doi.org/10.1038/s41598-021-89114-6
16. Zhang X, Lu D, Gao P et al (2020) Survival-relevant high-risk subregion identification for glioblastoma patients: the MRI-based multiple instance learning approach. Eur Radiol 30:5602–5610
17. Li Z, Zhao W, Shi F et al (2021) A novel multiple instance learning framework for COVID-19 severity assessment via data augmentation and self-supervised learning. Med Image Anal. https://doi.org/10.1016/j.media.2021.101978
18. Lu MY, Williamson DFK, Chen TY, Chen RJ, Barbieri M, Mahmood F (2021) Data-efficient and weakly supervised computational pathology on whole-slide images. Nat Biomed Eng 5:555–570
19. Maximilian Ilse JT, Welling M (2018) Attention-based deep multiple instance learning. Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, PMLR 80
20. Qi LL, Wu BT, Tang W et al (2020) Long-term follow-up of persistent pulmonary pure ground-glass nodules with deep learning-assisted nodule segmentation. Eur Radiol 30:744–755
21. Hanley JA, McNeil BJ (1983) A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 148:839–843
22. Garcia-Figueiras R, Baleato-Gonzalez S, Luna A et al (2020) Assessing immunotherapy with functional and molecular imaging and radiomics. Radiographics 40:1987–2010
23. Quellec G, Cazuguel G, Cochener B, Lamard M (2017) Multiple-instance learning for medical image and video analysis. IEEE Rev Biomed Eng 10:213–234
24. Whiteside TL (2008) The tumor microenvironment and its role in promoting tumor growth. Oncogene 27:5904–5912
25. Reck M, Remon J, Hellmann MD (2022) First-line immunotherapy for non-small-cell lung cancer. J Clin Oncol 40:586–597
26. Khorrami M, Prasanna P, Gupta A et al (2020) Changes in CT radiomic features associated with lymphocyte distribution predict overall survival and response to immunotherapy in non-small cell lung cancer. Cancer Immunol Res 8:108–119
27. Jiang Y, Wang H, Wu J et al (2020) Noninvasive imaging evaluation of tumor immune microenvironment to predict outcomes in gastric cancer. Ann Oncol 31:760–768
28. Vaidya P, Bera K, Patil PD et al (2020) Novel, non-invasive imaging approach to identify patients with advanced non-small cell lung cancer at risk of hyperprogressive disease with immune checkpoint blockade. J Immunother Cancer 8(2):e001343. https://doi.org/10.1136/jitc-2020-001343
29. Chen N, Fang W, Lin Z et al (2017) KRAS mutation-induced upregulation of PD-L1 mediates immune escape in human lung adenocarcinoma. Cancer Immunol Immunother 66:1175–1187
30. Mazieres J, Drilon A, Lusque A et al (2019) Immune checkpoint inhibitors for patients with advanced lung cancer and oncogenic driver alterations: results from the IMMUNOTARGET registry. Ann Oncol 30:1321–1328
31. Liu C, Zheng S, Jin R et al (2020) The superior efficacy of anti-PD-1/PD-L1 immunotherapy in KRAS-mutant non-small cell lung cancer that correlates with an inflammatory phenotype and increased immunogenicity. Cancer Lett 470:95–105
32. Bailly C (2020) Regulation of PD-L1 expression on cancer cells with ROS-modulating drugs. Life Sci 246:117403. https://doi.org/10.1016/j.lfs.2020.117403

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.