Journal of Cardiothoracic Surgery

RESEARCH

Open Access

Deep learning radiomics fusion model to predict visceral pleural invasion of clinical stage IA lung adenocarcinoma: a multicenter study

Jiabi Zhao[1,2†], Tingting Wang[1†], Bin Wang[2†], Bhuva Maheshkumar Satishkumar[3], Lumin Ding[4], Xiwen Sun[3*†] and Caizhong Chen[1*†]

Abstract

Aim To assess the predictive performance, risk stratification capabilities, and auxiliary diagnostic utility of radiomics, deep learning, and fusion models in identifying visceral pleural invasion (VPI) in lung adenocarcinoma.

Materials and methods A total of 449 patients (female:male, 263:186; 59.8 ± 10.5 years) diagnosed with clinical IA stage lung adenocarcinoma (LAC) from two distinct hospitals were enrolled in the study and divided into a training cohort ($n = 289$) and an external test cohort ($n = 160$). The fusion models were constructed from the feature level and the decision level respectively. A comprehensive analysis was conducted to assess the prediction ability and prognostic value of radiomics, deep learning, and fusion models. The diagnostic performance of radiologists of varying seniority with and without the assistance of the optimal model was compared.

Results The late fusion model demonstrated superior diagnostic performance (AUC = 0.812) compared to clinical (AUC = 0.650), radiomics (AUC = 0.710), deep learning (AUC = 0.770), and the early fusion models (AUC = 0.586) in the external test cohort. The multivariate Cox regression analysis showed that the VPI status predicted by the late fusion model were independently associated with patient disease-free survival (DFS) ($p = 0.044$). Furthermore, model assistance significantly improved radiologist performance, particularly for junior radiologists; the AUC increased by 0.133 ($p < 0.001$) reaching levels comparable to the senior radiologist without model assistance (AUC: 0.745 vs. 0.730, $p = 0.790$).

Conclusions The proposed decision-level (late fusion) model significantly reducing the risk of overfitting and demonstrating excellent robustness in multicenter external validation, which can predict VPI status in LAC, aid in prognostic stratification, and assist radiologists in achieving higher diagnostic performance.

Keywords Radiomics, Deep learning, Adenocarcinoma of lung, CT

†Jiabi Zhao, Tingting Wang and Bin Wang contributed equally to this work and share the first authorship.

†Xiwen Sun and Caizhong Chen contributed equally to this study as co-corresponding authors.

*Correspondence:
Xiwen Sun
sunxiwen5256@163.com
Caizhong Chen
chen_caizhong1967@163.com
Full list of author information is available at the end of the article

Introduction

In recent years, with the rise in low-dose CT screening programs, the detection rate of early lung cancer has significantly increased. Visceral pleural invasion (VPI) has emerged as a pivotal predictor of prognosis in non-small cell lung cancer (NSCLC) patients [1, 2], with its ability to forecast lymph node metastasis and postoperative recurrence [3, 4]. Consequently, the 8th edition of the TNM classification system stipulates that upon confirmation of VPI, regardless of the tumor's maximum diameter being less than or equal to 30 mm, the T descriptor is escalated from T1 to T2 [5]. Furthermore, for clinical stage IA patients, lobectomy coupled with lymph node dissection is the preferred treatment modality, compared to segmentectomy alone [6]. Thus, a precise preoperative assessment of VPI is imperative in determining the optimal therapeutic approach.

In the realm of medical imaging, specific CT scan characteristics, such as pleural tags or evidence of retraction, provide invaluable insight for predicting the existence of pathological visceral pleural invasion (pVPI) before surgery [7–10]. However, the application of CT imaging attributes as clinical indicators of T2 status remains a topic of ongoing discussion. While tumors with a diameter of 3 cm or less that exhibit pleural tags or retraction on CT scans are often categorized as clinical T2a in clinical practice, accurately diagnosing VPI preoperatively through imaging modalities alone poses a significant challenge. Notably, a recent study has highlighted the limitations of CT-based VPI diagnosis in terms of both diagnostic accuracy and prognostic significance in clinical stage I lung adenocarcinomas [11]. Additionally, addressing the substantial variability in interpretation among readers is of paramount importance in order to ensure the reliability and accuracy of such imaging-based assessments.

In the evaluation of VPI using preoperative CT images, radiomics features have played a pivotal role [12–14]. These studies have predominantly utilized diverse radiomics characteristics, such as shape, intensity, and texture, and subsequently integrated them into machine learning algorithms to ascertain the diagnosis of VPI. Furthermore, convolutional neural networks (CNNs) with varying network architectures have been extensively utilized in forecasting VPI [15–17]. As deep learning progresses rapidly, the features derived from it have become a crucial adjunct to the traditional handcrafted features in medical imaging. The integration of these extracted features with the conventional radiomics characteristics has the potential to augment the predictive prowess of radiomics even further [18–20]. Nevertheless, despite its significant promise, prior investigations have largely centered around feature-based early fusion methodologies. The influence of diverse fusion methodologies on model performance still remains a largely uncharted territory.

Although early fusion models have shown high predictive performance in some studies, they often suffer from overfitting and feature redundancy during external validation. In this study, the introduction of a decision-level (late fusion) strategy effectively overcomes these issues and significantly improves the model's generalizability. In our study, we aimed to compare the diagnostic efficiency and prognostic significance of radiomics, deep learning, and fusion models in forecasting VPI in IAC. To achieve this, we employed two distinct model fusion approaches: early fusion, which fused features at an initial stage, and late fusion, which integrated decisions at a later stage. Additionally, we assessed the diagnostic efficacy of radiologists of varying experience levels, both independently and with the aid of our models, to evaluate the utility of our models in augmenting diagnostic accuracy.

Materials and methods

The institutional review board of two hospitals approved this retrospective study and informed consent was waived.

Study participants

The current study encompassed patients who had undergone computed tomography (CT) imaging at two hospitals: center 1 from January 1, 2020, to June 30, 2020, and center 2 from January 1, 2015, to December 31, 2016. The inclusion criteria for the study were: (i) patients with pathologically confirmed invasive lung adenocarcinoma; (ii) patients diagnosed as clinical stage IA disease by preoperative examination; and (iii) patients who underwent CT scan within two weeks preceding the surgery. After applying these criteria, a total of consecutive 449 patients were included in the final cohort. Specifically, 289 patients from center 1 were designated as the training set, while 160 patients from center 2 comprised the external test set. The study's design and workflow are depicted in Fig. 1.

Image preprocessing

Patients from two institutions underwent comparable CT scan procedures utilizing distinct systems and parameters, as detailed in Supplementary Table 1. Both the image data and corresponding masks underwent isotropic resampling to a voxel size of $1\times 1\times 1$ mm^3, utilizing nearest neighbor interpolation. Standardization procedures were applied to the images, with a window width of 1200 and a window level of −450. Prior to feature extraction, Hounsfield Units (HUs) were discretized, employing a bin width of 25 HU. Regions of interest (ROIs) were manually delineated on each axial enhanced

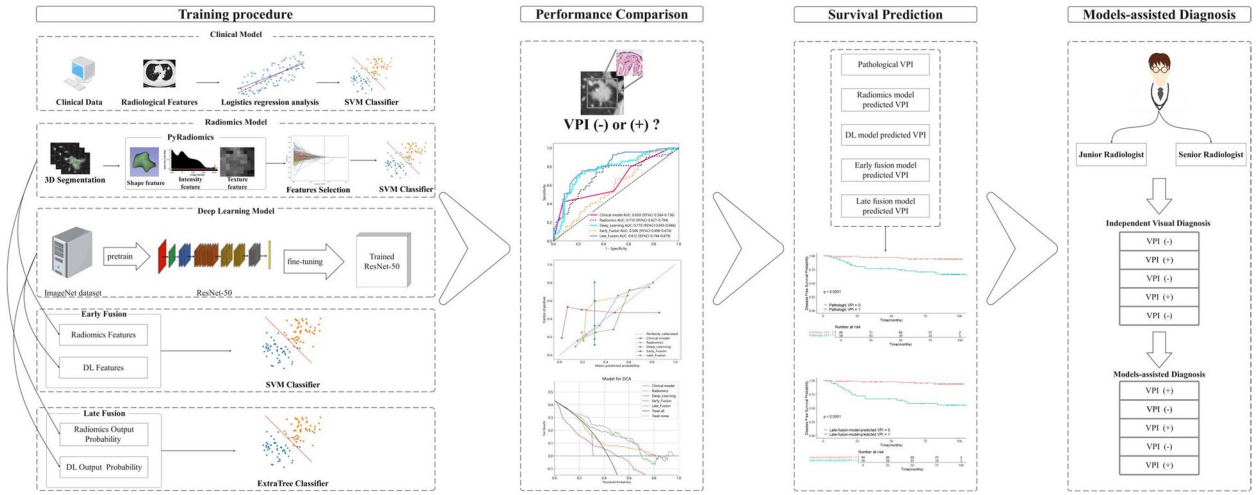

Fig. 1 The study design and pipeline

CT slice by an experienced radiologist, who possesses over three years of expertise and was blinded to tumor pathological information, utilizing 3D Slicer software (version 4.10.2). For more details, please refer to the supplementary materials.

Radiomics model construction

The features underwent standardization via z-score normalization. Initially, we employed the Pearson correlation coefficient to assess the correlation among features. If the correlation coefficient surpassed 0.9 for two specific features, we retained only one of them. Subsequently, we selected features exhibiting high stability, determined by intraclass correlation coefficients (ICCs) values exceeding 0.75 (more details see Supplementary Materials). Finally, we utilized the logistic regression model, coupled with an absolute shrinkage and selection operator, to further streamline the feature set and identify the most pertinent features.

In order to develop predictive models, we trained the SVM classifier using a five-fold cross-validation strategy. Subsequently, we fed the training dataset into the SVM classifier. The trained classifier model then calculated the likelihood of VPI in the tested patient, expressed as a probability ranging from 0 to 1.

Deep learning model development

In this study, we utilized ResNet50 as the CNN framework for extracting deep learning features. The ResNet50 model had been pre-trained on the ILSVRC-2012 dataset. The slice with the largest tumor area was selected to as representative. Subsequently, we normalized the gray values to a range of −1 to 1 using min–max transformation. The cropped sub-region images were resized to 224 × 224 pixels and then served as the input for our model.

We employed the Adam optimizer, set the batch size of 8 and initial learning rate of 0.05, implemented L2 regularization and early stopping strategies to prevent overfitting. The loss rate was used as the metric to evaluate model performance. Once the deep learning model training was completed, we extracted features from the avgpool layer as the deep learning features.

Construction of the fusion models

Two fusion strategies were adopt and compare in our research.

Early fusion, also referred to as feature-level fusion, consolidates features from different modalities into a single feature vector. Firstly, radiomics and DL features underwent z-score normalization. Subsequently, feature selection was conducted through Pearson correlation analysis and LASSO analysis. The SVM classifier was then trained to build the feature-based fusion model, termed "Early Fusion".

Late fusion, referred to as decision-level fusion, involves the consolidation of prediction probabilities stemming from distinct single modality models to arrive at a definitive prediction. We integrated the output probabilities derived from both radiomics and deep learning models. The performance of three machine learning algorithms, namely Random Forest, Support Vector Machine, and Extra Trees, was assessed on the training dataset. Following a comparative analysis of their respective outcomes (data not shown), the Extra Trees classifier emerged as the preferred choice for subsequent

investigation. Ultimately, the optimized model, designated as "Late Fusion" underwent evaluation in external test sets.

Reference standard for VPI

In our study, the assessment of VPI in lung cancer specimens relied on sections stained with hematoxylin–eosin and elastic-van Gieson [21]. Specifically, VPI was defined as infiltration extending beyond the elastic layer (PL1), encompassing invasion that reaches the visceral pleural surface (PL2) [21]. For more details, please refer to the supplementary materials.

Survival analyses

In this study, the primary focus was on disease-free survival (DFS), which we defined as the time elapsed from surgery until the initial imaging- or histologically verified recurrence of local or regional disease, distant metastasis, or death from any cause. Postoperatively, patients typically underwent CT follow-up evaluations every 6 to 12 months for a duration of two years, then followed by annual re-examination. Overall, the 5 years prognosis of 139 patients in the external test cohort were available. Firstly, univariate analysis was employed to identify statistically significant variables. Subsequently, these variables were incorporated into a multivariate Cox regression analysis to further validate the independent predictors of DFS. The Kaplan–Meier survival curve was constructed to visually represent the distribution of survival time among patients in the high- and low-risk groups, and the log-rank test was employed to assess the statistical significance of any differences between the curves.

Radiologists' visual evaluations and model–assisted diagnosis

The assessment of visual VPI was conducted using a 5-point grading system (1 indicating low likelihood of pVPI, 5 indicating high likelihood of pVPI), without knowledge of the histopathological results. The junior radiologist had six years of experience in lung imaging, while the senior radiologist had over thirty years of expertise. The scoring process for visual evaluations was performed without prior training or predefined standards, mirroring standard clinical practice. They evaluated the presence of VPI utilizing axial, coronal, and sagittal views, with the option of magnification whenever deemed necessary.

The late fusion model providing additional predicted probabilities indicating the likelihood of VPI presence (e.g., a 0.928 probability of presence and a 0.072 probability of absence). Additionally, we computed the Kappa values for both radiologists, first without and then with the inclusion of model assistance, to evaluate the clinical advantage.

Statistical analysis

Continuous variables were reported as means ± standard deviations, while categorical variables were expressed as percentages in this study. Statistical comparisons between the training and validation groups involved the use of appropriate tests such as Mann–Whitney U for continuous variables and Fisher's exact test or chi-square test for categorical ones. Agreement among readers was assessed through kappa statistic, while model fitness was evaluated using the Hosmer–Lemeshow goodness-of-fit test. Survival disparities were examined via Kaplan–Meier curves along with log-rank tests to determine statistical significance. The analyses employed Python 3.6.0, R 3.5 .3, or SPSS 26 .0, with a predefined significance level of $p < 0.05$.

Results

Baseline characteristics

Table 1 provides a comprehensive overview of the baseline characteristics of the participating subjects. This retrospective analysis encompassed a total of 449 patients, who were divided into the training and external validation cohorts, consisting of 289 and 160 patients, respectively. The average age among all participants stood at 59.8 years, with a standard deviation of 10.5. Out of these, 41.4% ($n = 186$) were males, while 35.0% ($n = 157$) were diagnosed with pathological VPI.

Supplementary Table 2 offers comprehensive details on clinical model construction. Supplementary Fig. 1 shows the definition of CT subtype of pleura.

Predictive performance of the models

Table 2 presents the performance metrics of each model in both the training set and the external test set. Notably, the late fusion model exhibited remarkable performance on the external test cohort, attaining an AUC 0.812, significantly surpassing the clinical model's AUC of 0.650 ($P < 0.001$), as depicted in Fig. 2.

In the training set, the early fusion model (included features are shown in supplementary Fig. 2) achieved an AUC of 0.988, surpassing all other prediction models, including the late fusion model (AUC = 0.910, $p < 0.001$), the DL model (AUC = 0.799, $p < 0.001$), and the radiomics model (AUC = 0.864, $P < 0.001$). Despite its outstanding performance in the training data, the early fusion model exhibited a notably diminished AUC of 0.586 on the external validation set, indicating a possible concern of overfitting. In comparison, the decision-based fusion model exhibited superiority in the external validation set, achieving an AUC of 0.812, significantly

Table 1 Clinical, pathological and radiological characteristics

Variable	Training cohot ($n=289$)		*P* value	External test cohort ($n=160$)		*P* value
	Negative	Positive		Negative	Positive	
Clinical characteristics						
Age[a]	59.88 ± 10.72	60.98 ± 10.20	0.380	58.32 ± 10.95	60.04 ± 9.60	0.292
Whole maximum diameter (mm)	18.91 ± 6.43	20.86 ± 6.78	**0.024**	18.67 ± 7.96	22.48 ± 6.60	**< 0.001**
Gender			0.830			0.496
Female	114(56.72)	48(54.55)		60(65.93)	41(59.42)	
Male	87(43.28)	40(45.45)		31(34.07)	28(40.58)	
Smoking history			0.788			0.394
No	162(80.60)	69(78.41)		84(92.31)	60(86.96)	
Yes	39(19.40)	19(21.59)		7(7.69)	9(13.04)	
Clinical T stage			**0.016**			**0.005**
cT1a	20(9.95)	4(4.55)		8(8.79)	1(1.45)	
cT1b	104(51.74)	35(39.77)		49(53.85)	26(37.68)	
cT1c	77(38.31)	49(55.68)		34(37.36)	42(60.87)	
Pathological characteristics						
Histologic subtype			**< 0.001**			0.430
Lepidic predominant	24(11.94)	1(1.14)		16(17.58)	9(13.04)	
Acinar predominant	145(72.14)	58(65.91)		57(62.64)	38(55.07)	
Papillary predominant	22(10.95)	11(12.50)		12(13.19)	12(17.39)	
Solid predominant	7(3.48)	13(14.77)		4(4.40)	7(10.14)	
Micropapillary predominant	3(1.49)	5(5.68)		2(2.20)	3(4.35)	
Radiological characteristics						
Consolidation maximum diameter (mm)	15.13 ± 7.11	19.54 ± 7.08	**< 0.001**	14.90 ± 7.67	20.76 ± 7.61	**< 0.001**
CTR (mm)[a]	0.81 ± 0.25	0.94 ± 0.14	**< 0.001**	0.82 ± 0.25	0.92 ± 0.18	**0.002**
Maximum diameter of contact with pleura (mm)[a]	6.04 ± 6.94	7.26 ± 6.90	0.094	2.75 ± 5.57	7.58 ± 7.17	**< 0.001**
Maximum diameter of consolidation contact with the pleura (mm)[a]	5.07 ± 6.58	7.20 ± 6.94	**0.005**	2.16 ± 4.82	7.10 ± 6.84	**< 0.001**
Minimum distance from the lesion to the pleura (mm)[a]	2.04 ± 3.52	1.80 ± 3.61	0.131	6.99 ± 7.93	1.32 ± 2.41	**< 0.001**
Location			0.778			0.461
LLL	27(13.43)	13(14.77)		13(14.29)	9(13.04)	
LUL	53(26.37)	27(30.68)		22(24.18)	16(23.19)	
RLL	50(24.88)	17(19.32)		18(19.78)	12(17.39)	
RML	12(5.97)	7(7.95)		7(7.69)	12(17.39)	
RUL	59(29.35)	24(27.27)		31(34.07)	20(28.99)	
Density			**< 0.001**			**< 0.001**
Subsolid	97(48.26)	20(22.73)		75(82.42)	38(55.07)	
Pure-solid	104(51.74)	68(77.27)		16(17.58)	31(44.93)	
CT subtype of pleura			0.320			**< 0.001**
Type 1	42(20.90)	26(29.55)		11(12.09)	23(33.33)	
Type 2	54(26.87)	25(28.41)		10(10.99)	17(24.64)	
Type 3	66(32.84)	21(23.86)		12(13.19)	18(26.09)	
Type 4	21(10.45)	11(12.50)		4(4.40)	9(13.04)	
Type 5	18(8.96)	5(5.68)		12(13.19)	1(1.45)	
Type 6	0(0)	0(0)		42(46.15)	1(1.45)	
Consolidation contact with the pleura			**0.010**			**< 0.001**
No	96(47.76)	27(30.68)		71(78.02)	21(30.43)	
Yes	105(52.24)	61(69.32)		20(21.98)	48(69.57)	

Unless otherwise noted, values are numbers of patients, with percentages in parentheses

CTR consolidation tumor ratio, *LLL* left lower lobe, *LUL* left upper lobe, *RLL* right lower lobe, *RML* right middle lobe, *RUL* right upper lobe, *VPI* visceral pleural invasion

[a] Data are means ± standard deviations

Table 2 Prediction performance of the models

Cohort	Model	AUC	95% CI	Accuracy	Sensitivity	Specificity	PPV	NPV
Training	Clinical model	0.55	0.472—0.617	0.696	0.000	1.000	0.000	0.696
	Radiomics	0.86	0.817—0.910	0.810	0.852	0.791	0.641	0.924
	Deep Learning model	0.80	0.743—0.854	0.789	0.511	0.910	0.714	0.810
	Early fusion model	0.99	0.973—1.000	0.955	0.966	0.950	0.895	0.966
	Late fusion model	0.91	0.877—0.943	0.785	0.854	0.751	0.603	0.926
External test	Clinical model	0.65	0.564—0.737	0.563	0.000	1.000	0.000	0.563
	Radiomics	0.71	0.627—0.794	0.681	0.604	0.604	0.600	0.785
	Deep Learning model	0.77	0.693—0.846	0.756	0.710	0.791	0.721	0.783
	Early fusion model	0.59	0.498—0.674	0.562	0.899	0.308	0.496	0.800
	Late fusion model	0.81	0.744—0.879	0.756	0.725	0.780	0.714	0.789

AUC area under the curve, *CI* confidence interval

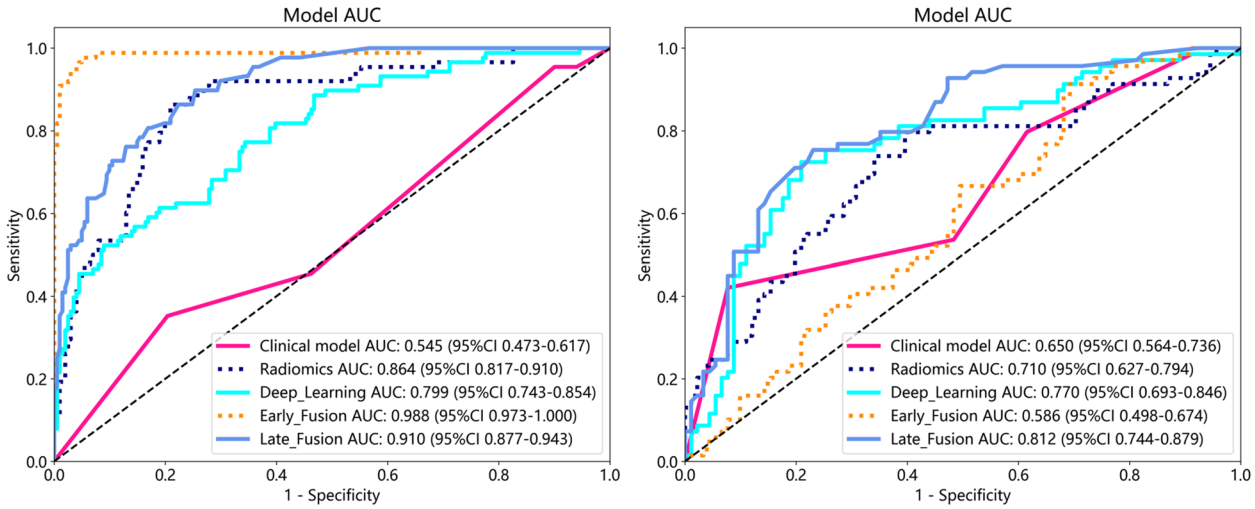

Fig. 2 The receiver operating characteristic curves (ROCs) of five distinct model in the trianing (left) and external test cohort (right)

outperforming the feature-based fusion model's AUC of 0.586 ($p < 0.001$).

As shown in Fig. 3, the late fusion model closely approximates the ideal model (represented by the diagonal dashed line) across a broader range of horizontal coordinates in the external validation set. Additionally, the DCA curves illustrated the superior net benefit offered by the late fusion model, as depicted in the same figure. The Hosmer–Lemeshow test revealed that only the DL model and the late fusion model exhibited good fitness in two study sets ($P > 0.005$). The DeLong's test results are presented in Supplementary Fig. 3.

The optimal cutoff values were determined by the maximum Youden index based on the training set, resulting in values of 0.305 for the clinical model, 0.228 for the radiomics model, 0.457 for the DL model, 0.256 for the early fusion model, and 0.259 for the late fusion model.

Survival prediction

One hundred thirty-nine patients with a minimum of 5 years follow-up from the external test cohort was incorporated into survival analysis. The DFS for these patients stood at 90.0 months, with 22 out of 160 (13.8%) occurred end-point event. The findings of the univariate Cox regression analyses, detailed in Supplementary Table 3, highlighted significant correlations between DFS and various factors, including smoking history, maximal whole tumor diameter, maximal consolidation diameter, density, pVPI, and VPI status predicted by radiomics, deep learning, and late fusion model (all with *p*-values less than 0.05).

After multivariate Cox regression analyses, smoking history, density, pVPI, and VPI status predicted by the late fusion model remained as significant predictors (as presented in Table 3). The survival curves of these patients are illustrated in Fig. 4. Specifically,

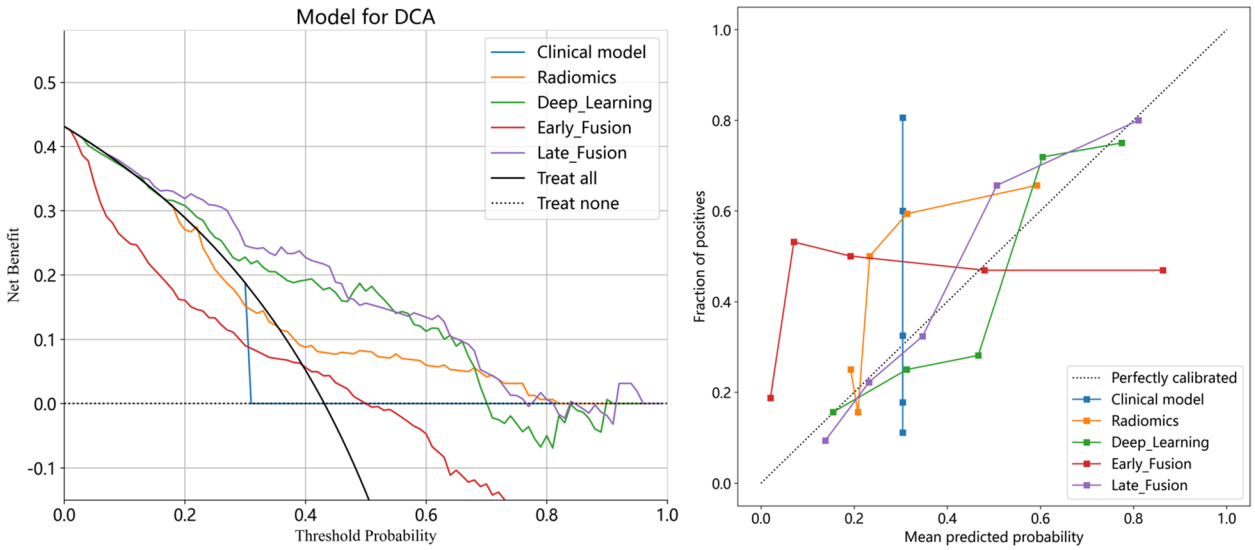

Fig. 3 A comparative analysis of decision curves (depicted on the left) and calibration curves (depicted on the right) across various models is presented for the external test cohort. The X-axis denotes the threshold probability, whereas the Y-axis represents the net benefit achieved. The thin grey line represents the hypothetical scenario where all LACs patients are classified as high-risk, and the thin black line corresponds to the scenario where all patients are designated as low-risk. Calibration curves depict the calibration of each model in terms of the agreement between the predicted status of visceral pleura and observed outcomes of visceral pleura. The y-axis represents the actual visceral pleural invasion rate. The x-axis represents the predicted visceral pleural invasion. The diagonal dotted line represents a perfect prediction by an ideal model

Table 3 Multivariate Cox regression analysis for disease-free survival in in clinical stage IA lung adenocarcinomas

Variable	Pathologic VPI		Radiomics model		Deep learning model		Late fusion model	
	HR (95%CI)	*P* value	HR (95%CI)	*P* value	HR (95%CI)	*P* value	HR (95%CI)	*P* value
Smoking history(ever smoker)	-	0.061	3.068(1.230–7.656)	**0.016**	3.068(1.230–7.656)	**0.016**	3.043(1.213–7.636)	**0.018**
Whole maximum diameter (< 21.15 mm)	-	0.151	-	0.139	-	0.139	-	0.220
Consolidation maximum diameter (< 18.95 mm)	-	0.113	-	0.064	-	0.064	-	0.097
Density(sub-solid)	11.686(3.882–35.179)	**< 0.001**	14.046(4.702–41.960)	**< 0.001**	14.046(4.702–41.960)	**< 0.001**	10.162(3.181–32.464)	**< 0.001**
VPI	4.461(1.485–13.395)	**0.008**	-	0.127		0.434	2.542(1.025–6.303)	**0.044**

CI confidence interval, *HR* hazard ratio, *VPI* visceral pleural invasion

patients with pathological VPI exhibited a median DFS of 88.0 months, while those without pathological VPI had a median DFS of 92.0 months, indicating a notable disparity ($p < 0.001$). Correspondingly, the predicted 5-year survival rates for these two groups were 72.6% and 94.5%, respectively. Furthermore, patients were stratified into two distinct categories using the late fusion model. Among these patients, those predicted to have VPI present using the late fusion model showed a median DFS of 87.0 months, while those predicted to be VPI-absent had a median DFS of 92.0 months ($p < 0.001$). The corresponding 5-year survival rates were 71.1% and 97.1%, respectively.

Diagnostic performance of the radiologist with or without model assistance

The diagnostic capabilities of different seniority radiologists in identifying the presence or absence of VPI, both independently and with the aid of the late fusion model, are comprehensively outlined in Table 4. Notably, the introduction of the model resulted in a substantial augmentation in the overall diagnostic proficiency of junior

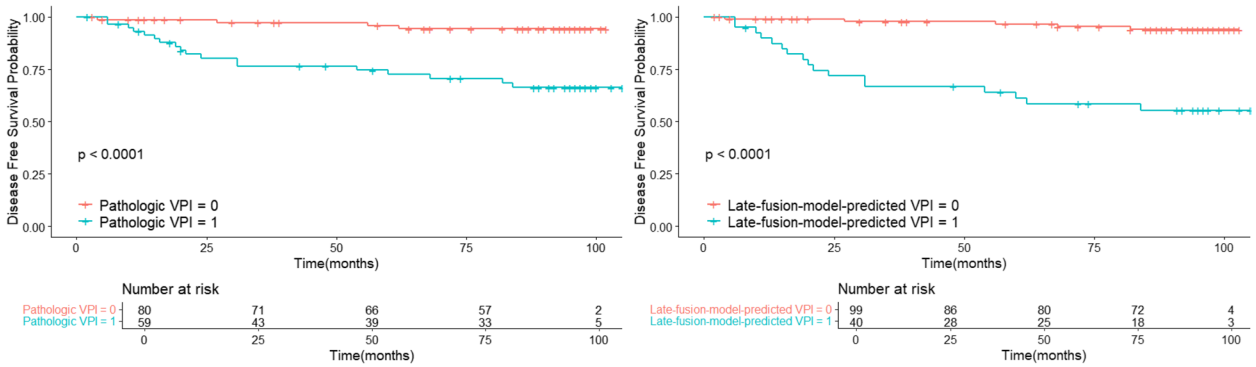

Fig. 4 Kaplan–Meier survival curves evaluating the outcomes of clinical stage IA lung adenocarcinoma grouped according to their pathological VPI status (left), or their VPI status as predicted by the late fusion model (right). DFS, disease-free survival; VPI, visceral pleural invasion

Table 4 Performance comparison between the late fusion model and radiologists and between radiologists with and without model assistance

	AUC	95% CI	Accuracy	Sensitivity	Specificity	PPV	NPV
Late fusion model	0.812	0.7443—0.8787	0.756	0.725	0.780	0.714	0.789
Junior radiologist	0.612	0.5265—0.6980	0.569	0.304	0.769	0.500	0.593
Senior radiologist	0.730	0.6510—0.8096	0.675	0.406	0.879	0.718	0.661
Late fusion model-assisted junior radiologist	0.745	0.6704—0.8204	0.669	0.362	0.901	0.735	0.651
Late fusion model-assisted senior radiologist	0.885	0.8350—0.9346	0.794	0.638	0.912	0.846	0.769

AUC area under the curve, *PPV* positive predictive value, *NPV* negative predictive value

radiologists. Specifically, in the external test set, the AUC augmented by 0.133 ($p < 0.001$), accompanied by increments in sensitivity by 0.100 and specificity by 0.132. These enhanced metrics approached the levels attained by senior radiologists without the model's assistance, as evidenced by the AUC values of 0.745 versus 0.730 ($p = 0.790$). Moreover, senior radiologists' diagnostic acumen also underwent a marked improvement with the model's assistance ($p < 0.001$), achieving AUC, accuracy, sensitivity, and specificity values of 0.885, 0.794, 0.638, and 0.912, respectively. Notably, the developed late fusion model displayed diagnostic efficacy akin to those of senior radiologists, significantly outperforming junior radiologists, as depicted in Fig. 5. Furthermore, the kappa value for VPI prediction among junior radiologists saw a substantial surge from 0.224 to 0.386, whereas for senior radiologists, it increased from 0.420 to 0.636, attributable to the assistance of the late fusion model.

Discussion

In the clinical setting, accurately identifying the presence of VPI in NSCLC patients, without invasive procedures, remains a critical priority. The current study endeavors to predict the VPI status among lung adenocarcinoma patients, leveraging a multimodal approach that integrates clinical data and radiological features. The evaluation of diverse predictive models underscores the significance of this methodology. Through development and evaluation of clinical, radiomics, and deep learning models, as well as feature- and decision-level fusion techniques, our findings highlight the performance of the decision-level fusion model, achieving an AUC score of 0.812 on an external test cohort. Notably, the VPI status predicted by our late fusion model exhibits a robust and independent association with patient DFS in survival analysis. Furthermore, the assistance rendered by this late fusion model not only enhances the diagnostic prowess of radiologists but particularly benefits junior practitioners, highlighting the potential of such integrated models to elevate diagnostic accuracy in clinical practice.

In this study, we observed that the early fusion model performed exceptionally well in the training set, with an AUC of 0.988. However, its AUC dropped to 0.586 in the external validation set, indicating a significant overfitting issue. Overfitting typically occurs when a model learns too many irrelevant details or noise from the training data, leading to poor generalization to unseen data, such as the external validation set. We believe that the excellent performance of the early fusion model in the training set may be partly due to its complex model structure, formed by directly fusing deep learning and conventional radiomics features at the feature level, which may have

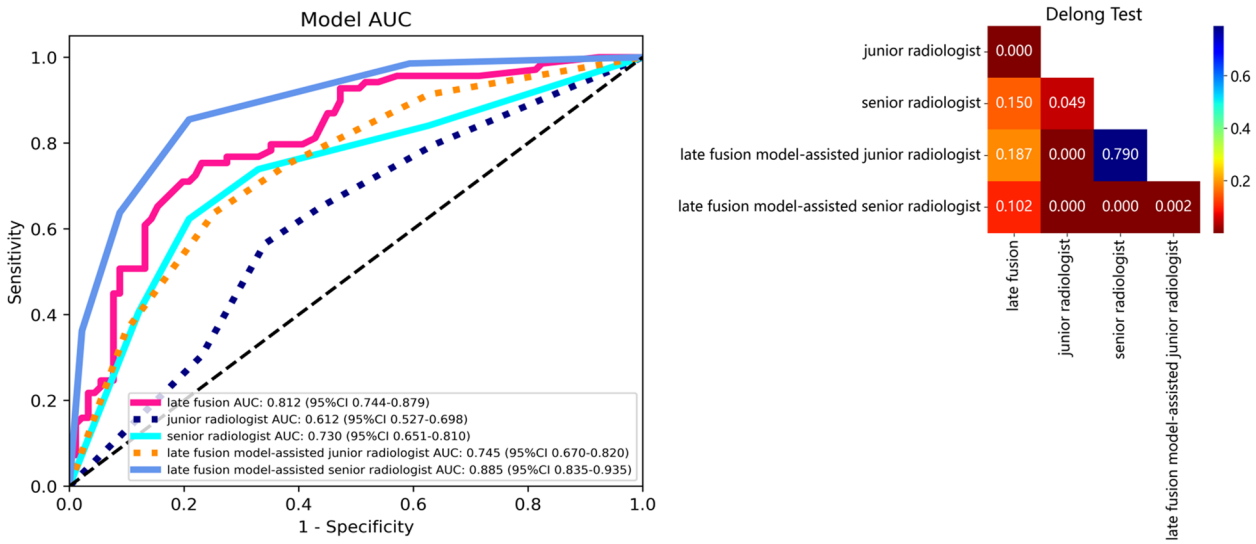

Fig. 5 Performance comparison between radiologists with or without the late fusion model assistance in assessing the VPI status for the external test cohort

led to overfitting on the training data. The efficacy of radiomics, deep learning (DL), and multi-domain fusion models were evaluated in a systematic research [22]. The analysis indicated that the fusion model surpassed 63% of the reviewed studies, fell behind in 25%, and matched other models' performance in 12%. Notably, the fusion model's performance appears to hinge on the fusion strategy adopted. Although the early fusion approach was preferred in most biomedical investigations, this methodology does not consistently produce the most optimal outcomes. Huang and colleagues conducted a comparative analysis of seven distinct fusion frameworks [23]. They concluded that the late fusion approach achieved superior performance, which concurs with our study. We assume that the interplay between deep learning and radiomics features could introduce instability in the estimation process, potentially degrading the efficacy of the feature-level fusion model. Consequently, we advise against relying solely on past experience in choosing fusion approach for a specific medical application. Instead, we recommend conducting preliminary experiments with multiple fusion strategies and selecting the most suitable one based on the findings.

In patients with a diagnosis of NSCLC, VPI has emerged as a significant adverse predictor of prognosis [6, 24, 25]. Notably, Kim et al. noted that semantic features derived from CT related to VPI were not independent predictors of DFS [10]. The morphological characteristics observed on CT scans are heavily contingent on the interpretive skills of radiologists. In our study, VPI served as a classification criterion for patient stratification, and we evaluated the predictive capability of a late fusion model in estimating DFS. Through the application of Kaplan–Meier analyses, our late fusion model successfully differentiated patients with clinical IA stage lung adenocarcinoma into distinct low- and high-risk groups. Additionally, multivariate analyses revealed that the model's predictions based on VPI status were independently associated with DFS.

Utilizing the late fusion model, the clinical diagnostic capabilities of the primary radiologist underwent a substantial improvement. Within the external validation dataset, notable gains were observed in key performance metrics, including a 0.133 augmentation in the area under the curve (AUC), a 0.100 rise in sensitivity, and a 0.132 increment in specificity. These advancements surpassed the baseline performance of senior radiologists without the model's assistance, as evidenced by the AUC comparison (0.745 vs. 0.730), with a p-value of 0.790. Additionally, the integration of this deep learning model revealed a positive trend in diagnostic alignment between junior and senior radiologists, indicating its potential as a valuable adjunct in enhancing radiological interpretations and minimizing errors stemming from inexperience.

The present study is limited by several factors. First and foremost, the retrospective nature of the study potentially gave rise to a selection bias, despite employing a large multicenter cohort for model development. However, to ensure the prediction model's reliability, a larger prospective study encompassing a broader patient population is imperative. Secondly, various CT scanners were utilized in the evaluation of patients, leading to inconsistent acquisition protocols across reports. Despite these limitations, we argue that the diversity of the data

is exactly reflective of the clinical practice encountered in real world. Consequently, any models trained on these data would be more appropriate for clinical implementation in real-world scenarios. Finally, for extracting deep learning features, we utilized the 2D slice displaying the largest tumor area, rather than 3D volume of entire lesion in our research. In fact, we conducted comparisons with 3D models during our study. However, the 3D model did not perform as well as the 2D fusion model (see Supplementary Fig. 4), likely due to the smaller dataset size and increased complexity of 3D feature extraction. Since the primary objective of our study was to compare the performance of different fusion models, we chose to focus on the 2D approach in our analysis.

Conclusion

In conclusion, the decision-level fusion model proposed in this study holds promise for predicting VPI presence and DFS in LAC. It can effectively aid radiologists in improving clinical outcomes systematically, thereby bridging the experience gap among different radiologists.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13019-025-03488-6.

Supplementary Material 1.

Authors' contributions

JZ, TW, and BW contributed equally to the study conception, design, data acquisition, analysis, and manuscript drafting. BMS and LD were responsible for data processing, interpretation, and figure preparation. XS and CC provided overall supervision, critically revised the manuscript, and approved the final version. All authors reviewed and approved the final manuscript.

Funding

This study has received funding by the Science and Technology Commission of Shanghai Municipality [grant Number: 21Y11910400] and National Natural Science Foundation of China [grant Number: 82302289].

Data availability

No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate

This retrospective study was approved by the Ethics Committee of Zhongshan Hospital and Shanghai Pulmonary Hospital, and the requirement to obtain informed consent from patients was waived.

Consent for publication

All authors have read the manuscript and have agreed to its publication in Journal of Cardiothoracic Surgery.

Competing interests

The authors declare no competing interests.

Author details

[1]Department of Radiology, Zhongshan Hospital, Fudan University, No.180 Fenglin Road, Xuhui District, Shanghai 200032, China. [2]Department of Radiology, Shuguang Hospital Affiliated to Shanghai University of Traditional Chinese Medicine, Shanghai 200021, China. [3]Department of Radiology, Shanghai Pulmonary Hospital, Tongji University School of Medicine, No.507 Zhengmin Road, Yangpu District, Shanghai 200433, China. [4]Department of Radiology, Shanghai Pudong New Area Mental Health Center, School of Medicine, Tongji University, Shanghai 200124, China.

Received: 6 August 2024 Accepted: 18 May 2025
Published online: 28 May 2025

References

1. Jiang L, Liang W, Shen J, Chen X, Shi X, He J, et al. The impact of visceral pleural invasion in node-negative non-small cell lung cancer: a systematic review and meta-analysis. Chest. 2015;148:903–11.
2. Yang X, Sun F, Chen L, Shi M, Shi Y, Lin Z, et al. Prognostic value of visceral pleural invasion in non-small cell lung cancer: a propensity score matching study based on the SEER registry. J Surg Oncol. 2017;116:398–406.
3. Inoue M, Minami M, Shiono H, Sawabata N, Ideguchi K, Okumura M. Clinicopathologic study of resected, peripheral, small-sized, non-small cell lung cancer tumors of 2 cm or less in diameter: pleural invasion and increase of serum carcinoembryonic antigen level as predictors of nodal involvement. J Thorac Cardiovasc Surg. 2006;131:988–93.
4. Gorai A, Sakao Y, Kuroda H, Uehara H, Mun M, Ishikawa Y, et al. The clinicopathological features associated with skip N2 metastases in patients with clinical stage IA non-small-cell lung cancer. Eur J Cardiothorac Surg. 2015;47:653–8.
5. Rami-Porta R, Bolejack V, Crowley J, Ball D, Kim J, Lyons G, et al. The IASLC lung cancer staging project: proposals for the revisions of the T descriptors in the forthcoming eighth edition of the TNM classification for lung cancer. J Thorac Oncol. 2015;10:990–1003.
6. Yu Y, Huang R, Wang P, Wang S, Ling X, Zhang P, et al. Sublobectomy versus lobectomy for long-term survival outcomes of early-stage non-small cell lung cancer with a tumor size ≤2 cm accompanied by visceral pleural invasion: a SEER population-based study. J Thorac Dis. 2020;12:592–604.
7. Yang Y, Xie Z, Hu H, Yang G, Zhu X, Yang D, et al. Using CT imaging features to predict visceral pleural invasion of non-small-cell lung cancer. Clin Radiol. 2023;78:e909–17.
8. Onoda H, Higashi M, Murakami T, et al. Correlation between pleural tags on CT and visceral pleural invasion of peripheral lung cancer that does not appear touching the pleural surface. Eur Radiol. 2021;31(12):9022–9. https://doi.org/10.1007/s00330-021-07869-y.
9. Sun Q, Li P, Zhang J, Yip R, Zhu Y, Yankelevitz DF, et al. CT predictors of visceral pleural invasion in patients with non-small cell lung cancers 30 mm or smaller. Radiology. 2024;310:e231611.
10. Kim H, Goo JM, Kim YT, Park CM. CT-defined visceral pleural invasion in T1 lung adenocarcinoma: lack of relationship to disease-free survival. Radiology. 2019;292:741–9.
11. Lim WH, Lee KH, Lee JH, et al. Diagnostic performance and prognostic value of CT-defined visceral pleural invasion in early-stage lung adenocarcinomas. Eur Radiol. 2024;34(3):1934–45. https://doi.org/10.1007/s00330-023-10204-2.
12. Zha X, Liu Y, Ping X, Bao J, Wu Q, Hu S, et al. A nomogram combined radiomics and clinical features as imaging biomarkers for prediction of visceral pleural invasion in lung adenocarcinoma. Front Oncol. 2022;12:876264.
13. Wei SH, Zhang JM, Shi B, Gao F, Zhang ZX, Qian LT. The value of CT radiomics features to predict visceral pleural invasion in ≤3 cm peripheral type early non-small cell lung cancer. J Xray Sci Technol. 2022;30:1115–26.
14. Yuan M, Liu JY, Zhang T, Zhang YD, Li H, Yu TF. Prognostic impact of the findings on thin-section computed tomography in stage I lung adenocarcinoma with visceral pleural invasion. Sci Rep. 2018;8:4743.
15. Lin X, Liu K, Li K, Chen X, Chen B, Li S, et al. A CT-based deep learning model: visceral pleural invasion and survival prediction in clinical stage IA lung adenocarcinoma. IScience. 2024;27:108712.
16. Shimada Y, Ojima T, Takaoka Y, et al. Prediction of visceral pleural invasion of clinical stage I lung adenocarcinoma using thoracoscopic images and deep learning. Surg Today. 2024;54(6):540–50. https://doi.org/10.1007/s00595-023-02756-z.
17. Choi H, Kim H, Hong W, Park J, Hwang EJ, Park CM, et al. Prediction of visceral pleural invasion in lung cancer on CT: deep learning model

achieves a radiologist-level performance with adaptive sensitivity and specificity to clinical needs. Eur Radiol. 2021;31:2866–76.
18. Zhong H, Wang T, Hou M, Liu X, Tian Y, Cao S, et al. Deep learning radiomics nomogram based on enhanced CT to predict the response of metastatic lymph nodes to neoadjuvant chemotherapy in locally advanced gastric cancer. Ann Surg Oncol. 2024;31:421–32.
19. Zhang M, Lu Y, Sun H, Hou C, Zhou Z, Liu X, et al. CT-based deep learning radiomics and hematological biomarkers in the assessment of pathological complete response to neoadjuvant chemoradiotherapy in patients with esophageal squamous cell carcinoma: a two-center study. Transl Oncol. 2024;39:101804.
20. Zhang J, Yin W, Yang L, Yao X. Deep Learning Radiomics Nomogram Based on Multiphase Computed Tomography for Predicting Axillary Lymph Node Metastasis in Breast Cancer. Mol Imaging Biol. 2024;26(1):90–100. https://doi.org/10.1007/s11307-023-01839-0.
21 Travis WD, Brambilla E, Rami-Porta R, Vallières E, Tsuboi M, Rusch V, et al. Visceral pleural invasion: pathologic criteria and use of elastic stains: proposal for the 7th edition of the TNM classification for lung cancer. J Thorac Oncol. 2008;3:1384–90.
22 Demircioğlu A. Are deep models in radiomics performing better than generic models? A systematic review. Eur Radiol Exp. 2023;7:11.
23. Huang S-C, Pareek A, Zamanian R, Banerjee I, Lungren MP. Multimodal fusion with deep neural networks for leveraging CT imaging and electronic health record: a case-study in pulmonary embolism detection. Sci Rep. 2020;10:22147.
24. Wang F, Li P, Li F. Nomogram for predicting the relationship between the extent of visceral pleural invasion and survival in non-small-cell lung cancer. Can Respir J. 2021;2021:8816860.
25. Yip R, Ma T, Flores RM, Yankelevitz D, Henschke CI. Survival with parenchymal and pleural invasion of non-small cell lung cancers less than 30 mm. J Thorac Oncol. 2019;14:890–902.

Publisher's Note