EJNMMI RESEARCH

ORIGINAL RESEARCH Open Access

A strategy for the automatic diagnostic pipeline towards feature-based models: a primer with pleural invasion prediction from preoperative PET/CT images

Xiangxing Kong[1,2†], Annan Zhang[3†], Xin Zhou[2], Meixin Zhao[3], Jiayue Liu[2], Xinliang Zhang[1], Weifang Zhang[3], Xiangxi Meng[2*], Nan Li[2*] and Zhi Yang[1,4*]

Abstract

Background This study aims to explore the feasibility to automate the application process of nomograms in clinical medicine, demonstrated through the task of preoperative pleural invasion prediction in non-small cell lung cancer patients using PET/CT imaging.

Results The automatic pipeline involves multimodal segmentation, feature extraction, and model prediction. It is validated on a cohort of 1116 patients from two medical centers. The performance of the feature-based diagnostic model outperformed both the radiomics model and individual machine learning models. The segmentation models for CT and PET images achieved mean dice similarity coefficients of 0.85 and 0.89, respectively, and the segmented lung contours showed high consistency with the actual contours. The automatic diagnostic system achieved an accuracy of 0.87 in the internal test set and 0.82 in the external test set, demonstrating comparable overall diagnostic performance to the human-based diagnostic model. In comparative analysis, the automatic diagnostic system showed superior performance relative to other segmentation and diagnostic pipelines.

Conclusions The proposed automatic diagnostic system provides an interpretable, automated solution for predicting pleural invasion in non-small cell lung cancer.

Keywords Pleural invasion, Non-small cell lung cancer, Radiomics, Deep learning, Predictive diagnostic modeling

[†] Xiangxing Kong and Annan Zhang have contributed equally to this work.

*Correspondence:
Xiangxi Meng
mengxiangxi@pku.edu.cn
Nan Li
Rainbow6283@sina.com
Zhi Yang
pekyz@163.com

[1] Institution of Medical Technology, Peking University Health Science Center, Beijing 100191, China

[2] Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education), Beijing Key Laboratory of Research, Investigation and Evaluation of Radiopharmaceuticals, NMPA Key Laboratory for Research and Evaluation of Radiopharmaceuticals (National Medical Products Administration), Department of Nuclear Medicine, Peking University Cancer Hospital and Institute, No. 52 Fucheng Rd., Haidian District, Beijing 100142, China

[3] Department of Nuclear Medicine, Peking University Third Hospital, Beijing 100191, China

[4] State Key Laboratory of Holistic Integrative Management of Gastrointestinal Cancers, Beijing Key Laboratory of Research, Investigation and Evaluation of Radiopharmaceuticals, NMPA Key Laboratory for Research and Evaluation of Radiopharmaceuticals (National Medical Products Administration), Department of Nuclear Medicine, Peking University Cancer Hospital and Institute, No. 52 Fucheng Rd., Haidian District, Beijing 100142, China

Background

Beyond subjective impressions and intuitive judgments, evidence-based decision-making in clinical practice often involves sophisticated mathematical models. Nomograms achieved a balance between efficiency and simplicity by utilizing graphical representations to elucidate complex interactions among various clinical and physiological variables. In medical imaging, a radiomics approach is often employed to extract quantitative image features for nomogram construction. This approach gains so much research interest that in the year 2024 alone, around 700 research papers exploring radiomics-based nomograms were indexed in the Web of Science database.

Although numerous nomograms have been developed, their integration into clinical workflows remains rare. One key limitation is the manual application of graphical analyses, which is labor-intensive and error-prone. This challenge is further exacerbated when radiomics features are involved, as these variables must first be extracted and processed using specialized computer systems, adding an additional layer of complexity to their practical use.

This challenge in the application of nomograms was virtually unsolvable until recent years, when advancements in artificial intelligence (AI), particularly deep learning algorithms, emerged as state-of-the-art solutions for various tasks in medical imaging [1, 2]. AI-driven approaches, such as convolutional neural networks (CNNs), offer automated diagnostic capabilities that bypass some of the limitations of traditional methods by learning representations directly from the imaging data [3–5]. However, these end-to-end deep learning-based diagnostics methods are not the ultimate solution, and faces multiple limitations [6]. The "black box" nature of these models often limits their interpretability, making it challenging for clinicians to understand and validate the underlying decision-making process [7, 8]. Additionally, most current AI methodologies function as standalone systems, seldom integrating established diagnostic tools such as radiomic features or clinically relevant parameters—elements that could enhance their clinical applicability, reliability, and acceptance.

In this research, we propose the integration of AI methodologies to address the longstanding challenges in the clinical application of nomograms. We demonstrate our approach through the development of a nomogram-based automated diagnostic pipeline designed to predict pleural invasion (PI) from preoperative PET/CT images. PI in lung cancer serves as a critical prognostic factor, significantly influencing clinical decision-making and treatment planning for patients with non-small cell lung cancer (NSCLC) [9, 10]. Accurate preoperative identification of PI is essential for determining surgical approaches and predicting disease progression [11]. Traditionally, predictive models such as nomograms have been developed to address this type of diagnostic need by incorporating clinical, radiological, and pathological data [12, 13]. These models offer individualized risk predictions by integrating parameters such as tumor size, pleural thickening, and the maximum standardized uptake value (SUVmax) from imaging. [14, 15] Despite their demonstrated utility, traditional methods rely on manual measurement and feature extraction, processes that are often time-intensive and prone to interobserver variability—challenges that are particularly pronounced in high-volume clinical settings where efficiency is a critical requirement [16, 17].

This study presents a hybrid diagnostic framework consisting of two key components: a feature-based diagnostic model (FDM), which utilizes radiomic features to generate interpretable clinical insights, and an AI-enabled diagnostic pipeline (AIDP), which integrates these insights into a fully automated diagnostic workflow. Building on this foundation, we developed an automatic diagnostic system (ADS) capable of predicting PI with high accuracy while simultaneously improving the interpretability and clinical usability of the results. This framework bridges traditional diagnostic practices and modern AI capabilities by incorporating interpretable radiomic features in alignment with clinical diagnostic habits, while also streamlining diagnostic processes through AI-driven automation. By combining these complementary strengths, our approach seeks to advance preoperative PI prediction in NSCLC, offering a solution that is not only accurate and efficient but also scalable and adaptable to clinical workflows.

Methods

Workflow overview

The computer-aided diagnostic strategies employing radiomics and clinical features, such as the FDM developed, traditionally find their clinical applications in the form of nomograms. However, this is a laborious multi-step procedure involving intensive human interference. As shown in Fig. 1, the current research explores an AIDP to substitute such interference. This AIDP contains (1) an automatic segmentation module to generate masks of the lesions and organs, (2) an automatic feature extraction module to generate various parameters based on the images, and (3) an automatic model evaluation module to calculate the model output. To demonstrate the effectiveness of such a strategy, we developed an ADS by combining an FDM for predicting PI with preoperative PET/CT images, and the AIDP for lung nodule.

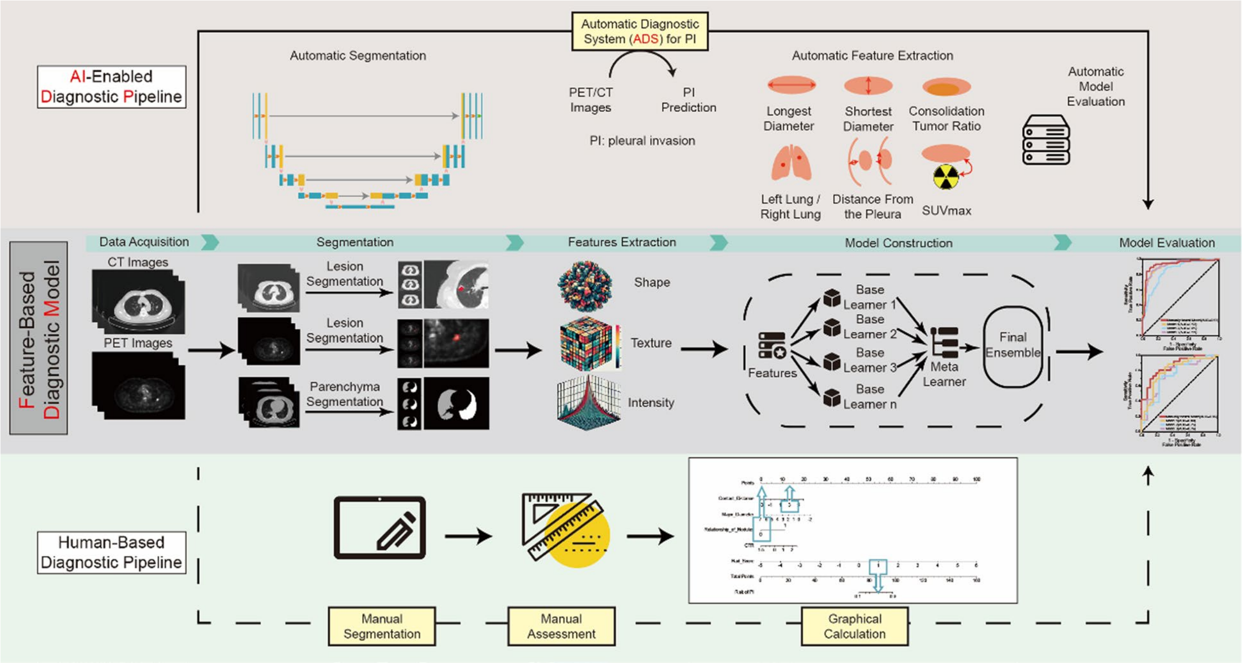

Fig. 1 The workflow of the proposed strategy

Clinical data acquisition

We retrospectively included 1116 NSCLC patients who underwent [^{18}F]FDG PET/CT scans for the construction of the Feature-Based Diagnostic model, with 1017 cases from Center 1 (Peking University Cancer Hospital) between January 2018 and January 2023; and 99 cases from Center 2 (Peking University Third Hospital) between August 2021 and August 2024. In the dataset from Center 1, 503 cases had manually delineated regions of interest (ROI), which were used for constructing the feature-based model and the AIDP. These cases were randomly divided into a training set of 403 cases and an internal test set of 100 cases. An additional 514 cases from Center 1 lacked manually delineated ROI information. These cases were used for constructing and evaluating the ADS, with 411 cases randomly assigned to the training set and 103 cases to the internal test set. The dataset from Center 2 was used as an external test set. The patient enrollment process is depicted in Fig. 2. Inclusion and exclusion criteria are detailed in the supplementary materials.

This retrospective study was approved by the Institutional Review Board of Center 1 (2018KT110) and the Institutional Review Board of Center 2 (LM2020001), and all methods were conducted in accordance with approved guidelines.

Feature-based diagnostic model

Manual segmentation

The segmentation of CT and PET images was performed manually using ITK-SNAP (version 3.8.0). To minimize interobserver variability, a reader with over five years of clinical diagnostic experience initially manually segmented the regions of interest. Subsequently, the second reader with over 15 years of clinical experience verified and confirmed the segmentation results. In cases of disagreement between the two physicians, discussions were held to reach a consensus on the ROI. Both readers were blinded to the clinical outcomes throughout the annotation process. To enhance efficiency, a drawing tablet with a stylus (Huion Kamvas 16, Huion, Shenzhen, China) was utilized during the manual segmentation.

Two readers independently analyzed the PET/CT images and extracted common clinical parameters for each patient. The following parameters were recorded: the major diameter of the tumor, the minor diameter of the tumor, the consolidation-to-tumor ratio (CTR), the tumor location, the shortest distance from the tumor to the pleura, and SUVmax. Detailed definitions of these parameters can be found in Table S1.

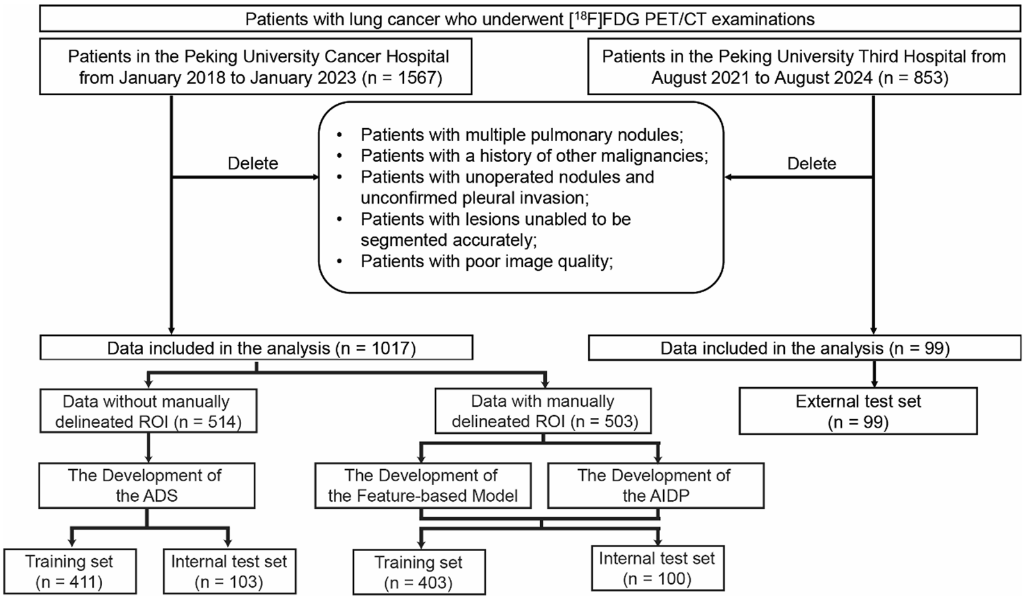
Patients with lung cancer who underwent [18F]FDG PET/CT examinations
Patients in the Peking University Cancer Hospital from January 2018 to January 2023 (n = 1567)
Patients in the Peking University Third Hospital from August 2021 to August 2024 (n = 853)
Delete
Delete
• Patients with multiple pulmonary nodules;
• Patients with a history of other malignancies;
• Patients with unoperated nodules and unconfirmed pleural invasion;
• Patients with lesions unabled to be segmented accurately;
• Patients with poor image quality;
Data included in the analysis (n = 1017)
Data included in the analysis (n = 99)
Data without manually delineated ROI (n = 514)
Data with manually delineated ROI (n = 503)
External test set (n = 99)
The Development of the ADS
The Development of the Feature-based Model
The Development of the AIDP
Training set (n = 411)
Internal test set (n = 103)
Training set (n = 403)
Internal test set (n = 100)

Fig. 2 Patient enrollment workflow

Feature selection

Using the Pyradiomics package (version 3.1.0) in Python (version 3.8.19), we extracted 200 radiomic features from the ROIs of the PET/CT images. Subsequently, z-score normalization was applied to the entire dataset to standardize the extracted features. Feature importance was assessed using SHapley Additive exPlanations (SHAP) analysis to identify the most diagnostically valuable features. SHAP-based importance ranking of all candidate features followed by incremental testing of top 1 to 15 features through accuracy evaluation on the internal test set. To mitigate batch effects introduced by different devices, experimental conditions, or scan batches, we applied the ComBat harmonization method [18] after feature extraction. The ComBat-harmonized data were then used for feature selection, model training, and evaluation.

Model construction and evaluation

In this study, random forest, k-nearest neighbors, logistic regression, decision tree, and adaptive boosting were employed as base learners, with the multi-layer perceptron model serving as the meta-learner in the stacking ensemble framework. The performance of the model was evaluated using various metrics, including accuracy, sensitivity, and specificity. The model's robustness was assessed through internal test using three-times five-fold cross-validation techniques, and its generalizability was further tested on the external test set.

AI-enabled diagnostic pipeline

Segmentation network

For lesion segmentation on CT images, we utilized the nnU-Net framework within the MONAI library [19, 20]. For PET images, lesion segmentation was performed using the ResU-Net framework [21]. The training process involved adjusting all image matrices to a size of 128 × 128 × 128, followed by normalization of the image data to enhance model performance.

The training process was conducted on a server equipped with an Intel Core i7-8700 CPU, 64 GB RAM, and an NVIDIA RTX 3090Ti GPU. For CT lesion segmentation using nnU-Net, the total training duration was 35 h over 300 epochs. The PET lesion segmentation model (ResU-Net) required 18 h for 300 epochs. The combined training time for both models was 53 h. We used Python version 3.11.9 along with MONAI version 1.3.2 and PyTorch version 2.4.1 for the implementation.

The training was conducted using the Adam optimizer, a variant of stochastic gradient descent. The initial learning rate was set to 10^{-3} and adjusted with a gamma value of 0.99 throughout the training process. The performance of the segmentation models was evaluated using the dice similarity coefficient (DSC) to quantify the overlap between the predicted and ground-truth segmentation masks, while Dice Loss was employed as the loss function during the backpropagation phase of the CNNs. Each dataset combination underwent training for 300 epochs to ensure robust learning.

For lung contour segmentation, we utilized the open-source tool TotalSegmentator. TotalSegmentator is an AI model trained using nnU-Net V2 on a large dataset, capable of automatically segmenting 117 organs across the body from CT data [22]. The implementation was carried out using Python version 3.11.9 and TotalSegmentator version 2.4.0, enabling efficient and accurate segmentation of lung contours.

Automatic feature extraction

Using the largest cross-section of the tumor as a standard, common clinical features were automatically extracted from the tumor ROIs and lung contours obtained through segmentation. The geometric relationships were primarily calculated using the Shapely package (version 2.0.1), OpenCV-Python (version 4.8.1), and NumPy (version 2.0.0). Detailed methodologies for feature extraction are provided in the supplementary materials. The clinical parameters extracted by the algorithm were compared with the manually labeled parameters using Spearman's rank correlation coefficient. The lesion ROIs and lung contours obtained from the automated segmentation network were used to extract the nine radiomic features selected for the feature-based model.

Automatic diagnostic system for PI prediction

ADS evaluation

The development of the ADS combines the FDM and the AIDP described. Based on the segmentation results of the lesions and lung contours, 6 automatically extracted features and 9 radiomic features were incorporated to construct the predictive model.

To compare the performance of models, confusion matrix analysis was performed, calculating key metrics such as accuracy, specificity, and sensitivity. Additionally, the receiver operating characteristic (ROC) curve was used to evaluate the model's ability to distinguish between lung cancer patients with and without PI.

Comparative analysis

To evaluate the performance of the ADS in predicting PI, a comparative analysis was conducted against existing methods. The performance of the default ADS (**Model 1**, MONAI nnU-Net for lesion segmentation, TotalSegmentator for organ segmentation) was benchmarked against two alternative models to evaluate its effectiveness in PI rediction. Other configuration unchanged, **Model 2** utilized the Moosez segmentation tool [23] for lesion delineation, while **Model 3** applied a maximum connected region algorithm for lung contour segmentation. Each model was evaluated on the same datasets for consistency and comparability.

Efficiency-efficacy comparison with human physicians

To validate the clinical utility and efficiency advantages of the ADS system, we conducted a comparative experiment involving three nuclear medicine physicians with varying clinical experience (Rater 1: in clinical training; Rater 2: 5 years' diagnostic experience; Rater 3: 10 years' diagnostic experience). Ten randomly selected preoperative cases were independently evaluated using a standardized protocol. The physicians were blinded from the diagnosis results and the individual identifier of the patient, and they used a previously developed set of rules for PI diagnosis. Detailed criteria can be found in the supplementary materials. Initially, all readers performed visual assessment of pleural invasion status for each case, with interpretation time and diagnostic conclusions recorded. The ADS system simultaneously processed identical cases through its automated pipeline, with total processing time and prediction outcomes documented. Subsequently, all physicians manually extracted three key imaging features previously established in publication [24] (jellyfish sign, pleural thickening, and pleural contact area), while the ADS executed automated feature extraction. To quantitatively analyze system efficiency, we conducted detailed time-motion analysis of ADS workflow components using Case 1 as a representative example. In the time analysis of ADS workflow components, we utilized a high-performance system featuring an AMD Ryzen 9 9900X (12-core/24-thread, 5.37 GHz), 64 GB DDR5 RAM, RTX 4060 Ti GPU (16 GB VRAM/Tensor Cores), and NVMe SSDs, integrated via PCIe 5.0 motherboard architecture to ensure accelerated computational parallelism and data throughput.

All evaluations were performed on standardized workstations to eliminate hardware variability. Diagnostic accuracy was calculated against pathological gold standards, with ADS performance validated against its internal test set labels.

Results

Patient characteristics

A total of 1,017 cases from Center 1 and 99 cases from Center 2 were included in this study. The baseline characteristics of the patient populations are summarized in Table 1. The prevalence of PI was 45.3% in Center 1 and 52.5% in Center 2. Although there was a time gap between the data collection periods at each center, no statistically significant differences in baseline clinical characteristics were observed between the two centers, indicating comparable populations for analysis.

Table 1 Clinical characteristics of the patients with pulmonary nodules in the different centers

Characteristic	Center 1 (n = 1017)	Center 2 (n = 99)
Gender		
Male	446 (43.9%)	35 (35.4%)
Female	571 (56.1%)	64 (64.6%)
Age		
Median (IQR) (y)	61 (54, 67)	65 (60, 72)
Smoke		
No	680 (66.9%)	71 (71.7%)
Yes	337 (33.1%)	28 (28.3%)
CEA		
≤ 5.0 ng/ml	807 (79.4%)	78 (78.8%)
> 5.0 ng/ml	210 (20.6%)	21 (21.2%)
CYFRA21-1		
≤ 3.3 ng/ml	827 (81.3%)	76 (76.8%)
> 3.3 ng/ml	190 (18.7%)	23 (23.2%)
NSE		
≤ 15.2 ng/ml	710 (69.8%)	79 (79.8%)
> 15.2 ng/ml	307 (30.2%)	20 (20.2%)
Thrombus		
No	887 (87.2%)	80 (80.8%)
Yes	130 (12.8%)	19 (19.2%)
Spiculation		
No	247 (24.3%)	31 (31.3%)
Yes	770 (75.7%)	68 (68.7%)
Lobulation		
No	158 (15.5%)	16 (16.2%)
Yes	859 (84.5%)	83 (83.8%)
Contact distance from the pleura		
Median (IQR) (cm)	0.14 (– 0.79, 1.51)	– 0.5 (– 0.8, 1.4)
Relationship of nodules to the pleura		
I/II	259 (25.5%)	24 (24.2%)
III/IV/V	758 (74.5%)	75 (75.8%)
CTR		
Median (IQR)	0.2 (0, 0.6)	0.7 (0.4, 0.9)
SUVmax		
Median (IQR)	3.6 (1.6, 7.5)	3.8 (1.6, 9.9)
The status of PI		
Negative	556 (54.7%)	47 (47.5%)
Positive	461 (45.3%)	52 (52.5%)

IQR interquartile range, *CEA* carcinoembryonic antigen, *CYFRA21-1* cytokeratin 19 fragment, *NSE* neuron-specific enolase, *CTR* consolidation-to-tumor ratio, *SUVmax* the maximum standardized uptake value

Table 2 Performance comparison between our feature-based diagnostic model and other models

	Proposed model	Previous model	Radiomics-based model
Internal test set			
Accuracy	**0.90** (90/100)	0.66 (66/100)	0.79 (79/100)
Sensitivity	**0.90** (43/48)	0.63 (30/48)	0.85 (41/48)
Specificity	**0.90** (47/52)	0.69 (36/52)	0.73 (38/52)
AUC	**0.95**	0.73	0.83
External test set			
Accuracy	**0.82** (81/99)	0.63 (62/99)	0.67 (66/99)
Sensitivity	**0.79** (41/52)	0.52 (27/52)	0.69 (36/52)
Specificity	**0.85** (40/47)	0.74 (35/47)	0.64 (30/47)
AUC	**0.90**	0.69	0.78

The highest metrics among the models are highlighted with bold font

Development and evaluation of the feature-based diagnostic model

Feature selection

From the PET/CT images, a total of 200 radiomic features were extracted. As shown in Figure S1, the model reached peak validation accuracy of 0.79 (± 0.01% fluctuation) when 9 radiomics features were included. These included one gray-level run-length matrix (GLRLM) feature, two gray-level co-occurrence matrix (GLCM) features, three first-order statistical features, and three shape-based features. Additionally, six clinical features were analyzed for importance. The importance ranking of clinical features is shown in Figure S2(a). Figure S2(b) illustrates the SHAP-derived importance ranking of these radiomic features, highlighting their relative contributions to the diagnostic model.

Model performance

The diagnostic performance of the FDM was evaluated across multiple datasets, demonstrating strong predictive capabilities for PI. Constructed using 9 selected radiomic features and 6 clinical features, the model achieved an area under the ROC curve (AUC) of 0.95 in the internal test set and 0.90 in the external test set, indicating robust discrimination in both settings. Furthermore, the stacked ensemble model, which combined multiple individual models, outperformed the single model, achieving accuracies of 0.90 and 0.82 on the internal and external test datasets, respectively. The performance comparison between the stacked ensemble model and individual machine learning models is shown in Table S2.

Incorporating clinical features into the model significantly enhanced its predictive performance compared to a radiomics-only approach. To further validate the

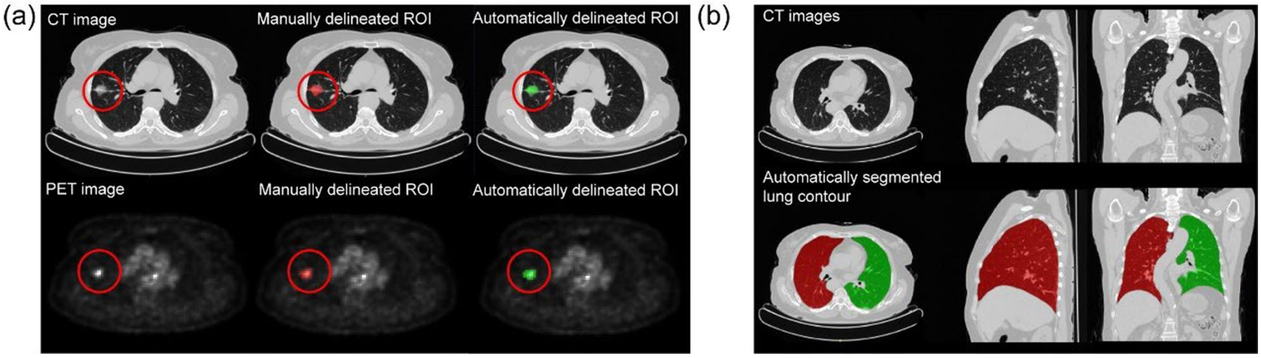
(a)
CT image
Manually delineated ROI
Automatically delineated ROI
PET image
Manually delineated ROI
Automatically delineated ROI
(b)
CT images
Automatically segmented lung contour

Fig. 3 Visualization of segmentation results. **a** Lesion segmentation; **b** Lung contour segmentation

Table 3 The Spearman correlation coefficient between the automated and manual clinical parameters

Variable	Internal test set		External test set	
	r	*p value*	*r*	*p value*
Major diameter	0.80	P < 0.05	0.88	P < 0.05
Minor diameter	0.87	P < 0.05	0.90	P < 0.05
Tumor Location	0.79	P < 0.05	0.94	P < 0.05
CTR	0.57	P < 0.05	0.58	P < 0.05
SUVmax	0.94	P < 0.05	0.96	P < 0.05
Shortest distance from the pleura	0.47	P < 0.05	0.45	P < 0.05

superiority of the proposed model [24], its performance was compared with a recently published method that predicts PI risk based on preoperative CT imaging. Using the same datasets, the proposed model demonstrated superior performance in terms of accuracy, sensitivity, specificity, and AUC across all sets, as detailed in Table 2.

Development and evaluation of the AIDP

Segmentation performance evaluation

The segmentation performance of the AIDP was evaluated on both PET and CT images, focusing on the automated delineation of lung lesions and lung contours. The nnU-Net and ResUNet models demonstrated high accuracy in segmenting lesions on CT and PET images, respectively. The learning curve of the neural network is shown in Figure S3. For CT images, the nnU-Net model achieved a mean DSC of 0.85, indicating a satisfactory level of agreement in delineating lung lesions. Figure 3a shows representative examples of the segmented lung tumors on CT, demonstrating the model's ability to capture tumor boundaries effectively.

For PET images, the ResU-Net model yielded a mean DSC of 0.89, reflecting a high degree of accuracy in segmenting regions of high [^{18}F]FDG uptake. The model was particularly effective at identifying metabolic tumor boundaries, as demonstrated in the PET examples shown in Fig. 3a.

The lung contours automatically segmented using TotalSegmentator were highly consistent with the actual lung boundaries, demonstrating excellent performance in capturing anatomical details. Figure 3b presents a visual example of the automatically segmented lung contours.

Feature extraction

Clinical features were automatically extracted from the segmented ROIs. Table 3 provides a summary of the correlation analysis between the automated and manual clinical parameters. These high correlation values indicate a strong agreement between the automated system and manual measurements, further validating the reliability of the automated feature extraction process. Figure S4 illustrates the visualization of the correlation analysis between manually delineated and automatically extracted features in the external test set.

Development and evaluation of the ADS for PI prediction

Model performance

The diagnostic performance of the ADS was evaluated across internal and external test datasets, demonstrating strong and consistent predictive ability for PI. This fully automated pipeline, integrating segmentation, feature extraction, and classification, achieved an AUC of 0.95 on the internal test set and 0.89 on the external test set, reflecting excellent discrimination across diverse patient populations.

The end-to-end system achieved an accuracy of 0.87, sensitivity of 0.90, and specificity of 0.85 in the internal test set, while achieving an accuracy of 0.82, sensitivity of 0.83, and specificity of 0.81 in the external set. Comparative analysis with the human-based diagnostic pipeline indicated that the ADS achieved comparable overall

Table 4 The performance comparison between the ADS and other models

	Model 1	Model 2	Model 3
Internal test set			
Accuracy	**0.87** (90/103)	0.77 (79/103)	0.86 (89/103)
Sensitivity	**0.90** (43/48)	**0.90** (43/48)	0.88 (42/48)
Specificity	**0.85** (47/55)	0.65 (36/55)	**0.85** (47/55)
AUC	**0.95**	0.85	0.93
External test set			
Accuracy	**0.82** (81/99)	0.73 (72/99)	0.78 (77/99)
Sensitivity	**0.83** (43/52)	**0.83** (43/52)	0.81 (42/52)
Specificity	**0.81** (38/47)	0.62 (29/47)	0.74 (35/47)
AUC	**0.89**	0.80	0.85

The highest metrics among the models are highlighted with bold font

diagnostic performance while demonstrating superior sensitivity.

Comparative analysis

On the different datasets, ADS (**Model 1**) demonstrated the highest diagnostic performance in terms of accuracy, sensitivity, and specificity, outperforming both **Model 2** and **Model 3**. These results consistently highlight ADS's advantage in providing reliable and robust predictions across diverse datasets. A detailed performance comparison is provided in Table 4.

In terms of discrimination ability, the AUC was 0.95 for **Model 1** on the internal dataset, significantly higher than 0.85 for **Model 2** and 0.93 for **Model 3**. On the external dataset, ADS maintained superior performance with an AUC of 0.89, compared to 0.80 for **Model 2** and 0.85 for **Model 3**. The ROC curves of the three models are shown in the Fig. 4.

Efficiency-efficacy comparison with human physicians

Figure 5a demonstrates the diagnostic performance and time distribution for PI assessment across three readers and the ADS system. The ADS achieved superior mean interpretation time (29.81 ± 0.47 s/case) compared to physicians' 104.3 ± 59.08 s (Reader 1), 65.4 ± 25.70 s (Reader 2), and 66.2 ± 10.54 s (Reader 3). Diagnostic accuracy analysis revealed ADS superiority (0.9) over human readers (0.7, 0.6, and 0.6 for Readers 1–3 respectively). Notably, ADS maintained consistent processing speed across all cases, while physicians' interpretation times exhibited significant case-dependent variability.

Figure 5b illustrates temporal differences in manual versus automated feature extraction. The ADS completed required feature extraction in 9.45 ± 0.43 s/case, substantially faster than physicians' manual measurements (109.3 ± 59.08 s, 42.6 ± 25.70 s, and 40.3 ± 10.54 s for Readers 1–3 respectively). Automated feature quantification demonstrated particular advantages in complex geometric calculations, as evidenced by significantly lower time variance (ADS SD = 0.43 s vs. physician SD range = 10.54–59.08 s).

Detailed workflow analysis for Case 1 (Fig. 5c) revealed total ADS processing time of 31.58 s, comprising automated segmentation (16.00 s, 50.66%) and feature extraction (9.85 s, 31.19%), with remaining time allocated to data preprocessing and classification. This granular temporal decomposition quantitatively demonstrates the system's efficiency optimization across computational stages.

Discussion

This study presents a strategy for an automatic diagnostic pipeline, demonstrated by the development and validation of an ADS for predicting PI in NSCLC patients using PET/CT imaging. By integrating radiomic and

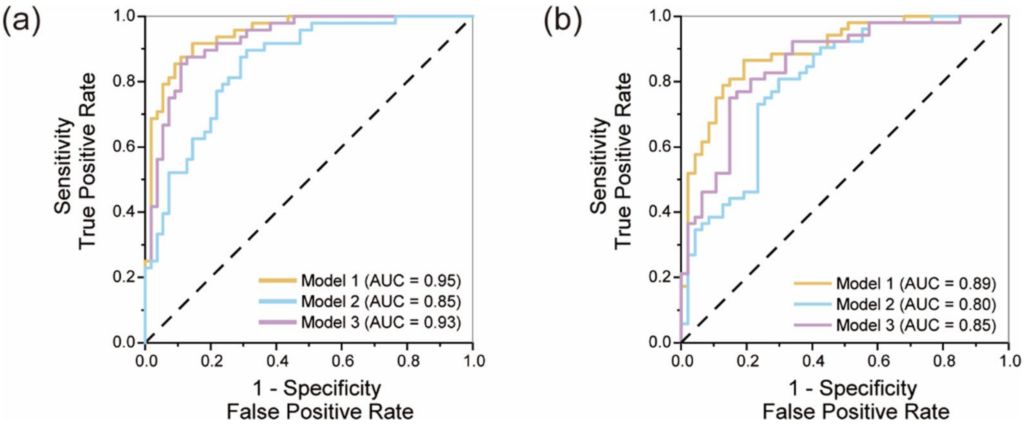

Fig. 4 The ROC curves of the three models in the internal test set (**a**) and external test set (**b**)

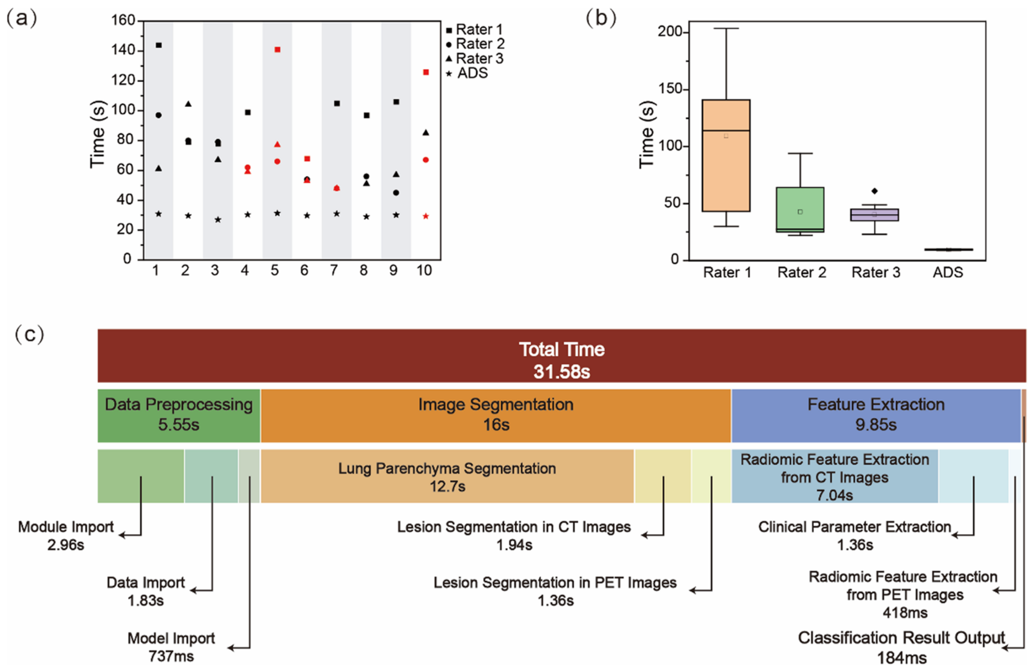

Fig. 5 Efficiency-effectiveness comparison between the ADS system and human readers. **a** Comparative performance of pleural invasion interpretation between physicians and the ADS system. Black and red markers denote correct and erroneous interpretations, respectively. The *x*-axis represents case ID, while the *y*-axis indicates interpretation time per case. **b** Box-and-whisker plots comparing manual versus automated feature extraction times across three readers and the ADS system. **c** Time decomposition analysis of the ADS workflow for Case 1, illustrating computational resource allocation across system modules

deep learning methodologies, the ADS achieves a balance between automation and interpretability, addressing existing gaps in clinical diagnostic workflows.

In this study, SUVmax and *CT GLRLM RunLengthNonUniformity* were identified as the most influential predictors for PI in lung cancer patients. SUVmax reflects the metabolic activity of the tumor, which is closely related to the tumor's aggressiveness and invasive potential. Previous studies have confirmed that SUVmax is an independent predictor of PI in lung cancer, demonstrating good stability [25]. Studies have also reported that NSCLC patients with SUVmax ≤ 1.3, regardless of tumor size, did not exhibit pleural invasion, and a positive correlation was observed between SUVmax and pleural infiltration ($r = 0.456$, $p < 0.001$) [26]. High SUVmax values are predictive of increased risks of lymph node involvement, distant metastasis, and recurrence, thereby guiding preoperative staging and therapeutic strategies [27]. Additionally, a high SUVmax was associated with a shorter overall survival and disease—free survival, suggesting that SUVmax could be used as a biomarker to predict the efficacy of neoadjuvant chemotherapy and patient prognosis [28]. On the other hand, *CT GLRLM RunLengthNonUniformity* captures tumor texture heterogeneity, which correlates with histopathological complexity and genomic instability. Tumors with higher heterogeneity are more likely to exhibit invasive growth patterns due to clonal diversity and adaptive resistance mechanisms [29]. These findings underscore the importance of integrating metabolic and texture-based data to improve the accuracy of PI prediction.

In the evaluation of the FDM in this study, we found that the stacked ensemble model outperformed individual machine learning models, a finding consistent with previous research [30, 31]. The stacked ensemble approach, by combining the predictions of multiple base learners, effectively reduces the bias and variance that individual models may have, thereby enhancing the generalization ability and accuracy of the model. Furthermore, the performance of the stacked ensemble model surpassed that of previously proposed single-modality CT models [24] and single-radiomics models, further demonstrating the importance of multimodal data in disease prediction. Single-modality models often rely solely on one type of feature, such as structural information from CT images or quantitative features from radiomics data. In our study, the supplementation of clinical features and PET/CT metabolic information significantly improved model performance. This highlights that PET/CT provides valuable biological activity information of the tumor, while clinical features contribute essential structural and locational information, and their integration greatly enhanced the model's accuracy in predicting pleural invasion.

In the correlation analysis between automatically extracted clinical features and manually delineated features, we observed certain discrepancies, particularly in the parameters of the CTR and the shortest distance from the pleura. The automatic extraction method relies on image processing algorithms, such as ellipse fitting and contour extraction, which are effective in capturing tumor regions and associated structural information. However, due to the sensitivity of these algorithms to image preprocessing, segmentation accuracy, and noise, certain features—such as the boundary of the solid component and the contact points between the tumor and the pleura—may not align perfectly with manual delineation, leading to discrepancies in the numerical values between the two methods. Despite these differences, the automatic extraction method does not diminish its clinical value. On the contrary, it offers a fast and standardized approach for feature extraction, which can assist clinicians in decision-making processes. The automation of feature extraction provides efficiency and consistency, significantly saving time when processing large datasets and reducing human biases.

The ADS demonstrated high diagnostic accuracy, comparable to experienced clinicians, by leveraging a combination of radiomic and clinical features. These features, being well-understood in clinical practice, enhance the interpretability of the model. Research has shown that compared to manual diagnostic methods, the ADS system demonstrates higher sensitivity in the external test set. This can be attributed to its ability to systematically analyze large datasets and detect subtle patterns that might be overlooked in manual evaluations. The automated segmentation and feature extraction processes ensure consistency, reducing observer variability, which is a significant limitation of manual methods. This transition holds promise for improving workflow efficiency in clinical settings and ensuring consistent diagnostic outcomes across institutions.

To address concerns regarding protocol variability, we systematically evaluated the impact of ComBat harmonization on model performance. Post-correction, the ADS demonstrated statistically significant improvements(Supplementary Table S3). These results underscore the necessity of batch effect correction in radiomic workflows, particularly when integrating multi-institutional data. While retrospective harmonization mitigates existing biases, future studies should prioritize real-time adaptive methods to address protocol heterogeneity in dynamic clinical settings.

In this research, we have demonstrated the feasibility of a strategy to circumvent the limitations of traditional, manual nomogram application. The AIDP we developed accommodates various FDMs, provided that the segmentation capabilities and generated features are compatible. This can be conceptualized as a set of optical filters, as illustrated in Fig. 6. The "optical apparatus" (AIDP) can accommodate an exchangeable set of "filters" (FDM), with each filter representing a distinct diagnostic model applicable to the same disease. Each filter can yield a specific diagnostic result (e.g., malignancy, pathological subtype),

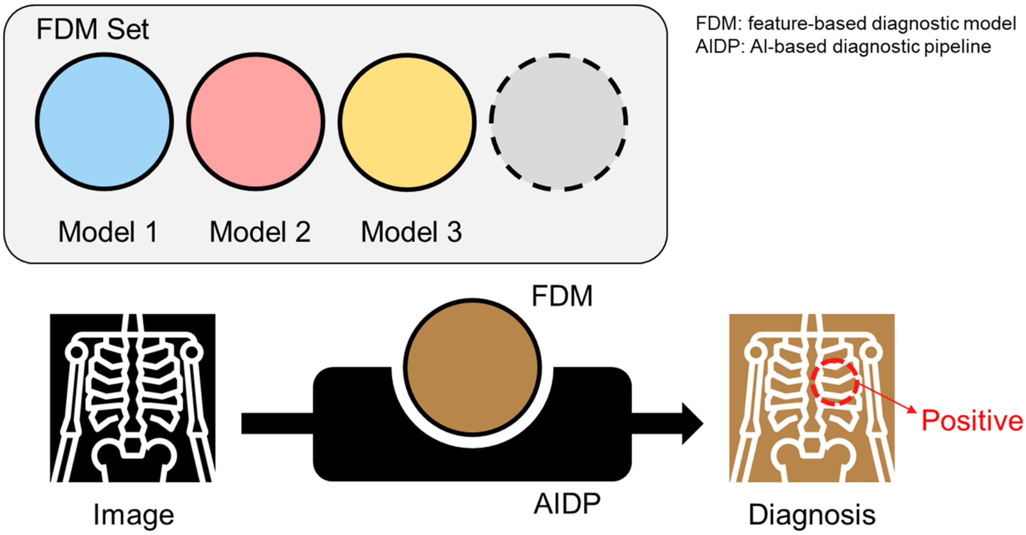
FDM Set
Model 1
Model 2
Model 3
FDM: feature-based diagnostic model
AIDP: AI-based diagnostic pipeline
FDM
Image
AIDP
Diagnosis
Positive

Fig. 6 The "optical filter" illustration of the proposed automatic diagnostic strategy. The AI-based diagnostic pipeline (AIDP, "apparatus") proposed may support a set of different feature-based diagnostic models (FDMs, "filters"). Inserting different FDMs may diagnose different aspects of the disease shown on the image. Icon by pmicon at flaticon.com

revealing a particular aspect of the input images. By implementing such a strategy, the practical application of various radiomics-based nomograms in clinical settings could be significantly enhanced.

Despite the promising results of this study, several limitations need to be acknowledged. First, the retrospective study design may introduce selection bias. Additionally, while automated feature extraction was employed to improve efficiency, the accuracy of the segmentation and feature extraction process is still heavily dependent on the preprocessing and segmentation algorithms. In some challenging cases, such as small or irregularly shaped lesions, the algorithm might struggle to achieve accurate tumor delineation, which could lead to discrepancies in extracted features.

Future work will focus on several improvements. Improving segmentation accuracy through the use of more deep learning techniques, such as attention mechanisms, will be essential to handle challenging tumor delineation cases. Additionally, large-scale multi-center validation with heterogeneous imaging protocols and tumor subtypes will be prioritized to ensure robustness across diverse clinical scenarios. Future work will prioritize prospective multi-center trials comparing the performance of the ADS against standard qualitative analyses conducted by experienced radiologists and clinicians. Such trials will assess diagnostic concordance, workflow efficiency, and interobserver variability reduction in real-world settings. Additionally, we aim to investigate the impact of ADS-guided predictions on clinical decision-making, including surgical planning, adjuvant therapy selection, and patient outcomes. Furthermore, we plan to explore the system's adaptability to other clinical scenarios (e.g., lymph node staging or recurrence prediction), leveraging its modular "optical filter" framework to support diverse diagnostic models. Through these efforts, we envision the ADS becoming a trusted tool for clinicians to optimize NSCLC management and, ultimately, improve patient outcomes.

Conclusion

In conclusion, this study highlights the potential of a fully automated diagnostic system to transform the assessment of PI in NSCLC patients. By combining AI technologies with clinical expertise, the ADS achieves high diagnostic accuracy, efficiency, and scalability. With further validation and optimization, the ADS has the potential to enhance clinical workflows and improve patient outcomes.

Abbreviations

AI	Artificial intelligence
CNNs	Convolutional neural networks
PI	Pleural invasion
NSCLC	Non-small cell lung cancer
SUVmax	The maximum standardized uptake value
FDM	Feature-based diagnostic model
AIDP	AI-enabled diagnostic pipeline
ADS	Automatic diagnostic system
ROI	Regions of interest
CTR	Consolidation-to-tumor ratio
SHAP	SHapley Additive exPlanations
DSC	Dice similarity coefficient
ROC	Receiver operating characteristic
GLRLM	Gray-level run-length matrix
GLCM	Gray-level co-occurrence matrix
AUC	Area under the ROC curve

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13550-025-01264-0.

Additional file 1

Acknowledgements

Not applicable.

Author contributions

All authors contributed to the study conception and design. All authors read and approved the final manuscript. Conception and design: XK, XM, and AZ; Development of methodology: XK, XM, ZY, and NL; Acquisition of data: XK, AZ, XZ, MZ, WZ, and JL; Analysis and interpretation of data: XM, XK; Writing, review, and/or revision of the manuscript: XK, XM, and AZ; Administrative, technical, or materials support: XZ, XM, NL, and ZY; Study supervision: XM, NL, and ZY.

Funding

The current research was financially supported by the National Natural Science Foundation of China (Grant No. 82372001), Key Clinical Project of Peking University Third Hospital (BYSYZD2024043) and the National Key R&D Program of China (No. 2022YFC2409405).

Availability of data and materials

Processed data for this study's results can be made available upon reasonable request from the corresponding author. The original data are not publicly available due to containing information that could compromise the privacy of research participants. The original source code is publicly available at https://github.com/KongXiangxing/AIDP-towards-FDM.

Declarations

Ethics approval and consent to participate

This retrospective study was approved by the Institutional Review Board of Peking University Cancer Hospital (2018KT110) and the Institutional Review Board of Peking University Third Hospital (LM2020001), and all methods were conducted in accordance with approved guidelines.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 18 February 2025 Accepted: 23 May 2025
Published online: 12 June 2025

References

1. Esteva A, Chou K, Yeung S, et al. Deep learning-enabled medical computer vision. NPJ Digit Med. 2021;4:5.
2. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. Med Image Anal. 2017;42:60–88.
3. Rahman H, Bukht TFN, Imran A, Tariq J, Tu S, Alzahrani A. A Deep learning approach for liver and tumor segmentation in CT images using ResUNet. Bioengineering (Basel). 2022;9:368.
4. Zhang L, Li H, Zhao S, et al. Deep learning model based on primary tumor to predict lymph node status in clinical stage IA lung adenocarcinoma: a multicenter study. J Natl Cancer Cent. 2024;4:233–40.
5. Leung KH, Rowe SP, Sadaghiani MS, et al. Deep semisupervised transfer learning for fully automated whole-body tumor quantification and prognosis of cancer on PET/CT. J Nucl Med. 2024;65:643–50.
6. Park SH, Han K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. Radiology. 2018;286:800–9.
7. Shen JY, Zhang CJP, Jiang BS, et al. Artificial intelligence versus clinicians in disease diagnosis: systematic review. JMIR Med Inform. 2019;7:15.
8. Mirbabaie M, Stieglitz S, Frick NRJ. Artificial intelligence in disease diagnostics: a critical review and classification on the current state of research guiding future direction. Health Technol. 2021;11:693–731.
9. Liu QX, Deng XF, Zhou D, Li JM, Min JX, Dai JG. Visceral pleural invasion impacts the prognosis of non-small cell lung cancer: a meta-analysis. Eur J Surg Oncol. 2016;42:1707–13.
10. Ryu JS, Ryu HJ, Lee SN, et al. Prognostic impact of minimal pleural effusion in non-small-cell lung cancer. J Clin Oncol. 2014;32:960–7.
11. Jiang L, Liang W, Shen 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.
12. Mamun M, Farjana A, Al Mamun M, Ahammed MS (2022) Lung cancer prediction model using ensemble learning techniques and a systematic review analysis. IEEE World AI IoT Congress (AIIoT). IEEE, Seattle, WA, pp 187–193
13. He NN, Xi Y, Yu DY, Yu CQ, Shen WY. Construction of IL-1 signalling pathway correlation model in lung adenocarcinoma and association with immune microenvironment prognosis and immunotherapy: multi-data validation. Front Immunol. 2023;14:12.
14. Liang WH, Zhang L, Jiang GN, et al. Development and validation of a nomogram for predicting survival in patients with resected non-small-cell lung cancer. J Clin Oncol. 2015;33:861.
15. Mao QX, Xia WJ, Dong GC, et al. A nomogram to predict the survival of stage IIIA-N2 non-small cell lung cancer after surgery. J Thorac Cardiovasc Surg. 2018;155:1784.
16. Santos J, Oliveira BC, Araujo JDB, et al. State-of-the-art in radiomics of hepatocellular carcinoma: a review of basic principles, applications, and limitations. Abdom Radiol. 2020;45:342–53.
17. Bandini M, Fossati N, Briganti A. Nomograms in urologic oncology, advantages and disadvantages. Curr Opin Urol. 2019;29:42–51.
18. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2006;8:118–27.
19. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods. 2021;18:203–11.
20. Cardoso MJ, Li W, Brown R et al (2022) MONAI: an open-source framework for deep learning in healthcare. arXiv:2211.02701
21. Zhang Z, Liu Q, Wang Y. Road extraction by deep residual U-Net. IEEE Geosci Remote Sens Lett. 2018;15:749–53.
22. Wasserthal J, Breit HC, Meyer MT, et al. TotalSegmentator: robust segmentation of 104 anatomic structures in CT images. Radiol Artif Intell. 2023;5: e230024.
23. Shiyam Sundar LK, Yu J, Muzik O, et al. Fully automated, semantic segmentation of whole-body (18)F-FDG PET/CT images based on data-centric artificial intelligence. J Nucl Med. 2022;63:1941–8.
24. Sun Q, Li P, Zhang J, et al. CT predictors of visceral pleural invasion in patients with non-small cell lung cancers 30 mm or smaller. Radiology. 2024;310: e231611.
25. Maeda R, Isowa N, Onuma H, et al. The maximum standardized 18F-fluorodeoxyglucose uptake on positron emission tomography predicts lymph node metastasis and invasiveness in clinical stage IA non-small cell lung cancer. Interact Cardiovasc Thorac Surg. 2009;9:79–82.
26. Tanaka T, Shinya T, Sato S, et al. Predicting pleural invasion using HRCT and 18F-FDG PET/CT in lung adenocarcinoma with pleural contact. Ann Nucl Med. 2015;29:757–65.
27. Ling T, Zhang L, Peng R, Yue C, Huang L. Prognostic value of (18)F-FDG PET/CT in patients with advanced or metastatic non-small-cell lung cancer treated with immune checkpoint inhibitors: a systematic review and meta-analysis. Front Immunol. 2022;13:1014063.
28. Zhuang F, Haoran E, Huang J, et al. Utility of 18F-FDG PET/CT uptake values in predicting response to neoadjuvant chemoimmunotherapy in resectable non-small cell lung cancer. Lung Cancer. 2023;178:20–7.
29. Sabaawy HE. Genetic heterogeneity and clonal evolution of tumor cells and their impact on precision cancer medicine. J Leuk (Los Angel). 2013;1:1000124.
30. Huang M-L, Liao Y-C. Stacking ensemble and ECA-EfficientNetV2 convolutional neural networks on classification of multiple chest diseases including COVID-19. Acad Radiol. 2023;30:1915–35.
31. Dai H, Wang Y, Fu R, et al. Radiomics and stacking regression model for measuring bone mineral density using abdominal computed tomography. Acta Radiol. 2021;64:028418512110681.

Publisher's Note