

RESEARCH

Open Access



Multicenter study of CT-based deep learning for predicting preoperative T staging and TNM staging in clear cell renal cell carcinoma

Wuchao Li^{1,2†}, Yin Xi^{3†}, Ming Lu^{4†}, Junjie He², Jianguo Zhu⁵, Haohan Li⁴, Tongyin Yang⁶, Xianchun Zeng², Xinfeng Liu², Rui Xu², Hui Huang⁷, Heng Liu⁸, Tijiang Zhang^{8*}, Xiangde Min^{9*} and Rongpin Wang^{2*}

Abstract

Background Accurate preoperative T and TNM staging of clear cell renal cell carcinoma (ccRCC) is crucial for diagnosis and treatment, but these assessments often depend on subjective radiologist judgment, leading to interobserver variability. This study aims to design and validate two CT-based deep learning models and evaluate their clinical utility for the preoperative T and TNM staging of ccRCC.

Methods Data from 1,148 ccRCC patients across five medical centers were retrospectively collected. Specifically, data from two centers were merged and randomly divided into a training set (80%) and a testing (20%) set. Data from two additional centers comprised external validation set 1, and data from the remaining independent center comprised external validation set 2. Two 3D deep learning models based on a Transformer-ResNet (TR-Net) architecture were developed to predict T staging (T1, T2, T3 + T4) and TNM staging (I, II, III, IV) using corticomedullary phase CT images. Gradient-weighted Class Activation Mapping (Grad-CAM) was used to generate heatmaps for improved model interpretability, and a human-machine collaboration experiment was conducted to evaluate clinical utility. Models' performance was evaluated using micro-average AUC (micro-AUC), macro-average AUC (macro-AUC), and accuracy (ACC).

Results Across the two external validation sets, the T staging model achieved micro-AUCs of 0.939 and 0.954, macro-AUCs of 0.857 and 0.894, and ACCs of 0.843 and 0.869, while the TNM staging model achieved micro-AUCs of 0.935 and 0.924, macro-AUCs of 0.817 and 0.888, and ACCs of 0.856 and 0.807. While the models demonstrated acceptable overall performance in preoperative ccRCC staging, performance was moderate for advanced subclasses (T3 + T4 AUC: 0.769 and 0.795; TNM III AUC: 0.669 and 0.801). Grad-CAM heatmaps highlighted key tumor regions, improving

[†]Wuchao Li, Yin Xi and Ming Lu contributed equally to this work.

*Correspondence:

Tijiang Zhang

tijzhang@163.com

Xiangde Min

minxiangde0129@126.com

Rongpin Wang

wangrongpin@126.com

Full list of author information is available at the end of the article



interpretability. The human-machine collaboration demonstrated improved diagnostic accuracy with model assistance.

Conclusion The CT-based 3D TR-Net models showed acceptable overall performance with moderate results in advanced subclasses in preoperative ccRCC staging, with interpretable outputs and collaborative benefits, making them potentially useful decision-support tools.

Keywords Clear cell renal cell carcinoma, Computed tomography, Deep learning, Tumor-Node-Metastasis staging

Background

Renal cell carcinoma (RCC), which originates in the renal cortex, accounts for 90% of primary renal tumors [1]. With over 400,000 new cases and more than 180,000 deaths annually worldwide, RCC ranks among the top 14 most common malignancies [1–3]. Clear cell renal cell carcinoma (ccRCC), accounting for over 80% of RCC cases, is the most common and biologically heterogeneous subtype, posing challenges in both treatment and prognosis due to its variable aggressiveness and response to therapies [1, 3, 4]. The tumor-node-metastasis (TNM) staging system evaluates tumor characteristics based on three components: tumor size and local invasion (T), regional lymph node involvement (N), and distant metastasis (M) [5]. Preoperative T staging focuses on assessing tumor size, depth of invasion, and involvement of surrounding tissues. Together, TNM staging plays a crucial role in determining the scope of surgery, systemic therapy needs, and overall prognosis. However, accurate staging often depends on postoperative pathology, limiting its applicability in early treatment planning. This underscores the significance of preoperative evaluation methods in providing timely and clinically actionable insights to guide and refine therapeutic decision-making [6].

Contrast-enhanced CT is widely used for preoperative evaluation of ccRCC, valued for its non-invasiveness, convenience, and consistency [7]. However, preoperative staging with CT is inherently observer-dependent, relying on radiologists' subjective visual assessments. This approach is not only time-consuming and labor-intensive but also subject to inter-observer variability, which can lead to inconsistencies in staging accuracy [8, 9]. These limitations highlight the pressing need for objective, automated, and accurate diagnostic tools for ccRCC staging.

Deep learning algorithms have shown remarkable accuracy in preoperative cancer staging, often matching or exceeding radiologist assessments in cancers such as lung, gastric, and colorectal [10–13]. Recent studies have explored applying deep learning to CT images for RCC preoperative staging. However, these studies often faced limitations such as reliance on single-center or public datasets, small sample sizes, unrepresentative case distributions, binary classification approaches, and insufficient predictive accuracy [7, 14–16]. Moreover, the

clinical interpretability of these models and their potential for collaboration with radiologists remain largely unexplored.

This study aims to address these challenges by developing and validating CT-based deep learning models using multicenter datasets comprising 1,148 ccRCC cases. Grad-CAM was integrated to explore model interpretability, and human-machine collaboration experiments were conducted to evaluate clinical utility. These comprehensive approaches enable precise and generalizable staging predictions across various clinical settings.

Methods

Patient collection

This study, approved by the Ethics Committee of the Guizhou Provincial People's Hospital (GPH) (Approval Number: KY2021165), waived the requirement for written informed consent due to its retrospective design and the anonymization of patient data. Data were retrospectively collected from five medical centers, and all cases were confirmed as ccRCC by postoperative pathology. Among them, data from GPH and Affiliated Hospital of Zunyi Medical University (ZMH) were merged and randomly divided into training set (80%) and testing set (20%). Data from Affiliated Hospital of Guizhou Medical University (GMH) and First Affiliated Hospital of Shihezi University (SUH) were combined to form external validation set 1, while data from the independent Guiqian International General Hospital (GQH) was used as external validation set 2. The data collection periods were as follows: GPH: July 2012 to April 2022; ZMH: April 2013 to February 2022; GMH: July 2012 to April 2022; SUH: April 2019 to April 2023; GQH: January 2021 to December 2024.

Inclusion criteria were: (1) surgical treatment, (2) postoperative pathological confirmation of ccRCC, and (3) contrast-enhanced CT scans conducted no more than thirty days prior to surgery. Exclusion criteria were: (1) lack of standard corticomedullary contrast-enhanced CT images, (2) poor-quality CT scans, including those with severe artifacts (e.g., motion artifacts, metal artifacts) or incomplete coverage of the lesion, (3) prior kidney biopsy or treatment before the CT scan, and (4) incomplete or inaccurate clinical and pathological data. The study flow-chart is presented in Fig. 1.

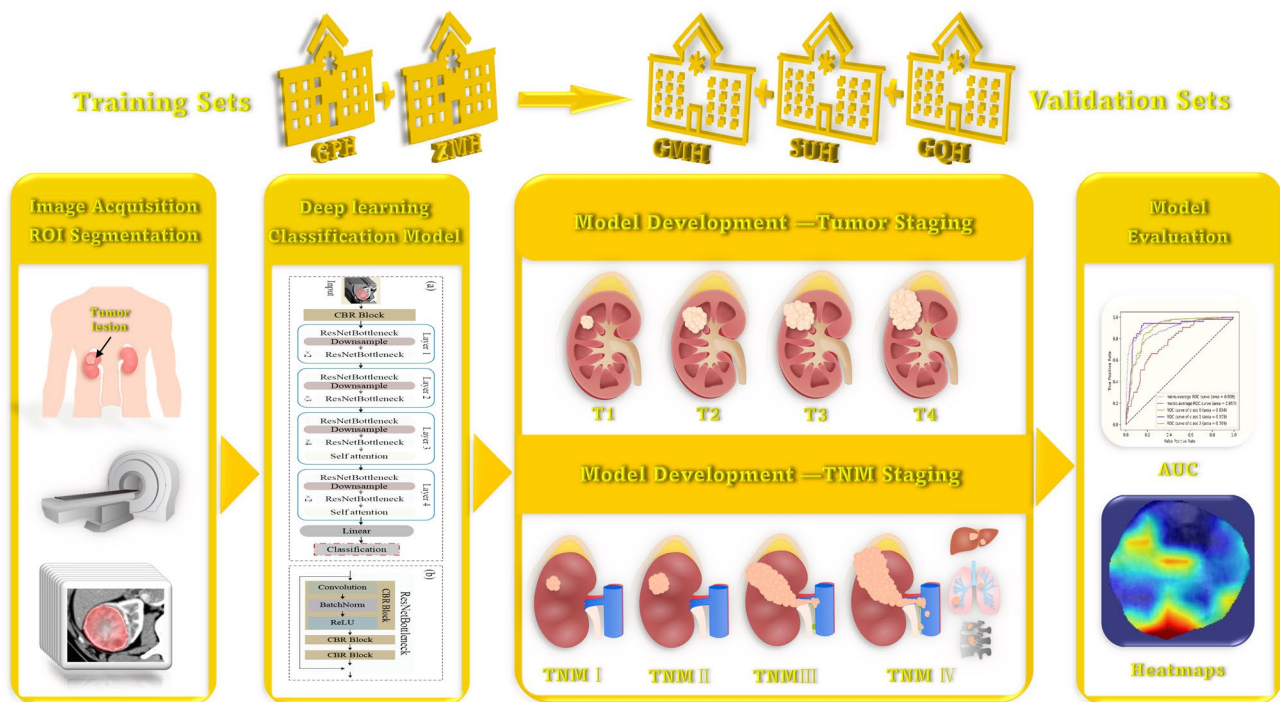


Fig. 1 The study flowchart

Clinical data collection

Preoperative corticomedullary contrast-enhanced CT images were retrieved from the Picture Archiving and Communication System at each center. Clinical and pathological data, including age, gender, tumor size, nuclear grade, T staging, N staging, M staging, and TNM staging, were collected through the clinical information management system. The primary predictors in this study were T staging and TNM staging, both crucial for determining treatment strategies and predicting outcomes. T staging was categorized into T1, T2, and T3 + T4, with T4 cases merged into T3 due to the small number of T4 cases. TNM staging was classified as stages I, II, III, or IV. At each center, two junior radiologists independently performed manual delineation of the region of interest (ROI) on corticomedullary contrast-enhanced CT images slice by slice using ITK-SNAP software for tumor segmentation. Any discrepancies between their delineations were reviewed and resolved by a senior radiologist to ensure consistency and minimize inter-rater variability.

Standardization of staging assessment

To ensure accuracy and consistency in T staging and TNM staging assessments, this study closely followed the AJCC 8th edition guidelines for renal cell carcinoma TNM staging [5]. A professional pathologist from each center independently re-evaluated the staging based on the standardized criteria to minimize variability and ensure uniformity across all data.

Network architecture

This study utilizes a Three Dimensional (3D) Transformer-ResNet (TR-Net) architecture to improve classification accuracy by progressively increasing network depth [17]. To address the challenge of capturing global features in Convolutional Neural Network (CNN) modules within the 3D ResNet, two Transformer modules were integrated into the model's backend to enhance its ability to extract long-range features in tumors [18]. The proposed 3D TR-Net model is illustrated in Fig. 2.

In the 3D model, convolutional layers utilize a $3 \times 3 \times 3$ kernel with a stride of $1 \times 1 \times 1$ and padding of $1 \times 1 \times 1$. Downsampling is achieved using convolution with a $1 \times 1 \times 1$ kernel and a stride of $2 \times 2 \times 2$. The final Linear layer decodes the 3D model features into classification results. The calculation of self-attention in the 3D context is defined as follows:

$$\text{Self-Attention} = \text{Softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V \quad (1)$$

where Q, K, and V are three feature matrices of the same size obtained by linear projection of the self-attention input. The softmax function is then applied to normalize the results [19]. d_k is the feature dimension used to scale the results, preventing gradient explosion during network training. Self-attention is a core module in the Transformer architecture, allowing the model to learn

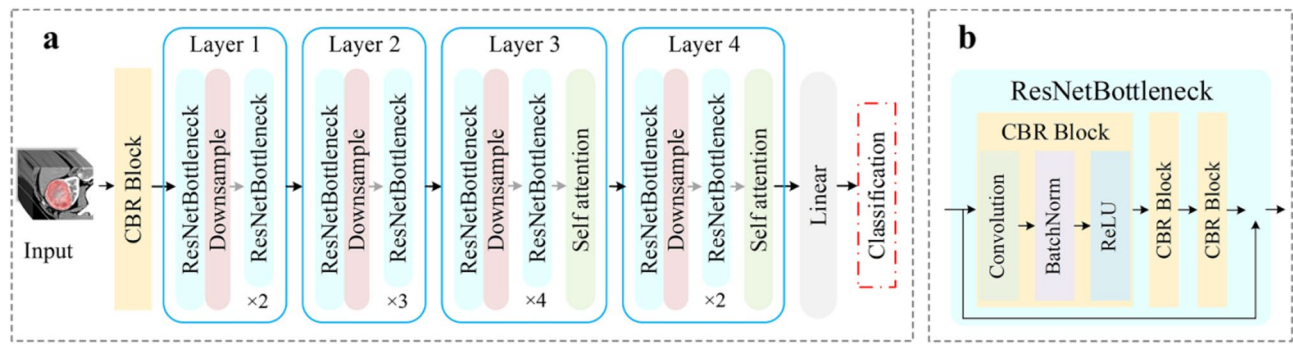


Fig. 2 Architecture of 3D TR-Net. (a) illustrates the workflow of 3D TR-Net. (b) depicts the structure of the ResNetBottleneck within the workflow

the global features of the data. However, due to its high computational complexity, 3D TR-Net employs two modules only in the backend of the network.

Data preprocessing

Preprocessing operations were conducted on the collected 3D CT images to improve data quality, enhance usability, and increase model classification accuracy. Given the use of multicenter data in this study, it was essential to integrate data from various sources for effective management and utilization. As previously mentioned, data from GPH and ZMH were used for model development, including training and test (80%/20%), while data from GMH and SUH were used as external validation set 1, and GQH data served as an independent external validation set 2. To provide transparency on the CT protocols across the five centers, key acquisition parameters including manufacturer, model, kVp, contrast dose, and slice thickness, are summarized in Supplementary Table S1.

To convert raw CT data into a manageable and analyzable format, data transformation, including dimensionality reduction and numerical normalization, was performed. Tumor ROI (Region of Interest) segmentation results were used as the reference to extract data blocks of size $128 \times 128 \times 64$ from the center of the tumor ROI outward. This cropping method, rather than resizing the entire tumor ROI, preserves the morphological and textural information essential for model learning. Following extraction, a masking operation retained only the histological information of the tumor and its surrounding tissues. Standardization was then applied as follows:

$$\text{Model}_{\text{in}} = \frac{\text{cropped} - \text{mean}(\text{cropped})}{\text{std}(\text{cropped})} \quad (2)$$

where cropped refers to the tumor ROI data that has been trimmed or extracted from the original image, mean refers to the average value, and std refers to the standard deviation. To mitigate inter-center variability from diverse CT scanners and acquisition settings, we

applied per-sample z-score normalization to the cropped tumor ROI data. This approach standardizes each sample individually, enabling the model to prioritize relative intensity and structural patterns over scanner-dependent absolute values, thus preserving critical tumor features for classification.

Network training

Cross-entropy loss was utilized to train the 3D network model for classification tasks. Given the uneven distribution of data classes [20], weighted focal loss was applied to improve the network's performance on small-sample classes [21]. Focal Loss is defined as:

$$\mathcal{L}_F(p_t) = -\alpha_t(1 - p_t)^\gamma \log(p_t) \quad (3)$$

where p_t denotes the predicted probability of the sample by the model, α_t represents the class weights of the sample, and γ is the modulation factor. The cross-entropy loss also incorporates sample weights and can be represented as:

$$\mathcal{L}_{CE}(x, y) = \{\mathcal{L}_1, \dots, \mathcal{L}_N\}^T, \mathcal{L}_n = -w_{y_n} \log \frac{\exp(x_{n,y_n})}{\sum_{c=1}^C \exp(x_{n,c})} \quad (4)$$

where x represents the model output, y represents the labels, w represents the weights, C represents the number of classes, and N represents the number of samples. Finally, the total training loss of the model is obtained by combining the two losses using the weight parameter α :

$$\mathcal{L}_{\text{Total}} = \alpha \mathcal{L}_F + (1 - \alpha) \mathcal{L}_{CE} \quad (5)$$

In this study, α is set to 0.5. The model parameters were optimized using the AdamW optimizer with an initial learning rate of $1e-4$. The model was trained for 50 epochs on the training set, with performance monitored on the internal testing set. The model parameters that achieved the best performance on this internal testing set were then selected for final evaluation on two external

validation sets. All experiments were conducted on an NVIDIA A100 Graphics Processing Unit.

Visualizing model attention with Gradient-weighted class activation mapping (Grad-CAM)

To enhance the interpretability of the 3D TR-Net model, the Grad-CAM technique was employed on the training set. This method visualizes the regions of CT images that the model focuses on during T staging and TNM staging predictions. Grad-CAM generates heatmaps by computing the gradients of the final convolutional layer and overlaying them onto the original images. The color bar, ranging from blue to red, indicates activation intensity from low to high. This approach provides an intuitive understanding of the model's decision-making process and its attention to tumor-related features.

Model evaluation

This study assessed the multiclass performance of the T staging and TNM staging models quantitatively, primarily using micro-averaged AUC (micro-AUC), macro-averaged AUC (macro-AUC), and accuracy (ACC). To provide a comprehensive evaluation of the models' diagnostic efficacy, supplementary metrics like precision, recall, and F1-score were also employed. To mitigate potential bias from relying on a single dataset, this study employed two independent external validation sets to ensure robustness and generalizability of the models.

Human-machine collaboration evaluation

To evaluate the clinical utility of the proposed T staging model, an additional blinded evaluation was conducted using external validation set 2. Two junior radiologists (5 and 7 years of experience) and two senior radiologists (14 and 20 years of experience), all blinded to pathology results, independently performed T staging assessments. Subsequently, the radiologists reviewed the model's staging predictions along with Grad-CAM visualizations, and revised their assessments accordingly. ACC was calculated to objectively compare performance between independent and human-machine collaborative assessments.

Statistical analysis

Analyses were performed using R software (version 4.2.2) and Python (version 3.11.11). Categorical data were analyzed using chi-square or Fisher's exact tests, depending on sample size, while continuous data were assessed using t-tests or the Mann-Whitney U test based on the data distribution. A threshold of $p < 0.05$ was set to denote statistical significance. Paired categorical accuracy data (unaided vs. with model assistance) were evaluated with McNemar's test. To ensure adequate sensitivity, post-hoc power analyses were conducted for each

McNemar comparison, targeting $\geq 80\%$ power, with a significance threshold of $\alpha = 0.05$ ($p < 0.05$).

Results

This study included 1,148 ccRCC patients from five centers: GPH ($n = 344$, 30%), ZMH ($n = 221$, 19%), GMH ($n = 288$, 25%), SUH ($n = 161$, 14%), and GQH ($n = 134$, 12%). This multicenter approach provided a diverse and representative dataset, ensuring robust generalizability of the models. Data from GPH and ZMH ($n = 565$) were merged and divided into a training set (80%, $n = 452$) and a testing set (20%, $n = 113$). Two external validation sets were defined: external validation set 1 (GMH+SUH, $n = 449$) and external validation set 2 (GQH, $n = 134$). Except for nuclear grade and size ($P < 0.05$), no significant differences were observed in age, sex, or staging distributions (T, N, M, TNM) across the three sets (all $P > 0.05$), indicating no systematic bias among the groups. Baseline clinicopathological characteristics are detailed in Table 1.

For Age and Size, data are expressed as mean $\pm$ standard deviation. All other data are presented in the form of count and percentage (n (%)).

T Staging, Tumor Staging; N Staging, Node Staging; N0, No regional lymph node metastasis; Nx, Regional lymph nodes cannot be assessed; N1, Regional lymph node metastasis present; M Staging, Metastasis Staging; M0, No distant metastasis; M1, Distant metastasis present.

For T staging prediction, the 3D TR-Net model demonstrated acceptable overall performance across all sets. In the training, testing, external validation 1, and external validation 2 sets, The micro-AUCs were 0.974, 0.946, 0.939, and 0.954, respectively; macro-AUCs were 0.943, 0.839, 0.857, and 0.894; and ACCs were 0.929, 0.882, 0.843, and 0.869. Although overall performance was acceptable, the AUC for T3+T4 was moderate (external validation 1: AUC=0.769; external validation 2: AUC=0.795), likely due to class imbalance, while the model maintained strong discrimination for T1 (AUC=0.894 and 0.933) and T2 (AUC=0.909 and 0.955) in external validation 1 and 2 sets. Precision, recall, and F1-score for each stage further confirmed performance across stages, though with variability in advanced subclasses (as shown in Supplementary Table S2), for instance, F1-scores in external validation set 1 were 0.863 for T1, 0.629 for T2, and 0.634 for T3+T4; in set 2 they were 0.899 for T1, 0.750 for T2, and 0.716 for T3+T4, highlighting stronger results for the majority T1 class but moderate-to-lower metrics for minority classes (Table 2; Fig. 3).

For TNM staging prediction, the 3D TR-Net model also exhibited acceptable overall performance. In the training, testing, external validation 1, and external validation 2 sets, the micro-AUCs were 0.955, 0.918, 0.935, and

Table 1 Patient characteristics and clinicopathological variables of ccRCC patients across different sets

Characteristics	GPH + ZMH(<i>n</i> = 565)	GMH + SUH(<i>n</i> = 449)	GQH (<i>n</i> = 134)	<i>p</i> -value
Age	56.35 ± 12.36	56.87 ± 12.81	58.25 ± 11.65	0.138
Gender				0.337
Female	213 (37.70%)	157 (34.97%)	42(31.34%)	
Male	352 (62.30%)	292 (65.03%)	92(68.66%)	
Size	4.92 ± 2.38	4.77 ± 2.35	4.57 ± 2.34	0.024
Nuclear Grading				< 0.001
I	88 (15.58%)	96 (21.38%)	16(11.94%)	
II	390 (69.03%)	295 (65.70%)	80(59.70%)	
III	77 (13.63%)	48 (10.69%)	32(23.88%)	
IV	10 (1.77%)	10 (2.23%)	6(4.48%)	
T staging				0.469
T1	439 (77.70%)	344 (76.61%)	105 (78.36%)	
T2	77 (13.63%)	52 (11.58%)	15 (11.19%)	
T3 + T4	49 (8.67%)	53 (11.80%)	14 (10.45%)	
T staging				0.487
T1	439 (77.70%)	344 (76.61%)	105 (78.36%)	
T2	77 (13.63%)	52 (11.58%)	15 (11.19%)	
T3	45 (7.96%)	49 (10.91%)	13 (9.70%)	
T4	4 (0.71%)	4 (0.89%)	1 (0.75%)	
N staging				0.155
N0+Nx	559 (98.94%)	440 (98.00%)	134(100%)	
N1	6 (1.06%)	9 (2.00%)	0(0%)	
M staging				0.414
M0	535 (94.69%)	433 (96.44%)	128(95.52%)	
M1	30 (5.31%)	16 (3.56%)	6(4.48%)	
TNM staging				0.487
I	421 (74.51%)	337 (75.06%)	104(77.61%)	
II	69 (12.21%)	47 (10.47%)	13(9.70%)	
III	41 (7.26%)	46 (10.24%)	11(8.21%)	
IV	34 (6.02%)	19 (4.23%)	6(4.48%)	

Table 2 Performance of the 3D TR-Net model for preoperative T staging of ccRCC across different sets

Category	Training set	Testing set	External validation set 1	External validation set 2
Micro AUC	0.974 (0.964–0.982)	0.946 (0.916–0.967)	0.939 (0.924–0.951)	0.954 (0.921–0.978)
Macro AUC	0.943 (0.918–0.962)	0.839 (0.758–0.901)	0.857 (0.822–0.888)	0.894 (0.838–0.940)
T1 AUC	0.948 (0.923–0.967)	0.846 (0.766–0.907)	0.894 (0.861–0.921)	0.933 (0.873–0.977)
T2 AUC	0.968 (0.947–0.982)	0.870 (0.794–0.926)	0.909 (0.878–0.934)	0.955 (0.909–0.989)
T3 + T4 AUC	0.913 (0.884–0.938)	0.800 (0.714–0.869)	0.769 (0.728–0.808)	0.795 (0.665–0.890)
Accuracy	0.929 (0.901–0.955)	0.882 (0.813–0.946)	0.843 (0.806–0.880)	0.869 (0.803–0.935)
Precision	0.917 (0.883–0.948)	0.785 (0.659–0.909)	0.725 (0.642–0.817)	0.746 (0.616–0.860)
Recall	0.914 (0.881–0.944)	0.850 (0.761–0.930)	0.796 (0.749–0.845)	0.836 (0.750–0.905)
F1-score	0.916 (0.882–0.946)	0.816 (0.706–0.919)	0.759 (0.691–0.830)	0.788 (0.676–0.882)

0.924, respectively, the macro-AUCs were 0.883, 0.782, 0.817, and 0.888; and the ACCs were 0.891, 0.841, 0.856, and 0.807. Although overall performance was acceptable, the AUC for TNM III was moderate (external validation 1: AUC = 0.669; external validation 2: AUC = 0.801), likely due to class imbalance, while the model maintained strong discriminative power for TNM I (AUC = 0.864 and 0.936), TNM II (AUC = 0.907 and 0.963), and TNM IV (AUC = 0.826 and 0.852) in the two external validation sets. Precision, recall, and F1-score for each stage further confirmed consistent performance (Table 3; Fig. 4).

Grad-CAM visualizations

Grad-CAM heatmaps generated from the training set illustrate the model's attention regions in T staging and TNM staging predictions. For T staging (Fig. 5), T1 presents a deep blue low-intensity heatmap, T2 shows a broader distribution with moderate intensity, and T3 + T4 displays significant intensity concentrated in the tumor center and high-density areas. For TNM staging (Fig. 6), the heatmap for TNM I is predominantly deep

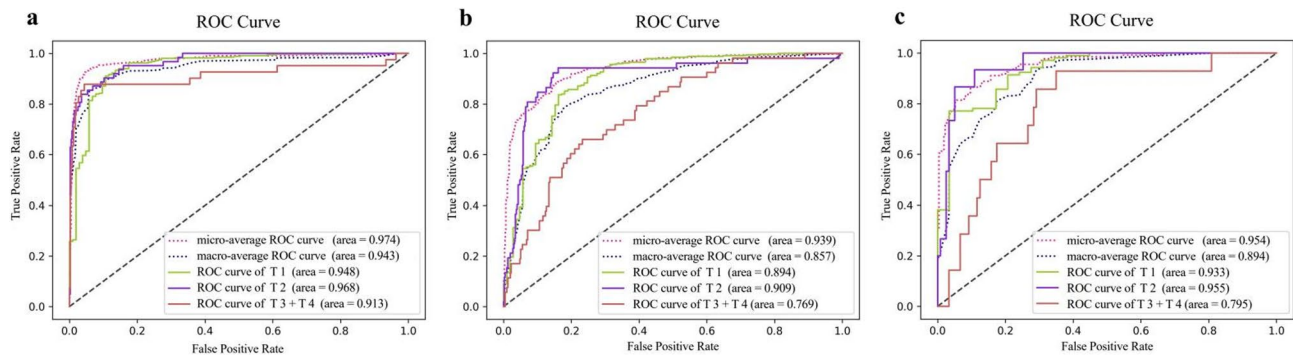


Fig. 3 ROC curves for the 3D TR-Net model for predicting T staging in the training set (a), the external validation set 1 (b) and the external validation set 2 (c)

Table 3 Performance of the 3D TR-Net model for preoperative TNM staging of ccRCC across different sets

Category	Training set	Testing set	External validation set 1	External validation set 2
Micro AUC	0.955 (0.945–0.965)	0.918 (0.889–0.942)	0.935 (0.922–0.946)	0.924 (0.890–0.957)
Macro AUC	0.883 (0.850–0.911)	0.782 (0.697–0.853)	0.817 (0.780–0.850)	0.888 (0.846–0.928)
TNM I	0.901 (0.870–0.927)	0.839 (0.758–0.901)	0.864 (0.829–0.894)	0.936 (0.884–0.976)
TNM II	0.927 (0.899–0.949)	0.871 (0.795–0.927)	0.907 (0.877–0.932)	0.963 (0.913–0.996)
TNM III	0.865 (0.830–0.895)	0.699 (0.605–0.782)	0.669 (0.624–0.713)	0.801 (0.707–0.898)
TNM IV	0.840 (0.802–0.872)	0.721 (0.629–0.802)	0.826 (0.788–0.860)	0.852 (0.758–0.937)
Accuracy	0.891 (0.862–0.919)	0.841 (0.773–0.908)	0.856 (0.822–0.889)	0.807 (0.743–0.871)
Precision	0.729 (0.666–0.792)	0.704 (0.547–0.845)	0.716 (0.638–0.790)	0.603 (0.460–0.735)
Recall	0.828 (0.786–0.870)	0.761 (0.662–0.859)	0.789 (0.742–0.834)	0.775 (0.679–0.857)
F1-score	0.776 (0.721–0.829)	0.731 (0.599–0.852)	0.750 (0.686–0.811)	0.679 (0.549–0.792)

blue, with low-intensity features uniformly distributed within the tumor and occasional high-weight regions at the edges. TNM II displays a blue-green heatmap with scattered yellow high-weight regions at the edges. TNM III exhibits unevenly distributed yellow and red high-weight regions, while TNM IV is characterized by concentrated red and yellow high-intensity areas in the tumor center and high-density regions. These visualizations demonstrate the model's focus on clinically relevant regions, thereby enhancing its interpretability and clinical credibility.

Human–Machine collaboration results

On external validation set 2, the accuracy of T staging by the model alone was 0.869. The junior radiologists achieved an accuracy of 0.813 (0.767–0.860) unaided, which improved to 0.873 (0.833–0.913) with model assistance, representing an absolute improvement of 5.96% points ($p=0.001$, McNemar's test; post-hoc power=98.3%). Similarly, the senior radiologists' accuracy increased from 0.854 (0.812–0.897) unaided to 0.896 (0.859–0.932) with model assistance, representing an absolute improvement of 4.20% points ($p=0.009$, McNemar's test; post-hoc power=88.4%). These results demonstrate statistically significant improvement in diagnostic accuracy with sufficient power to detect the

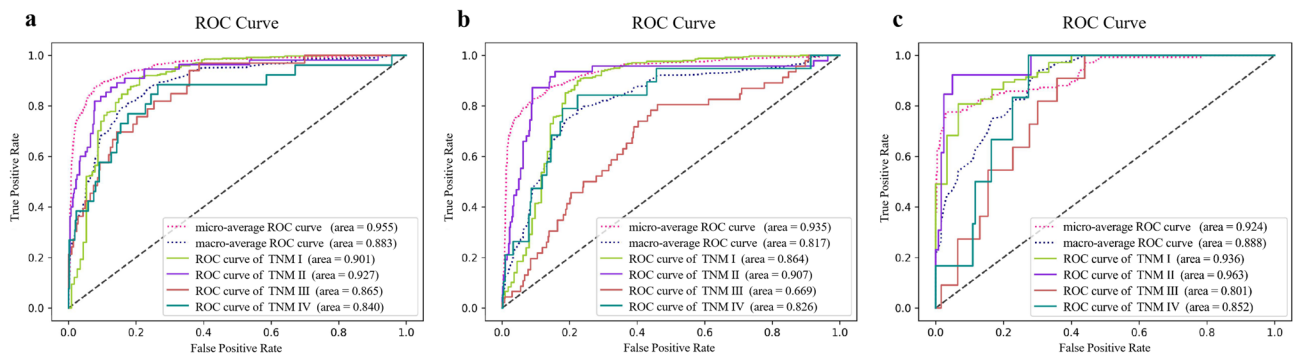


Fig. 4 ROC curves for the 3D TR-Net model for predicting TNM staging in the training set (a), the external validation set 1 (b) and the external validation set 2 (c)

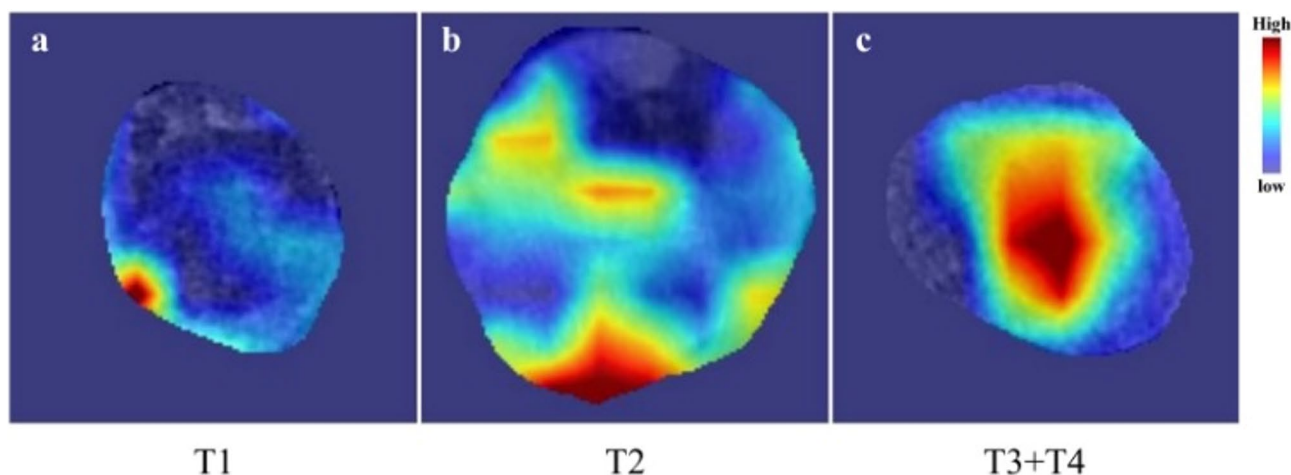


Fig. 5 Representative Grad-CAM heatmaps illustrating T stage predictions of ccRCC on CT images. T1 tumors exhibit uniformly low activation—blue (a); T2 tumors show moderate activation primarily along the lesion edges—green to yellow (b); while T3 + T4 tumors display high central activation intensity—yellow to red (c). The color bar indicates increasing activation strength from blue to red

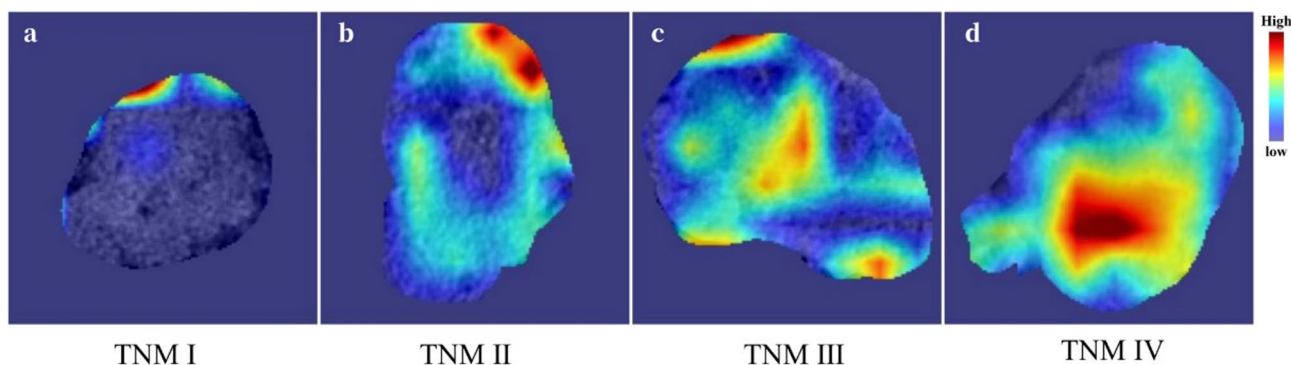


Fig. 6 Representative Grad-CAM heatmaps for TNM staging of ccRCC on CT images. TNM I tumors show diffuse low activation—blue (a); TNM stage II demonstrates moderate activation concentrated at lesion boundaries—green to yellow (b); TNM stage III presents heterogeneous activation both centrally and peripherally—yellow to red (c); and TNM stage IV reveals intense central activation—yellow to red (d). Color gradients denote activation strength from low to high

observed effects, particularly benefiting less-experienced radiologists.

Discussion

This study developed and externally validated two CT-based 3D TR-Net models for preoperative T staging and TNM staging in ccRCC. In independent validation cohorts, the T-staging model demonstrated acceptable discrimination and accuracy, paralleled by similarly acceptable performance from the TNM-staging model. Grad-CAM heatmaps highlighted the key tumor regions driving each prediction, enhancing model interpretability. In a human-machine collaboration study, radiologists assisted by 3D TR-Net improved their staging accuracy. Together, these results demonstrate that 3D TR-Net offers acceptable overall performance with robust generalizability in majority classes, and practical interpretability, making it a promising tool for prospective clinical evaluation in ccRCC staging despite moderate results in

advanced subclasses, as it reflects real-world class distributions and supports radiologist assistance.

This study addresses several Limitations in prior work through the following. First, it adhered to the latest AJCC 8th edition renal cancer TNM staging guidelines, resolving inconsistencies found in previous studies [5]. This approach improves the reliability and significantly enhances the clinical applicability of the results. Secondly, the 3D TR-Net hybrid neural network model effectively integrates the global feature extraction capabilities of the Transformer module, the local detail recognition strengths of the CNN module, and the depth feature representation of ResNet to capture 3D tumor structures [17, 21, 22]. This combination provides a robust solution for predicting preoperative T and TNM staging in ccRCC. Lastly, this study addresses limitations in prior research that relied on single-center or public datasets, which often led to discrepancies between dataset case distributions and real-world epidemiological patterns

[14, 16, 23]. Incorporating data from 1,148 ccRCC cases across five medical centers allowed the creation of a large and diverse dataset, enhancing the clinical relevance of the findings and supporting more accurate staging predictions.

In early explorations of deep learning for T staging prediction in ccRCC, Hadjiyski et al. pioneered the application of this technique using CT images, achieving an AUC of 0.90 [24]. Subsequently, Wu et al. utilized CT texture features in a multicenter study, attaining an AUC of 0.80 [16]. While these studies underscored the potential of intelligent diagnosis in T staging, they were limited to binary classification, failing to capture the full complexity of T staging. To overcome this limitation, Tian et al. developed a multi-class model for pathological T1-T3 staging based on data from two centers [23]. Although effective on internal test sets, the model demonstrated limited generalizability on external sets, with micro- and macro-AUCs of 0.72 and 0.78, respectively. In contrast, the 3D TR-Net model presented in this study exhibited improved performance on external validation—likely due to our large multicenter dataset—though it showed moderate results in advanced subclasses. Its overall diagnostic accuracy slightly exceeded that of previous radiologist assessments (ACC: 0.80) [25].

However, despite these improvements, the T staging model faces challenges from class imbalance and limitations in capturing complex tissue invasion. It focuses solely on the tumor, overlooking key features like tumor thrombus and perinephric invasion, which are vital for differentiating localized from advanced ccRCC, especially in T3 staging. These shortcomings limit its ability to accurately stage advanced ccRCC, particularly in T3 and T4 subgroups. Still, its performance, backed by a large multicenter dataset, demonstrates clinical potential for ccRCC T staging prediction. Future inclusion of tumor thrombus and perinephric invasion features could enhance its accuracy in advanced ccRCC stages (T3 and T4).

Regarding TNM staging, the results of this study also showed advantages over previous research, such as improved generalizability from multicenter data. Ökmen et al. laid the foundation for intelligent preoperative TNM staging prediction, achieving an accuracy of 0.85, which was slightly lower than the overall accuracy of 0.856 demonstrated by this study [26]. Talaat et al. later developed a radiomics-based model for multi-class TNM staging, reporting an accuracy of 0.99 in the validation set [27]. However, their feature selection was conducted simultaneously on both the training and testing sets, which increased the risk of overfitting. In contrast, this study utilized a large, multicenter dataset to ensure the model's generalizability and reliability. Furthermore, in studies on binary classification models for early and mid-late RCC TNM staging, Hussain et al. achieved

an accuracy of 0.83 in a single-center study [28], while Demirjian et al. reported an AUC of 0.80 in a multicenter study [29]. Neither of these models matched the accuracy and generalizability demonstrated by this study.

Although overall performance was acceptable, TNM stage III cases exhibited a noticeably lower AUC, suggesting diminished discriminative capacity for this category. This may be partially attributed to class imbalance, which can bias the model toward majority classes and reduce its sensitivity to underrepresented stages such as TNM stage III [30]. Additionally, TNM stage III often presents CT features that substantially overlap with those of TNM stages II and IV, making accurate classification difficult even for experienced radiologists [31]. Furthermore, the current architecture focuses solely on localized CT features and lacks mechanisms to incorporate broader anatomical context—such as peripheral invasion, tumor thrombus, and distant metastases—which are essential for accurate staging of advanced TNM cases. Future enhancements may include the use of context-aware modules (e.g., attention mechanisms or graph-based networks), and multimodal fusion of CT image with clinical and radiology report to improve the model's ability to distinguish intermediate and advanced stage cases.

Grad-CAM visualizations offer an intuitive depiction of the tumor regions that 3D TR-Net exploits for staging prediction. Prior renal cancer studies have similarly validated its clinical utility: Zhu et al. applied Grad-CAM to a multimodal B-mode and contrast-enhanced ultrasound network, revealing distinct modality dependencies when differentiating low versus high nuclear grade RCC [32]; Moon et al. used Grad-CAM with a ResNet-18 classifier to delineate the model's focus regions across normal, benign, and malignant renal pathological tissues [33]. In this study, Grad-CAM heatmaps exhibited a clear progression as tumor T and TNM stage increased—from low intensity, diffuse activations in T1/TNM I cases to high intensity, centrally concentrated activations in T3+T4/TNM IV cases. This visualization confirms 3D TR-Net's ability to capture key CT imaging features related to heterogeneity, invasiveness, and metastatic potential, enhancing model interpretability and potentially aiding diagnosis.

Human-machine collaboration using the 3D TR-Net model with Grad-CAM visualization significantly improves ccRCC T-staging accuracy, particularly benefiting junior radiologists and reducing experience-based variability in diagnosis. Similar benefits have been demonstrated in other radiological applications, such as a CT-based clinico-radiomics model that improved junior radiologists' diagnostic performance when differentiating mediastinal lymphomas from thymic epithelial tumors [34], and a breast cancer classification model that provides interpretable insights to radiologists via Grad-CAM

heatmaps, highlighting critical regions in mammograms influencing its predictions [35]. This synergistic approach enhances diagnostic workflow safety, consistency, and efficiency in clinical practice.

Despite the acceptable performance of the proposed models, several limitations remain. First, reliance on manual ROI delineation is time-consuming and subjective, limiting scalability. Future work will explore automated segmentation frameworks to improve efficiency. Second, the study used only corticomedullary contrast-enhanced CT images; integrating multiphase or multi-modal imaging and clinical data may further enhance model performance. Third, due to the limited number of T4 cases, T3 and T4 stages were combined to ensure model robustness, inevitably reducing clinical granularity. Fourth, the model focused solely on intratumoral features, without fully incorporating peritumoral invasion characteristics such as vascular or renal sinus fat infiltration, which are essential for distinguishing advanced ccRCC, especially in T3 staging. Prior research has highlighted the value of integrating intratumoral and peritumoral radiomics for predicting nuclear grading and survival outcomes in ccRCC, suggesting that such peri-tumoral analysis has the potential to substantially improve predictive accuracy for higher T stages in future iterations [36, 37]. Fifth, while per-sample z-score normalization mitigated inter-center CT variability without advanced methods like HU calibration or ComBat (as routine scanner HU calibration ensured absolute consistency and ComBat suits radiomics features, not end-to-end deep learning from raw images [38]), its efficacy is evidenced by robust external validation and comparable multicenter CT studies [39]. Future studies with more T4 data should aim for finer subclassification. Lastly, although Grad-CAM offers some interpretability, the decision-making process of deep models remains largely opaque, potentially limiting clinical trust and adoption.

Conclusion

This multicenter study validates the CT-based 3D TR-Net models for preoperative ccRCC staging, demonstrating acceptable overall performance on external cohorts but moderate results in advanced subclasses and offering potential to enhance clinical decision-making and radiologist diagnostic accuracy, especially when augmented with future improvements addressing class imbalance and extratumoral features.

Abbreviations

RCC	Renal Cell Carcinoma
ccRCC	Clear Cell Renal Cell Carcinoma
AJCC	American Joint Committee on Cancer
T staging	Tumor staging
N staging	Lymph node staging
M staging	Metastasis staging
TNM	Tumor-Node-Metastasis

CNN	Convolutional Neural Network
3D	Three Dimensional
ResNet	Residual Network
TR-Net	Transformer-ResNet
ROI	Region of Interest
AUC	Area Under the Curve
ACC	Accuracy
micro-AUC	Micro-averaged AUC
macro-AUC	Macro-averaged AUC
Grad-CAM	Gradient-weighted Class Activation Mapping
GPH	Guizhou Provincial People's Hospital
ZMH	Zunyi Medical University
GMH	Guizhou Medical University
SUH	Shihezi University
GQH	Guiqian International General Hospital

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12885-025-14836-z>.

Supplementary Material 1.

Acknowledgements

Not applicable.

Authors' contributions

WL contributed to Study Design, Data Analysis, and Manuscript Writing. JH was responsible for Algorithm Development and Model Design. YX, TZ, and ML provided data. JZ and XM were involved in Study Design. TY, XZ, HL and HH carried out Data Organization. XL and RX handled Data Analysis and Statistics. HL was responsible for Manuscript Revision. RW acquired funding and supervised the project. All authors have read and approved the final manuscript.

Funding

The study was supported by the National Natural Science Foundation of China (No. 82160349, 82460344, 82060314), and Guizhou Province High level Innovative Talent Project (QKHPTRC-GCC[2022]041 – 1, GZWJRS[2023]004), Science and Technology Foundation of Guizhou Province (QKHPTRC[2019]5803), Guizhou Provincial International Cooperation Bases (QKH-GHJD[2025]004) and Natural Scientific Project of Guizhou Province: Qiankehe Foundation -ZK[2022]General 263.

Data availability

The datasets analyzed during the current study were collected retrospectively from five centers. Due to patient privacy and institutional policies, the datasets are not publicly available but can be requested from the corresponding author under a data use agreement.

Declarations

Ethics approval and consent to participate

This study approved by the Ethics Committee of the corresponding hospital (Approval Number: KY2021165), waived the requirement for written informed consent due to its retrospective design and the anonymization of patient data. All methods were carried out in accordance with relevant guidelines and regulations in the study.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

¹Medical College, Guizhou University, Guiyang, China

²Department of Radiology, Guizhou Provincial People's Hospital, Guiyang, China

³Medical Image Center, First Affiliated Hospital of Shihezi University, Shihezi, China

⁴Department of Radiology, Guiqian International General Hospital, Guiyang, China

⁵Department of Urinary, Guizhou Provincial People's Hospital, Guiyang, China

⁶Department of Pathology, Guizhou Provincial People's Hospital, Guiyang, China

⁷Department of Radiology, Affiliated Hospital of Guizhou Medical University, Guiyang, China

⁸Department of Radiology, Affiliated Hospital of Zunyi Medical University, Zunyi, China

⁹Department of Radiology, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China

Received: 15 February 2025 / Accepted: 8 August 2025

Published online: 17 October 2025

References

- Bukavina L, Bensalah K, Bray F, Carlo M, Challacombe B, Karam JA, Kassouf W, Mitchell T, Montironi R, O'Brien T. Epidemiology of renal cell carcinoma: 2022 update. *Eur Urol*. 2022;82(5):529–42.
- Makino T, Kadamoto S, Izumi K, Mizokami A. Epidemiology and prevention of renal cell carcinoma. *Cancers*. 2022;14(16):4059.
- Young M, Jackson-Spence F, Beltran L, Day E, Suarez C, Bex A, Powles T, Szabados B. Renal cell carcinoma. *Lancet*. 2024;404(10451):476–91.
- Motzer RJ, Jonasch E, Agarwal N, Alva A, Baine M, Beckermann K, Carlo MI, Choueiri TK, Costello BA, Derweesh IH. Kidney cancer, version 3.2022, NCCN clinical practice guidelines in oncology. *J Natl Compr Canc Netw*. 2022;20(1):71–90.
- Edition S, Edge S, Byrd D. AJCC cancer staging manual. *AJCC cancer staging manual* 2017.
- Motzer RJ, Jonasch E, Agarwal N, Alva A, Bagshaw H, Baine M, Beckermann K, Carlo MI, Choueiri TK, Costello BA. NCCN guidelines® insights: kidney cancer, version 2.2024: featured updates to the NCCN guidelines. *J Natl Compr Canc Netw*. 2024;22(1):4–16.
- Pinto PVA, Coelho FMA, Schuch A, Zapparoli M, Baroni RH. Pre-operative imaging evaluation of renal cell carcinoma. *Revista Da Associação Médica Brasileira*. 2024;70(suppl 1):e2024S2107.
- Elkassam AA, Allen BC, Sharbidre KG, Rais-Bahrami S, Smith AD. 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*. 2021;217(3):541–55.
- Shah PH, Moreira DM, Patel VR, Gaunay G, George AK, Alom M, Kozel Z, Yaskiv O, Hall SJ, Schwartz MJ. Partial nephrectomy is associated with higher risk of relapse compared with radical nephrectomy for clinical stage T1 renal cell carcinoma pathologically up staged to T3a. *J Urol*. 2017;198(2):289–96.
- Najjar R. Redefining radiology: a review of artificial intelligence integration in medical imaging. *Diagnostics*. 2023;13(17):2760.
- Zheng X, He B, Hu Y, Ren M, Chen Z, Zhang Z, Ma J, Ouyang L, Chu H, Gao H. Diagnostic accuracy of deep learning and radiomics in lung cancer staging: a systematic review and meta-analysis. *Front Public Health*. 2022;10:938113.
- Liu F, Xie Q, Wang Q, Li X. Application of deep learning-based CT texture analysis in TNM staging of gastric cancer. *J Radiation Res Appl Sci*. 2023;16(3):100635.
- Hou M, Zhou L, Sun J. Deep-learning-based 3D super-resolution MRI radiomics model: superior predictive performance in preoperative T-staging of rectal cancer. *Eur Radiol*. 2023;33(1):1–10.
- Tian L, Li Z, Wu K, Dong P, Liu H, Wu S, Zhou F. The clinical significance of computed tomography texture features of renal cell carcinoma in predicting pathological T1–3 staging. *Quant Imaging Med Surg*. 2023;13(4):2415.
- Liang S, Gu Y. SRENet: a Spatiotemporal relationship-enhanced 2D-CNN-based framework for staging and segmentation of kidney cancer using CT images. *Appl Intell*. 2023;53(13):17061–73.
- Wang N, Bing X, Li Y, Yao J, Dai Z, Yu D, Ouyang A. Study of radiomics based on dual-energy CT for nuclear grading and T-staging in renal clear cell carcinoma. *Medicine*. 2024;103(10):e37288.
- He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*: 2016; 2016: 770–778.
- Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. *Advances in neural information processing systems*. *Adv Neural Inf Process Syst* 2017, 30(2017).
- Liang X, Wang X, Lei Z, Liao S, Li SZ. Soft-margin softmax for deep classification. In: *International Conference on Neural Information Processing: 2017*: Springer; 2017: 413–421.
- Zhang Z, Sabuncu M. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems* 2018, 31.
- Li T, Zhang Z, Zhu M, Cui Z, Wei D. Combining transformer global and local feature extraction for object detection. *Complex Intell Syst* 2024;1–24.
- Singh SP, Wang L, Gupta S, Goli H, Padmanabhan P, Gulyás B. 3D deep learning on medical images: a review. *Sensors*. 2020;20(18):5097.
- Wu K, Wu P, Yang K, Li Z, Kong S, Yu L, Zhang E, Liu H, Guo Q, Wu S. A comprehensive texture feature analysis framework of renal cell carcinoma: pathological, prognostic, and genomic evaluation based on CT images. *Eur Radiol* 2022;1–11.
- Hadjjiski N. Kidney cancer staging: Deep learning neural network based approach. In: *2020 International Conference on e-Health and Bioengineering (EHB): 2020*: IEEE; 2020: 1–4.
- Fateh SM, Arkawazi LA, Tahir SH, Rashid RJ, Rahman DH, Aghaways I, Kakamad FH, Salih AM, Bapir R, Fakhralddin SS. Renal cell carcinoma T staging: diagnostic accuracy of preoperative contrast-enhanced computed tomography. *Mol Clin Oncol*. 2023;18(2):1–7.
- Ökmen HB, Uysal H, Guvenis A. Automated prediction of TNM stage for clear cell renal cell carcinoma disease by analyzing CT images of primary tumors. In: *2019 11th International Conference on Electrical and Electronics Engineering (ELECO): 2019*: IEEE; 2019: 456–459.
- Talaat D, Zada F, Kadry R. Staging of clear cell renal cell carcinoma using random forest and support vector machine. In: *Journal of Physics: Conference Series*: 2020: IOP Publishing; 2020: 012012.
- Hussain MA, Hamarneh G, Garbi R. Learnable image histograms-based deep radiomics for renal cell carcinoma grading and staging. *Comput Med Imaging Graph*. 2021;90:101924.
- Demirjian NL, Varghese BA, Cen SY, Hwang DH, Aron M, Siddiqui I, Fields BK, Lei X, Yap FY, Rivas M. CT-based radiomics stratification of tumor grade and TNM stage of clear cell renal cell carcinoma. *Eur Radiol* 2022;1–12.
- Liu S, Roemer F, Ge Y, Bedrick EJ, Li Z-M, Guermazi A, Sharma L, Eaton C, Hochberg MC, Hunter DJ. Comparison of evaluation metrics of deep learning for imbalanced imaging data in osteoarthritis studies. *Osteoarthritis Cartil*. 2023;31(9):1242–8.
- Abou Elkassam AM, Lo SS, Gunn AJ, Shuch BM, Dewitt-Foy ME, Abouassaly R, Vaidya SS, Clark JI, Louie AV, Siva S. Role of imaging in renal cell carcinoma: a multidisciplinary perspective. *Radiographics*. 2021;41(5):1387–407.
- Zhu Y, Wu J, Long Q, Li Y, Luo H, Pang L, Zhu L, Luo H. Multimodal deep learning with MUF-net for noninvasive WHO/ISUP grading of renal cell carcinoma using CEUS and B-mode ultrasound. *Front Physiol*. 2025;16:1558997.
- Moon SW, Kim J, Kim YJ, Kim SH, An CS, Kim KG, Jung CK. Leveraging explainable AI and large-scale datasets for comprehensive classification of renal histologic types. *Sci Rep*. 2025;15(1):1745.
- Talaat FM, Gamel SA, El-Balka RM, Shehata M, ZainEldin H. Grad-cam enabled breast cancer classification with a 3d inception-resnet v2: empowering radiologists with explainable insights. *Cancers*. 2024;16(21):3668.
- Xia H, Yu J, Nie K, Yang J, Zhu L, Zhang S. CT radiomics and human-machine hybrid system for differentiating mediastinal lymphomas from thymic epithelial tumors. *Cancer Imaging*. 2024;24(1):163.
- Li X, Lin J, Qi H, Dai C, Guo Y, Lin D, Zhou J. Radiomics predict the WHO/ISUP nuclear grade and survival in clear cell renal cell carcinoma. *Insights into Imaging*. 2024;15(1):175.
- Li X, Guo Y, Huang S, Wang F, Dai C, Zhou J, Lin D. A CT-based intratumoral and peritumoral radiomics nomogram for postoperative recurrence risk stratification in localized clear cell renal cell carcinoma. *BMC Med Imaging*. 2025;25(1):167.
- Mali SA, Ibrahim A, Woodruff HC, Andrearczyk V, Müller H, Primakov S, Salahuddin Z, Chatterjee A, Lambin P. Making radiomics more reproducible across scanner and imaging protocol variations: a review of harmonization methods. *J Personalized Med*. 2021;11(9):842.
- Wang X, Zhang A, Yang H, Zhang G, Ma J, Ye S, Ge S. Multicenter development of a deep learning radiomics and dosiomics nomogram to predict radiation pneumonia risk in non-small cell lung cancer. *Sci Rep*. 2025;15(1):17106.

Publisher's Note

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