BMC Cancer

SYSTEMATIC REVIEW

Open Access

Diagnostic performance of artificial intelligence in detection of renal cell carcinoma: a systematic review and meta-analysis

Mahdi Gouravani[1†], Mohammad Shahrabi Farahani[2†], Mohammad Amin Salehi[3*†], Shayan Shojaei[3], Sina Mirakhori[4], Hamid Harandi[3], Soheil Mohammadi[5] and Ramy R. Saleh[6,7]

Abstract

Objectives The detection of renal cell carcinoma (RCC) tumors in the earlier stages is of great importance for more effective treatment. Encouraged by the key role of imaging in the management of RCC, we conducted a systematic review and meta-analysis of the studies that made use of artificial intelligence (AI) for the detection of RCC to quantitatively determine the performance of AI for distinguishing related renal lesions.

Materials and methods PubMed, Scopus, CENTRAL, and Embase electronic databases were systematically searched in November 2024 to identify studies that applied AI for the detection or classification of RCC. We conducted a meta-analysis to evaluate the diagnostic performance of utilized algorithms. Moreover, meta-regression was conducted over suspected covariates to evaluate potential sources of inter-study heterogeneity. Publication bias and quality assessment were also done for the included studies.

Results Sixty-four studies were included in this systematic review, of which 31 studies were selected for meta-analysis. The studies assessing algorithms' performance on internal validation showed pooled sensitivity and specificity of 85% (95% confidence interval [CI], 82 to 87) and 76% (95% CI, 70 to 80), respectively. Moreover, externally validated AI algorithms had a pooled sensitivity and specificity of 80% (95% CI, 73 to 84) and 90% (95% CI, 84 to 93), respectively. Studies that performed internal validation for clinician performance had a pooled sensitivity of 79% (95% CI, 72 to 85) and specificity of 60% (95% CI, 49 to 70).

Conclusion The findings of the present study validate the acceptable performance of AI algorithms when contrasted with medical professionals in the identification and categorization of RCC. Nevertheless, the presence of heterogeneity between studies and the absence of coherence in the results underscore the necessity for the cautious interpretation of these results and additional prospective studies.

Keywords Renal cell carcinoma, Artificial intelligence, Deep learning, Meta-analysis, Imaging

†Mahdi Gouravani, Mohammad Shahrabi Farahani and Mohammad Amin Salehi contributed equally equally to this work.

*Correspondence:
Mohammad Amin Salehi
mohamsa@gmail.com
Full list of author information is available at the end of the article

Introduction

As the most prevalent primary renal neoplasm, renal cell carcinoma (RCC) has shown improving case fatality and survival rate in the past decades where the five-year survival rate has increased from 34 to 75% over the past 60 years [1–3]. This is mostly attributed to the earlier detection of tumors especially those with a diameter of less than 4 cm in addition to the introduction of curative surgical treatments [4].

Although previously marked as benign adenoma, recent evidence strongly suggests that even lesions smaller than 3 cm can indicate carcinoma [5, 6]. Whereas localized (stage I to III) clear cell RCC and non-clear cell RCC have a considerable chance of cure following ablation or surgical resection, until recently, patients with advanced or metastatic RCC only received palliative care even though newer treatment options like stereotactic ablative radiotherapy (SABR) have emerged [7–9]. The detection of RCC tumors in the earlier stages is of great importance for clinicians to have wider and more effective treatment options to offer and an abdominal triple-phase computed tomography (CT) scan remains the primary and most accurate modality for staging RCC [10, 11].

Recently, artificial intelligence (AI) has gained abundant attention in all aspects of human life including medical science and practice [12, 13]. Although there is still a serious debate over the application of AI in medical practice, several AI-powered software have been approved by The US Food and Drug Administration (FDA) for administration [14]. As the essential components of AI, machine learning (ML) and its subset, deep learning (DL), deal with data analysis, pattern recognition, and decision-making just as a medical practitioner does regularly [15].

Radiomics have been applied in RCC, especially to differentiate benign renal lesions from malignant ones which is important to prevent overtreatment [16]. Although radiomics examines extensive data collections, it remains constrained by the requirement to manually specify quantitative metrics. The incorporation of ML algorithms and DL facilitates additional automation, as images are scrutinized utilizing AI pattern recognition [17]. Encouraged by the key role of imaging in the management of RCC, we conducted a systematic review and meta-analysis of the studies that made use of AI for the detection of RCC to quantitatively determine the performance of AI for distinguishing related renal lesions.

Materials and methods

Protocol and registration

The Preferred Reporting Items for Systematic Reviews and Meta-analysis (PRISMA) statement guidelines were followed to conduct this systematic review and meta-analysis [18]. This review was submitted to the International Prospective Register of Systematic Reviews (PROSPERO) website (CRD42023410777).

Search strategy and study selection

PubMed, Scopus, CENTRAL, and Embase electronic databases were systematically searched in October 2023 and updated in November 2024 to identify studies that applied AI for the detection or classification of RCC amid all renal lesions published since the conception of the databases. For this purpose, we used the combination of key terms “renal cell carcinoma” and “artificial intelligence” and their synonyms and related Medical Subject Headings (MeSH) terms (Table E1). Two separate reviewers (S.S and M.A.S) were responsible for the primary screening of the identified records based on their titles and abstracts followed by a full-text screening. All of the original human research studies that developed or validated an AI algorithm to detect RCC amid other renal lesions or compared to healthy controls meeting following PICOS criteria were included without any publication time, or study setting limitations. P: Patients with RCC; I: Image analysis based on AI algorithm; C: Benign renal lesions or healthy controls; O: Performance metrics such as sensitivity, specificity, accuracy, etc.; S; Single studies, comparative studies, and original studies on humans. The Exclusion criteria were as follows: (1) non-original articles; (2) non-English articles. Any disagreements in this phase were resolved by discussion.

Data extraction

The following demographic, methodologic, and analytic data were extracted from the included studies by two independent reviewers (M.S, S.S, and S.M): name of first author; year of publication; study location; the number of cases for each trait; modality of imaging; algorithm and architecture applied; imaging dimension and view; evaluation metrics; and training size; type and size of validation; the number of TN, FN, TP, and FP; sensitivity and specificity; positive predictive value (PPV); negative predictive value (NPV); AUC; accuracy. We made a contingency table where possible using the required data to extract calculated sensitivity and specificity. Discrepancies in data extraction were resolved through consensus.

Statistical analysis

Stata version 17 software (Stata Corp, College Station, TX) was employed for all quantitative data analyses to evaluate the diagnostic performance of utilized algorithms [19]. We conducted a meta-analysis if at least two eligible studies were identified to determine the diagnostic performance of AI algorithms along with radiologists.

A random-effects model was applied to measure the pooled sensitivity and specificity with their corresponding 95% confidence intervals (CI) via a bivariate approach, in which, logit-transformed sensitivity and specificity are measured separately and put together across the studies. Summary receiver operating characteristic (ROC) curves were also applied to figuratively demonstrate the pooled sensitivities and specificities. ROC curves were computed by merging the effect sizes and their corresponding variances or standard errors from each study, and by means of calculating summary sensitivity and specificity at various cut-off points along the curve, in order to produce a representation of the summary ROC curve.

Moreover, to evaluate potential sources of inter-study heterogeneity, meta-regression with a logistic regression model was conducted over covariates of the data augmentation and the modality used in each study. All statistical tests were two-sided and a P value less than 0.05 was considered statistically significant.

Quality and risk of bias assessment

The Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) checklist consisting of 22 recommendations for transparent reporting in studies about the development or validation of prediction models was used for quality assessment [20]. A modified version of TRIPOD (Table E2) was used for quality assessment since several items in the original checklist were unsuitable for AI studies such as follow-up time which is irrelevant in diagnostic accuracy studies. Furthermore, we applied the Prediction Model Study Risk of Bias Assessment Tool (PROBAST) checklist to assess the presence of bias and evaluate the applicability of the included articles in four domains of participants, predictors, outcomes, and analyses (Table E3) [21]. Both testing and training images were considered to evaluate the participants' domain. Quality assessment was separately done by three reviewers (M.S and S.S and S.M) and disagreement was resolved by a consult with a third reviewer (S.M).

Publication bias

We conducted a thorough examination of the bibliographies of the studies that were included, in an attempt to mitigate the potential influence of publication bias. A regression analysis utilizing diagnostic log odds ratios and asymmetry testing was to evaluate publication bias [22].

Results

Study selection and characteristics

As depicted in Fig. 1, an initial search of electronic databases identified 3167 records (PubMed: 766, Scopus: 1768, Embase: 608, CENTRAL: 25), of which 933 records were duplicates and thus removed. The remaining articles underwent primary screening based on their title and abstract which resulted in the exclusion of 1974 studies. In the next step, the full texts of 260 studies were evaluated for eligibility criteria, of which 195 articles were excluded for the following reasons: irrelevant to the title ($n=81$); non-original ($n=59$); not using imaging as its diagnosis method ($n=25$); unavailable full-text ($n=19$); non-English ($n=5$); not reporting diagnostic performance measures ($n=6$). Eventually, 64 studies were included in the current systematic review [23–86].

Included studies' characteristics including target condition and comparator, imaging modality, radiologic view, and reference standard were summarized in Tables 1 and 2. Nine studies used MRI [25, 40, 43, 57, 67, 71, 74, 76, 78], 2 studies used ultrasound [29, 55], and the rest of the studies used CT as their imaging modality. MRI protocol and radiologic views of the studies can be found in Tables 1 and 2. Moreover, 61 studies [23–71, 73–80, 82, 83, 85, 86] used internal validation for their outcomes and 8 studies [43, 49, 51, 59, 69, 72, 81, 84] used external validation method (Tables E5 and E6). Regarding comparison groups, 26 studies compared the results between multiple algorithms [23, 24, 29, 34–37, 40, 41, 43, 45, 47, 49, 50, 52, 53, 56, 60, 61, 65, 70, 71, 76, 78, 85, 86], 25 studies with expert clinicians [25, 27, 28, 31, 32, 39, 42, 46, 51, 54, 55, 59, 62–64, 67, 68, 73–75, 77, 80, 82–84], 4 studies with both multiple algorithms and expert clinicians [33, 57, 72, 79], and 2 studies between different phases of CT scan [69, 81]. In the study by Coy et al. [44], the performance of AI was compared between different processing modes. Six studies did not specifically mention their comparison group [26, 30, 38, 48, 58, 66]. Only the study by Zhu et al. reported 'expert consensus' as their reference standard [26].

Study participants

The number of participants in each study ranged considerably (median, 161.5; range, 38–1795; interquartile range, 78.75–310.5). Likewise, the percentage of participants with RCC also varied substantially between the included studies considerably (median, 71; range, 14.91–95.3; interquartile range, 58.97–80.85). Five studies did not mention the proportion of RCC cases among the study participants [26, 54, 56, 78, 83].

Algorithm advancement and model output

The number of images used for model training in the included studies differed widely (Tables 1 and 2): training set (median, 234.5; range, 47–48832; interquartile range, 117–423.25), tuning set (median, 116; range, 20–238; interquartile range, 30–217.5), testing set (median, 124.5;

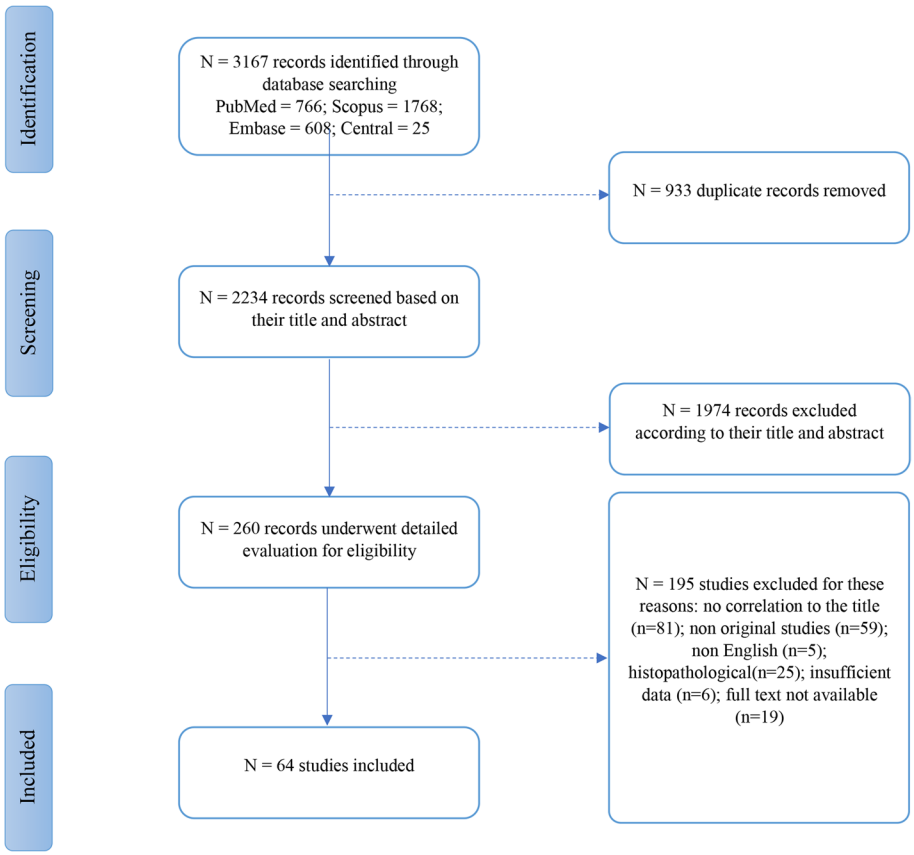

Fig. 1 Flowchart of the study selection process

range, 18–20000; interquartile range, 54.5–288). Twenty-one studies did not mention the number of these data sets separately. Notably, only 8 studies reported the number of images per tuning set. In addition, Table E4 shows data augmentation use, transfer learning utilization, and the validation method in each study. Results of the developed models were reported in the form of specificity by 36 studies, sensitivity by 39 studies, accuracy by 35 studies, AUC by 34 studies, NPV by 12 studies, PPV by 19 studies, and F1 by 7 studies (Tables E5 and E6).

Quality assessment

Included studies had various levels of adherence to TRIPOD guidelines (Fig. 2). Only six items of this guideline were addressed by less than half of the included studies: sample size estimation by 5% of the studies, handling missing data by 6%, model-updating description by 42%, difference between development and validation data by 13%, model specification by 13%, and reporting model-updating results by 36%.

Evaluation of the studies with the PROBAST checklist was summarized in Fig. 2. Results indicated that 46 studies (72%) had a high risk of bias and 6 studies (9%) showed a high applicability concern. While participants' selection bias was high in 25 studies (39%), applicability concern was high in none of the studies. The risk of bias in the outcome domain and respected applicability concern were low in nearly all of the studies (95% and 91%, respectively). Finally, in the analysis domain, the risk of bias was high in 24 studies (38%). The PROBAST evaluation was primarily influenced by two distinct categories of research, namely studies lacking external validation and studies featuring internal validation but employing small sample sizes.

Meta-analysis

Thirty-one studies were selected for meta-analysis as they provided desirable data to obtain contingency tables used for RCC detection and classification [25, 27, 30–36, 39, 42, 43, 45, 51, 56, 57, 59–63, 68, 71–75, 78, 82, 84, 85]. As brought in Table 3, 109 contingency tables were made from 27 studies assessing algorithm performance on internal validation [25, 27, 30–36, 42, 45, 51, 56, 57, 59–63, 68, 71, 73–75, 78, 82, 85], the analysis of which showed pooled sensitivity and specificity of 85% (95% CI, 82 to 87) and 76% (95% CI, 70 to

Table 1 Study characteristics, internal validation

Author	Year	Country	Modality	MRI protocol	View	Target condition	Comparator	Comparison group	No. of Images per training Set	No. of Images per validation Set	No. of Images per testing Set	Model output	Reference standard
Philip S. Maclin	1991	USA	Ultrasound	NA	NA	RCC	Benign	Expert clinicians	52	-	47	Binary classification	Pathology report
Han Sang Lee	2017	Republic of Korea	CT scan	NA	axial	ccRCC	AML	Multiple Algorithms	-	-	-	Binary classification	Pathology report
Roberto Magherini	2017	Italy	CT scan	NA	NA	ccRCC	oncocytoma	Expert clinicians	-	-	-	Binary classification	Pathology report
Zhichao Feng	2017	China	CT scan	NA	axial	RCC	AMLwvf	Multiple Algorithms	-	-	-	Binary classification	expert consensus
HeiShun Yu	2017	USA	CT scan	NA	axial	ccRCC, pRCC, chRCC	oncocytoma	Expert clinicians	-	-	-	Binary classification	Pathology report
Hansang Lee	2018	Republic of Korea	CT scan	NA	axial	ccRCC	AMLwvf	Multiple Algorithms	-	-	-	Binary classification	Pathology report
Yajuan Li	2019	China	CT scan	NA	NA	chRCC	oncocytoma	Multiple Algorithms	-	-	-	Binary classification	Pathology report
Takashi Tanaka	2019	Japan	CT scan	NA	NA	ccRCC	AML	Expert clinicians	48,832	-	281	Binary classification	Pathology report
Ruimeng Yang	2019	China	CT scan	NA	axial	ccRCC, chRCC, pRCC	AML	Multiple Algorithms	163	-	-	Binary classification	Pathology report
Markus M. Obmann	2019	Germany—Switzerland	CT scan	NA	axial	RCC	Healthy control	Multiple Algorithms	106	-	46	Binary classification	Pathology report
En-Ming Cui	2019	China	CT scan	NA	axial	RCC	AMLwvf	Expert clinicians	-	-	-	Binary classification	Pathology report
Heidi Coy	2019	USA	CT scan	NA	axial, coronal, sagittal	ccRCC	oncocytoma	Multiple views	-	-	-	Binary classification	Pathology report
Xue-Ying Sun	2019	China	CT scan	NA	NA	ccRCC, pRCC, chRCC	AMLwvf, Oncocytoma	Expert clinicians	-	-	Image set 1 = 100 Image set 2 = 226	Classification	Pathology report
Yang Yi	2020	China	CT scan	NA	NA	ccRCC	AML	Multiple Algorithms	-	-	-	Binary classification	Pathology report
Mikkel Pedersen	2020	Denmark	CT scan	NA	axial	RCC	oncocytoma	Multiple Algorithms	14,001	238	948	Binary classification	Pathology report
Amir Baghdadi	2020	USA	CT scan	NA	axial	chRCC	Oncocytoma	Multiple Algorithms	172	20	-	Binary classification	Pathology report

Table 1 (continued)

Author	Year	Country	Modality	MRI protocol	View	Target condition	Comparator	Comparison group	No. of Images per training Set	No. of Images per validation Set	No. of Images per testing Set	Model output	Reference standard
Fatemeh Zabihollahy	2020	Canada	CT scan	NA	axial	ccRCC, pRCC, chRCC	AMLwvf, Oncocytoma	Multiple Algorithms	155	-	160	Binary classification	Pathology report
Daniela Said	2020	USA	MRI	Using 1.5 T ($n=95$; Aera, Espree, Symphony,Amira, Siemens Healthineers; Signa HD, HDxt and Optima, GE medical systems) or 3 T ($n=30$; Skyra and Verio, Siemens Healthineers; Signa HDxt, Discovery 750w, GE Medical Systems)	axial, coronal	ccRCC, pRCC, chRCC	fat poor AML, oncocytoma, solitary fibrous tumor, mixed epithelial and stromal tumor	Multiple Algorithms	88	37	-	Binary classification	Pathology report
Assad Oberai	2020	USA	CT scan	NA	axial	ccRCC, chRCC, pRCC	oncocytoma, AML	NA	120	23	-	Binary classification	Pathology report
Nicola Schieda	2020	Canada	CT scan	NA	axial	Malignant (ccRCC, pRCC, chRCC)	benign (oncocytoma, fat-poor angiomyolipoma)	Expert clinicians	-	-	-	Binary classification	Pathology report
Johannes Uhlig 1	2020	Germany	CT scan	NA	axial	ccRCC, chRCC, pRCC	oncocytoma, AML	Expert clinicians, multiple algorithms	-	-	-	Binary classification	Pathology report
Johannes Uhlig 2	2020	Germany	CT scan	NA	axial	ccRCC, pRCC	oncocytoma, AML	multiple algorithms	-	-	-	Binary classification	Pathology report
Hersh Sagreiya	2020	USA	Ultrasound point shear wave elastography	NA	NA	RCC	AML	Multiple algorithms	-	-	NA	Binary classification	Pathology report
Felix Y. Yap	2020	USA	CT scan	NA	axial, coronal, sagittal	Benign	Malignant	Expert clinicians	-	-	453	Binary classification	Pathology report

Table 1 (continued)

Author	Year	Country	Modality	MRI protocol	View	Target condition	Comparator	Comparison group	No. of Images per training Set	No. of Images per validation Set	No. of Images per testing Set	Model output	Reference standard
Camila Lopes Vendrami	2020	USA	Multiparametric MRI (DWI and DCE)	Using 1.5 T (Magnetom Espree, Avanto, or Aera, Siemens Healthcare, Malvern, PA, USA) and 3 T MR (Skyra and Verio) scanners	axial, coronal	RCC	AML, oncocytoma	Multiple algorithms	264	-	66	Binary classification	Pathology report
Cagri Erdim	2020	Turkey	CT scan	NA	NA	ccRCC, chRCC, pRCC	AML, oncocytoma	Multiple algorithms	84	-	84	Binary classification	Pathology report
Ianto Lin Xi	2020	USA	MRI		coronal, sagittal, axial	ccRCC, chRCC, pRCC, Clear cell papillary RCC, Multilocular cystic RCC, Unclassified RCC	Oncocytoma, AML, Mixed epithelial and stromal tumor, Metanephric adenoma, Renal adenoma	Multiple algorithm, Expert clinicians	816	234	112		Pathology report
Kwang-Hyun Uhm	2021	Republic of Korea	CT scan	NA	NA	ccRCC, chRCC, pRCC	oncocytoma, AML	Expert clinicians	258	-	50	Classification	Pathology report
Aleksandra Maria Osowska-Kurczab	2021	Poland	CT scan	NA	NA	ccRCC, chRCC, pRCC, Multilocular Cystic RCC	Angiomyolipoma, oncocytoma, Urothelial Carcinoma, Renal Cyst	Multiple algorithms	-	-	4213	eightfold classification	Pathology report
Mohamed Shehata	2021	USA	Contrast-enhanced CT scan	NA	NA	RCC	AML	Expert clinicians	-	-	NA	Binary classification	Pathology report
Moozhan Nikpanah	2021	USA	MRI	Using 1.5 T (Aera, Siemens Healthcare)	NA	ccRCC	oncocytoma	Expert clinicians	192	-	48, 80	Expert clinicians	Pathology report
Andrew L. Wentland	2022	USA	CT scan	NA	NA	ccRCC, chrRCC, pRCC	oncocytoma, AML	Expert clinicians	103	-	45	Binary classification	Pathology report

Table 1 (continued)

Author	Year	Country	Modality	MRI protocol	View	Target condition	Comparator	Comparison group	No. of Images per training Set	No. of Images per validation Set	No. of Images per testing Set	Model output	Reference standard
Ruben Ngnitewe Massa'a	2022	USA	MRI	Using 1.5 Tesla (70.6% of subjects on GE Signa Artist, HDxt, or Optima; GE Healthcare, Milwaukee, WI, USA) or 3 Tesla (29.4% of subjects on GE Signa Artist, Architect, or Discovery 750; GE Healthcare, Milwaukee, WI, USA) MR scanners	NA	ccRCC, chrRCC, pRCC	oncocytoma, AML	Multiple algorithms	127.4	-	54.6	Binary classification	Imaging (for angiomyolipoma)/ biopsy or surgical resection (for oncocytomas or renal cell carcinomas)
Jingya Liu	2022	USA	CT scan	NA	NA	ccRCC, chRCC, pRCC	Oncocytoma	NA	210	-	210	Classification	Pathology report
Li Zhang	2022	China	CT scan	NA	NA	SRCC	sfp-AML	NA	399	-	171	Binary classification	Pathology report
Xi-Liang Zhu	2022	China	CT scan	NA	NA	ccRCC, pRCC, multilocular cystic renal cell carcinoma	AML	NA	210	-	90	Binary classification	Expert consensus
Tao Dai	2022	China	CT scan	NA	NA	RCC	AML	Multiple algorithms	3653	-	1707	Binary classification	NA
Abeer J. Alhussaini	2022	UK	CT scan	NA	axial, coronal, sagittal	chRCC	Oncocytoma	Multiple algorithms	-	-	-	Binary classification	Pathology report
Nima Nassiri	2022	USA	CT scan	NA	NA	ccRCC, pRCC, chRCC, sarcomatoid RCC, unclassified RCC	AML, Oncocytoma	Multiple Algorithm, Expert clinicians	-	-	-	Binary classification	Pathology report
Naoto Sassa	2022	Japan	CT scan	NA	NA			Expert clinicians	99	-	-	NA	Pathology report
Gianluca Carlini	2023	Italy	CT scan	NA	axial	ccRCC	Oncocytoma	NA	80% of the data	-	20% of the data	Probability	Pathology report

Table 1 (continued)

Author	Year	Country	Modality	MRI protocol	View	Target condition	Comparator	Comparison group	No. of Images per training Set	No. of Images per validation Set	No. of Images per testing Set	Model output	Reference standard
Nityanand Miskin	2023	USA	CT scan	NA	Axial	RCC	benign cysts, mixed epithelial and stromal tumors, cystic oncocytomas, cystic neoplasms of low malignant potential, papillary adenoma, cystic nephroma, cystic AML	Multiple algorithms	1296	-	144	Binary classification	Pathology report0
Sakib Mahmud	2023	Qatar	CT scan	NA	NA	ccRCC, chRCC, pRCC	Oncocytoma	Multiple algorithms	240		60	Classification	Pathology report
Cassandre Garnier	2023	France	CT scan	NA	axial, coronal, sagittal	ccRCC, chRCC, pRCC, other malignant tumor	AML, oncocytoma, and other benign tumor	Multiple algorithms	-	-	-	Binary classification	Pathology report
Haohua Yao	2023	China	CT scan	NA	Axial	ccRCC, chrRCC, pRCC	AMLwvf	Expert clinicians	267	-	53	Binary classification	Pathology report
Shengxing Feng	2023	China	CT scan	NA	axial, coronal, sagittal	Malignant SRM (small renal masses)	Benign	Expert clinicians	108	-	48	Binary classification	Pathology report
Jianyi Qu	2023	China	CT scan	NA	Axial	ccRCC, chRCC, pRCC	Oncocytoma	Expert clinicians	47	-	18	Binary classification	Pathology report
Rohini Gaikar	2023	Canada	MRI	Using Siemens, GE Healthcare, and Phillips MRI systems with magnetic field strengths 1.5 T and 3 T	axial	Malignant (RCC, ccRCC, pRCC, and chRCC)	benign (fpAML, oncocytomas)	Multiple algorithms	80% of the data	-	20% of the data	Binary classification	Pathology report
Mateusz Glembin	2023	Poland	CT scan	NA	axial	ccRCC, chRCC, pRCC, other malignant	AML, Oncocytoma, and Other benign	Expert clinicians	7207	-	-	Binary classification	Pathology report

Table 1 (continued)

Author	Year	Country	Modality	MRI protocol	View	Target condition	Comparator	Comparison group	No. of Images per training Set	No. of Images per validation Set	No. of Images per testing Set	Model output	Reference standard
María Aymerich	2023	Spain	CT scan	NA	axial	chRCC	Oncocytoma	Expert clinicians	80% of the data	-	20% of the data	Binary classification	Pathology report
Huancheng Yang	2023	China	Contrast-enhanced CT scan	NA	axial, coronal, sagittal	ccRCC, chRCC, pRCC	angiomyolipoma, oncocytoma, other types	CT phases, Internal VS External	680	201	201	Binary classification	Pathology report
Thibault Toffoli	2023	France	MRI	Using an Achieva 1.5 T scanner (Philips Medical Systems, Amsterdam, The Netherlands) or 1.5 T Avanto scanner (SIEMENS, Erlingen, Germany), and a Discovery 3 T MR750W scanner (GE Medical Systems, Chicago, IL, USA)	axial	ccRCC	Oncocytoma	Multiple Algorithms	-	-	-	Binary classification	Pathology report
Vlad-Octavian Bolocan	2023	Romania	CT scan	NA	NA	RCC	Benign	Expert clinicians	70% of the data	20% of the data	10% of the data	Binary classification	Pathology report
Mary R. Myers	2024	India	CT scan	NA	axial	ccRCC, pRCC,chRCC	oncocytoma	NA	50	-	20,000	Classification	Pathology report
Roberto Francischello	2024	Italy	CT scan	NA	NA	ccRCC	oncocytoma	Expert clinicians	-	-	-	Binary classification	Pathology report
Huanhuan Kang	2024	China	MRI	Using 3.0 Tesla (T) and 1.5 T MRI scanners (mainly Discovery MR750 3.0 T, GE Healthcare	axial	Malignant	Benign	multiple algorithms	278	-	48	Binary classification	Pathology report
Haytham Shebel	2024	Egypt	Contrast-enhanced CT scan	NA	axial, coronal, sagittal	ccRCC	Oncocytoma	multiple algorithms	351	116	-	Binary classification	Pathology report
Michail E.Klontzas	2024	Sweden	CT scan	NA	axial	RCC	Benign	Expert clinicians	394	-	170	Binary classification	Pathology report

Table 1 (continued)

Author	Year	Country	Modality	MRI protocol	View	Target condition	Comparator	Comparison group	No. of Images per training Set	No. of Images per validation Set	No. of Images per testing Set	Model output	Reference standard
Pouria Yazdian Anari	2024	USA	MRI	NA	axial	hybrid, ccRCC, HLRCC, pRCC, ChRCC	oncocytoma, angiomyolipoma	Expert clinicians	-	-	-	Classification	Pathology report
Ruiting Wang	2024	China	MRI	Using 3.0 T MRI scanners (Magnetom Aera; Siemens Healthineers	axial	non- ccRCCs (pRCCs and chRCCs)	benign renal tumors without visible fat component (AMLwvfs and oncocytoma)	Expert clinicians	-	-	-	Binary classification	Pathology report
Shuanbao Yu	2024	China	Contrast-enhanced CT scan	NA	NA	ccRCC, chrRCC, pRCC, MiT Family translocation, uncommon, unclassified RCC	AML, oncocytoma, benign cyst, uncommon benign renal tumor	Expert clinicians	1184	-	611	Binary classification	Pathology report

Abbreviations: *RCC* Renal cell carcinoma, *pRCC* papillary renal cell carcinoma, *ccRCC* chromophobe renal cell carcinoma, *ccRCC* clear cell renal cell carcinoma, *AML* Angiomyolipoma, *AMLwvf* Angiomyolipoma without visible fat, *sRCC* sarcomatoid renal cell carcinoma, *sfp-AML* small fat-poor angiomyolipoma, *MRI* Magnetic resonance imaging, *NR* Not reported, *N/A* Not applicable, *CT* Computed tomography

Table 2 Study characteristics, external validation

Author	year	country	modality	MRI protocol	view	Target condition	Comparator	Comparison group	No. of Images per training Set	No. of Images per tuning Set	No. of Images per testing Set	Model output	Reference standard
Mikkel Pedersen	2020	Denmark	CT scan	NA	Axial	RCC	oncocytoma	Multiple Algorithms	14,001	-	356	Binary classification	Pathology report
Camila Lopes Vendrami	2020	USA	Multiparametric MRI (DWI and DCE)	Using 1.5 T (Magnetom Espree, Avanto, or Aera, Siemens Healthcare, Malvern,PA, USA) and 3 T MR (Skyra and Verio) scanners	axial, coronal	ccRCC, chRCC, pRCC	classic AML, AMLwvf, oncocytoma	Multiple Algorithms	-	-	-	Binary classification	Pathology report
Kwang-Hyun Uhm	2021	Republic of Korea	CT scan	NA	NA	ccRCC, chRCC, pRCC	oncocytoma, AML	Expert clinicians	258	-	40	Classification	Pathology report
Tao Zhou	2023	China	CT scan	NA	NA	ccRCC, chRCC, pRCC	oncocytoma, AML	Expert clinicians	496	-	309	Binary classification	Pathology report
Haohua Yao	2023	China	CT scan	NA	Axial	ccRCC, chRCC, pRCC	AMLwvf	Expert clinicians	267	-	132	Binary classification	Pathology report
Huancheng Yang	2023	China	Contrast-enhanced CT scan	NA	axial, coronal, sagittal	ccRCC, chRCC, pRCC	angiomyolipoma, oncocytoma, other types	CT phases, Internal vs External	680	201	201	Binary classification	Pathology report
Annemarie Uhlig	2024	Germany	Contrast-enhanced CT scan	NA	axial	ccRCC, chRCC, pRCC	angiomyolipoma, oncocytoma, other types	CT phases	297	-	121	Binary classification	Pathology report
Jiayue Han	2024	China	CT scan	NA	NA	ccRCC	indolent renal tumours	Multiple algorithms, Expert clinicians	229	-	81	Binary classification	Pathology report

Abbreviations: *RCC* Renal cell carcinoma, *pRCC* papillary renal cell carcinoma, *chRCC* chromophobe renal cell carcinoma, *ccRCC* clear cell renal cell carcinoma, *AML* Angiomyolipoma, *AMLwvf* Angiomyolipoma without visible fat, *sRCC* sarcomatoid renal cell carcinoma, *sfp-AML* small fat-poor angiomyolipoma, *MRI* Magnetic resonance imaging, *NR* Not reported, *N/A* Not applicable, *CT* Computed tomography

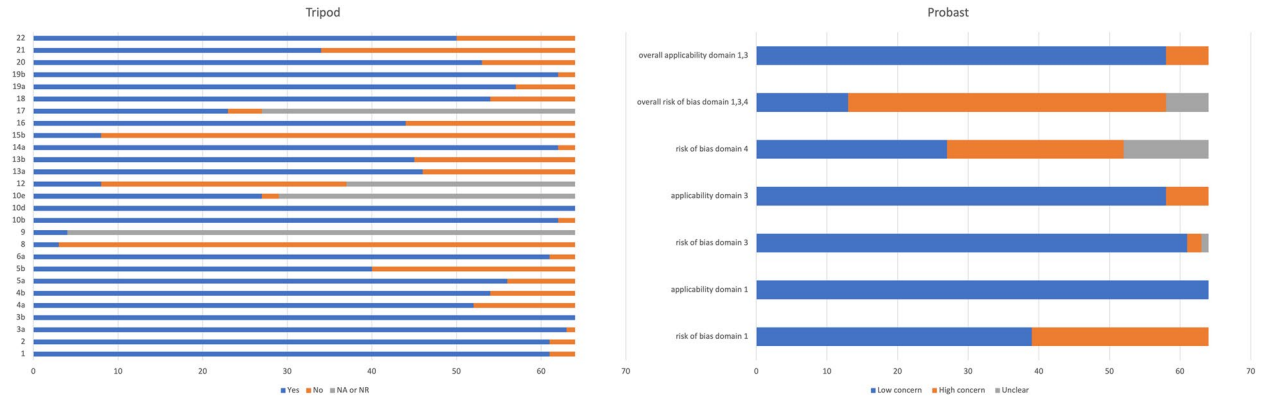

Fig. 2 Adherence of the included studies to TRIPOD and PROPAST guidelines

80), respectively (Figure E1). Three observations from two different studies had an accuracy of less than 60% and the studies themselves had not recommended these algorithms [31, 57]; thus, as strong confounders, they were not included in this analysis.

Moreover, 5 studies [43, 51, 59, 72, 84] with 20 contingency tables that externally validated AI algorithms had a pooled sensitivity and specificity of 80% (95% CI, 73 to 84) and 90% (95% CI, 84 to 94), respectively (Figure E2).

Analysis of 7 studies [25, 32, 33, 39, 51, 57, 68] with 22 contingency tables that performed internal validation for clinician performance resulted in the pooled sensitivity of 79% (95% CI, 72 to 85) and specificity of 60% (95% CI, 49 to 70; Figure E3). Since one study reported externally validated clinician performance with contingency tables [51], no analysis was done for this group. However, 3 studies with 8 contingency tables assessed clinician performance with AI assistance [39, 72, 84] that had a pooled sensitivity of 86% (95% CI, 79 to 91) and specificity of 72% (95% CI, 66 to 77; Figure E4). Other analyzed metrics comprising AUC, diagnostic odds ratio, and positive and negative likelihood ratio for aforementioned groups were collectively brought in Table 3. Moreover, hierarchical summary receiver operating characteristic (HSROC) curves for the internal and external validation as well as clinicians with AI assistance test sets were depicted in Figs. 3, 4, and 5, respectively.

Furthermore, in the subgroup analysis based on comparator group, pooled results of 13 contingency tables from 5 studies [25, 56, 61, 62, 73] showed that internally-validated AI algorithms differentiated RCC from oncocytoma with the pooled sensitivity of 75% (95% CI, 56 to 88) and specificity of 73% (95% CI, 64 to 81). Pooled results from 5 studies [27, 32, 42, 45, 85] with 24 contingency tables demonstrated that AI algorithms with internal validation differentiated RCC from angiomyolipoma with the pooled sensitivity of 80% (95% CI, 75 to 84) and specificity of 81% (95% CI, 73 to 86).

We conducted meta-regression on three variables including the presence of DL, application of transfer learning, and usage of data augmentation which were reported by more than three studies (Table E7). Results demonstrated that studies implementing DL in their algorithms had higher sensitivity than those without DL (87% [95%CI, 83 to 90] vs. 84% [95%CI, 81 to 87]). Likewise, specificity was higher in the studies that used DL than those without DL in their algorithms (86% [95%CI, 77 to 92] vs. 70% [95%CI, 65 to 75]). Moreover, studies that applied transfer learning had a higher sensitivity than those without transfer learning (88% [95%CI, 82 to 92] vs. 84% [95%CI, 81 to 87]), as well as higher specificity (96% [95%CI, 89 to 99] vs. 71% [95%CI, 66 to 76]). Eventually, studies with data augmentation had lower sensitivity than the group of studies that did not use this method (83% [95%CI, 79 to 87] vs. 86% [95%CI, 82 to 89]); However, the specificity was higher in the former group (81% [95%CI, 72 to 88] vs. 72% [95%CI, 66 to 77]).

Publication bias

Analysis for publication bias was conducted for studies assessing AI algorithms with internal validation and external validation as well as studies evaluating clinician performance on internal validation tests and studies that measured clinician performance with AI assistance (Table E8). Statistically significant publication bias was only detected in the analysis of studies assessing developed algorithms with internal validation datasets which corresponded to a slope coefficient of -42.07 (95% CI, -47.82 to -36.32; P value < 0.001).

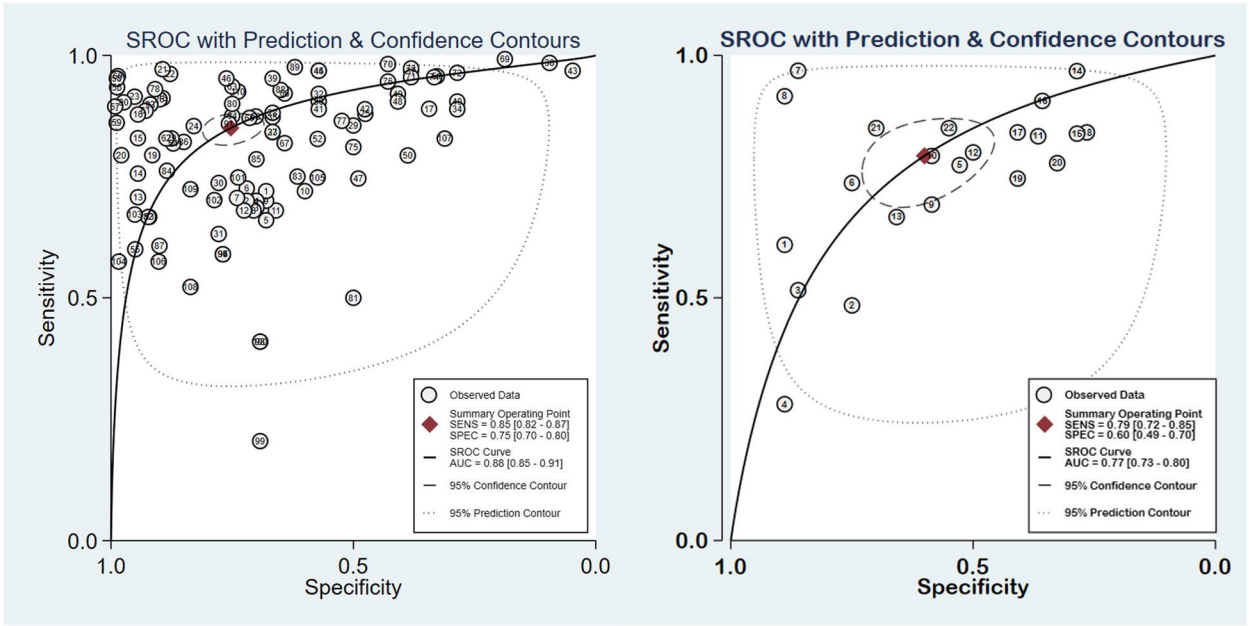

Fig. 3 HSROC curves for artificial intelligence algorithms (**A**) and clinicians (**B**) at the internal validation test. Legend: The Hierarchical summary receiver operating characteristic (HSROC) curve serves as a comprehensive representation of the overall diagnostic accuracy across various studies, plotting sensitivity against specificity. The confidence contour visually presents the confidence intervals for both sensitivity and specificity

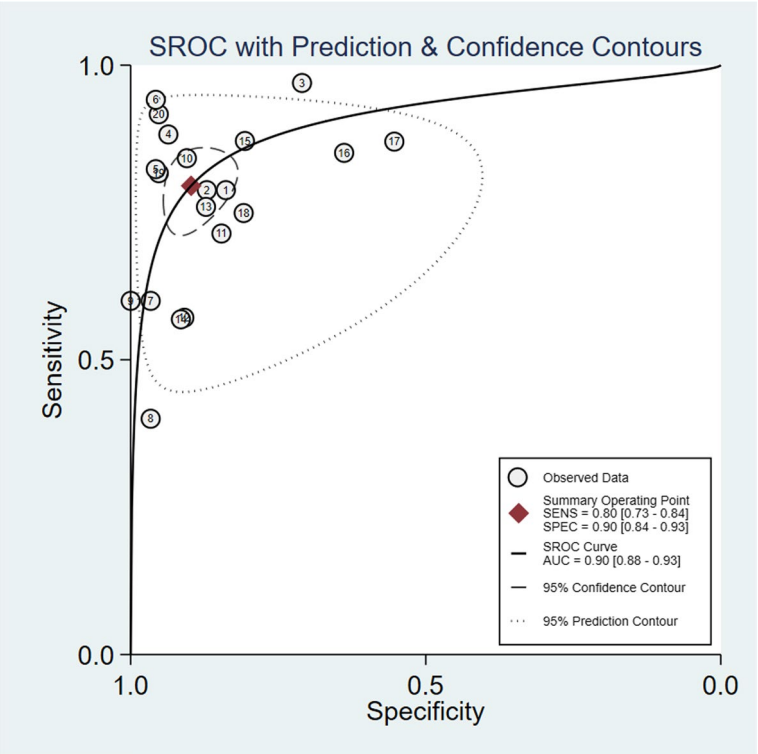

Fig. 4 HSROC curves for artificial intelligence algorithms at the external validation test sets. Legend: The Hierarchical summary receiver operating characteristic (HSROC) curve serves as a comprehensive representation of the overall diagnostic accuracy across various studies, plotting sensitivity against specificity. The confidence contour visually presents the confidence intervals for both sensitivity and specificity

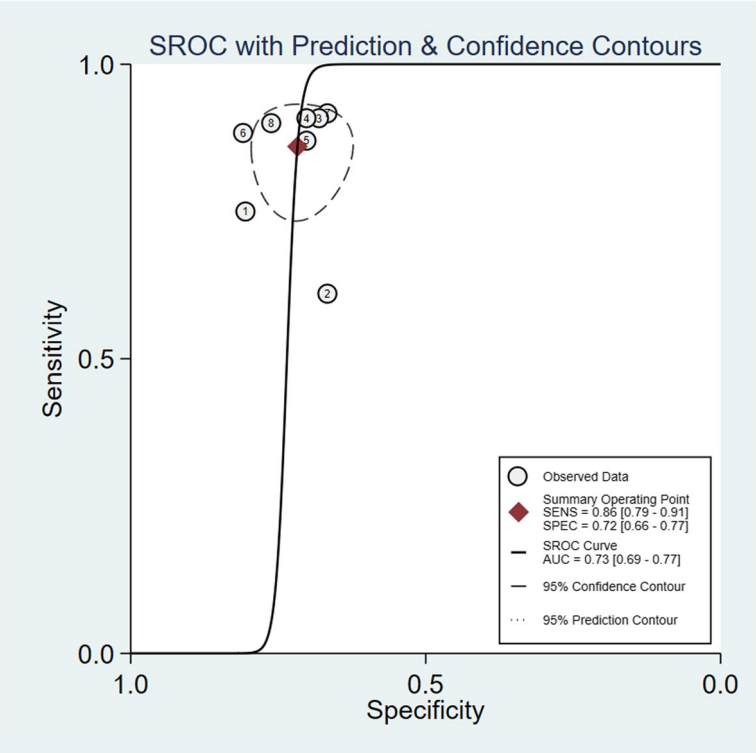

Fig. 5 HSROC curves for clinicians performance with artificial intelligence assistance. Legend: The Hierarchical summary receiver operating characteristic (HSROC) curve serves as a comprehensive representation of the overall diagnostic accuracy across various studies, plotting sensitivity against specificity. The confidence contour visually presents the confidence intervals for both sensitivity and specificity

Table 3 Pooled sensitivities, specificities, and area under the curve, positive likelihood ratio, negative likelihood ratio, diagnostic odds ratio

Parameter	Sensitivity	Specificity	AUC	Positive Likelihood Ratio	Negative Likelihood Ratio	Diagnostic Odds Ratio	
Algorithms, Internal validation, All modalities	0.85 [0.82, 0.87]	0.76 [0.70, 0.80]	0.88 [0.85—0.91]	3.5 [2.8, 4.2]	0.20 [0.17, 0.24]	17 [13, 24]	109
Algorithms, External validation	0.80 [0.73, 0.84]	0.90 [0.84, 0.93]	0.90 [0.88 - 0.93]	7.8 [5.1, 11.8]	0.23 [0.18, 0.29]	34 [21, 56]	20
Clinicians, Internal validation	0.79 [0.72, 0.85]	0.60 [0.49, 0.70]	0.77 [0.73 - 0.80]	2.0 [1.6, 2.5]	0.34 [0.25, 0.47]	6 [4, 9]	22
Algorithms + clinicians	0.86 [0.79, 0.91]	0.72 [0.66, 0.77]	0.73 [0.69—0.77]	3.0 [2.5, 3.7]	0.19 [0.13, 0.30]	16 [9, 27]	8

Abbreviations: *AUC* Area under the curve

Discussion

Results of the current systematic review and meta-analysis indicated that AI algorithms with internal validation had higher performance than clinicians with internal validation in terms of pooled sensitivity (85% vs. 79%) and specificity (76% vs. 60%). Furthermore, externally validated AI algorithms were also at comparable levels with a pooled sensitivity of 80% in addition to a pooled specificity of 90%. Moreover, pooled sensitivity and specificity for clinician performance with AI assistance were 86% and 72%, respectively. Therefore, among the analyzed groups, clinicians with AI assistance showed the highest sensitivity (86%), and AI algorithms with external validation represented the highest specificity (90%). However, clinicians' performance with external validation was not surveyed in this study since only one study reported this parameter [51, 84]. Externally validated clinician performance in the study by Uhm et al. at its highest level had a sensitivity of 92% which was in the classification of clear cell RCC, and

a specificity of 90% which was in the classification of chromophobe RCC [51].

The utilization of AI and ML has seen a growing trend in the assessment and treatment of RCC. These technologies have been investigated in several diagnostic areas, such as differentiation between benign and malignant renal tumors, categorizing subtypes, determining the grade and stage of RCC, and characterizing small renal masses [87]. As observed in our included studies, AI has been extensively used by researchers for distinguishing RCC from other renal masses such as angiomyolipoma, oncocytoma, and renal cysts. AI algorithms, particularly deep learning models like Res-UNet, have been developed to automatically segment kidney and renal mass regions in 3D CT images. This reduces the need for manual segmentation, which is both time-consuming and costly for clinical radiology [48]. In the urology setting, AI can enhance surgical procedures by providing 3D anatomical models, which facilitate better pre-surgical planning and reduce operation times [88]. AI algorithms can also identify patients at high risk of recurrence, enabling tailored follow-up strategies and personalized treatment plans [89].

It has been well-established that CT scan plays a crucial role as the primary imaging modality in the assessment of renal neoplasms [90]. A systematic review showed that CT had a median sensitivity and specificity of 88% (IQR, 81–94) and 75% (IQR, 51–90), and for MRI, the aforementioned metrics were 87.5% (IQR, 75.25–100) and 89% (IQR, 75–96) [91]. Additionally, neural networks are employed to analyze imaging data and extract quantitative features of lesions, including contour, heterogeneity, and gray zone features [92]. Thus, it has been promised that by incorporating clinical data, utilizing ML techniques, performing texture analysis, and retraining ML algorithms, the accuracy of detection and classification by imaging modalities such as CT and MRI can be significantly improved. The findings of the current meta-analysis revealed that AI algorithms have comparable and even higher performance in comparison to clinicians. Nevertheless, the considerable heterogeneity between the included studies urged us to perform a meta-regression analysis to evaluate the influence of possible confounders comprising the application of DL, transfer learning, and data augmentation. The results indicated that the use of DL led higher sensitivity and specificity. In addition, transfer learning as a technique that involves utilizing knowledge learned from one task to improve performance on another related task and is particularly useful when there is limited labeled data available [93], improved the performance of internally-validated AI in terms of sensitivity and specificity. The data augmentation, in which new training data were generated by manipulating existing data to increase the robustness of the model and improve the generalization and performance of models in DL [94], resulted in a higher pooled specificity, but a lower sensitivity. Subgroup analysis based on comparator group showed that internally-validated AI algorithms performed better in differentiation of RCC from angiomyolipoma than in differentiation of RCC from onvocytoma. These inconsistent results and the evident impact of methodologic diversities between studies underscore the necessity of a cautious interpretation of our results.

In the publication bias analysis, the slope coefficient denotes the correlation between the diagnostic odds ratio (DOR) and the inverse of the square root of the effective sample size (ESS). In the absence of publication bias, the slope coefficient should approximate zero. A positive slope coefficient implies that studies with elevated DORs possess lower ESS, thereby indicating a plausible publication bias favoring studies with positive outcomes. There was a statistically significant publication bias in the analysis of algorithms with internal validation that should be noted.

Validation methods are generally categorized into two types: internal and external. Internal validation methods like cross-validation and bootstraps involve splitting one dataset into parts, using some for training and the rest for validation. External validation, on the other hand, tests the model's generalizability on a different dataset that reflects the target population but comes from another source. External validation is crucial for checking the model's generalizability and transportability, while internal validation provides an estimate of the model's performance using the available data [95]. This denotes the necessity for the development of more algorithms with external validation datasets than we currently have.

Furthermore, the current study is subject to several limitations. Out of 64 studies included in the qualitative synthesis of this systematic review, 33 studies, which is approximately half of the studies, did not provide sufficient data to extract contingency tables and consequently did not enter the meta-analysis. Accuracy was documented by multiple studies as a measure of diagnostic performance, despite the fact that this statistical measure is contingent upon the prevalence of the disease among different study populations, even when the sensitivity and specificity are identical between them [96]. Moreover, several studies failed to disclose essential study details, such as the quantity of images included in both the training and testing datasets. A significant publication bias was also detected in the most extensive analysis conducted in the current study, i.e., the performance of AI algorithms at internal validation. This indicates a potential bias in publishing may exist favoring

studies that report positive outcomes. Likewise, evaluation with the PROBAST checklist demonstrated that 72% of studies had a high risk of bias. Moreover, 6 items of the TRIPOD checklist were complied with by less than half of the studies and most studies had unsatisfying status in the method and result domains. The performance of clinicians may be undervalued, given that the studies rarely employed clinicians with a substantial clinical background. Unexpectedly, externally-validated AI algorithms had a superior pooled specificity than internally-validated algorithms which may be due to lower analyzed contingency tables in external validation subset of studies leading to contradictory results. Lastly, the applied meta-analysis model did not take into consideration the clustered nature of the data as multiple contingency tables were drawn from each study. The conventional practice in meta-analysis studies is to include one effect size from each study, as it is widely recognized and recommended. However, when conducting a meta-analysis of diagnostic performance for AI algorithms, a different situation arises. In most studies, multiple algorithms are tested, and various diagnostic performance measures are reported. In such cases, it is not appropriate to include only a single measure in the analysis, as it may not provide a comprehensive evaluation of the algorithms' performance. Recently, this approach of including multiple measures has been adopted in high-quality meta-analyses of diagnostic performance for AI algorithms [13, 97, 98]. Multiple calculated contingency tables are extracted from each study in this approach to determine the overall performance of AI algorithms developed in the studies that are included. The interstudy differences in the architecture of various algorithms and the applied training and testing sets which lies in the nature of AI studies caused a vast heterogeneity in the accuracy reported by these studies. Moreover, different imaging modalities used in the studies is another source of heterogeneity. We tried to address this issue by performing meta-regression analysis, however, the use of more standardized methodologies with more robust methods such as external validation in the future studies can improve the quality of subsequent meta-analyses.

Conclusion

The results of the current study confirmed the acceptable and comparable performance of AI algorithms compared to clinicians in the detection and classification of RCC, for which early diagnosis is of great importance. However, inter-study heterogeneities and inconsistency in several findings underpin the need for further prospective studies and clinical trials with a larger study population to carefully evaluate the performance of AI in comparison to experienced clinicians in real clinical settings. Thus, the successful translation of AI systems to healthcare depends not only on their performance but also on factors such as interpretability, safety, and trust from clinicians. Moreover, designing platforms optimized for intuitive human-AI interactions and ensuring the understandability of AI image classifiers to human observers are crucial for building trust. AI algorithms for RCC detection in imaging are transforming clinical radiology and urology by enhancing diagnostic accuracy, increasing efficiency, and providing valuable decision support. However, AI is currently seen as a tool to augment human expertise rather than replace it entirely.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12885-025-13547-9.

Supplementary Material 1.
Supplementary Material 2.
Supplementary Material 3.
Supplementary Material 4.
Supplementary Material 5.
Supplementary Material 6.
Supplementary Material 7.
Supplementary Material 8.
Supplementary Material 9.
Supplementary Material 10.
Supplementary Material 11.
Supplementary Material 12.

Acknowledgements
Not applicable.

Authors' contributions
MAS, takes responsibility for the integrity of the work as a whole, from inception to finished article. SM, SS, HH, MG, SM, and MSF did the literature search, developed the protocol, and contributed to screening and data extraction. MG and MSF wrote the original draft. MAS designed the project, performed the review, and contributed to the protocol development and final manuscript. RRS encouraged and supervised the project and reviewed the manuscript.

Funding
Not applicable.

Data availability
All data supporting the findings of this study are available within the paper and its Supplementary Information. If any more data is required, they will be provided upon reasonable request.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Musculoskeletal Imaging Research Center (MIRC), Tehran University of Medical Sciences, Tehran, Iran. [2]Medical Students Research Committee, Shahed University, Tehran, Iran. [3]School of Medicine, Tehran University of Medical Sciences, Dr. Qarib St, Keshavarz Blvd, Tehran 14194, Iran. [4]Faculty of Medicine, Tabriz University of Medical Sciences, Tabriz, Iran. [5]Mallinckrodt Institute of Radiology, Washington University in St. Louis, Saint Louis, USA. [6]Department of Oncology, McGill University, Montreal, QC H3A 0G4, Canada. [7]Division of Medical Oncology, McGill University Health Centre, Montreal, QC H4A 3J1, Canada.

Received: 7 September 2024 Accepted: 17 January 2025
Published online: 27 January 2025

References

1. Jemal A, Siegel R, Xu J, Ward E. Cancer statistics, 2010. CA Cancer J Clin. 2010;60:277–300.
2. Pantuck AJ, Zisman A, Belldegrun AS. The changing natural history of renal cell carcinoma. J Urol. 2001;166:1611–23.
3. Scelo G, Larose TL. Epidemiology and Risk Factors for Kidney Cancer. J Clin Oncol Off J Am Soc Clin Oncol. 2018;36:JCO2018791905.
4. Hollingsworth JM, Miller DC, Daignault S, Hollenbeck BK. Rising incidence of small renal masses: a need to reassess treatment effect. J Natl Cancer Inst. 2006;98:1331–4.
5. Schlomer B, Figenshau RS, Yan Y, Venkatesh R, Bhayani SB. Pathological features of renal neoplasms classified by size and symptomatology. J Urol. 2006;176 4 Pt 1:1317–20; discussion 1320.
6. Pahernik S, Ziegler S, Roos F, Melchior SW, Thüroff JW. Small renal tumors: correlation of clinical and pathological features with tumor size. J Urol. 2007;178:414–7.
7. Krabbe L-M, Bagrodia A, Margulis V, Wood CG. Surgical management of renal cell carcinoma. Semin Intervent Radiol. 2014;31:27–32.
8. Tenold M, Ravi P, Kumar M, Bowman A, Hammers H, Choueiri TK, et al. Current Approaches to the Treatment of Advanced or Metastatic Renal Cell Carcinoma. Am Soc Clin Oncol Educ book Am Soc Clin Oncol Annu Meet. 2020;40:1–10.
9. Yang DX, Kwon YS, Timmerman R, Hannan R. Stereotactic ablative radiotherapy for primary renal cell carcinoma. Clin Transl Radiat Oncol. 2024;44:100705.
10. Nazim SM, Ather MH, Hafeez K, Salam B. Accuracy of multidetector CT scans in staging of renal carcinoma. Int J Surg. 2011;9:86–90.
11. Johnson CD, Dunnick NR, Cohan RH, Illescas FF. Renal adenocarcinoma: CT staging of 100 tumors. AJR Am J Roentgenol. 1987;148:59–63.
12. McCarthy J. From here to human-level AI. Artif Intell. 2007;171:1174–82.
13. Kuo RYL, Harrison C, Curran T-A, Jones B, Freethy A, Cussons D, et al. Artificial Intelligence in Fracture Detection: A Systematic Review and Meta-Analysis. Radiology. 2022;304:50–62.
14. Benjamens S, Dhunnoo P, Meskó B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. npj Digit Med. 2020;3:118.
15. Han S-H, Kim KW, Kim S, Youn YC. Artificial Neural Network: Understanding the Basic Concepts without Mathematics. Dement neurocognitive Disord. 2018;17:83–9.
16. Kowalewski K-FF, Egen L, Fischetti CE, Puliatti S, Juan GR, Taratkin M, et al. Artificial intelligence for renal cancer: From imaging to histology and beyond. Asian J Urol. 2022;9:243–52.
17. Xu Q, Zhu Q, Liu H, Chang L, Duan S, Dou W, et al. Differentiating Benign from Malignant Renal Tumors Using T2- and Diffusion-Weighted Images: A Comparison of Deep Learning and Radiomics Models Versus Assessment from Radiologists. J Magn Reson Imaging. 2022;55:1251–9.
18. Higgins J, Thomas J, Chandler J, Cumpston M, Li T, MJ P, et al. Cochrane Handbook for Systematic Reviews of Interventions version 6.3 (updated February 2022); 2022.
19. Harbord RM, Whiting P. Metandi: Meta-analysis of Diagnostic Accuracy Using Hierarchical Logistic Regression. Stata J. 2009;9:211–29.
20. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. BMC Med. 2015;13:1.
21. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. Ann Intern Med. 2019;170:51–8.
22. Deeks JJ, Macaskill P, Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. J Clin Epidemiol. 2005;58:882–93.
23. Miskin N, Qin L, Silverman SG, Shinagare AB. Differentiating Benign From Malignant Cystic Renal Masses: A Feasibility Study of Computed Tomography Texture-Based Machine Learning Algorithms. J Comput Assist Tomogr. 2023;47:376–81.
24. Dai T, Zhu S, Han F, Ye M, Xiang W, Tan W, et al. Benchmarking Machine Learning Algorithms for Diagnosis of Renal Cell Carcinoma. Iran J Radiol. 2022;19:1.
25. Nikpanah M, Xu Z, Jin D, Farhadi F, Saboury B, Ball MW, et al. A deep-learning based artificial intelligence (AI) approach for differentiation of clear cell renal cell carcinoma from oncocytoma on multi-phasic MRI. Clin Imaging. 2021;77:291–8.
26. Zhu X-L, Shen H-B, Sun H, Duan L-X, Xu Y-Y. Improving segmentation and classification of renal tumors in small sample 3D CT images using transfer learning with convolutional neural networks. Int J Comput Assist Radiol Surg. 2022;17:1303–11.
27. Tanaka T, Huang Y, Marukawa Y, Tsuboi Y, Masaoka Y, Kojima K, et al. Differentiation of small (≤ 4 cm) renal masses on multiphase contrast-enhanced CT by deep learning. Am J Roentgenol. 2020;214:605–12.
28. Yu H, Scalera J, Khalid M, Touret A-S, Bloch N, Li B, et al. Texture analysis as a radiomic marker for differentiating renal tumors. Abdom Radiol (New York). 2017;42:2470–8.
29. Sagreiya H, Akhbardeh A, Li D, Sigrist R, Chung BI, Sonn GA, et al. Point Shear Wave Elastography Using Machine Learning to Differentiate Renal Cell Carcinoma and Angiomyolipoma. Ultrasound Med Biol. 2019;45:1944–54.
30. Zhang L, Sun K, Shi L, Qiu J, Wang X, Wang S. Ultrasound Image-Based Deep Features and Radiomics for the Discrimination of Small Fat-Poor Angiomyolipoma and Small Renal Cell Carcinoma. Ultrasound Med Biol. 2023;49:560–8.
31. Schieda N, Nguyen K, Thornhill RE, McInnes MDF, Wu M, James N. Importance of phase enhancement for machine learning classification of solid renal masses using texture analysis features at multi-phasic CT. Abdom Radiol (New York). 2020;45:2786–96.
32. Cui E-M, Lin F, Li Q, Li R-G, Chen X-M, Liu Z-S, et al. Differentiation of renal angiomyolipoma without visible fat from renal cell carcinoma by machine learning based on whole-tumor computed tomography texture features. Acta Radiol. 2019;60:1543–52.
33. Uhlig J, Biggemann L, Nietert MM, Beißbarth T, Lotz J, Kim HS, et al. Discriminating malignant and benign clinical T1 renal masses on computed tomography: A pragmatic radiomics and machine learning approach. Medicine (Baltimore). 2020;99:e19725–e19725.
34. Osowska-Kurczab AM, Markiewicz T, Dziekiewicz M, Lorent M. Multi-feature ensemble system in the renal tumour classification task. Bull Polish Acad Sci Tech Sci. 2021;69:e136749.
35. Erdim C, Yardimci AH, Bektas CT, Kocak B, Koca SB, Demir H, et al. Prediction of Benign and Malignant Solid Renal Masses: Machine Learning-Based CT Texture Analysis. Acad Radiol. 2020;27:1422–9.
36. Obmann MM, Cosentino A, Cyriac J, Hofmann V, Stieltjes B, Boll DT, et al. Quantitative enhancement thresholds and machine learning algorithms for the evaluation of renal lesions using single-phase split-filter dual-energy CT. Abdom Radiol (New York). 2020;45:1922–8.
37. Yang Y, Qian X-S, Zhou Z-Y, Zhu J-B, Shen J-K, Dai Y-K. Classification of renal tumor histology subtypes using radiomics. Zhejiang Daxue Xuebao (Gongxue Ban)/J Zhejiang Univ Eng Sci. 2019;53:2381–8.
38. Carlini G, Gaudiano C, Golfieri R, Curti N, Biondi R, Bianchi L, et al. Effectiveness of Radiomic ZOT Features in the Automated Discrimination of Oncocytoma from Clear Cell Renal Cancer. J Pers Med. 2023;13:478.
39. Sun X-Y, Feng Q-X, Xu X, Zhang J, Zhu F-P, Yang Y-H, et al. Radiologic-Radiomic Machine Learning Models for Differentiation of Benign and Malignant Solid Renal Masses: Comparison With Expert-Level Radiologists. AJR Am J Roentgenol. 2020;214:W44-54.

40. Massa'a RN, Stoeckl EM, Lubner MG, Smith D, Mao L, Shapiro DD, et al. Differentiation of benign from malignant solid renal lesions with MRI-based radiomics and machine learning. Abdom Radiol. 2022;47:2896–904.
41. Li Y, Huang X, Xia Y, Long L. Value of radiomics in differential diagnosis of chromophobe renal cell carcinoma and renal oncocytoma. Abdom Radiol (New York). 2020;45:3193–201.
42. Shehata M, Alksas A, Abouelkheir RT, Elmahdy A, Shaffie A, Soliman A, et al. A Comprehensive Computer-Assisted Diagnosis System for Early Assessment of Renal Cancer Tumors. Sensors (Basel). 2021;21:4928.
43. Lopes Vendrami C, McCarthy RJ, Villavicencio CP, Miller FH. Predicting common solid renal tumors using machine learning models of classification of radiologist-assessed magnetic resonance characteristics. Abdom Radiol (New York). 2020;45:2797–809.
44. Coy H, Hsieh K, Wu W, Nagarajan MB, Young JR, Douek ML, et al. Deep learning and radiomics: the utility of Google TensorFlow™ Inception in classifying clear cell renal cell carcinoma and oncocytoma on multiphasic CT. Abdom Radiol (New York). 2019;44:2009–20.
45. Feng Z, Rong P, Cao P, Zhou Q, Zhu W, Yan Z, et al. Machine learning-based quantitative texture analysis of CT images of small renal masses: Differentiation of angiomyolipoma without visible fat from renal cell carcinoma. Eur Radiol. 2018;28:1625–33.
46. Yap FY, Varghese BA, Cen SY, Hwang DH, Lei X, Desai B, et al. Shape and texture-based radiomics signature on CT effectively discriminates benign from malignant renal masses. Eur Radiol. 2021;31:1011–21.
47. Alhussaini AJ, Steele JD, Nabi G. Comparative Analysis for the Distinction of Chromophobe Renal Cell Carcinoma from Renal Oncocytoma in Computed Tomography Imaging Using Machine Learning Radiomics Analysis. Cancers (Basel). 2022;14:3609.
48. Liu J, Yildirim O, Akin O, Tian Y. AI-Driven Robust Kidney and Renal Mass Segmentation and Classification on 3D CT Images. Bioengineering. 2023;10:116.
49. Pedersen M, Andersen MB, Christiansen H, Azawi NH. Classification of renal tumour using convolutional neural networks to detect oncocytoma. Eur J Radiol. 2020;133:109343.
50. Said D, Hectors SJ, Wilck E, Rosen A, Stocker D, Bane O, et al. Characterization of solid renal neoplasms using MRI-based quantitative radiomics features. Abdom Radiol (New York). 2020;45:2840–50.
51. Uhm K-H, Jung S-W, Choi MH, Shin H-K, Yoo J-I, Oh SW, et al. Deep learning for end-to-end kidney cancer diagnosis on multi-phase abdominal computed tomography. NPJ Precis Oncol. 2021;5:54.
52. Yang R, Wu J, Sun L, Lai S, Xu Y, Liu X, et al. Radiomics of small renal masses on multiphasic CT: accuracy of machine learning-based classification models for the differentiation of renal cell carcinoma and angiomyolipoma without visible fat. Eur Radiol. 2020;30:1254–63.
53. Zabihollahy F, Schieda N, Ukwatta E. Patch-based convolutional neural network for differentiation of cyst from solid renal mass on contrast-enhanced computed tomography images. IEEE Access. 2020;8:8595–602.
54. Sassa N, Kameya Y, Takahashi T, Matsukawa Y, Majima T, Tsuruta K, et al. Creation of synthetic contrast-enhanced computed tomography images using deep neural networks to screen for renal cell carcinoma. Nagoya J Med Sci. 2023;85:713–24.
55. Maclin PS, Dempsey J, Brooks J, Rand J. Using neural networks to diagnose cancer. J Med Syst. 1991;15:11–9.
56. Baghdadi A, Aldhaam NA, Elsayed AS, Hussein AA, Cavuoto LA, Kauffman E, et al. Automated differentiation of benign renal oncocytoma and chromophobe renal cell carcinoma on computed tomography using deep learning. BJU Int. 2020;125:553–60.
57. Xi IL, Zhao Y, Wang R, Chang M, Purkayastha S, Chang K, et al. Deep Learning to Distinguish Benign from Malignant Renal Lesions Based on Routine MR Imaging. Clin cancer Res. 2020;26:1944–52.
58. Oberai A, Varghese B, Cen S, Angelini T, Hwang D, Gill I, et al. Deep learning based classification of solid lipid-poor contrast enhancing renal masses using contrast enhanced CT. Br J Radiol. 2020;93:20200002.
59. Yao H, Tian L, Liu X, Li S, Chen Y, Cao J, et al. Development and external validation of the multichannel deep learning model based on unenhanced CT for differentiating fat-poor angiomyolipoma from renal cell carcinoma: a two-center retrospective study. J Cancer Res Clin Oncol. 2023;149:15827–38.
60. Garnier C, Ferrer L, Vargas J, Gallinato O, Jambon E, Le Bras Y, et al. A CT-Based Clinical, Radiological and Radiomic Machine Learning Model for Predicting Malignancy of Solid Renal Tumors (UroCCR-75). Diagnostics. 2023;13:2548.
61. Mahmud S, Abbas TO, Mushtak A, Prithula J, Chowdhury MEH. Kidney Cancer Diagnosis and Surgery Selection by Machine Learning from CT Scans Combined with Clinical Metadata. Cancers (Basel). 2023;15:3189.
62. Qu J, Zhang Q, Song X, Jiang H, Ma H, Li W, et al. CT differentiation of the oncocytoma and renal cell carcinoma based on peripheral tumor parenchyma and central hypodense area characterisation. BMC Med Imaging. 2023;23:16.
63. Glembin M, Obuchowski A, Klaudel B, Rydziński B, Karski R, Syty P, et al. Enhancing Renal Tumor Detection: Leveraging Artificial Neural Networks in Computed Tomography Analysis. Med Sci Monit. 2023;29:e939462.
64. Feng S, Gong M, Zhou D, Yuan R, Kong J, Jiang F, et al. A CT-based radiomics nomogram for differentiation of benign and malignant small renal masses (≤4 cm). Transl Oncol. 2023;29:101627.
65. Uhlig J, Leha A, Delonge LM, Haack A-M, Shuch B, Kim HS, 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). 2020;12:3010.
66. Myers MR, Ravipati C, Thangam V. Artificial Intelligence-Based Non-invasive Differentiation of Distinct Histologic Subtypes of Renal Tumors With Multiphasic Multidetector Computed Tomography. Cureus. 2024;16:e57959.
67. Anari PY, Lay N, Zahergivar A, Firouzabadi FD, Chaurasia A, Golagha M, et al. Deep learning algorithm (YOLOv7) for automated renal mass detection on contrast-enhanced MRI: a 2D and 2.5D evaluation of results. Abdom Radiol (New York). 2024;49:1194–201.
68. Wentland AL, Yamashita R, Kino A, Pandit P, Shen L, Brooke Jeffrey R, et al. Differentiation of benign from malignant solid renal lesions using CT-based radiomics and machine learning: comparison with radiologist interpretation. Abdom Radiol (New York). 2023;48:642–8.
69. Yang H, Liu H, Lin J, Xiao H, Guo Y, Mei H, et al. An automatic texture feature analysis framework of renal tumor: surgical, pathological, and molecular evaluation based on multi-phase abdominal CT. Eur Radiol. 2024;34:355–66.
70. Shebel H, Abou El Atta HM, El-Diasty T, Sharaf DE. Predictive quantitative multidetector computed tomography models for characterization of renal cell carcinoma subtypes and differentiation from renal oncocytoma: nomogram algorithmic approach analysis. Egypt J Radiol Nucl Med. 2024;55:138.
71. Kang H, Xie W, Wang H, Guo H, Jiang J, Liu Z, et al. Multiparametric MRI-Based Machine Learning Models for the Characterization of Cystic Renal Masses Compared to the Bosniak Classification, Version 2019: A Multicenter Study. Acad Radiol. 2024;31:3223–34.
72. Han J, Chen B, Cheng C, Liu T, Tao Y, Lin J, et al. Development and Validation of a Diagnostic Model for Identifying Clear Cell Renal Cell Carcinoma in Small Renal Masses Based on CT Radiological Features: A Multicenter Study. Acad Radiol. 2024;31:4085–95.
73. Francischello R, Fanni SC, Chiellini M, Febi M, Pomara G, Bandini C, et al. Radiomics-based machine learning role in differential diagnosis between small renal oncocytoma and clear cells carcinoma on contrast-enhanced CT: A pilot study. Eur J Radiol Open. 2024;13:100604.
74. Wang R, Zhong L, Zhu P, Pan X, Chen L, Zhou J, et al. MRI-based radiomics machine learning model to differentiate non-clear cell renal cell carcinoma from benign renal tumors. Eur J Radiol open. 2024;13:100608.
75. Klontzas ME, Kalarakis G, Koltsakis E, Papathomas T, Karantanas AH, Tzortzakakis A. Convolutional neural networks for the differentiation between benign and malignant renal tumors with a multicenter international computed tomography dataset. Insights Imaging. 2024;15:26.
76. Toffoli T, Saut O, Etchegaray C, Jambon E, Le Bras Y, Grenier N, et al. Differentiation of Small Clear Renal Cell Carcinoma and Oncocytoma through Magnetic Resonance Imaging-Based Radiomics Analysis: Toward the End of Percutaneous Biopsy. J Pers Med. 2023;13:1444. https://doi.org/10.3390/jpm13101444.
77. Aymerich M, García-Baizán A, Franco PN, Otero-García M. Exploratory Analysis of the Role of Radiomic Features in the Differentiation of Oncocytoma and Chromophobe RCC in the Nephrographic CT Phase. Life (Basel). 2023;13:1950. https://doi.org/10.3390/life13101950.
78. Gaikar R, Azad A, Schieda N, Ukwatta E. Fully automated cascaded approach for renal mass detection on T2 weighted MRI images. In: Proc. SPIE; 2023. p. 124651M. https://doi.org/10.1117/12.2653286.

79. Nassiri N, Maas M, Cacciamani G, Varghese B, Hwang D, Lei X, et al. A Radiomic-based Machine Learning Algorithm to Reliably Differentiate Benign Renal Masses from Renal Cell Carcinoma. Eur Urol Focus. 2022;8:988–94.
80. Bolocan VO, Secareanu M, Sava E, Medar C, Manolescu LSC, Cătălin Raşcu AŞ, et al. Convolutional Neural Network Model for Segmentation and Classification of Clear Cell Renal Cell Carcinoma Based on Multiphase CT Images. J Imaging. 2023;9:280. https://doi.org/10.3390/jimaging9120280. Erratum in: J Imaging. 2024;10:35. https://doi.org/10.3390/jimaging10020035.
81. Uhlig A, Uhlig J, Leha A, Biggemann L, Bachanek S, Stöckle M, et al. Radiomics and machine learning for renal tumor subtype assessment using multiphase computed tomography in a multicenter setting. Eur Radiol. 2024;34:6254–63.
82. Yu S, Yang Y, Wang Z, Zheng H, Cui J, Zhan Y, et al. CT-based conventional radiomics and quantification of intratumoral heterogeneity for predicting benign and malignant renal lesions. Cancer imaging Off Publ Int Cancer Imaging Soc. 2024;24:130.
83. Magherini R, Servi M, Volpe Y, Campi R, Buonamici F. Distinguishing Kidney Tumor Types Using Radiomics Features and Deep Features. IEEE Access. 2024;12:84241–52. https://doi.org/10.1109/ACCESS.2024.3412655.
84. Zhou T, Guan J, Feng B, Xue H, Cui J, Kuang Q, et al. Distinguishing common renal cell carcinomas from benign renal tumors based on machine learning: comparing various CT imaging phases, slices, tumor sizes, and ROI segmentation strategies. Eur Radiol. 2023;33:4323–32.
85. Lee HS, Hong H, Jung DC, Park S, Kim J. Differentiation of fat-poor angiomyolipoma from clear cell renal cell carcinoma in contrast-enhanced MDCT images using quantitative feature classification. Med Phys. 2017;44:3604–14.
86. Lee H, Hong H, Kim J, Jung DC. Deep feature classification of angiomyolipoma without visible fat and renal cell carcinoma in abdominal contrast-enhanced CT images with texture image patches and hand-crafted feature concatenation. Med Phys. 2018;45:1550–61.
87. Shehata M, Abouelkheir RT, Gayhart M, Van Bogaert E, Abou El-Ghar M, Dwyer AC, et al. Role of AI and Radiomic Markers in Early Diagnosis of Renal Cancer and Clinical Outcome Prediction: A Brief Review. Cancers (Basel). 2023;15:2835.
88. Shaker G. 175 Beyond Human Capabilities: The Potential of AI in Robotic Urology. Br J Surg. 2024;111 Supplement_6:znae163.721.
89. Zhu M, Gu Z, Chen F, Chen X, Wang Y, Zhao G. Application of artificial intelligence in the diagnosis and treatment of urinary tumors. Front Oncol. 2024;14:1440626.
90. Nahar S, Mizanur Rahman M, Mahmud MS, Obaedul H. Rafiq SM, Sharmin S. Diagnostic accuracy of computed tomography scan in renal cell carcinoma. Int J Res Med Sci. 2023;11:1895–9. https://doi.org/10.18203/23206012.ijrms20231598.
91. Vogel C, Ziegelmüller B, Ljungberg B, Bensalah K, Bex A, Canfield S, et al. Imaging in Suspected Renal-Cell Carcinoma: Systematic Review. Clin Genitourin Cancer. 2019;17:e345–55.
92. Ferro M, Crocetto F, Barone B, Del Giudice F, Maggi M, Lucarelli G, et al. Artificial intelligence and radiomics in evaluation of kidney lesions: a comprehensive literature review. Ther Adv Urol. 2023;15:17562872231164804.
93. Xu H, Li W, Cai Z. Analysis on methods to effectively improve transfer learning performance. Theor Comput Sci. 2023;940:90–107.
94. Garcea F, Serra A, Lamberti F, Morra L. Data augmentation for medical imaging: A systematic literature review. Comput Biol Med. 2023;152:106391.
95. Ho S, Phua K, Wong L, Goh W. Extensions of the External Validation for Checking Learned Model Interpretability and Generalizability. Patterns. 2020;1:100129.
96. Šimundić A-M. Measures of Diagnostic Accuracy: Basic Definitions. EJIFCC. 2009;19:203–11.
97. Mohammadi S, Salehi MA, Jahanshahi A, Shahrabi Farahani M, Zakavi SS, Behrouzieh S, et al. Artificial intelligence in osteoarthritis detection: A systematic review and meta-analysis. Osteoarthr Cartil. 2023. https://doi.org/10.1016/j.joca.2023.09.011.
98. Zheng Q, Yang L, Zeng B, Li J, Guo K, Liang Y, et al. Artificial intelligence performance in detecting tumor metastasis from medical radiology imaging: A systematic review and meta-analysis. EClinicalMedicine. 2021;31:100669.

Publisher's Note