La radiologia medica (2024) 129:239–251
https://doi.org/10.1007/s11547-024-01770-6

CHEST RADIOLOGY

Computer-aided diagnosis of distal metastasis in non-small cell lung cancer by low-dose CT based radiomics and deep learning signatures

Xiaoyi Song[1,2] · Xiaobei Duan[3] · Xinghua He[4] · Yubo Wang[4] · Kunwei Li[5] · Bangxuan Deng[4] · Xiangmeng Chen[6] · Ying Wang[4] · Man Li[1,2,8,9] · Hong Shan[1,2,7]

Received: 28 July 2023 / Accepted: 3 January 2024 / Published online: 12 January 2024

Abstract

Background This study aimed to develop and validate radiomics and deep learning (DL) signatures for predicting distal metastasis (DM) of non-small cell lung cancer (NSCLC) in low-dose computed tomography (LDCT).

Methods Images and clinical data were retrospectively collected for 381 NSCLC patients and prospectively collected for 114 patients at the Fifth Affiliated Hospital of Sun Yat-Sen University. Additionally, we enrolled 179 patients from the Jiangmen Central Hospital to externally validate the signatures. Machine-learning algorithms were employed to develop radiomics signature while the DL signature was developed using neural architecture search. The diagnostic efficiency was primarily quantified with the area under receiver operating characteristic curve (AUC). We interpreted the reasoning process of the radiomics signature and DL signature by radiomics voxel mapping and attention weight tracking.

Results A total of 674 patients with pathologically-confirmed NSCLC were included from two institutions, with 143 of them having DM. The radiomics signature achieved AUCs of 0.885, 0.854, and 0.733 in the internal validation, prospective validation, and external validation while those for DL signature were 0.893, 0.786, and 0.780. The proposed signatures achieved a promising performance in predicting the DM of NSCLC and outperformed the approaches proposed in previous studies. Interpretability analysis revealed that both radiomics and DL signatures could detect the variations among voxels inside tumors, which helped in identifying the DM of NSCLC.

Conclusions Our study demonstrates the potential of LDCT-based radiomics and DL signatures for predicting DM in NSCLC. These signatures could help improve lung cancer screening regarding further diagnostic tests and treatment strategies.

Keywords Non-small cell lung cancer · Low dose computer tomography · Deep learning · Radiomics · Cancer screening · Tumor metastasis

Xiaoyi Song and Xiaobei Duan contributed equally to this work.

✉ Ying Wang
wangy9@mail.sysu.edu.cn

✉ Man Li
liman26@mail.sysu.edu.cn

✉ Hong Shan
shanhong_2022@126.com

[1] Guangdong Provincial Engineering Research Center of Molecular Imaging, the Fifth Affiliated Hospital, Sun Yat-Sen University, Zhuhai 519000, Guangdong Province, China

[2] Guangdong-Hong Kong-Macao University Joint Laboratory of Interventional Medicine, the Fifth Affiliated Hospital, Sun Yat-Sen University, Zhuhai 519000, China

[3] Department of Nuclear Medicine, Jiangmen Central Hospital, Jiangmen 529030, China

[4] Department of Nuclear Medicine, the Fifth Affiliated Hospital, Sun Yat-Sen University, Zhuhai 519000, Guangdong Province, China

[5] Department of Radiology, the Fifth Affiliated Hospital, Sun Yat-Sen University, Zhuhai 519000, China

[6] Department of Radiology, Jiangmen Central Hospital, Jiangmen 529030, China

[7] Department of Interventional Medicine, the Fifth Affiliated Hospital, Sun Yat-Sen University, Zhuhai 519000, China

[8] Department of Information Technology and Data Center, the Fifth Affiliated Hospital, Sun Yat-Sen University, Zhuhai 519000, China

[9] Biobank of the Fifth Affiliated Hospital, Sun Yat-Sen University, Zhuhai 519000, China

Background

Lung malignancy is the second most common cancer and one of the most deadly cancers in the world [1], in which non-small cell lung cancer (NSCLC) occupies around 85% [2]. Accurate staging plays a crucial role in guiding treatment strategies and predicting prognosis in the management of NSCLC [3]. Nevertheless, staging NSCLC remains challenging due to the inability of conventional imaging modalities to meet clinical requirements [4].

Low-dose computed tomography (LDCT) is the most common modalities recommended for cancer screening. It has been shown to significantly reduce mortality from lung cancer [5]. Nevertheless, LDCT has limited diagnostic performance in identifying tumor metastasis. To complement LDCT, additional examinations, such as ^{18}F-FDG-PET/CT or endobronchial ultrasound-guided transbronchial needle aspiration are recommended [6]. Among these additional examinations, ^{18}F-FDG-PET/CT is the preferred modality for diagnosing distal metastasis (DM) in NSCLC due to high accuracy [7]. However, it comes with the drawback of being costly and time-consuming due to the need for a whole-body scan. This creates a clinical demand for convenient, economical, and reliable non-invasive imaging parameters that can improve preliminary screening for DM in NSCLC patients [8].

Recently, there has been a growing interest in utilizing radiomics and deep learning (DL) techniques to analyze medical images. These methods have demonstrated their ability to learn and decipher the representative radiologic phenotypes of tumors [9–12]. Notably, previous studies have highlighted their success in predicting lymph node metastasis in various cancers, including lung [13, 14], breast [15], gastric [16], and thyroid [17]. Despite these achievements, both radiomics and DL methods have shown limited efficacy in predicting DM in patients with NSCLC [18]. Moreover, the learning pattern and mechanism underlying the prediction of these methods remain unclear. Unraveling these mechanisms could contribute to enhancing the practical applicability of artificial intelligence in real clinical settings.

Herein, the aim of this study was to develop and validate LDCT-based radiomics and DL signatures to improve the preliminary screening for DM in patients with NSCLC using LDCT images. Furthermore, we sought to investigate the learning pattern and mechanism involved underlying these prediction methods.

Materials and methods

Study population

This study was approved by our institutional review board. For patients in the retrospective cohort, the written informed consent was waived while those of patients in the prospective cohort were obtained in this study. Patients who underwent the LDCT examination and ^{18}F-FDG PET/CT scan from November 2017 to July 2020 were retrospectively recruited from the Fifth Affiliated Hospital of Sun Yat-Sen University. The included patients should satisfy the demand for (1) pathologically-confirmed primary NSCLC; (2) single lesion; (3) no histories of other cancers; (4) the interval time from CT examination to ^{18}F-FDG PET/CT scan within 2 weeks. The exclusion criteria were as follows: (a) prior puncture biopsy, chemotherapy, or radiotherapy before PET/CT scanning; (b) unsatisfactory image quality; (c) inability to delineate the lesion on CT; (d) incomplete clinicopathologic data. Stratified random sampling was performed to allocate the study population into the development cohort (n = 337) and internal validation cohort (n = 44) at a ratio of 7:3 based on the case group. Samples in the control group within the internal validation cohort were allocated equally to those in the case group to balance the data.

Following the same admission criteria, eligible patients at the Fifth Affiliated Hospital of Sun Yat-Sen University from August 2020 to October 2022 were prospectively collected to form a prospective validation cohort (n = 114). Additionally, we recruited patients from January 2020 to July 2022 from the Jiangmen Central Hospital to create an external validation cohort (n = 179). The flowchart illustrating the patient recruitment process is presented in Fig. 1.

Clinicopathological characteristics, such as age, gender, smoking history, pathologic type, TNM staging, and ^{18}F-FDG PET/CT parameters were extracted from the medical records. The status of DM was confirmed based on the diagnostic report of the ^{18}F-FDG PET/CT examination.

Acquisition of CT images and interpretation of radiologic signs

The imaging process and acquisition parameters of the CT scanner and PET scanner are detailed in the Supplementary Information. Radiologic signs were manually extracted, including tumor size and location, pleural tag, pleural lesions, air bronchogram, calcification, cavitation, well-defined, lobulation, spiculation, vessel convergence, and vascular involvement. Detailed definitions of these CT

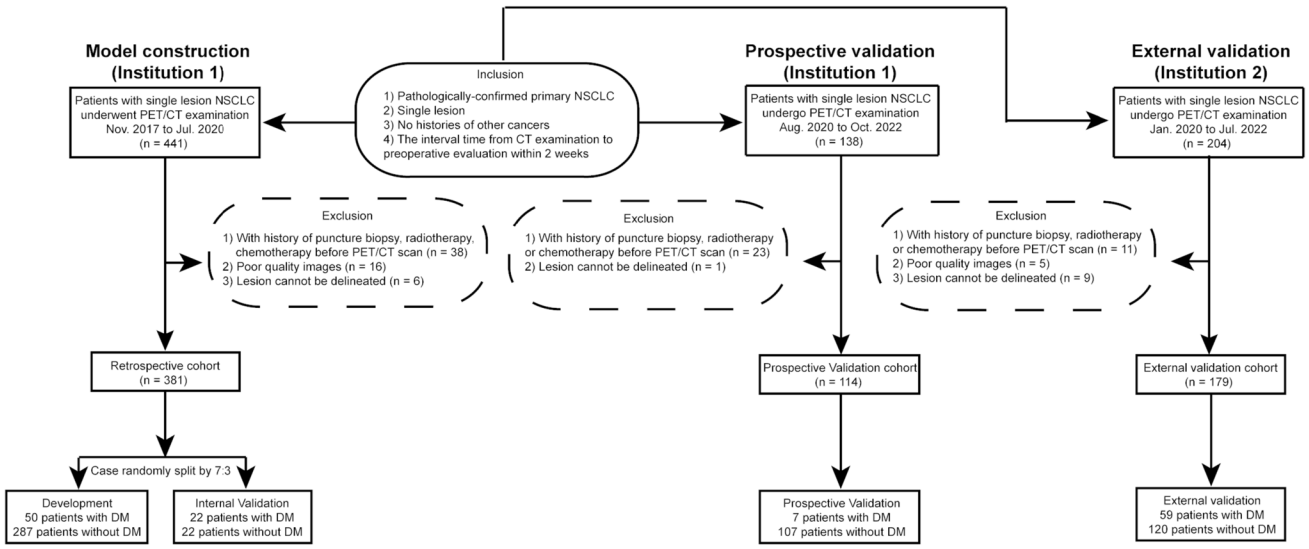

Fig. 1 Flowchart of the process of patient enrollment and grouping. *Institution 1*, the Fifth Affiliated Hospital of Sun Yat-Sen University, *Institution 2*, Jiangmen Central Hospital

radiologic signs and the interpretive process can be found in the Supplementary Information. The average of continuous variables was calculated as the final value, while for categorical variables, consensus was reached through discussion in the event of any discrepancies. Cohen's Kappa coefficient or intraclass correlation coefficient was calculated for each radiologic sign, and the interobserver agreement distribution is presented in Table S1.

Signature development and interpretation

Machine-learning (ML) algorithms were utilized to develop the radiomics signature while the DL signature was developed using neural architecture search (NAS) [19]. The 3D region of interest of the primary tumor was delineated from LDCT using ITK-SNAP software (Version 3.8.0) and the procedure was detailed in Supplementary Information. Figure 2 provides an abstract representation of the development workflows.

For the development of radiomics signature, radiomics features were extracted from the 3D region of interest of tumors using PyRadiomics [20] and further selected by the random forest regressor. A total of 7 ML algorithms (Logistic Regression, BernoulliNB, KNeighborsClassifier, RandomForestClassifier, XGBClassifier, DecisionTreeClassifier, SVM) were trained to constitute the radiomics signature. To interpret the reasoning process, we used radiomics voxel mapping to reflect the contribution of each voxel to the calculation of a certain radiomics feature. The whole development process is detailed in Supplementary Information.

Based on NAS, we selected and trained the top 10 DL architectures to construct the DL signature. To track the attention weight of each voxel in the reasoning process, convolutional block attention modules [21] were added. The detailed development process is described in Supplementary Information.

For both signatures, an ensemble strategy was used. Radiomics signature was determined by averaging the predicted probabilities generated by the seven ML algorithms, while the DL signature was determined by averaging the predicted probabilities generated by the 10 candidate DL classifiers. To verify the ensemble strategy, n models were randomly picked to perform an ensemble prediction. For the radiomics signature, n ranged from 1 to 7, and for the DL signature, n ranged from 1 to 10. The progress was repeated 10 times, and the average performance indices were calculated.

Construction of the clinical-radiologic model and combined models

In the development cohort, the logistic regression model was used to construct a clinical-radiologic (CR) model by incorporating clinical characteristics and radiologic signs. Based on CR model, additional combined models were constructed by integrating the radiomics and DL signatures with CR model. The aim of constructing the combined models was to investigate the ability of the signatures to improve the CR model and identify the optimal prediction model.

Statistical analysis

Statistical analysis was performed using R software (version 4.1.0) and SPSS software (IBM, version 23.0). Continuous data were compared using the Student's t-test or

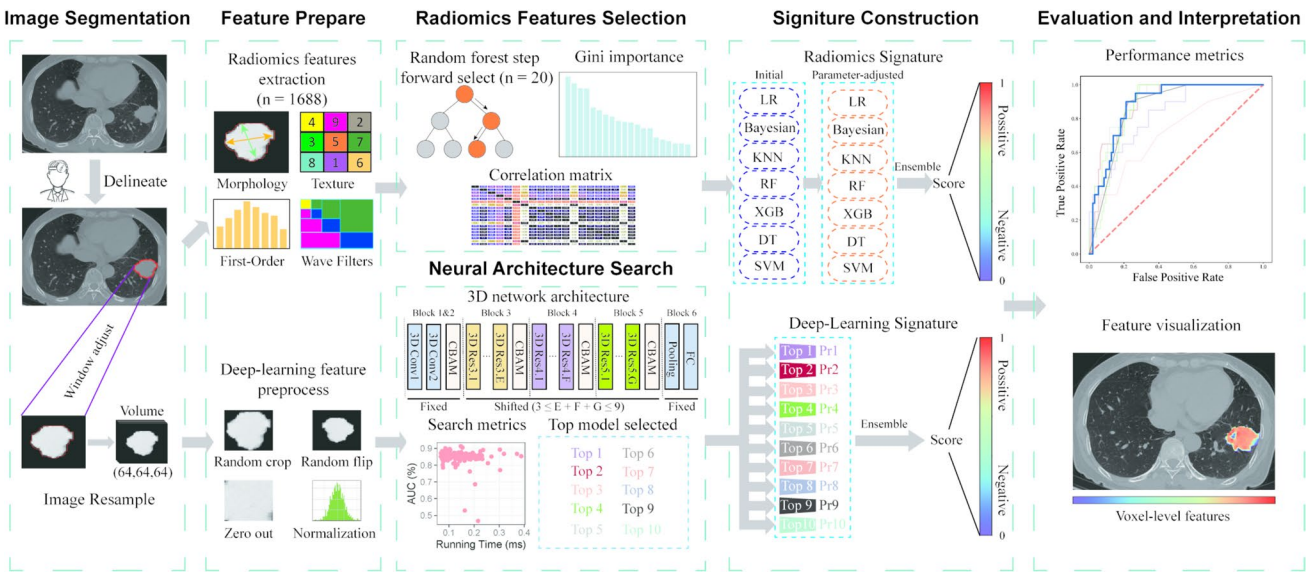

Fig. 2 Workflow of the radiomics and deep learning signatures building process. For radiomics signature, a total of 1688 features were extracted and the top 20 features were further selected by the random forest regressor. Following parameter adjustments, seven machine learning models were trained to create an ensemble prediction. Regarding the deep learning signature, a 3D network architecture search was conducted. The number of shifted frameworks ranged from 3 to 9 ($3 \le E+F+G \le 9$). The top 10 architectures that exhibited an excellent trade-off between performance and speed were selected and trained to make an ensemble prediction. As for the evaluation and interpretation, receiver operating characteristic (ROC) curves were used and the features selected to construct the radiomics and deep learning signatures were visualized

Kruskal–Wallis test. Categorical data were analyzed using the Chi-square test or Fisher's exact test. The Spearman correlation coefficient was used to evaluate the correlation between variables. The diagnostic efficiency of models was mainly quantified by the area under the receiver operating characteristic curve (AUC). The confidence interval (CI) of AUC was calculated using 10000 bootstrap replicates. DeLong test was employed to compare AUCs of different methods. Additional performance metrics including accuracy, sensitivity, specificity, positive predictive value, and negative predictive value were also reported. Net Reclassification Index and Integrated Discrimination Improvement were utilized to quantify the ability of signatures to improve the CR model. For all statistical tests, a $P < 0.05$ indicated a statistically significant difference.

Results

Baseline characteristics

A total of 674 patients were included in the study, divided into the development cohort (n = 337), internal validation cohort (n = 44), prospective validation cohort (n = 114) and external validation cohort (n = 179). The mean age of the study population was 60 years ± 11 (± standard deviation), and 51% (n = 343) of the patients were male. Adenocarcinoma (n = 592) accounted for 88% of the total population, while squamous cell carcinoma (n = 74) accounted for 11%. Among the patients, 21% (n = 143) had DM. The incidence rates of DM across different T-staging were as follows: T1 (9%), T2 (32%), T3 (37%), and T4 (60%). The baseline characteristics of the patients are presented in Table 1.

The development of radiomics signature and DL signature

For the radiomics signature, we extracted a total of 1688 features (Table S2). After applying the Gini importance ranking using a random forest regressor, we selected 20 features (Fig. S1) for further analysis. These 20 features were used to train the seven classifiers, as described in the Supplementary Information. In the case of the DL signature, we employed the NAS approach to select and train the top 10 candidate model architectures, as explained in the Supplementary Information. The development cohort's AUC for each single model used to create the radiomics signature and DL signature is presented in Fig. 3a–b. It is evident that these single models exhibited varying performance. However, as demonstrated in Fig. 3c–d, larger numbers of models used for ensemble prediction generally led to improved performance.

Table 1 The distribution of clinicopathological characteristics and radiologic signs across the cohorts

Characteristic	Development cohort (n = 337)	Internal validation cohort (n = 44)	Prospective validation cohort (n = 114)	External validation cohort (n = 179)	*P* Value
Age					0.051
> 65 years	109 (32.3%)	22 (50.0%)	38 (33.3%)	73 (40.8%)	
≤ 65 years	228 (67.6%)	22 (50.0%)	76 (66.7%)	106 (59.2%)	
Mean age (y)*	60 ± 11	62 ± 11	60 ± 11	61 ± 11	0.375
Gender					0.673
M	165 (48.9%)	24 (54.5%)	57 (50.0%)	97 (54.2)	
F	172 (51.1%)	20 (45.5%)	57 (50.0%)	82 (45.8)	
Smoking history					0.907
Ever smoke	106 (31.5%)	14 (31.8%)	32 (28.1%)	53 (29.6%)	
Never smoke	231 (68.5%)	30 (68.2%)	82 (71.9%)	126 (70.3%)	
Tumor size (cm)*	2.60 ± 1.81	3.22 ± 2.19	2.15 ± 1.20	2.96 ± 1.65	< 0.001
Tumor location					0.003
LUL	87 (25.8%)	7 (15.9%)	24 (21.1%)	35 (19.6%)	
LLL	47 (13.9%)	4 (9.1%)	22 (19.3%)	28 (15.6%)	
RUL	111 (32.9%)	14 (31.8%)	38 (33.3%)	52 (29.1%)	
RML	36 (10.7%)	5 (11.4%)	8 (7.0%)	9 (5.0%)	
RLL	49 (14.6%)	13 (29.5%)	22 (19.3%)	42 (23.5%)	
Central	7 (2.1%)	1 (2.3%)	0 (0%)	13 (7.3%)	
Pathologic type					0.683
LUSC	43 (12.8%)	4 (9.1%)	11 (9.6%)	16 (8.9%)	
LUAD	291 (86.4%)	39 (88.6%)	102 (89.5%)	160 (89.4%)	
Other	3 (0.8%)	1 (2.3%)	1 (0.9%)	3 (1.7%)	
Radiologic signs					
Pleural tag	209 (62.0%)	31 (70.5%)	67 (58.8%)	61 (34.1%)	< 0.001
Pleural lesions	62 (18.4%)	12 (27.3%)	3 (2.6%)	32 (17.9%)	< 0.001
Air bronchogram	42 (12.5%)	2 (4.5%)	22 (19.3%)	46 (25.7%)	< 0.001
Calcification	3 (0.9%)	1 (2.3%)	0 (0%)	18 (10.1%)	< 0.001
Cavitation	63 (18.7%)	9 (20.5%)	17 (14.9%)	42 (23.5%)	< 0.001
Well defined	35 (10.4%)	4 (9.1%)	10 (8.8%)	40 (22.3%)	< 0.001
Lobulation	272 (80.7%)	35 (79.5%)	112 (98.2%)	68 (37.9%)	< 0.001
Spiculation	201 (59.6%)	26 (59.1%)	74 (64.9%)	77 (43.0%)	< 0.001
Vessel convergence	43 (12.8%)	6 (13.6%)	39 (34.2%)	48 (26.8%)	< 0.001
Vascular involvement	39 (11.6%)	7 (15.9%)	10 (8.8%)	29 (16.2%)	0.225
AJCC T stage					< 0.001
T1	208 (61.7%)	20 (45.5%)	89 (78.1%)	107 (59.8%)	
T2	57 (16.9%)	7 (15.9%)	17 (14.8%)	52 (29.0%)	
T3	29 (8.6%)	4 (9.1%)	6 (5.3%)	15 (8.4%)	
T4	43 (12.8%)	13 (29.5%)	2 (1.8%)	5 (2.8%)	
AJCC N stage					< 0.001
N0	226 (67.1%)	21 (47.8%)	94 (82.5%)	94 (52.5%)	
N1	23 (6.8%)	2 (4.5%)	5 (4.4%)	17 (9.5%)	
N2	50 (14.8%)	4 (9.1%)	8 (7.0%)	18 (10.1%)	
N3	38 (11.3%)	17 (38.6%)	7 (6.1%)	50 (27.9%)	
AJCC M stage					< 0.001
M0	287 (85.2%)	22 (50.0%)	107 (93.9%)	115 (64.2%)	
M1	50 (14.8%)	22 (50.0%)	7 (6.1%)	64 (35.7%)	
*^{18}F-FDG PET/CT parameters**					
SUVmax	7.19 ± 7.43	8.08 ± 7.46	3.97 ± 4.76	6.66 ± 6.38	< 0.001
SUVmin	0.37 ± 0.17	0.37 ± 0.19	0.31 ± 0.18	0.88 ± 1.39	< 0.001
SUVavg	4.25 ± 4.29	4.90 ± 4.54	2.43 ± 2.91	2.82 ± 2.58	< 0.001

Unless otherwise noted, values are numbers of patients and compared using the Chi-square test or Fisher's exact test, with percentages in parentheses. *LUL* left upper lobe, *LLL* left lower lobe, *RUL* right upper lobe, *RML* right middle lobe, *RLL* right lower lobe, *LUSC* lung squamous carcinoma, *LUAD* lung adenocarcinoma, *AJCC* American Joint Committee on Cancer

Table 1 (continued)

*Data were presented as means ± standard deviations and compared using the Kruskal–Wallis test

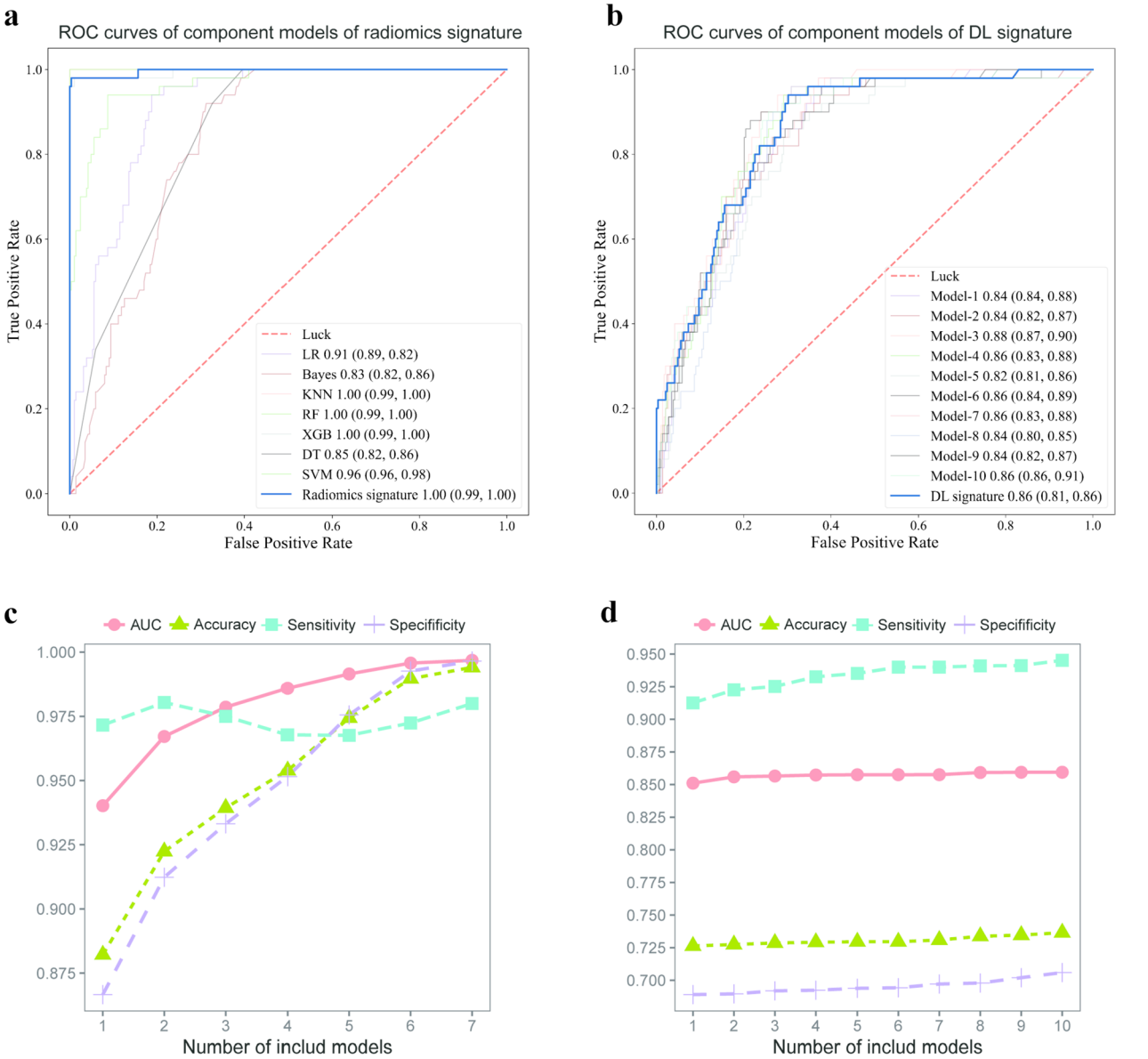

Fig. 3 The performance of single models and the evaluation of the proposed ensemble method using different numbers of single models. **a**–**b** Receiver operating characteristic (ROC) curves of each single model used to create the radiomics signature and DL signature in the development cohort. **a** radiomics signature and **b** deep learning signature. **c**–**d** The trend of different criteria when using different numbers of models to make an ensemble prediction. **c** radiomics signature and **d** deep learning signature

Evaluation of the predictive efficiency of the signature

In the internal validation (Fig. 4a), radiomics signature achieved an AUC of 0.89 [95% CI 0.87, 0.90], which was comparable to the AUC of DL signature. For prospective validation (Fig. 4b), radiomics signature obtained an AUC of 0.85 [95% CI 0.80, 0.88], while DL signature yielded an AUC of 0.79 [95% CI 0.76, 0.83]. However, Delong test indicated no statistical difference ($P = 0.184$) between the two. In external validation (Fig. 4c), DL signature achieved a significantly higher AUC (0.78 [95% CI 0.75, 0.80]) compared to radiomics signature (0.73 [95% CI 0.72, 0.77]) ($P = 0.043$).

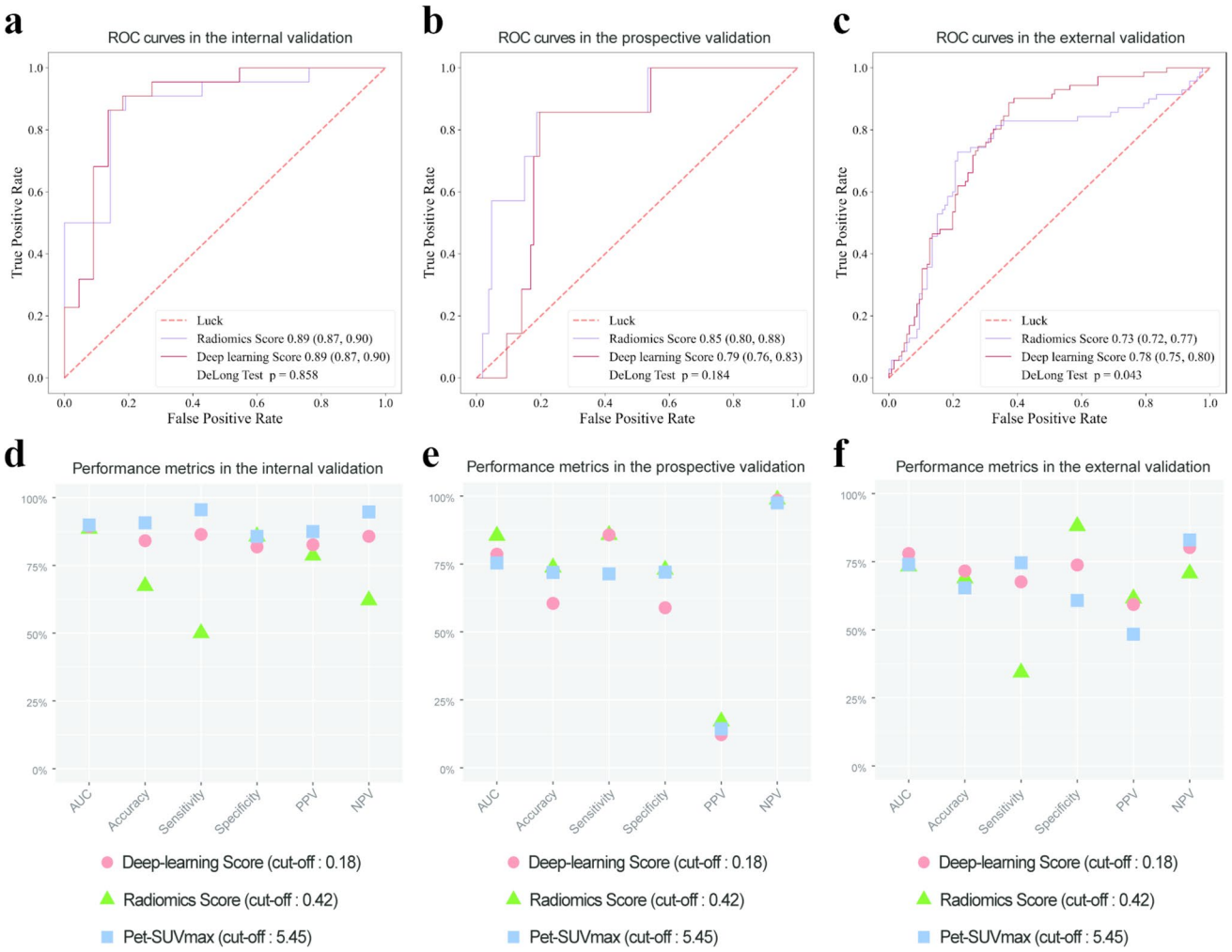

Fig. 4 Predictive performance of the radiomics and deep learning signatures across the validation cohorts. **a**–**c**) Receiver operating characteristic (ROC) curves of the radiomics signature in the **a** internal validation cohort, **b** prospective validation cohort, and **c** external validation cohort. **d**–**f** The comparison among radiomics signature, deep learning signature, and SUVmax on various performance indices in **d** internal validation set and **e** prospective validation cohort, and **f** external validation cohort

To comprehensively assess the diagnostic efficiency, the performance of the signatures was compared to that of SUVmax, a well-established ^{18}F-FDG PET/CT parameter known to be strongly linked to tumor metastasis [22, 23]. As depicted in Fig. 4e–f, both signatures exhibited similar discriminability to SUVmax in predicting DM across validation cohorts. Additionally, radiomics signature demonstrated superior specificity, whereas DL signature displayed better sensitivity across validation cohorts.

Signature reasoning pattern interpretation

To elucidate the reasoning pattern of radiomics signature, we focused on the “Original first order Energy”, which demonstrated the strongest association with DM based on both the Gini importance ranking (Fig. S1) and Spearman correlation index (Fig. S2). By employing radiomics voxel mapping, we visualized the contribution of individual voxels in calculating this radiomics feature on the CT scan. Interestingly, we observed discrepancies in voxel-level contribution between patients with and without DM, with a higher number of voxels exhibiting substantial contribution in tumors of patients with DM (Fig. 5a). Furthermore, statistical analysis confirmed that patients with DM exhibited significantly higher levels of the "Original first order Energy" (Fig. 5c).

To interpret the reasoning pattern of the DL signature, we extracted the voxel-level attention weights from the first convolutional block attention module layer of the trained DL models. As depicted in Fig. 5b, the tumor voxels of patients with DM exhibited higher attention weights compared to those of patients without DM. Additionally, the average voxel-level attention weights were significantly higher in tumors of patients with DM (Fig. 5d).

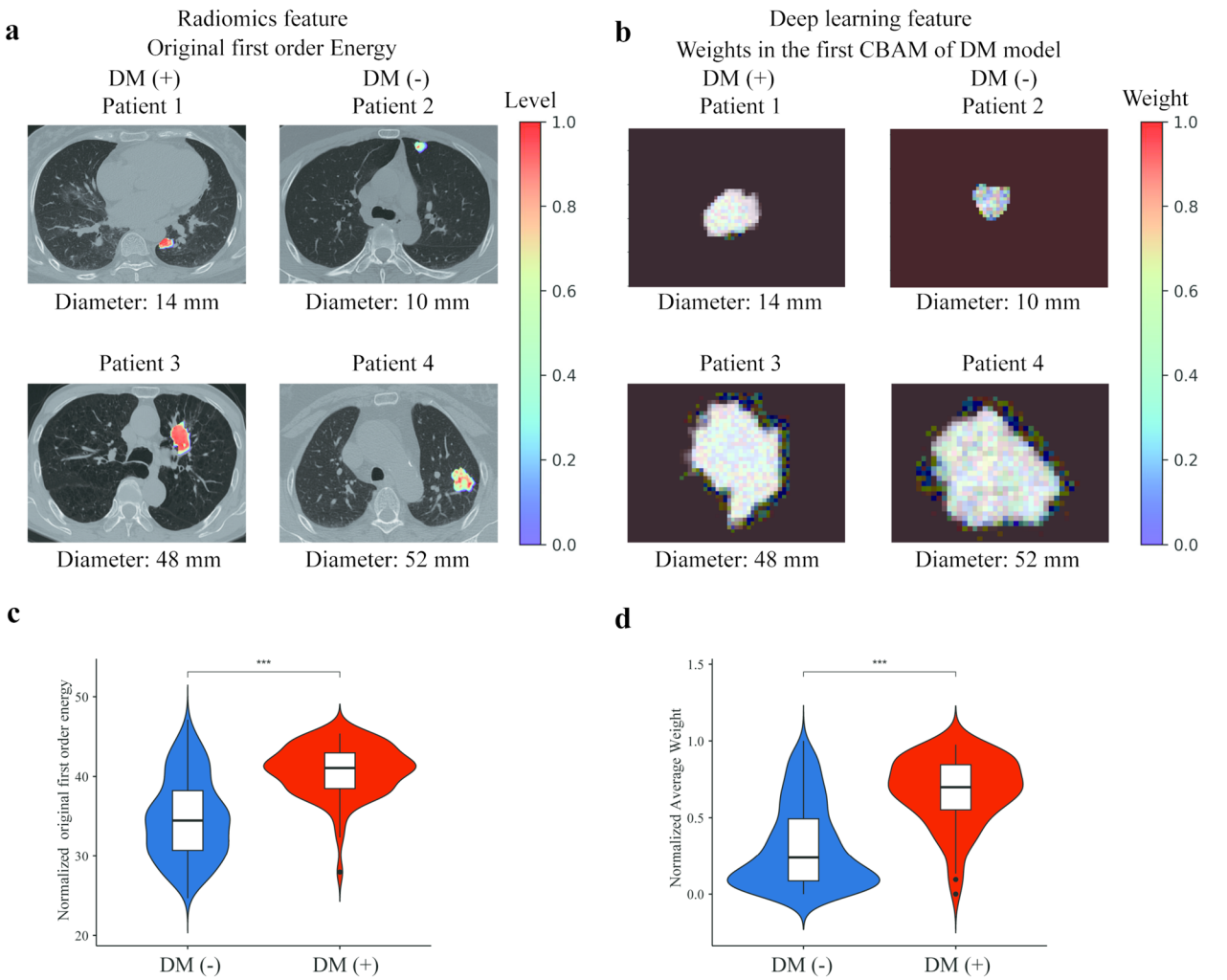

Fig. 5 The comparison between cases and controls on intra-nodular radiomics feature and deep learning attention weights at CT (axial view). **a** The radiomics voxel mapping technique visualized the most relevant radiomics feature, "Original first order Energy", in tumors of various sizes to identify distal metastasis (DM). **b** The visualization of attention weights of the deep learning approach in tumors of various sizes facilitated the identification of DM. The color bar illustrated the strength of these features. **c** The comparison of the value of Original first-order Energy between patients with DM or without. **d** The comparison of the average attention weights between patients with DM or without. ***$P < 0.001$

These findings demonstrate that both radiomics and DL techniques have the capability to identify variations among voxels within CT.

Construction and evaluation of clinical-radiologic models and combined models

To develop the CR model, we initially identified the candidate covariates that were significantly associated with DM through univariate analysis. The Spearman correlation coefficients among the candidate covariates are shown in Fig. S3. Subsequently, the final model was constructed using multivariate logistic regression analysis (Table 2). Noted that pathological metrics and ^{18}F-FDG PET/CT parameters were waived in model construction because of the inherent design for preliminary screening. The result of multivariate logistic regression analysis revealed Pleural invasion (OR: 5.00; 95% CI 1.92, 13.36; $P = 0.001$) and Cavitation (OR:0.18; 95% CI 0.03, 0.65; $P = 0.024$) were two radiologic signs that were independent risk predictors for DM in patients with NSCLC.

The AUC for CR model in internal validation was 0.874 [95% CI 0.862, 0.892], in prospective validation it was 0.533 [95% CI 0.516, 0.537], and in external validation it was 0.712 [95% CI 0.670, 0.721] (Table 3). In internal validation, there was no statistical difference in the discriminative abilities of the CR model, radiomics signature, and DL signature ($P > 0.05$). In prospective validation cohort, CR model performed inferiorly to both radiomics signature (0.533 vs. 0.854, $P < 0.001$) and DL signature (0.533 vs. 0.786, $P < 0.001$). Moreover, in external validation, CR

Table 2 Univariate and multivariable logistic regression analysis to construct the clinical-radiologic model for predicting distal metastasis

Variables	Univariate analysis			Multivariable analysis		
	β	OR (95% CI)	*P* Value	β	OR (95% CI)	*P* Value
Age	0.03	1.03 (0.99, 1.06)	0.065	–	–	–
Gender						
M	0.52	1.68 (0.91, 3.15)	0.099	–	–	–
F	−0.52	0.59 (0.32, 1.09)	0.099	–	–	–
Smoking	0.53	1.70 (0.91, 3.14)	0.091	–	–	–
Tumor size (cm)	0.39	1.48 (1.29, 1.71)	< 0.001*	0.07	1.08 (0.84, 1.37)	0.561
Tumor location						
LUL	−0.53	0.59 (0.26, 1.22)	0.179	–	–	–
LLL	0.53	1.69 (0.75, 3.57)	0.181	–	–	–
RUL	−0.38	0.68 (0.33,1.31)	0.267	–	–	–
RML	1.09	2.97 (1.31, 6.39)	0.007*	0.69	2.00 (0.74, 5.11)	0.158
RLL	−0.78	0.46 (0.13, 1.19)	0.152	–	–	–
Central	1.51	4.53 (0.87, 21.19)	0.053	–	–	–
Radiologic signs						
Pleural tag	0.90	0.90 (0.49, 1.68)	0.737	–	–	–
Pleural invasion	9.51	9.51 (4.93, 18.67)	< 0.001*	1.61	5.00 (1.92, 13.36)	0.001*
Air bronchogram	0.12	0.12 (0.01, 0.59)	0.041*	−1.95	0.14 (0.01, 0.83)	0.080
Calcification	2.92	2.92 (0.13, 31.03)	0.386	–	–	–
Cavitation	0.15	0.15 (0.02, 0.51)	0.010*	−1.74	0.18 (0.03, 0.65)	0.024*
Well defined	0.32	0.32 (0.05,1.11)	0.128	–	–	–
Lobulation	2.42	2.42 (1.00, 7.21)	0.073	–	–	–
Spiculation	1.54	1.54 (0.82, 2.98)	0.185	–	–	–
Vessel convergence	2.25	2.25 (1.02, 4.74)	0.037*	−0.16	0.86 (0.33, 2.12)	0.743
Vascular involvement	4.71	4.71 (2.23, 9.79)	< 0.001*	−0.72	0.49 (0.15, 1.51)	0.219
AJCC T stage						
T1 or T2	0.11	0.11 (0.06, 0.22)	< 0.001*	–	–	–
T3 or T4	8.79	8.79 (4.61, 17.13)	< 0.001*	1.23	3.43 (1.20, 9.67)	0.020*

OR odds ratio, *CI* confidence interval, *LUL* left upper lobe, *LLL* left lower lobe, *RUL* right upper lobe, *RML* right middle lobe, *RLL* right lower lobe, *LUSC* lung squamous carcinoma, *LUAD* lung adenocarcinoma, *AJCC* American Joint Committee on Cancer

*$P < 0.05$

model performed similarly to radiomics signature (0.712 vs. 0.733, $P = 0.398$) but worse than DL signature (0.712 vs. 0.780, $P = 0.039$).

Based on CR model, combined models were developed (Table S3). In prospective validation, the CR-radiomics model, CR-DL model, and CR-radiomics-DL model achieved an AUC of 0.876, 0.813, and 0.876, respectively. In external validation, the aforementioned combined models achieved an AUC of 0.707, 0.721, and 0.705, respectively. The analysis of the Net Reclassification Index and Integrated Discrimination Improvement showed CR model was significantly improved by integrating radiomics and DL signatures in the development cohort, internal validation cohort, and prospective validation cohort (Table S4). However, in external validation, there were only limited improvements observed for CR model, and all of the combined models performed inferiorly to DL signature (CR-radiomics model: 0.707 vs. 0.780, $P = 0.026$; CR-DL model: 0.721 vs. 0.780, $P = 0.046$; CR-radiomics-DL model: 0.705 vs. 0.780, $P = 0.020$).

Discussion

In this study, we have several notable strengths. Firstly, we have developed two signatures that outperform previous methods in accurately predicting DM in NSCLC. Secondly, we have provided valuable insights into the interpretation of radiomics and DL signatures, shedding light on their capability to identify DM in NSCLC. Furthermore, these signatures were specifically developed using LDCT, suggesting their potential for broad applicability as a preliminary screening tool for DM in NSCLC patients.

Table 3 Performance of all models constructed in this study in the development cohort, internal validation cohort, prospective validation cohort, and external validation cohort

Models	AUC (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
Development cohort			
Radiomics signature	0.983 (0.973, 0.989)	0.940 (0.880, 0.999)	0.951 (0.978, 0.999)
DL signature	0.859 (0.824, 0.878)	0.940 (0.825, 0.984)	0.698 (0.641, 0.750)
CR model	0.798 (0.783, 0.845)	0.700 (0.552, 0.817)	0.854 (0.807, 0.892)
CR-radiomics model	0.985 (0.978, 0.993)	0.940 (0.825, 0.984)	0.976 (0.948, 0.989)
CR-DL model	0.877 (0.843, 0.894)	0.920 (0.799, 0.974)	0.736 (0.681, 0.785)
CR-radiomics-DL model	0.991 (0.980, 0.992)	0.980 (0.879, 0.999)	0.944 (0.909, 0.967)
Internal validation cohort			
Radiomics signature	0.885 (0.867, 0.899)	0.500 (0.289, 0.712)	0.857 (0.626, 0.962)
DL signature	0.893 (0.869, 0.902)	0.864 (0.640, 0.964)	0.818 (0.589, 0.940)
CR model	0.874 (0.862, 0.892)	0.727 (0.496, 0.884)	0.905 (0.682, 0.983)
CR-radiomics model	0.883 (0.869, 0.897)	0.455 (0.251, 0.673)	0.952 (0.741, 0.997)
CR-DL model	0.922 (0.913, 0.937)	0.864 (0.640, 0.964)	0.857 (0.626, 0.962)
CR-radiomics-DL model	0.885 (0.863, 0.899)	0.500 (0.288, 0.712)	0.952 (0.741, 0.998)
Prospective validation cohort			
Radiomics signature	0.854 (0.806, 0.889)	0.857 (0.420, 0.992)	0.729 (0.633, 0.808)
DL signature	0.786 (0.757, 0.822)	0.857 (0.420, 0.992)	0.589 (0.489, 0.682)
CR model	0.533 (0.516, 0.537)	0.000 (0.000, 0.000)	0.907 (0.953, 1.000)
CR-radiomics model	0.876 (0.849, 0.921)	0.857 (0.420, 0.992)	0.748 (0.653, 0.824)
CR-DL model	0.813 (0.784, 0.841)	0.857 (0.420, 0.992)	0.673 (0.569, 0.753)
CR-radiomics-DL model	0.876 (0.802, 0.887)	0.857 (0.420, 0.992)	0.710 (0.613, 0.792)
External validation cohort			
Radiomics signature	0.733 (0.722, 0.772)	0.343 (0.236, 0.467)	0.881 (0.808, 0.930)
DL signature	0.780 (0.754, 0.798)	0.676 (0.553, 0.779)	0.738 (0.651, 0.810)
CR model	0.712 (0.670, 0.721)	0.268 (0.173, 0.388)	0.897 (0.827, 0.942)
CR-radiomics model	0.707 (0.689, 0.745)	0.296 (0.196, 0.418)	0.889 (0.817, 0.936)
CR-DL model	0.721 (0.696, 0.747)	0.563 (0.441, 0.679)	0.786 (0.702, 0.852)
CR-radiomics-DL model	0.705 (0.671, 0.724)	0.437 (0.321, 0.559)	0.865 (0.789, 0.917)

CR clinical-radiologic, *CI* confidence interval, *DL* deep learning. The best cut-off pretest probability thresholds were identified in training cohort for each model, with 0.42 for Radiomics signature, 0.18 for DL signature, 0.22 for CR model, 0.38 for CR-radiomics model, 0.09 for CR-DL model, and 0.14 for CR-radiomics-DL model

Preoperative staging plays a crucial role in determining the appropriate management strategy for patients with clinical stage I NSCLC [24]. Our study population exhibited high incidence rates of DM in patients with T1 (9%) or T2 (32%) tumors, as evidenced by their baseline characteristics. Consequently, additional evaluation, such as an ^{18}F-FDG PET/CT scan, is essential in such cases [25]. Given that our signatures were developed and validated using a population comprising 83% (n = 557) of patients with T1 or T2 tumors, they could serve as a valuable tool for clinicians in conducting preliminary screening for DM using LDCT during lung cancer screening. This aids in triaging patients who require further examination.

The success of DL hinges on the efficacy of its carefully crafted neural architectures. These architectures are meticulously designed by experts with extensive professional experience, involving a time-consuming process [26]. However, a new automated method called neural architecture search (NAS) has emerged, showcasing its superiority over manually designed architectures in various tasks [27–29]. Previously, radiomics served as the vital connection between medical imaging and personalized medicine [30]. It is commonly used in conjunction with ML algorithms for a range of tasks, and it's important to note that there is no universally applicable ML model that fits every specific task [31]. In this study, we employed an ensemble strategy to develop signatures that effectively leveraged the strengths of multiple algorithms, thus complementing each other. This fusion resulted in an improved performance of the ensemble prediction.

Early radiomics studies showed limited predictive capabilities of primary tumor features for DM, with an AUC ranging from 0.64 to 0.71 [32, 33]. Even when DL methods were applied, they failed to significantly improve performance, achieving an AUC ranging from 0.65 to 0.71 [18]. These studies also had notable limitations, such as lacking

external validation and producing unexplainable predictions. In our study, we introduced cutting-edge radiomics and DL signatures that outperformed previous research in internal validation, achieving AUCs ranging from 0.786 to 0.893. Furthermore, in external validation, our signatures demonstrated superior diagnostic efficiency with AUCs of 0.73 and 0.78.

Based on our analysis of reasoning patterns, we discovered the radiomics and the DL methods operate in a similar manner by identifying voxel-level differences within the tumor. These differences are then leveraged to assess the strength of correlation between individual voxels and DM. Notably, we found that tumors of NSCLC patients with DM tend to exhibit a higher prevalence of voxels demonstrating a strong association with DM. This finding validates the efficacy of radiomics and DL methods in detecting DM, surpassing what human observers can achieve. Additionally, our interpretation of reasoning pattern reveals that utilizing a 2D approach may not be appropriate for this task. This insight can help explain why previous DL studies, which mainly utilized a 2D approach, yielded underwhelming results.

The performance of radiomics and DL signatures in this study showed variability across the validation cohorts, which could be attributed to inherent differences in the baseline characteristics among these cohorts. Furthermore, the performance might have been affected by variations in acquisition parameters across CT scanners from different vendors and institutions [34]. To improve the generalization of signatures when applied to new situations, transfer learning has been proposed as a potential solution [35]. Notably, both signatures demonstrated superior generalization ability compared to the CR model. When integrated with the CR model, these signatures significantly enhanced its discriminability. Ultimately, during external validation, the DL signature outperformed other models in terms of performance and sensitivity, suggesting its potential for optimizing the clinical workflow.

Our study has several limitations. Firstly, the data were collected from only two institutions, which resulted in a relatively small number of NSCLC patients with DM (143 patients). To ensure the reproducibility and generalizability of the findings, further validation of the signatures with a larger, multi-institution study is necessary. Secondly, there is a possibility of overfitting during the model development phase. However, it's important to note that our signatures are integrated models comprised of multiple single models. This ensemble strategy helps to partially offset the impact of overfitting of the single models. Moreover, the performance of our models in the external validation set highlights their superior generalization ability. In addition, it is worth mentioning that a comprehensive explanation of the underlying rationale behind these radiomics and DL signatures was not provided. Further investigation is required to enhance our understanding. For instance, exploring the involvement of specific genes or proteins may contribute to providing genomic biological interpretability for the signatures.

In conclusion, we developed and validated explainable LDCT-based radiomics and DL models for identifying DM in patients with NSCLC. Our models have demonstrated a high level of predictive efficiency, indicating their potential for effectively screening DM in NSCLC patients using routine LDCT scans in real-world clinical practice for lung cancer screening.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11547-024-01770-6.

Author contributions Guarantors of integrity of entire study, XYS, XBD, XHH, YBW, KWL, BXD, XMC, YW, ML, HS; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, XYS, XBD, WY, ML, HS; clinical studies, XYS, XBD, XHH, YBW, KWL, BXD, XMC, YW, HS; experimental studies, XYS, XBD, XHH, YBW, ML, HS; statistical analysis, XYS, XBD, ML; and manuscript editing, XYS, XBD, ML, HS.

Funding This study was funded by grants from the National Natural Science Foundation of China (grants number 82000628 and 62176104), the Medical Scientific Research Foundation of Guangdong Province of China (grants number B2022113), and the Guangdong-Hong Kong-Macao University Joint Laboratory of Interventional Medicine Foundation of Guangdong Province (2023LSYS001).

Availability of data and materials The datasets analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethics approval This study was approved by our institutional review board: Ethics Committee of the Fifth Affiliated Hospital of Sun Yat-sen University (approval number: [2021] K01-1). For patients in retrospective cohort, the written informed consents were waived while those of patients in prospective cohort were obtained in this study.

Consent to participate For patients in retrospective cohort, the written informed consents were waived while those of patients in prospective cohort were obtained in this study.

References

1. Sung H, Ferlay J, Siegel RL et al (2021) Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 71:209–249. https://doi.org/10.3322/caac.21660

2. Duma N, Santana-Davila R, Molina JR (2019) Non-small cell lung cancer: epidemiology, screening, diagnosis, and treatment. Mayo Clin Proc 94:1623–1640. https://doi.org/10.1016/j.mayocp.2019.01.013
3. National Cancer Institute (2023) The Surveillance E, and End Results (SEER) Program. Cancer Stat Facts: Lung and Bronchus Cancer. https://seer.cancer.gov/statfacts/html/lungb.html. Published December 11, 2011. Updated March 30, 2022. Accessed June 30
4. Ashok A, Jiwnani SS, Karimundackal G et al (2021) Controversies in mediastinal staging for nonsmall cell lung cancer. Indian J Med Paediatr Oncol 42:406–414. https://doi.org/10.1055/s-0041-1739345
5. Lam S, Bai C, Baldwin D et al (2023) Current and future perspectives on CT screening for lung cancer: a road map for 2023–2027 from the IASLC. J Thorac Oncol. https://doi.org/10.1016/j.jtho.2023.07.019
6. Lv XY, Wu ZG, Cao JL et al (2021) A nomogram for predicting the risk of lymph node metastasis in T1–2 non-small-cell lung cancer based on PET/CT and clinical characteristics. Transl Lung Cancer Res 10:430–438. https://doi.org/10.21037/tlcr-20-1026
7. Manafi-Farid R, Askari E et al (2022) [18F]FDG-PET/CT radiomics and artificial intelligence in lung cancer: technical aspects and potential clinical applications. Semin Nucl Med 52:759–780. https://doi.org/10.1053/j.semnuclmed.2022.04.004
8. Kandathil A, Kay FU, Butt YM, Wachsmann JW, Subramaniam RM (2018) Role of FDG PET/CT in the eighth edition of TNM staging of non-small cell lung cancer. Radiographics 38:2134–2149. https://doi.org/10.1148/rg.2018180060
9. Esteva A, Kuprel B, Novoa RA et al (2017) Dermatologist-level classification of skin cancer with deep neural networks. Nature 542:115–118. https://doi.org/10.1038/nature21056
10. Li X, Zhang S, Zhang Q et al (2019) Diagnosis of thyroid cancer using deep con-volutional neural network models applied to sonographic images: a retro-spective, multicohort, diagnostic study. Lancet Oncol 20:193–201. https://doi.org/10.1016/S1470-2045(18)30762-9
11. Rajpurkar P, Irvin J, Ball RL et al (2018) Deep learning for chest radiograph diagnosis: a retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLoS Med 15:e1002686. https://doi.org/10.1371/journal.pmed.1002686
12. Bi WL, Hosny A, Schabath MB et al (2019) Artificial intelligence in cancer imaging: clinical challenges and applications. CA Cancer J Clin 69:127–157. https://doi.org/10.3322/caac.21552
13. Ma X, Xia L, Chen J, Wan W, Zhou W (2023) Development and validation of a deep learning signature for predicting lymph node metastasis in lung adenocarcinoma: comparison with radiomics signature and clinical-semantic model. Eur Radiol 33:1949–1962. https://doi.org/10.1007/s00330-022-09153-z
14. Cong M, Feng H et al (2020) Development of a predictive radiomics model for lymph node metastases in pre-surgical CT-based stage IA non-small cell lung cancer. Lung Cancer 139:73–79. https://doi.org/10.1016/j.lungcan.2019.11.003
15. Sun Q, Lin X, Zhao Y et al (2020) Deep learning vs. radiomics for predicting axillary lymph node metastasis of breast cancer using ultrasound images: don't forget the peritumoral region. Front Oncol 10:53. https://doi.org/10.3389/fonc.2020.00053
16. Dong D, Fang MJ, Tang L et al (2020) Deep learning radiomic nomogram can predict the number of lymph node metastasis in locally advanced gastric cancer: an international multicenter study. Ann Oncol 31:912–920. https://doi.org/10.1016/j.annonc.2020.04.003
17. Lee JH, Ha EJ, Kim D et al (2020) Application of deep learning to the diagnosis of cervical lymph node metastasis from thyroid cancer with CT: external validation and clinical utility for resident training. Eur Radiol 30:3066–3072. https://doi.org/10.1007/s00330-019-06652-4
18. Tau N, Stundzia A, Yasufuku K, Hussey D, Metser U (2020) Convolutional neural networks in predicting nodal and distant metastatic potential of newly diagnosed non-small cell lung cancer on FDG PET images. AJR Am J Roentgenol 215:192–197. https://doi.org/10.2214/AJR.19.22346
19. Liu Y, Sun Y et al (2021) A survey on evolutionary neural architecture search. IEEE Trans Neural Netw Learn Syst. https://doi.org/10.1109/TNNLS.2021.3100554
20. van Griethuysen JJM et al (2017) Computational radiomics system to decode the radiographic phenotype. Can Res 77:e104–e107. https://doi.org/10.1158/0008-5472.CAN-17-0339
21. Woo S, Park J, Lee JY, Kweon IS, (2018) CBAM: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 3–19. https://doi.org/10.1007/978-3-030-01234-2_1
22. Lee SM, Park CM, Paeng JC et al (2012) Accuracy and predictive features of FDG-PET/CT and CT for diagnosis of lymph node metastasis of T1 non-small-cell lung cancer manifesting as a subsolid nodule. Eur Radiol 22:1556–1563. https://doi.org/10.1007/s00330-012-2395-4
23. Li CJ, Tian YL et al (2023) Usefulness of [68Ga]FAPI-04 and [18F]FDG PET/CT for the detection of primary tumour and metastatic lesions in gastrointestinal carcinoma: a comparative study. Eur Radiol 33:2779–2791. https://doi.org/10.1007/s00330-022-09251-y
24. Gao SJ, Kim AW et al (2017) Indications for invasive mediastinal staging in patients with early non-small cell lung cancer staged with PET-CT. Lung Cancer 109:36–41. https://doi.org/10.1016/j.lungcan.2017.04.018
25. Qi YM, Wu SS et al (2021) Development of nomograms for predicting lymph node metastasis and distant metastasis in newly diagnosed T1–2 non-small cell lung cancer: a population-based analysis. Front Oncol 11:683282. https://doi.org/10.3389/fonc.2021.683282
26. Real E, Aggarwal A, Huang Y, Le QV (2019) Aging evolution for image classifier architecture search. In: AAAI Conference on Artificial Intelligence, 2, p 2
27. Faes L, Wagner SK, Fu DJ et al (2019) Automated deep learning design for medical image classification by health-care professionals with no coding experience: a feasibility study. Lancet Digit Health 1:e232–e242. https://doi.org/10.1016/S2589-7500(19)30108-6
28. Yu Q, Yang D, et al. (2020) C2FNAS: Coarse–to-fine neural architecture search for 3D medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 4126–4135
29. Jiang H, Shen F, Gao F, Han W (2021) Learning efficient, explainable and discriminative representations for pulmonary nodules classification. Pattern Recogn 113:107825. https://doi.org/10.1016/j.patcog.2021.107825
30. Lambin P, Leijenaar RTH, Deist TM et al (2017) Radiomics: the bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol 14:749–762. https://doi.org/10.1038/nrclinonc.2017.141
31. Bradshaw TJ, Boellaard R, Dutta J et al (2022) Nuclear medicine and artificial intelligence: best practices for algorithm development. J Nucl Med 63:500–510. https://doi.org/10.2967/jnumed.121.262567
32. Coroller T, Yip S, Kim J et al (2016) SU-D-207B-03: A PET-CT radiomicscomparison to predict distant metastasis in lung adenocarcinoma. J Med Phys 43:3349–3349. https://doi.org/10.1118/1.4955671
33. Wu J, Aguilera T, Shultz D et al (2016) Early-stage non-small cell lung cancer: quantitative imaging characteristics of (18)F

fluorodeoxyglucose PET/CT allow prediction of distant metastasis. Radiology 281:270–278
34. Zhong Y, She Y, Multi-omics Classifier for Pulmonary Nodules (MISSION) Collaborative Group et al (2022) Deep learning for prediction of N2 metastasis and survival for clinical stage I non-small cell lung cancer. Radiology 302:200–211. https://doi.org/10.1148/radiol.2021210902
35. Rajpurkar P, Lungren MP (2023) The current and future state of AI interpretation of medical images. N Engl J Med 388:1981–1990. https://doi.org/10.1056/NEJMra2301725