npj | precision oncology

Article

Published in partnership with The Hormel Institute, University of Minnesota

https://doi.org/10.1038/s41698-025-00927-4

Biopsy image-based deep learning for predicting pathologic response to neoadjuvant chemotherapy in patients with NSCLC

Yibo Zhang[1,6], Shuaibo Wang[2,3,6], Xinying Liu[4,6], Yang Qu[5,6], Zijian Yang[1], Yang Su[1], Bin Hu[2], Yousheng Mao[3] ✉, Dongmei Lin[4] ✉, Lin Yang[5] ✉ & Meng Zhou[1] ✉

Neoadjuvant chemotherapy (NAC) is a widely used therapeutic strategy for patients with resectable non-small cell lung cancer (NSCLC). However, individual responses to NAC vary widely among patients, limiting its effective clinical application. In this study, we propose a weakly supervised deep learning model, DeepDrRVT, which integrates self-supervised feature extraction and attention-based deep multiple instance learning, to improve NAC decision making from pretreatment biopsy images. DeepDrRVT demonstrated superior predictive performance and generalizability, achieving AUCs of 0.954, 0.872 and 0.848 for complete pathologic response, and 0.968, 0.893 and 0.831 for major pathologic response in the training, internal validation and external validation cohorts, respectively. The DeepDrRVT digital assessment of residual viable tumor correlated significantly with the local pathologists' visual assessment (Pearson r = 0.98, 0.80, and 0.59; digital/visual slope = 1.0, 0.8 and 0.55) and was also associated with longer disease-free survival (DFS) in all cohorts (HR = 0.455, 95% CI 0.234–0.887, $P = 0.018$; HR = 0.347, 95% CI 0.135–0.892, $P = 0.021$ and HR = 0.446, 95% CI 0.193–1.027, $P = 0.051$). Furthermore, DeepDrRVT remained an independent prognostic factor for DFS after adjustment for clinicopathologic variables (HR = 0.456, 95% CI 0.227–0.914, $P = 0.027$; HR = 0.358, 95% CI 0.135–0.949, $P = 0.039$ and HR = 0.419, 95% CI 0.181–0.974, $P = 0.043$). Thus, DeepDrRVT holds promise as an accessible and reliable tool for clinicians to make more informed treatment decisions prior to the initiation of NAC.

Non-small cell lung cancer (NSCLC) accounts for approximately 85% of all lung cancers and remains a leading cause of cancer-related mortality worldwide[1]. Neoadjuvant treatments, including neoadjuvant chemotherapy (NAC), chemoimmunotherapy, and targeted therapy, are important therapeutic options for patients with resectable disease, as recommended by the National Comprehensive Cancer Network (NCCN) guidelines[2]. Patients who achieve a major pathologic response (MPR) or complete pathologic response (CPR), regardless of the neoadjuvant regimen, tend to have better long-term survival outcomes. However, response rates to NAC are highly variable due to tumor heterogeneity, with only a subset of patients achieving a significant pathologic response[3]. Pre-treatment response assessment is essential for early identification of treatment-sensitive groups, thereby

[1]Institute of Genomic Medicine, School of Biomedical Engineering, Wenzhou Medical University, Wenzhou, 325027, P. R. China. [2]Department of Thoracic Surgery, Beijing Institute of Respiratory Medicine and Beijing Chao-Yang Hospital, Capital Medical University, Beijing, 100020, P. R. China. [3]Department of Thoracic Surgery, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, 100021, P. R. China. [4]Department of Pathology, Peking University Cancer Hospital & Institute, Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education), Beijing, 100142, P. R. China. [5]Department of Pathology, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, 100021, P. R. China. [6]These authors contributed equally: Yibo Zhang, Shuaibo Wang, Xinying Liu, Yang Qu.
✉e-mail: youshengmao@gmail.com; lindm3@163.com; yanglin@cicams.ac.cn; zhoumeng@wmu.edu.cn

maximizing clinical benefit with minimal drug exposure, and enables the exploration of different strategies for treatment-resistant patients. Unlike neoadjuvant targeted therapies and immunotherapy, NAC currently lacks effective pre-treatment histologic biomarkers, which limits its precise clinical application[3]. Therefore, there is a significant clinical challenge and an urgent need for reliable pretreatment assessment of pathologic response to optimize individualized treatment strategies and improve patient outcomes.

Fiberoptic bronchoscopy and CT-guided biopsy are widely used standard diagnostic procedures to obtain accessible tissue for pathologic diagnosis and molecular subtyping of lung cancer[2,4]. Pre-treatment biopsy slides provide additional insights into the tumor microenvironment and offer potential predictive markers for response to NAC. Recent advances in artificial intelligence (AI), particularly in deep learning, have begun transforming the landscape of medical imaging and pathology[5]. Using digital histopathology, AI-driven models have shown promise in predicting treatment outcomes by analyzing histological and spatial features within the tumor microenvironment[6–10]. However, the application of deep learning to pretreatment biopsy specimens to predict NAC outcomes alone without immunotherapy in NSCLC has been less extensively explored.

In this multicenter, retrospective study, we aimed to develop and validate a deep learning model, DeepDrRVT, to predict the pathologic response to NAC in patients with NSCLC using haematoxylin and eosin (H&E)-stained whole slide images (WSIs) from pretreatment biopsies. In addition, we examined the association between visual and digital assessments of the percentage of residual viable tumor, and evaluated the prognostic value of DeepDrRVT. Finally, we visualized the model outputs and analyzed the underlying histopathologic features driving its predictions to assess the reliability and interpretability of DeepDrRVT.

Results

Study design and patient characteristics

The overall study design is shown in Fig. 1. We conducted a discovery and validation study to develop and independently validate a deep learning model, DeepDrRVT, for the pretreatment prediction of pathologic response to NAC in patients with NSCLC using biopsy slides. A total of 385 patients were screened for eligibility, and 331 patients were enrolled in this study. The enrolled patients were divided into three cohorts: the training cohort ($n = 113$), the internal test cohort ($n = 76$), and the external validation cohort ($n = 142$). Baseline characteristics are summarized in Table 1. The median age of the entire cohort was 59 years (range: 34–75). The cohort was predominantly male, with 79.46% males and 20.54% females. In addition, 76.74% of the patients were smokers and 23.26% were non-smokers. Of the 331 participants, 59 (17.82%) achieved MPR and 22 (6.65%) achieved CPR after NAC. There were no statistically significant differences in baseline characteristics between patients who achieved MPR and those who did not.

Deep learning model for pathologic response prediction

We developed and trained the DeepDrRVT model for the digital assessment of pathologic response to NAC using pretreatment biopsy WSIs by separating the patients from the PUCH into a training cohort (60%, $n = 113$) and an internal testing cohort (40%, $n = 76$) (Fig. 2). The DeepDrRVT demonstrated strong predictive performance, achieving AUCs of 0.954 (95% CI 0.912–0.995) for CPR and 0.968 (95% CI 0.939–0.998) for MPR in the PUCH training cohort (Fig. 3A and Supplementary Table 1). In the PUCH test cohort, the model achieved AUCs of 0.872 (95% CI: 0.764–0.981) for CPR and 0.893 (95% CI: 0.799–0.987) for MPR (Fig. 3B and Supplementary Table 1). To assess the generalizability of DeepDrRVT, we validated the model in an independent external cohort from CHCAMS. DeepDrRVT consistently demonstrated robust prediction performance, achieving AUCs of 0.848 (95% CI 0.673–1.000) for CPR and 0.831 (95% CI 0.700–0.962) for MPR in the CHCAMS cohort (Fig. 3C and Supplementary Table 1).

Using the Youden index, digital cutoff values were determined from the ROC curves in the PUCH-train cohort. Patients were stratified into Dig.MPR+ or Dig.MPR- groups (cutoff of 0.1935), and Dig.CPR+ or Dig.CPR- groups (cutoff of 0.1005) based on the digital %RVT predicted by DeepDrRVT. DeepDrRVT demonstrated strong discriminatory ability, accurately distinguishing MPR+ from MPR- patients: 25/25 (100%) of MPR patients in the PUCH-train cohort, 11/15 (73.33%) in the PUCH-test cohort, and 13/19 (68.42%) in the CHCAMS cohort were correctly identified. In addition, 78/88 (88.64%) of non-MPR patients in the PUCH-train cohort, 51/61 (83.61%) in the PUCH-test cohort, and 119/123 (96.75%) in the CHCAMS cohort were accurately classified by DeepDrRVT (Supplementary Fig. 1).

Furthermore, the digital %RVT derived from DeepDrRVT showed a strong positive correlation with the visual %RVT assessed by specialized pathologists, with Spearman's correlation coefficients of 0.97, 0.81, and 0.58 in the PUCH-train, PUCH-test and CHCAMS cohorts, respectively (Fig. 3D). The predicted probability distributions effectively distinguished between MPR+ and MPR- cases and CPR+ and CPR- cases (Supplementary Fig. 2). DCA analysis revealed that the DeepDrRVT provided a greater net benefit in identifying responders (MPR/CPR) and non-responders across a wide range of threshold probabilities (Supplementary Fig. 3). The calibration curve of the DeepDrRVT for %RVT demonstrated good agreement between prediction and observation in all cohorts (Supplementary Fig. 4).

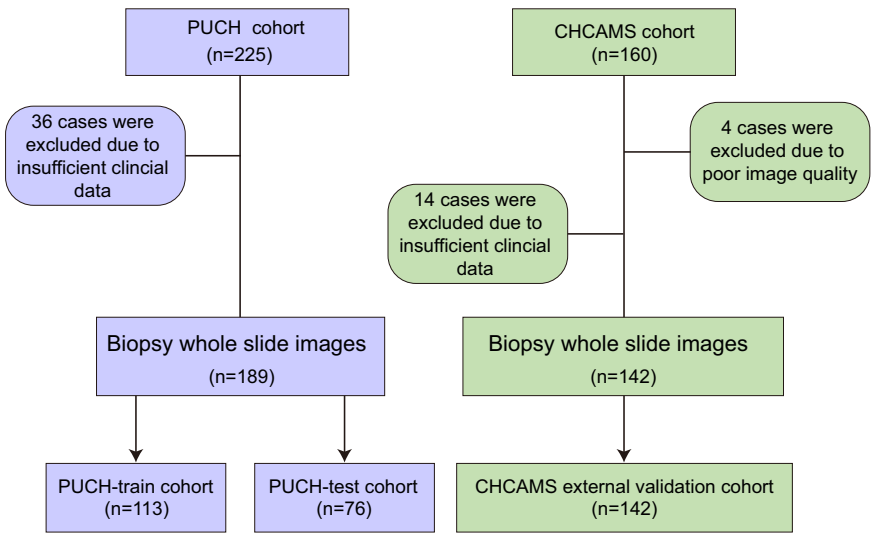

Fig. 1 | Flowchart of the patient enrollment process and cohort division. 331 patients from two medical centers (PUCH and CHCAMS) were enrolled in this study. The patients were divided into three cohorts: the training cohort ($n = 113$), the internal testing cohort ($n = 76$), and the external validation cohort ($n = 142$). PUCH, Peking University Cancer Hospital; CHCAMS, Cancer Hospital Chinese Academy of Medical Sciences.

Table 1 | Baseline demographic and clinical characteristics of the cohorts in this study

Characteristics	All (n = 331)	PUCH-train cohort		p value	PUCH-test cohort		p value	CHCAMS cohort		p value
		(MPR, n = 25)	(non-MPR, n = 88)		(MPR, n =15)	(non-MPR, n = 61)		(MPR, n = 19)	(non-MPR, n = 123)	
[a] Age, median (range)	59 (34–75)	60 (39–73)	58.5 (34–75)	0.4675	60 (45–75)	61 (46–75)	0.393	58 (43–72)	58 (37–73)	0.6989
[b] Gender, n (%)				0.0601			0.3276			0.1953
Male	263 (79.46)	23 (92.00)	65 (73.86)		13 (86.67)	43 (70.49)		14 (73.68)	105 (85.37)	
Female	68 (20.54)	2 (8.00)	23 (26.14)		2 (13.33)	18 (29.51)		5 (26.32)	18 (14.63)	
[b] Smoking, n (%)				0.0582			0.3276			1
Yes	254 (76.74)	23 (92.00)	64 (72.73)		13 (86.67)	43 (70.49)		15 (78.95)	96 (78.05)	
No	77 (23.26)	2 (8.00)	24 (27.27)		2 (13.33)	18 (29.51)		4 (21.05)	27 (21.95)	
[b] Stage, n (%)				0.2722			0.4783			1
I	12 (3.63)	0 (0.00)	0 (0.00)		0 (0.00)	0 (0.00)		1 (5.26)	11 (8.94)	
II	66 (19.94)	3 (12.00)	21 (23.86)		4 (26.67)	11 (18.03)		4 (21.05)	23 (18.70)	
III	253 (76.43)	22 (88.00)	67 (76.14)		11 (73.33)	50 (81.97)		14 (73.68)	89 (72.36)	
[b] Clinical T, n (%)				0.769			0.2409			0.1663
T1	54 (16.31)	5 (20.00)	14 (15.91)		0 (0.00)	12 (19.67)		1 (5.26)	22 (17.89)	
T2	152 (45.92)	11 (44.00)	41 (46.59)		6 (40.00)	23 (37.70)		10 (52.63)	61 (49.59)	
T3	92 (27.80)	5 (20.00)	23 (26.14)		7 (46.67)	20 (32.79)		8 (42.11)	29 (23.58)	
T4	33 (9.97)	4 (16.00)	10 (11.36)		2 (13.33)	6 (9.84)		0 (0.00)	11 (8.94)	
[b] Clinical N, n (%)				0.5804			1			0.492
N0	52 (15.71)	3 (12.00)	5 (5.68)		1 (6.67)	7 (11.48)		5 (26.32)	31 (25.20)	
N1	72 (21.75)	8 (32.00)	30 (34.09)		4 (26.67)	18 (29.51)		0 (0.00)	12 (9.76)	
N2	207 (62.54)	14 (56.00)	53 (60.23)		10 (66.67)	36 (59.02)		14 (73.68)	80 (65.04)	
[b] Histological type, n (%)				0.002			0.040			0.618
ADC	131 (39.58)	4 (16.00)	45 (51.14)		2 (13.33)	26 (42.62)		6 (31.58)	48 (39.02)	
SCC	200 (60.42)	21 (84.00)	43 (48.86)		13 (86.67)	35 (57.38)		13 (68.42)	75 (60.98)	
[b] NAC regimens, n (%)				0.017			0.136			0.516
Platin + Gemcitabine	88 (26.59)	13 (52.00)	25 (28.41)		4 (26.67)	20 (32.79)		4 (21.05)	22 (17.89)	
Platin + Pemetrexed	100 (30.21)	2 (8.00)	29 (32.95)		2 (13.33)	21 (34.42)		4 (21.05)	42 (34.15)	
Platin + taxel	143 (43.20)	10 (40.00)	34 (38.64)		9 (60.00)	20 (32.79)		11 (57.90)	59 (47.97)	
[b] NAC cycles, n (%)				1.000			0.478			0.430
≤2	249 (75.23)	21 (84.00)	71 (80.68)		11 (73.33)	50 (81.97)		11 (57.90)	85 (69.11)	
>2	82 (24.77)	4 (16.00)	17 (19.32)		4 (26.67)	11 (18.03)		8 (42.10)	38 (30.89)	

[a]Fisher's exact test for categorical variables. [b]T-test for continuous variables. *MPR* major pathological response. *non-MPR* non-major pathological response. *NAC* neoadjuvant chemotherapy.

DeepDrRVT-inferred digital %RVT associated with disease-free survival

The prognostic value of the DeepDrRVT in evaluating the clinical efficacy of NAC was assessed through disease-free survival (DFS) analysis. DeepDrRVT effectively stratified patients into high-risk and low-risk groups for DFS. The Dig.MPR+ group had significantly longer DFS compared to the Dig.MPR- group (log-rank $p = 0.018$ in the PUCH-train cohort; $p = 0.021$ in the PUCH-test cohort; and $p = 0.051$ in the CHCAMS cohort) (Fig. 4A). Comparative analysis showed a similar ability in the predictive performance of DFS between the Vis.MPR and the Dig.MPR (Supplementary Fig. 5). Multivariate Cox regression analysis showed that DeepDrRVT was significantly associated with DFS after adjusting for other clinical factors (HR = 0.456, 95% CI: 0.227–0.914, $p = 0.027$ for the PUCH-train cohort; HR = 0.358, 95% CI: 0.135–0.949, $p = 0.039$ for the PUCH-test cohort; and HR = 0.419, 95% CI: 0.181–0.974, $p = 0.043$ for the CHCAMS cohort), confirming that DeepDrRVT is an independent prognostic factor for DFS. Subgroup analyses further validated the robust prognostic stratification and MPR prediction ability of DeepDrRVT for DFS across histologic subtypes (HR = 0.488, 95% CI: 0.291–0.817, $p = 0.005$ and AUC = 0.914 for the SCC; HR = 0.310, 95% CI: 0.113–0.852, $p = 0.016$ and AUC = 0.879 for the ADC) and treatment duration (HR = 0.372, 95% CI: 0.209–0.664, $p < 0.001$ and AUC = 0.906 for the treatment cycles ≤2; HR = 0.557, 95% CI: 0.271–1.141, $p = 0.11$ and AUC = 0.920 for the treatment cycles >2) (Supplementary Fig. 6).

Interpretation of DeepDrRVT

We generated attention heatmaps using the attMIL architecture within the trained DeepDrRVT model to identify tiles with different attention scores. As shown in Fig. 5A, patients with higher predicted %RVT scores tended to show increased model attention in tumor regions with higher tumor purity

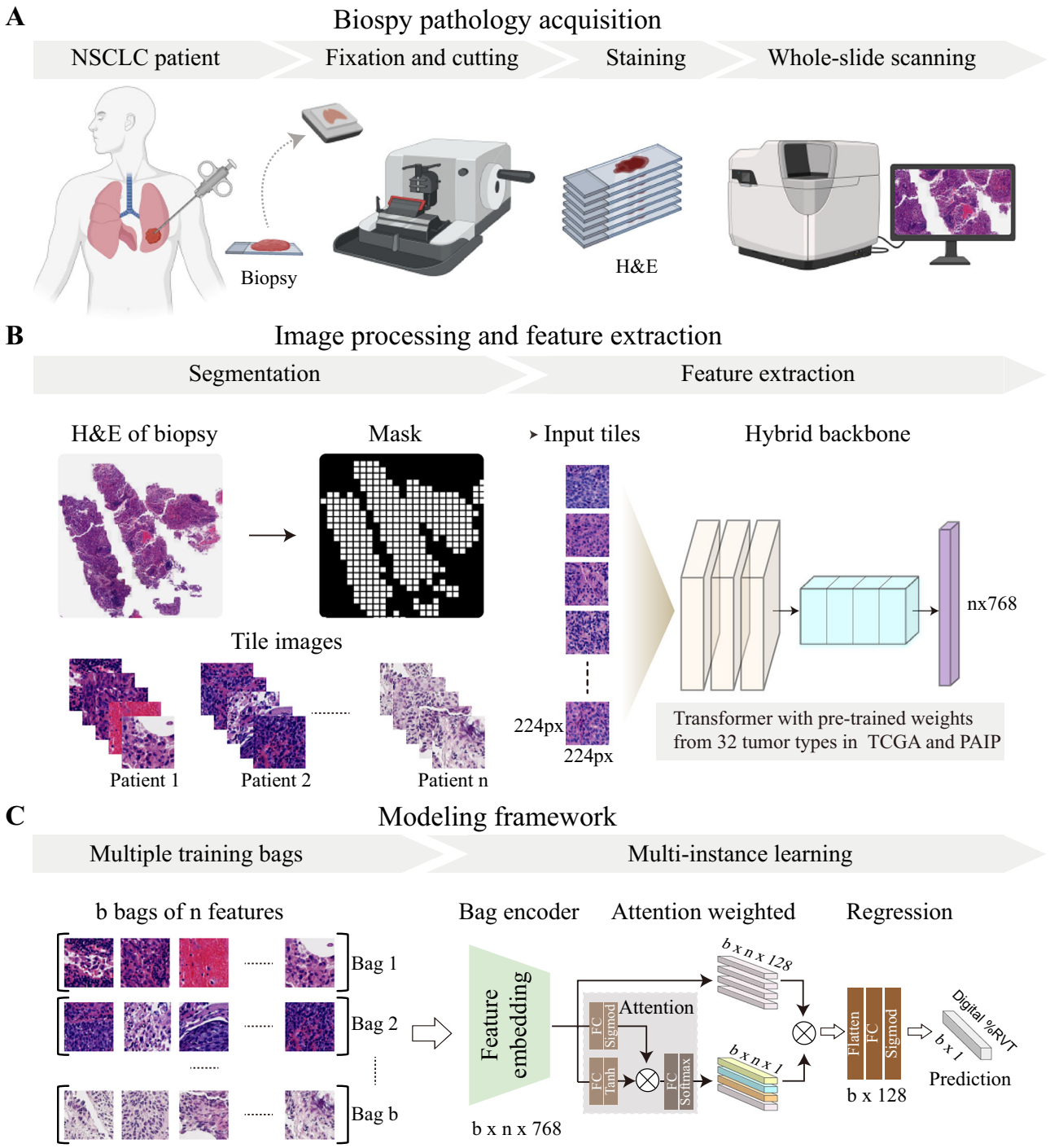

Fig. 2 | Overview of the DeepDrRVT framework and architecture. A Biopsy pathology acquisition: NSCLC biopsy specimens undergo fixation, sectioning, staining, and whole-slide scanning to generate digital H&E images (Created with BioRender.com). **B** Image preprocessing and feature extraction: The preprocessing pipeline involves segmentation of H&E-stained biopsy images to generate input tiles. Feature extraction is performed using a hybrid backbone model, CtransPath, with pre-trained weights from 32 tumor types based on the TCGA and PAIP datasets, which converts tile images into feature embeddings. **C** Modeling framework: Each patient's biopsy image is treated as a "bag" of multiple "tiles". Modeling uses attention-based multiple instance learning (attMIL) to aggregate tile-level features for patient-level %RVT prediction. The bag encoder processes the features, followed by attention-weighted feature selection and regression analysis.

(tumor/stroma ratio), nuclear atypia and nuclear pleomorphism, while patients with lower %RVT scores were associated with a focus on tumor regions with lower tumor purity (Fig. 5A and Supplementary Fig. 7). We also compared the histopathologic features of the post-treatment specimens between patients in the Dig.MPR+ and Dig.MPR- groups as predicted by DeepDrRVT. The results showed that, compared to the Dig.MPR- group, the Dig.MPR+ group had significantly higher proportions of multi-nucleated giant cells (17.32% vs. 41.18%), foamy macrophages (30.71% vs. 58.82%) and cholesterol clefts (25.20% vs. 47.06%) in the tumor stroma, but significantly lower proportions of necrosis (54.33% vs. 29.41%), plasmacytes (62.99% vs. 41.18%), neutrophils (20.47% vs. 0.00%), and peritumoral lymphocytes (78.74% vs. 58.82%) (Fig. 5B).

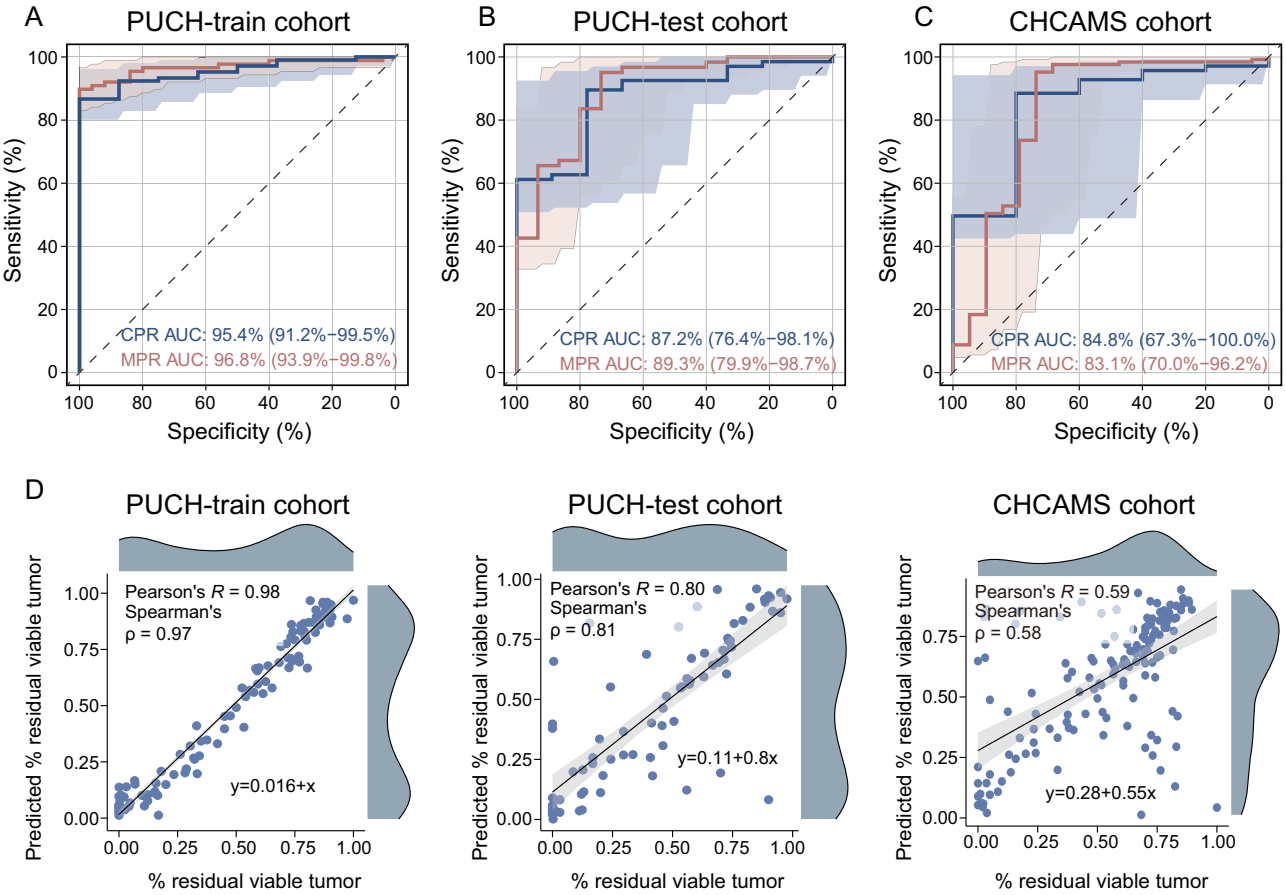

Fig. 3 | Performance of DeepDrRVT for predicting pathologic response in different cohorts. ROC curves showing the predictive performance of DeepDrRVT for MPR and CPR prediction in the PUCH-train cohort (**A**), PUCH-test cohort (**B**), and CHCAMS cohort (**C**). **D** Correlation between digital assessment of the percentage of viable tumor by DeepDrRVT and the visual assessment by the local pathologists in three cohorts. Scatter plots with linear regression lines show Pearson's correlation coefficients (R) and Spearman's rank correlation coefficient (ρ). MPR, major pathologic response; CPR, complete pathologic response; ROC, receiver operating characteristic; ROC, receiver operating characteristic curve.

Discussion

In this study, we developed and validated a deep learning model, DeepDrRVT, which utilizes biopsy images and an attention-based multiple instance learning framework to predict pathologic response to NAC in patients with NSCLC. Our results demonstrate that DeepDrRVT accurately predicts MPR and CPR, providing clinicians with a valuable tool for informed treatment decision-making prior to NAC. Studies such as KEYNOTE-671 and NeoR-world have shown that patients with MPR from immunochemotherapy or chemotherapy have comparable DFS, but patients without MPR benefit more from immunochemotherapy[11,12]. By probabilistically stratifying the MPR likelihood, DeepDrRVT enables a risk-adapted therapeutic framework in which patients predicted to have MPR may benefit from chemotherapy de-escalation to minimize immune-related adverse events, while non-MPR patients may require early immunotherapy intensification or combinatorial regimens. To mitigate the risk of undertreatment (false positives) or overtreatment (false negatives), we propose an adaptive trial design that incorporates pretreatment prediction with radiographic re-evaluation after two cycles.

Previous research into predicting treatment response in NSCLC has focused on radiomics and molecular omics. Radiomics extracts quantitative features from imaging scans (e.g., CT, MRI), whereas molecular omics uses genomic, transcriptomic, or proteomic data to identify biomarkers associated with treatment outcomes[13–17]. However, radiomics relies on imaging acquisition standardization, variations in analysis that may introduce bias and reduce generalizability across different patient cohorts[18]. In addition, many molecular markers identified by omics studies have yet to be validated in multicenter cohorts[19]. In contrast, biopsy specimens are routinely available.

While deep learning models are highly effective in pattern recognition, they often function as "black boxes". However, the attention-based deep learning model in DeepDrRVT localizes relevant regions in biopsy slides that are most predictive of treatment response, revealing insights that might be overlooked by traditional methods. Our analysis of prediction-related regions revealed that DeepDrRVT focused primarily on tumor/stroma ratio, nuclear atypia and nuclear pleomorphism. These results suggest that the model effectively captures epithelial, stromal, immune, and spatial interaction features that not typically assessed in routine practice, supporting the use of biopsy specimens for digital pathology assessment. To ensure analytical validity, pathologists excluded sections with technical artifacts (e.g., tissue folding, scanner-induced color drift) that could confound computational analysis. While this quality control enhanced model reliability, it also reflects a key technical limitation: the exclusion of real-world variability in slide preparation may reduce generalizability to settings with differing quality thresholds. Furthermore, the observed differences in histopathologic features between the Dig.MPR+ and Dig.MPR- groups suggest that DeepDrRVT not only predicts %RVT but also reflects the underlying tumor microenvironment, which is driven by variations in immune cell infiltration, stromal composition, and necrosis—factors known to correlate with long-term survival[20]. The model achieved reliable performance in binary MPR classification but showed limited accuracy in predicting %RVT during external validation. False-positive diagnoses arose when the model focused predominantly on lymphocytes and degenerated

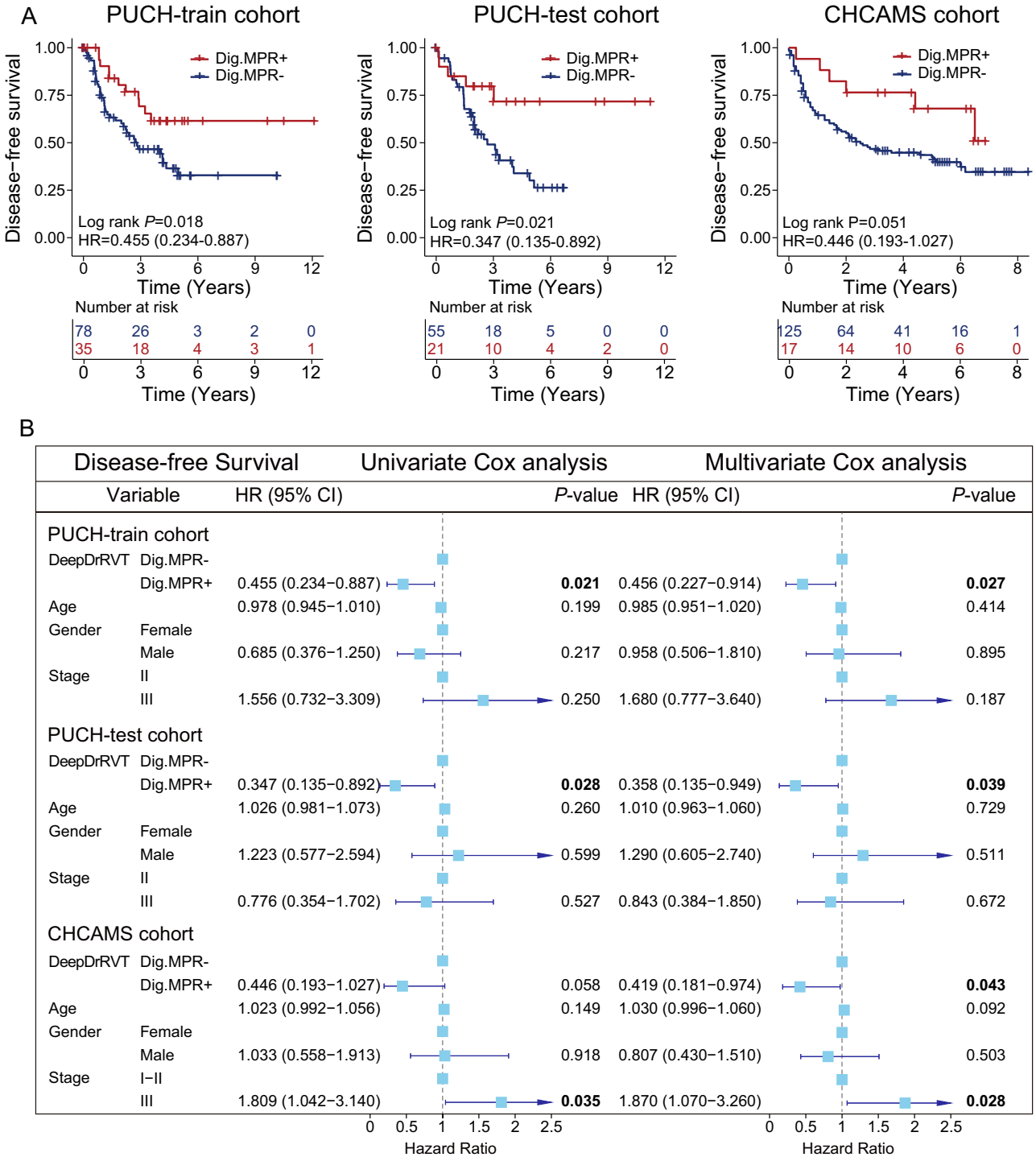

Fig. 4 | Prognostic evaluation of DeepDrRVT in different cohorts. A Kaplan-Meier curves showing disease-free survival for patients who received neoadjuvant therapy, categorized into Dig.MPR+ and Dig.MPR- groups in the PUCH-train, PUCH-test, and CHCAMS cohorts. **B** Univariate and multivariate Cox regression analyses for disease-free survival in the PUCH-train, PUCH-test, and CHCAMS cohorts. Dig.MPR+ , digital MPR positive; Dig.MPR-, digital MPR negative; HR, hazard ratios.

stromal fibers in specimens containing viable tumor with degenerative cellular changes, while false-negative errors were associated with staining variability impairing nuclear atypia assessment (Supplementary Fig. 8). These limitations likely stem from two factors: technical challenges in stain standardization across institutions, and biological complexity introduced by heterogeneous tumor regression patterns. Additionally, faded/low-quality histological slides may compromise nuclear feature clarity, highlighting the need for standardized image preprocessing.

Despite the promising results, this study has several limitations. First, our study was retrospective in nature, which introduces potential biases. Future studies should focus on the prospective validation of DeepDrRVT in larger, multicenter cohorts to confirm its generalizability across different patient populations and clinical settings. Second, although biopsy specimens provide a convenient and minimally invasive option for obtaining tumor tissue, they may not fully capture the spatial heterogeneity of the tumor microenvironment. Future models could integrate molecular data

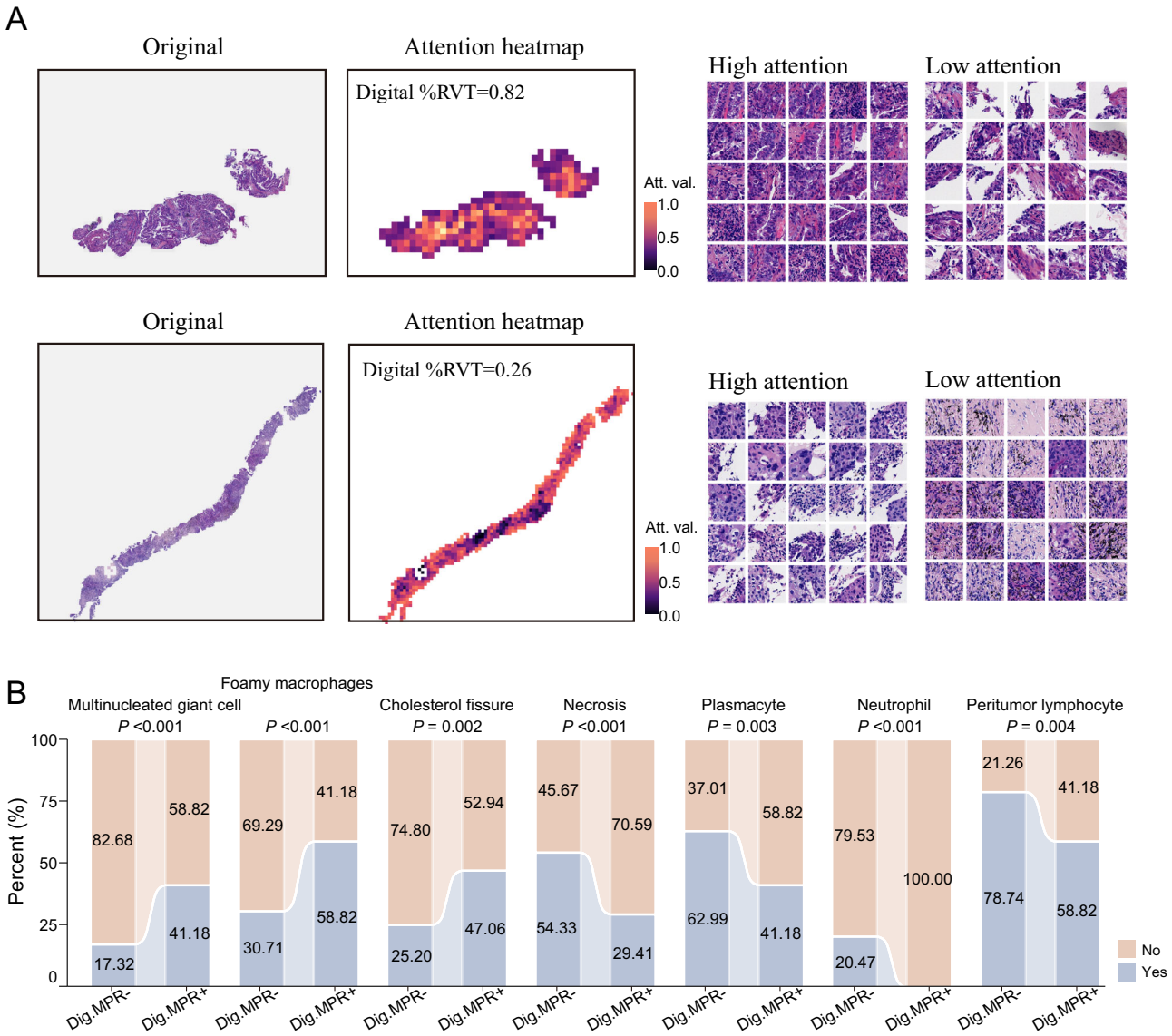

Fig. 5 | Visualization of architectural features and histopathologic features in biopsy specimens predicted by DeepDrRVT. A Visualization of architectural features for biopsy specimens from two representative patients generated using attMIL within the DeepDrRVT. **B** Comparative analysis of histopathologic features in on-treatment biopsy specimens between patients in the Dig.MPR+ and Dig.MPR- groups, as predicted by DeepDrRVT. attMIL, attention-based deep multiple instance learning; Dig.MPR +, digital MPR positive; Dig.MPR-, digital MPR negative.

from biopsies or radiomic features to improve model robustness and predictive power. Third, ICI plus platinum-based therapy has now become the standard care for NAC, and therefore DeepDrRVT, which is based on solely on H&E-stained slides, may be inadequate for predicting response to immunotherapy.

In conclusion, DeepDrRVT demonstrates the potential to predict pathologic response to NAC in NSCLC patients using biopsy-derived WSIs. This deep learning model provides a reliable, minimally invasive approach for pre-treatment assessment of MPR and CPR, enabling clinicians to make personalized and informed treatment decisions.

Methods

Ethics approval

Ethics approvals were obtained from the Institutional Review Boards of all participating centers (23/394-4137 and 2018YJZ19). Informed consent was waived due to the retrospective nature of this study. All data were analyzed anonymously.

Patients and data collection

We retrospectively reviewed clinicopathologic and pathologic data of 385 patients with primary stage IB-III NSCLC who received radical lung cancer surgery after neoadjuvant chemotherapy at the National Cancer Center/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College (CHCAMS) and Peking University Cancer Hospital (PUCH), between January 2009 and December 2020. Patients were eligible for inclusion if they had pre-treatment biopsy H&E sections available for pathological diagnosis at their respective centers, with all pathological tissue samples used in the analysis obtained from primary tumor sites. Exclusion criteria included: (i) presence of distant metastatic lesions; (ii) only exploratory surgery, biopsy, or wedge resection; (iii) perioperative death within 30 days after surgery; and, (iv) insufficient clinical data and poor image quality. The detailed enrollment process and cohort division are illustrated in Fig. 1. Finally, 331 patients were enrolled from the two medical centers.

Neoadjuvant chemotherapy regimens and response assessment

All patients received at least one cycle of NAC, with most receiving between two to four cycles. Chemotherapy regimens were based on a platinum doublet: platinum plus paclitaxel for squamous histology and platinum plus pemetrexed for non-squamous histology. After NAC, radical surgery was performed to remove the primary tumor and the regional draining lymph nodes. Pathologic evaluation of lung cancer resection specimens was conducted according to the IASLC multidisciplinary recommendations[21]. Archival slides of all enrolled patients were retrospectively reviewed to assess the percentage of residual viable tumor (RVT), necrosis, and stroma percentages from each slide, with the overall RVT percentage (visual %RVT) calculated as the arithmetic mean. Major pathologic response (MPR) was defined as no more than 10% RVT, while complete pathologic response (CPR) was defined as the absence of viable tumor cells. Specific components of stromal inflammation within the tumor bed were reassessed according to consensus definitions[22].

Image acquisition and processing

Formalin-fixed, paraffin-embedded (FFPE) pretreatment biopsies before NAC were stained with hematoxylin and eosin (H&E) and digitized into WSIs at 20x resolution. After quality control review by a professional pathologist, unqualified sections, including include those with artifacts such as tissue folding, air bubbles, sectioning defects, blurred regions, compression artifacts, or scanner-induced color drift, were removed. The digital slides were then processed into non-overlapping 224 × 224 pixel tiles at a 20x magnification using the OpenSlide library. The Otsu method was used to exclude tiles with insufficient tissue coverage (less than 40%). The average number of tiles per slide was approximately 820, with a standard deviation of 790 and a range of 71 to 6499 tiles, depending on the tissue area.

Feature extraction from pathology images

The preprocessed image tiles were transformed into 768-dimensional feature vectors using CtransPath, a transformer-based contrastive learning model fine-tuned on more than 30,000 WSIs from different organs and cancer types. For each patient, the feature extraction yielded an n×768 feature matrix, where n represents the number of tiles.

Development of the deep learning model

We developed a weakly supervised deep learning framework, DeepDrRVT, to predict pathologic responses to NAC before treatment. The model development workflow is shown in Fig. 2. The proposed model used an attention-based deep multiple instance learning (attMIL) architecture to aggregate patch-level features into patient-level predictions. The attMIL architecture allowed the network to focus on the most relevant information for %RVT prediction. Specifically, a bag X= {X1, …, Xn} consists of n image tiles, where n varies for different bags. The bag representation of patients was input into a fully connected (FC) and ReLU layer as a feature embedding to generate n-dimensional features of size M, denoted as Z. Two sub-attention layers were applied through a fully connected layer with tanh and sigmoid activation functions. According to Eq.(1) and obtained an attention weight A:

$$A = soft\,max(w_3\ \tanh(w_1 Z + b_1) * sigmoid(w_2 Z + b_2) + b_3) \quad (1)$$

where Z is the input feature, $w_i Z + b_i (i = 1, 2, 3)$ and softmax are the activation functions. The outputs from two sub-attention networks are multiplied and passed through a third FC layer with a softmax function to obtain the final attention with a size of n × 1, where n is the number of image tiles.

The attention weights were then multiplied by the outputs from the bag encoder layer, and the resulting product was flattened. This was followed by another fully connected layer with a sigmoid activation function to produce the predicted output, Dig.RVT.

$$Dig.RVT = sig\,\mathrm{mod}(w_4(Z * A) + b_4) \quad (2)$$

Batch-based Monte Carlo Loss (BMC Loss) was used as the loss function to predict the continuous value of %RVT.

$$\text{BMC Loss} = 2\sigma^2 \times CrossEntropy\left(-\frac{\left(Dig.RVT - Vis.RVT^T\right)^2}{2\sigma^2}, Vis.RVT\right) \quad (3)$$

where σ^2 is the noise variance, $Vis.RVT$ represents the target values and *CrossEntropyLoss* was used to calculate the cross-entropy loss.

The proposed deep learning framework was implemented using the open-source PyTorch library and trained on an NVIDIA GeForce RTX 4080 GPU. The model was optimized using the AdamW optimizer with a learning rate of 1e-3. Neither weight decay nor learning rate schedulers were used. The batch size was set to 8, and the hyperparameter σ^2 was configured to 0.6.

Visualization and explanation

To elucidate the decision-making process of the deep learning model using clinically relevant features, the attention component of the attMIL architecture was used. High-resolution attention heatmaps were generated by loading the attMIL model and the corresponding weights obtained from the training process. These heatmaps visually highlighted the areas that the model considered most relevant in its predictions. The Box-Cox transformation was used to account for the skewness of the attention weights.

Statistical analysis

All statistical analyses were performed using Python (version 3.7) and R (version 4.1.3). For comparisons between two groups, continuous variables were analyzed using Student's t-test, while categorical variables were compared using Fisher's exact test or chi-squared test. Survival curves were generated using the Kaplan-Meier method, and the log-rank test for comparisons using the R package 'survminer' (version 0.4.9). Cox regression analyses were performed for both univariate and multivariate models to estimate hazard ratios (HR) and their corresponding 95% confidence intervals (CI). The Youden index was used to determine the optimal threshold for the classification of responders and non-responders. Model performance was assessed using several metrics, including area under the curve (AUC), accuracy (ACC), sensitivity (SEN), specificity (SPE), F1-score, and confusion matrix. The clinical utility of the predictive model was evaluated using decision curve analysis (DCA). Calibration curves were used to assess the goodness of fit of the model. A two-tailed $p < 0.05$ was considered statistically significant.

Data availability

The WSI datasets generated and/or analyzed during the current study are not publicly available due to hospital regulations, but could be available from the corresponding author on reasonable request.

Code availability

The source code used in this study is publicly available at https://github.com/ZhoulabCPH/DeepDrRVT.

Received: 13 January 2025; Accepted: 28 April 2025;
Published online: 07 May 2025

References

1. Siegel, R. L., Giaquinto, A. N. & Jemal, A. Cancer statistics, 2024. *CA Cancer J. Clin.* **74**, 12–49 (2024).
2. Riely, G. J. et al. Non-Small Cell Lung Cancer, Version 4.2024, NCCN Clinical Practice Guidelines in Oncology. *J. Natl. Compr. Canc Netw.* **22**, 249–274 (2024).
3. Spicer, J. D. et al. Neoadjuvant and Adjuvant Treatments for Early Stage Resectable NSCLC: Consensus Recommendations From the International Association for the Study of Lung Cancer. *J. Thorac. Oncol.* https://doi.org/10.1016/j.jtho.2024.06.010 (2024).

4. Fox, A. H., Alexander, M., Forcucci, J. A. & Silvestri, G. A. Biomarker Testing for Guiding Precision Medicine for Patients with Non-Small Cell Lung Cancer. *Chest* https://doi.org/10.1016/j.chest.2024.08.006 (2024).
5. Perez-Lopez, R., Ghaffari Laleh, N., Mahmood, F. & Kather, J. N. A guide to artificial intelligence for cancer researchers. *Nat. Rev. Cancer* **24**, 427–441 (2024).
6. Hoang, D. T. et al. A deep-learning framework to predict cancer treatment response from histopathology images through imputed transcriptomics. *Nat. Cancer* **5**, 1305–1317 (2024).
7. McCaffrey, C. et al. Artificial intelligence in digital histopathology for predicting patient prognosis and treatment efficacy in breast cancer. *Expert Rev. Mol. Diagn.* **24**, 363–377 (2024).
8. Shao, Y. et al. Prostate Cancer Risk Stratification by Digital Histopathology and Deep Learning. *JCO Clin. Cancer Inf.* **8**, e2300184 (2024).
9. Yang, Z. et al. Prediction of prognosis and treatment response in ovarian cancer patients from histopathology images using graph deep learning: a multicenter retrospective study. *Eur. J. Cancer* **199**, 113532 (2024).
10. Zhang, Y. et al. Histopathology images-based deep learning prediction of prognosis and therapeutic response in small cell lung cancer. *NPJ Digit Med.* **7**, 15 (2024).
11. Spicer, J. D. et al. Neoadjuvant pembrolizumab plus chemotherapy followed by adjuvant pembrolizumab compared with neoadjuvant chemotherapy alone in patients with early-stage non-small-cell lung cancer (KEYNOTE-671): a randomised, double-blind, placebo-controlled, phase 3 trial. *Lancet* **404**, 1240–1252 (2024).
12. Yang, Z. et al. Treatment patterns and clinical outcomes of patients with resectable non-small cell lung cancer receiving neoadjuvant immunochemotherapy: A large-scale, multicenter, real-world study (NeoR-World). *J. Thorac. Cardiovasc. Surg.* **168**, 1245–1258.e1217 (2024).
13. Castello, A. et al. Predictive and Prognostic Role of Metabolic Response in Patients With Stage III NSCLC Treated With Neoadjuvant Chemotherapy. *Clin. Lung Cancer* **21**, 28–36 (2020).
14. Cerfolio, R. J., Bryant, A. S., Winokur, T. S., Ohja, B. & Bartolucci, A. A. Repeat FDG-PET after neoadjuvant therapy is a predictor of pathologic response in patients with non-small cell lung cancer. *Ann. Thorac. Surg.* **78**, 1903–1909 (2004).
15. Huang, M. et al. Aneuploid Circulating Tumor Cells as a Predictor of Response to Neoadjuvant Chemotherapy in Non-Small Cell Lung Cancer. *Int J. Gen. Med.* **14**, 6609–6620 (2021).
16. Jiang, Z. et al. Machine Learning-Based Prediction of Pathological Responses and Prognosis After Neoadjuvant Chemotherapy for Non-Small-Cell Lung Cancer: A Retrospective Study. *Clin. Lung Cancer* **25**, 468–478.e463 (2024).
17. William, W. N. Jr. et al. Computed tomography RECIST assessment of histopathologic response and prediction of survival in patients with resectable non-small-cell lung cancer after neoadjuvant chemotherapy. *J. Thorac. Oncol.* **8**, 222–228 (2013).
18. Stefano, A. Challenges and limitations in applying radiomics to PET imaging: Possible opportunities and avenues for research. *Comput. Biol. Med.* **179**, 108827 (2024).
19. Zhou, Y. et al. Tumor biomarkers for diagnosis, prognosis and targeted therapy. *Signal Transduct. Target Ther.* **9**, 132 (2024).
20. Deutsch, J. S. et al. Association between pathologic response and survival after neoadjuvant therapy in lung cancer. *Nat. Med.* **30**, 218–228 (2024).
21. Travis, W. D. et al. IASLC Multidisciplinary Recommendations for Pathologic Assessment of Lung Cancer Resection Specimens After Neoadjuvant Therapy. *J. Thorac. Oncol.* **15**, 709–740 (2020).
22. Wang, S. et al. Pathological response and tumor stroma immunogenic features predict long-term survival in non-small cell lung cancer after neoadjuvant chemotherapy. *Cell Oncol.* **47**, 1005–1024 (2024).

Acknowledgements

The study was funded by the Beijing Chao-Yang Hospital Golden Seeds Foundation (CYJZ202410) and the Beijing Clinical Key Specialty Construction Project.

Author contributions

M.Z., L.Y., D.M.L. and Y.S.M. designed the study and developed the conceptual ideas. S.B.W., X.Y.L., Y.Q. and B.H. collected all the image sources and additional data. Y.B.Z., Z.J.Y. and Y.S. implemented the main algorithms, computational analysis and analyzed the results. Y.B.Z. and S.B.W. wrote the manuscript. M.Z. and L.Y discussed and reviewed the manuscript. All the authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41698-025-00927-4.

Correspondence and requests for materials should be addressed to Yousheng Mao, Dongmei Lin, Lin Yang or Meng Zhou.