npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-01723-x

Enhanced staging of renal cell carcinoma using tumor morphology features: model development and multi-source validation

Enyu Yuan[1,5], Yuntian Chen[1,5], Lei Ye[1,5], Ben He[2], ChunLei He[1,3], Junchao Ma[1], Ting Yang[1], Hao Zeng[4], Ling Yang[1,6] ✉, Jin Yao[1,6] ✉ & Bin Song[1,3,6] ✉

Preoperative detection of pT3a invasion in non-metastatic renal cell carcinoma (RCC) remains challenging with CT. This study developed and validated radiomic models using preoperative CT to identify pT3a invasions. Six models were trained and internally validated via nested cross-validation on 999 patients from one hospital. External validation included 313 patients from two hospitals and 204 patients from four TCIA datasets. A multi-reader multi-case study with seven radiologists evaluated the model's incremental value. The morphology model achieved the highest internal AUC (0.867, 95% CI: 0.866–0.869) and maintained performance in external validations (AUC = 0.895 and 0.842). When used as a second reader, it significantly improved junior radiologists' sensitivity and discrimination (AUC: 0.790 vs. 0.831, $p < 0.001$) without compromising specificity. This study demonstrates that CT-based radiomic models, particularly the morphology model, can reliably detect pT3a invasion and enhance diagnostic accuracy for junior radiologists, offering potential clinical utility in preoperative staging.

Renal cell carcinoma (RCC), the most common kidney cancer in adults, accounts for roughly 3% of all cancers[1]. For patients with non-metastatic RCC, surgical resection remains the primary and potentially curative treatment. Accurate preoperative staging is essential to assess recurrence risk and determine the appropriate surgical approach. Locally-advanced RCCs, staged over pT3a, refer to RCCs that invade into surrounding structures including perinephric fat, renal sinus fat, renal vein, or pelvicalyceal system[2]. These RCCs are more susceptible to metastasis and recurrence, and radical nephrectomy (RN) is required to ensure complete resection of the tumor due to the poor prognosis. Conversely, localized RCCs that are confined to the renal parenchyma can be managed with partial nephrectomy (PN), allowing for complete tumor removal while preserving renal function and reducing the incidence of long-term complications.

CT has been recommended as the first-line diagnostic and staging tool for RCC by European Association of Urology, National Comprehensive Cancer Network, and American Urological Association guidelines. Despite its widespread use, CT shows variable accuracy in distinguishing between tumor stages. The reported accuracy for differentiate pT3–pT4 from pT1–pT2 ranged from 0.70 to 0.97[3–6], while the accuracy for identifying the invasion of perinephric fat, renal sinus fat, renal vein, and pelvicalyceal system varied from 0.348 to 0.915, 0.354 to 0.937, 0.597 to 0.95, and 0.481 to 0.932, respectively[6–10]. These inconsistencies underscore the need for more reliable methods to identify invasive characteristics.

Radiomics is a mathematic approach that can encode the image information within the region of interest into quantitative morphology, intensity, and texture features, allowing an in-depth characterization of the image. Previous studies suggest that radiomics models performs well in prediction of nuclear grade and cancer-specific survival in patients with RCC[11,12]. However, few studies have investigated the value of radiomics in staging RCC[13–15]. These studies were single-center designed and the findings were limited by the small sample size. Large sample size with multi-center external validation studies are required to establish the robustness of the radiomics approach for identifying the T3a invasions.

Therefore, the aim of our study is to develop radiomic models based on preoperative CT for identifying pT3a invasions in RCC and to test the models in internal and external datasets against a pathologic reference

[1]Department of Radiology, West China Hospital, Sichuan University, Chengdu, 610041, China. [2]Department of Urology, The Third People's Hospital of Chengdu/The Affiliated Hospital of Southwest Jiaotong University, Chengdu, 610014, China. [3]Department of Radiology, Sanya People's Hospital, Sanya, China. [4]Department of Urology, West China Hospital, Sichuan University, Chengdu, 610041, China. [5]These authors contributed equally: Enyu Yuan, Yuntian Chen, Lei Ye. [6]These authors jointly supervised this work: Ling Yang, Jin Yao, Bin Song. ✉e-mail: florialinda@163.com; yaojin@wchscu.edu.cn; songlab_radiology@163.com

standard. We also seek to assess the added value of radiomics as a second reader to aid radiologists in a multi-reader multi-case study setup (Fig. 1).

Results

Baseline characteristics

A total of 1516 eligible patients from three centers and four public datasets were included in this study. Among these, 999 patients (340 with pathological stage T3a or higher) were from the development dataset; 313 patients (45 with pathological stage T3a or higher) were from the bi-center validation dataset; and 204 patients (73 with pathological stage T3a or higher) were from the TCIA dataset. The demographic and clinical characteristics of the patients are summarized in Table 1.

Feature statistics

The Dice Similarity Coefficient between the two segmentations was 0.928 ± 0.070 in data of center A. Among 1470 tumor features, the intraclass correlation coefficient (ICC) was 0.951 ± 0.108, while 1403 features had ICC > 0.8. Among 1456 peritumor features, the ICC was 0.955 ± 0.051, while 1426 features had ICC > 0.8. The modeling pipeline finally selected four, five, seven, five, five, and four features for M, IT, M-IT, PIT, M-PIT, and M-IT-PIT models, respectively. The frequency of being selected in nested cross-validation and the feature coefficient in logistic regression model were depicted in Fig. 1. In M model, Maximum 3D Diameter exhibited the highest absolute coefficient (1.448), while Sphericity showed the highest absolute coefficient in M-IT, M-PIT, M-IT-PIT model (−1.242, −1.052, and −1.055). Within the IT model, the Gray Level Run Length Matrix derived tumor Gray Level Non-Uniformity after applying exponential filter had the highest coefficient in IT model (0.703), while the Gray Level Run Length Matrix derived peritumor Gray Level Non-Uniformity after applying square filter had the highest coefficient in PIT model (0.686). Feature selection pathway and final features were detailed in Supplementary Tables 1 and 2.

Internal validation

Nested cross-validation was employed for internal validation, with the joint distribution of training and test AUC values over 1000 outer loops displayed in Fig. 2. The figure indicates minimal overfitting concerns, as training and test AUCs were closely aligned along the counter-diagonal. The mean performance in 1000 outer loops were summarized in Table 2. Among the pipelines, the M pipeline achieved the highest mean test AUC (0.867 [0.866–0.869]), followed by the M-IT-PIT pipeline (0.865 [0.864–0.867], $p < 0.001$), M-PIT pipeline (0.865 [0.864–0.867], $p < 0.001$), M-IT pipeline (0.864 [0.863–0.866], $p < 0.001$), PIT pipeline (0.824 [0.822–0.826], $p < 0.001$), and IT pipeline (0.823 [0.821–0.824], $p < 0.001$). The calibration curve demonstrated excellent consistency between predicted and observed probabilities across all six pipelines (Supplementary Fig. 1). Decision curve analysis also showed that each pipeline provided greater net benefit than both the 'treat-all' (RN for all) and 'treat-non' (PN for all) strategies (Supplementary Fig. 1).

External validation

The performance of each model was presented in Table 3 and Fig. 3. In the bi-center validation dataset, the M-PIT model had the highest AUC (0.911 [0.867–0.956]), but the differences were minimal and non-significant between the AUCs of M model (0.895 [0.840–0.949], $p = 0.09$), IT model (0.887 [0.837–0.937], $p = 0.15$), M-IT model (0.908 [0.860–0.955], $p = 0.69$), PIT model (0.902 [0.854–0.950], $p = 0.57$), and M-IT-PIT model (0.911 [0.863–0.958], $p = 0.90$). The calibration curve showed slightly under-calibration for all models. The calibration curve indicated slight under-calibration across all models, while the decision curve showed all models offered better clinical benefit than 'treat-none' or 'treat-all' approaches.

In the TCIA validation dataset, the M model again had the highest AUC (0.842 [0.789–0.895]), with similar performance observed for the IT model (0.816 [0.759–0.873], $p = 0.05$), M-IT model (0.841 [0.789–0.894], $p = 0.90$), M-PIT model (0.829 [0.774–0.884], $p = 0.18$), and M-IT-PIT model (0.831 [0.777–0.886], $p = 0.08$). The PIT model had the lowest AUC (0.806 [0.747–0.865], $p = 0.02$) compared to the M model. The calibration curve showed good consistency between the predicted probability and observed probability for all models. The decision curve also showed all the model had favorable clinical benefits than either the 'treat-none' or 'treat-all' strategies.

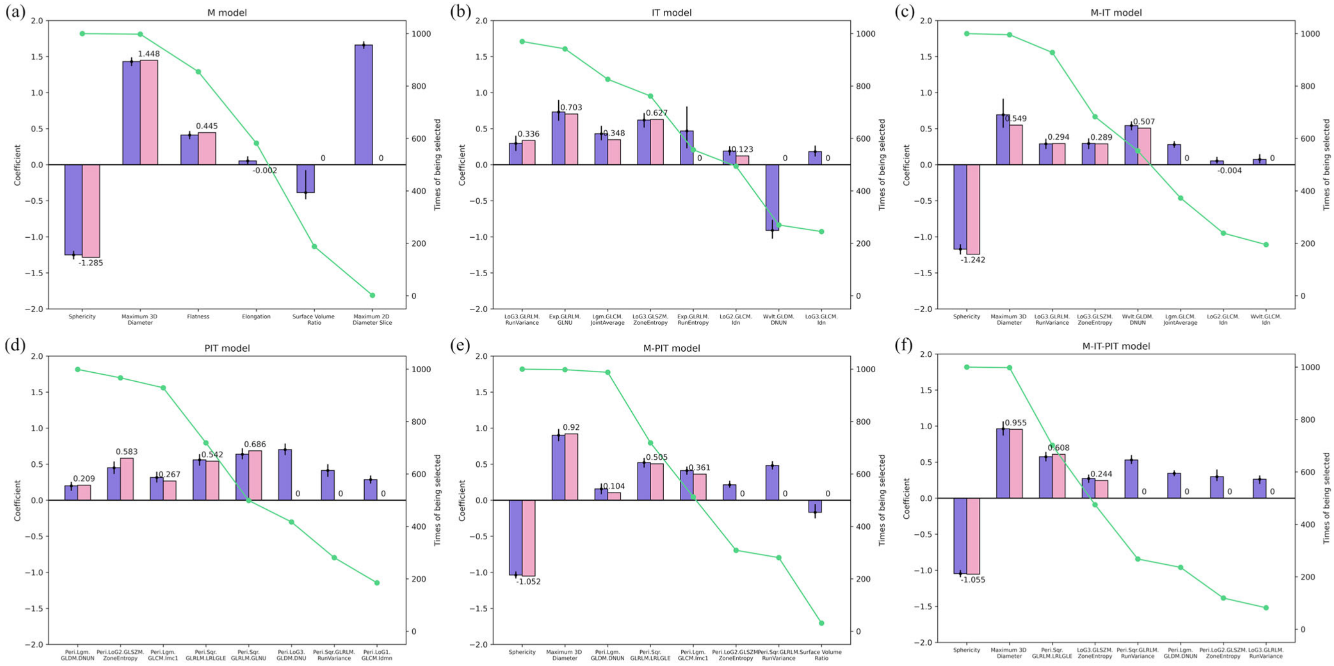

Fig. 1 | Feature importance of the radiomic models. The purple bar denotes the median feature coefficient along with lower and upper quartiles in the logistic regression model in 1000 outer loops. The green line denotes the frequency of being selected as the final feature in 1000 outer loops. The pink bar denotes the feature coefficient in the final model. **a** Morphology model, **b** tumor intensity and texture model, **c** morphology and tumor intensity and texture model, **d** peritumor intensity and texture model, **e** morphology and peritumor intensity and texture model, **f** morphology, tumor and peritumor intensity and texture model.

Table 1 | Baseline patient characteristics

	Development dataset (n = 999)			Bi-center validation dataset (n = 313)			TCIA validation dataset (n = 204)		
	T3a−	T3a +	p Value	T3a−	T3a +	P Value	T3a−	T3a +	P Value
Age (y)[a]	54.9 ± 13.1 (19–86)	60.4 ± 11.9 (24–87)	<0.001	57.7 ± 12.7 (22–89)	62.7 ± 12.4 (21–85)	0.007	57.4 ± 12.2 (26–86)	62.6 ± 10.5 (34–85)	0.001
Sex			0.40			0.58			0.39
Male	432 (65.6)	232 (68.2)		167 (62.3)	30 (66.7)		92 (70.2)	47 (64.4)	
Female	227 (34.4)	108 (31.8)		101 (37.7)	15 (33.3)		39 (29.8)	26 (35.6)	
Operation			<0.001			<0.001			/
PN	374 (56.8)	20 (5.9)		180 (67.2)	2 (4.4)		NA	NA	
RN	285 (43.2)	320 (94.1)		88 (32.8)	43 (95.6)		NA	NA	
Subtype			0.06			0.02			0.16
ccRCC	492 (74.7)	276 (81.2)		227 (84.7)	37 (82.2)		115 (87.8)	70 (95.9)	
pRCC	22 (3.3)	12 (3.5)		20 (7.5)	2 (4.4)		10 (7.6)	2 (2.7)	
chRCC	40 (6.1)	10 (2.9)		12 (4.5)	0 (0.0)		6 (4.6)	1 (1.4)	
other RCC[b]	105 (15.9)	42 (12.4)		9 (3.4)	6 (13.3)		0 (0.0)	0 (0.0)	
ISUP Grade			<0.001			<0.001			<0.001
G1	31 (4.7)	1 (0.3)		71 (26.5)	1 (2.2)		2 (1.5)	1 (1.4)	
G2	392 (59.5)	93 (27.4)		121 (45.1)	10 (22.2)		64 (48.9)	15 (20.5)	
G3	125 (19.0)	149 (43.8)		39 (14.6)	15 (33.3)		39 (29.8)	33 (45.2)	
G4	14 (2.1)	61 (17.9)		8 (3.0)	14 (31.1)		9 (6.9)	21 (28.8)	
NA	97 (14.7)	36 (10.6)		29 (10.8)	5 (11.1)		17 (13.0)	3 (4.1)	
Pathologic T Stage			<0.001			<0.001			<0.001
pT1	617 (93.6)	0 (0.0)		263 (98.1)	0 (0.0)		109 (83.2)	0 (0.0)	
pT2	42 (6.4)	0 (0.0)		5 (1.9)	0 (0.0)		22 (16.8)	0 (0.0)	
pT3	0 (0.0)	321 (94.4)		0 (0.0)	40 (88.9)		0 (0.0)	69 (94.5)	
pT4	0 (0.0)	19 (5.6)		0 (0.0)	5 (11.1)		0 (0.0)	4 (5.5)	

Unless otherwise noted, data are numbers of participants and data in parentheses are the percentages.

CIA The Cancer Imaging Archive, *PN* partial nephrectomy, *RN* radical nephrectomy, *ccRCC* clear cell renal cell carcinoma, *pRCC* papillary renal cell carcinoma, *chRCC* chromophobe renal cell carcinoma, *ISUP* International Society of Urological Pathology.

[a] Data are means ± SDs; data in parentheses are the range.

[b] Other RCC includes: collecting duct carcinoma, mucinous tubular and spindle cell carcinoma, eosinophilic solid and cystic renal cell carcinoma, TFE3 rearranged renal cell carcinomas, TFEB altered renal cell carcinomas, ELOC mutated renal cell carcinoma, Fumarate hydratase deficient renal cell carcinoma, Succinate dehydrogenase deficient renal cell carcinoma, renal medullary carcinoma, and renal cell carcinoma NOS.

The M model exhibited the lowest coefficient of variation (0.030) for AUCs across the cross-validation, bi-center validation, and TCIA validation datasets (Supplementary Fig. 2). In clear cell RCC (ccRCC) subgroup, all six models performed well across development, bi-center validation, and TCIA validation datasets (AUC > 0.8). Performance of the IT, PIT, and M-PIT models slightly decreased in non-ccRCC subgroup, though the differences were not statistically significant (Supplementary Table 3). The sensitivity analysis demonstrated that removal of individual model did not significantly reduce model performance, suggesting robustness of the morphological features stable across CT scanners (Supplementary Figs. 3 and 4).

Clinical evaluation

Table 4 and Fig. 4 show the performance of individual radiologists, their combined average, and the model-augmented performance. Individual AUCs ranged from 0.780 (0.715–0.845) to 0.865 (0.818–0.913), with an overall radiologist average AUC of 0.810 (0.757–0.863). The average AUC of senior radiologists was 0.861 (0.820–0.902), while the average AUC of junior radiologists was 0.790 (0.740–0.839). When assisted by the M model as a second reader, junior radiologists showed a notable improvement (AUC: 0.790 [0.740–0.839] vs. 0.831 [0.784–0.878], p < 0.001). Using a cutoff score of 3 (scores of three, four, and five indicate positive T3a invasion), the model significantly improved the sensitivity of the junior readers (from 0.759 [0.628–0.890] to 0.841 [0.760–0.922], p = 0.03) without sacrificing specificity (0.698 [0.591–0.804] vs. 0.719 [0.628–0.810], p = 0.19). The average AUC of senior radiologists plus M model was similar to senior radiologists alone (0.861 [0.820–0.902] vs. 0.864 [0.821–0.908], p = 0.60).

Discussion

Identifying T3a invasion in RCC before surgery is crucial for risk assessment and surgical planning. In our study, we developed and validated six radiomic models using nephrogenic phase CT images to detect T3a invasion. Among these, the morphology model, leveraging four visually interpretable features, achieved the highest AUC of 0.867 (0.866–0.869) in internal validation. It also showed stable AUCs of 0.895 (0.840–0.949) and 0.842 (0.789–0.895) across two external validation datasets, with a coefficient of variation of 0.030. Furthermore, the clinical evaluation indicated that the morphology model, when used as a second reader, enhanced junior radiologists' diagnostic performance, improving both sensitivity (0.759 [0.628–0.890] vs. 0.841 [0.760–0.922], p = 0.03) and overall discrimination (0.790 [0.740–0.839] vs. 0.831 [0.784–0.878], p < 0.001).

Previous research on CT-based machine learning models for RCC staging is limited. Wang et al. used dual-energy CT images from 200 patients to construct radiomic models, achieving AUCs of 0.64–0.97 during training but only 0.61–0.66 in testing, reflecting suboptimal performance of the models[13]. In another study, Orton et al. evaluated radiomic models in 101 consecutive ccRCC patients, finding a 'shape' model can differentiate overall

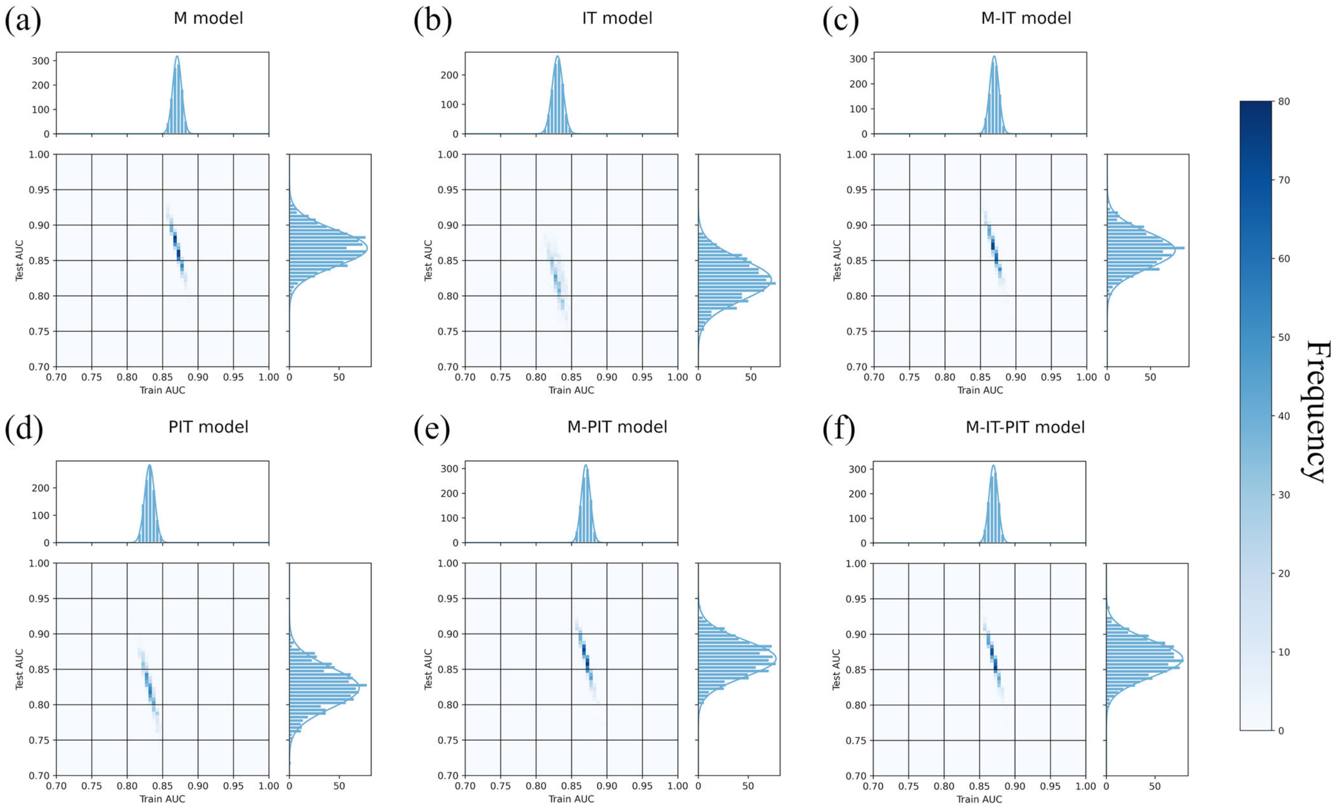

Fig. 2 | The joint distribution of area under receiver operating characteristic curve of cross validation training and test sets in 1000 outer loops. a Morphology model, **b** tumor intensity and texture model, **c** morphology and tumor intensity and texture model, **d** peritumor intensity and texture model, **e** morphology and peritumor intensity and texture model, **f** morphology, tumor and peritumor intensity and texture model.

TNM stage 1/2 and 3/4 with an AUC of 0.893[14]. Demirjian et al. reported AUC values of 0.8 in training and 0.77 in external validation for a CT-based radiomic model[15]. However, in their study, the morphology features were not extracted and the modeling pipelines were not internally validated. In the study by Shimada et al., tumor shape irregularity, quantified by the radiomic feature SHAPE_Sphericity, showed strong impact in predicting pT3a upstaging in localized cT1b-cT2 RCCs[16]. Nevertheless, this finding was limited by single center design and relatively small sample size. These studies highlight the need for robust modeling pipelines and external validation.

Our study included data from 999 patients for model development and 517 for external validation, totaling 1516 cases. Internal validation employed nested cross-validation to minimize bias, and external validation used cases from two medical centers and four public datasets to assess generalizability. Although only the M model was selected for clinical evaluation, all six models displayed minimal overfitting in internal validation and strong generalizability in external validation (maximum coefficient of variation = 0.06 for the PIT model). Additionally, our study did not restrict the pathology subtype to ccRCC, but included as many subtypes as possible. Indeed, non-ccRCC tumors encompass a diverse group of indolent subtypes (e.g., chromophobe RCC) and aggressive subtypes (e.g., fumarate hydratase-deficient RCC). Subgroup analysis confirmed robust performance in ccRCC and non-ccRCC groups alike.

Image signs for T3a invasion in RCC have been extensively researched. For example, RCC tumors with irregular margin, rather than round or lobular margin, were found to be more likely to invade the perinephric fat and the renal sinus fat[10,17–23]. Consistent with this, our models identified ‘Sphericity’, a three-dimensional measure of tumor roundness relative to a sphere, as a key feature in identifying T3a invasion. Our results showed this feature was frequently select in nested cross-validation and exhibited a high coefficient in M, M-IT, M-PIT, and M-IT-PIT models. This finding aligns with a prior study[16]. This convergence not only validated tumor shape irregularity as a biomarker of invasive behavior—where reduced sphericity may indicate extra-renal growth—but also demonstrated radiomics’ ability to objectively quantify traditionally subjective assessments (i.e., ‘irregular margin’ to ‘Sphericity’), enabling standardized invasion evaluation. Additionally, the feature ‘3D maximum diameter’ also demonstrated high importance, consistent with previous findings[17,19,24]. The two features, along with ‘elongation’ and ‘flatness’, exhibited high reproducibility and interobserver agreement, reinforcing their value in minimizing variability in qualitative assessments. These four features formed the foundation of our morphology model, which improved sensitivity in junior radiologists while preserving specificity.

Our study had several limitations. First, given that perinephric stranding, sinus invasion, and other peritumor image signs can be challenging to capture with a uniformly dilated mask, targeting suspicious subregions (via attention-based or handcrafted segmentation) might increase performance of models incorporating peritumor radiomic features[3–10,17,18]. Second, while manual segmentation with high Dice coefficients ensured reproducible region-of-interest delineation in this study, its labor-intensive nature may hinder large-scale clinical adoption. Future research should explore integrating automated deep learning-based segmentation tools (e.g., nnU-Net) to improve workflow efficiency. Thirdly, the small sample sizes for individual pathologic subtypes in certain cohorts limited the statistical power and reliability of subtype-specific validation. Our model demonstrated stable performance between ccRCC and non-ccRCC groups, which preliminarily proves the robustness of the model. However, future prospective studies with larger subtype-specific cohorts are still needed to validate radiomic models for distinct RCC subtypes. Fourthly, incomplete follow-up data restricted our ability to assess the prognostic value of the radiomic models. Lastly, our clinical validation revealed the morphology model improved the performance of junior radiologists.

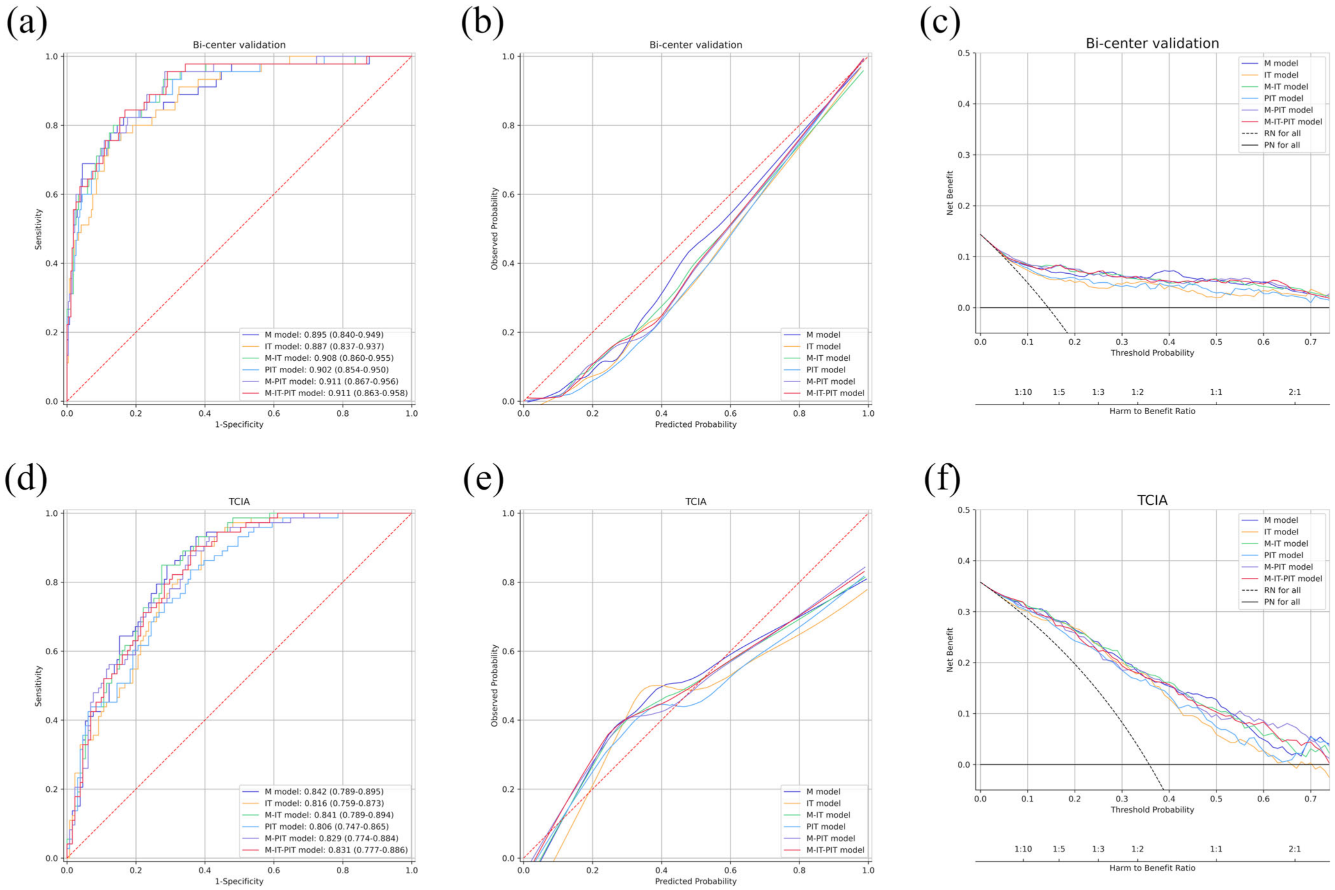

Fig. 3 | The discrimination, calibration, and clinical utility of the developed models in two external validation datasets. **a** The receiver operating characteristic (ROC) curves in bi-center validation dataset, **b** the calibration curves in bi-center validation dataset, **c** the decision curves in bi-center validation dataset, **d** the receiver operating characteristic (ROC) curves in The Cancer Imaging Archive (TCIA) validation dataset, **e** the calibration curves in TCIA dataset, and **f** the decision curves in TCIA dataset.

Table 2 | Internal validation performance

Model	Cross-validation Training AUC[a]	P Value	Cross-validation Test AUC[a]	P Value
M	0.871 (0.870–0.871)	/	0.867 (0.866–0.869)	/
IT	0.830 (0.830–0.831)	<0.001	0.823 (0.821–0.824)	<0.001
M-IT	0.869 (0.869–0.870)	<0.001	0.864 (0.863–0.866)	<0.001
PIT	0.832 (0.831–0.832)	<0.001	0.824 (0.822–0.826)	<0.001
M-PIT	0.870 (0.870–0.871)	<0.001	0.865 (0.864–0.867)	<0.001
M-IT-PIT	0.870 (0.869–0.870)	<0.001	0.865 (0.864–0.867)	<0.001

P values correspond to comparisons between the M model and the other models.
AUC area under the receiver operating characteristic curve.
[a]Data in parentheses are 95% CIs.

However, the lack of clear guidelines for adjusting original ratings after seeing the model's output hinders its clinical adoption, especially for senior radiologists who may be reluctant to change assessments without understanding the model's reasoning. To address this, model explanation methods such as Shapley Additive Explanation and Local Interpretable Model-agnostic Explanations can be used. These methods highlight high-impact radiomic features driving predictions, making the model's decisions more transparent and helping to align the model's rationale with the radiologists' expertise. Additionally, decision support systems analogous to the R.E.N.A.L. score can provide a framework for adjusting ratings based on model outputs (e.g., consider extra-renal invasion if perinephric fat stranding is present alongside a high model prediction). Moreover, our results showed variability in senior radiologists' diagnostic criteria (favoring high sensitivity or specificity). To further investigate these diverse diagnostic criteria and accurately estimate the average sensitivity and specificity, prospective MRMC studies with a larger sample of both senior and junior radiologists are needed.

In conclusion, using nephrogenic phase CT images, we developed and validated six radiomic models that reliably identified T3a invasion in renal cell carcinoma across internal and external validations. Integrating the morphology model as a second reader improved junior radiologists' sensitivity, enhancing the overall diagnostic performance. These models hold potential to identify cases with extra-renal invasion, aiding in treatment planning and potentially reducing unnecessary radical nephrectomy for localized renal cell carcinoma.

Methods

Study design

This article was prepared following the Checklist for Evaluation of Radiomics research (Supplemental Data 1)[25]. This retrospective study design was approved by the institutional review board of West China Hospital (Approval No. 2021-1333). Written informed consent was waived because of the retrospective nature of our study. The patient-sensitive information was removed from all type of data before analysis.

This study analyzed data collected from three tertiary academic hospitals and four public imaging datasets. The inclusions and exclusions criteria are summarized in Fig. 5. We reviewed consecutive patients who underwent nephrectomy at the West China Hospital (center A) from

Table 3 | External validation performance

Model	Cutoff	AUC[a]	P Value	Sensitivity[a]	P Value	Specificity[a]	P Value
Bi-center validation							
M-PIT	0.332	0.911 (0.867–0.956)	/	0.756 (0.630–0.881)	/	0.888 (0.850–0.926)	/
M	0.358	0.895 (0.840–0.949)	0.09	0.689 (0.554–0.824)	0.38	0.918 (0.885–0.951)	0.02
IT	0.306	0.887 (0.837–0.937)	0.15	0.778 (0.656–0.899)	>0.99	0.840 (0.796–0.883)	0.02
M-IT	0.399	0.908 (0.860–0.955)	0.69	0.644 (0.505–0.784)	0.06	0.922 (0.889–0.954)	0.01
PIT	0.352	0.902 (0.854–0.950)	0.57	0.800 (0.683–0.917)	0.69	0.847 (0.804–0.890)	0.05
M-IT-PIT	0.328	0.911 (0.863–0.958)	0.90	0.756 (0.630–0.881)	>0.99	0.888 (0.850–0.926)	>0.99
TCIA validation							
M	0.358	0.842 (0.789–0.895)	/	0.699 (0.593–0.804)	/	0.771 (0.699–0.843)	/
IT	0.306	0.816 (0.759–0.873)	0.05	0.781 (0.686–0.876)	0.11	0.702 (0.624–0.781)	0.01
M-IT	0.399	0.841 (0.789–0.894)	0.90	0.726 (0.624–0.828)	0.50	0.771 (0.699–0.843)	>0.99
PIT	0.352	0.806 (0.747–0.865)	0.02	0.712 (0.608–0.816)	>0.99	0.718 (0.640–0.795)	0.09
M-PIT	0.332	0.829 (0.774–0.884)	0.18	0.740 (0.639–0.840)	0.25	0.756 (0.682–0.829)	0.63
M-IT-PIT	0.358	0.831 (0.777–0.886)	0.08	0.740 (0.639–0.840)	0.25	0.733 (0.657–0.809)	0.06

P values correspond to comparisons between the model with the highest AUC and the other models. For bi-center validation dataset, the comparisons were between the M-PIT model and the other models. For TCIA validation dataset, the comparisons were between the M model and the other models. The sensitivity and specificity values were calculated according to the cutoff value.

AUC area under the receiver operating characteristic curve.

[a] Data in parentheses are 95% CIs.

January 2012 to May 2023, the West China Tianfu Hospital (center B) from March 2022 to September 2023, and the Third People's Hospital of Chengdu (center C) from January 2017 to September 2023. This study included patients who met the following criteria: aged beyond 18 years, underwent RN or PN, were diagnosed with pathology-confirmed RCC, and underwent preoperative CT within three months before surgery. Patients were excluded if they underwent any treatment for RCC before surgery, had inadequate documentation of key pathologic features (i.e., status of T3a invasions), or had suboptimal CT quality for analyses (e.g., incomplete images or severe artifacts). Data from center B and center C were combined to form a bi-center validation dataset.

In addition to the hospital data, we searched The Cancer Imaging Archive (TCIA) for public imaging collections of RCC. We identified four collections: CPTAC-CCRCC, TCGA-KIRC, TCGA-KIRP, and TCGA-KICH[26–29]. Patients from these collections were excluded if they had incomplete documentation of pathologic T stage, lacked preoperative axial nephrogenic-phase CT, or had suboptimal CT quality for analysis. All the eligible cases were combined create the TCIA dataset.

Using these datasets, we designated data from center A for developing and internally cross-validating the proposed models. The bi-center validation dataset and the TCIA dataset were employed for external validation to assess the generalizability of the models. All private data were not used in previous publications. To maintain transparency and reproducibility, we offer access to the source code, configuration files, and trained models through open-source platforms (https://github.com/DOCT-Y/RCC-T3a_Invasion-radiomics).

Image acquisition and preprocessing

All patients received unenhanced and CT scans at 30–40 s and 80–100 min post-intravenous contrast injection (corticomedullary and nephrogenic phases, respectively). Detailed scanner specifications and imaging protocols are available in Supplementary Table 4. For analysis, nephrogenic phase images were used, with voxel dimensions standardized to 0.75 × 0.75 × 5 mm via three-dimensional linear interpolation.

Outcome definition

The sampling and pathological diagnosis were performed by at least two urological pathologists who were aware of the clinical data. The status of T3a invasions were assessed in accordance with the International Society of Urological Pathology Consensus recommendations[30]. Tumors presenting with any form of invasion were categorized as invasion-positive, while those without any invasion were classified as invasion-negative.

Image annotation

The three-dimensional region of interest of tumor area was annotated using ITK-SNAP (version 3.8)[31]. For multifocal lesions, only the largest mass was considered for image analysis and segmentation. The annotators were blinded to the clinical, surgical, and pathological data. A urologist (B.H. with 12 years of experience) performed tumor annotation for all cases, while a radiologist (L.Y. with 4 years of experience) randomly re-annotated one third of the cases from center A to assess reproducibility of annotation and feature.

Annotations from center A were reviewed by two senior radiologists (J.Y. and L.Y., with 15 and 15 years of experience) Any discrepancies were resolved through consensus-based discussions to ensure consistent data, which were then used for model development. The tumor mask was dilated by 5 mm using Segment Editor module in 3D Slicer (version 5.2.2)[32], and the peritumor mask was the subtraction of the original mask from the dilated mask.

Feature extraction

Image Biomarker Standardization Initiative–compliant features were extracted from the segmentations using Pyradiomics (version 3.0.1)[33]. For the tumor mask, 14 morphology features were extracted. Then, the image filters were applied, including Laplacian of Gaussian filter (sigma = 1, 2, and 3), wavelet filter, square filter, square root filter, logarithm filter, and exponential filter. A set of 91 intensity-texture features were extracted from the original and filtered images, including: first order statistics (n = 18), gray level co-occurrence matrix (n = 22), gray level run length matrix (n = 16), gray level size zone matrix (n = 16), gray level dependence matrix (n = 14) and neighboring gray tone difference matrix (n = 5), resulting in a total of 1470 tumor features. For the peritumor mask, only intensity-texture features were extracted, resulting in a total of 1456 peritumor features. After providing all modified parameters of pre-processing and radiomic feature extraction, all other parameters remained as a default configuration.

Modeling pipeline

Using data from center A, we developed three single models and three combined models, all employing the same computational pipeline but

Table 4 | Clinical validation performance

Reader	AUC[a]		P Value	Sensitivity[a]		P Value	Specificity[a]		P Value
	Without AI	With AI		Without AI	With AI		Without AI	With AI	
All radiologist average	0.810 (0.757–0.863)	0.841 (0.794–0.888)	0.008	0.793 (0.678–0.907)	0.847 (0.778–0.917)	0.11	0.689 (0.585–0.793)	0.727 (0.649–0.806)	0.14
Senior radiologist average	0.861 (0.820–0.902)	0.864 (0.821–0.908)	0.60	0.877 (0.000–1.000)	0.863 (0.758–0.968)	0.88	0.668 (0.000–1.000)	0.748 (0.526–0.970)	0.499
Junior radiologist average	0.790 (0.740–0.839)	0.831 (0.784–0.878)	<0.001	0.759 (0.628–0.890)	0.841 (0.760–0.922)	0.03	0.698 (0.591–0.804)	0.719 (0.628–0.810)	0.19
Senior radiologist 1	0.865 (0.818–0.913)	0.869 (0.821–0.917)	0.70	0.781 (0.686–0.876)	0.836 (0.751–0.921)	0.22	0.794 (0.725–0.863)	0.794 (0.725–0.863)	>0.99
Senior radiologist 2	0.857 (0.808–0.906)	0.859 (0.811–0.908)	0.87	0.973 (0.935–1.010)	0.890 (0.819–0.962)	0.07	0.542 (0.457–0.627)	0.702 (0.624–0.781)	<0.001
Junior radiologist 1	0.791 (0.731–0.852)	0.843 (0.790–0.897)	0.02	0.616 (0.505–0.728)	0.740 (0.639–0.840)	0.049	0.794 (0.725–0.863)	0.817 (0.751–0.883)	0.55
Junior radiologist 2	0.780 (0.715–0.845)	0.817 (0.759–0.874)	0.03	0.767 (0.670–0.864)	0.849 (0.767–0.931)	0.03	0.687 (0.608–0.766)	0.687 (0.608–0.766)	>0.99
Junior radiologist 3	0.783 (0.719–0.847)	0.835 (0.781–0.888)	0.007	0.753 (0.655–0.852)	0.863 (0.784–0.942)	0.02	0.718 (0.640–0.795)	0.740 (0.665–0.816)	0.55
Junior radiologist 4	0.784 (0.724–0.845)	0.829 (0.773–0.885)	0.02	0.740 (0.639–0.840)	0.849 (0.767–0.931)	0.02	0.725 (0.649–0.802)	0.718 (0.640–0.795)	>0.99
Junior radiologist 5	0.809 (0.752–0.866)	0.832 (0.777–0.886)	0.07	0.918 (0.855–0.981)	0.904 (0.837–0.972)	>0.99	0.565 (0.480–0.650)	0.634 (0.551–0.716)	0.004

P values correspond to comparisons between radiologist(s) without the M model the radiologist(s) with the M model as second reader. The sensitivity and specificity values were calculated using scores larger than 2 as the cutoff value.

AI artificial intelligence, *AUC* area under the receiver operating characteristic curve.

[a] Data in parentheses are 95% CIs.

differing in the radiomic features used as input. These models include: 1) the Morphology (M) model, utilizing tumor morphology features; 2) the Tumor Intensity-Texture (IT) model, leveraging tumor intensity-texture features; 3) the Peritumor Intensity-Texture (PIT) model, incorporating peritumor intensity-texture features; 4) the Tumor Morphology-Intensity-Texture (M-IT) model, combining tumor morphology and intensity-texture features; 5) the Tumor Morphology and Peritumor Intensity-Texture (M-PIT) model, integrating tumor morphology and peritumor intensity-texture features; and 6) the Tumor Morphology-Intensity-Texture and Peritumor Intensity-Texture (M-IT-PIT) model, encompassing tumor morphology, tumor intensity-texture, and peritumor intensity-texture features.

The modeling was performed on scikit-learn (version 1.2.2)[34]. Model development followed a six-step process. First, features with zero variance were removed. The remaining features were standardized by median and interquartile range. Next, to reduce dimensionality, we retained the top 50 features based on Analysis of Variance F-values, with an exception for the M model, which originally had only 14 shape features. Highly correlated feature pairs (Pearson correlation > 0.85) were analyzed, and the feature with the lower F-value was excluded. The remaining features were entered into a Least Absolute Shrinkage and Selection Operator regression model, retaining only those with non-zero weights. The optimal lambda was determined through five-fold cross-validation, repeated 20 times. Logistic regression with an L2 penalty was applied using these final features, with the cutoff for positive samples set by maximizing the Youden index on the development data.

Model validation

The performance was evaluated by area under receiver operating characteristic curve (AUC, with 95% confidence interval) and confusion-matrix-derived statistics. Calibration and decision curves provided visual insights into model calibration and clinical applicability. We used nested cross-validation for internal validation: a five-fold cross-validation, repeated 20 times, optimized lambda in the inner loop, while a five-fold cross-validation, repeated 200 times, evaluated model performance in the outer loop[35]. All six models then underwent external validation. The model with the highest mean test AUC during internal cross validation was selected for the subsequent clinical evaluation.

To explore performance across subgroups, data were further stratified by pathological subtype, and metrics were calculated for each subgroup to gauge model robustness. To assess the model sensitivity to different scanner models, a one-at-a-time sensitivity analysis was also performed on the two external validation cohorts.

Clinical evaluation

To evaluate the radiomic model in clinical practice, a multi-reader multi-case study using case-by-case sequential design was conducted with images from the TCIA dataset. Two senior abdominal radiologists and five junior abdominal radiologists independently reviewed nephrogenic phase images (see Supplementary Table 5). The reviewers knew all patients had been diagnosed with RCC but were blinded to other clinical, surgical, and pathological details.

For each case, the reviewers initially rated the likelihood of T3a invasion on a five-point Likert scale based on imaging features described in prior studies. Afterward, they were shown the model's predictions and then asked to adjust their invasion likelihood ratings. The following signs were suggestive of perinephric fat invasion: a) perinephric fat stranding, b) perinephric fat soft-tissue nodules, c) perinephric vascularity, d) ill-defined tumor margin, and e) irregular tumor margin. The following signs were suggestive of renal sinus fat invasion: a) tumor abutted or bulged into the renal sinus, b) finger-like projection, and c) irregular tumor margin. The following signs were suggestive of renal vein invasion: a) venous enlargement with tumor thrombus enhancement and b) intraluminal filling defect in a segmental or main renal vein. The following signs were suggestive of

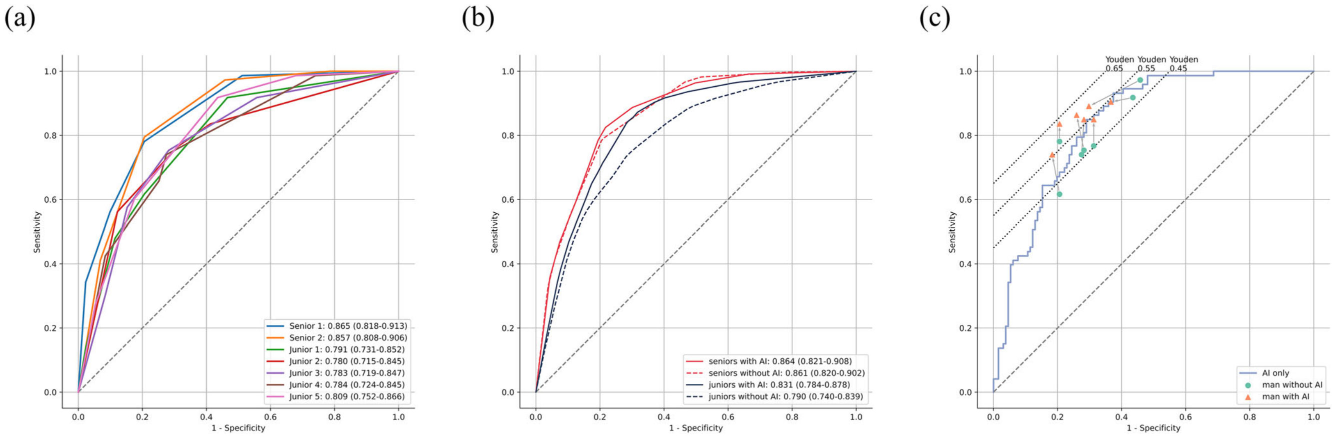

Fig. 4 | Key results of clinical evaluation. a The receiver operating characteristic (ROC) curves of individual radiologists, **b** the average ROC curves for senior and junior radiologists, and **c** using a cutoff score of 2, the change in sensitivity and specificity of each radiologists were depicted. AI artificial intelligence.

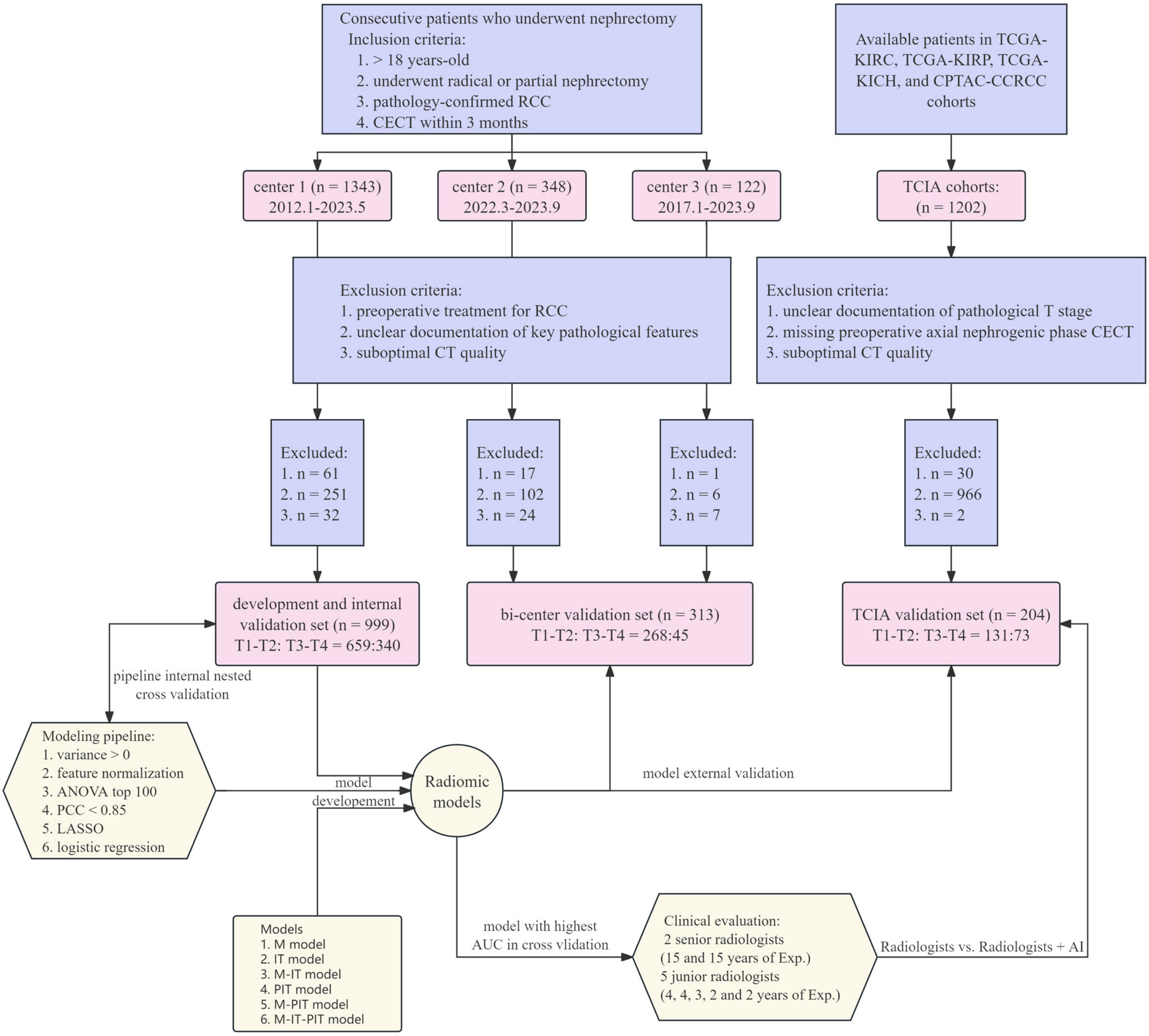
Consecutive patients who underwent nephrectomy
Inclusion criteria:
1. > 18 years-old
2. underwent radical or partial nephrectomy
3. pathology-confirmed RCC
4. CECT within 3 months
Available patients in TCGA-KIRC, TCGA-KIRP, TCGA-KICH, and CPTAC-CCRCC cohorts
center 1 (n = 1343) 2012.1-2023.5
center 2 (n = 348) 2022.3-2023.9
center 3 (n = 122) 2017.1-2023.9
TCIA cohorts: (n = 1202)
Exclusion criteria:
1. preoperative treatment for RCC
2. unclear documentation of key pathological features
3. suboptimal CT quality
Exclusion criteria:
1. unclear documentation of pathological T stage
2. missing preoperative axial nephrogenic phase CECT
3. suboptimal CT quality
Excluded: 1. n = 61 2. n = 251 3. n = 32
Excluded: 1. n = 17 2. n = 102 3. n = 24
Excluded: 1. n = 1 2. n = 6 3. n = 7
Excluded: 1. n = 30 2. n = 966 3. n = 2
development and internal validation set (n = 999) T1-T2: T3-T4 = 659:340
bi-center validation set (n = 313) T1-T2: T3-T4 = 268:45
TCIA validation set (n = 204) T1-T2: T3-T4 = 131:73
pipeline internal nested cross validation
Modeling pipeline:
1. variance > 0
2. feature normalization
3. ANOVA top 100
4. PCC < 0.85
5. LASSO
6. logistic regression
model developement
Radiomic models
model external validation
Models
1. M model
2. IT model
3. M-IT model
4. PIT model
5. M-PIT model
6. M-IT-PIT model
model with highest AUC in cross vlidation
Clinical evaluation:
2 senior radiologists
(15 and 15 years of Exp.)
5 junior radiologists
(4, 4, 3, 2 and 2 years of Exp.)
Radiologists vs. Radiologists + AI

Fig. 5 | Flowchart of patient selection at the three institutions and four public datasets. RCC renal cell carcinoma, CECT contrast-enhanced computed tomography, TCIA The Cancer Imaging Archive, ANOVA analysis of variance, PCC Pearson correlation coefficient, LASSO Least absolute shrinkage and selection operator, M morphology, IT intensity and texture, PIT peritumor intensity and texture, Exp experience, AI artificial intelligence.

pelvicalyceal system invasion: a) tumor reaches renal pelvis and b) ill-defined tumor margin.

Statistical analysis

For model development, the samples were included as much as possible, so there was no need for sample size estimation. For model performance validation, the sample size was calculated to determine the minimum number of cases required to evaluate the diagnostic performance of the test using receiver operating characteristic curve analysis. The parameters were as follows: The null hypothesis was that the AUC is 0.5, while the alternative hypothesis was that the AUC is not equal to 0.5. The significance level was set at 0.05, and the power of the study was set at 0.9. The expected AUC was 0.8. Additionally, the ratio of negative to positive cases was set at 3:1 according to previous studies[13,14]. The sample size was calculated employing the method described by Hanley and McNeil[36]. To ensure an adequate number of subjects in each group, we doubled the sample size to account for potential attrition or incomplete data. As a result, a total of 96 subjects of each dataset were determined to be necessary to achieve the desired statistical power.

For baseline characteristics, differences were compared with the Student t test or Mann-Whitney U test for continuous variables and with the $\chi 2$ Test or Fisher exact test for categorical variables, when applicable. During internal validation, pipeline metrics were compared using paired Student's t-tests or Wilcoxon signed-rank tests. For external validation, the AUCs for different models were compared using Delong's test[37]. The sensitivity and specificity were compared using McNemar's test[38]. Clinical evaluation results were averaged and assessed using the multi-reader multi-case method[39,40]. All analyses were performed by E.Y.Y., with seven years of programming and statistical analysis experience, using Python (versions 3.7.16 and 3.8.10) and R (version 4.2.1). The significance level was set at 0.05.

Data availability

The datasets generated or analyzed during the study are available from the corresponding author on reasonable request.

Code availability

To maintain transparency and reproducibility, we offer access to the source code, configuration files, and trained models through open-source platforms (https://github.com/DOCT-Y/RCC-T3a_Invasion-radiomics).

Received: 6 March 2025; Accepted: 15 May 2025;
Published online: 24 May 2025

References

1. Capitanio, U. et al. Epidemiology of renal cell carcinoma. *Eur. Urol.* **75**, 74–84 (2019).
2. Ljungberg, B. et al. European Association of Urology Guidelines on renal cell carcinoma: the 2022 update. *Eur. Urol.* **82**, 399–410 (2022).
3. Johnson, C. D., Dunnick, N. R., Cohan, R. H. & Illescas, F. F. Renal adenocarcinoma: CT staging of 100 tumors. *Am. J. Roentgenol.* **148**, 59–63 (1987).
4. Türkvatan, A. et al. Preoperative staging of renal cell carcinoma with multidetector CT. *Diagn. Interv. Radiol.* **15**, 22–30 (2009).
5. Liu, Y., Song, T., Huang, Z., Zhang, S. & Li, Y. The accuracy of multidetector computed tomography for preoperative staging of renal cell carcinoma. *Int. Braz. J. Urol.* **38**, 627–636 (2012).
6. Renard, A. S. et al. Is multidetector CT-scan able to detect T3a renal tumor before surgery?. *Scand. J. Urol.* **53**, 350–355 (2019).
7. Hallscheidt, P. et al. Multislice computed tomography in planning nephron-sparing surgery in a prospective study with 76 patients: comparison of radiological and histopathological findings in the infiltration of renal structures. *J. Comput. Assist. Tomogr.* **30**, 869–874 (2006).
8. Tsili, A. C. et al. Perirenal fat invasion on renal cell carcinoma: evaluation with multidetector computed tomography-multivariate analysis. *J. Comput. Assist. Tomogr.* **37**, 450–457 (2013).
9. Landman, J. et al. Preoperative computed tomography assessment for perinephric fat invasion: comparison with pathological staging. *J. Comput. Assist. Tomogr.* **41**, 702–707 (2017).
10. Fateh, S. M. et al. Renal cell carcinoma T staging: diagnostic accuracy of preoperative contrast-enhanced computed tomography. *Mol. Clin. Oncol.* **18**, 11 (2023).
11. Li, X. et al. Radiomics predict the WHO/ISUP nuclear grade and survival in clear cell renal cell carcinoma. *Insights Imaging* **15**, 175 (2024).
12. Nie, P. et al. A preoperative CT-based deep learning radiomics model in predicting the stage, size, grade and necrosis score and outcome in localized clear cell renal cell carcinoma: a multicenter study. *Eur. J. Radiol.* **166**, 111018 (2023).
13. Wang, N. et al. Study of radiomics based on dual-energy CT for nuclear grading and T-staging in renal clear cell carcinoma. *Medicine***103**, E37288 (2024).
14. Orton, M. R. et al. Interpretability of radiomics models is improved when using feature group selection strategies for predicting molecular and clinical targets in clear-cell renal cell carcinoma: insights from the TRACERx Renal study. *Cancer Imaging* **23**, 76 (2023).
15. Demirjian, N. L. et al. CT-based radiomics stratification of tumor grade and TNM stage of clear cell renal cell carcinoma. *Eur. Radiol.* **32**, 2552–2563 (2022).
16. Shimada, W. et al. Significance of tumor shape irregularity: radiomics analysis based on dynamic computed tomography for predicting pT3a upstaging in cT1b-2N0M0 renal cell carcinoma. *Int. J. Urol.* **29**, 1387–1389 (2022).
17. Kim, C., Choi, H. J. & Cho, K. S. Diagnostic performance of multidetector computed tomography in the evaluation of perinephric fat invasion in renal cell carcinoma patients. *J. Comput. Assist. Tomogr.* **38**, 268–273 (2014).
18. Sokhi, H. K., Mok, W. Y. & Patel, U. Stage T3a renal cell carcinoma: Staging accuracy of CT for sinus fat, perinephric fat or renal vein invasion. *Br. J. Radiol.* **88**, 20140504 (2015).
19. Elkassem, A. A., Allen, B. C., Sharbidre, K. G., Rais-Bahrami, S. & Smith, A. D. Update on the role of imaging in clinical staging and restaging of renal cell carcinoma based on the AJCC 8th edition, from the AJR special series on cancer staging. *Am. J. Roentgenol.* **217**, 541–555 (2021).
20. Damgaci, L., Özer, H. & Rona, G. Diagnostic value of MDCT in determining the perinephric fat tissue and renal sinus invasion in patients with clear cell renal cell carcinoma. *Niger. J. Clin. Pract.* **24**, 489–495 (2021).
21. Kim, J. K. J. H., Park, K. J., Kim, M. H. & Kim, J. K. J. H. Preoperative assessment of renal sinus invasion by renal cell carcinoma according to tumor complexity and imaging features in patients undergoing radical nephrectomy. *Korean J. Radiol.* **22**, 1323–1331 (2021).
22. Teishima, J. et al. Impact of radiological morphology of clinical T1 renal cell carcinoma on the prediction of upstaging to pathological T3. *Jpn. J. Clin. Oncol.* **50**, 473–478 (2020).
23. Tanaka, H. et al. Defining tumour shape irregularity for preoperative risk stratification of clinically localised renal cell carcinoma. *Eur. Urol. Open Sci.* **48**, 36–43 (2023).
24. Kim, C., Choi, H. J. & Cho, K. S. Diagnostic value of multidetector computed tomography for renal sinus fat invasion in renal cell carcinoma patients. *Eur. J. Radiol.* **83**, 914–918 (2014).
25. Kocak, B. et al. CheckList for EvaluAtion of Radiomics research (CLEAR): a step-by-step reporting guideline for authors and reviewers endorsed by ESR and EuSoMII. *Insights Imaging* **14**, (2023).
26. National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (CPTAC). The Clinical Proteomic Tumor Analysis Consortium Clear Cell Renal Cell Carcinoma Collection (CPTAC-CCRCC). https://doi.org/10.7937/k9/tcia.2018.oblamn27 (2018).
27. Akin, O. et al. The Cancer Genome Atlas Kidney Renal Clear Cell Carcinoma Collection (TCGA-KIRC). https://doi.org/10.7937/K9/TCIA.2016.V6PBVTDR (2016).

28. Linehan, M. et al. The Cancer Genome Atlas Cervical Kidney Renal Papillary Cell Carcinoma Collection (TCGA-KIRP). https://doi.org/10.7937/K9/TCIA.2016.ACWOGBEF (2016).
29. Linehan, M. W., Gautam, R., Sadow, C. A. & Levine, S. The Cancer Genome Atlas Kidney Chromophobe Collection (TCGA-KICH). https://doi.org/10.7937/K9/TCIA.2016.YU3RBCZN (2016).
30. Trpkov, K. et al. Handling and staging of renal cell carcinoma: the International Society of Urological Pathology Consensus (ISUP) conference recommendations. *Am. J. Surg. Pathol.* **37**, 1505–1517 (2013).
31. Yushkevich, P. A. et al. User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. *Neuroimage* **31**, 1116–1128 (2006).
32. Fedorov, A. et al. 3D Slicer as an image computing platform for the Quantitative Imaging Network. *Magn. Reson. Imaging* **30**, 1323–1341 (2012).
33. Zwanenburg, A. et al. The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology* **295**, 328–338 (2020).
34. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
35. Bradshaw, T. J., Huemann, Z., Hu, J. & Rahmim, A. A guide to cross-validation for artificial intelligence in medical imaging. *Radiol. Artif. Intell.* **5**, e220232 (2023).
36. Hanley, J. A. & McNeil, B. J. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology* **148**, 839–843 (1983).
37. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44**, 837–845 (1988).
38. McNEMAR, Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* **12**, 153–157 (1947).
39. Gallas, B. D. One-shot estimate of MRMC variance: AUC. *Acad. Radiol.* **13**, 353–362 (2006).
40. Smith, B. J. & Hillis, S. L. Multi-reader multi-case analysis of variance software for diagnostic performance comparison of imaging modalities. *Proc. SPIE Int. Soc. Opt. Eng.* **11316**, 113160K (2020).

Acknowledgements

This study was funded by the Sichuan Science and Technology Program (Grant number 2025ZNSFSC1759) and the 1.3.5 project for disciplines of excellence, West China Hospital, Sichuan University (Grant number ZYAI24048). The funder played no role in study design, data collection, analysis and interpretation of data, or the writing of this manuscript.

Author contributions

Conceptualization: E.Y., Y.C., Lei.Y., B.H., H.Z., L.Y., J.Y., B.S. Data curation: E.Y., Y.C. Formal analysis: E.Y., Y.C., Lei.Y., B.H., C.H. Funding acquisition: Y.C. Investigation: E.Y., Y.C., Lei.Y., B.H., C.H., J.M., T.Y., H.Z., L.Y., J.Y., B.S. Methodology: E.Y., Y.C., Lei.Y., B.H., H.Z., L.Y., J.Y., B.S. Project administration: L.Y., J.Y., B.S. Resources: H.Z., L.Y., J.Y., B.S. Supervision: L.Y., J.Y., B.S. Validation: E.Y., Y.C., Lei.Y., B.H., C.H., J.M., T.Y., H.Z., L.Y., J.Y., B.S. Visualization: E.Y. Writing—original draft: E.Y., Y.C., H.Z., L.Y., J.Y., B.S. Writing—review & editing: E.Y., Y.C., H.Z., L.Y., J.Y., B.S.

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/s41746-025-01723-x.

Correspondence and requests for materials should be addressed to Ling Yang, Jin Yao or Bin Song.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.