Original Article

Performance of deep learning model and radiomics model for preoperative prediction of spread through air spaces in the surgically resected lung adenocarcinoma: a two-center comparative study

Xiang Wang[1#], Chao Ma[2#], Qinling Jiang[1#^], Xuebin Zheng[3], Jun Xie[3], Chuan He[3], Pengchen Gu[3], Yanyan Wu[1*], Yi Xiao[1*], Shiyuan Liu[1*]

[1]Department of Radiology, Second Affiliated Hospital of Naval Medical University, Shanghai, China; [2]Department of Radiology, Frist Affiliated Hospital of Naval Medical University, Shanghai, China; [3]Shanghai Aitrox Technology Corporation Limited, Shanghai, China

Contributions: (I) Conception and design: Y Wu, Y Xiao, S Liu; (II) Administrative support: S Liu; (III) Provision of study materials or patients: X Wang, C Ma; (IV) Collection and assembly of data: X Wang, C Ma; (V) Data analysis and interpretation: Q Jiang; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

[#]These authors contributed equally to this work as co-first authors.

*These authors contributed equally to this work.

Correspondence to: Shiyuan Liu, PhD; Yi Xiao, PhD; Yanyan Wu, BA. Department of Radiology, Second Affiliated Hospital of Naval Medical University, No. 415, Fengyang Road, Huangpu District, Shanghai 200003, China. Email: shczyy_lsy@163.com; xiaoyi@188.com; huxiyaya2014@163.com.

Background: Spread through air spaces (STAS) in lung adenocarcinoma (LUAD) is a distinct pattern of intrapulmonary metastasis where tumor cells disseminate within the pulmonary parenchyma beyond the primary tumor margins. This phenomenon was officially included in the World Health Organization (WHO)'s classification of lung tumors in 2015. STAS is characterized by the spread of tumor cells in three forms: single cells, micropapillary clusters, and solid nests. Clinical studies have linked STAS to a poorer prognosis, higher recurrence risk, and more advanced clinicopathological staging in LUAD patients. In this study, we constructed radiomics models and deep learning models based on computed tomography (CT) for predicting preoperative STAS status in LUAD.

Methods: A total of 395 (57.19±11.40 years old) patients with pathologically confirmed LUAD from two centers were enrolled in this retrospective study, in which STAS was detected in 146 patients (36.96%). The general clinical data, preoperative CT images, and the results of pathology reports of all patients were collected. Two experienced radiologists independently segmented the lesions by medical imaging interaction toolkit (MITK) software. The CT-based models only, the clinical-based models only, and the fusion model based on the two were constructed using radiomics and deep learning methods, respectively. The diagnostic performance of the different models was evaluated by comparing the area under the curve (AUC) of the subjects' receiver operating characteristics (ROCs).

Results: The deep learning model based on CT images achieved satisfactory discriminative performance in predicting STAS and outperformed the radiomics model and the clinical-radiomics model. The AUC of deep learning model was 0.918 for the internal test set and 0.766 for the external test set. The radiomics model had an AUC of 0.851 for the internal test set and an AUC of 0.699 for the external test set. The clinical-radiomics deep learning model was slightly less effective than the deep learning model (internal AUC =0.915, external AUC =0.773).

Conclusions: The constructed deep learning model based on preoperative chest CT can be used to determine the STAS status of LUAD patients with good diagnostic performance and is superior to radiomics models.

^ ORCID: 0009-0007-2930-6598.

Keywords: Spread through air spaces (STAS); computed tomography (CT); radiomics; deep learning

Submitted Jul 24, 2024. Accepted for publication Nov 26, 2024. Published online Dec 27, 2024.
doi: 10.21037/tlcr-24-646
View this article at: https://dx.doi.org/10.21037/tlcr-24-646

Introduction

Background

Lung cancer, a prevalent malignancy worldwide, holds the highest mortality rate of all tumors (1). According to the most recent data from the International Agency for Research on Cancer (IARC), lung cancer accounted for 11.4% of all cancer cases globally in 2020, claiming more than 1.8 million lives (2). The incidence of lung adenocarcinoma (LUAD) is surpassing squamous cell carcinoma, firmly establishing itself as the most diagnosed subtype of non-small-cell lung cancer (3). In 2015, the World Health Organization (WHO) classified 'spread through air spaces' (STAS) as a novel invasive pattern pertaining to adenocarcinoma (4). STAS impacts both postoperative recurrence and survival rates (5). Kadota *et al.* have shown that pulmonary lobectomy can significantly reduce postoperative recurrence in STAS-positive patients (6), thereby improving surgical outcomes for these patients. Multiple research studies support the notion that STAS indicates an adverse prognostic element for LUAD (7,8).

Rationale and knowledge gap

The factors affecting the occurrence of STAS continue to be a debated topic in numerous studies, highlighting the variability in incidence rates among different pathological subtypes of lung cancer. The current gold standard for assessing STAS status is postoperative pathological evaluation, characterized by the presence of tumor cells in the form of micropapillary clusters, solid nests, or individual cells within the alveolar spaces adjacent to the primary tumor (9). Nevertheless, the invasive nature and inherent limitations of postoperative sampling constrain its utility in informing preoperative surgical planning. Previous research demonstrates that intraoperative frozen section analysis for the detection of STAS exhibits low sensitivity and a low negative predictive value (10). Consequently, a precise preoperative evaluation of STAS status would significantly enhance the effectiveness of treatment planning.

Computed Tomography (CT) has become the imaging modality of choice for thoracic diseases. Recent studies underscore the predictive value of specific traditional CT morphological characteristics in determining STAS status. Research by de Margerie-Mellon *et al.* has identified nodule size and the proportion of the solid component within a nodule as potential indicators of STAS status (11). Furthermore, the research conducted by Toyokawa *et al.* has demonstrated a positive correlation between the size of ground glass opacity (GGO) on CT scans, the percentage of solid components, and the status of STAS (12). However, the evaluation of these conventional morphological characteristics is often subjective, leading to variability in radiologists' interpretations. Consequently, it is essential to develop an objective and precise model for the preoperative prediction of STAS status.

Objective

Radiomics and artificial intelligence (AI) utilize advanced image analysis algorithms, which can reflect the information of complete and heterogeneous tumors pixel by pixel by mining a large number of image features (13). In previous studies, our team has established AI-based models for the diagnosis of benign and malignant pulmonary nodules, infiltration grading of LUAD, prediction of lymph node metastasis, gene mutation prediction, and prognosis assessment (14-19), which have played a significant role in the diagnosis and treatment of lung cancer. In this research, we studied the performance of deep learning and machine learning models in multi-modal data and variable fusion strategies to predict STAS status in LUAD. We present this article in accordance with the STARD reporting checklist (available at https://tlcr.amegroups.com/article/view/10.21037/tlcr-24-646/rc).

Methods

Patients

The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The study was approved by ethics board of Changzheng Hospital (No.

2024SL005). Changhai Hospital was informed and agreed with this study. Individual consent for this retrospective analysis was waived. Two cohorts of clinical, imaging, and pathological data of LUAD patients were collected from June 2014 to December 2022. The inclusion criteria included the followings: (I) postoperative pathological confirmation of invasive adenocarcinoma or minimally invasive adenocarcinoma; (II) patients who underwent CT examination within one month before surgery (layer thickness of ≤1 mm); (III) availability of complete clinical and pathological information. The exclusion criteria involved the followings: (I) serious artifacts in the image caused by misalignment, metal, respiratory motion, etc., making it difficult to accurately identify lesions; (II) preoperative antitumor treatment such as radiotherapy and chemotherapy.

CT examinations

All patients were scanned in supine position at the end of normal inspiration and breath-hold using CT machines including the Philips iCT 256CT (Philips, Eindhoven, the Netherlands), Philips Ingenuity (Philips), GE LightSpeed VCT (GE Healthcare, Piscataway, New Jersey, USA), TOSHIBA 512CT and TOSHIBA Aquilion (Toshiba, Tokyo, Japan). The scanning parameters of routing CT: tube voltage, 120 kV; tube current, 50 to 150 mAs or auto mAs. The reconstruction algorithms were standard and sharp with 0.625 to 1 mm slice thickness. The scan range covered from the lung apex to the lower margin of the posterior costophrenic angle.

The radiological features were independently evaluated by two experienced thoracic radiologists with 6 and 12 years of experience in chest CT interpretation, respectively, who were blinded to the pathological results. Any discrepancies in interpretation were resolved through consensus, and in case of persistent disagreement, a more senior radiologist with greater expertise in the field was consulted to reach a consensus. The CT morphological features analyzed for each pulmonary tumor included (I) tumor-lung interface (clear/unclear); (II) lobulation (absent/present); (III) spiculation (absent/present); (IV) vacuole sign (absent/present); (V) peritumoral emphysema (absent/present); (VI) air bronchogram (absent/present); (VII) pleural indentation (absent/present); (VIII) maximum diameter; and (IX) density (pure ground-glass density/mixed ground-glass density/solid density). The CT interpretation criteria of primary tumors refer to our previous research (18). All CT findings were assessed based on high-resolution CT (HRCT) images using a reconstruction lung window (width: 1,500 HU, level: −450 HU) and a mediastinal window (width: 300 HU, level: 60 HU).

Tumor segmentation

Tumor segmentation was performed utilizing the freely available open-source software medical imaging interaction toolkit (MITK; v2021.02, https://www.mitk.org/). The chest CT scan medical digital imaging and communications in medicine (DICOM) format images were imported into the software for delineation. Two radiologists with 6 years of experience in chest CT interpretation manually delineated the volumes of interest (VOIs) at the pixel level. Subsequently, the VOIs were confirmed by a third radiologist with 13 years of experience in chest CT interpretation.

Examination of STAS

The surgically resected tumor specimen was fixed in 10% formaldehyde solution, followed by embedding in paraffin wax. Subsequently, paraffin sections with a thickness of 5 micrometers were prepared. These sections were then stained with hematoxylin-eosin (HE) for histopathological examination by two senior pathologists. The pathological diagnosis was conducted in accordance with the WHO lung tumor classification guidelines. The detection of STAS is defined as the appearance of tumor cells in the normal alveolar space away from the main lesion in the form of microcell clusters, small cancer nests, or single cells (9). Disseminated tumor cells must not exhibit direct attachment to the main lesion. To avoid misidentification of the cells artificially isolated during tumor cleavage as STAS, it is imperative to observe a minimum of three slides per case under a microscope, ensuring the validity of the boundaries observed in each slide.

Radiomics feature extract

In this study, all acquired CT images were in DICOM format. The mask DICOM files containing 512×512 CT images for each patient were reconstructed into n×512×512 MetaImage Header and Archive (MHA) format datasets, with 'n' representing the number of image slices. We use Python to implement PyRadiomics and extract features from CT images. The MHA datasets for both the CT scans and masks were then subjected to the Pyradiomics feature extraction suite. By configuring specific filters and feature types via ImageTypes and Features settings, an advanced feature computation was performed. A comprehensive analysis was conducted using a suite of seven filters—original, wavelet, log, square, square root, exponential, and logarithm—resulting in the generation of 1,251 conventional radiomics features per mask, which delineated the lesion areas.

Following the feature extraction, a refined selection process was implemented using the sklearn.feature_selection.SelectKBest function to streamline the extensive feature set based on their relevance to STAS positivity. Through this process, five radiomics features were identified as the most significant correlates: original firstorder_Median, wavelet-LLL firstorder RootMeanSquared, wavelet-LLL firstorder_Median, original_firstorder_RootMeanSquared, and original_firstorder_Mean. These salient features were selected based on the model.scores and model.pvalues criteria, where higher model.scores_ indicate greater importance and smaller model.pvalues_ reflect higher confidence, aligning the order of feature importance derived from both parameters. These five features were subsequently used as inputs for the construction of the radiomics models to effectively predict the STAS status.

Clinical data analysis

In the present study, a total of 11 clinical data for each patient were collected from each subject, as detailed in *Table 1*. Through a multivariate analysis, we narrowed down the significant information to four key factors: age, spiculation, vacuole sign, and density (with further details provided in Appendix 1). Conventionally, clinical data is widely utilized in radiomics models. However, a notable oversight among researchers is the screening of clinical data. In this regard, our investigation delved into the impact of employing both the complete clinical data set and the statistically significant subset.

STAS status prediction with conventional radiomics

Three widely employed machine learning classifiers, namely logistic regression, random forest, and support vector machine (SVM) were chosen for the purpose of constructing conventional radiomics models. Four sets of features were identified for analysis, encompassing the five most pertinent radiomics features associated with STAS positivity, eleven clinical features, four additional clinical features, and a combination of both radiomics and clinical data. For each set of features, the three selected classifier models were individually established. The model was trained in the training cohort, and a 5-fold cross-validation was used to determine the parameters of the model.

STAS status prediction with 3D Deep Convolutional Neural Networks (DCNN) model

A CT scan was conducted on each patient, generating a set of 512×512 n slices, where n denotes the number of slices per patient. Subsequently, the CT images were converted into the MetaImage (MHA) format, a widely accepted standard for storing medical imagery data. The tumor's locations and dimensions in each case were obtained according to the annotated mask. To extract the CT image data associated corresponding to the tumor's position, a cubic volume centered on the tumor was constructed, encompassing both the tumor and the surrounding tissue. The cubic volume's size was established to be 1.2 times the tumor's size, to guarantee that enough contextual information was included for precise analysis. Given the disparities in tumor dimensions, all intercepted tumors were standardized to a uniform dimension of 32×64×64 by torchio.transforms.resize function. Furthermore, data augmentation was achieved by implementing flipping and introducing random Gaussian noise to double the training data size.

Four 3D DCNN architectures were employed in the DCNN models to assess the utilization and integration of various data modalities. These architectures encompassed 3D ResNet18 (Figure S1), 3D ResNet50 (Figure S2), 3D DenseNet121 (Figure S3), and 3D DenseNet201 (Figure S4). During the optimization of each model, a batch size of 16 was maintained for training purposes. The models' parameters were iteratively updated using the Adam optimizer with exponential decay. The training process was terminated after 150 epochs. Additionally, the dropout layer parameter within the network was 0.5.

STAS status prediction with 2.5D DCNN model

The DenseNet architecture, which has demonstrated high performance in processing medical images, was employed in training the 2.5D DCNN models. Consistent with its application in 3D DCNN models, the cubic volume's dimensions were set to be 1.2 times the size of the tumor.

Table 1 Baseline characteristics of the participants

Characteristics	All patients	Training	Internal test	P value
Number	315	250	65	
STAS				0.61
Positive	120	97	23	–
Negative	195	153	42	–
Gender				0.87
Female	187	149	38	–
Male	128	101	27	–
Age, years (mean ± SD)	57.19±11.40	57.09±11.87	57.55±9.35	0.77
Maximum diameter, mm	18.85±10.52	18.44±10.35	20.41±11.00	0.18
Tumor-lung interface				0.23
Clear	289	227	62	–
Unclear	26	23	3	–
Pericystic emphysema				0.27
Positive	12	8	4	–
Negative	303	242	61	–
Pleural Indentation				0.22
Positive	120	91	29	–
Negative	195	159	36	–
Spiculation				0.49
Positive	42	35	7	–
Negative	273	215	58	–
Lobulation				0.87
Positive	259	206	53	–
Negative	56	44	12	–
Vacuole sign				0.4
Positive	61	46	15	–
Negative	254	204	50	–
Air bronchogram				0.03
Positive	138	102	36	–
Negative	177	148	29	–
Density				0.87
pGGO	99	79	20	–
mGGO	132	103	29	–
Solid	84	68	16	–

STAS, spread through air spaces; SD, standard deviation; pGGO, pure ground grass opacity; mGGO, mixed ground grass opacity.

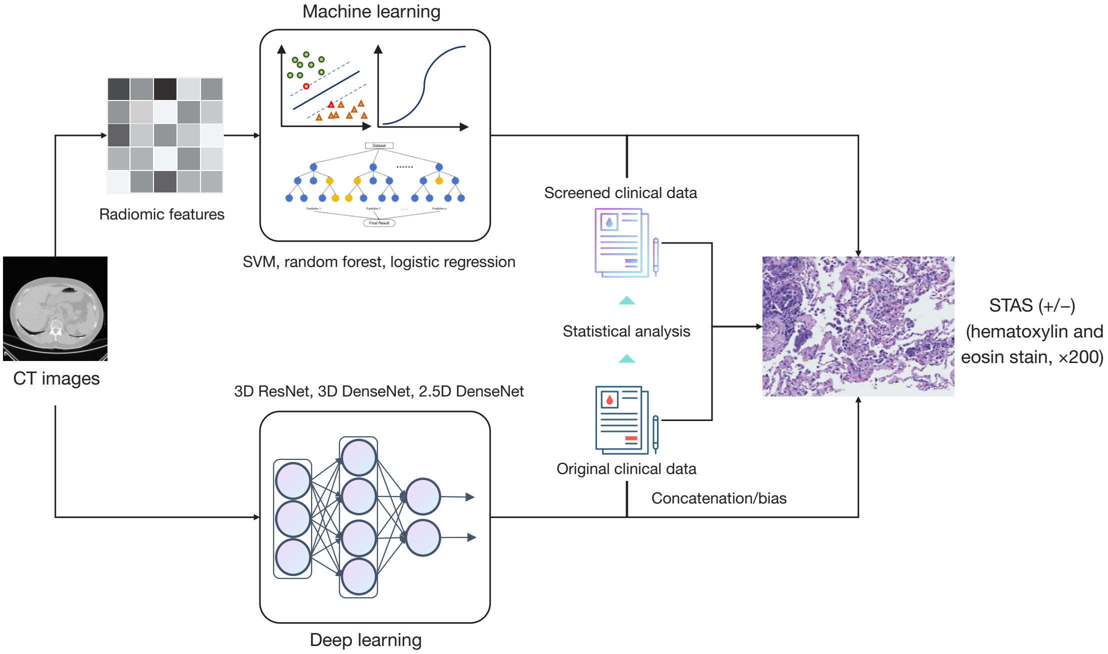

Figure 1 The flow chart of our study design. SVM, support vector machine; STAS, spread through air spaces; CT, computed tomography; D, dimension.

Across the center of the cube, perpendicular planar tumor images were captured from the axial, coronal, and sagittal planes. Furthermore, two oblique intersections (45° and −45°) were included for each plane. These three vertical planar images constituted a set of three-channel image data, and each tumor was capable of yielding three such sets for training the 2.5D DenseNet121 model (Figure S5). Additionally, to augment the training set, samples were duplicated by flipping and introducing random Gaussian noise. During the optimization process, the models use a batch size of 16 for training. The Adam optimizer, which will reduce the learning rate by 0.8 times when the loss function does not decrease for 5 consecutive epochs, was used to update the model parameters. Training was terminated after 150 epochs.

Multi-modal features fusion strategies

By integrating clinical features into deep learning models, it is possible to achieve more precise classification outcomes. However, the integration of clinical features with the basic block of DCNN models can be both complex and costly. Therefore, we incorporated the clinical features into the feature space, which can be accomplished through two methods: concatenation and bias.

In the concatenation method (Figure S6), the clinical features are concatenated to the feature vectors of deep learning models after the dropout layer of the main network channel. Subsequently, the classification outcomes are generated by the fully connected layer. Prior to concatenation, the clinical data undergoes batch normalization to mitigate the influence of diverse distributions, considering the inclusion of variables such as age and maximum diameter, and Boolean values like gender and vacuole sign. In the method of bias (Figure S7), the clinical data is transformed into the same size of the deep learning feature vectors through the fully connected layer, and the clinical features are added to the features layer in the form of bias. The model classification results are then derived through the subsequent dropout layer and fully connected layer. The flow chart of our study design is demonstrated in *Figure 1*.

Statistical analysis

The difference of clinical data from different datasets was compared by *t*-test or Mann-Whitney *U* test. The receiver operating characteristic (ROC) curve and the area under the curve (AUC) were introduced as graphical representations to showcase the balance between true positive and false

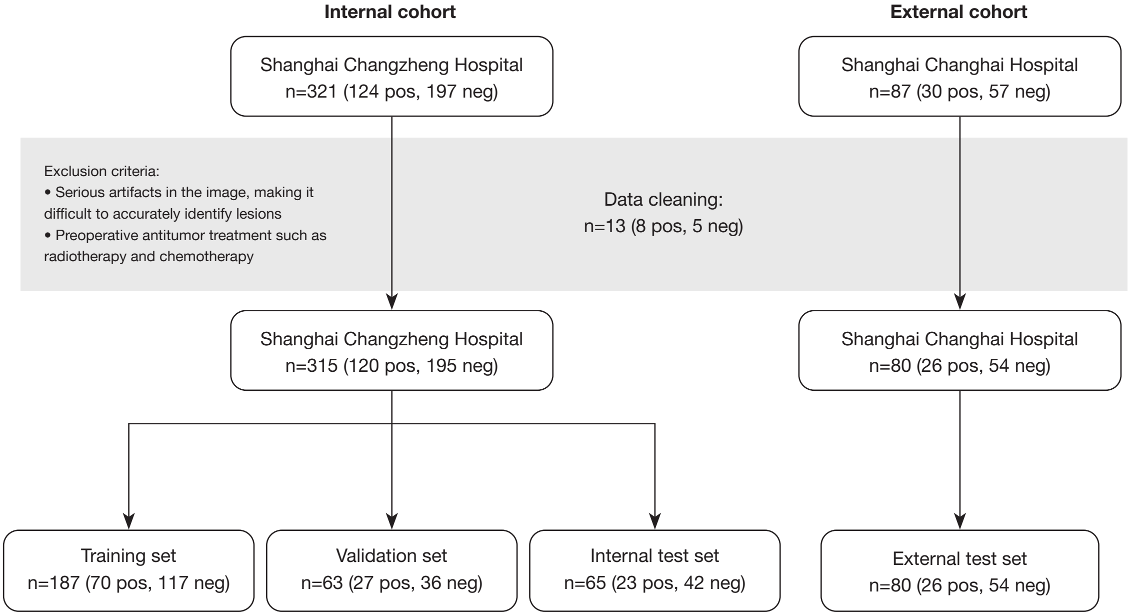

Figure 2 Data organization workflow. pos, positive; neg, negative.

positive rates. In addition, accuracy, recall, precision, and F1 scores were calculated to comprehensively assess the performance of the model. Cohen's kappa was also used to measure the agreement between human experts and the AI system with respect to the predictions made on the test sets. All statistics tests were two side and a P value less than 0.05 was considered statistically significant. The statistical analysis was conducted using MedCalc (Version 22.009, MedCalc Software Ltd., Ostend, Belgium).

Results

Baseline characteristics

This study included a total of 395 patients, who were divided into STAS-negative group (249 cases) and STAS-positive group (146 cases) based on postoperative pathological data. Cohort 1 included data from 315 cases in Shanghai Changzheng Hospital, and Cohort 2 included 80 cases in Shanghai Changhai Hospital. Cohort 1 were randomly divided into training set (n=187), validation set (n=63), and internal test set (n=65) according to 3:1:1 proportion. The data organization workflow is demonstrated in *Figure 2*. The clinical data of the participants, including age, gender, maximum tumor diameter, density, tumor-lung interface, peritumoral pulmonary emphysema, pleural indentation sign, spiculation, lobulation, vacuole sign, air bronchogram, and histopathological evidence were recorded in *Table 2*. There was no significant difference between the detailed characteristics of the two cohorts (all P>0.05 except for air bronchogram, *t*-test or Mann-Whitney *U* test).

Prediction of STAS status of models with CT images only

In the internal test set, 3R18DODA (3D Resnet18 architecture with dropout 0.5 and data augmentation) achieved the highest AUC of 0.918. This model demonstrates superior accuracy, precision, recall, and F1 score compared to most other methods, suggesting its effectiveness in handling the given task, while other 3D deep learning based models (3D ResNet50, 3D DenseNet121, and 3D DenseNet201) achieved AUCs ranging from 0.85 to 0.89 (all the abbreviations and counterparts of models are summarized in Table S1 of Appendix 2). All machine learning radiomics demonstrated relatively lower performance with AUCs below 0.85. Notably, RR (random forest with radiomics) achieved an AUC of 0.790, which is significantly lower than that of 3R18DODA [P=0.0063, Delong *et al.* (20)]. Similarly, 2.5D DenseNet also exhibited a relatively lower prediction performance than the highest one as well (AUC: 0.804, P=0.0154, Delong *et al.*).

In the external test set, the highest AUC is achieved by 3R50 (3D Resnet50) with a value of 0.789. However, it is noteworthy that there is no significant difference in

Table 2 The prediction of deep learning models and radiomics models with CT images only

Model	Internal test set					External test set				
	Precision	Recall	F1 score	Acc	AUC (95% CI)	Precision	Recall	F1 score	Acc	AUC (95% CI)
RF + radiomics	0.74	0.69	0.69	0.72	0.790[a] (0.671, 0.881)	0.59	0.6	0.59	0.62	0.715 (0.603, 0.810)
2.5D densenet121 + DA	0.74	0.50	0.60	0.65	0.804[b] (0.687, 0.892)	0.38	0.45	0.42	0.65	0.722 (0.611, 0.817)
SVM + radiomics	0.83	0.79	0.8	0.82	0.847 (0.736, 0.924)	0.73	0.74	0.74	0.76	0.697 (0.584, 0.795)
3D ResNet50 + DA	0.61	0.67	0.64	0.75	0.850 (0.740, 0.926)	0.69	0.62	0.65	0.76	0.759 (0.651, 0.848)
LR + radiomics	0.82	0.80	0.81	0.82	0.851 (0.741, 0.927)	0.70	0.72	0.70	0.73	0.699 (0.587, 0.797)
3D ResNet50	0.61	0.67	0.64	0.75	0.863 (0.756, 0.936)	0.54	0.61	0.57	0.74	0.789 (0.684, 0.872)
3D ResNet50 + dropout 0.5 + DA	0.61	0.78	0.68	0.8	0.871 (0.764, 0.941)	0.50	0.59	0.54	0.73	0.759 (0.651, 0.848)
3D DenseNet121 + DA	0.57	0.72	0.63	0.77	0.884 (0.780, 0.950)	0.65	0.59	0.62	0.74	0.769 (0.662, 0.856)
3D ResNet18	0.65	0.68	0.67	0.77	0.885 (0.782, 0.951)	0.65	0.59	0.62	0.74	0.768 (0.660, 0.855)
3D DenseNet201	0.52	0.67	0.59	0.74	0.886 (0.783, 0.952)	0.69	0.62	0.65	0.76	0.775 (0.668, 0.861)
3R18DA + DA	0.61	0.70	0.65	0.77	0.890 (0.788, 0.954)	0.65	0.57	0.61	0.73	0.767 (0.659, 0.854)
3D ResNet18 + dropout 0.5 + DA	0.78	0.75	0.77	0.83	0.918 (0.823, 0.972)	0.73	0.61	0.67	0.76	0.766 (0.658, 0.854)

[a], indicates P=0.0063, Delong *et al.* in comparison with 3D ResNet18 + dropout 0.5 + DA in internal test set. [b], indicates P=0.0154, Delong *et al.* in comparison with 3D ResNet18 + dropout 0.5 + DA in internal test set. CT, computed tomography; Acc, accuracy; AUC, area under the receiver operating characteristic curve; CI, confidence interval; RF, random forest; DA, data augmentation; SVM, support vector machine; LR, logistic regression; D, dimension.

performance among all the models tested (P>0.05, Delong *et al.*). The detailed statistical results were summarized in *Table 2*.

Prediction of STAS status of models with clinical data only

During the data collection stage, a total of 11 clinical data were collected from each patient. Subsequently, a multivariate analysis was conducted to identify four key clinical data points from the initial dataset, as detailed in Appendix 1. Then the performance of different radiomics models using entire clinical data and the narrowed-down one was compared.

In the internal test set, two models with 4 significant clinical data (SVMs and logistic regression) achieved AUC of 0.899 and 0.882, respectively, exhibiting superior performance. In contrast, R4C (random forest with 4 significant clinical data) demonstrated an AUC of 0.814, which is significantly different with the highest one (P=0.0224). Models with 11 clinical data demonstrated generally lower performance, with AUC values ranging from 0.768 to 0.878. Notably, the AUC scores of models in the external test set did not exhibit significant differences, with minimal variance observed, ranging from 0.726 to 0.771. The detailed statistical results were summarized in *Table 3*.

Prediction of STAS status of models with multi-modal data

Multi-modal data is widely regarded as a crucial factor for improving the efficacy of deep learning. The present research conducted a comparison between various deep learning and radiomics models with different fusion strategies. In the internal test set, the performance of all the 4 deep learning models—3R18DODA11CC (3D Resnet18 + dropout 0.5 + data augmentation + CT images + 11 clinical data with concatenation method), 3R18DODA11CB (3D Resnet18 + dropout 0.5 + data augmentation + CT images + 11 clinical data with bias method), 3R18DODA4CB (3D Resnet18 + dropout 0.5 + data augmentation + CT images + 4 significant clinical data with bias method), and 3R18DODA4CB (3D

Table 3 The prediction of models with clinical data only

Model	Internal test set					External test set				
	Precision	Recall	F1 score	Acc	AUC (95% CI)	Precision	Recall	F1 score	Acc	AUC (95% CI)
SVM + 11 clinical data	0.81	0.69	0.69	0.74	0.768 (0.647, 0.864)	0.65	0.59	0.59	0.70	0.726 (0.615, 0.820)
RF + 4 significant clinical data	0.78	0.79	0.78	0.80	0.814[a] (0.698, 0.899)	0.65	0.65	0.56	0.56	0.75 (0.641, 0.840)
RF + 11 clinical data	0.82	0.71	0.71	0.75	0.866 (0.759, 0.938)	0.70	0.73	0.70	0.71	0.762 (0.653, 0.850)
LR + 11 clinical data	0.86	0.83	0.83	0.85	0.878 (0.773, 0.946)	0.68	0.71	0.67	0.68	0.769 (0.662, 0.856)
LR + 4 significant clinical data	0.85	0.84	0.85	0.86	0.882 (0.778, 0.949)	0.73	0.76	0.71	0.71	0.771 (0.663, 0.857)
SVM + 4 significant clinical data	0.83	0.83	0.83	0.85	0.899 (0.798, 0.960)	0.69	0.70	0.69	0.73	0.762 (0.654, 0.850)

[a], indicates P=0.0224, Delong *et al.* in comparison with SVM + 4 significant clinical data in internal test set. AUC, area under the receiver operating characteristic curve; Acc, accuracy; CI, confidence interval; SVM, support vector machine; RF, random forest; LR, logistic regression.

Table 4 The prediction of models with multi-modal data

Model	Internal test set					External test set				
	Precision	Recall	F1 score	Acc	AUC (95% CI)	Precision	Recall	F1 score	Acc	AUC (95% CI)
LR + radiomics + 11 clinical data	0.8	0.76	0.76	0.78	0.839[a] (0.726, 0.918)	0.68	0.7	0.66	0.68	0.752 (0.643, 0.842)
SVM + radiomics + 11 clinical data	0.83	0.79	0.8	0.82	0.847 (0.736, 0.924)	0.73	0.74	0.74	0.76	0.697 (0.584, 0.795)
RF + radiomics + 11 clinical data	0.79	0.77	0.78	0.79	0.850 (0.740, 0.926)	0.65	0.67	0.63	0.64	0.761 (0.653, 0.850)
3D ResNet18 + dropout 0.5 + data augmentation + 4 significant clinical data (concatenation method)	0.87	0.8	0.83	0.88	0.893 (0.792, 0.956)	0.65	0.55	0.60	0.71	0.748 (0.638, 0.838)
3D ResNet18 + dropout 0.5 + data augmentation + 4 significant clinical data (bias method)	0.70	0.76	0.73	0.82	0.898 (0.797, 0.959)	0.69	0.58	0.63	0.74	0.750 (0.641, 0.840)
3D ResNet18 + dropout 0.5 + data augmentation + 11 clinical data (bias method)	0.78	0.78	0.78	0.85	0.913 (0.817, 0.969)	0.65	0.61	0.63	0.75	0.750 (0.641, 0.840)
3D ResNet18 + dropout 0.5 + data augmentation + 11 clinical data (concatenation method)	0.65	0.79	0.71	0.82	0.915 (0.819, 0.970)	0.58	0.60	0.59	0.74	0.773 (0.665, 0.859)

[a], indicates P=0.0364, Delong *et al.* in comparison with 3R18DODA11CC in internal test set. Acc, accuracy; AUC, area under the receiver operating characteristic curve; CI, confidence interval; LR, logistic regression; SVM, support vector machine; RF, random forest; D, dimension.

Resnet18 + dropout 0.5 + data augmentation + CT images + 4 significant clinical data with bias method)—in predicting STAS status was highly impressive, with AUCs ranging from 0.893 to 0.915. Nevertheless, the radiomics models only achieved AUCs not exceeding 0.850. Notably, the AUC of LRC (logistic regression + radiomics + 11 clinical data) was significantly lower than the highest performing radiomics model. This situation was not observed in the external test set that the AUCs exhibited a slight variance, with no statistically significant differences observed. The detailed statistical results were summarized in *Table 4*.

Comparison of deep learning models and radiomics models in predicting STAS

The ROC curves of 6 models with highest performance for both deep learning and radiomics are presented in *Figure 3*. Radiomics achieved AUCs exceeding 0.850 and

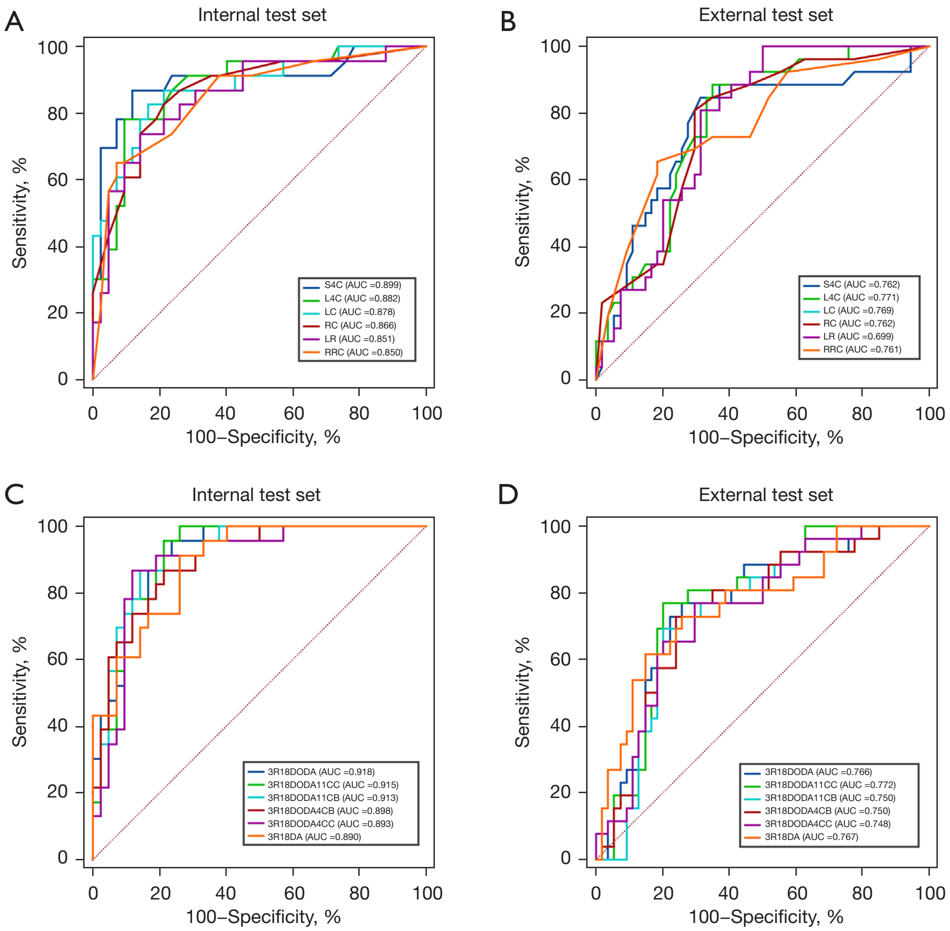

Figure 3 Performances for STAS prediction. The ROC curves of high-performance machine learning models in internal test set (A), external test set (B), and high-performance deep learning models in internal test set (C), and external test set (D). AUC, area under the curve; STAS, spread through air spaces; ROC, receiver operating characteristic.

0.699 in the internal test set (*Figure 3A*) and external test set (*Figure 3B*), respectively. Notably, SVM and logistic regression, both utilizing four significant clinical data, exhibited the top 2 highest performance in the internal test set, and the models with clinical data only occupy the top 4 performance, outperforming the models with CT images only and the models with multi-modal data. Similarly, the top 4 performance also occurred in the models with clinical data only in the external test set.

Deep learning models, on the other hand, demonstrated generally higher performance than the radiomics models in both the internal test set (*Figure 3C*) and external test set (*Figure 3D*). Interestingly, the highest AUC was achieved by 3D Resnet18 + dropout 0.5 + data augmentation, which is without the inclusion of clinical data in the internal test set. However, the multi-modal data models also demonstrated impressive AUCs, which are higher than 0.91 in both 3R18DODA11CC (3D Resnet18 + dropout 0.5 + data augmentation + CT images + 11 clinical data with concatenation method) and 3R18DODA11CB (3D Resnet18 + dropout 0.5 + data augmentation + CT images + 11 clinical data with bias method). 3R18DODA11CC achieved the highest AUC in the external test set. Unlike the trend in radiomics, the inclusion of four significant clinical data did not result in enhanced model performance in either the internal or external test sets. The ROC curves of the models besides those with highest performance for both deep learning and machine learning are presented in Figure S8.

Discussion

Key findings

This investigation aims to assess the efficacy of the CT-based models only, the clinical-based models only, and the fusion model based on the two using radiomics and deep learning methods, respectively, in predicting the STAS in LUAD. Our findings suggest that deep learning models exhibit promising utility in preoperatively evaluating STAS in LUAD patients, displaying robust diagnostic accuracy superior to radiomics models. Among the deep learning models, those solely utilizing CT images demonstrate the highest predictive capability, followed by the image-clinical fusion models. These findings are further substantiated by a comparison of human and AI performance (see section 'Comparison of AI performance with that of human readers' in Appendix 2 and Figure S9), in which deep learning model with CT images demonstrates the best performance to serve as a reliable tool for STAS diagnosis. Conversely, in the realm of radiomics models, the model relying solely on clinical features demonstrates optimal performance, with the subsequent radiomics model based solely on CT image features trailing closely behind.

Explanations of findings

The AUC of the radiomics model based exclusively on CT images registers at 0.851 in the internal test set and 0.699 in the external test set, underscoring the robust predictive capacity of radiomics features for lung cancer STAS. Chen *et al.* (21) analyzed CT images from 233 LUAD patients, yielding a model with an AUC of 0.63. While radiomics demonstrated some predictive capability in this context, the model's AUC reflected suboptimal performance. Onozato *et al.* (22) leveraged radiomics to forecast tumor sizes ≤2 cm in non-small cell lung cancer patients, achieving a model AUC of 0.77. Bassi *et al.* (23) augmented the predictive efficiency of STAS by amalgamating radiological signs and radiomics features, notably attaining a collective AUC of 0.79 in a prospective validation cohort of 50 cases, further affirming the dependability of radiomics. Nevertheless, these aforementioned studies solely relied on CT image data without integrating clinical features, and the highest AUC attained exhibited only moderate accuracy.

In terms of radiomics models, the predictive model based on four clinical features (age, spiculation, vacuole sign, and density) demonstrated the highest predictive efficiency, achieving an AUC of 0.899 in the internal test set and 0.762 in the external test set. These clinical findings regarding STAS prediction are consistent with prior studies. Qi *et al.* (24) corroborated CT image features' predictive role in STAS, encompassing tumors, satellites, ground glass ribbon sign, pleural attachment, and unclear tumor-lung interface. Among these, spiculation displayed robust discriminatory ability, along with notable correlations observed for consolidation tumor ratio (CTR) and tumor density. Gu *et al.* (25) further illustrated the association of spiculation with STAS+ tumors, contrasting STAS– adenocarcinomas. Jiang *et al.* (14) revealed a significant age-STAS correlation, developing a radiomics-age prediction model with an AUC of 0.754. Notably, models utilizing all 11 clinical data points generally exhibited lower performance, suggesting that additional clinical data may not necessarily enhance predictive accuracy and could even introduce noise, hindering accurate predictions.

The efficacy of the radiomics model based on clinical features (AUC of 0.899 in the internal test set and 0.762 in the external test set) surpassed that of the image-based radiomics model (AUC of 0.851 in the internal test set and 0.699 in the external test set). This aligns with Suh *et al.*'s findings (26), where the clinical feature model (AUC of 0.784) outperformed the radiomics model (AUC of 0.731) in STAS prediction. However, Han *et al.*'s study contradicted these results (27). Analyzing 395 stage IA LUAD patients, their radiomics CT model (AUC of 0.812) exhibited satisfactory performance in early LUAD preoperative STAS prediction, surpassing the clinical CT model (AUC of 0.721). Zhuo *et al.* (28) proposed an imaging-radiomics nomogram, showing promising predictive performance, but overshadowed by the high predictive ability of clinical features. Variations in results could stem from differences in subject inclusion criteria and clinical feature selection.

In terms of deep learning models, the 3R18DODA model showcased the highest AUC in the internal testing set, alongside superior accuracy, precision, recall, and F1 scores compared to other methods. This suggests deep learning's effectiveness in extracting CT image information for STAS prediction. Conversely, 2.5D DenseNet exhibited relatively lower performance, emphasizing the importance of utilizing 3D information for accurate STAS prediction. Notably, the number of researches on predicting STAS status using deep learning remains limited. Tao *et al.* (29) developed five models, with the 3D Convolutional Neural Network (CNN) model outperforming others, albeit lacking external validation for generalization assessment.

In our study, we incorporated an external test set, yielding AUC values of 0.918 and 0.766 in the internal test set and external test set, respectively, indicative of good predictive performance. However, the AUC of the deep learning fusion model was lower than that of the CT image-based deep learning model, potentially due to significant spatial redundancy between the two features, negatively impacting performance. Similar observations were noted in Tao *et al.*'s (29) study.

Strengths and limitations

There are several limitations in this study. Firstly, being retrospective, it only encompasses patients undergoing surgical tumor resection, potentially introducing selection bias. Secondly, as CT scans were conducted across different centers and scanners, slight variations in image acquisition protocols may exist, possibly leading to biases. Thirdly, due to the small sample size, a dedicated classification study of density or size focusing on tumours was lacking. Lastly, while some studies have indicated the effectiveness of STAS in predicting early recurrence of stage I LUAD (30), the absence of follow-up data precluded further evaluation of STAS's impact on patient prognosis.

Conclusions

This study underscores the remarkable performance of deep learning models leveraging CT images in predicting STAS status, outperforming radiomics-based predictions. These findings hold potential for preoperative STAS prediction in LUAD patients, facilitating informed clinical decision-making regarding surgical interventions and thereby advancing the clinical integration of precision medicine.

Acknowledgments

Funding: This work was supported by Special Project for Promoting High-Quality Development of Industries in Shanghai, 2022-2023 (Artificial Intelligence Topic) (Grant No. 2023-GZL-RGZN-01014), and the National Natural Science Foundation of China (No. 82271994).

Footnote

Reporting Checklist: The authors have completed the STARD reporting checklist. Available at https://tlcr.amegroups.com/article/view/10.21037/tlcr-24-646/rc

Data Sharing Statement: Available at https://tlcr.amegroups.com/article/view/10.21037/tlcr-24-646/dss

Peer Review File: Available at https://tlcr.amegroups.com/article/view/10.21037/tlcr-24-646/prf

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://tlcr.amegroups.com/article/view/10.21037/tlcr-24-646/coif). All the authors report that this work was supported by Special Project for Promoting High-Quality Development of Industries in Shanghai, 2022-2023 (Artificial Intelligence Topic, Grant No. 2023-GZL-RGZN-01014), and the National Natural Science Foundation of China (No. 82271994). X.Z., J.X., C.H., and P.G. are from Shanghai Aitrox Technology Co. Ltd. The authors have no other conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The study was approved by ethics board of Changzheng Hospital (No. 2024SL005). Changhai Hospital was informed and agreed with this study. Individual consent for this retrospective analysis was waived.

References

1. Zheng RS, Chen R, Han BF, et al. Cancer incidence and mortality in China, 2022. Zhonghua Zhong Liu Za Zhi 2024;46:221-31.
2. Siegel RL, Miller KD, Fuchs HE, et al. Cancer statistics, 2022. CA Cancer J Clin 2022;72:7-33.
3. IARC. Cancer Today. 2022. Available online: https://gco.iarc.fr/today/online-analysis-pie
4. Travis WD, Brambilla E, Nicholson AG, et al. The

2015 World Health Organization Classification of Lung Tumors: Impact of Genetic, Clinical and Radiologic Advances Since the 2004 Classification. J Thorac Oncol 2015;10:1243-60.

5. Lee JS, Kim EK, Kim M, et al. Genetic and clinicopathologic characteristics of lung adenocarcinoma with tumor spread through air spaces. Lung Cancer 2018;123:121-6.
6. Kadota K, Kushida Y, Kagawa S, et al. Limited Resection Is Associated With a Higher Risk of Locoregional Recurrence than Lobectomy in Stage I Lung Adenocarcinoma With Tumor Spread Through Air Spaces. Am J Surg Pathol 2019;43:1033-41.
7. Shiono S, Endo M, Suzuki K, et al. Spread Through Air Spaces Is a Prognostic Factor in Sublobar Resection of Non-Small Cell Lung Cancer. Ann Thorac Surg 2018;106:354-60.
8. Warth A, Muley T, Kossakowski CA, et al. Prognostic Impact of Intra-alveolar Tumor Spread in Pulmonary Adenocarcinoma. Am J Surg Pathol 2015;39:793-801.
9. Yokoyama S, Murakami T, Tao H, et al. Tumor Spread Through Air Spaces Identifies a Distinct Subgroup With Poor Prognosis in Surgically Resected Lung Pleomorphic Carcinoma. Chest 2018;154:838-47.
10. Kameda K, Lu S, Eguchi T, et al. MA12.05 Can Tumor Spread through Air Spaces (STAS) in Lung Adenocarcinomas Be Predicted Pre- and Intraoperatively? J Thoracic Oncol 2017;12:S411-2.
11. de Margerie-Mellon C, Onken A, Heidinger BH, et al. CT Manifestations of Tumor Spread Through Airspaces in Pulmonary Adenocarcinomas Presenting as Subsolid Nodules. J Thorac Imaging 2018;33:402-8.
12. Toyokawa G, Yamada Y, Tagawa T, et al. Significance of Spread Through Air Spaces in Resected Pathological Stage I Lung Adenocarcinoma. Ann Thorac Surg 2018;105:1655-63.
13. Binczyk F, Prazuch W, Bozek P, et al. Radiomics and artificial intelligence in lung cancer screening. Transl Lung Cancer Res 2021;10:1186-99.
14. Jiang Q, Sun H, Deng W, et al. Super Resolution of Pulmonary Nodules Target Reconstruction Using a Two-Channel GAN Models. Acad Radiol 2024;31:3427-37.
15. Wang X, Zhao X, Li Q, et al. Can peritumoral radiomics increase the efficiency of the prediction for lymph node metastasis in clinical stage T1 lung adenocarcinoma on CT? Eur Radiol 2019;29:6049-58.
16. Tu W, Sun G, Fan L, et al. Radiomics signature: A potential and incremental predictor for EGFR mutation status in NSCLC patients, comparison with CT morphology. Lung Cancer 2019;132:28-35.
17. Wang X, Li Q, Cai J, et al. Predicting the invasiveness of lung adenocarcinomas appearing as ground-glass nodule on CT scan using multi-task learning and deep radiomics. Transl Lung Cancer Res 2020;9:1397-406.
18. Zhao X, Wang X, Xia W, et al. A cross-modal 3D deep learning for accurate lymph node metastasis prediction in clinical stage T1 lung adenocarcinoma. Lung Cancer 2020;145:10-7.
19. Zhao X, Wang X, Xia W, et al. 3D multi-scale, multi-task, and multi-label deep learning for prediction of lymph node metastasis in T1 lung adenocarcinoma patients' CT images. Comput Med Imaging Graph 2021;93:101987.
20. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44:837-45.
21. Chen D, She Y, Wang T, et al. Radiomics-based prediction for tumour spread through air spaces in stage I lung adenocarcinoma using machine learning. Eur J Cardiothorac Surg 2020;58:51-8.
22. Onozato Y, Nakajima T, Yokota H, et al. Radiomics is feasible for prediction of spread through air spaces in patients with nonsmall cell lung cancer. Sci Rep 2021;11:13526.
23. Bassi M, Russomando A, Vannucci J, et al. Role of radiomics in predicting lung cancer spread through air spaces in a heterogeneous dataset. Transl Lung Cancer Res 2022;11:560-71.
24. Qi L, Xue K, Cai Y, et al. Predictors of CT Morphologic Features to Identify Spread Through Air Spaces Preoperatively in Small-Sized Lung Adenocarcinoma. Front Oncol 2020;10:548430.
25. Gu Y, Zheng B, Zhao T, et al. Computed Tomography Features and Tumor Spread Through Air Spaces in Lung Adenocarcinoma: A Meta-analysis. J Thorac Imaging 2023;38:W19-29.
26. Suh YJ, Han K, Kwon Y, et al. Computed Tomography Radiomics for Preoperative Prediction of Spread Through Air Spaces in the Early Stage of Surgically Resected Lung Adenocarcinomas. Yonsei Med J 2024;65:163-73.
27. Han X, Fan J, Zheng Y, et al. The Value of CT-Based Radiomics for Predicting Spread Through Air Spaces in Stage IA Lung Adenocarcinoma. Front Oncol 2022;12:757389.

28. Zhuo Y, Feng M, Yang S, et al. Radiomics nomograms of tumors and peritumoral regions for the preoperative prediction of spread through air spaces in lung adenocarcinoma. Transl Oncol 2020;13:100820.
29. Tao J, Liang C, Yin K, et al. 3D convolutional neural network model from contrast-enhanced CT to predict spread through air spaces in non-small cell lung cancer. Diagn Interv Imaging 2022;103:535-44.
30. Wang Y, Ding Y, Liu X, et al. Preoperative CT-based radiomics combined with tumour spread through air spaces can accurately predict early recurrence of stage I lung adenocarcinoma: a multicentre retrospective cohort study. Cancer Imaging 2023;23:83.

Cite this article as: Wang X, Ma C, Jiang Q, Zheng X, Xie J, He C, Gu P, Wu Y, Xiao Y, Liu S. Performance of deep learning model and radiomics model for preoperative prediction of spread through air spaces in the surgically resected lung adenocarcinoma: a two-center comparative study. Transl Lung Cancer Res 2024;13(12):3486-3499. doi: 10.21037/tlcr-24-646