He et al. Mil Med Res 2026
https://doi.org/10.1016/j.mmr.2026.100001

MMR MILITARY MEDICAL RESEARCH

RESEARCH Open Access

Non-invasive evaluation of muscle invasion and survival prognosis in bladder cancer using enhanced CT-based deep learning radiomics: a multi-center real-world cohort study

Yun-Bo He[1,2†], Jiao Hu[1,2†], Zhi Liu[1,3], Zi-Cheng Xiao[1,2], Jin-Hui Liu[1,2], Hai-Su Liang[1,2], Wen-Zhi Deng[4], Zhi-Wei Li[5], Jun Zhang[6], Jia-Quan Long[7], Ning Gao[8], Bin Huang[8], Xi Guo[9], Zhen-Yu Ou[1,2], Jin-Bo Chen[1,2], Pei-Hua Liu[1,2], Min-Feng Chen[1,2], Hui-Huang Li[1,2], Rui-Zhe Wang[1,2], Xiao Guan[1,2], Shi-Yu Tong[1,2], Yang-Le Li[1,2], Wei He[1,2], Yan-Hua Zhao[1,2], Zhi-Yong Cai[1,2], Yu Gan[1,2], Cheng Zhao[1,2], Yu Cui[1,2], Yuan-Qing Dai[1,2], Yi Cai[1,2], Zhen-Yu Nie[1,2], Wei-Min Zhou[1,2], Bo-Han Zhou[1,2], Ming-Hui Hu[1,2], Ben-Yi Fan[1,2*], Ding-Shan Deng[9*] and Xiong-Bing Zu[9*]

Abstract

Background: Bladder cancer (BLCA) is a prevalent malignancy characterized by high recurrence and poor prognosis, particularly muscle-invasive bladder cancer (MIBC). Histopathology, the gold standard for assessing muscle invasion, often suffers from sampling errors and operator dependency, underscoring the need for non-invasive, accurate preoperative assessment methods. This study aimed to develop and validate a hybrid artificial intelligence (AI) model based on computed tomography (CT) radiomics and deep learning (DL) to predict MIBC and overall survival (OS) preoperatively in BLCA patients.

Methods: A total of 1370 patients from 6 academic medical centers were retrospectively included. Preoperative contrast-enhanced CT scans were analyzed to extract handcrafted radiomic features using PyRadiomics and DL features using ResNet101, followed by machine learning (ML)-based modeling for prediction. A hybrid model combining radiomic and DL features was constructed and validated in internal and external cohorts. Model performance was evaluated using metrics such as the area under the curve (AUC) and Cox proportional hazards analysis for OS prediction.

Results: The DL radiomics nomogram (DLRN) model demonstrated superior diagnostic performance, achieving an AUC of 0.807 in the internal validation cohort and 0.783 in the external multi-center validation cohort for predicting muscle invasion. The DLRN generated an imaging-derived risk score (DLRN score), which was subsequently incorporated as one covariate into a multivariable Cox proportional hazards model together with clinicopathological variables to evaluate OS. Using this approach, patients were effectively stratified into high- and low-risk groups for OS, showing robust generalizability across diverse clinical settings. AI-assisted diagnostics significantly improved the sensitivity and accuracy of urologists, particularly among less experienced clinicians.

Conclusion: The DLRN model provides a reliable, non-invasive tool for preoperative assessment of muscle invasion and prognosis in BLCA. Addressing histopathology limitations, it offers valuable insights for personalized treatment strategies, paving the way for precision oncology in real-world clinical applications.

Key words Bladder cancer (BLCA), Deep learning (DL), Multi-center study, Artificial intelligence (AI), Radiomics

Background

Bladder cancer (BLCA) is one of the most prevalent malignancies worldwide [1]. According to the American Cancer Society, in 2023, new cases and deaths from BLCA are projected to rank 4th and 8th, respectively, among all male cancers in the United States [2]. Clinically, BLCA is characterized by rapid progression, high recurrence rates, multidrug resistance, and poor prognosis. For instance, the recurrence rate of non-muscle-invasive bladder cancer (NMIBC) within 1-year of diagnosis can reach 60%, whereas

†Yun-Bo He and Jiao Hu contribute equally to this work
*Correspondence to: Xiong-Bing Zu, zuxbxy@csu.edu.cn; Ding-Shan Deng, dds15116217256@163.com; Ben-Yi Fan, fanbenyi2009@yeah.net
[1]Department of Urology, Xiangya Hospital, Central South University, Changsha 410008, China
[9]Department of Urology, Hunan Provincial People's Hospital/the First Affiliated Hospital of Hunan Normal University, Changsha 410005, China
Full list of author information is available at the end of the article

Received: 25 May 2025 Accepted: 05 March 2026

the 5-year survival rate for muscle-invasive bladder cancer (MIBC) is approximately 50% [3,4]. Despite significant advances in medical research providing diverse treatment options, such as surgery, radiotherapy, chemotherapy, immunotherapy, and targeted therapy [5], the 5-year survival rate for patients with distant metastases remains below 15% [6]. Although histopathology remains the gold standard for diagnosing muscle invasion in clinical management, biopsy-based approaches are operator-dependent and may fail to adequately sample all tumor regions [7]. Furthermore, thermal damage during transurethral resection of bladder tumor (TURBT) can compromise sample quality, potentially leading to understaging, with MIBC being mistaken for NMIBC [8]. These limitations highlight the need for novel preoperative methods to assess muscle invasion in BLCA and improve treatment planning and prognosis.

Artificial intelligence (AI) refers to computational methods that enable machines to perform tasks typically requiring human intelligence, such as perception, language understanding, pattern recognition, and decision-making [9], through techniques including machine learning (ML) and deep learning (DL) [10]. ML and DL are major subfields of AI, with ML focusing on algorithms that learn patterns from data, and DL representing a subset of ML that employs multi-layer neural networks to automatically extract hierarchical features for predictive tasks [11]. Currently, various DL methods that integrate genomic, transcriptomic, or histopathological data are used in clinical oncology and translational cancer research for tumor diagnosis, prognosis, treatment selection, and drug discovery [12]. These substantially influenced advancements in clinical diagnostic and therapeutic approaches, particularly in medical imaging, where AI's advanced image recognition algorithms enable non-invasive prediction of disease occurrence, progression, treatment response, and prognosis. These capabilities provide a foundation for early and precise clinical decision-making [13,14].

Computed tomography (CT)-based radiomics has shown promise in diagnosing various diseases [15,16]. Traditional radiomics studies predominantly employ ML algorithms, as handcrafted radiomic features are typically modeled using ML classifiers [17-19]. Recent advances in DL, however, have enabled automatic extraction of high-level imaging representations and achieved strong performance in a variety of medical imaging tasks [20,21]. Previous research has demonstrated the feasibility of CT radiomics, mainly ML-based, in distinguishing low- from high-grade BLCA [22]. However, the integration of CT-based radiomics with DL for preoperative assessment of muscle invasion and survival prognosis in BLCA remains limited. Therefore, this study aims to develop a hybrid prediction model combining CT-based radiomics and DL features to evaluate muscle invasion in BLCA preoperatively. This approach is intended to support clinicians in developing therapeutic strategies, especially in cases where biopsy samples are insufficient or compromised, thereby functioning as a valuable adjunct to clinical decision-making.

Materials and methods

Study design and patients

We retrospectively analyzed data from 2909 consecutive BLCA patients who underwent surgery at 6 academic medical centers (center 1–6). Eligible patients had histologically confirmed BLCA based on tissue obtained from TURBT or radical cystectomy (RC), along with preoperative pelvic CT scans and complete clinicopathological data. Patients with other synchronous malignancies or whose CT scans did not clearly identify the primary tumor were excluded. The detailed inclusion and exclusion criteria are presented in Fig. 1 and Additional file 1: Methods.

A total of 1370 patients from 7 independent cohorts satisfying these criteria were included in the study. The training cohort (n=827) and internal validation cohort (n=209) were derived from center 1. Patients in the training cohort were enrolled consecutively from January 1, 2016, to December 31, 2022, while those in the internal validation cohort were enrolled from January 1, 2023, to December 31, 2023. Center 2–6 consisted of patients from 5 independent Chinese academic medical centers enrolled from January 1, 2018, to December 31, 2023, while the 7th cohort was derived from The Cancer Genome Atlas (TCGA)-BLCA imaging dataset (https://doi.org/10.7937/K9/TCIA.2016.8LNG8XDR) [23], representing predominantly North American patients. The 6 external validation cohorts were consolidated into a single external multi-center validation cohort (n=334) after adjusting for center effects (Fig. 1). This design facilitated a preliminary assessment of the model's robustness across different ethnic and geographic populations. A temporal sampling strategy was used to divide the training and internal validation cohorts [24], enabling a pseudo-prospective evaluation in which the model is trained on earlier cases and tested on newly collected samples. By consolidating multiple medical centers into a single external multi-center validation cohort, we aimed to rigorously evaluate the model's generalizability, given the diverse regions, medical practices, and imaging equipment across the external centers.

Clinical variables collected included gender, age, BMI (kg/m^2), smoking history, alcohol consumption history, past

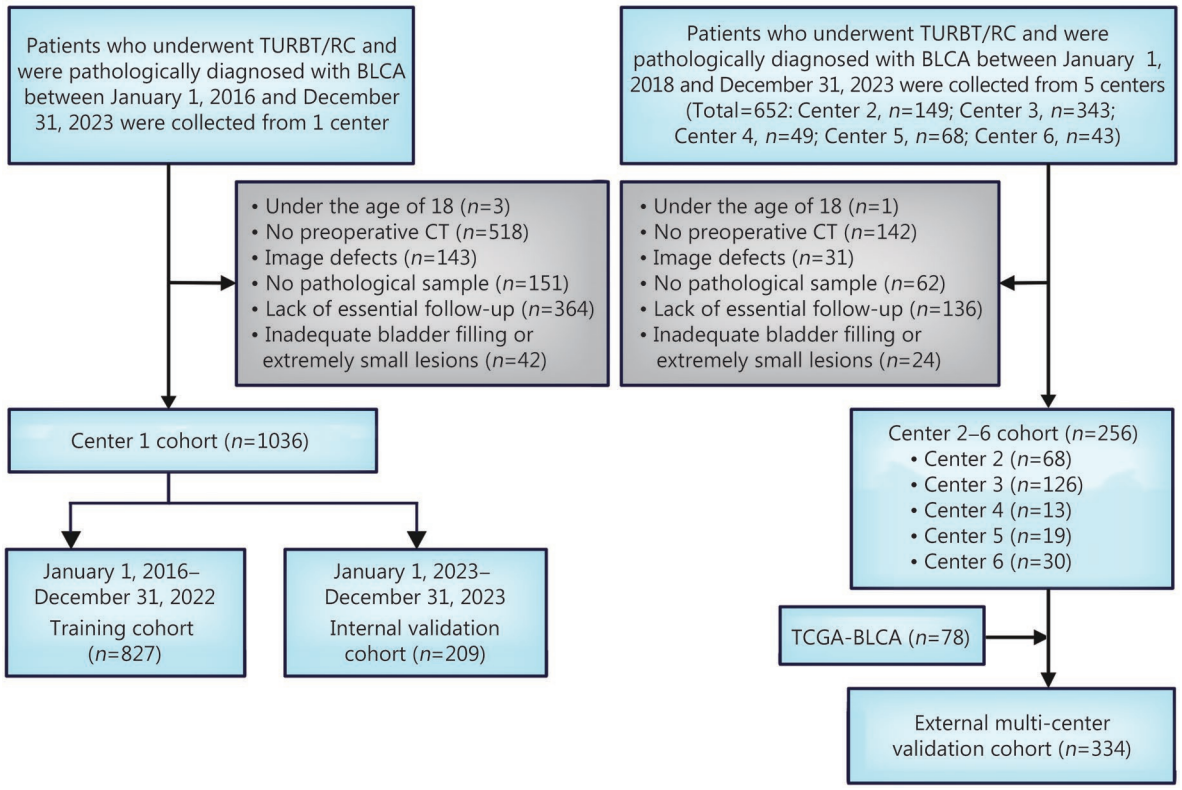

Fig. 1 The inclusion and exclusion process of patients in this study.
Center 1: Xiangya Hospital of Central South University; Center 2: Xiangya Boai Rehabilitation Hospital; Center 3: the Third Xiangya Hospital of Central South University; Center 4: the Second Affiliated Hospital of Guizhou Medical University; Center 5: the First People's Hospital of Kaili; Center 6: the Second Hospital of University of South China. Details of the CT data for TCGA-BLCA: https://doi.org/10.7937/K9/TCIA.2016.8LNG8XDR. BLCA. Bladder cancer; TURBT. Transurethral resection of bladder tumor; RC. Radical cystectomy; CT. Computed tomography

medical history (e.g., hypertension and diabetes), and drug allergy history. In addition to these general clinical data, tumor-related clinical and pathological characteristics were also recorded, including tumor-induced hydronephrosis, surgical treatment, and tumor size (mm). Tumor size was defined based on the postoperative pathological examination of the resected specimen and recorded as the maximum diameter (mm) of the largest tumor; in patients with multifocal disease, the largest lesion was used for this measurement. All patients were restaged according to the 8th edition of the American Joint Committee on Cancer (AJCC) staging system [25]. This study was approved by the Clinical Research Ethics Committee of Xiangya Hospital, Central South University (202304064), with informed consent waived. The workflow involving data preprocessing, model construction, and prospective validation is illustrated in Fig. 2.

Image acquisition and processing

All patients underwent preoperative contrast-enhanced CT examinations of the bladder region. The delayed (excretory) phase contrast-enhanced CT images used for model training were retrieved from the institutional Picture Archiving and Communication System (PACS). The detailed imaging acquisition protocols used at each center are summarized in Additional file 1: Table S1. Bladder tumors were delineated on CT images using ITK-SNAP software (version 3.8.0) by two radiologists (radiologist A with 5 years and radiologist B with 8 years of clinical experience in pelvic CT interpretation) and one senior urologist (urologist C with 25 years of clinical experience in urology). All 3 physicians collaborated during the delineation process and reached a consensus on tumor descriptions to ensure accuracy and consistency. The DL radiomics nomogram (DLRN) model was trained exclusively on tumor-specific regions of interest (ROIs), focusing on the segmented tumor areas rather than the entire bladder, to capture discriminative morphological and textural features associated with muscle invasion.

All images were filtered using a window level of –75 to 175 Hounsfield units (HUs) and subsequently resampled to a standardized in-plane resolution of 512×512 pixels. To

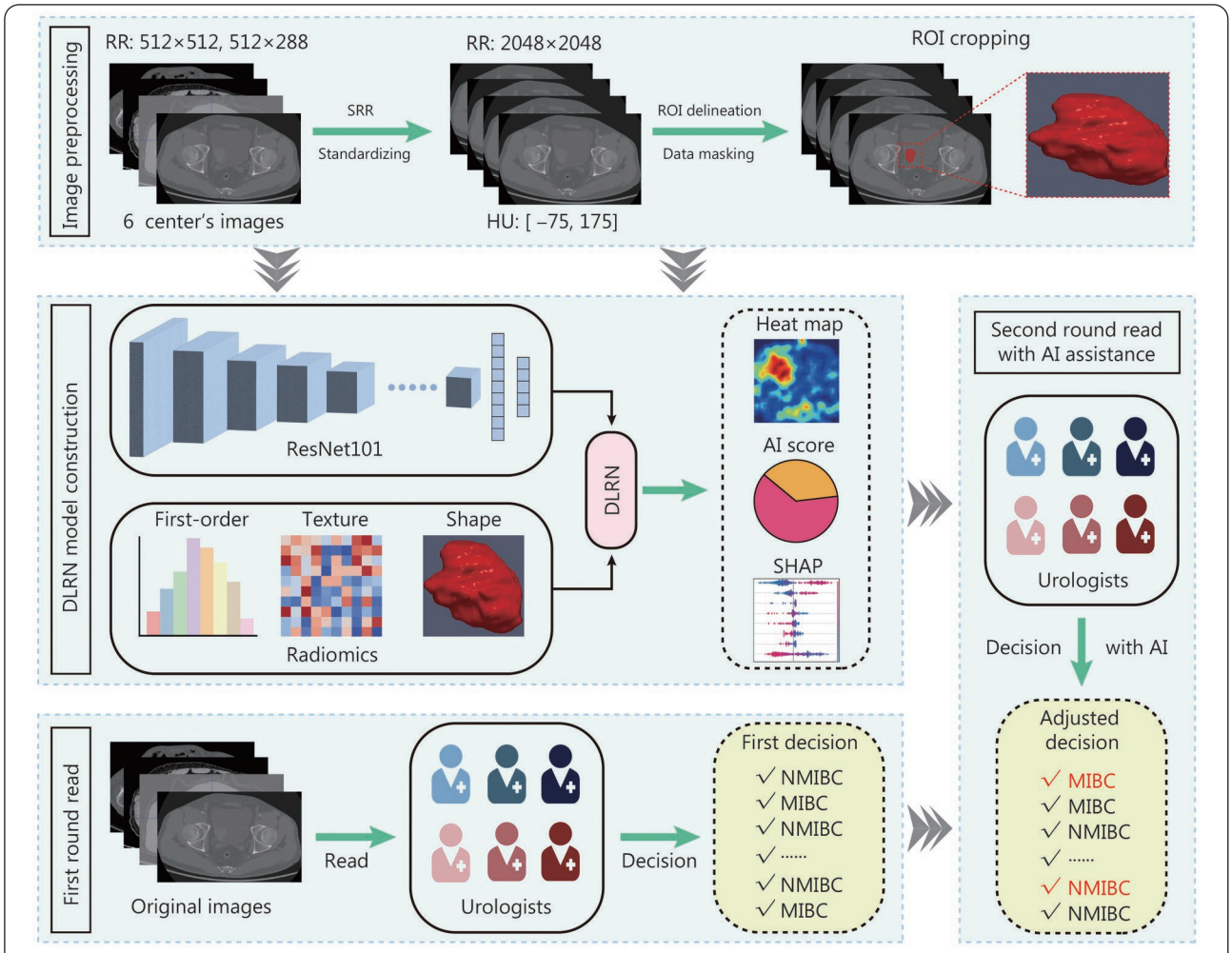

Fig. 2 Workflow of the deep learning radiomics nomogram (DLRN) model development and AI-assisted clinical evaluation. Image preprocessing: CT images from 6 centers were resampled to a reconstruction resolution (RR) of 512×512 and 512×288 and intensity-standardized within a Hounsfield unit (HU) range of −75 to 175. A super-resolution reconstruction (SRR) algorithm was applied to generate 2048×2048 high-resolution images, enhancing detail uniformity across scanners. Tumor region of interests (ROIs) were manually delineated, masked, and cropped for subsequent analysis. DLRN model construction: the hybrid DLRN combined radiomic features (first-order, texture, and shape descriptors) with DL features extracted via ResNet101. The integrated model output included a heatmap, AI score, and feature attribution (SHAP) for interpretability. First-round read: urologists independently reviewed original CT images to classify patients as non-muscle-invasive bladder cancer (NMIBC) or muscle-invasive bladder cancer (MIBC). Second-round read: urologists re-evaluated cases with DLRN assistance, incorporating AI-generated predictions and heatmaps, leading to adjusted and more accurate decisions. SHAP. SHapley Additive exPlanations; AI. Artificial intelligence

enhance image quality and reduce potential information loss from heterogeneous scanners, we applied a generative adversarial network (GAN)-based super-resolution method to resample CT images [26,27]. Specifically, images were resampled from 512×512 to 2048×2048 resolution. This GAN-based super-resolution significantly improved model area under the curve (AUC) in the training (P=0.008) and external multi-center validation cohorts (P=0.031), with a consistent trend in the internal validation cohort (P=0.403) (Additional file 1: Fig. S1). To ensure that the super-resolution reconstruction did not introduce distortion or artifacts, we validated the consistency of radiomic features before and after super-resolution in a randomly selected subset of 100 patients drawn from all 7 cohorts. The intraclass correlation coefficients (ICCs) of key radiomic features exceeded 0.85. We also inspected representative cases with two radiologists to ensure that no artificial tumor structures were introduced. For patients with multiple lesions, the three-dimensional (3D) CT volume of the largest tumor was used as the representative lesion for imaging feature extraction and subsequent survival analysis.

Model procedures and selection

Handcrafted radiomic features were extracted via an in-house feature analysis program implemented with Pyradiomics

(https://github.com/Radiomics/pyradiomics) [18]. Feature selection in the training cohort was conducted using ICC analysis, *t*-tests, Pearson correlation coefficients, least absolute shrinkage and selection operator (LASSO), logistic regression, and Cox regression (Additional file 1: Methods). The radiomic score (RL) (Rad-score) was calculated by combining the selected features with their respective weights. Several ML algorithms, including logistic regression (LR), support vector machine (SVM), k-nearest neighbors (KNN), random forests, extra trees, extreme gradient boosting (XGBoost), light gradient boosting machine (LightGBM), and multilayer perceptron (MLP), were employed to construct radiomic models and scores. The training cohort was divided into training and validation subsets at a 7:3 ratio, and these subsets were used to build models with the specified ML methods. Metrics such as accuracy, sensitivity, and the AUC of the receiver operating characteristic (ROC) curve were calculated to evaluate the suitability of each ML algorithm (Additional file 1: Fig. S2). Among the tested algorithms, KNN and XGBoost exhibited relatively higher median AUC values, whereas SVM demonstrated the poorest performance (Additional file 1: Fig. S2a). In the training cohort (Additional file 1: Fig. S2b), KNN achieved the highest AUC (0.877), followed by XGBoost (AUC=0.858). However, the accuracy of KNN decreased markedly in the test cohort (Additional file 1: Fig. S2c), indicating potential overfitting. In contrast, XGBoost maintained stable performance between the training and test cohorts. In the ROC analysis of the test cohort (Additional file 1: Fig. S2d), XGBoost also demonstrated well-balanced predictive performance across metrics, and was therefore selected as the primary algorithm for constructing the radiomic model and scores.

To compare their predictive performance, we evaluated 15 representative DL models. A 7:3 split of the training cohort was used to train and validate deep neural network models, including ResNet50 and ResNet101 (Additional file 1: Fig. S3). ResNet101 outperformed other models and was chosen as the final feature extraction model. Features extracted from the average pooling layer of ResNet101 were utilized as DL features. For DL, 2048 features were extracted from the ResNet101 average pooling layer. To reduce redundancy and overfitting risk, we applied dimension correlation analysis (DCA), which summarized these into 8 representative dimensions (DL_0–7) capturing the major variance. Subsequent feature selection (*t*-tests, Pearson correlation coefficients, LASSO) retained 3 DL features (DL_0, DL_1, DL_3), which were combined with 13 radiomics features (5 GLSZM features, 4 first-order features, 2 NGTDM features, 1 GLRLM feature, and 1 shape feature) to construct the final DLRN model.

Detailed hyperparameter settings related to model training and preprocessing are provided in the configuration script (config.py) and summarized in Additional file 1: Tables S2 and S3. These include learning rates, number of estimators (ntrees), batch sizes, epochs, and optimizer details for each ML and DL model.

To leverage the complementary strengths of radiomics and deep learning, we integrated handcrafted radiomic features with DL-derived representations to develop a hybrid model for predicting muscle invasion in BLCA. The XGBoost algorithm was also employed to construct a clinical model (Clinic) based on clinicopathological factors. We developed a comprehensive nomogram that incorporated the hybrid model, the DL signature (DL_Signature), and clinicopathological factors to provide individualized assessments of overall survival (OS) for post-surgical BLCA patients.

Survival analysis

For OS prediction, clinical and pathological variables including age, gender, tumor size, T stage, surgical approaches, pathological grade, presence of hydronephrosis, and lymph node status were incorporated into multivariate Cox regression. For survival analysis, the DLRN score was not applied as an independent predictive model but was incorporated as a covariate, alongside clinicopathological factors, into a multivariable Cox proportional hazards model and nomogram for OS estimation. In addition, the DLRN score derived from preoperative CT was included to assess its incremental prognostic value.

Human-AI interaction experiment

A total of 100 patients with BLCA who underwent standard diagnostic evaluation and preoperative management at Xiangya Hospital between January 1 and April 30, 2024, were prospectively enrolled as a validation cohort. All patients underwent contrast-enhanced CT imaging before receiving definitive surgical treatment (TURBT or radical cystectomy); because surgery occurred after CT acquisition, the treatment regimen did not influence the imaging-based assessment. A prospective observer study was conducted involving 6 radiologists and 6 urologists. Observers interpreted CT scans both with AI assistance (With-AI) and without AI assistance (No-AI) using a counterbalanced crossover design. Detailed procedures regarding observer blinding, randomization, second-read bias mitigation, and statistical procedures are provided in Additional file 1: Methods.

Outcomes

The primary outcome was the presence of muscle invasion in BLCA, determined by pathological confirmation, which the models aimed to predict using preoperative CT imaging. The secondary outcome was overall survival (OS), defined as the time from surgery to death from any cause or last follow-up.

Statistical analysis

Data were expressed as mean±standard deviation (SD) for normally distributed continuous variables and as median (interquartile range, IQR) for non-normally distributed data. Categorical variables were presented as n(%). We used t-tests to compare continuous variables and χ^2 tests or Fisher's exact tests for categorical variables between the two groups. Survival curves were generated using the Kaplan-Meier method. Univariate and multivariate analyses were performed with the Cox proportional hazards model. For all two-tailed analyses, a P-value<0.05 was considered statistically significant. The delong test was applied to assess statistical differences between the AUC of different predictive models. Statistical analyses were conducted using Python (version 3.11.7), R (version 4.3.3), and SPSS (version 21.0).

Results

Patient characteristics

The detailed clinicopathological characteristics and treatment outcomes of patients from the 6 medical centers included in this study are summarized in Table 1. A total of 827 patients were included in the training cohort, 209 in the internal validation cohort, and 334 in the external multi-center validation cohort. The distributions of MIBC and NMIBC were as follows: 376 MIBC and 451 NMIBC cases in the training cohort; 81 MIBC and 128 NMIBC cases in the internal validation cohort; and 180 MIBC and 154 NMIBC cases in the external multi-center validation cohort.

In the training cohort, patients with MIBC had significantly larger tumors than those with NMIBC [(26.54±15.74) mm vs. (23.52±14.35) mm, P=0.004] and were slightly younger [(62.32±11.04) years vs. (64.11±11.29) years, P=0.022]. No significant difference in body mass index (BMI) was observed between the groups (P=0.761). The proportion of patients with multiple tumors was comparable between the MIBC and NMIBC groups (63.30% vs. 65.19%, P=0.572). With respect to clinical T stage, the overall distribution of stages differed significantly between MIBC and NMIBC (P<0.001), with a higher proportion of patients presenting with ≥cT2 disease in the MIBC group than in the NMIBC group (28.72% vs. 12.42%), consistent with their disease classification. No significant differences were observed in gender distribution between MIBC and NMIBC groups across all cohorts (P=0.293). Similarly, the surgical approach (TURBT vs. RC) showed no significant difference between the two groups in the training or validation cohorts (P=0.911). For tumor grade, high-grade lesions were more common in MIBC patients than in NMIBC patients in the training cohort (68.74% vs. 39.63%), and the overall distribution of pathological grades differed significantly between the two groups (P<0.001), consistent with the more aggressive biological behavior of MIBC. However, pathological grade distributions were relatively balanced in the validation cohorts. Regarding histological subtype, urothelial carcinoma was predominant in all cohorts (>93%), with few cases of squamous cell carcinoma, adenocarcinoma, or small cell carcinoma, and no significant difference was found between MIBC and NMIBC groups (P>0.2). In the internal validation cohort, tumor size, age, and BMI did not differ significantly between MIBC and NMIBC groups. However, in the external multi-center validation cohort, BMI was significantly higher in MIBC patients than NMIBC patients [(24.95±4.10) kg/m^2 vs. (23.23±2.35) kg/m^2, P<0.001], while tumor size and age remained statistically similar. A significantly higher proportion of multiple tumors was observed in MIBC patients than in NMIBC patients (76.11% vs. 59.74%, P=0.001). These findings indicate mild heterogeneity across cohorts and emphasize the clinical distinctions between MIBC and NMIBC subgroups.

Performance of the radiomic model

On the basis of the evaluation of 8 ML algorithms (Additional file 1: Fig. S2), the XGBoost algorithm was selected to construct the prediction model using the identified features. Radiomic features within the tumor ROI were extracted using the Pyradiomics tool (version 2.1.2) in Python [18]. A total of 1834 features were extracted (Additional file 1: Fig. S4a, b), and all features were normalized using Z-scores to minimize variability across different scales. During feature selection, t-tests were performed to identify features with significant differences between the two risk groups (P<0.05). Pearson correlation coefficients were calculated to assess inter-feature correlations, and the LASSO logistic regression model was applied to further reduce the feature set, retaining the most predictive features for model construction (Additional file 1: Fig. S4c, d). The XGBoost classifier's feature importance scores highlighted the most influential features for predicting muscle invasion in BLCA (Additional file 1: Fig. S4e). Spectral visualization of dimensionality reduction demonstrated the discriminative ability of the selected features in differentiating

Table 1 Baseline characteristics of the training, internal validation, and external multi-center validation cohorts

Charateristics	Training cohort (*n*=827)			Internal validation cohort (*n*=209)			External multi-center validation cohort (*n*=334)		
	MIBC (*n*=376)	NMIBC (*n*=451)	*P*-value	MIBC (*n*=81)	NMIBC (*n*=128)	*P*-value	MIBC (*n*=180)	NMIBC (*n*=154)	*P*-value
Tumor size (mm, mean±SD)	26.54±15.74	23.52±14.35	0.004	27.62±19.80	26.84±19.43	0.779	28.66±13.11	29.87±14.11	0.417
Age (year, mean±SD)	62.32±11.04	64.11±11.29	0.022	65.77±11.31	63.13±12.73	0.130	66.43±10.28	67.14±10.11	0.526
BMI (kg/m^2, mean±SD)	23.32±3.10	23.39±3.04	0.761	22.99±2.40	23.35±2.83	0.343	24.95±4.10	23.23±2.35	<0.001
Multiple tumors [*n*(%)]			0.572			0.066			0.001
Yes	238(63.30)	294(65.19)		49(60.49)	93(72.66)		137(76.11)	92(59.74)	
No	138(36.70)	157(34.81)		32(39.51)	35(27.34)		43(23.89)	62(40.26)	
cT [*n*(%)]			<0.001			<0.001			<0.001
cT1	268(71.28)	395(87.58)		44(54.32)	109(85.16)		61(33.89)	139(90.26)	
cT2	82(21.81)	52(11.53)		31(38.27)	14(10.94)		52(28.89)	14(9.09)	
cT3	19(5.05)	2(0.44)		4(4.94)	3(2.34)		42(23.33)	1(0.65)	
cT4	7(1.86)	2(0.44)		2(2.47)	2(1.56)		25(13.89)	0	
Gender [*n*(%)]			0.293			0.367			0.765
Male	302(80.32)	375(83.15)		63(77.78)	106(82.81)		149(82.78)	126(81.82)	
Female	74(19.68)	76(16.85)		18(22.22)	22(17.19)		31(17.22)	28(18.18)	
Surgical treatment [*n*(%)]			0.911			0.804			0.246
TURBT	278(73.94)	335(74.28)		52(64.20)	80(62.50)		167(92.78)	137(88.96)	
RC	98(26.06)	116(25.72)		29(35.80)	48(37.50)		13(7.22)	17(11.03)	
Pathological grade [*n*(%)]			<0.001			0.770			0.002
Low	149(39.63)	310(68.74)		45(55.56)	64(50.00)		129(71.67)	132(85.71)	
High	220(58.51)	127(28.16)		35(43.21)	62(48.44)		50(27.78)	20(12.99)	
Other	7(1.86)	14(3.10)		1(1.23)	2(1.56)		1(0.56)	2(1.30)	
Hydronephrosis [*n*(%)]			<0.001			0.011			0.015
No	310(82.45)	412(91.35)		68(83.95)	121(94.53)		162(90.00)	149(96.75)	
Yes	66(17.55)	39(8.65)		13(16.05)	7(5.47)		18(10.00)	5(3.25)	
Multiple tumors [*n*(%)]			<0.001			0.424			0.009
No	181(48.14)	359(79.60)		54(66.67)	92(71.88)		133(73.89)	93(60.39)	
Yes	195(51.86)	92(20.40)		27(33.33)	36(28.12)		47(26.11)	61(39.61)	
Intravesical therapy [*n*(%)]			0.399			0.809			0.997
No	354(94.15)	418(92.68)		76(93.83)	119(92.97)		173(96.11)	148(96.10)	
Yes	22(5.85)	33(7.32)		5(6.17)	9(7.03)		7(3.89)	6(3.90)	

(Continued)

Charateristics	Training cohort (*n*=827)			Internal validation cohort (*n*=209)			External multi-center validation cohort (*n*=334)		
	MIBC (*n*=376)	NMIBC (*n*=451)	*P*-value	MIBC (*n*=81)	NMIBC (*n*=128)	*P*-value	MIBC (*n*=180)	NMIBC (*n*=154)	*P*-value
Hypertension [*n*(%)]			0.627			0.671			<0.001
No	284(75.53)	334(74.06)		56(69.14)	92(71.88)		158(87.78)	107(69.48)	
Yes	92(24.47)	117(25.94)		25(30.86)	36(28.12)		22(12.22)	47(30.52)	
Diabetes [*n*(%)]			0.641			0.951			0.022
No	333(88.56)	404(89.58)		70(86.42)	111(86.72)		168(93.33)	132(85.71)	
Yes	43(11.44)	47(10.42)		11(13.58)	17(13.28)		12(6.67)	22(14.29)	
Allergy history [*n*(%)]			0.395			0.831			0.567
No	348(92.55)	410(90.91)		74(91.36)	118(92.19)		176(97.78)	148(96.10)	
Yes	28(7.45)	41(9.09)		7(8.64)	10(7.81)		4(2.22)	6(3.90)	
Smoking status [*n*(%)]			0.259			0.136			0.001
No	214(56.91)	239(52.99)		44(54.32)	56(43.75)		71(39.44)	88(57.14)	
Yes	162(43.09)	212(47.01)		37(45.68)	72(56.25)		109(60.56)	66(42.86)	
Alcohol consumption [*n*(%)]			0.020			0.502			<0.001
No	298(79.26)	326(72.28)		58(71.60)	86(67.19)		165(91.67)	115(74.68)	
Yes	78(20.74)	125(27.72)		23(28.40)	42(32.81)		15(8.33)	39(25.32)	
Histological subtype [*n*(%)]			0.046			0.718			0.223
Urothelial carcinoma	361(96.78)	436(97.10)		76(93.83)	124(96.88)		177(98.88)	149(96.75)	
Squamous cell carcinomas	7(1.88)	1(0.22)		1(1.23)	1(0.78)		0	1(0.65)	
Adenocarcinoma	3(0.80)	9(2.00)		1(1.23)	0		1(0.56)	3(1.95)	
Small cell carcinoma	2(0.54)	2(0.45)		2(2.47)	2(1.56)		0	1(0.65)	
Sarcomatoid carcinoma	0	1(0.22)		1(1.23)	1(0.78)		1(0.56)	0	

χ^2 or Fisher's exact tests were used to test whether the variable composition varied significantly between NMIBC and MIBC patients. NMIBC. Non-muscle-invasive bladder cancer; MIBC. Muscle-invasive bladder cancer; TURBT. Transurethral resection of bladder tumor; RC. Radical cystectomy; cT. Clinical T stag

MIBC from NMIBC, although the separation was not perfect (Additional file 1: Fig. S4f).

The radiomic model exhibited good diagnostic performance in the internal validation cohort, achieving an AUC of 0.718 (95% CI 0.638–0.798). In the center 2–6 cohorts, the model's AUC varied significantly, likely due to the small sample sizes and varying scales of the external datasets. The model performed worst in the TCGA-BLCA cohort, with an AUC of 0.519. The model's AUC in the external multi-center validation cohort was 0.618 (95% CI 0.554–0.681) (Additional file 1: Fig. S5). This poor performance may reflect multiple factors: 1) the imaging data spanned 1998–2009, during which CT technology and acquisition protocols differed substantially from modern practice; 2) patient demographics and clinical practices in TCGA (mainly North American) differed from our Chinese cohorts, indicating potential population and distribution shifts; and 3) The radiomic feature distributions of the TCGA-BLCA cohort (North American patients) showed noticeable distributional shifts compared with the combined Chinese multi-center cohorts (center 1–6), although the PCA plot indicates that their feature spaces still share substantial overlap (Additional file 1: Fig. S6), suggesting a domain shift that impaired generalizability. These findings highlight that model fragility can emerge when applied to markedly heterogeneous data, underscoring the need for harmonization and domain-adaptation strategies in future studies.

Performance of the DL model and model interpretability

We evaluated 15 DL networks and ultimately selected ResNet101 because it demonstrated the most favorable balance between predictive accuracy and generalization performance. Among all evaluated architectures, ResNet101 achieved one of the highest validation AUCs while maintaining relatively small discrepancies between the training and validation ROC curves, indicating stable model behavior with limited overfitting (Additional file 1: Fig. S3). This DL model demonstrated a good ability to differentiate MIBC from NMIBC in the internal validation cohort with an AUC of 0.694 (95% CI 0.604–0.783) (Additional file 1: Fig. S7a), although its performance was slightly lower than that of the ML model. However, in the external multi-center validation cohort, the DL model demonstrated better generalization than the ML model, as reflected by a higher overall AUC (0.707 vs. 0.618) and less performance fluctuation across centers. Although the small sample size of some external cohorts contributed to variability in AUC estimates, the DL model nonetheless showed more consistent performance across sites. Notably, the DL model outperformed the ML model in the TCGA-BLCA cohort, indicating that ResNet101 effectively extracts deeper features from non-standardized CT imaging data from multiple centers, resulting in more accurate predictions. The ResNet101 model's AUC in the external multi-center cohort was similar to that in the internal validation cohort (Additional file 1: Fig. S7a), with an AUC of 0.707 (95% CI 0.659–0.755), demonstrating good generalizability and stable predictive capability across diverse clinical settings.

The distributions of risk scores generated by the DL model and the gradient-weighted class activation map (Grad-CAM) of the original image are shown in Fig. 3. The red-highlighted regions are primarily concentrated in the basal and tumor areas, indicating that the DL model focuses on these regions to extract valuable information.

Performance of the DLRN

Both the XGBoost ML model, which is based on handcrafted features, and the ResNet101 DL model, which relies on DL features, demonstrated unique strengths and weaknesses, excelling in different aspects of prediction. To enhance prediction accuracy and clinical utility, we developed a hybrid model that combines the advantages of both techniques. After final feature selection, 13 handcrafted features were retained for constructing the Rad_Signature (Radimoics_Signature), and 8 DL features (derived from DCA-reduced features) were selected for constructing the DL_Signature. By integrating handcrafted and DL features through LASSO logistic regression, *t*-tests, and Pearson correlation coefficients, a Combined_Signature was constructed using 13 handcrafted features and 3 DL features. Additionally, we developed a clinical model (Clinic) in which muscle invasion was predicted using patients' clinicopathological factors, including demographic information, medical history, and pathological tumor characteristics.

SHapley Additive exPlanations (SHAP) interpretability analysis ranked feature importance based on average SHAP values. The results indicated that the DL_0 feature had the highest weight in influencing the model's output, while the handcrafted features had the greatest overall impact on the model's predictions (Fig. 4a, b). These findings highlight that the DLRN successfully integrates the strengths of both the DL model and radiomic models.

Both theRad model and the DL model outperformed the clinical model (Clinic) in the internal validation cohort and the external multi-center validation cohort (Fig. 4c; Additional file 1: Table S4). In the training cohort, the DLRN model demonstrated superior performance, achieving an AUC of 0.895 (95% CI 0.871–0.919), sensitivity of 0.8918, and

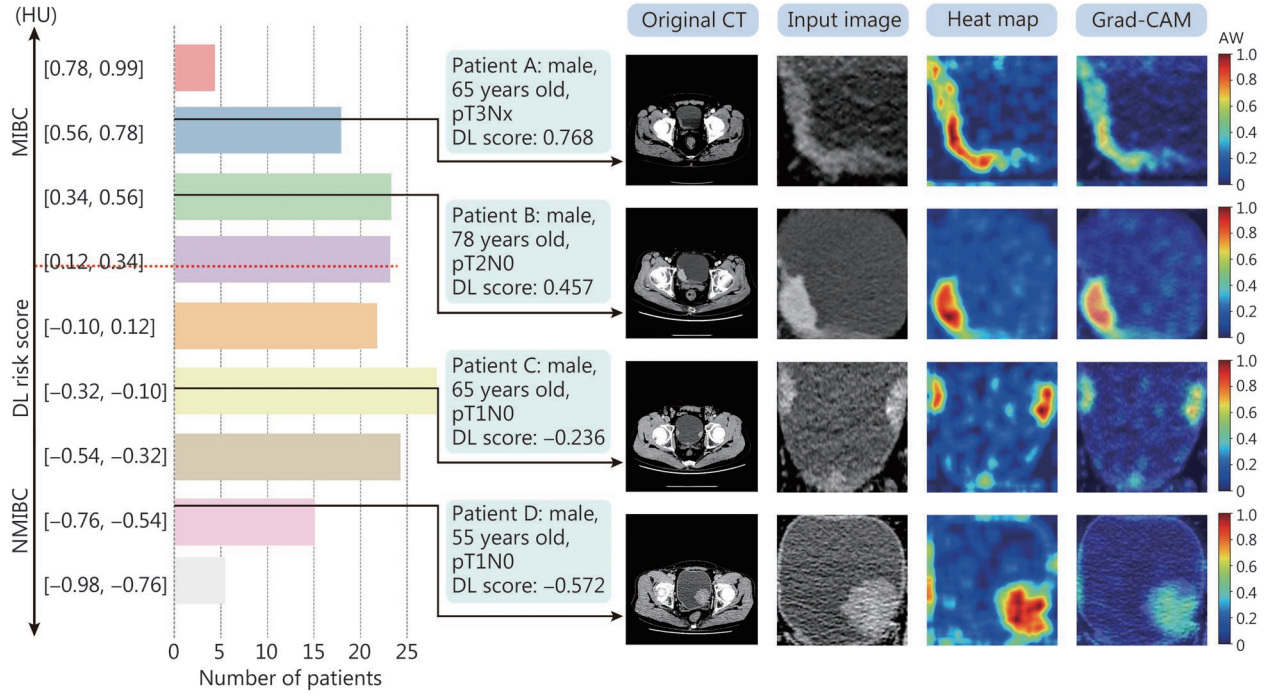

Fig. 3 Distribution of feature risk scores calculated by the deep learning (DL) model in the training cohort and representative CT image examples.

The left panel shows a histogram of risk score distributions generated by the DL model, with the red dashed line representing the cut-off value (0.283). Original CT: the raw CT image containing the tumor with the maximum diameter. Input image: the cropped CT image used as input for the model. Heat map: the heat map generated by the DL model based on Grad-CAM. Grad-CAM: the gradient-weighted class activation map corresponding to the input image. The attention-based DL model focuses predominantly on the tumor base and the entire tumor region within the image. CT. Computed tomography; NMIBC. Non-muscle-invasive bladder cancer; MIBC. Muscle-invasive bladder cancer; HU. Hounsfield unit; AW. Attention weight

specificity of 0.8496. The positive predictive value (PPV) was 0.8318, and the negative predictive value (NPV) was 0.9040. The F1-score was 0.8608, indicating an excellent balance between precision and recall. In the internal validation cohort, the DLRN model achieved the highest overall performance, with an AUC of 0.807 (95% CI 0.772–0.843), sensitivity of 0.8687, specificity of 0.6523, PPV of 0.7042, and NPV of 0.8427. The F1-score was 0.7651, reflecting a favorable balance between precision and recall. In comparison, the Clinic model yielded an AUC of 0.559 with lower overall discriminability (sensitivity: 0.6522, specificity: 0.3983, F1-score: 0.5012). The Rad model achieved moderate performance (AUC: 0.751, sensitivity: 0.6707, specificity: 0.7455, F1-score: 0.6469), while the DL model showed a higher sensitivity of 0.8746 but lower specificity of 0.5787 (F1-score: 0.6888). In the external multi-center validation cohort, the DLRN model maintained the highest discriminative ability with an AUC of 0.783 (95% CI 0.743–0.822), sensitivity of 0.8176, and specificity of 0.6405. The PPV and NPV were 0.7052 and 0.7380, respectively, resulting in an F1-score of 0.7573. In comparison, the Clinic model showed limited discriminability with an AUC of 0.550 (sensitivity: 0.4781, specificity: 0.4263, F1-score: 0.4856). The Rad model achieved moderate performance with an AUC of 0.648 (specificity: 0.6379, sensitivity: 0.4826, F1-score: 0.6130), while the DL model yielded an AUC of 0.707 (specificity: 0.4648, sensitivity: 0.7426, F1-score: 0.6750). These results demonstrate the superiority of the DLRN in capturing complementary clinical, radiomic, and deep semantic imaging features, resulting in significantly improved predictive accuracy for MIBC (Fig. 4c; Additional file 1: Table S4). The model also demonstrated improved effectiveness in individual external centers (center 2–6, TCGA-BLCA) (Additional file 1: Fig. S7b, Table S5).

Finally, we constructed a nomogram combining the hybrid model score (Combined) with clinicopathological characteristics (e.g., gender, BMI, and medical history) to predict the 1-, 3-, and 5-year survival rates for BLCA patients (Additional file 1: Figs. S8-S9, Table S6). All patients were stratified into a high-risk group (n=735) and a low-risk group (n=635) based on the DLRN score (cut-off value=0.283) from the training cohort. The model demonstrated strong performance in the training cohort

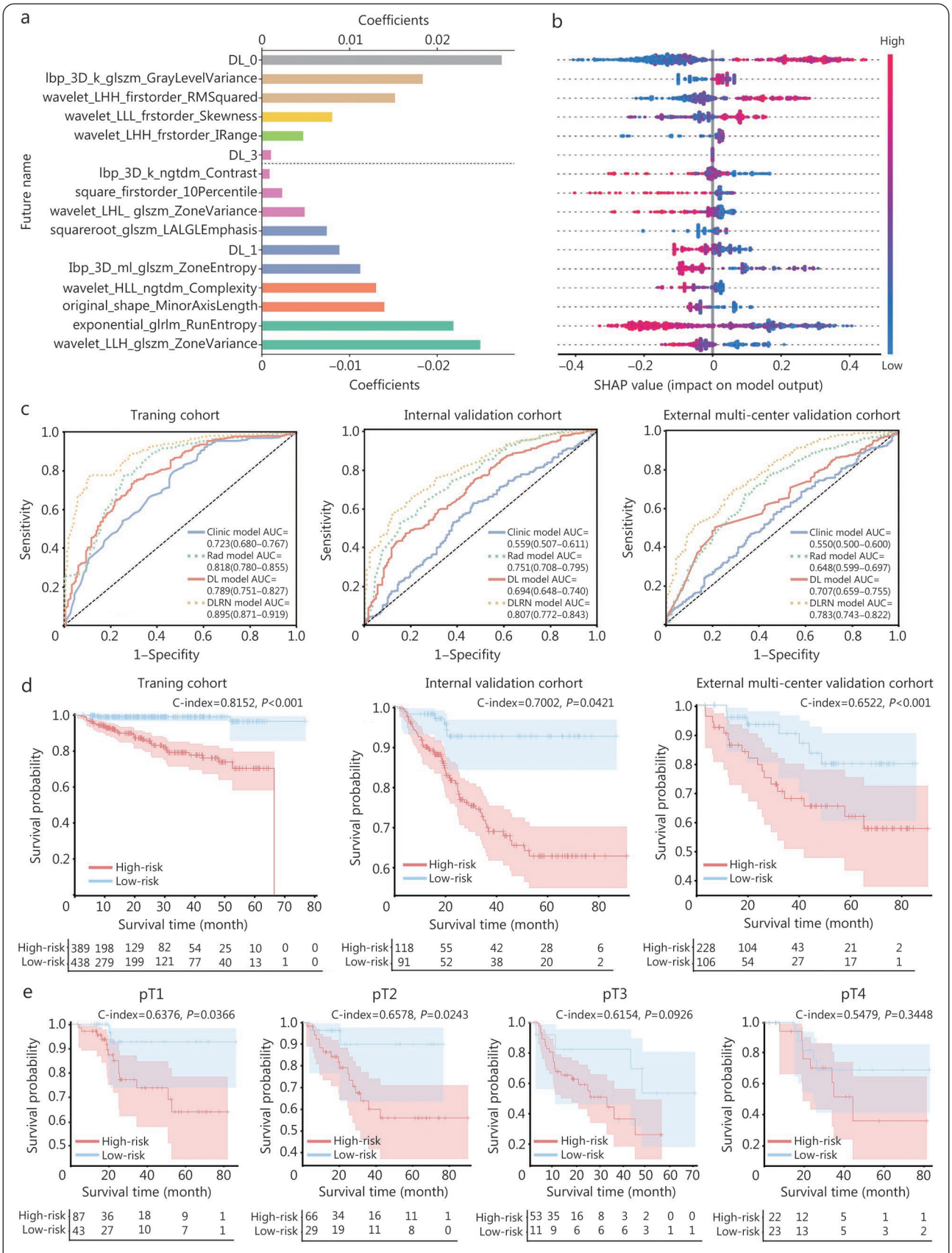

Fig. 4 (See legend on next page.)

(See figure on previous page.)

Fig. 4 SHapley Additive exPlanations (SHAP) interpretability analysis and model performance of deep learning radiomics nomogram (DLRN).

a Overall impact weights of 13 handcrafted features and 3 deep learning (DL) features on the model's output. **b** Distribution of SHAP values across samples, illustrating the specific contribution of each feature to the model's output. **c** AUCs of different models in the training cohort, internal validation cohort, and external multi-center validation cohort. **d** Kaplan-Meier curves for overall survival (OS) in the training cohort, internal validation cohort, and external multi-center validation cohort. **e** Kaplan-Meier curves for OS in patients with pT1–pT4 stages from the external validation cohort. Patients are stratified into high-risk (red line) and low-risk (blue line) groups based on DLRN risk scores, with a cut-off value of 0.283. AUC. Area under the curve; Rad. Radiomics; DL. Deep learning

(C-index=0.8152, P<0.001), internal validation cohort (C-index=0.7002, P=0.0421), and external multi-center validation cohort (C-index=0.6522, P<0.001). Patients in the high-risk group had significantly worse OS than those in the low-risk group (Fig. 4d). Within the subgroups of patients staged as pathological T stage (pT1–pT4), DLRN showed better predictive performance for patients at the pT1 stage (C-index=0.6376, P=0.0366) and the pT2 stage (C-index=0.6578, P=0.0243). For the pT3 (C-index=0.6154, P=0.0926) and pT4 (C-index=0.5479, P=0.3448) subgroups, although statistical significance was not reached, the observed trends suggested potential prognostic utility (Fig. 4e). These results indicate that larger, adequately powered cohorts will be necessary to confirm the value of DLRN in advanced-stage patients.

AI assists doctors in diagnosing MIBC and NMIBC

Three junior urologists and three senior urologists independently reviewed the original imaging data of these patients and provided preliminary diagnoses for MIBC. The diagnostic performance of the AI model and the 6 urologists is illustrated in Fig. 5a and b. Junior urologists demonstrated a limited ability to differentiate between MIBC and NMIBC, whereas senior urologists exhibited moderate diagnostic capability. However,

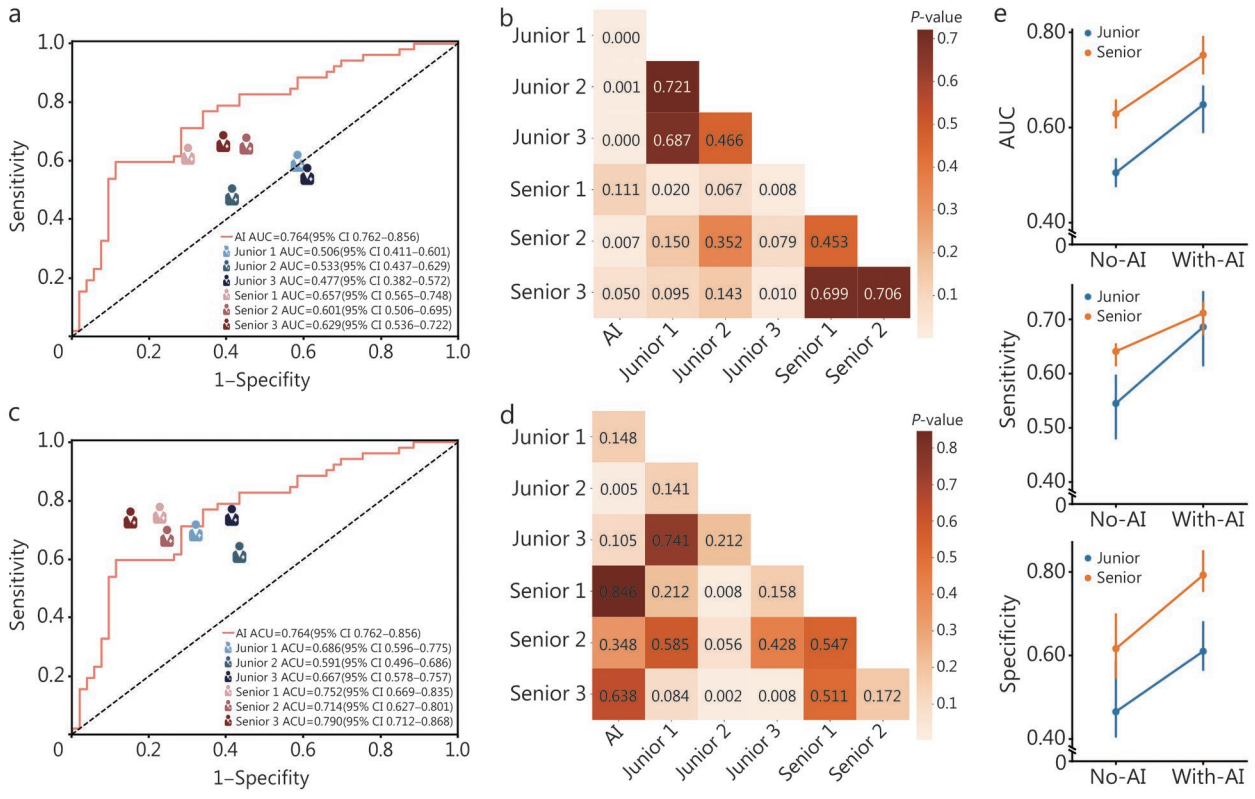

Fig. 5 Comparison of diagnostic performance: human, AI, and AI-assisted human.

a AUC comparison between the first diagnoses by 6 doctors and the AI diagnosis. **b** De-long test results comparing the first diagnoses by 6 doctors with the AI diagnosis. **c** AUC comparison between the 6 doctors' second diagnoses, after reviewing AI results, and the AI diagnosis. **d** De-long test results comparing AI-assisted diagnoses with standalone AI diagnoses. **e** Changes in accuracy, sensitivity, and specificity for the 6 doctors' second diagnoses after reviewing AI results. AI. Artificial intelligence

the AI model significantly outperformed both groups. With the assistance of risk scores and visualized heatmaps generated by the AI model, the diagnostic performance of all 6 urologists improved significantly (Fig. 5c, d). Notably, junior urologists showed greater improvement in diagnostic capability, particularly in sensitivity (Fig. 5e). Moreover, senior urologists achieved higher diagnostic accuracy using AI-assisted results compared to relying solely on the AI model.

To ensure the robustness and clinical relevance of our evaluation, we initially invited 3 junior urologists and 3 senior urologists to independently interpret the CT scans and diagnose MIBC based on their clinical experience and surgical judgment. This choice was based on the fact that, unlike magnetic resonance imaging (MRI), there is currently no universally accepted radiological scoring system for assessing muscularis propria invasion on CT. Therefore, in real-world clinical settings, CT-based assessments of MIBC heavily rely on the urologists' clinical expertise and integrated evaluation of tumor size, shape, interface with the bladder wall, and signs of extravesical extension.

To further validate our findings and address potential subjectivity, we additionally recruited 6 board-certified radiologists (3 juniors and 3 seniors) to independently review the same CT scans, both with and without AI assistance (Additional file 1: Fig. S10). This dual-reader design, incorporating both urologists and radiologists, ensures comprehensive and multidisciplinary validation of the AI model's assistive value in real clinical workflows.

A detailed error analysis identified 5 cases that were correctly diagnosed by the AI model but misdiagnosed by all 3 senior urologists. Two MIBC cases were misdiagnosed as NMIBC due to subtle lesions, while 3 NMIBC cases were misdiagnosed as MIBC due to large, poorly defined lesions (Additional file 1: Fig. S11a, b). Interestingly, 5 cases misdiagnosed by the AI model were correctly diagnosed by all 3 senior urologists, including two cases where catheter-related artifacts in the bladder caused errors (Additional file 1: Fig. S11c).

Discussion

The DLRN model developed in this study demonstrated high diagnostic performance, with AUCs of 0.807 (95% CI 0.772–0.843) in the internal validation cohort and 0.783 (95% CI 0.743–0.822) in the external multi-center validation cohort. These results underscore the potential of combining DL technology with traditional radiomic features to enhance diagnostic accuracy. Currently, TURBT is the preferred treatment for NMIBC patients, whereas MIBC requires more aggressive treatment strategies, such as RC with pelvic lymph node dissection or combined neoadjuvant therapy. Accurate preoperative assessment of BLCA muscle invasion is therefore critical for determining appropriate treatment strategies [28].

Early diagnosis of BLCA primarily relies on CT and MRI. A total of 2909 patients were included in the initial cohort from 6 centers. The vast majority (n=2249) received preoperative CT, whereas only 221 underwent MRI, confirming CT as the primary imaging technique. However, the limited soft-tissue resolution of CT hampers accurate differentiation between NMIBC and MIBC, often leaving clinicians dependent on experience. Recently, the MRI-based vesical imaging reporting and data system (VI-RADS) has shown good accuracy for staging [29,30], but its clinical use is constrained by cost, scan time, and contraindications. Thus, improving CT-based assessment is of particular importance, as enhanced diagnostic precision could optimize treatment planning, reduce inappropriate management, and improve outcomes.

Radiomics combined with AI has been increasingly applied in oncology, showing value in diagnosis, recurrence monitoring, and prognosis prediction [31-36]. In BLCA, prior CT- or MRI-based radiomics studies achieved promising AUCs [22,37-39], but most were limited to small, single-center cohorts and lacked external validation, restricting generalizability. In contrast, our study leveraged a large multi-center cohort (n=1370) with both internal and external validation, including a prospective component. The DLRN maintained strong performance despite heterogeneity, achieving an AUC of 0.895 (95% CI 0.871–0.919), sensitivity of 0.8918, and specificity of 0.8496, underscoring its robustness and clinical potential. To contextualize our findings, we compared the DLRN with existing methods (Additional file 1: Table S7). MRI-based VI-RADS achieves good accuracy for MIBC (AUC=0.87) but is constrained by cost, contraindications, and scan time [29]. Prior radiomics models also reported AUCs of 0.78–0.90 but were based on single-center datasets with limited validation [33,37,39]. By contrast, our DLRN achieved comparable or superior accuracy in a large multi-center setting, with substantially improved generalizability. Moreover, unlike invasive cystoscopy, our model relies on routine CT imaging, highlighting its potential as a practical, non-invasive decision-support tool.

Our study prospectively compared human and AI diagnostic performance, showing that AI-assisted decision-making significantly improved clinicians' accuracy in distinguishing MIBC from NMIBC. Unlike prior work on standalone AI models [40,41], this highlights the complementary role of AI in reducing diagnostic variability and enhancing sensitivity,

particularly for less experienced clinicians. These findings underscore the clinical value of human-AI collaboration in achieving more consistent diagnoses. Interestingly, in our study, AI assistance benefited urologists more than radiologists. This discrepancy likely reflects baseline expertise: radiologists already achieved moderate accuracy in pelvic CT interpretation, leaving limited room for improvement, whereas urologists, who typically integrate imaging with surgical context rather than detailed radiological assessment, gained substantial benefit from AI outputs, especially in sensitivity. Clinically, this suggests AI may have its greatest immediate impact as a decision-support tool for urologists, while for radiologists, it may function more effectively as a second-reader system to reduce oversight rather than dramatically increase accuracy.

While our study successfully developed AI models for predicting BLCA muscle invasion and prognosis, several limitations should be acknowledged. The first concerns imaging heterogeneity across institutions. Variations in scanners, contrast agents, and acquisition protocols inevitably influenced model performance across centers. We tried to minimize this effect through voxel resampling, intensity normalization, GAN-based super-resolution, and stringent feature selection to reduce overfitting [42]. Even so, variability remained, reminding us that prospective studies with standardized imaging protocols will be essential. A second limitation lies in the algorithms themselves. The DL backbones we employed, such as ResNet101 and GoogLeNet, were originally developed for natural image recognition [43,44]. While their transferability is impressive, they are not fully optimized for CT imaging. We believe future progress will depend on architectures specifically designed for medical applications, which may better capture the nuances of BLCA CT. Third, the generalizability of the study population is another important consideration. Most of our training and validation cohorts came from Chinese academic centers, representing largely East Asian patients. The inclusion of TCGA-BLCA, which primarily involves North American patients, provided a preliminary test of cross-ethnic applicability. Yet, a limited number of cases cannot stand in for true diversity. Broader validation in large, multi-ethnic international cohorts remains necessary to confirm generalizability across diverse genetic and healthcare contexts.

Beyond BLCA, the methodology used in this study has the potential to be extended to other tumor types. Since the proposed framework integrates handcrafted radiomics features with DL representations extracted from CT images, it can, in principle, be adapted to different malignancies. With the application of transfer learning, pretrained models on BLCA CT data could be fine-tuned using relatively smaller datasets from other tumor types, thereby reducing the need for large-scale annotation and training. Such adaptability underscores the translational relevance of our approach and paves the way for broader applications in oncology imaging research.

Despite harmonization attempts such as voxel resampling, intensity normalization, and GAN-based super-resolution, residual heterogeneity in imaging and clinical variables almost certainly persisted. This serves as a reminder that harmonization can help, but will not completely erase inter-center differences. Standardized prospective imaging protocols remain the gold standard.

The poor performance in the TCGA-BLCA cohort (AUC=0.519) deserves special attention. We identified a clear distribution shift in radiomic and DL features compared with our training data (Additional file 1: Fig. S6), the markedly earlier time of data acquisition in TCGA-BLCA may also have contributed to this decline. Differences in imaging protocols, scanner generations, and clinical workflows across acquisition periods are known to introduce systematic domain shifts, which could impair model generalizability. In my view, this is one of the fundamental challenges for AI in medicine. Domain adaptation strategies and inclusion of multi-regional, multi-ethnic cohorts will be necessary to improve robustness.

Conclusions

The proposed CT-based DLRN demonstrated reliable performance in predicting muscle invasion and OS in patients with BLCA. By integrating DL with handcrafted radiomic features, the model provides a non-invasive and interpretable tool for preoperative risk stratification as well as individualized clinical decision-making.

Abbreviations

AI: Artificial intelligence
AJCC: American Joint Committee on Cancer
AUC: Area under the curve
BLCA: Bladder cancer
BMI: Body mass index
CI: Confidence interval
CT: Computed tomography
DCA: Dimension correlation analysis
DL: Deep learning
DLRN: Deep learning radiomics nomogram
GAN: Generative adversarial network
Grad-CAM: Gradient-weighted Class Activation Mapping
HR: Hazard ratio
ICCs: Intraclass correlation coefficients
KNN: k-nearest neighbors
LASSO: Least absolute shrinkage and selection operator
LightGBM: Light gradient boosting machine
LR: Logistic regression

MIBC: Muscle-invasive bladder cancer
ML: Machine learning
MLP: Multilayer perceptron
MRI: Magnetic resonance imaging
NMIBC: Non-muscle-invasive bladder cancer
OS: Overall survival
RC: Radical cystectomy
ROC: Receiver operating characteristic
ROI: Region of interest
SD: Standard deviation
SHAP: SHapley Additive exPlanations
SVM: Support vector machine
TCGA: The Cancer Genome Atlas
TURBT: Transurethral resection of bladder tumor
VI-RADS: Vesical imaging reporting and data system
XGBoost: Extreme gradient boosting

Supplementary information

The online version contains supplementary material available at https://doi.org/10.1016/j.mmr.2026.100001.

Additional file 1. Methods. Table S1 The CT protocols of the 6 centers. **Table S2** Hyperparameters for the deep learning model. **Table S3** Hyperparameters for machine learning models. **Table S4** Performance evaluation of models. **Table S5** Quantitative assessment of feature distribution shift between the multi-center cohorts and TCGA-BLCA cohort. **Table S6** Baseline multivariate Cox proportional hazards model of clinicopathological variables for overall survival. **Table S7** Comparison of methods. **Fig. S1** Comparison of model performance before and after GAN-based super-resolution. **Fig. S2** Performance evaluation of 8 machine learning (ML) algorithms. **Fig. S3** Performance evaluation of 15 deep learning models. **Fig. S4** Construction of a radiomic model and score for predicting muscle invasion in bladder cancer (BLCA). **Fig. S5** The AUC of the XGBoost model in each validation cohort. **Fig. S6** Principal component analysis (PCA) visualization of feature distribution shift between multi-center (center 1–6) and TCGA-BLCA cohorts. **Fig. S7** The AUC of the ResNet101 model (a) and the combined model (b) in each validation cohort. **Fig. S8** Forest plot of multivariate Cox regression analysis for overall survival in bladder cancer patients. **Fig. S9** A nomogram to predict 1, 3, and 5-year survival after bladder cancer surgery. **Fig. S10** Performance comparison of radiologists with and without AI assistance. **Fig. S11** Representative cases.

Acknowledgements

The results shown here are in part based upon data generated by the TCGA Research Network: https://www.cancergenome.nih.gov/. The authors would like to acknowledge that the manuscript was polished using the AI-based language editing service (AJE's Automated Grammar Check Tool, https://www.aje.com/). Special thanks go to Ms. Juliet Matsika (matsikajuliet@yahoo.com) for her professional language polishing and editing of the manuscript. Her meticulous work has greatly improved the clarity and quality of the writing. We acknowledge that all authors are participating in this study for data collection, preparation, and quality control.

Authors' contributions

YBH and JH contributed equally to this work and are regarded as co-first authors. BYF, DSD, and XBZ are co-corresponding authors. YBH, JH, BYF, DSD, and XBZ conceived and designed the study. ZL, ZCX, JHL, HSL, WZD, ZWL, JZ, JQL, NG, BH, and XG contributed to data collection and curation. ZYO, JBC, PHL, MFC, HHL, RZW, XG, SYT, and YLL performed image processing and model development. WH, YHZ, ZYC, YG, CZ, YC, YQD, YC, and ZYN conducted statistical analysis and visualization. YBH and JH drafted the manuscript, while BYF, DSD, and XBZ critically revised it. XBZ supervised and administered the overall project. All authors read and approved the final manuscript.

Funding

This work was supported by the Noncommunicable Chronic Diseases-National Science and Technology Major Project (2024ZD0525700), the National Natural Science Foundation of China (81902592, 82070785, 82303760, 82373337), the China Postdoctoral Innovation Talents Support Program (BX20230431), the China Postdoctoral Science Foundation (2023M733951), the Hunan Natural Science Foundation (2023JJ40946, 2024JJ2093), the Hunan Provincial Key Area Research Plan (2023SK2016), the Hunan Province Young Talents Program (2023RC3073), the Changsha Natural Science Foundation (kq2208377), the Youth Science Foundation of Xiangya Hospital (2022Q20), the National Clinical Research Center for Geriatric Disorders (2022LNJJ04), the Central South University Research Programme of Advanced Interdisciplinary Studies (2023QYJC029), the Health Research Project of Hunan Provincial Health Commission (202204054669), and the Scientific Research Program of FuRong Laboratory (2024PT5102).

Availability of data and materials

All data generated from this study are available upon request to the corresponding author.

Declarations

Ethics approval and consent to participate

This study was reviewed and approved by the Clinical Research Ethics Committee of Xiangya Hospital, Central South University (202304064), and written informed consent was not required from patients due to the retrospective nature.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of Urology, Xiangya Hospital, Central South University, Changsha 410008, China. [2]National Clinical Research Center for Geriatric Disorders, Xiangya Hospital, Central South University, Changsha 410008, China. [3]Department of Urology, the Second Affiliated Hospital, Guizhou Medical University, Guiyang 550000, China. [4]Department of Pathology, the Third Xiangya Hospital of Central South University, Changsha 410013, China. [5]Department of Urology, the Second Hospital of University of South China, Hengyang 421001, Hunan, China. [6]Department of Imaging, the First People's Hospital of Kaili, Kaili 556000, Guiyang, China. [7]Department of Imaging, the Second Affiliated Hospital, Guizhou Medical University, Kaili 556000, Guiyang, China. [8]Department of Urology, Xiangya Boai Rehabilitation Hospital, Changsha 410146, China. [9]Department of Urology, Hunan Provincial People's Hospital/the First Affiliated Hospital of Hunan Normal University, Changsha 410005, China.

References

1. Zi H, Liu MY, Luo LS, Huang Q, Luo PC, Luan HH, *et al*. Global burden of benign prostatic hyperplasia, urinary tract infections, urolithiasis, bladder cancer, kidney cancer, and prostate cancer from 1990 to 2021. Mil Med Res. 2024;11(1):64.
2. Siegel RL, Miller KD, Wagle NS, Jemal A. Cancer statistics, 2023. CA Cancer J Clin. 2023;73(1):17-48.
3. Mancini M, Righetto M, Zumerle S, Montopoli M, Zattoni F. The bladder EpiCheck test as a non-invasive tool based on the identification of DNA methylation in bladder cancer cells in the urine: a review of published evidence. Int J Mol Sci. 2020;21(18): 6542.
4. Kamoun A, De Reyniès A, Allory Y, Sjödahl G, Robertson AG, Seiler R, *et al*. A consensus molecular classification of muscle-invasive bladder cancer. Eur Urol. 2020;77(4):420-33.
5. Tran L, Xiao JF, Agarwal N, Duex JE, Theodorescu D. Advances in bladder cancer biology and therapy. Nat Rev Cancer. 2021;21(2): 104-21.
6. Schafer EJ, Jemal A, Wiese D, Sung H, Kratzer TB, Islami F, *et al*. Disparities and trends in genitourinary cancer incidence and mortality in the USA. Eur Urol. 2023;84(1):117-26.
7. Messina E, Proietti F, Laschena L, Flammia RS, Pecoraro M, Cipollari S, *et al*. MRI for risk stratification of muscle invasion by upper tract urothelial carcinoma: a feasibility study. Eur Radiol Exp. 2024;8(1):9.
8. Miyake M, Hirao S, Mibu H, Tanaka M, Takashima K, Shimada K, *et al*. Clinical significance of subepithelial growth patterns in non-muscle invasive bladder cancer. BMC Urol. 2011;11:17.
9. Peng HT, Siddiqui MM, Rhind SG, Zhang J, da Luz LT, Beckett A. Artificial intelligence and machine learning for hemorrhagic trauma care. Mil Med Res. 2023;10(1):6.
10. Hamet P, Tremblay J. Artificial intelligence in medicine. Metabolism. 2017;69S:S36-S40.
11. Van Der Velden BHM, Kuijf HJ, Gilhuijs KGA, Viergever MA. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. Med Image Anal. 2022;79:102470.
12. Stahlschmidt SR, Ulfenborg B, Synnergren J. Multimodal deep learning for biomedical data fusion: a review. Brief Bioinform. 2022;23(2):bbab569.
13. Suarez-Ibarrola R, Hein S, Reis G, Gratzke C, Miernik A. Current and future applications of machine and deep learning in urology: a review of the literature on urolithiasis, renal cell carcinoma, and bladder and prostate cancer. World J Urol. 2020;38(10):2329-47.
14. Li M, Jiang Z, Shen W, Liu H. Deep learning in bladder cancer imaging: a review. Front Oncol. 2022;12:930917.
15. Arendt CT, Leithner D, Mayerhoefer ME, Gibbs P, Czerny C, Arnoldner C, *et al*. Radiomics of high-resolution computed tomography for the differentiation between cholesteatoma and middle ear inflammation: effects of post-reconstruction methods in a dual-center study. Eur Radiol. 2021;31(6):4071-8.
16. Martini K, Baessler B, Bogowicz M, Blüthgen C, Mannil M, Tanadini-Lang S, *et al*. Applicability of radiomics in interstitial lung disease associated with systemic sclerosis: proof of concept. Eur Radiol. 2021;31(4):1987-98.
17. Mühlbauer J, Egen L, Kowalewski KF, Grilli M, Walach MT, Westhoff N, *et al*. Radiomics in renal cell carcinoma-a systematic review and meta-analysis. Cancers (Basel). 2021;13(6):1348
18. Tang Y, Li S, Zhu L, Yao L, Li J, Sun X, *et al*. Improve clinical feature-based bladder cancer survival prediction models through integration with gene expression profiles and machine learning techniques. Heliyon. 2024;10(20):e38242.
19. Xiong S, Fu Z, Deng Z, Li S, Zhan X, Zheng F, *et al*. Machine learning-based CT radiomics enhances bladder cancer staging predictions: a comparative study of clinical, radiomics, and combined models. Med Phys. 2024;51(9):5965-77.
20. She Y, He B, Wang F, Zhong Y, Wang T, Liu Z, *et al*. Deep learning for predicting major pathological response to neoadjuvant chemoimmunotherapy in non-small cell lung cancer: a multi-centre study. EBioMedicine. 2022;86:104364.
21. Jiang Y, Zhang Z, Yuan Q, Wang W, Wang H, Li T, *et al*. Predicting peritoneal recurrence and disease-free survival from CT images in gastric cancer with multitask deep learning: a retrospective study. Lancet Digit Health. 2022;4(5):e340-e50.
22. Zhang G, Xu L, Zhao L, Mao L, Li X, Jin Z, *et al*. CT-based radiomics to predict the pathological grade of bladder cancer. Eur Radiol. 2020;30(12):6749-56.
23. Kirk SLY, Lucchesi FR, Aredes ND, Gruszauskas N, Catto J, Garcia K, *et al*. The Cancer Genome Atlas Urothelial Bladder Carcinoma Collection (TCGA-BLCA) (Version 8). The Cancer Imaging Archive; 2016. https://doi.org/10.7937/K9/TCIA.2016.8LNG8XDR.
24. Jiang Y, Liang X, Han Z, Wang W, Xi S, Li T, *et al*. Radiographical assessment of tumour stroma and treatment outcomes using deep learning: a retrospective, multicohort study. Lancet Digit Health. 2021;3(6):e371-e82.
25. In H, Solsky I, Palis B, Langdon-Embry M, Ajani J, Sano T. Validation of the 8th edition of the AJCC TNM staging system for gastric cancer using the National Cancer Database. Ann Surg Oncol. 2017;24(12):3683-91.
26. You C, Li G, Zhang Y, Zhang X, Shan H, Li M, *et al*. CT super-resolution GAN constrained by the identical, residual, and cycle learning ensemble (GAN-CIRCLE). IEEE Trans Med Imaging. 2020; 39(1):188-203.
27. Guerreiro J, Tomás P, Garcia N, Aidos H. Super-resolution of magnetic resonance images using generative adversarial networks. Comput Med Imaging Graph. 2023;108:102280.
28. Denlinger CS, Sanft T, Baker KS, Baxi S, Broderick G, Demark-Wahnefried W, *et al*. Survivorship, Version 2.2017, NCCN Clinical Practice Guidelines in Oncology. J Natl Compr Canc Netw. 2017; 15(9):1140-63.
29. Ueno Y, Tamada T, Takeuchi M, Sofue K, Takahashi S, Kamishima Y, *et al*. VI-RADS: multi-institutional multireader diagnostic accuracy and interobserver agreement study. AJR Am J Roentgenol. 2021; 216(5):1257-66.
30. Panebianco V, Narumi Y, Altun E, Bochner BH, Efstathiou JA, Hafeez S, *et al*. Multiparametric magnetic resonance imaging for bladder cancer: development of VI-RADS (Vesical Imaging-Reporting And Data System). Eur Urol. 2018;74(3):294-306.
31. Bhinder B, Gilvary C, Madhukar NS, Elemento O. Artificial intelligence in cancer research and precision medicine. Cancer Discov. 2021;11(4):900-15.
32. Bi WL, Hosny A, Schabath MB, Giger ML, Birkbak NJ, Mehrtash A, *et al*. Artificial intelligence in cancer imaging: clinical challenges and applications. CA Cancer J Clin. 2019;69(2):127-57.
33. Wu S, Zheng J, Li Y, Yu H, Shi S, Xie W, *et al*. A radiomics nomogram for the preoperative prediction of lymph node metastasis in bladder cancer. Clin Cancer Res. 2017;23(22):6904-11.
34. Zhang GM, Sun H, Shi B, Jin ZY, Xue HD. Quantitative CT texture analysis for evaluating histologic grade of urothelial carcinoma. Abdom Radiol (NY). 2017;42(2):561-8.
35. Lucas M, Jansen I, Van Leeuwen TG, Oddens JR, De Bruin DM,

Marquering HA. Deep learning-based recurrence prediction in patients with non-muscle-invasive bladder cancer. Eur Urol Focus. 2022;8(1):165-72.
36. Wang H, Zhang M, Miao J, Hou F, Chen Y, Huang Y, *et al*. Deep learning signature based on multiphase enhanced CT for bladder cancer recurrence prediction: a multi-center study. E Clinical Medicine. 2023;66:102352.
37. Zhang G, Wu Z, Zhang X, Xu L, Mao L, Li X, *et al*. CT-based radiomics to predict muscle invasion in bladder cancer. Eur Radiol. 2022;32(5):3260-8.
38. Garapati SS, Hadjiiski L, Cha KH, Chan HP, Caoili EM, Cohan RH, *et al*. Urinary bladder cancer staging in CT urography using machine learning. Med Phys. 2017;44(11):5814-23.
39. Wang H, Xu X, Zhang X, Liu Y, Ouyang L, Du P, *et al*. Elaboration of a multisequence MRI-based radiomics signature for the preoperative prediction of the muscle-invasive status of bladder cancer: a double-center study. Eur Radiol. 2020;30(9):4816-27.
40. Borhani S, Borhani R, Kajdacsy-Balla A. Artificial intelligence: a promising frontier in bladder cancer diagnosis and outcome prediction. Crit Rev Oncol Hematol. 2022;171:103601.
41. Ren Y, Wang G, Wang P, Liu K, Liu Q, Sun H, *et al*. MM-SFENet: multi-scale multi-task localization and classification of bladder cancer in MRI with spatial feature encoder network. Phys Med Biol. 2024;69(2).
42. Mayerhoefer ME, Materka A, Langs G, Häggström I, Szczypiński P, Gibbs P, *et al*. Introduction to Radiomics. J Nucl Med. 2020;61(4):488-95.
43. Haennah JHJ, Christopher CS, King GRG. Prediction of the COVID disease using lung CT images by deep learning algorithm: DETS-optimized Resnet 101 classifier. Front Med (Lausanne). 2023;10:1157000.
44. Balagourouchetty L, Pragatheeswaran JK, Pottakkat B, G R. GoogLeNet-Based Ensemble FCNet Classifier for Focal Liver Lesion Diagnosis. IEEE J Biomed Health Inform. 2020;24(6):1686-94.

https://doi.org/10.1016/j.mmr.2026.100001
Cite this article as: He YB, Hu J, Liu Z, Xiao ZC, Liu JH, Liang HS, *et al*. Non-invasive evaluation of muscle invasion and survival prognosis in bladder cancer using enhanced CT-based deep learning radiomics: a multi-center real-world cohort study. *Mil Med Res*. 2026;13(1):100001.