Journal of Imaging Informatics in Medicine (2025) 38:1323–1333
https://doi.org/10.1007/s10278-024-01276-7

ORIGINAL PAPER

A Robust Deep Learning Method with Uncertainty Estimation for the Pathological Classification of Renal Cell Carcinoma Based on CT Images

Ni Yao[1] · Hang Hu[1] · Kaicong Chen[2] · Huan Huang[1] · Chen Zhao[3,6] · Yuan Guo[5] · Boya Li[2] · Jiaofen Nan[1] · Yanting Li[1] · Chuang Han[1] · Fubao Zhu[1] · Weihua Zhou[3,4] · Li Tian[2]

Received: 15 May 2024 / Revised: 4 September 2024 / Accepted: 15 September 2024 / Published online: 23 September 2024

Abstract

This study developed and validated a deep learning–based diagnostic model with uncertainty estimation to aid radiologists in the preoperative differentiation of pathological subtypes of renal cell carcinoma (RCC) based on computed tomography (CT) images. Data from 668 consecutive patients with pathologically confirmed RCC were retrospectively collected from Center 1, and the model was trained using fivefold cross-validation to classify RCC subtypes into clear cell RCC (ccRCC), papillary RCC (pRCC), and chromophobe RCC (chRCC). An external validation with 78 patients from Center 2 was conducted to evaluate the performance of the model. In the fivefold cross-validation, the area under the receiver operating characteristic curve (AUC) for the classification of ccRCC, pRCC, and chRCC was 0.868 (95% CI, 0.826–0.923), 0.846 (95% CI, 0.812–0.886), and 0.839 (95% CI, 0.802–0.88), respectively. In the external validation set, the AUCs were 0.856 (95% CI, 0.838–0.882), 0.787 (95% CI, 0.757–0.818), and 0.793 (95% CI, 0.758–0.831) for ccRCC, pRCC, and chRCC, respectively. The model demonstrated robust performance in predicting the pathological subtypes of RCC, while the incorporated uncertainty emphasized the importance of understanding model confidence. The proposed approach, integrated with uncertainty estimation, offers clinicians a dual advantage: accurate RCC subtype predictions complemented by diagnostic confidence metrics, thereby promoting informed decision-making for patients with RCC.

Keywords Deep learning · Renal cell carcinoma · Pathological classification · Computed tomography

Introduction

Renal cell carcinoma (RCC) accounts for approximately 90–95% of renal cancers [1] and is often classified into several histological subtypes based on their morphological characteristics. Clear cell RCC (ccRCC) is the most common type, accounting for 75% of all cases. This subtype is notably aggressive and has a poor prognosis. Following ccRCC, the second and third most frequent subtypes are papillary RCC (pRCC) and chromophobe RCC (chRCC), making up 10–15% and 5% of the cases, respectively [2]. RCC subtypes range from indolent to aggressive with associated treatment

✉ Weihua Zhou
whzhou@mtu.edu

✉ Li Tian
tianli@sysucc.org.cn

[1] School of Computer Science and Technology, Zhengzhou University of Light Industry, Zhengzhou 450002, Henan, China

[2] Department of Medical Imaging, State Key Laboratory of Oncology in South China, Guangdong Provincial Clinical Research Center for Cancer, Collaborative Innovation Center for Cancer Medicine, Sun Yat-Sen University Cancer Center, Guangzhou 510060, Guangdong, China

[3] Department of Applied Computing, Michigan Technological University, Houghton, MI, USA

[4] Center for Biocomputing and Digital Health, Institute of Computing and Cybersystems, and Health Research Institute, Michigan Technological University, Houghton, MI, USA

[5] Department of Radiology, The First People's Hospital of Guangzhou, Guangzhou 510180, Guangdong, China

[6] Department of Computer Science, Kennesaw State University, Marietta, GA, USA

considerations [3]. In addition, the selection of targeted drug therapy and immunotherapy for advanced tumors is based on RCC subtypes [4, 5]. Consequently, accurate pathological classification is crucial for patients with RCC, in terms of both prognosis and therapeutic strategies [6].

Percutaneous renal biopsy provides diagnostic value for classifying the most common types of RCC [7]. However, biopsy is invasive, associated with complications, and may be limited by tumor location and the optimal timing for the procedure. Growing evidence suggests that CT plays a crucial role in RCC treatment, from diagnosis and staging the disease to assessing the response to treatment [8]. However, significant overlaps exist in image-level features between renal tumor subtypes, which complicates subtype classification and introduces interobserver variability [9]. These clinical challenges highlight the need for automated systems to reduce misdiagnosis costs and assist radiologists in medical image interpretation [10].

In recent years, deep learning, particularly convolutional neural networks (CNNs), has shown great potential in enhancing the efficiency of radiologists for complex medical image analysis, especially in the field of RCC [11–13]. Many studies have focused on the detection and segmentation of organs in medical images, achieving high accuracy rates, often exceeding 90% [14–16], with specific performance in tumor detection reaching an accuracy of 87.5% for tumors smaller than 4 cm [17]. Additionally, deep learning models are employed for classification tasks, such as distinguishing between benign and malignant tumors or performing differential diagnosis of RCC histological subtypes. For example, Xi et al. developed a residual CNN classifier to differentiate benign renal tumors from RCC on MR imaging, achieving an AUC of 0.99 on the training set and 0.77 on the validation set [18]. This study demonstrates the potential of deep learning models for RCC classification but also highlights limitations, such as the dependency on large annotated datasets and the risk of overfitting. Liu et al. [19] employed ResNet-50 as the backbone network for RCC subtype classification, achieving AUCs of 0.87 for ccRCC, 0.86 for pRCC, and 0.85 for chRCC. Another study by Han et al. proposed a deep learning framework for classifying RCC into three major subtypes [20]. This framework simplified the classification task into binary comparison: chRCC vs. non-chRCC, pRCC vs. non-pRCC, and ccRCC vs. non-ccRCC, achieving AUCs of 0.87, 0.91, and 0.93, respectively. However, such simplification may fail to capture the complex interrelationships between the three subtypes. Additionally, reliance on manually labeled regions of interest (ROIs) introduces variability and potential subjectivity, which may affect overall model performance.

Moreover, while many studies report high accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve, they often overlook the importance of uncertainty estimation in their outputs. Uncertainty estimation provides a measure of reliability for a model's predictions, allowing clinicians to better understand and trust the model's classification decisions, especially in complex or borderline cases. A recent study on lung cancer detection utilized an uncertainty-aware framework incorporating data preprocessing and uncertainty quantification techniques, such as Monte Carlo Dropout, Deep Ensemble, and Ensemble Monte Carlo Dropout, achieving an accuracy of up to 0.959 [21]. However, the study did not fully explore the implications of uncertainty on clinical decision-making outcomes.

Therefore, from an evidence-based medicine perspective, our goal is not only to develop an efficient diagnostic model but also to ensure that the model can sensitively identify difficult cases and reduce the workload of radiologists by providing reliable confidence metrics for each classification decision. Against this background, we aimed to propose a framework for the automatic pathological classification of RCC based on CT images by combining uncertainty estimation and deep learning, including a target detection model specifically designed to identify areas of renal tumors and a classification model to distinguish the three major pathological subtypes of RCC.

Materials and Methods

Study Population

This retrospective study was approved by the Ethics Committee of both the Sun Yat-sen University Cancer Center and the First People's Hospital of Guangzhou, and informed consent was waived for all patients. All data have been deposited at Sun Yat-sen University Cancer Center (https://www.researchdata.org.cn/; Research Data Deposit; No. RDDA2023739917).

We reviewed the medical records from January 2016 to November 2022 at the Sun Yat-sen University Cancer Center (Center 1) and from February 2017 to June 2023 at the First People's Hospital of Guangzhou (Center 2). The included patients had histopathologically confirmed diagnoses of ccRCC, pRCC, or chRCC (Fig. 1). The following criteria were applied for data exclusion: (1) incomplete 4-phase CT scanning (unenhanced, corticomedullary, nephrographic, and excretory phases); (2) poor image quality (such as motion artifacts and metal artifacts); and (3) the presence of ambiguous or contradictory clinical data. Finally, 668 consecutive patients from Center 1 (ccRCC, $n = 395$; pRCC, $n = 167$; chRCC, $n = 106$) and 78 patients from Center 2 (ccRCC, $n = 49$; pRCC, $n = 11$; chRCC, $n = 17$) were included.

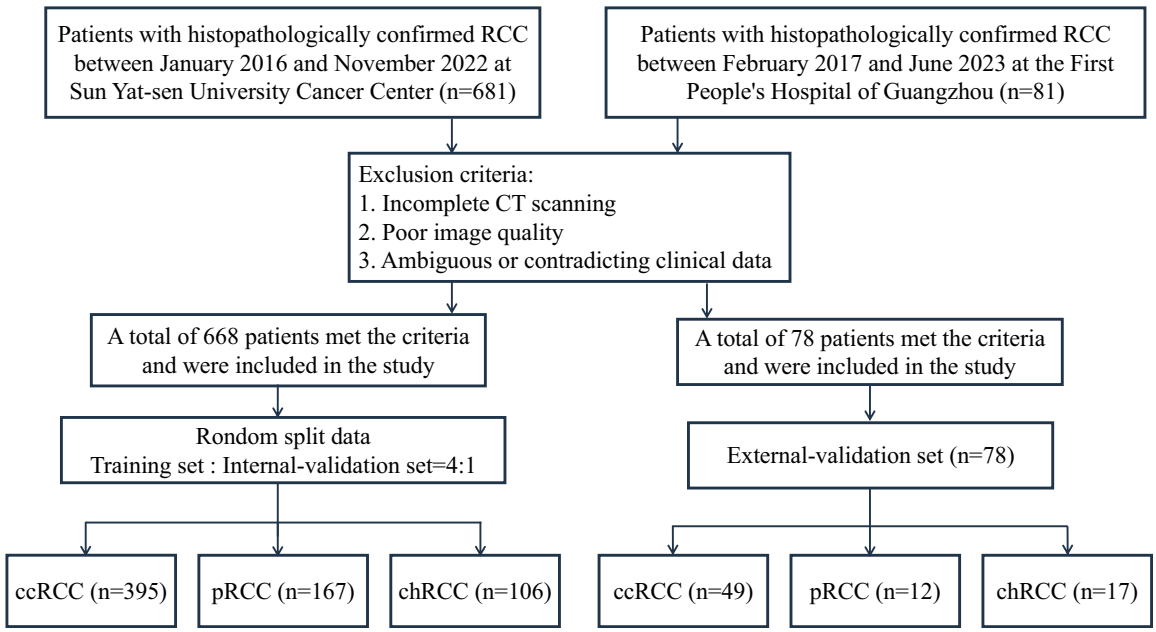

Fig. 1 Flowchart of patient recruitment. ccRCC, clear cell renal cell carcinoma; pRCC, papillary renal cell carcinoma; chRCC, chromophobe renal cell carcinoma

Image Acquisition

All patients underwent CT examinations using multi-slice spiral CT scanners (Center 1: Somatom Force, Siemens Healthineers, Forchheim, Germany; Center 2: Aquilion 64, Toshiba, Tokyo, Japan). Data acquisition was performed in three phases: the precontrast phase (PCP), corticomedullary phase (CMP; 30-s delay after contrast injection), and nephrographic phase (NP; 90-s delay after contrast injection). After the unenhanced scan, an intravenous bolus injection of nonionic contrast media (350 mg/mL; 1.5 mL/kg) was administered at a rate of 4 mL/s.

Data Preprocessing

The CT slices for each patient were preprocessed to better analyze the renal parenchymal regions, with a window width and level set of 300 and 40, respectively. The CT images were then resampled to obtain uniform images with a voxel size of $1 \times 1 \times 1\ mm^3$. The voxel values were then normalized to the range of 0 to 1.

Tumor Detection

We utilized an automated method to acquire the bounding boxes of the tumors in our dataset. This method allows the model to concentrate on CT image RoIs for the subsequent classification of subtypes. A YOLO object detection model using PyTorch 1.7 framework and Python was built to produce renal tumor RoIs [22]. This model was trained and validated using the publicly available KiTS19 dataset [23]. Data augmentation techniques, including rotation, flipping, and changes in scale to augment the training dataset, were used to reduce overfitting. The model achieved a mean average precision of 0.87 using a threshold of 0.5 for the intersection over union (IoU), which is a commonly used criterion to evaluate whether the predicted bounding box sufficiently overlaps with the true tumor region.

CT slices without kidney tumors were excluded from further analysis. The detection results for each slice were merged to cover the entire tumor in three-dimensional volumes of interest (VoIs). These cropped VoIs were reviewed by a trained technician for validation. They were confirmed by an experienced radiologist and manually corrected if there were mistakes. These VoIs were then inputted into the classification model.

Tumor Classification

The structure of the renal cell carcinoma subtype classification network is shown in Fig. 2. A 3D convolutional neural network adapted from the architecture proposed by Wang et al. [24] was used. This model incorporates two residual blocks designed for efficient feature extraction and gradient propagation. Each convolutional layer contains multiple channels to extract distinct features from the input and preserve spatial information from the CT volume. Instead of using a softmax layer, our architecture outputs nonnegative

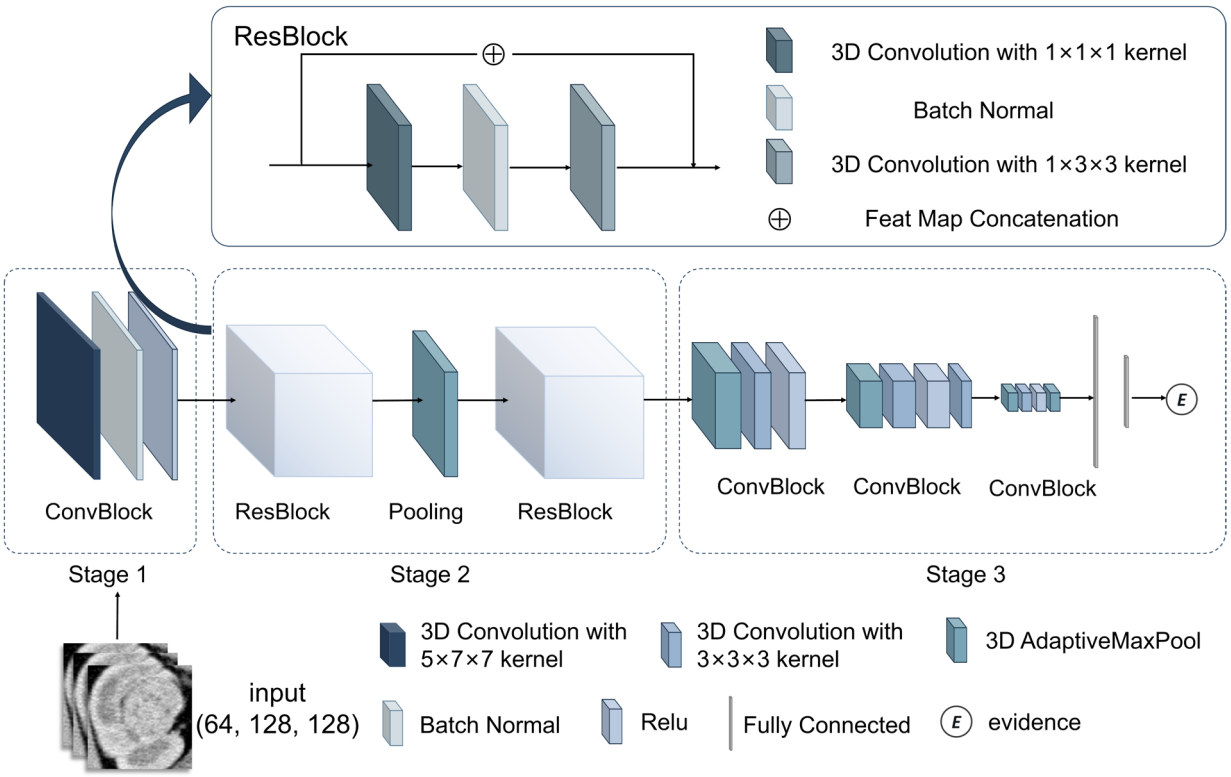

Fig. 2 Structure of the 3D convolutional neural network for the renal cell carcinoma subtype classification. Stage 1, 3D convolutional block layer; Stage 2, two 3D residual blocks with a pooling layer; Stage 3, output layer producing nonnegative values for uncertainty estimation

continuous values. These values are treated as evidence, which provides a basis for the Dirichlet distribution parameters in our evidence-based uncertainty estimation [25].

The model was trained using a Tesla V100 GPU with 32 GB of GPU memory. The training was conducted using the PyTorch framework. Optimization was performed with an adaptive moment optimizer (Adam) and a learning rate of 0.0001. The batch size was set to 32, and the number of training epochs was set to 300. To improve the model's generalization, dropout regularization with a rate of 0.5 was applied. Additionally, real-time data augmentation techniques, including random rotations, flips, and shifts, were employed during training to enhance model robustness and prevent overfitting.

Uncertainty Quantification and Loss Function

We employed an approach based on Dirichlet distribution [25] to quantify the uncertainty in our model predictions. Central to this method is the translation of our model's outputs, which we term as evidence, into parameters for the Dirichlet distribution. The Dirichlet distribution relies on strictly positive parameters to meaningfully represent the strength or support for different classes. To achieve this, especially when the evidence obtained from our model approached zero, we added a constant of 1 to the evidence. Based on this foundation, the uncertainty u is quantified by the equation:

$$u = \frac{K}{\sum_{j=1}^{K}(e_j+1)}$$

where e_j corresponds to the evidence associated with the j -th class.K representing the total number of classes serves to normalize the uncertainty measure, ensuring that it remains proportionate and avoids inflate with an increase in the number of classes. In our study, K is 3.

Another crucial component of our uncertainty quantification method is the design of the loss function. The selected loss function serves a dual purpose: it aligns the model's output with the Dirichlet distribution for uncertainty estimation, while addressing and mitigating issues related to class imbalances. Class imbalances arise when certain classes have considerably fewer samples than others, potentially causing the model to exhibit a bias toward predicting classes with larger sample sizes. In our datasets, there is a notable imbalance with a greater number of ccRCC samples compared to the other two classes. The loss function is expressed as follows:

$$L_i = \sum_{j=1}^{K} y_{ij}\left(\psi\left(\sum_{j=1}^{K}(e_j+1)\right) - \psi(e_{ij}+1)\right)w_j$$

where i represents the i-th subject and y_{ij} is the one-hot encoding of the actual class for the i-th subject.

w_j represents the weight of the j-th class of the loss function. In this method, it is chosen to be the reciprocal of the number of samples for its respective class. The function $\psi(\cdot)$ represents the logarithmic derivative of the gamma function, which transforms the model's outputs into probabilistic estimates, thereby facilitating the expression of the prediction uncertainty.

Evaluation

The evaluation metrics consisted of two parts: the evaluation metrics of RCC subtyping and the metrics used to analyze the effectiveness of the introduced uncertainty grades.

The primary classification evaluation metric for both internal and external datasets is the AUC of the receiver operating characteristic curve. AUC values from fivefold cross-validation are presented with 95% confidence intervals (CI). In addition to the AUC, we also report accuracy, sensitivity, and specificity. All results are reported as the mean ± SD based on fivefold cross-validation SD standard deviation. The performance metric values were evaluated using the following equations.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

$$Sensitivity = \frac{TP}{TP+FN}$$

$$Specificity = \frac{TN}{TN+FP}$$

where TP, TN, FP, and FN are the true-positive, true-negative, false-positive, and false-negative counts, respectively.

Given that the range of uncertainty spans from 0 to 1, we evenly distributed this range to define the five uncertainty grades. To evaluate the effectiveness of incorporating uncertainty into the model, the correct rate for each grade was calculated based on the number of accurate predictions relative to the total number of predictions within each respective grade.

Statistical Analysis

Statistical analyses were conducted using Python (version 3.10, https://www.python.org/) with the SciPy package. The Kruskal–Wallis test was used to compare continuous variables across different groups, while the chi-squared test was applied to categorical variables. Effect sizes were computed using epsilon squared (ε^2) for the Kruskal–Wallis test. ε^2 values indicate the proportion of variance in the dependent variable explained by the independent variable, with small effects typically considered as $\varepsilon^2 < 0.01$, medium effects as $0.01 \le \varepsilon^2 < 0.06$, and large effects as $\varepsilon^2 \ge 0.06$. Performance metrics such as sensitivity, specificity, and accuracy were compared between the internal and external validation sets using t-tests, and two-sided p-values were calculated. Statistical significance was defined as a p-value less than 0.05.

Results

Population

Table 1 details the demographics and kidney tumor subtype distribution for 746 patients (mean age 53.3 ± 12.5 years) used in our study. The datasets were sourced from Sun Yat-sen University Cancer Center (training and internal validation sets) and the First People's Hospital of Guangzhou (external validation set). In these datasets, RCC subtypes were consistently distributed, with ccRCC being the predominant subtype. The overall tumor diameter was 53.8 ± 30.3 mm, and subtype-specific diameters were 51.9 ± 28.3 mm for ccRCC, 61.2 ± 36.9 mm for pRCC, and 54.1 ± 30.2 mm for chRCC. The maximum tumor diameter averaged 53.6 ± 31.0 mm in the training set, 55.3 ± 28.8 mm in the internal validation set, and 51.9 ± 27.5 mm in the external validation set. Group comparisons were performed using the Kruskal–Wallis test for continuous variables and the chi-squared test for categorical variables. The last column presents the corresponding p-values from these tests. Effect sizes were calculated using epsilon squared (ε^2) for the Kruskal–Wallis test.

Classification Performance

In the internal validation dataset, the deep learning model achieved AUCs of 0.874 ± 0.039 for ccRCC, 0.849 ± 0.03 for pRCC, and 0.841 ± 0.031 for chRCC. In the external validation set, the AUCs were 0.860 ± 0.018, 0.787 ± 0.025, and 0.795 ± 0.029 for ccRCC, pRCC, and chRCC, respectively. Figure 3 presents the receiver operating characteristic curves for each RCC subtype, based on the model's discrimination for each classification. These curves represent outcomes from the fivefold cross-validation, showing the performance of the model across different partitions of the dataset.

Additional performance metrics, including accuracy, sensitivity, and specificity for each subtype, are listed in Table 2. These metrics further provide insight into the model's ability to differentiate RCC subtypes across the internal and external datasets.

Table 1 Patient characteristics and tumor size distribution on training and validation sets

Characteristic	All	Training	Internal validation	External validation	*p*-value
Patients,*n*	746	534	134	78	
Age,*n*(%)					0.001*
<40	117 (16)	92 (21)	22 (16)	3 (4)	
40–60	421 (56)	295 (68)	74 (55)	52 (67)	
>60	208 (28)	147 (11)	38 (28)	23 (29)	
Gender, *n* (%)					0.650
Female	260 (35)	181 (34)	51 (38)	28 (36)	
Male	486 (65)	353 (66)	83 (62)	50 (64)	
Subtype,*n*(%)					0.146
ccRCC	444 (59)	316 (59)	79 (59)	49 (63)	
pRCC	118 (16)	85 (16)	21 (16)	12 (15)	
chRCC	184 (25)	133 (25)	34 (25)	17 (22)	
Maximum tumor diameter (mm)					0.472
<30	151 (20)	111 (21)	26 (19)	14 (18)	
30–50	252 (34)	185 (34)	39 (29)	28 (36)	
50–70	168 (23)	112 (21)	34 (25)	22 (28)	
>70	175 (23)	126 (24)	35 (26)	14 (18)	

*The training and internal validation datasets are from fold 1 in the fivefold cross-validation procedure. Although there are statistically significant differences in age across the groups ($p=0.001$), the effect size ($\varepsilon^2=0.016$) suggests that the practical impact of these differences is moderate but limited

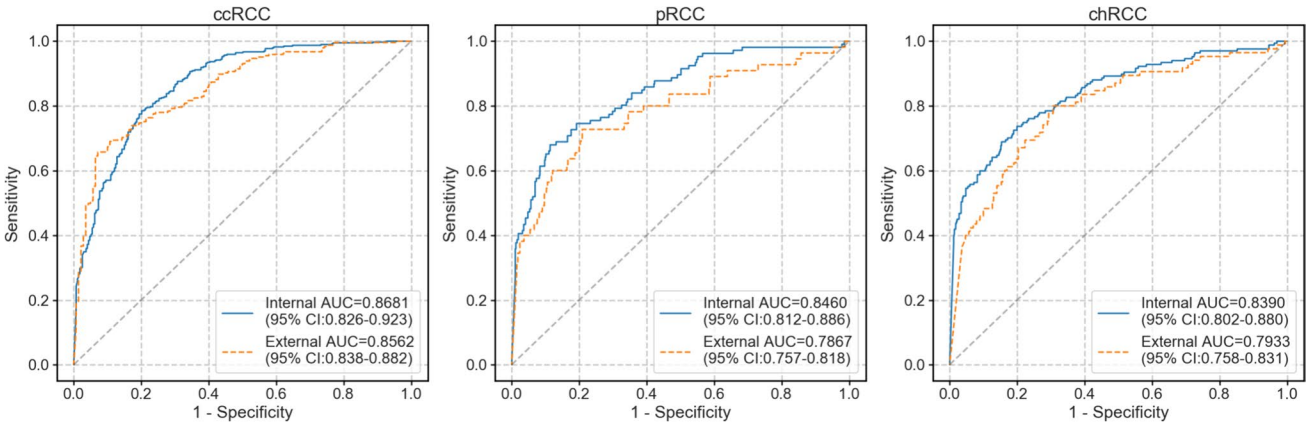

Fig. 3 Receiver operating characteristic curves for fivefold cross-validation of internal and external sets

Table 2 The performance of the model in classifying renal cell carcinoma subtype classification on internal and external validation sets

		Accuracy	Sensitivity	Specificity	AUC
Internal validation set	ccRCC	0.791 ± 0.016	0.828 ± 0.082*	0.736 ± 0.111	0.874 ± 0.039
	pRCC	0.871 ± 0.024	0.566 ± 0.042	0.929 ± 0.032*	0.849 ± 0.030*
	chRCC	0.815 ± 0.030*	0.634 ± 0.106	0.874 ± 0.064	0.841 ± 0.031*
External validation set	ccRCC	0.763 ± 0.028	0.727 ± 0.043	0.829 ± 0.064	0.860 ± 0.018
	pRCC	0.826 ± 0.040	0.600 ± 0.050	0.864 ± 0.043	0.787 ± 0.025
	chRCC	0.771 ± 0.027	0.600 ± 0.097	0.820 ± 0.046	0.795 ± 0.029
p-value	ccRCC	0.098	<0.05	0.143	0.460
	pRCC	0.060	0.279	<0.05	<0.05
	chRCC	<0.05	0.609	0.165	<0.05

**T*-tests were used to compare the differences in performance metrics between the internal and external validation sets

Table 3 Performance metrics of the model for low-uncertainty grade predictions (uncertainty < 0.2) on internal and external validation sets

		Accuracy	Sensitivity	Specificity	AUC
Internal validation set	ccRCC	0.953 ± 0.046	0.948 ± 0.071	0.954 ± 0.069	0.964 ± 0.037
	pRCC	0.957 ± 0.026*	0.935 ± 0.093*	0.960 ± 0.040	0.948 ± 0.068*
	chRCC	0.939 ± 0.017*	0.876 ± 0.099	0.964 ± 0.033*	0.913 ± 0.040
External validation set	ccRCC	0.893 ± 0.041	0.837 ± 0.105	0.950 ± 0.049	0.922 ± 0.044
	pRCC	0.902 ± 0.038	0.753 ± 0.143	0.931 ± 0.054	0.820 ± 0.067
	chRCC	0.860 ± 0.050	0.856 ± 0.094	0.863 ± 0.085	0.844 ± 0.060
p-value	ccRCC	0.063	0.084	0.922	0.141
	pRCC	< 0.05	< 0.05	0.353	< 0.05
	chRCC	< 0.05	0.752	< 0.05	0.064

**T*-tests were used to compare the differences in performance metrics between the internal and external validation sets

To further evaluate the model's reliability, we analyzed the performance metrics for cases with uncertainty < 0.2, as shown in Table 3. For these low-uncertainty cases, the accuracy, sensitivity, specificity, and AUC were calculated separately for each RCC subtype in both the internal and external validation sets. This analysis aims to highlight the model's predictive performance when uncertainty grade is minimal, providing clinicians with more reliable decision-making information.

In comparison with all cases presented in Table 2, the performance metrics for low-uncertainty cases show an overall improvement across all indicators. Specifically, the accuracy, sensitivity, specificity, and AUC for each RCC subtype are higher in these low-uncertainty cases, reflecting the model's enhanced predictive performance when the uncertainty grade is minimal.

Uncertainty Analysis

In our deep learning model, uncertainty estimation is crucial for assessing the reliability of the subtype predictions. This evaluation was based on the best model obtained through fivefold cross validation.

Figure 4 illustrates the kernel density estimates (KDE) of uncertainty distributions across different validation sets. For the internal validation set, the median uncertainties were 0.24 for ccRCC, 0.22 for pRCC, and 0.19 for chRCC. In the external validation set, these values were slightly higher: 0.26 for ccRCC, 0.24 for pRCC, and 0.23 for chRCC. Figure 5 displays the uncertainty grades and their correlation with the correct rates for both the validation sets. Overall, as the uncertainty grade decreases, the rate of accurate predictions tends to increase. Figure 6 shows the overall correct rate between the two validation sets across varying uncertainty grades. The correct rate for grade 1 was 89.29% ($n = 134$) in the internal validation set and 77.78% ($n = 78$) in

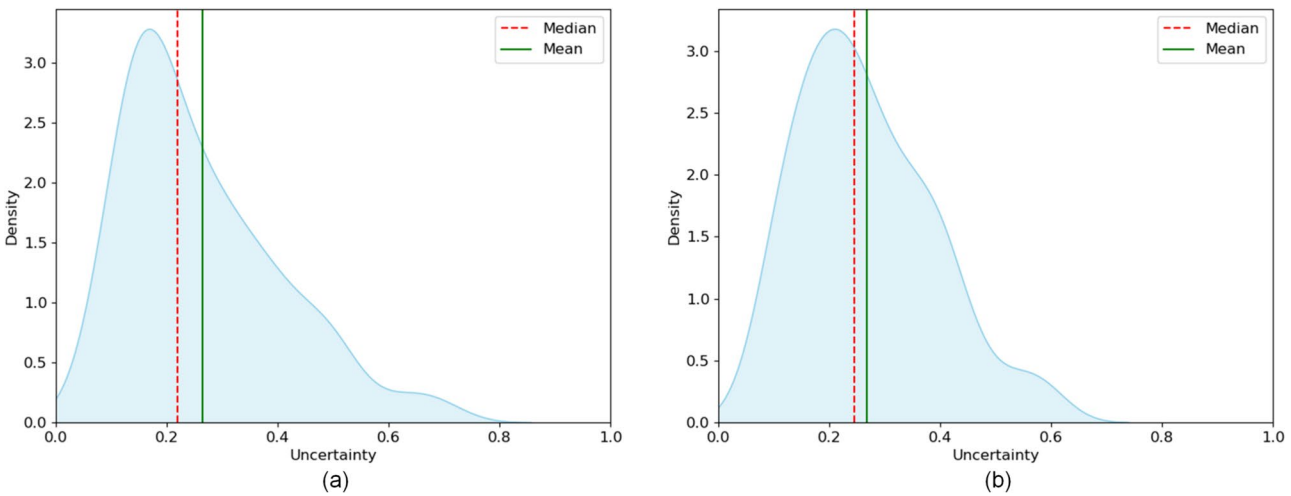

Fig. 4 Kernel density estimates of uncertainty distributions. **a** Internal validation set. **b** External validation set

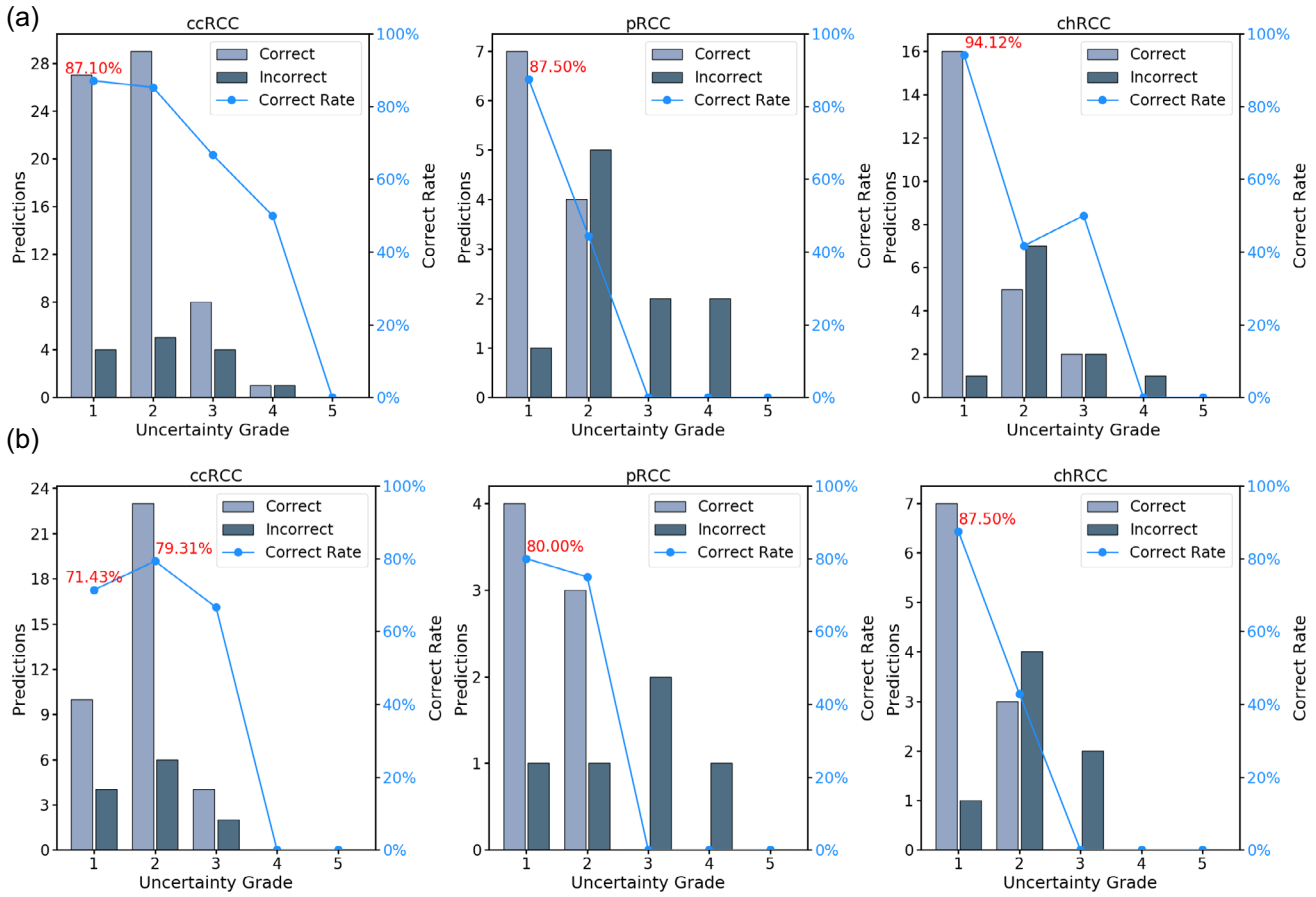

Fig. 5 Uncertainty-graded prediction performance for kidney cancer subtypes. **a** Internal validation set. **b** External validation set. Uncertainty grades 1–5 have an uncertainty range of 0–0.2, 0.2–0.4, 0.4–0.6, 0.6–0.8, and 0.8–1, respectively

the external validation set. Predictions with high uncertainty grades (4–5, $n = 5$) showed an incorrect rate of 80%, indicating the importance of uncertainty in identifying potentially challenging diagnostic scenarios. Several anomalous cases are shown in Fig. 7.

Discussion

Accurately classifying the pathological subtypes of RCC has profound implications for patient management and therapeutic strategies. In clinical practice, although the accuracy of deep learning–based diagnostic tools is necessary, it is insufficient. If a model only provides predictions without indicating its uncertainty, it can mislead radiologists, particularly when diagnosing intractable diseases [26]. This omission can result in radiologists overlooking the need for further tests, potentially affecting patient diagnosis and treatment. In response to this, our study introduced a deep learning-based model with uncertainty estimation for RCC subtype classification.

Incorporating uncertainty into our model distinctly contrasts with traditional models, which often provide a sole classification in medical imaging [27–29]. Such an approach can lead to overreliance on the model, potentially resulting in misdiagnoses, especially when RCC subtypes have similar phenotypic manifestations. Our method not only offers classification but also sheds light on the model's confidence in its prediction. This dual output offers a new dimension to clinical practice, providing clinicians with a more comprehensive understanding of the RCC subtype prediction, thereby aiding in more informed decisions. Moreover, by presenting both subtype prediction and its associated uncertainty, the model effectively reduces the manual evaluation burden on clinicians. This streamlined process not only makes the diagnostic procedure more efficient, but also bolsters the clinician's confidence in adopting the model's recommendations, thereby striking a balance between automation and human expertise.

In our results, a lower uncertainty grade was consistently associated with higher correct rates; a trend was also observed by Song et al. [30] and Ahsan et al. [31]. This demonstrates the advantage of incorporating uncertainty

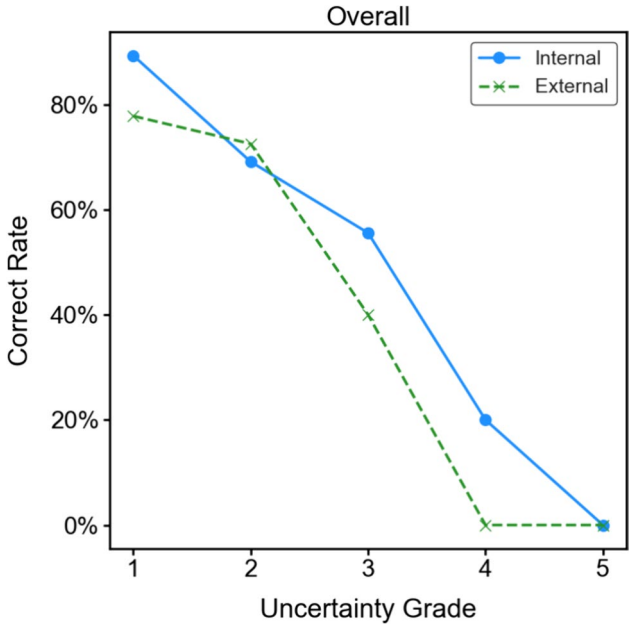

Fig. 6 Comprehensive correct rate comparison of all subtypes on internal and external validation sets for different grades of uncertainty. Uncertainty grades 1–5 have an uncertainty range of 0–0.2, 0.2–0.4, 0.4–0.6, 0.6–0.8, and 0.8–1, respectively

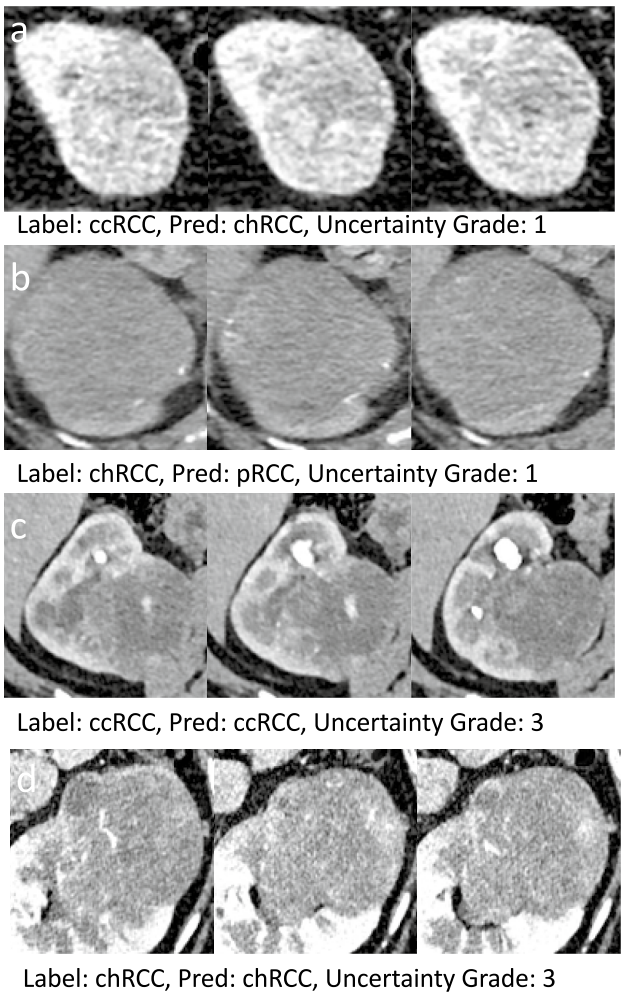

Fig. 7 Four anomalous cases. The cases have an incorrect prediction with a low uncertainty in **a** and **b** and a correct prediction with a high uncertainty in **c** and **d**. Uncertainty grades 1–5 have an uncertainty range of 0–0.2, 0.2–0.4, 0.4–0.6, 0.6–0.8, and 0.8–1, respectively

estimation to enhance the diagnostic performance. At the lowest uncertainty grade (grade 1), the correct rate for the internal validation set reached 89.29%, and for the external validation set, 77.78%. Thiagarajan et al. [32] employed a low-dimensional visualization techniques and observed that the low uncertainty images revealed distinct class separations, a manifestation of high classification confidence, reinforcing our own observations. Moreover, we calculated the optimal uncertainty threshold using the Youden Index [33], which revealed a value of 0.2. This aligns with uncertainty grade 1, reinforcing the efficacy of this uncertainty grade in predicting subtypes accurately. However, at the highest uncertainty grade (grade 5), the correct rate for both validation sets dropped to 0. Research by Rączkowska et al. suggested that images with the highest grades of uncertainty often display features that are pathologically challenging to classify [34], providing a potential explanation for the observations. Analyzing the data by subtype, the model performed the best for chRCC with correct rates of 94.12% for the internal validation set and 87.50% for the external validation set at grade 1. The rates of pRCC were 87.5% and 80%, respectively. In contrast, the performance for ccRCC at grade 1 was lower, even though it showed the highest AUC among the three subtypes. One potential factor contributing to this could be the weight settings introduced during the model training. Intended to balance the influence of each subtype, weight adjustments might have unintentionally increased the prediction uncertainty for ccRCC. Specifically, the reduced weight for ccRCC, despite its larger dataset, could have made the model more cautious in its predictions for this subtype. This cautiousness might manifest as higher uncertainty, particularly when the model encounters ccRCC cases that resemble to other subtypes.

Our deep learning model exhibited a robust performance across diverse RCC subtypes and validation sets, demonstrating its potential reliability for clinical deployment. While deep learning has been extensively applied in medical imaging, research specifically targeting RCC subtype classification is sparse. Most existing studies primarily focus on distinguishing between benign and malignant tumors, or broader classifications such as differentiating ccRCC from non-ccRCC [35]. Traditionally, RCC classification has been largely dependent on machine learning methods, which often require significant manual oversight, spanning from data preprocessing to feature extraction [36, 37]. In contrast, our

model provides an end-to-end solution, substantially reducing the need for human intervention, enhancing the model's applicability in practice through uncertainty estimation, and serves as a significant step toward the clinical application of automated RCC pathological subtype diagnosis using deep learning. Although the model in the tumor detection stage achieved a mean average precision of 0.87 using a threshold of 0.5 for the IoU, a high IoU is not necessary since the aim is to define a bounding box that approximately covers the tumor. Manual correction can be applied when the bounding box does not cover the entire tumor.

Our study had several limitations. First, our study was intended to use CT images from the CMP, variations in scanning delays occasionally yielded images that leaned toward PCP or NP. The inclusion of these subjects with deviation could potentially influence the model's performance, given that deep learning outcomes are closely tied to the quality and consistency of the training data. Second, while we introduced a measure of reliability by estimating uncertainty, the "black box" nature of deep learning models remains a challenge to interpret. Third, to mitigate data imbalance, we adjusted weights based on the size of each class. While a common approach, this might have unintentionally affected the model's confidence. Last, our research primarily focuses on a specific imaging modality. Future work could benefit from incorporating diverse modalities, such as MRI and histology images, to further enhance our model's diagnostic capabilities.

Funding This study received support from the National Natural Science Foundation of China (Grant Numbers: 62476255, 82370513, 62303427, and 62106233), the Henan Science and Technology Development Plan (Grant Number: 242102211018, 232102210010, 232102210062, 232102211003, and 222102210219), and the Young Backbone Teacher Program of Zhengzhou University of Light Industry.

Declarations

Ethics Approval Institutional review board approval was obtained.

Consent to Participate Written informed consent was waived by the institutional review board.

Consent for Publication The authors affirm that human research participants provided informed consent for publication of the images in Fig. 7.

Conflict of Interest The authors declare no competing interests.

References

1. Ghidini M, Petrelli F, Ghidini A, et al.: Clinical development of mTor inhibitors for renal cancer. Expert Opin Investig Drugs 26(11):1229–1237, 2017.
2. Liu H, Cao H, Chen L, et al.: The quantitative evaluation of contrast-enhanced ultrasound in the differentiation of small renal cell carcinoma subtypes and angiomyolipoma. Quant Imaging Med Surg 12(1):106–118, 2022.
3. Robila V, Kraft AO, Smith SC: New entities, new technologies, new findings: A review of the cytologic features of recently established subtypes of renal cell carcinoma. Cancer Cytopathol 127(2):79–97, 2019.
4. Ghali F, Patel SH, Derweesh IH: Current Status of Immunotherapy for Localized and Locally Advanced Renal Cell Carcinoma. J Oncol 2019:7309205, 2019.
5. Berquist SW, Yim K, Ryan ST, et al.: Systemic therapy in the management of localized and locally advanced renal cell carcinoma: Current state and future perspectives. Int J Urol 26(5):532–542, 2019.
6. Yunker A, Holder L, Nething J: Newly described eosinophilic, solid and cystic renal cell carcinoma: a case report and review of the literature. Arch Nephrol Urol 3(2):38–45, 2020.
7. Rasmussen R, Sanford T, Parwani AV, et al.: Artificial Intelligence in Kidney Cancer. Am Soc Clin Oncol Educ Book 42:1–11, 2022.
8. Tian L, Li Z, Wu K, et al.: The clinical significance of computed tomography texture features of renal cell carcinoma in predicting pathological T1–3 staging. Quant Imaging Med Surg 13(4):2415–2425, 2023.
9. Uhm K-H, Jung S-W, Choi MH, et al.: Deep learning for end-to-end kidney cancer diagnosis on multi-phase abdominal computed tomography. NPJ Precis Oncol 5(1):54, 2021.
10. Kaur R, Juneja M, Mandal AK: Computer-aided diagnosis of renal lesions in CT images: a comprehensive survey and future prospects. Comput Electr Eng 77:423–434, 2019.
11. Comes MC, Fucci L, Mele F, et al.: A deep learning model based on whole slide images to predict disease-free survival in cutaneous melanoma patients. Sci Rep 12(1):20366, 2022.
12. Suarez-Ibarrola R, Hein S, Reis G, et al.: Current and future applications of machine and deep learning in urology: a review of the literature on urolithiasis, renal cell carcinoma, and bladder and prostate cancer. World J Urol 38(10):2329–2347, 2020.
13. Hussain MA, Hamarneh G, Garbi R: Learnable image histograms-based deep radiomics for renal cell carcinoma grading and staging. Comput Med Imaging Graph 90:101924, 2021.
14. Yang E, Kim CK, Guan Y, et al.: 3D multi-scale residual fully convolutional neural network for segmentation of extremely large-sized kidney tumor. Comput Methods Programs Biomed 215:106616, 2022.
15. Zhao C, Keyak JH, Tang J, et al.: ST-V-Net: incorporating shape prior into convolutional neural networks for proximal femur segmentation. Complex Intell Syst 9(3):2747–2758, 2023.
16. Zhao C, Xu Y, He Z, et al.: Lung segmentation and automatic detection of COVID-19 using radiomic features from chest CT images. Pattern Recognit 119:108071, 2021.
17. Toda N, Hashimoto M, Arita Y, et al.: Deep Learning Algorithm for Fully Automated Detection of Small (≤4 cm) Renal Cell Carcinoma in Contrast-Enhanced Computed Tomography Using a Multicenter Database. Invest Radiol 57(5):327–333, 2022.
18. Xi IL, Zhao Y, Wang R, et al.: Deep Learning to Distinguish Benign from Malignant Renal Lesions Based on Routine MR Imaging. Clin Cancer Res 26(8):1944–1952, 2020.
19. Liu J, Yildirim O, Akin O, et al.: AI-Driven Robust Kidney and Renal Mass Segmentation and Classification on 3D CT Images. Bioengineering 10(1):116, 2023.
20. Han S, Hwang SI, Lee HJ: The Classification of Renal Cancer in 3-Phase CT Images Using a Deep Learning Method. J Digit Imaging 32(4):638–643, 2019.
21. Zahari R, Cox J, Obara B: Uncertainty-aware image classification on 3D CT lung. Comput. Biol. Med 172:108324, 2024.
22. Jocher G, Chaurasia A, Stoken A, et al.: ultralytics/yolov5: v3.0. Zenodo, 2020. https://doi.org/10.5281/zenodo.7347926.

23. KiTS19 Challenge. Available at https://kits19.grand-challenge.org/, Accessed 14 May 2024.
24. Wang X, Deng X, Fu Q, et al.: A Weakly-Supervised Framework for COVID-19 Classification and Lesion Localization From Chest CT. IEEE Trans Med Imaging 39(8):2615–2625, 2020.
25. Sensoy M, Kaplan L, Kandemir M: Evidential Deep Learning to Quantify Classification Uncertainty. arXiv preprint https://arxiv.org/abs/1806.01768, 2018.
26. Seoni S, Jahmunah V, Salvi M, et al.: Application of uncertainty quantification to artificial intelligence in healthcare: A review of last decade (2013–2023). Comput Biol Med 165:107441, 2023.
27. Yu Q, Ning Y, Wang A, et al.: Deep learning–assisted diagnosis of benign and malignant parotid tumors based on contrast-enhanced CT: a multicenter study. Eur Radiol 33(9):6054–6065, 2023.
28. Panthakkan A, Anzar SM, Jamal S, et al.: Concatenated Xception-ResNet50 - A novel hybrid approach for accurate skin cancer prediction. Comput Biol Med 150:106170, 2022.
29. Yang X, Xi X, Yang L, et al.: Multi-modality relation attention network for breast tumor classification. Comput Biol Med 150:106210, 2022.
30. Song B, Sunny S, Li S, et al.: Bayesian deep learning for reliable oral cancer image classification. Biomed Opt Express 12(10):6422–6430, 2021.
31. Ahsan MA, Qayyum A, Razi A, et al.: An active learning method for diabetic retinopathy classification with uncertainty quantification. Med Biol Eng Comput 60(10):2797–2811, 2022.
32. Thiagarajan P, Khairnar P, Ghosh S: Explanation and Use of Uncertainty Quantified by Bayesian Neural Network Classifiers for Breast Histopathology Images. IEEE Trans Med Imaging 41(4):815–825, 2022.
33. YOUDEN WJ: Index for rating diagnostic tests. Cancer 3(1):32–35, 1950.
34. Rączkowska A, Możejko M, Zambonelli J, et al.: ARA: accurate, reliable and active histopathological image classification framework with Bayesian deep learning. Sci Rep 9(1):14347, 2019.
35. Shehata M, Alksas A, Abouelkheir RT, et al.: A Comprehensive Computer-Assisted Diagnosis System for Early Assessment of Renal Cancer Tumors. Sensors (Basel) 21(14):4928, 2021.
36. Kocak B, Yardimci AH, Bektas CT, et al.: Textural differences between renal cell carcinoma subtypes: Machine learning-based quantitative computed tomography texture analysis with independent external validation. Eur J Radiol 107:149–157, 2018.
37. Uhlig J, Leha A, Delonge LM, et al.: Radiomic Features and Machine Learning for the Discrimination of Renal Tumor Histological Subtypes: A Pragmatic Study Using Clinical-Routine Computed Tomography. Cancers (Basel) 12(10):3010, 2020.