

Delta-Enhanced CT-Based 2.5D Deep Learning Model for Noninvasive Prediction of Leptomeningeal Metastasis in Lung Adenocarcinoma

Jiachun Sun

Henan Key Laboratory of Cancer Epigenetics, Cancer Institute, The First Affiliated Hospital, College of Clinical Medicine, Henan University of Science and Technology, Luoyang, Henan, China.

Renke Yu

Department of General Practice, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou, Zhejiang, China.

Jingxiang Su

Department of Radiation Oncology, Tianjin Medical University Cancer Institute & Hospital, Tianjin, China.

Bo Sun

Henan Key Laboratory of Cancer Epigenetics, Cancer Institute, The First Affiliated Hospital, College of Clinical Medicine, Henan University of Science and Technology, Luoyang, Henan, China.

Yitao Fan

Henan Key Laboratory of Cancer Epigenetics, Cancer Institute, The First Affiliated Hospital, College of Clinical Medicine, Henan University of Science and Technology, Luoyang, Henan, China.

Yingda Xie

Henan Key Laboratory of Cancer Epigenetics, Cancer Institute, The First Affiliated Hospital, College of Clinical Medicine, Henan University of Science and Technology, Luoyang, Henan, China.

Hanliang Jiang

Department of Respiratory and Critical Care Medicine, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou, Zhejiang, China.

Lu Zheng

zhenglu_cank@163.com

Department of Medical Oncology, The Affiliated Cancer Hospital of Zhengzhou University and Henan Cancer Hospital, Zhengzhou, Henan, China.

Weidong Han

hanwd@zjcc.org.cn

Department of Colorectal Medical Oncology, Zhejiang Cancer Hospital, Chinese Academy of Sciences, Hangzhou, Zhejiang, China.

Research Article

Keywords: leptomeningeal metastasis, lung adenocarcinoma, delta radiomics, 2.5D deep learning model, multi-instance learning

Posted Date: March 16th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9115524/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Delta-Enhanced CT-Based 2.5D Deep Learning Model for Noninvasive Prediction of Leptomeningeal Metastasis in Lung Adenocarcinoma

Jiachun Sun^{1#}, Renke Yu^{2#}, Jingxiang Su³, Bo Sun¹, Yitao Fan¹, Yingda Xie¹, Hanliang Jiang⁴, Lu Zheng^{5*}, Weidong Han^{6*}

1. Henan Key Laboratory of Cancer Epigenetics, Cancer Institute, The First Affiliated Hospital, College of Clinical Medicine, Henan University of Science and Technology, Luoyang, Henan, China.
2. Department of General Practice, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou, Zhejiang, China.
3. Department of Radiation Oncology, Tianjin Medical University Cancer Institute & Hospital, Tianjin, China.
4. Department of Respiratory and Critical Care Medicine, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou, Zhejiang, China.
5. Department of Medical Oncology, The Affiliated Cancer Hospital of Zhengzhou University and Henan Cancer Hospital, Zhengzhou, Henan, China.
6. Department of Colorectal Medical Oncology, Zhejiang Cancer Hospital, Chinese Academy of Sciences, Hangzhou, Zhejiang, China.

#These authors contributed equally to this work.

* Corresponding authors: Weidong Han, Zhejiang Cancer Hospital, Chinese Academy of Sciences, 38#, Guangji Road, Hangzhou, Zhejiang, China 310022. Phone: +86-571-88144401; E-mail: hanwd@zjcc.org.cn; and Lu Zheng, Department of Medical Oncology, The Affiliated Cancer Hospital of Zhengzhou University and Henan Cancer Hospital, Zhengzhou, Henan, China. E-mail: zhenglucank@163.com.

Abstract

Leptomeningeal metastasis (LM) is a devastating complication of lung adenocarcinoma, the current diagnosis of which relies on invasive cerebrospinal fluid cytology or costly MRI, both of which are limited in sensitivity and accessibility. To enable early, noninvasive LM risk prediction, we developed a 2.5D delta-radiomics deep learning framework based on routinely acquired contrast-enhanced CT scans. The proposed architecture integrates multi-instance learning to aggregate slice-level features into patient-level representations while preserving spatial context and controlling computational complexity. Delta features, derived by comparing pre-LM and post-LM scans, quantified subtle longitudinal changes within intratumoral and peritumoral regions. Multiple classifiers were evaluated, with XGBoost selected for its optimal performance, and Grad-CAM visualization was employed to assess spatial attention, revealing biologically plausible regions that drove model predictions. A nomogram was constructed to facilitate individualized clinical application, and decision curve analysis (DCA) demonstrated a consistent net benefit across a broad range of decision thresholds. In the test set, the delta-based multi-instance learning signature (MILDelta) achieved an AUC of 0.871, outperforming models trained solely on pre-LM (AUC 0.574) or post-LM (AUC 0.661) images, whereas a combined model incorporating imaging signatures and clinical variables, including EGFR mutation status, yielded the highest predictive accuracy, with an AUC of 0.910. This interpretable delta-informed model leverages standard contrast-enhanced CT to achieve individualized, noninvasive LM risk stratification with demonstrated generalizability and clinical utility, offering a scalable tool for timely intervention in precision oncology workflows.

Key words

leptomeningeal metastasis; lung adenocarcinoma; delta radiomics; 2.5D deep learning model; multi-instance learning

Lung cancer remains the leading cause of cancer-related mortality worldwide, and its high fatality rate is largely attributed to distant metastases [1]. Among its subtypes, lung adenocarcinoma is the most prevalent [1], and the incidence of leptomeningeal metastasis (LM) has shown a rising trend in recent years [2, 3]. LM represents a devastating form of cancer, characterized by an extremely poor prognosis [4] and a median survival time of only a few months [5]. Currently, the diagnostic gold standard depends on cerebrospinal fluid cytology (CFC) or magnetic resonance imaging (MRI) [6]. However, these methods have notable limitations, including low sensitivity, invasive procedures, and high costs [7]. Furthermore, in accordance with

the follow-up requirements of the National Comprehensive Cancer Network (NCCN) guidelines for non-small cell lung cancer (version 2026 V3) [8], since there are no symptoms related to the central nervous system, routine cranial MRI examinations are not recommended. Therefore, before central nervous system symptoms occur, LM is often overlooked [9]. Therefore, developing a noninvasive and early predictive tool for LM risk is highly clinically important for improving patient management and guiding precision therapy.

In recent years, radiomics and deep learning have demonstrated great potential in medical image analysis by extracting high-dimensional quantitative features from conventional enhanced computed tomography (CT) images, that are often imperceptible to the human eye [10, 11]. These techniques have been widely applied to tumor diagnosis, prognosis evaluation, and treatment response prediction [11]. Nevertheless, most existing studies treat tumors as homogeneous entities and overlook their intrinsic spatial heterogeneity and temporal evolution. Conventional two-dimensional (2D) or three-dimensional (3D) deep learning models are often constrained either by the loss of spatial context information or by excessive computational complexity and data requirements [12]. Furthermore, limited attention has been given to the subtle imaging alterations of lung primary lesions before LM development, which may harbor critical information indicative of metastatic potential.

To address these limitations, this study proposes an innovative 2.5D deep learning framework based on enhanced CT imaging for the noninvasive prediction of LM in patients with lung cancer. The 2.5D approach integrates the tumor's largest cross-sectional image with several adjacent slices, effectively preserving essential spatial information while substantially reducing computational complexity and overfitting risk [13]. Additionally, a delta-based (temporal difference) analytical framework is introduced that leverages pre- and post-LM CT scans to capture dynamic imaging changes in tumor heterogeneity over time [14]. The delta-radiomics has been used for predicting the major pathological response to neoadjuvant chemoimmunotherapy in non-small cell lung cancer (NSCLC) by contrasting the contrast-enhanced CT before and after the treatment [15]. This study aims to develop and validate a delta-enhanced CT-based 2.5D deep learning model capable of noninvasively predicting the risk of LM among lung cancer patients. We hypothesize that this model can identify subtle intratumoral variations associated with metastatic potential, thereby providing a robust imaging biomarker to support early risk stratification and individualized clinical intervention.

Methods

Patients

This multicenter retrospective study included patients who were diagnosed between January 2012 and January 2024 across three separate large medical centers: the First Affiliated Hospital of Henan University of Science and Technology (Henan, China), Sir Run Run Shaw Hospital, School of Medicine, Zhejiang University (Zhejiang, China), and Henan Cancer Hospital (Henan, China). Eligible participants met the following inclusion criteria: (1) histopathological confirmation of lung adenocarcinoma; (2) cytological confirmation of LM by cerebrospinal fluid (CSF) analysis; (3) availability of two sets of enhanced CT images: one obtained at the initial lung adenocarcinoma diagnosis and another at suspected LM diagnosis; and (4) complete clinical records and follow-up information. The exclusion criteria were (1) the absence of pathological or CSF cytological confirmation; (2) incomplete radiological imaging data; (3) LM cases in limited-stage lung adenocarcinoma or patients whose primary tumor could not be evaluated because of prior targeted therapy or radiotherapy; and (4) missing clinical or follow-up data.

Clinical and pathological data collected from eligible patients included age, sex, smoking history, date of initial lung adenocarcinoma diagnosis, date of suspected LM diagnosis, TNM stage at first diagnosis according to the latest Union for International Cancer Control (UICC) and American Joint Committee on Cancer (AJCC) editions, epidermal growth factor receptor (EGFR) mutation status (exon 21 L858R and exon 19 deletion), history of radiotherapy or targeted therapy, and overall survival (OS) time.

All patients provided written informed consent to participate. Ethical approval was obtained from the institutional review boards of all three participating centers (2024-03-K0140 for the First Affiliated Hospital of Henan University of Science and Technology, 2025-2561-01 for Sir Run Run Shaw Hospital, 2025-775-001 for Henan Cancer Hospital), and the study was conducted in accordance with the Declaration of Helsinki and applicable privacy protection regulations.

To protect patient privacy, all Digital Imaging and Communications in Medicine (DICOM) files from CT scans were subjected to metadata cleansing before analysis. Using the PyDICOM library, original DICOM files were read, and all protected health information (PHI) and personally identifiable elements were removed while technical parameters essential for medical image processing, such as scan sequences and pixel spacing, were retained. The cleansed files were saved in a designated secure directory. This process ensured the removal of sensitive patient information while preserving necessary technical attributes for subsequent image analysis, thereby maintaining both

regulatory compliance and the protection of patient confidentiality.

Imaging Acquisition and Preprocessing

Two sets of contrast-enhanced chest CT scans were obtained for each patient: at the initial diagnosis of lung adenocarcinoma and at the suspected LM diagnosis. All CT scans were performed using multidetector CT scanners (64–128 slices) from major vendors (including GE Medical Systems, Siemens Healthineers, and Philips Healthcare) at the three participating centers. Scans were acquired in the supine position during full inspiration, covering the entire thoracic cavity from the lung apex to the diaphragm. Imaging parameters were standardized across centers to ensure reproducibility, with a slice thickness of 1.0–1.25 mm, a reconstruction matrix size of 512×512 , a tube voltage of 120 kVp, and automatic tube current modulation. Contrast enhancement was achieved through intravenous administration of nonionic iodinated contrast agents (1.5 mL/kg), with arterial-phase acquisition initiated approximately 25–30 seconds after injection using bolus-tracking techniques.

For image preprocessing, the original DICOM files were first anonymized as described above and then imported for segmentation. Prior to feature extraction and deep learning analysis, all the images were resampled to an isotropic voxel size of $1 \times 1 \times 1 \text{ mm}^3$ using trilinear interpolation to eliminate differences in resolution. Intensity normalization was applied by clipping the voxel intensities to a fixed Hounsfield unit range (–400 to 1200 HU for the lung window), followed by z score normalization to account for interscan intensity variations. Tumor regions of interest (ROIs), delineated via ITK-SNAP, were cropped from full CT volumes with a predefined margin to preserve peritumoral information. The segmentation masks and corresponding cropped CT images were saved in NIfTI format and further processed into 2D slices and slice stacks for 2.5D deep learning input. This preprocessing pipeline ensured spatial consistency, intensity normalization, and standardized data representation across all the cases, allowing for robust feature extraction and model training.

Tumor segmentation

Three experienced radiologists performed the tumor segmentation on arterial-phase images under lung window settings using ITK-SNAP (version 3.8; <http://www.itksnap.org>). The workload was evenly distributed, and each radiologist delineated the tumor boundary across all the slices to create volumetric segmentation masks. An internal verification protocol was implemented, wherein each case was independently reviewed by a second radiologist. Discrepancies were discussed to reach a consensus, with adjudication by a third radiologist if needed.

Data allocation and validation

The dataset was randomly divided into a training set (70%) and an internal validation set (30%). An independent external test cohort was used to evaluate model generalizability and ensure robustness across unseen data. Stratified sampling was applied to maintain balanced class distributions between cohorts.

2.5D Deep Learning Radiomics

Data Preparation and Cropping

A 2.5D deep learning framework was designed to incorporate both spatial and temporal information from pre- and post-LM CT data. For each patient, multiple cross-sectional slices centered on the largest tumor section were extracted, including adjacent slices (± 1 , ± 2 , and ± 4), to preserve spatial continuity. The pre- and post-LM images were assigned to the first and second channels, respectively, while a third channel was constructed to represent the delta information between the two time points, which was calculated as the elementwise difference:

$$\Delta I = I_{\text{post}} - I_{\text{pre}} \quad (1)$$

These three channels (pre, post, and delta) were concatenated to form a unified 2.5D input tensor. This multitemporal design enabled the model to jointly learn baseline imaging characteristics and metastasis-induced changes, thereby enhancing the representation capability for subsequent predictive modeling. The methodological workflow, illustrated in Figure 1, demonstrates the effectiveness of integrating radiomics and deep learning-based delta modeling for improved diagnostic accuracy.

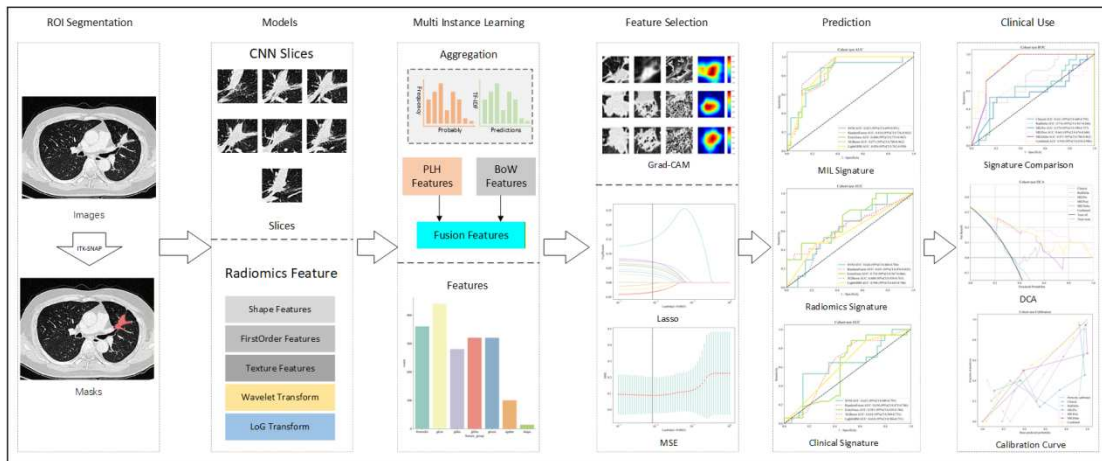


Figure 1. Overall workflow of this study

Slice-Level Model Training

Each slice was treated as an independent training instance, with patient-level

labels propagated to all corresponding slices. Deep feature extraction was conducted using convolutional neural networks, including DenseNet121, ResNet18, ResNet50, ResNet101, and Inception_v3, all initialized with ImageNet pretrained weights to improve convergence and model stability in limited-sample medical imaging tasks (details in **Supplementary 1A**).

Multi-Instance Learning

To aggregate the slice-level information into a patient-level representation, a multi-instance learning (MIL) strategy was applied. Features extracted from individual slices were combined through probabilistic aggregation methods, including Predictive Likelihood Histograms and Bag-of-Words representations with TF-IDF weighting. This approach preserved discriminative slice-level features while forming a comprehensive patient-level representation. Details are provided in **Supplementary 1B**.

Signature Construction

- **Multi-Instance Learning 2.5D Signature (MIL):** The fused features derived from the MIL framework were standardized and subjected to sequential feature selection, including univariate analysis, correlation filtering, mRMR (top 32 features), and LASSO. First, we applied minimum Redundancy–Maximum Relevance (mRMR) to rank features by mutual information with the outcome while penalizing redundancy; the top 32 features were retained as candidates. Second, we used LASSO logistic regression with 10-fold cross-validation (λ selected by the 1-SE rule) to obtain a sparse subset; features with non-zero coefficients were carried forward to model training. The optimized features were used to train support vector machine, random forest, and ExtraTrees classifiers. Model optimization was conducted through grid search and cross-validation, and the final signature demonstrated strong discrimination and generalizability.
- **Delta Radiomics Signature (RadDelta):** Derived from radiomic features extracted at the pre- and post-LM stages and transformed through delta modeling. The other parameters in the machine learning models were kept the same as those in the MIL signature.
- **Clinical Signature:** Clinical variables were used to model CSF. Clinical modeling is performed using the same classifiers to provide a reference for comparison with imaging-derived models.
- **Combined Signature:** A multimodal model was constructed by integrating the most informative imaging-based features with significant clinical factors. The final integrated model was visualized using a nomogram to enhance

interpretability and facilitate clinical application.

Model Evaluation: Model performance was assessed using an independent test cohort. Discrimination ability was measured by the area under the receiver operating characteristic curve (AUC). Calibration was evaluated through calibration plots and the Hosmer–Lemeshow goodness-of-fit test. Decision curve analysis (DCA) was used to assess clinical utility by estimating the net benefit across a range of threshold probabilities.

Statistical Analysis

The Shapiro–Wilk test was used to assess the normality of continuous variables. Depending on the data distribution, either the independent-samples t test or the Mann–Whitney U test was applied. Categorical variables were compared using the chi-square (χ^2) test. All analyses were performed on the OnekeyAI platform (v3.5.12) using Python (v3.7.12). Statistical analyses were conducted using Statsmodels (v0.13.2), radiomic features were extracted using PyRadiomics (v3.0.1), and machine learning models were implemented using Scikit-learn (v1.0.2).

Results

Clinical Features

A total of 231 patients with lung adenocarcinoma were included in this study: 126 from the First Affiliated Hospital of Henan University of Science and Technology (the training cohort, $n = 126$), 54 from Sir Run Run Shaw Hospital (the internal validation cohort, $n = 54$), and 51 from Henan Cancer Hospital (the external validation cohort, $n = 51$). Among them, 73 patients were diagnosed with LM by CFS cytology, while 158 patients had no LM. Baseline demographic and clinical characteristics are summarized in **Table 1**. No significant differences were observed between the cohorts (all $p > 0.05$), indicating effective randomization and balanced group allocation. The median age of all patients was 61 years (range: 31–84 years), including 121 (55.0%) females and 110 males (45.0%), and 72 patients (31.2%) had a history of smoking. A total of 147 patients (63.6%) had EGFR mutations, including 61 (41.5%) with exon 21 L858R mutations and 46 (31.3%) with exon 19 deletion mutations.

Characteristics	Train($n=126$)	Validation($n=54$)	Test($n=51$)	<i>P</i> -value
	Mean $\pm$ SD/ $N(\%)$	Mean $\pm$ SD/ $N(\%)$	Mean $\pm$ SD/ $N(\%)$	
Age	61.98 $\pm$ 10.86	61.35 $\pm$ 9.41	59.52 $\pm$ 9.74	0.71
Sex				0.72

Characteristics	Train(n=126) Mean ± SD/N(%)	Validation(n=54) Mean ± SD/N(%)	Test(n=51) Mean ± SD/N(%)	<i>P</i> -value
Female	66(52.38)	26(48.15)	29(56.86)	1
Male	60(47.62)	28(51.85)	22(43.14)	
Smoking				
No	83(65.87)	36(66.67)	40(78.43)	
Yes	43(34.13)	18(33.33)	11(21.57)	
pT stage				0.401
0	38(30.16)	16(29.63)	12(23.53)	0.723
1	40(31.75)	20(37.04)	14(27.45)	
2	22(17.46)	4(7.41)	6(11.76)	
3	25(19.84)	14(25.93)	15(29.41)	
4	1(0.79)	null	4(7.84)	
pN stage				0.723
0	16(12.70)	10(18.52)	3(5.88)	0.125
1	7(5.56)	3(5.56)	1(1.96)	
2	57(45.24)	25(46.30)	16(31.37)	
3	44(34.92)	16(29.63)	24(47.06)	
4	2(1.59)	null	7(13.73)	
pM stage				0.125
0	39(30.95)	10(18.52)	3(5.88)	0.199
1	87(69.05)	44(81.48)	48(94.12)	
Stage				
0	10(7.94)	3(5.56)	1(1.96)	
1	5(3.97)	2(3.70)	null	
2	24(19.05)	4(7.41)	2(3.92)	0.558
3	87(69.05)	45(83.33)	48(94.12)	
EGFR				
mutation	75(59.52)	35(64.81)	37(72.55)	
unkown	2(1.59)	null	null	
wild	49(38.89)	19(35.19)	14(27.45)	0.26
L858R21				

Characteristics	Train(n=126)	Validation(n=54)	Test(n=51)	<i>P</i> -value
	Mean ± SD/N(%)	Mean ± SD/N(%)	Mean ± SD/N(%)	
wild	91(72.22)	44(81.48)	35(68.63)	0.411
mutation	35(27.78)	10(18.52)	16(31.37)	
DEL19				
wild	104(82.54)	41(75.93)	40(78.43)	0.626
mutation	22(17.46)	13(24.07)	11(21.57)	
TKIAll				
combined	1(0.79)	null	null	0.604
none	105(83.33)	43(79.63)	47(92.16)	
yes	20(15.87)	11(20.37)	4(7.84)	
Efficacy of TKI				0.548
PD	16(12.70)	10(18.52)	3(5.88)	
PR	37(29.37)	16(29.63)	5(9.80)	
SD	33(26.19)	10(18.52)	41(80.39)	0.293
none	40(31.75)	18(33.33)	2(3.92)	
Radiation				
none	106(84.13)	48(88.89)	48(94.12)	0.293
yes	20(15.87)	6(11.11)	3(5.88)	
Efficacy of radiation				
PD	6(4.76)	1(1.85)	2(3.92)	
PR	9(7.14)	1(1.85)	1(1.96)	
SD	5(3.97)	4(7.41)	null	
none	106(84.13)	48(88.89)	null	

EGFR: Epidermal Growth Factor Receptor; TKIAll: the use of Tyrosine Kinase Inhibitor; PR: partial response; PD: progressive disease; SD: stable disease.

Table 1. Baseline characteristic of our cohorts.

We conducted a detailed univariate analysis of all clinical features (Table 2), focusing on calculating the odds ratio (OR) and associated p value for each feature. Notably, features with p values less than 0.05 were considered to be statistically significant. Consequently, these significant features were selected for further multivariate analysis and served as critical clinical comparison factors in our subsequent evaluations.

Characteristics	OR UNI	95% CI UNI	p-value UNI	OR MULTI	95% CI MULTI	p-value MULTI
Age	0.981	0.9750- 0.9860	<0.05	0.972	0.9420- 1.0030	0.137
Smoking	0.194	0.0990- 0.3840	<0.05	0.52	0.1490- 1.8130	0.389
Sex	0.250	0.1470- 0.4250	<0.05	2.037	0.6550- 6.3280	0.302
Radiation	0.429	0.1920- 0.9560	0.082			
TKIAll	0.432	0.3240- 0.5760	<0.05	1.187	0.5080- 2.7730	0.74
T	0.492	0.3850- 0.6290	<0.05	0.765	0.5240- 1.1160	0.244
N	0.627	0.5370- 0.7310	<0.05	0.909	0.5220- 1.5860	0.778
M	0.381	0.2570- 0.5650	<0.05	1.086	0.1550- 7.5990	0.944
Stage	0.681	0.5990- 0.7750	<0.05	1.479	0.4600- 4.7590	0.582
PR	0.524	0.4300- 0.6380	<0.05	0.98	0.6140- 1.5640	0.944
Efficacy	0.658	0.5800- 0.7460	<0.05	1.148	0.6900- 1.9080	0.656
EGFR	0.169	0.0840- 0.3400	<0.05	0.228	0.1030- 0.5060	<0.05
L858R21	0.667	0.3780- 1.1760	0.24			
DEL19	1.000	0.4960- 2.0160	1.0			

OR: odds ratio; CI:confidence interval; UNI: Univariable analysis; MULTI: multivariable analysis; EGFR: Epidermal Growth Factor Receptor; TKIAll: the use of Tyrosine Kinase Inhibitor; PR: partial response.

Table 2. Univariable and Multivariable Analysis of clinical features.

Results of the MIL Signature

Slice-level Results

Across the nested cross-validation, mRMR reduced the candidate feature space to

32 features per training fold, after which LASSO with stability selection yielded a final subset of 12 features for the radiomics-only model. The final number of features entering the classifier in the model using only imaging features is 12, while in the delta deep learning model, there are 12 features. The combined model includes the imaging Signature, the delta deep learning Signature, and one clinical variable. The Inception_v3 model achieved an accuracy of 0.800 and an AUC of 0.855 (95% CI: 0.8273–0.8825), with a sensitivity of 0.724, a specificity of 0.824, a PPV of 0.563, and an NPV of 0.905. In the validation cohort, it maintained robust generalization performance, with an accuracy of 0.706 and an AUC of 0.827 (95% CI: 0.7842–0.8699), along with high sensitivity (0.929) and satisfactory specificity (0.656). On the independent test set, the model achieved an accuracy of 0.711 and an AUC of 0.809 (95% CI: 0.7624–0.8549), outperforming the ResNet18, ResNet50, ResNet101, and DenseNet201 models across most evaluation metrics, which supported its stable and consistent discriminative capability across all the datasets. The details are listed in Table 3.

Model Name	Acc	AU C	95% CI	Sensitivity	Specificity	PPV	NPV	Cohort
resnet18	0.844	0.918	0.8981-0.9383	0.824	0.850	0.631	0.939	train
resnet18	0.722	0.737	0.6782-0.7967	0.643	0.740	0.360	0.901	validation
resnet18	0.543	0.698	0.6426-0.7532	0.958	0.336	0.419	0.941	test
resnet50	0.647	0.805	0.7720-0.8372	0.876	0.576	0.392	0.937	train
resnet50	0.661	0.662	0.5945-0.7287	0.586	0.679	0.293	0.878	validation
resnet50	0.619	0.736	0.6841-	0.874	0.492	0.462	0.886	test

Model Name	Acc	AU C	95% CI	Sensitivity	Specificity	PPV	NPV	Cohort
			0.7881					
resnet101	0.837	0.901	0.8778-0.9233	0.805	0.847	0.621	0.933	train
resnet101	0.582	0.772	0.7185-0.8259	0.957	0.497	0.302	0.981	validation
resnet101	0.622	0.668	0.6107-0.7250	0.706	0.580	0.457	0.798	test
densenet201	0.883	0.948	0.9323-0.9636	0.881	0.884	0.703	0.960	train
densenet201	0.619	0.783	0.7311-0.8356	0.886	0.558	0.313	0.956	validation
densenet201	0.599	0.656	0.5981-0.7142	0.689	0.555	0.436	0.781	test
inception_v3	0.800	0.855	0.8273-0.8825	0.724	0.824	0.563	0.905	train
inception_v3	0.706	0.827	0.7842-0.8699	0.929	0.656	0.380	0.976	validation
inception_v3	0.711	0.809	0.7624-0.8549	0.832	0.651	0.544	0.886	test

Acc: accuracy; AUC: area under the curve; CI: confidence interval; PPV: positive predictive value; NPV: negative predictive value.

Table 3. Slice level results of different CNN models.

Among all the candidate architectures (Figure 2), Inception_v3 demonstrated the most balanced performance in terms of accuracy, sensitivity, and AUC, reflecting superior feature extraction and generalization ability for the classification task. Therefore, features derived from the Inception_v3 backbone were selected for multi-instance learning-based aggregation to construct the final patient-level predictive model.

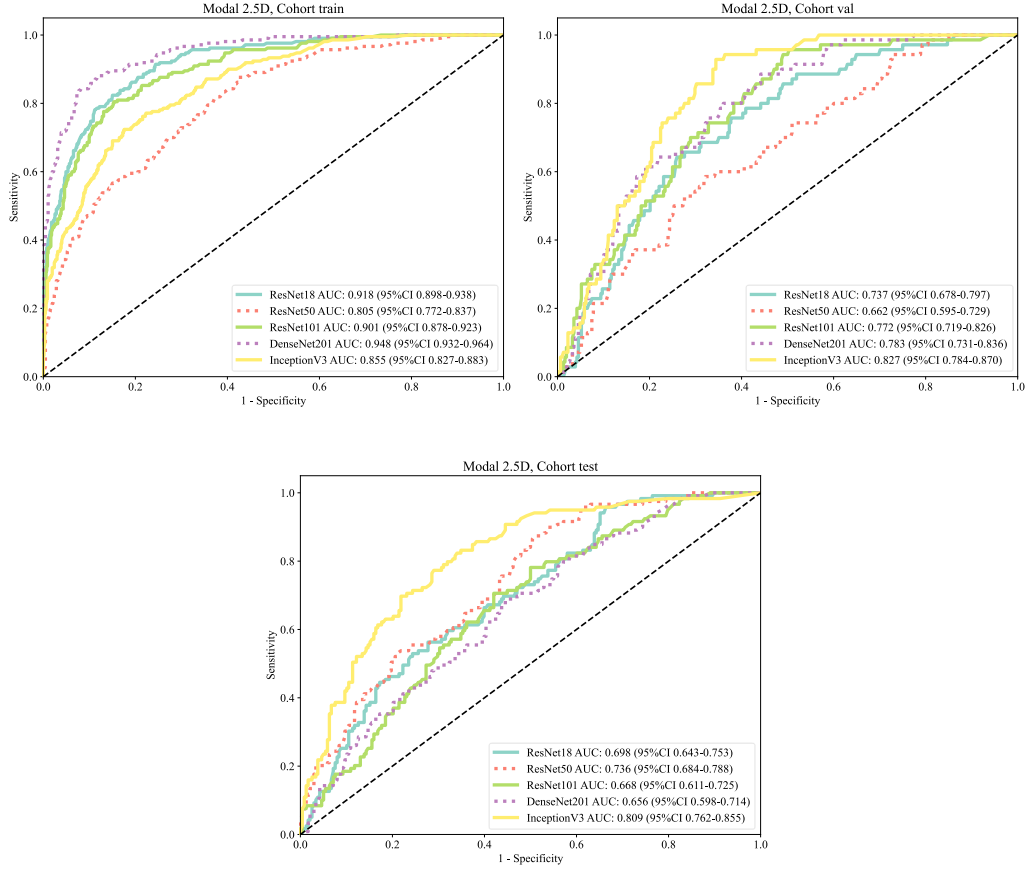


Figure 2. ROC of different models in slice level prediction.

Grad-CAM

To explore the recognition abilities of deep learning models across different samples, we utilized the Gradient-weighted Class Activation Mapping (Grad-CAM) technique. This method visualizes the activations in the final convolutional layer related to class predictions, as shown in Figure 3. Grad-CAM highlights the image areas that are most influential in the model's decisions, thereby improving our understanding of model interpretability.

The Grad-CAM visualizations for two representative samples are displayed in Figure 3, which demonstrates how the model selectively focuses on different regions

of the images to make predictions. This visualization is crucial for understanding the attention mechanism of the model in practical applications.

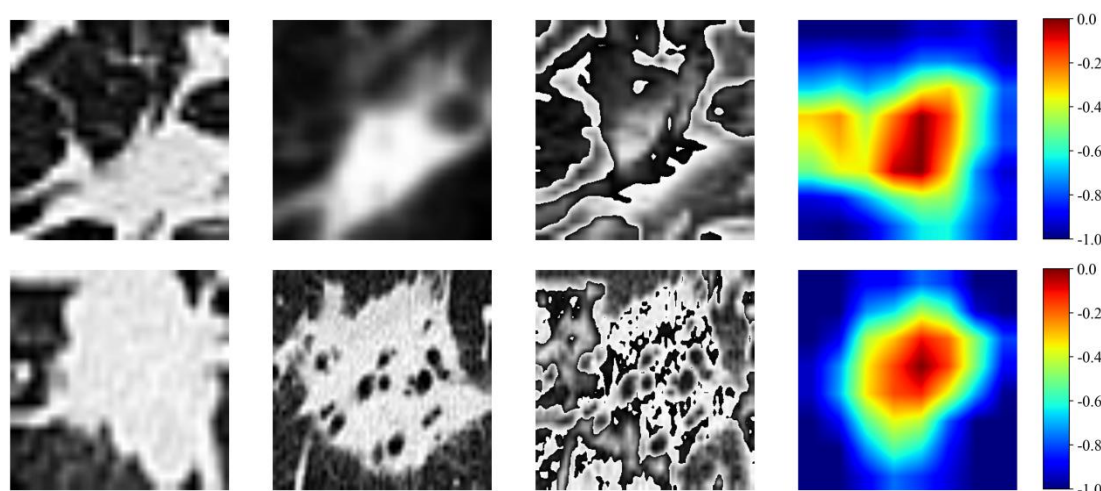


Figure 3. Grad-CAM visualizations for two representative samples

MIL Fusion Results

In the training cohort, the XGBoost model achieved an accuracy of 0.929 and an AUC of 0.956 (95% CI: 0.910–1.000), with a sensitivity of 0.967, a specificity of 0.917, a PPV of 0.784, and an NPV of 0.989. In the validation set, the model maintained an accuracy of 0.722 and an AUC of 0.910 (95% CI: 0.830–0.991), achieving perfect sensitivity (1.000) and acceptable specificity (0.659). In the independent test cohort, the model achieved an accuracy of 0.745 and an AUC of 0.871 (95% CI: 0.780–0.962), with balanced sensitivity (1.000) and specificity (0.618). Compared with the other classifiers, XGBoost consistently demonstrated higher AUC values and robust sensitivity across all the datasets, confirming its strong discriminative ability and generalizability. The details are listed in Table 4.

Model Name	Accuracy	AUC	95% CI	Sensitivity	Specificity	PPV	NPV	Cohort
SVM	0.913	0.975	0.950 - 1.000	0.933	0.906	0.757	0.978	train
SVM	0.815	0.839	0.671 - 1.000	0.900	0.795	0.500	0.972	validation
SVM	0.784	0.823	0.695 - 0.951	0.882	0.735	0.625	0.926	test
Random Forest	0.921	0.978	0.953 - 1.000	0.967	0.906	0.763	0.989	train
Random Forest	0.815	0.905	0.814 - 0.995	0.900	0.795	0.500	0.972	validation

Model Name	Accuracy	AUC	95% CI	Sensitivity	Specificity	PPV	NPV	Cohort
Random Forest	0.745	0.834	0.726 - 0.942	1.000	0.618	0.567	1.000	test
ExtraTrees	0.937	0.983	0.964 - 1.000	0.933	0.937	0.824	0.978	train
ExtraTrees	0.907	0.898	0.803 - 0.992	0.800	0.932	0.727	0.953	validation
ExtraTrees	0.745	0.840	0.733 - 0.947	1.000	0.618	0.567	1.000	test
XGBoost	0.929	0.956	0.910 - 1.000	0.967	0.917	0.784	0.989	train
XGBoost	0.722	0.910	0.830 - 0.991	1.000	0.659	0.400	1.000	validation
XGBoost	0.745	0.871	0.780 - 0.962	1.000	0.618	0.567	1.000	test
LightGBM	0.857	0.949	0.905 - 0.993	0.967	0.823	0.630	0.987	train
LightGBM	0.722	0.908	0.831 - 0.985	1.000	0.659	0.400	1.000	validation
LightGBM	0.745	0.856	0.762 - 0.950	1.000	0.618	0.567	1.000	test

AUC: area under the curve; CI: confidence interval; PPV: positive predictive value; NPV: negative predictive value.

Table 4. Evaluation Results of Different Models on the 2.5D MIL models

Overall, XGBoost exhibited the best balance between accuracy, sensitivity, and AUC among all the evaluated machine learning models (Figure 4a). Its adaptive boosting mechanism effectively captured nonlinear relationships and feature interactions, making it the most suitable model for final feature fusion and predictive modeling (Figure 4b–d). Therefore, the XGBoost-based signature was selected as the optimal classifier for subsequent analysis and interpretation.

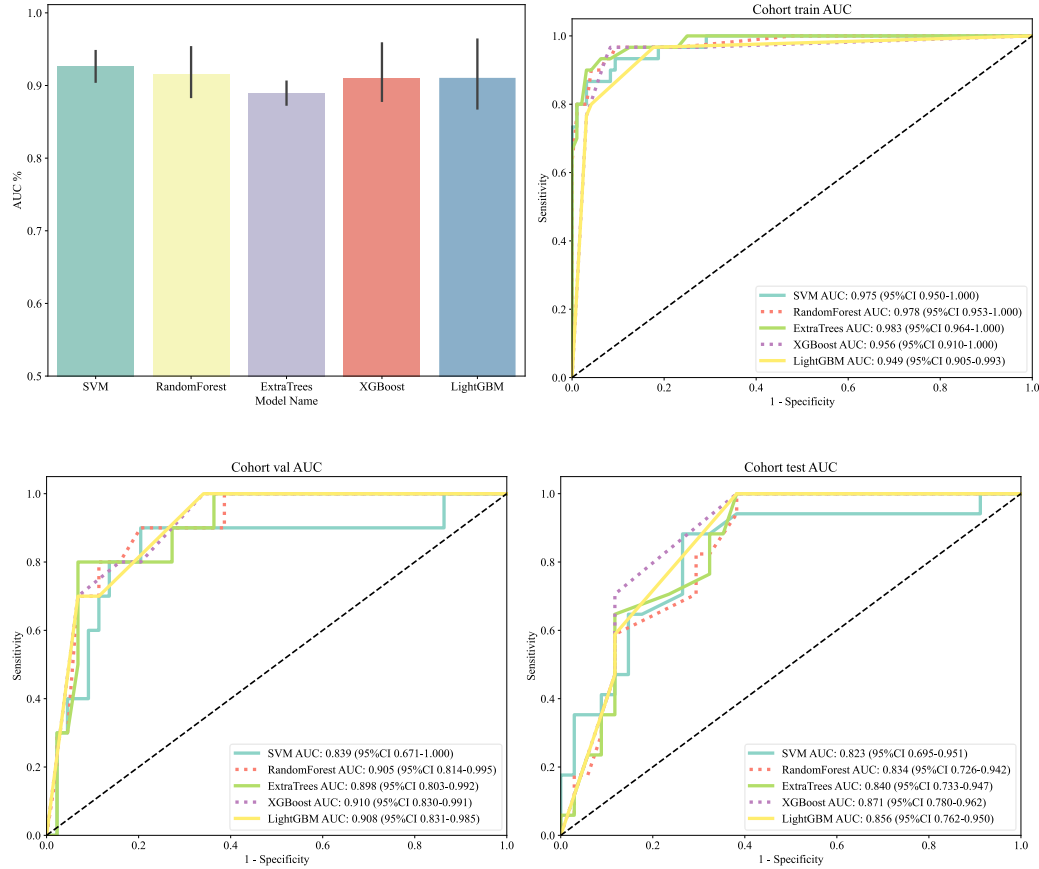


Figure 4a: Results of the 5-fold cross-validation used for hyperparameter tuning via GridSearch. Figures 4b-d: Comparative ROC curves of different models across various cohorts for the 2.5D MIL Signature.

Signature Comparison

In the training cohort, the delta-based deep learning model (MILDelta) achieved the best overall performance, with an accuracy of 0.929 and an AUC of 0.956 (95% CI: 0.9099–1.0000), indicating high sensitivity (0.967), specificity (0.917), PPV (0.784), and NPV (0.989). In the validation cohort, the MILDelta score was strongly generalized, with an accuracy of 0.722 and an AUC of 0.910 (95% CI: 0.8299–0.9906), indicating perfect sensitivity (1.000) and acceptable specificity (0.659). On the independent test set, MILDelta again demonstrated superior discriminative ability, with an accuracy of 0.745 and an AUC of 0.871 (95% CI: 0.7798–0.9624). In comparison, both the pre-LM (MILPre) and post-LM (MILPost) models yielded lower AUC values (0.574 and 0.661, respectively) in the test cohort, confirming the advantage of modeling temporal changes through delta features (Listed in Table 5).

Signature	Accuracy	AUC	95% CI	Sensitivity	Specificity	PPV	NPV	Cohort
Clinical	0.754	0.855	0.7813 - 0.9277	0.867	0.719	0.491	0.945	train
RadDelta	0.825	0.830	0.7336 - 0.9265	0.767	0.844	0.605	0.920	train
MILPre	0.643	0.780	0.7010 - 0.8597	1.000	0.531	0.400	1.000	train
MILPost	0.889	0.959	0.9286 - 0.9888	0.967	0.865	0.690	0.988	train
MILDelta	0.929	0.956	0.9099 - 1.0000	0.967	0.917	0.784	0.989	train
Combined	0.952	0.977	0.9544 - 0.9998	0.933	0.958	0.875	0.979	train
Clinical	0.574	0.711	0.5556 - 0.8671	0.900	0.500	0.290	0.957	validation
RadDelta	0.759	0.719	0.5217 - 0.9169	0.700	0.773	0.412	0.919	validation
MILPre	0.667	0.700	0.5458 - 0.8542	0.900	0.614	0.346	0.964	validation
MILPost	0.296	0.514	0.3181 - 0.7092	1.000	0.136	0.208	1.000	validation

Signature	Accuracy	AUC	95% CI	Sensitivity	Specificity	PPV	NPV	Cohort
MILDelta	0.722	0.910	0.8299 - 0.9906	1.000	0.659	0.400	1.000	validation
Combined	0.907	0.920	0.8386 - 1.0000	0.800	0.932	0.727	0.953	validation
Clinical	0.745	0.621	0.4492 - 0.7930	0.529	0.853	0.643	0.784	test
RadDelta	0.745	0.716	0.5667 - 0.8658	0.471	0.882	0.667	0.769	test
MILPre	0.706	0.574	0.3900 - 0.7570	0.529	0.794	0.562	0.771	test
MILPost	0.765	0.661	0.4736 - 0.8482	0.353	0.971	0.857	0.750	test
MILDelta	0.745	0.871	0.7798 - 0.9624	1.000	0.618	0.567	1.000	test
Combined	0.784	0.910	0.8340 - 0.9861	1.000	0.676	0.607	1.000	test

AUC: area under the curve; CI: confidence interval; PPV: positive predictive value; NPV: negative predictive value.

Table 5. Prediction performance of Habitat Region based rad signatures

Among all the models, the deep learning-based delta signature (MILDelta) exhibited the most stable and discriminative performance across cohorts. When combined with clinical features, the integrated model (Combined) achieved the

highest accuracy (0.784) and AUC (0.910, 95% CI: 0.8340–0.9861) on the test set, outperforming both the individual imaging and clinical models (Figure 5). These results indicate that the fusion of delta-based deep learning features with clinical variables provides complementary predictive information, resulting in a more robust and generalizable model for outcome prediction.

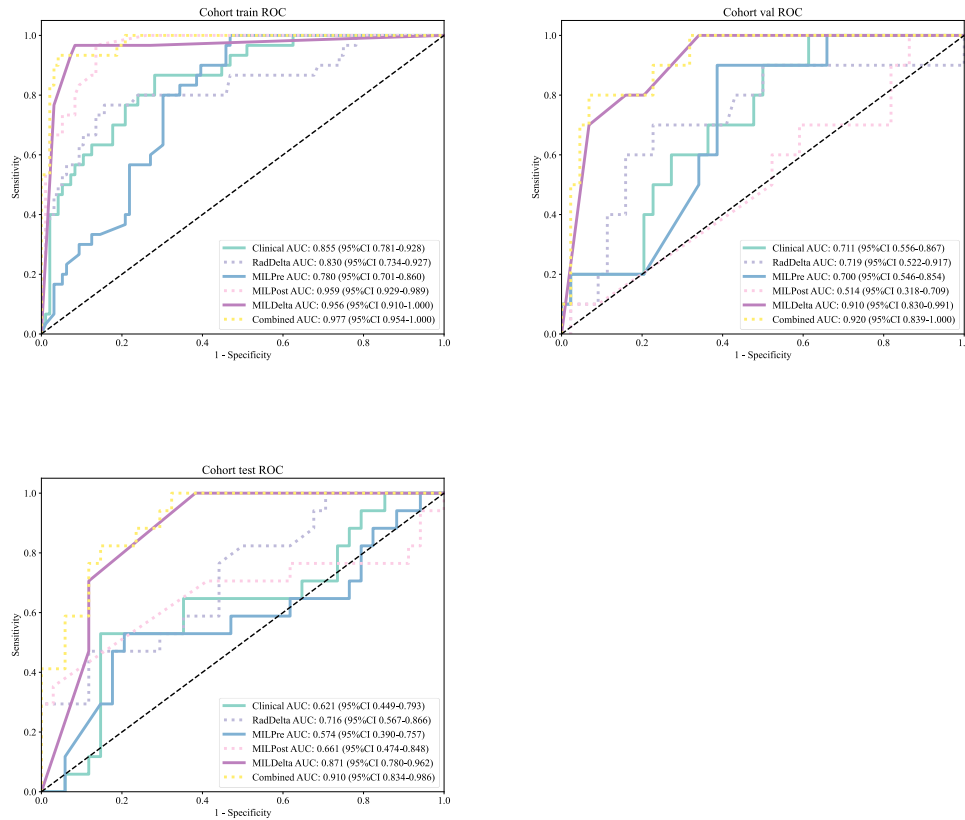


Figure 5. Different signatures' AUROC on all cohort.

Calibration Curve Analysis

The Hosmer–Lemeshow (HL) test was employed to evaluate the agreement between the predicted probabilities and actual observations, where smaller HL values indicate better calibration. The proposed nomogram model exhibited strong calibration performance, with HL values of 0.126, 0.644, and 0.747 in the training, validation, and test cohorts, respectively (Figure 6). These results demonstrated that the model maintained reliable predictive consistency and accurate probability estimation across all the datasets.

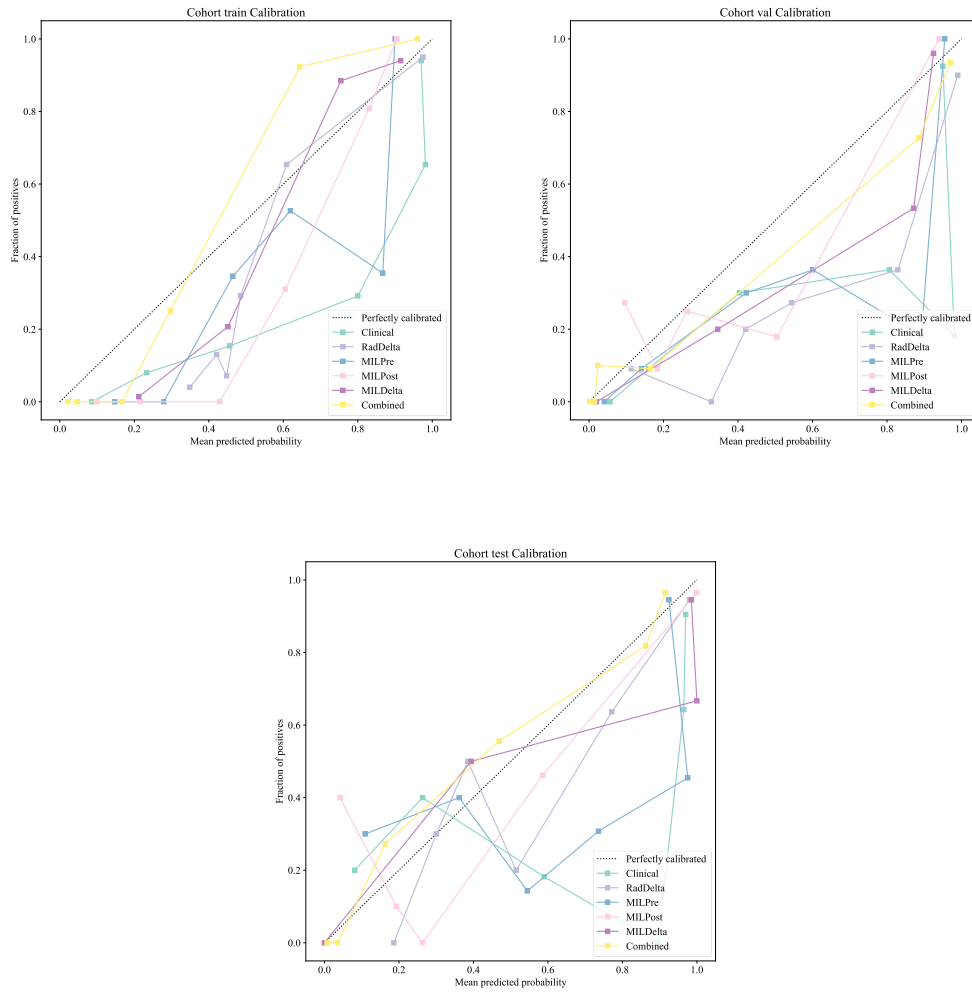


Figure 6. Different signatures' calibration curve on test cohort.

DeLong Test

The figure below compares the significance of improvement among different signatures across datasets, revealing that the combined model consistently exhibited notable enhancements over most other individual models. In our experiments, the combined model consistently outperformed the traditional radiomic and clinical models, with significant improvements, as indicated by p values less than 0.05 in the training and validation cohorts (Figure 7).

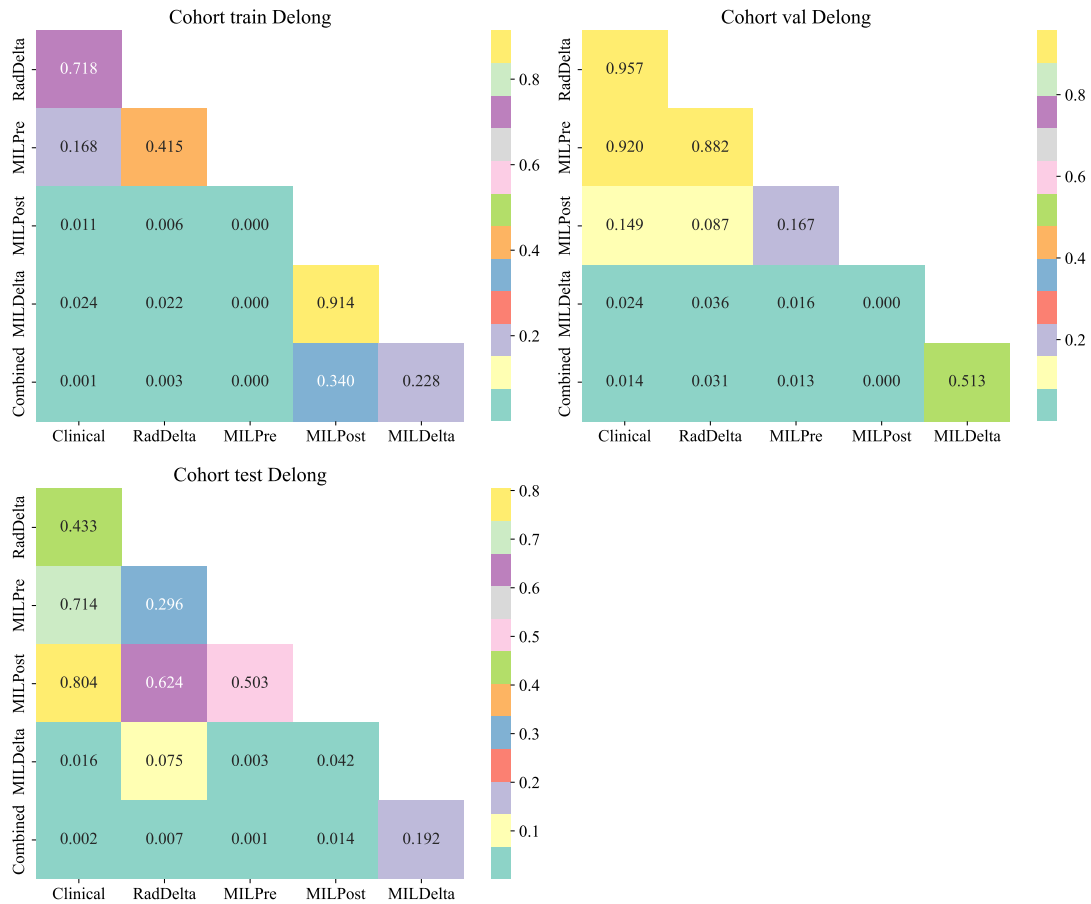


Figure 7. Delong et. al. of different signature

Clinical Use

Decision Curve Analysis (DCA): DCA demonstrated that the combined model yielded the greatest net benefit across a broad range of clinically relevant threshold probabilities, exceeding both the treat-all and treat-none strategies in the training cohort (0.02–0.71), internal validation cohort (0.02–0.72), and external test cohort (0.03–0.97) (Figure 8). Within these ranges, the combined model also showed higher net benefit than other models, indicating fewer avoidable interventions at comparable detection rates. At representative thresholds ($p_t \approx 0.30$), the net benefit advantage of the combined model suggests a meaningful reduction in unnecessary workups while maintaining or improving identification of true-positive cases. Outside these ranges, the net benefit curves for all strategies converged, implying limited clinical advantage of model-guided decisions at very low or very high-risk thresholds. These findings support the clinical utility and generalizability of the combined model across datasets. Figure 8 displays the DCA curves for both the training and testing cohorts. The analysis of these curves indicated that our combined model offered a notable advantage in terms of the net benefit derived from the predicted probabilities.

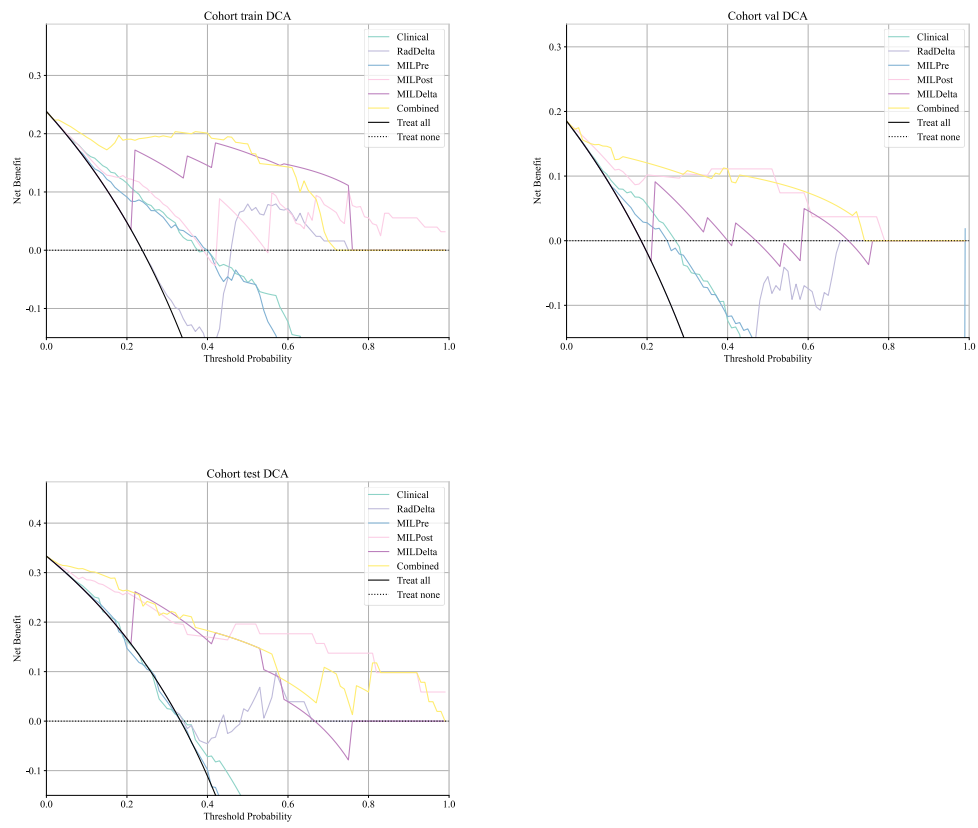


Figure 8. Different signatures' decision curve on test cohort.

Nomogram: We employed a nomogram to visualize the results of our combined model (Figure 9).

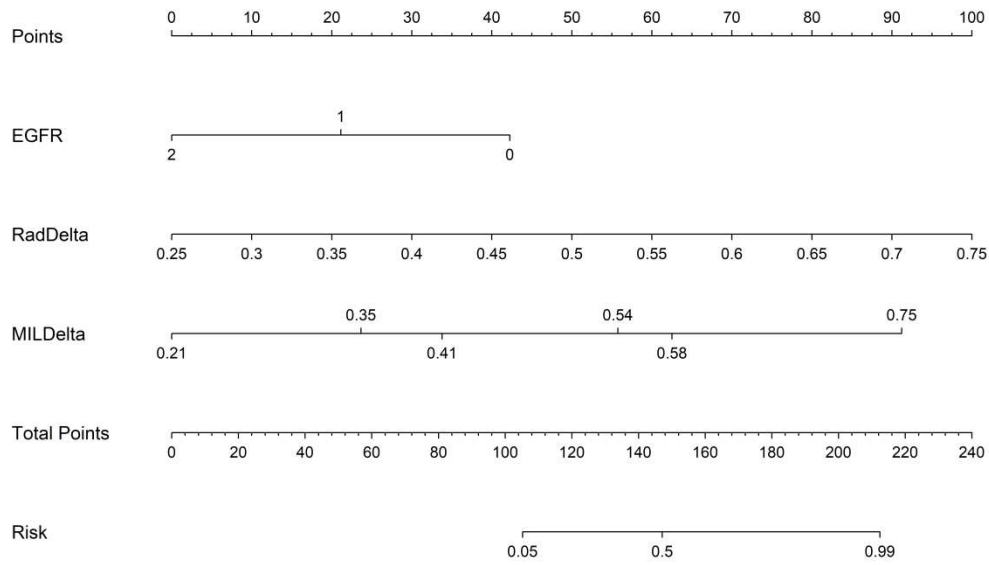


Figure 9. shows the nomogram for clinical use.

Discussion

In this multicenter study, a novel 2.5D deep learning framework integrating contrast-enhanced CT imaging and delta (temporal difference) radiomics features was proposed and validated, and clinical variables were selected to noninvasively predict the risk of LM in lung adenocarcinoma patients. The model consistently demonstrated high discriminative power and satisfactory calibration in the training, internal validation, and external independent test cohorts, outperforming models based solely on single-time point imaging or clinical information.

The proposed 2.5D deep learning architecture effectively preserved the essential spatial context by integrating the largest cross-sectional tumor slice with its adjacent slices, thereby retaining critical three-dimensional anatomical cues while maintaining manageable computational costs. This hybrid design bridged the gap between conventional 2D models—which often fail to capture spatial heterogeneity because of their reliance on single-plane information—and full 3D approaches, which, although rich in volumetric context, demand substantially higher data volumes and computational resources [16, 17]. Building on this foundation, Zewen Han and colleagues validated an MRI-based delta-radiomic deep learning framework for the preoperative prediction of glypican-3-positive hepatocellular carcinoma, underscoring the translational potential of temporal imaging feature analysis in oncology [18].

In our study, the integration of delta imaging features—derived from comparing pre- and post-LM CT scans—enabled precise quantification of subtle intratumoral and peritumoral changes over time, thereby capturing the temporal heterogeneity that was closely linked to metastatic progression. This dynamic representation proved to be highly informative, as evidenced by the delta-based multi-instance learning signature (MILDelta), which achieved an AUC of 0.871 in the independent test cohort. This performance was markedly superior to that of models trained solely on pre- or post-LM images (AUCs of 0.574 and 0.661, respectively), demonstrating that longitudinal dynamics deliver substantial incremental predictive value beyond static imaging. Notably, this advantage was consistent across the internal and external validation cohorts, further supporting the robustness of temporal feature integration in LM risk assessment. These findings highlight the innovative nature of our approach, positioning delta-enhanced 2.5D modeling as a promising direction for translating temporal radiomics into clinical decision support systems.

Prior MRI-based investigations have largely targeted LM detection on contrast-enhanced brain MRI, showing clinically useful discrimination while depending on specialized neuroimaging workflows and predominantly addressing overt LM rather than preemptive risk stratification [19, 20]. In parallel, radiomics studies built from primary tumor imaging (CT or MRI) have explored LM risk prediction and reported encouraging performance; however, many are single-center, cross-sectional analyses that use conventional classifiers and lack longitudinal feature modeling, multi-instance aggregation, external validation, or clinical utility assessment by decision-curve analysis [19, 21]. Compared with these efforts, our framework integrates delta-radiomics to capture temporal evolution, a 2.5D deep learning backbone with multi-instance learning for patient-level representation, external validation across cohorts, and explicit evaluation of net benefit. These design choices aim to improve robustness and facilitate clinical translation.

The multi-instance learning (MIL) strategy [22] played a pivotal role in our framework by enabling the systematic aggregation of slice-level feature representations into holistic patient-level embeddings [23]. This approach accommodates the inherent heterogeneity present across different tumor slices [24], ensuring that salient local features are not lost while the broader anatomical and pathological context is simultaneously captured. By doing so, the MIL mechanism enhances both model robustness and generalizability across diverse imaging conditions and patient populations, mitigating potential performance degradation due to slice-level variability.

When alternative machine learning classifiers were compared, XGBoost [25]

consistently outperformed the others in terms of sensitivity and overall area under the curve (AUC) across the training, internal validation, and external test datasets. A previous study also indicated the excellent predictive ability of XGBoost [26]. This observation reflects the algorithm's proven capacity to model complex, nonlinear relationships and feature interactions—a characteristic that is particularly well suited to the high-dimensional, heterogeneous nature of radiomics data derived from medical imaging.

Model interpretability was further enriched through Grad-CAM visualization, which revealed that predictive attention was predominantly concentrated within intratumoral regions and their immediate peritumoral microenvironments [27]. This spatial focus aligns closely with clinical observations that both the tumor core and surrounding tissue microstructure harbor prognostically relevant information on metastatic risk. Such visual evidence not only supports the biological plausibility of the model's decision-making process but also fosters clinician confidence, an essential factor for successful AI integration into routine oncology workflows.

Importantly, increasing the imaging-derived signature with selected clinical variables (such as the EGFR mutation status) further improved predictive performance, achieving an AUC of 0.910 in the independent test cohort. This synergy underscores the complementary value of multimodal input, wherein radiomic phenotypes provide quantitative descriptors of tumor morphology and dynamics, while clinical parameters contribute to the establishment of prognostic markers rooted in tumor biology. These findings are consistent with prior research [3], which suggests that integrating imaging and clinical domains enables more accurate and individualized risk assessment, ultimately supporting precision medicine initiatives in advanced lung adenocarcinoma.

Clinically, the proposed noninvasive predictive framework holds substantial potential to transform early detection strategies for LM in patients with lung adenocarcinoma. Current LM diagnostic practices rely primarily on invasive CSF cytology[28], which, despite being considered the diagnostic gold standard, often suffers from low sensitivity, frequent false negatives [29], and patient discomfort because of its invasive nature. Alternatively, contrast-enhanced brain or spinal MRI offers valuable anatomical visualization but remains costly, less accessible in resource-limited settings, and subject to variability in interpretation, particularly in early or subtle LM presentations[29].

By capitalizing on routinely acquired contrast-enhanced thoracic CT scans, which is already integrated into standard lung cancer follow-up protocols, our framework

proactively flags patients at high risk of LM prior to the occurrence of overt neurological symptoms or radiologically evident meningeal involvement. This capability provides a critical window for earlier clinical intervention, potentially including prophylactic intrathecal chemotherapy, systemic targeted therapies tailored to molecular profiles such as EGFR mutations, or intensified surveillance protocols[30]. Such preemptive measures could delay LM onset, mitigate symptom burden, and ultimately improve quality of life and survival outcomes.

Furthermore, the translation of model outputs into a clinically friendly nomogram enhances usability by providing an individualized [31], quantitative LM risk estimate that can be interpreted at the point of care without requiring deep technical expertise [32]. Decision curve analysis confirmed that applying this tool in clinical workflows would yield net clinical benefit across a wide range of threshold probabilities, underscoring its practical value for guiding intervention timing and intensity [33]. Collectively, these attributes position our approach not only as a research innovation but also as a viable candidate for integration into precision oncology pipelines, bridging the gap between advanced computational modeling and actionable patient-centered care.

Methodologically, coupling the “delta concept” with a 2.5D deep learning paradigm introduces a new analytic route for addressing spatiotemporal tumor heterogeneity in radiomics. This strategy balances 3D spatial context with computational efficiency, making it well suited for multitimepoint longitudinal imaging studies. The delta features reflect a shift from static cross-sectional imaging analyses to dynamic trajectory-based modeling, unlocking opportunities to detect subtle changes in tumor phenotype and behavior.

Limitations and Future Directions

This study has several limitations that should be carefully considered when interpreting the results.

First, although the retrospective design enables efficient utilization of existing data, it carries inherent risks of selection and information bias. Despite the application of propensity score matching and other statistical methods to balance baseline characteristics across cohorts, the causal inference strength achieved through randomization cannot be fully replicated. Potential unmeasured confounding factors not included in the analysis may still influence the findings.

Second, the overall sample size was relatively small, primarily because all of the included patients with lung adenocarcinoma with LM were confirmed through cerebrospinal fluid (CSF) cytology rather than being diagnosed solely via cranial

MRI. Patients diagnosed by cranial MRI may include false positives or cases that are actually brain metastases, and the number of clinical patients undergoing CSF puncture is limited, reflecting a real-world challenge.

Third, with respect to temporal imaging analysis, the time interval between pre-LM and post-LM CT scans lacked standardization. Disease progression of LM, changes during treatment breaks, and other clinical events could occur within this window. Such heterogeneity directly affects the consistency of delta-radiomics feature calculations, introducing variability that may reflect not only LM-related changes but also tumor natural history or short-term progression, thereby weakening their reliability as dynamic biomarkers.

Fourth, in terms of technical reproducibility, this study faced common challenges associated with multicenter imaging research. Although DICOM metadata were rigorously anonymized to protect patient privacy and ComBat harmonization was applied to mitigate batch effects, differences across participating centers in CT scanner models, acquisition protocols (e.g., slice thickness, tube voltage), and reconstruction parameters (e.g., convolution kernel) remained major sources of technical noise affecting radiomics feature stability. These variations limited the plug-and-play applicability of the model across different hardware and scanning settings.

Finally, in terms of model generalizability, further validation is warranted. While external validation was performed, the data were derived exclusively from three domestic medical centers with patient populations sharing relatively homogeneous genetic backgrounds, living environments, and clinical practice patterns. Accordingly, the generalizability of the findings to broader populations, such as other ethnic groups, different health care systems, or varied treatment standards, remains unconfirmed.

Future research should focus on conducting prospective, multicenter cohort studies with prestandardized imaging acquisition protocols and follow-up time points. Actively incorporating international and multiethnic independent validation cohorts will be crucial for testing model robustness and clinical translational potential. Moreover, exploring advanced feature engineering techniques or deep learning architectures that are less sensitive to scanning parameter variations represents an important pathway toward enhancing the reliability and real-world applicability of radiomics in clinical practice.

Conclusion

We developed a 2.5D delta-radiomics deep learning model with a multi-instance learning strategy to predict LM risk in patients with lung adenocarcinoma using routine contrast-enhanced CT. By capturing subtle longitudinal intratumoral and peritumoral changes, the model achieved superior performance over static imaging approaches and demonstrated robust generalizability. Grad-CAM visualization and a nomogram interface enhanced the interpretability, whereas decision curve analysis confirmed the tangible clinical benefit. This noninvasive, cost-effective tool shows promise for early LM risk stratification and timely intervention within precision oncology workflows.

Data availability

This manuscript does not report data generation or analysis.

Acknowledgments

The authors wish to thank all the patients who participated and the clinical teams at the three collaborating medical centers for their invaluable contributions to data collection and image acquisition. We are grateful to the radiologists who performed and verified the tumor segmentation, as well as to the engineers and data scientists who supported the preprocessing and model development workflows. This work was supported by the Central Guidance for Local Scientific and Technological Development Funds (grant nos. 2024ZY01018 to Weidong Han) and the grants from the National Natural Science Foundation of China (Grant No. 62176230 to Hanliang Jiang).

Author Contributions Statement

W.H. conceived the study. J.S.(Jiachun Sun) and R.Y. collected clinical data and wrote the main manuscript text; J.S.(Jingxiang Su), B.S. and Y.X. also participated in the clinical data collection and data analysis; R.Y. and Y. F. prepared figures and tables. H.J. and W.H. provided funds and L.Z. provided technical support. All authors reviewed the manuscript.

Competing Interests Statement

The authors declare that they have no competing financial or non-financial interests related to this work. No conflicts exist regarding employment, consultancies, stock ownership, honoraria, paid expert testimony, patent applications, or any other potential sources of bias. All authors have read and approved the final manuscript and agree with its submission to International Journal of Surgery.

References

1. Siegel, R.L., A.N. Giaquinto, and A. Jemal, *Cancer statistics, 2024*. CA Cancer J Clin, 2024. **74**(1): p. 12-49.
2. Remon, J., E. Le Rhun, and B. Besse, *Leptomeningeal carcinomatosis in non-small cell lung cancer patients: A continuing challenge in the personalized treatment era*. Cancer Treat Rev, 2017. **53**: p. 128-137.
3. Ozcan, G., M. Singh, and J.J. Vredenburg, *Leptomeningeal Metastasis from Non-Small Cell Lung Cancer and Current Landscape of Treatments*. Clin Cancer Res, 2023. **29**(1): p. 11-29.
4. Gao, X., et al., *Leptomeningeal enhancement in magnetic resonance imaging predicts poor prognosis in lung adenocarcinoma patients with leptomeningeal metastasis*. Thorac Cancer, 2022. **13**(7): p. 1059-1066.
5. Yang, J.T., et al., *Randomized Phase II Trial of Proton Craniospinal Irradiation Versus Photon Involved-Field Radiotherapy for Patients With Solid Tumor Leptomeningeal Metastasis*. J Clin Oncol, 2022. **40**(33): p. 3858-3867.
6. Hong, J., et al., *Advances in the research of leptomeningeal metastases in non-small cell lung cancer: a narrative review*. Transl Lung Cancer Res, 2025. **14**(8): p. 3216-3232.
7. Huijs, J.W.J., et al., *Overall survival in patients with non-small cell lung cancer and leptomeningeal metastases: Data from the Dutch cancer Registry*. Lung Cancer, 2025. **209**: p. 108806.
8. (NCCN)., N.C.C.N., *Clinical Practice Guidelines in Oncology: Non-Small Cell Lung Cancer*. 2026. **Version 1**.
9. Cheng, H. and R. Perez-Soler, *Leptomeningeal metastases in non-small-cell lung cancer*. Lancet Oncol, 2018. **19**(1): p. e43-e55.
10. Lambin, P., et al., *Radiomics: extracting more information from medical images using advanced feature analysis*. Eur J Cancer, 2012. **48**(4): p. 441-6.
11. Thawani, R., et al., *Radiomics and radiogenomics in lung cancer: A review for the clinician*. Lung Cancer, 2018. **115**: p. 34-41.
12. Çiçek, Ö., et al. *3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation*. in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*. 2016. Cham: Springer International Publishing.
13. Wang, W., et al., *Predicting Radiation Pneumonitis Integrating Clinical Information, Medical Text, and 2.5D Deep Learning Features in Lung Cancer*. Int J Radiat Oncol Biol Phys, 2025.
14. Nardone, V., et al., *Delta radiomics: an updated systematic review*. Radiol Med, 2024. **129**(8): p. 1197-1214.
15. Han, X., et al., *Delta-radiomics features for predicting the major pathological response to neoadjuvant chemoimmunotherapy in non-small cell lung cancer*. Eur Radiol, 2024. **34**(4): p. 2716-2726.
16. Ram, U.S., et al., *Assessing Quantitative Performance and Expert Review of Multiple Deep Learning-Based Frameworks for Computed Tomography-based Abdominal Organ Auto-Segmentation*. Intell Oncol, 2025. **1**(2): p. 160-171.
17. Seong, M., et al., *Diagnostic accuracy of MR imaging of patients with*

- leptomeningeal seeding from lung adenocarcinoma based on 2017 RANO proposal: added value of contrast-enhanced 2D axial T2 FLAIR.* J Neurooncol, 2020. **149**(2): p. 367-372.
18. Han, Z., et al., *Delta-radiomics models based on multi-phase contrast-enhanced magnetic resonance imaging can preoperatively predict glypican-3-positive hepatocellular carcinoma.* Front Physiol, 2023. **14**: p. 1138239.
 19. Bashir, S., et al., *Development and validation of a predictive nomogram for leptomeningeal metastasis risk in NSCLC brain metastases: role of tumor location, driver mutations, and stereotactic radiosurgery.* J Neurooncol, 2025. **175**(3): p. 1377-1390.
 20. Jin, K., et al., *A survival prediction model for leptomeningeal metastasis patients with non-small cell lung cancer based on deep learning.* BMC Cancer, 2025. **26**(1): p. 175.
 21. Nakagawa, K., et al., *Prognostic impact of clinical and radiological factors on leptomeningeal metastasis from solid cancers.* J Neurooncol, 2024. **167**(3): p. 397-406.
 22. Quellec, G., et al., *Multiple-Instance Learning for Medical Image and Video Analysis.* IEEE Rev Biomed Eng, 2017. **10**: p. 213-234.
 23. Zhao, W., et al., *Lung Cancer Screening Classification by Sequential Multi-Instance Learning (SMILE) Framework With Multiple CT Scans.* IEEE Trans Med Imaging, 2025. **44**(8): p. 3151-3161.
 24. Campanella, G., et al., *Clinical-grade computational pathology using weakly supervised deep learning on whole slide images.* Nat Med, 2019. **25**(8): p. 1301-1309.
 25. Tianqi Chen, C.G., *XGBoost: A scalable tree boosting system.* Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining., 2016: p. 785-794.
 26. Guan, X., et al., *Construction of the XGBoost model for early lung cancer prediction based on metabolic indices.* BMC Med Inform Decis Mak, 2023. **23**(1): p. 107.
 27. Vanitha, K., et al., *Deep learning ensemble approach with explainable AI for lung and colon cancer classification using advanced hyperparameter tuning.* BMC Med Inform Decis Mak, 2024. **24**(1): p. 222.
 28. Wang, X., et al., *CEACAM6 serves as a biomarker for leptomeningeal metastasis in lung adenocarcinoma.* Cancer Med, 2023. **12**(4): p. 4521-4529.
 29. Hyun, J.W., et al., *Leptomeningeal metastasis: Clinical experience of 519 cases.* Eur J Cancer, 2016. **56**: p. 107-114.
 30. Wang, Y., et al., *Leptomeningeal metastases in non-small cell lung cancer: Diagnosis and treatment.* Lung Cancer, 2022. **174**: p. 1-13.
 31. Iasonos, A., et al., *How to build and interpret a nomogram for cancer prognosis.* J Clin Oncol, 2008. **26**(8): p. 1364-70.
 32. Topol, E.J., *High-performance medicine: the convergence of human and artificial intelligence.* Nat Med, 2019. **25**(1): p. 44-56.
 33. Vickers, A.J. and E.B. Elkin, *Decision curve analysis: a novel method for*

evaluating prediction models. Med Decis Making, 2006. **26**(6): p. 565-74.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [supplementary.docx](#)