Diagn Interv Radiol 2024; DOI: 10.4274/dir.2023.232386

ARTIFICIAL INTELLIGENCE AND INFORMATICS

ORIGINAL ARTICLE

Cystic renal mass screening: machine-learning-based radiomics on unenhanced computed tomography

Lesheng Huang[1*]
Yongsong Ye[2*]
Jun Chen[1,2]
Wenhui Feng[3]
Se Peng[4]
Xiaohua Du[5]
Xiaodan Li[6]
Zhixuan Song[7]
Tianzhu Liu[1]

[1]Guangdong Provincial Hospital of Chinese Medicine, Department of Radiology, Zhuhai, China

[2]Guangdong Provincial Hospital of Chinese Medicine, Department of Radiology, Guangzhou, China

[3]Zhuhai People's Hospital, Department of Radiology, Zhuhai, China

[4]Guangdong Provincial Hospital of Chinese Medicine, Department of Laboratory Medicine, Zhuhai, China

[5]Guangdong Provincial Hospital of Chinese Medicine, Department of Pathology, Guangzhou, China

[6]Guangdong Provincial Hospital of Chinese Medicine, Department of Gynaecology, Zhuhai, China

[7]Philips Healthcare, Clinical and Technical Support, Guangzhou, China

PURPOSE

The present study compares the diagnostic performance of unenhanced computed tomography (CT) radiomics-based machine learning (ML) classifiers and a radiologist in cystic renal masses (CRMs).

METHODS

Patients with pathologically diagnosed CRMs from two hospitals were enrolled in the study. Unenhanced CT radiomic features were extracted for ML modeling in the training set (Guangzhou; 162 CRMs, 85 malignant). Total tumor segmentation was performed by two radiologists. Features with intraclass correlation coefficients of >0.75 were screened using univariate analysis, least absolute shrinkage and selection operator, and bidirectional elimination to construct random forest (RF), decision tree (DT), and k-nearest neighbor (KNN) models. External validation was performed in the Zhuhai set (45 CRMs, 30 malignant). All images were assessed by a radiologist. The ML models were evaluated using calibration curves, decision curves, and receiver operating characteristic (ROC) curves.

RESULTS

Of the 207 patients (102 women; 59.1 ± 11.5 years), 92 (41 women; 58.0 ± 13.7 years) had benign CRMs, and 115 (61 women; 59.8 ± 11.4 years) had malignant CRMs. The accuracy, sensitivity, and specificity of the radiologist's diagnoses were 85.5%, 84.2%, and 91.1%, respectively [area under the (ROC) curve (AUC), 0.87]. The ML classifiers showed similar sensitivity (94.2%–100%), specificity (94.7%–100%), and accuracy (94.3%–100%) in the training set. In the validation set, KNN showed better sensitivity, accuracy, and AUC than DT and RF but weaker specificity. Calibration and decision curves showed excellent and good results in the training and validation set, respectively.

CONCLUSION

Unenhanced CT radiomics-based ML classifiers, especially KNN, may aid in screening CRMs.

KEYWORDS

Cystic renal mass, diagnosis, radiomics, machine-learning

Corresponding author: Tianzhu Liu

E-mail: hadesfantasy012@21cn.com

Received 27 Jun 2023; revision requested 30 July 2023; last revision received 23 November 2023; accepted 30 November 2023.

Epub: 02.01.2024

Publication date: 08.07.2024

DOI: 10.4274/dir.2023.232386

Cystic renal masses (CRMs) are defined as renal lesions with <25% enhancing tissue, and they are often identified incidentally on abdominal computed tomography (CT) scans.[1] The majority of CRMs are benign, but a minority are diagnosed as renal cell carcinoma or other rare malignant renal tumors.[2,3] The proposed 2019 version of the Bosniak classification stratifies CRMs according to their risk of malignancy;[1] however, the diagnostic accuracy of this classification is low when applied to unenhanced CT scans because of the poor ability to visually judge gray-scale features with the naked eye.[4] Unfortunately, plain CT scans are commonly used in many situations, such as renal insufficiency, night-time emergencies, and especially annual CT examinations. Thus, a technique that enables the use of unenhanced CT scans for the accurate stratification of CRMs could assist radiologists and surgeons in screening to differentiate between malignant and benign CRMs.

You may cite this article as: Huang L, Ye Y, Chen J, et al. Cystic renal mass screening: machine-learning-based radiomics on unenhanced computed tomography. *Diagn Interv Radiol*. 2024;30(4):236-247.

Radiomic features have the potential to aid in the classification of lesion characteristics.[5] This quantitative approach to analyzing microscopic differences represents an emerging method in the pursuit of better understanding and identifying tumor phenotypes, although further research is required to establish specific feature-to-property correlations and standardize methodologies. Multiple supervised machine learning (ML) classifiers, such as the support vector machine, random forest (RF), decision tree (DT), and k-nearest neighbor (KNN), can be used to build diagnostic models based on radiomic features. Numerous studies on renal cell carcinoma have confirmed the excellent diagnostic efficacy of radiomics-based ML methods.[6-10] Recently, several ML algorithms were applied to classify CRMs into benign or malignant masses by using CT-based radiomic features.[11-13] While these studies are important and indispensable, further research on CRMs and ML is required for a number of reasons. First, previous algorithms were trained with arterial-phase (AP) and venous-phase (VP) scans; unenhanced CT features were either not used at all or only used as a supplementary part during model construction.[11-13] Second, some studies[11] lacked external data validation or validation in other centers to verify the diagnostic effectiveness of the models constructed. Finally, the above studies did not compare the diagnostic effectiveness of the ML-based models with that of manual diagnosis by experienced radiologists. To overcome the above shortcomings, the present authors aimed to build diagnostic ML models of CRMs based on unenhanced CT radiomic features; these models were verified with external data from a different center, and the diagnostic efficiency of the ML classifiers was compared with that of manual diagnosis.

Main points

- Several machine learning (ML) algorithms have been used to classify cystic renal masses (CRMs) into benign or malignant masses using computed tomography (CT)-based radiomic features, but previous algorithms were trained with arterial-phase and venous-phase scans.
- The present study showed that ML algorithms with unenhanced-CT radiomics features also presented acceptable diagnostic efficiency. The k-nearest neighbor (KNN) model presented satisfactory sensitivity and accuracy and was similar to the radiologist's performance, and the decision tree and random forest models presented satisfactory specificity.
- Due to its satisfactory sensitivity, the KNN model could be a potential screening method for patients with CRMs.

Methods

Ethics approval and case selection

This retrospective study was approved by the Medical Ethics Committee of Guangdong Provincial Hospital of Chinese Medicine (no: ZE2023-090-01), and the requirement for written informed consent was waived. Patients with CRMs who were treated at Guangdong Provincial Hospital of Chinese Medicine in either Guangzhou or Zhuhai (Center 1: Guangzhou and Center 2: Zhuhai) between January 2018 and February 2022 were eligible for this study. The inclusion criteria were as follows: (a) unenhanced and enhanced CT scans, including AP and VP images, were completed for the stratification of CRMs with the Bosniak classification; (b) complete clinical data were available, including age, sex, location of the lesions, intact operation and/or biopsy records, and histopathological results (obtained from the pathological retrieval systems of the two centers); and (c) good-quality CT images were stored in the Picture Archiving and Communications System. The exclusion criteria were (a) low-quality or incomplete CT data and (b) masses belonging to category II or lower according to the Bosniak classification. After the application of the above selection criteria, a total of 207 cases (92 benign and 115 malignant CRMs) were included in the study. The cases from Center 1 (77 benign and 85 malignant CRMs) were allocated to the training set, while the cases from Center 2 (15 benign and 30 malignant CRMs) were assigned to the validation set for external validation. The workflow of the ML approach is shown in Figure 1, and a flow chart of the case selection is shown in Figure 2.

Computed tomography examinations

All patients underwent unenhanced and dual-phase contrast-enhanced CT. The CT scanning was performed using three CT scanners: Definition Flash (Siemens, Forchheim, Germany) and IQon Spectral (Philips Healthcare, Amsterdam, Netherlands) in Center 1, and Aquilion One 750 W (Canon, Tokyo, Japan) in Center 2. Images obtained in three phases (unenhanced, AP, and VP) were used for the Bosniak classification, and unenhanced images were used for radiomic-feature extraction. The following scanning parameters were applied for all images: tube voltage = 120 kV; tube current = 250 mA; section interval = 5 mm; section thickness = 5 mm; and matrix size = 512 × 512 mm. After conventional unenhanced scanning, 100–120 mL of the contrast medium, iopromide (Ultravist 370, Bayer Schering Pharma, Germany) was injected into the median cubital vein via a pump injector (MEDRAD Stellant CT, Ulrich Medical, Ulm, Germany) at a flow rate of 3–4 mL/s. The AP was scanned using an aortic monitoring trigger, and the VP was scanned after approximately 60 s of delay after the contrast medium injection.

A single radiologist (J.C.) with 18 years of experience analyzed all the CT images to (a) check that all cases met the standard of <25% enhancing tissue, (b) confirm the Bosniak class (version 2019), and (c) measure the size of the CRMs.

Mass segmentation and radiomic-feature extraction

The open-source software platform, 3D Slicer (version 5.2.1, www.slicer.org), was applied for mass segmentation and calculation of radiomic features. Masses were delineated on the original CT images using 3D Slicer. Segmentation of whole masses was performed by associate chief radiologists (T.L. and L.H.) with more than 15 years of experience in abdominal radiography; to outline the shape and edges of the masses more accurately, the radiologists were allowed to observe the enhanced CT images. In each case, the entire CRM was carefully and manually segmented to avoid beyond-boundary or insufficient filling. Following tumor segmentation, 855 radiomic features were extracted using the "PyRadiomics" package with 3D Slicer. The extracted features were classified into seven categories as follows: first-order features, two-dimensional features, gray-level co-occurrence matrix, gray-level dependence matrix, gray-level size-zone matrix, gray-level run-length matrix, and neighboring gray tone difference matrix. Additionally, the following 14 filters were applied to the original images: exponential, gradient, square, square root, logarithm, lbp2D, wavelet-HLH, wavelet-HLL, wavelet-LHL, wavelet-LLL, wavelet-LHH, wavelet-LLH, wavelet-HHL, and wavelet-HHH. The images thus derived were analyzed for each patient. All classes of features were computed on both the original images and the derived images.

To ensure the stability of the radiomic features extracted from the CT images, the segmentation and feature-extraction process was repeated in 80 randomly selected patients with CRMs from the training set.

Intraclass correlation coefficients (ICCs) were used to evaluate consistency across the radiomic features; features with ICCs >0.75 were considered stable and were included in this analysis.

After meeting the standard of consistency, the features were further selected to avoid overfitting. The least absolute shrinkage and selection operator (LASSO) method was applied to select the most suitable radiomic features to develop a radiomic signature with the "glmnet" package. First, 10-fold cross-validation was performed to obtain the optimal parameter λ[14] by 1,000 iterations. Second, the LASSO method based on the optimal parameter λ was used to calculate the coefficient of each feature, and features with non-zero coefficients were selected.[14] Finally, bidirectional elimination was used to further filter the radiomic features selected using the LASSO method;[15] the "mass" package in the R software (version 4.2.2) was used for bidirectional elimination (Figure 1).

Statistical analysis

The χ^2 test was used to compare categorical data, and the independent-samples t-test was used to compare inter-group differences in clinical data. Statistical analysis was conducted using SPSS (version 26.0, IBM, Armonk, NY, USA) and R (version 4.2.2). A two-sided *P* value of <0.05 was considered statistically significant.

Machine learning algorithms

The radiomic features selected using the above steps were standardized to a mean of 0 and an standard deviation of 1 before ML algorithm construction. Supervised learning was achieved using three supervised learning classifiers: RF, DT, and KNN. A 10-fold cross-validation strategy was applied to assess the performance of the classification models. Under this strategy, the data were divided into 10 parts; nine parts were used for training in turn, and the remaining part was used to estimate the efficacy of the models. During the process of fine-tuning the models, the grid search method was employed to select the best combination of hyperparameter values.

Patients from Center 1 (77 benign and 85 malignant CRMs) were allocated to the training set, and patients from Center 2 (15 benign and 30 malignant CRMs) were allocated to the validation set for external validation to estimate the performance of the models. The discriminative performance of different models was quantified using area under the [receiver operating characteristic (ROC)] curve (AUC), accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). The AUCs of the ML models were also compared using the Delong test. The SHapley Additive exPlanations (SHAP) values, which indicate the importance of radiomic features, were derived for the RF and DT models (SHAP values are not suitable for the KNN model).

The ML algorithm creation was performed using the "Caret" package. Calibration curves were plotted using the "rms" package. Decision curve analysis (DCA) was performed using the "rmda" and "ggDCA" packages. The "pROC" package was used for ROC curve analysis. The "reportROC" package was used to present the sensitivities, specificities, accuracies, PPVs, NPVs, and 95% confidence intervals (CIs) of the AUCs obtained using ROC curve analysis.

Manual diagnosis by radiologist

The study authors also assessed the diagnostic performance of an attending radiologist (W.F.) with more than 7 years of experience in radiology diagnoses. This radiologist used the open-source DICOM viewer MicroDicom (https://www.microdicom.com/) for image evaluation. The radiologist was from a hospital not involved in this study and was blinded to the patient demographic and clinical characteristics. The radiologist independently reviewed the unenhanced CT images and established a diagnosis according to the Bosniak classification, based on the morphological features of the lesions.

Quality evaluation of research

To evaluate the quality of research, the study authors used the CheckList for EvaluAtion of Radiomics research (CLEAR)[16] and the radiomics quality score (RQS).[17] The datasets and source code generated and/or analyzed during the current study are available on GitHub (https://github.com/elliiesong/CRM-screening-with-machine-learning-unenhanced-CT).

Results

Patient characteristics

This study included 207 patients (105 men, 102 women; mean age: 59.1 ± 11.5 years) with CRMs. Of these, 92 patients (51 men, 41 women; mean age: 58.0 ± 13.7 years) had benign CRMs, and 115 patients (54 men, 61 women; mean age: 59.8 ± 11.4 years) had malignant CRMs (Figure 2). There were no significant differences in age, sex, or mass location or size between patients with benign or malignant CRMs (Table 1). All benign CRMs were simple kidney cysts, except for one case of angiomyolipoma. All malignant CRMs were clear cell carcinoma, except for one case of mixed epithelial and stromal tumor of the kidney.

Radiomic-feature selection

Following univariate analyses, 216 radiomic features were extracted from the unenhanced CT images, and LASSO and 10-fold cross-validation were used to screen and select radiomic features. Finally, the following four features screened out from unenhanced CT images were selected: Original_glcm_Maximum_Probability, Wavelet.LHH_firstorder_Median, Wavelet.LLL_firstorder_90Percentile, and Wavelet.LLL_firstorder_Median.

Diagnostic performance of machine learning algorithms

Four features (Original_glcm_Maximum_Probability, Wavelet.LHH_firstorder_Median, Wavelet.LLL_firstorder_90Percentile, and Wavelet.LLL_firstorder_Median) were used to construct the ML models. The diagnostic efficiencies of the ML classifiers are summarized in Table 2 and Figure 3. In the training set, the accuracy, specificity, sensitivity, and AUC of RF, DT, and KNN (*k*-value: 4) were satisfactory and similar to each other. A confusion matrix was prepared from the verification set, and the accuracy of RF, DT, and KNN in this set was 77.3% (95% CI: 76.5%–78.1%), 79.5% (95% CI: 78.8%–80.3%), and 84.1% (95% CI: 83.5%–84.7%), respectively. The specificity of KNN (73.3%, 95% CI: 51.0%–95.7%) was significantly weaker than that of RF (80.6%, 95% CI: 60.7%–100%) and DT (80.0%, 95% CI: 59.8%–100%). The sensitivity of KNN (89.7%, 95% CI: 78.6%–100%) was significantly better than that of RF (65.5%, 95% CI: 48.2%–82.8%) and DT (79.3%, 95% CI: 64.6%–94.1%). The AUC of KNN (0.86, 95% CI: 0.74–0.98) was slightly better than that of RF (0.77, 95% CI: 0.61–0.92) and DT (0.80, 95% CI: 0.67–0.93). None of the ML classifiers significantly differed from manual diagnosis (Supplemantary Table S1). The results of the Delong test showed that there was no statistical difference between the ML classifiers (KNN and RF: $P = 0.205$; KNN and DT: $P = 0.061$; RF and DT: $P = 0.586$). The SHAP values of DT and RF (Supplemantary Figure S1) showed that the feature Wavelet.LLL_firstorder_Median held absolute weight in the two models, especially in the DT model.

Calibration curve analysis and DCA of the ML classifiers were performed in the training and validation sets (Figure 3c-f). The calibration curves were excellent and close to the ideal line in the training set but showed some degree of deviation from the ideal line in the validation set. The KNN and DT lines were above the ideal line but became close to and intersected the ideal line in the latter half, and the RF line was below the ideal line in the first half and above it in the second half. The DCA showed excellent results in the training set and revealed a greater net benefit than all positive and negative lines when the risk threshold was more than approximately 0.3 in the validation set; the KNN, DT, and RF lines were similar.

Efficiency of manual diagnosis

The manual diagnosis results are summarized in Table 2 and Figure 3b. The radiologist's diagnoses using unenhanced CT images presented an accuracy, sensitivity, and specificity of 85.5%, 84.2%, and 91.1%, respectively, with an AUC of 0.866.

Radiomics quality score

The quality of this study was evaluated using CLEAR[16] and RQS.[17] The results of the CLEAR evaluation were 43/9/6 (Yes/No/n/a, total: 58), and the RQS was 47.22% (17/36). The details of the RQS and CLEAR are summarized in Supplementary Tables S2, S3.

Discussion

In this bicentric study, the authors attempted to create multiple ML classifiers to distinguish between benign and malignant CRMs on unenhanced CT images. The results indicated that the accuracy and AUC of the ML classifiers were satisfactory (accuracy: 77.3%–84.1%; AUC: 0.77–0.86) and similar to that of the radiologist's diagnoses. The KNN

Table 1. Clinical and pathological characteristics of the included CRM patients

	Training set (n = 162)			Validation set (n = 45)		
Characteristic	Benign (n = 77)	Malignant (n = 85)	*P* value	Benign (n = 15)	Malignant (n = 30)	*P* value
Age (years), mean ± SD	57.4 ± 9.8	60.3 ± 12.6	0.746	58.2 ± 10.6	61.6 ± 14.0	0.633
Gender			0.281			0.831
Male	41 (19.8%)	37 (17.87%)		9 (4.35%)	17 (8.21%)	
Female	36 (12.56%)	48 (23.18%)		6 (2.90%)	13 (6.28%)	
Mass size (cm), mean ± SD	4.80 ± 1.32	5.06 ± 1.85	0.790	4.77 ± 1.69	6.10 ± 1.22	0.509
Location						
Right kidney	41 (19.80%)	53 (25.60%)	0.311	6 (2.90%)	13 (6.28%)	0.831
Left kidney	36 (17.39%)	32 (15.46%)		9 (4.35%)	17 (8.21%)	
Histological subtype			<0.0001			<0.0001
Simple kidney cyst	77 (37.19%)	0 (0%)		14 (6.76%)	0 (0%)	
Clear cell carcinoma	0 (0%)	84 (40.57%)		0 (0%)	30 (14.50%)	
Other	0 (0%)	1 (0.48%)		1 (0.48%)	0 (0%)	
Bosniak classification			<0.0001			<0.0001
IIF	63 (30.43%)	16 (7.73%)		12 (5.80%)	3 (1.45%)	
III	14 (6.76%)	21 (10.14%)		3 (1.45%)	10 (4.83%)	
IV	0 (0%)	48 (23.18%)		0 (0%)	17 (8.21%)	

Data are expressed as mean ± SD, median (interquartile range), or frequency (constituent ratio). CRM, cystic renal mass; SD, standard deviation.

Table 2. Diagnostic efficiency of three computed tomography radiomic feature–based machine learning algorithms in differentiating benign from malignant cystic renal masses (n = 207) in the training and validation sets

Machine learning algorithm/ manual analysis	Sensitivity (%), (95% CI)	Specificity (%), (95% CI)	Accuracy (%), (95% CI)	PPV (%), (95% CI)	NPV (%), (95% CI)	AUC (95% CI)
Training set						
RF	100 (99.2–100)	100 (98–100)	100 (98.9–100)	100 (99.4–100)	100 (99.4–100)	1.00 (0.98–1.00)
DT	94.2 (89.2–99.1)	94.7 (89.7–99.8)	94.4 (94.4–94.5)	95.3 (90.8–99.8)	93.5 (88.0–99.0)	0.95 (0.91–0.98)
KNN	94.2 (89.2–99.1)	95.0 (90.1–100)	94.3 (94.0–94.7)	95.3 (91.0–99.8)	93.5 (88.0–99.0)	0.97 (0.95–0.99)
Validation set						
RF	65.5 (48.2–82.8)	80.6 (60.7–100)	77.3 (76.5–78.1)	87.4 (73.2–100)	57.2 (36.9–78.3)	0.77 (0.61–0.92)
DT	79.3 (64.6–94.1)	80.0 (59.8–100)	79.5 (78.8–80.3)	88.5 (76.2–100)	66.7 (44.9–88.4)	0.80 (0.67–0.93)
KNN	89.7 (78.6–100)	73.3 (51.0–95.7)	84.1 (83.5–84.7)	86.7 (74.5–98.8)	78.6 (57.1–100)	0.86 (0.74–0.98)
Radiologist	84.2	91.1	85.5	90.9	83.6	0.87

CI, confidence interval; PPV, positive predictive value; NPV, negative predictive value; AUC, area under the curve; RF, random forest; DT, decision tree; KNN, k-nearest neighbor.

presented the highest sensitivity and accuracy, and the DT and RF presented the highest specificity.

The Bosniak classification is the standard stratification method used to estimate the risk of malignancy in CRMs; however, this classification does have some limitations. First, ambiguous definitions, such as "cystic," "solid," "walls," and "septa," are difficult to quantify.[18-23] Second, the Bosniak classification is limited by considerable variability between radiologists,[24] especially for Bosniak classes II, IIF, and III, for which absolute disagreement ranges from 6% to 75%.[25] Finally, most CRMs are found incidentally, owing to which the scanning procedure is not planned for imaging the entire mass and may not include enhanced CT scans; hence, the Bosniak classification often cannot be applied.[26]

Compared with visual analysis, ML classifiers of radiomic features could more comprehensively and objectively reflect the phenotypic properties of masses, which may represent the underlying microscopic pathological changes and heterogeneity of the disease. The ML classifiers have potential benefits in screening CRMs: first, they are objective and not subject to reader interpretation, although segmentation by readers can still be needed; however, automatic segmentation has been used in some situations. Second, unlike the Bosniak classification, which depends on enhanced scanning, the ML classifiers can be applied to single-phase CT scans and may obviate additional radiological examinations.

Other diagnostic models based on radiomic features have also been studied. A decision algorithm used by Dana et al.[12] was built by combining consensus radiological readings of Bosniak categories and radiomics-based risks; the results showed excellent diagnostic performance (AUC: 0.96). He et al.[13] applied deep learning and a radiomic feature-based blending ensemble classifier to predict the malignancy risk of CRMs and obtained satisfactory diagnostic performance (AUC: 0.934). However, both these models were based on CT images obtained in the three phases or in the arterial phase. The following inferences can be drawn from the above findings: first, radiomic features play a valuable role in the diagnosis of CRMs; second, unenhanced CT scan-based radiomic features of CRMs were underappreciated in previous studies. Unlike other studies, the present study focused on unenhanced CT scan-based radiomic features and presented acceptable diagnostic efficiency (RF: AUC = 0.77; DT: AUC = 0.80; KNN: AUC = 0.86) in the absence of other CT phases.

Building on prior studies,[6,8,11] this study applied unenhanced CT-based ML classifiers independent of the Bosniak classification and compared their performance in the diagnosis of pathologically proven masses. Each of the three ML classifiers (RF, DT, and KNN) showed a similar high accuracy in distinguishing between benign and malignant CRMs. Although prior work has demonstrated the ability of ML classifiers to differentiate between benign and malignant

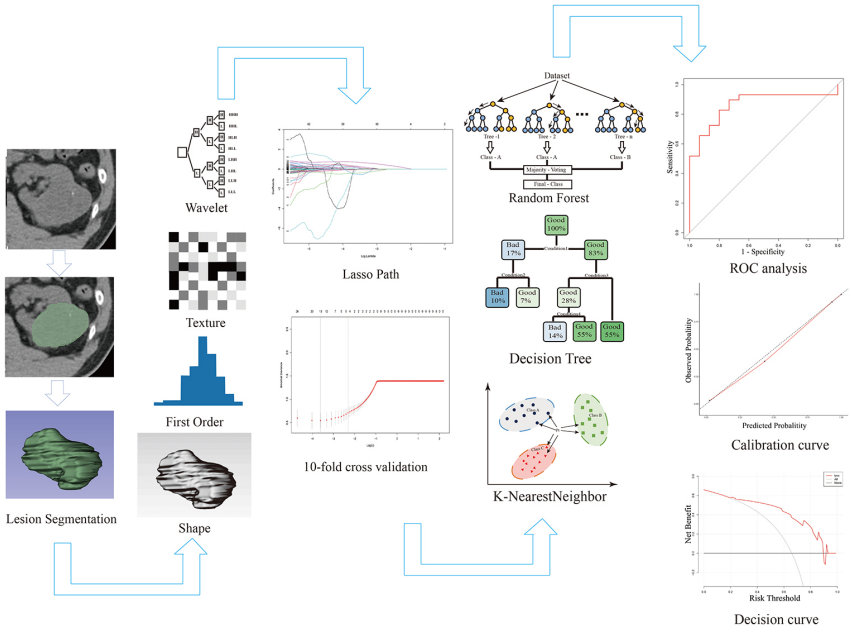


Figure 1. Workflow of the machine learning approach.

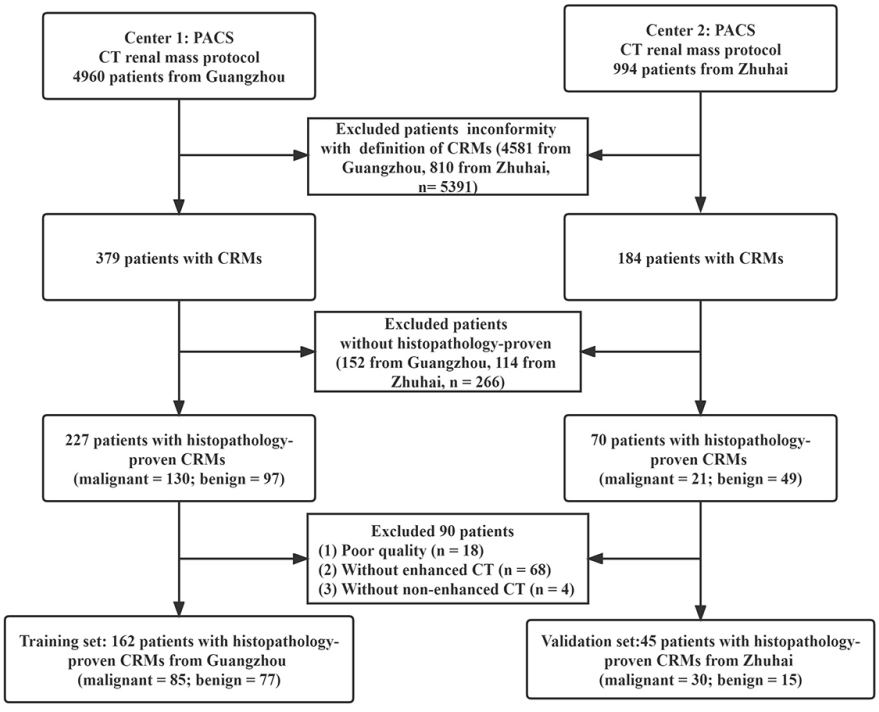


Figure 2. Flow chart of patient inclusion and exclusion criteria. PACS, Picture Archiving and Communications System; CT, computed tomography; CRMs, cystic renal masses.

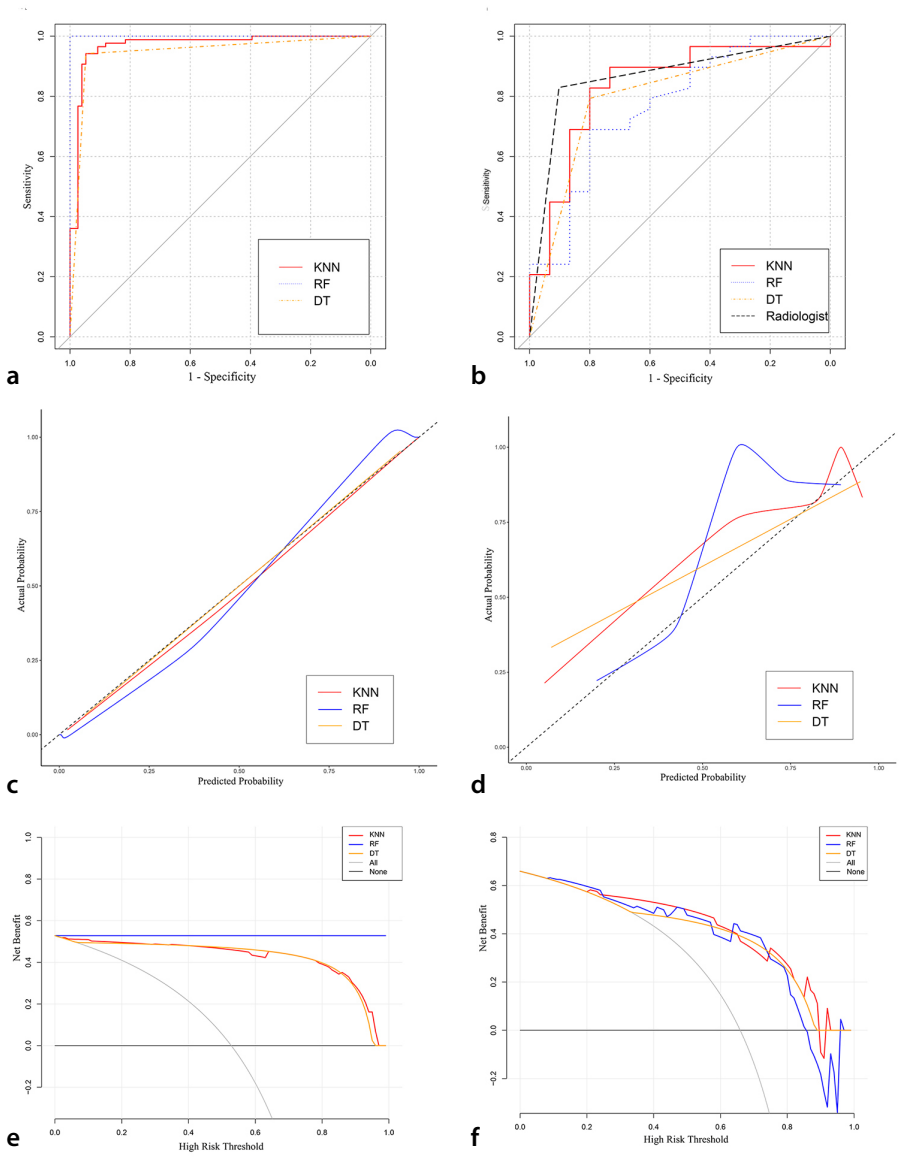


Figure 3. (a, b) Receiver operating characteristic curves of the machine learning (ML) classifiers for cystic renal masses (CRMs) in the training **(a)** and validation **(b)** sets. **(c, d)** Calibration curves of the ML classifiers for CRM prediction in the training **(c)** and validation **(d)** sets. **(e, f)** Decision curve analysis of the ML classifiers for CRMs in the training **(e)** and validation **(f)** sets.

solid or CRMs,[11] to the best of the authors' knowledge, this study is the first to develop ML classifiers to distinguish between benign and malignant CRMs based on unenhanced CT images, as well as compare the diagnostic effectiveness of ML classifiers with that of manual diagnosis by a radiologist.

The ML classifiers showed acceptable-to-high sensitivity (65.5%–89.7%) and specificity (73.3%–80.6%) in the validation set in this study. The authors considered satisfactory sensitivity of single-phase radiomics models, especially unenhanced models, important for clinical application because most CRMs are found incidentally, and an unenhanced model could provide a preliminary diagnosis to help clinicians make the next decision. In this study, KNN presented the highest sensitivity among the ML classifiers, which was better than that of manual diagnosis (KNN vs. radiologist: 89.7% and 84.2%, respectively). This indicates that KNN could screen malignant CRMs at a greater probability. Compared with the increased detection of suspected malignant masses that need further examination, such as enhanced CT or MR scanning, the misdiagnosis of malignant CRMs is a greater disadvantage and may cause patients to miss the optimal time window for treatment. An unenhanced CT-based KNN classifier could be a valuable diagnostic method for CRMs in clinical and radiological practice. Compared with the linear pattern of the DT line and the sigmoid pattern of the RF line, the KNN line in the calibration curve analysis was close to the ideal line in the second half. This may mean that the KNN classifier exhibited more adaptability in the positive diagnosis of CRMs. On the other hand, the composition and importance of features are also noteworthy points. In this study, three of the four radiomic features used for model predictions were computed with wavelet filters. Thus, radiomic features derived using wavelet filters dominated the models and may have had a significant impact on the predictive performance of the models.[27]

The drawbacks of ML classifiers need to be acknowledged. The ML classifiers used in this study are supervised methods that require a reader to segment the masses and extract the features; thus, the performance of the models may be affected by the segmentation process, unless an automatic segmentation is applied.

There are several limitations to this study. First, although this study is a bicentric study, the two hospitals share a set of CT scanning and image-reconstruction standards, although the CT scanning equipment is different; hence, the images can still have relatively high consistency. Verification with scans from other hospitals with different scanning parameters is required to confirm the diagnostic efficiency of the ML models from this study. Second, the composition of the validation set was not balanced (15 benign and 30 malignant CRMs), which may have led to potential risks and affected the validation results. Third, KNN is a simple classifier and has the potential risk of overfitting; hence, even though the diagnostic efficiency of the models was satisfactory in both the training and validation sets, more data are needed for verification. Fourth, the majority of patients were pathologically diagnosed with renal cysts and clear cell carcinomas; the diagnostic performance of the models on other pathological types of CRMs, such as papillary and tubular renal cell carcinomas, remains unconfirmed. To truly understand the models' capabilities across all pathological types, further comprehensive research is essential. Fifth, the radiologists were allowed to observe the enhanced CT images to delineate the boundaries of the masses, which may have led to bias in practical applications. Sixth, although identical CT acquisition and reconstruction settings were used in both centers, there is still a concern that the radiomic feature values may have been

affected by the use of different scanners (two scanners in the training cohort center, and one scanner in the validation cohort center). Thus, it may be necessary to apply a data harmonization procedure, such as ComBat and modified ComBat, for non-single center radiomics studies. Finally, there were some unusual findings for the RF model, such as the widening gap between the AUCs of this model in the training and validation sets and the parallel line in the DCA in the training set. The authors consider the RF model to possibly have the risk of overfitting.

In conclusion, ML classifiers based on unenhanced CT scans showed acceptable diagnostic efficiencies in the diagnosis of CRMs. Furthermore, KNN may be used as a potential screening method in patients with CRMs.

Acknowledgement

We thank Philips Healthcare Inc. and Medjaden Inc. for their technical support and assistance in the preparation of this manuscript.

Data sharing statement

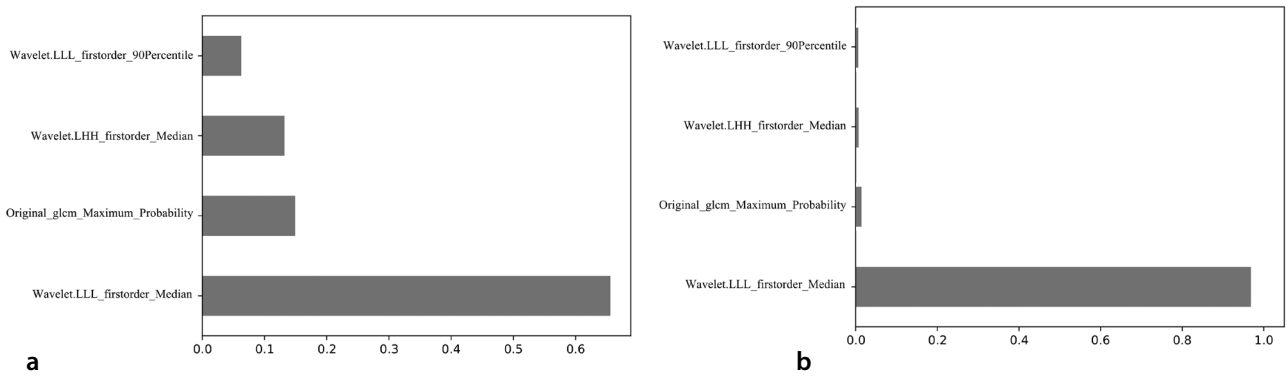
The source code, datasets, and models generated and/or analyzed during the current study are available on GitHub (https://github.com/elliiesong/CRM-screening-with-machine-learning-unenhanced-CT).

Conflict of interest disclosure

The authors declared no conflicts of interest.

References

1. Silverman SG, Pedrosa I, Ellis JH, et al. Bosniak classification of cystic renal masses, version 2019: an update proposal and needs assessment. *Radiology.* 2019;292(2):475-488. [CrossRef]
2. Terada N, Arai Y, Kinukawa N, Yoshimura K, Terai A. Risk factors for renal cysts. *BJU Int.* 2004;93(9):1300-1302. [CrossRef]
3. Carrim ZI, Murchison JT. The prevalence of simple renal and hepatic cysts detected by spiral computed tomography. *Clin Radiol.* 2003;58(8):626-629. [CrossRef]
4. Silverman SG, Israel GM, Herts BR, Richie JP. Management of the incidental renal mass. Radiology. 2008;249(1):16-31. [CrossRef]
5. Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. *Radiology.* 2016;278(2):563-577. [CrossRef]
6. Lubner MG. Radiomics and artificial intelligence for renal mass characterization. *Radiol Clin North Am.* 2020;58(5):995-1008. [CrossRef]
7. Mühlbauer J, Egen L, Kowalewski KF, et al. Radiomics in renal cell carcinoma-a systematic review and meta-analysis. *Cancers (Basel).* 2021;13(6):1348. [CrossRef]
8. Ursprung S, Beer L, Bruining A, et al. Radiomics of computed tomography and magnetic resonance imaging in renal cell carcinoma-a systematic review and meta-analysis. *Eur Radiol.* 2020;30(6):3558-3566. [CrossRef]
9. Li S, Liu J, Xiong Y, et al. Application values of 2D and 3D radiomics models based on CT plain scan in differentiating benign from malignant ovarian tumors. *Biomed Res Int.* 2022;2022:5952296. [CrossRef]
10. Yang X, He J, Wang J, et al. CT-based radiomics signature for differentiating solitary granulomatous nodules from solid lung adenocarcinoma. *Lung Cancer.* 2018;125:109-114. [CrossRef]
11. 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;47:376-381. [CrossRef]
12. Dana J, Lefebvre TL, Savadjiev P, et al. Malignancy risk stratification of cystic renal lesions based on a contrast-enhanced CT-based machine learning model and a clinical decision algorithm. *Eur Radiol.* 2022;32(6):4116-4127. [CrossRef]
13. He QH, Feng JJ, Lv FJ, Jiang Q, Xiao MZ. Deep learning and radiomic feature-based blending ensemble classifier for malignancy risk prediction in cystic renal lesions. *Insights Imaging.* 2023;14(1):6. [CrossRef]
14. Chen M, Cao J, Hu J, et al. Clinical-radiomic analysis for pretreatment prediction of objective response to first transarterial chemoembolization in hepatocellular carcinoma. *Liver Cancer.* 2021;10(1):38-51. [CrossRef]
15. Hu MJ, Yu YX, Fan YF, Hu CH. CT-based radiomics model to distinguish necrotic hepatocellular carcinoma from pyogenic liver abscess. *Clin Radiol.* 2021;76(2):161. [CrossRef]
16. Kocak B, Baessler B, Bakas S, et al. CheckList for EvaluAtion of Radiomics research (CLEAR): a step-by-step reporting guideline for authors and reviewers endorsed by ESR and EuSoMII. *Insights Imaging.* 2023;14(1):75. [CrossRef]
17. Lambin P, Leijenaar RTH, Deist TM, et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol.* 2017;14(12):749-762. [CrossRef]
18. Corica FA, Iczkowski KA, Cheng L, et al. Cystic renal cell carcinoma is cured by resection: a study of 24 cases with long-term followup. *J Urol.* 1999;161(2):408-411. [CrossRef]
19. Webster WS, Thompson RH, Cheville JC, Lohse CM, Blute ML, Leibovich BC. Surgical resection provides excellent outcomes for patients with cystic clear cell renal cell carcinoma. *Urology.* 2007;70(5):900-904. [CrossRef]
20. Jhaveri K, Gupta P, Elmi A, et al. Cystic renal cell carcinomas: do they grow, metastasize, or recur? *AJR Am J Roentgenol.* 2013;201(2):292-296. [CrossRef]
21. Cooperberg MR, Mallin K, Kane CJ, Carroll PR. Treatment trends for stage I renal cell carcinoma. *J Urol.* 2011;186(2):394-399. [CrossRef]
22. Daskivich TJ, Tan HJ, Litwin MS, Hu JC. Life expectancy and variation in treatment for early stage kidney cancer. *J Urol.* 2016;196(3):672-677. [CrossRef]
23. Kane CJ, Mallin K, Ritchey J, Cooperberg MR, Carroll PR. Renal cell cancer stage migration: analysis of the National Cancer Data Base. *Cancer.* 2008;113(1):78-83. [CrossRef]
24. Siegel CL, McFarland EG, Brink JA, Fisher AJ, Humphrey P, Heiken JP. CT of cystic renal masses: analysis of diagnostic performance and interobserver variation. *AJR Am J Roentgenol.* 1997;169(3):813-818. [CrossRef]
25. Schoots IG, Zaccai K, Hunink MG, Verhagen PCMS. Bosniak classification for complex renal cysts reevaluated: a systematic review. *J Urol.* 2017;198(1):12-21. [CrossRef]
26. Silverman SG, Israel GM, Trinh QD. Incompletely characterized incidental renal masses: emerging data support conservative management. *Radiology.* 2015;275(1):28-42. [CrossRef]
27. Demircioğlu A. The effect of preprocessing filters on predictive performance in radiomics. *Eur Radiol Exp.* 2022;6(1):40. [CrossRef]



Supplemantary Figure S1. **(a)** SHapley Additive exPlanations (SHAP) value of RF. **(b)** SHAP value of the decision tree.

Supplemantary Table S1. Comparison of the diagnostic efficiencies of 3 CT radiomic feature–based machine learning algorithms with that of a radiologist's diagnosis

Machine learning algorithm	Sensitivity	Specificity	Accuracy	PPV	NPV
		P value vs. radiologist			
RF	0.35	0.75	0.50	0.87	0.35
DT	0.86	0.75	0.73	0.90	0.60
KNN	0.84	0.61	0.90	0.84	0.78

PPV, positive predictive value; NPV, negative predictive value; RF, random forest; DT, decision tree; KNN, k-nearest neighbor; CT, computed tomography.

Supplemantary Table S2. Radiomics quality score of this research study

Criteria		Points
1	Image protocol quality - well-documented image protocols (for example, contrast, slice thickness, energy, etc.) and/or usage of public image protocols allow reproducibility/replicability	+ 2
2	Multiple segmentations - possible actions are: segmentation by different physicians/algorithms/software, perturbing segmentations by (random) noise, segmentation at different breathing cycles. Analyze feature robustness to segmentation variabilities	+ 1
3	Phantom study on all scanners - detect inter-scanner differences and vendor-dependent features. Analyze feature robustness to these sources of variability	**+ 0**
4	Imaging at multiple time points - collect images of individuals at additional time points. Analyze feature robustness to temporal variabilities (for example, organ movement, organ expansion/shrinkage)	+ 0
5	Feature reduction or adjustment for multiple testing - decreases the risk of overfitting. Overfitting is inevitable if the number of features exceeds the number of samples. Consider feature robustness when selecting features	+ 3
6	Multivariable analysis with non-radiomics features (for example, EGFR mutation) - is expected to provide a more holistic model. Permits correlating/inferencing between radiomics and non-radiomics features	+ 0
7	Detect and discuss biological correlates - demonstration of phenotypic differences (possibly associated with underlying gene–protein expression patterns) deepens understanding of radiomics and biology	+ 0
8	Cut-off analyses - determine risk groups by either the median, a previously published cut-off or report a continuous risk variable. Reduces the risk of reporting overly optimistic results	+ 0
9	Discrimination statistics - report discrimination statistics (for example, C-statistic, ROC curve, AUC) and their statistical significance (for example, *P* values, confidence intervals). One can also apply resampling method (for example, bootstrapping, cross-validation)	+ 1
10	Calibration statistics - report calibration statistics (for example, calibration-in-the-large/slope, calibration plots) and their statistical significance (for example, *P* values, confidence intervals). One can also apply resampling method (for example, bootstrapping, cross-validation)	+ 1
11	Prospective study registered in a trial database - provides the highest level of evidence supporting the clinical validity and usefulness of the radiomics biomarker	+ 0
12	Validation - the validation is performed without retraining and without adaptation of the cut-off value, provides crucial information with regard to credible clinical performance	**+ 3**
13	Comparison to 'gold standard' - assess the extent to which the model agrees with/is superior to the current 'gold standard' method (for example, TNM-staging for survival prediction). This comparison shows the added value of radiomics	+ 2
14	Potential clinical utility - report on the current and potential application of the model in a clinical setting (for example, decision curve analysis)	+ 2
15	Cost-effectiveness analysis - report on the cost-effectiveness of the clinical application (for example, QALYs generated)	+ 0
16	Open science and data - make code and data publicly available. Open science facilitates knowledge transfer and reproducibility of the study	+ 2
	Total	**17**

Total points (36 = 100%). ROC, receiver operating characteristic; AUC, area under the curve; TNM, tumor, node and metastasis.

Supplemantary Table S3.

CLEAR Checklist v1.0

Note: Use the checklist in conjunction with the main text for clarification of all items.
Yes, details provided; No, details not provided; n/e, not essential; n/a, not applicable; Page, page number

Section	No.	Item	Yes	No	n/a	Page
Title						
	1	Relevant title, specifying the radiomic methodology	■	☐	☐	1
Abstract						
	2	Structured summary with relevant information	■	☐	☐	2
Keywords						
	3	Relevant keywords for radiomics	■	☐	☐	2
Introduction						
	4	Scientific or clinical background	■	☐	☐	3
	5	Rationale for using a radiomic approach	■	☐	☐	3
	6	Study objective(s)	■	☐	☐	3-4
Method						
Study Design	7	Adherence to guidelines or checklists (e.g., CLEAR checklist)	■	☐	☐	Attachment
	8	Ethical details (e.g., approval, consent, data protection)	■	☐	☐	4
	9	Sample size calculation	☐	■	☐	
	10	Study nature (e.g., retrospective, prospective)	■	☐	☐	4
	11	Eligibility criteria	■	☐	☐	4
	12	Flowchart for technical pipeline	■	☐	☐	Fig 1
Data	13	Data source (e.g., private, public)	■	☐	☐	4
	14	Data overlap	☐	■	☐	
	15	Data split methodology	■	☐	☐	4
	16	Imaging protocol (i.e., image acquisition and processing)	■	☐	☐	4
	17	Definition of non-radiomic predictor variables	☐	☐	■	4-5
	18	Definition of the reference standard (i.e., outcome variable)	■	☐	☐	4
Segmentation	19	Segmentation strategy	■	☐	☐	5
	20	Details of operators performing segmentation	■	☐	☐	5
Pre-processing	21	Image pre-processing details	■	☐	☐	5
	22	Resampling method and its parameters	☐	■	☐	
	23	Discretization method and its parameters	☐	■	☐	

Supplemantary Table S3. continued

Section	No.	Item	Yes	No	n/a	Page
	24	Image types (e.g., original, filtered, transformed)	■	☐	☐	5
Feature extraction	25	Feature extraction method	■	☐	☐	5
	26	Feature classes	■	☐	☐	5
	27	Number of features	■	☐	☐	5
	28	Default configuration statement for remaining parameters	■	☐	☐	5
Data preparation	29	Handling of missing data	☐	☐	■	
	30	Details of class imbalance	☐	☐	■	
	31	Details of segmentation reliability analysis	■	☐	☐	5-6
	32	Feature scaling details (e.g., normalization, standardization)	■	☐	☐	6
	33	Dimension reduction details	☐	■	☐	
Modeling	34	Algorithm details	■	☐	☐	6
	35	Training and tuning details	■	☐	☐	6
	36	Handling of confounders	☐	■	☐	
	37	Model selection strategy	■	☐	☐	6
Evaluation	38	Testing technique (e.g., internal, external)	■	☐	☐	6
	39	Performance metrics and rationale for choosing	■	☐	☐	7
	40	Uncertainty evaluation and measures (e.g., confidence intervals)	■	☐	☐	
	41	Statistical performance comparison (e.g., DeLong's test)	■	☐	☐	7
	42	Comparison with non-radiomic and combined methods	■	☐	☐	7
	43	Interpretability and explainability methods	☐	■	☐	
Results						
	44	Baseline demographic and clinical characteristics	■	☐	☐	7-8
	45	Flowchart for eligibility criteria	■	☐	☐	Fig 2
	46	Feature statistics (e.g., reproducibility, feature selection)	■	☐	☐	8
	47	Model performance evaluation	■	☐	☐	8-9
	48	Comparison with non-radiomic and combined approaches	■	☐	☐	Table 2
Discussion						
	49	Overview of important findings	■	☐	☐	9
	50	Previous works with differences from the current study	■	☐	☐	10
	51	Practical implications	■	☐	☐	11
	52	Strengths and limitations (e.g., bias and generalizability issues)	■	☐	☐	11-12

Supplemantary Table S3. continued

Section	No.	Item	Yes	No	n/a	Page
Open Science						
Data availability	53	Sharing images along with segmentation data [n/e]	☐	■	☐	
	54	Sharing radiomic feature data	■	☐	☐	14
Code availability	55	Sharing pre-processing scripts or settings	☐	☐	■	
	56	Sharing source code for modeling	■	☐	☐	14
Model availability	57	Sharing final model files	☐	■	☐	
	58	Sharing a ready-to-use system [n/e]	☐	☐	■	

Kocak B, Baessler B, Bakas S, Cuocolo R, Fedorov A, Maier-Hein L, Mercaldo N, Müller H, Orlhac F, Pinto Dos Santos D, Stanzione A, Ugga L, Zwanenburg A. CheckList for EvaluAtion of Radiomics research (CLEAR): a step-by-step reporting guideline for authors and reviewers endorsed by ESR and EuSoMII. Insights Imaging. 2023 May 4;14(1):75. doi: 10.1186/s13244-023-01415-8