European Radiology (2021) 31:1011–1021
https://doi.org/10.1007/s00330-020-07158-0

IMAGING INFORMATICS AND ARTIFICIAL INTELLIGENCE

Shape and texture-based radiomics signature on CT effectively discriminates benign from malignant renal masses

Felix Y. Yap[1] · Bino A. Varghese[1] · Steven Y. Cen[1] · Darryl H. Hwang[1] · Xiaomeng Lei[1] · Bhushan Desai[1] · Christopher Lau[1] · Lindsay L. Yang[1] · Austin J. Fullenkamp[1] · Simin Hajian[1] · Marielena Rivas[1] · Megha Nayyar Gupta[1] · Brian D. Quinn[1] · Manju Aron[2] · Mihir M. Desai[3] · Monish Aron[3] · Assad A. Oberai[4] · Inderbir S. Gill[3] · Vinay A. Duddalwar[1,3]

Received: 1 April 2020 / Revised: 15 June 2020 / Accepted: 5 August 2020 / Published online: 15 August 2020

Abstract

Objectives Using a radiomics framework to quantitatively analyze tumor shape and texture features in three dimensions, we tested its ability to objectively and robustly distinguish between benign and malignant renal masses. We assessed the relative contributions of shape and texture metrics separately and together in the prediction model.

Materials and methods Computed tomography (CT) images of 735 patients with 539 malignant and 196 benign masses were segmented in this retrospective study. Thirty-three shape and 760 texture metrics were calculated per tumor. Tumor classification models using shape, texture, and both metrics were built using random forest and AdaBoost with tenfold cross-validation. Sensitivity analyses on five sub-cohorts with respect to the acquisition phase were conducted. Additional sensitivity analyses after multiple imputation were also conducted. Model performance was assessed using AUC.

Results Random forest classifier showed shape metrics featuring within the top 10% performing metrics regardless of phase, attaining the highest variable importance in the corticomedullary phase. Convex hull perimeter ratio is a consistently high-performing shape feature. Shape metrics alone achieved an AUC ranging 0.64–0.68 across multiple classifiers, compared with 0.67–0.75 and 0.68–0.75 achieved by texture-only and combined models, respectively.

Conclusion Shape metrics alone attain high prediction performance and high variable importance in the combined model, while being independent of the acquisition phase (unlike texture). Shape analysis therefore should not be overlooked in its potential to distinguish benign from malignant tumors, and future radiomics platforms powered by machine learning should harness both shape and texture metrics.

Key Points

- *Current radiomics research is heavily weighted towards texture analysis, but quantitative shape metrics should not be ignored in their potential to distinguish benign from malignant renal tumors.*
- *Shape metrics alone can attain high prediction performance and demonstrate high variable importance in the combined shape and texture radiomics model.*
- *Any future radiomics platform powered by machine learning should harness both shape and texture metrics, especially since tumor shape (unlike texture) is independent of the acquisition phase and more robust from the imaging variations.*

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00330-020-07158-0) contains supplementary material, which is available to authorized users.

✉ Bino A. Varghese
bino.varghese@med.usc.edu

[1] Department of Radiology, Keck School of Medicine, University of Southern California, 1441 Eastlake Ave, Los Angeles, CA 90033, USA

[2] Department of Pathology, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA

[3] Institute of Urology and the Catherine & Joseph Aresty, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA

[4] Department of Aerospace and Mechanical Engineering, Viterbi School of Engineering, University of Southern California, Los Angeles, CA, USA

Keywords Tomography, X-ray computed · Kidney neoplasms · Data analysis · Machine learning · Image Processing, radiomics

Abbreviations

CHA	Convex hull area ratio
CHP	Convex hull perimeter ratio
DICOM	Digital imaging and communications in medicine
EC	Elliptic compactness
FFT	Fast Fourier transform
GLCM	Gray-level co-occurrence matrix
GLDM	Gray-level difference matrix
HIPAA	Health Insurance Portability and Accountability Act
PACS	Picture archiving and communication system
RD	Radial distance
ROC	Receiver operating characteristic
ZC	Zero-crossing count

Introduction

Most renal tumors are incidentally diagnosed on routine imaging. Nevertheless, accurate preoperative characterization of renal masses carries a significant error rate [1, 2]. Currently, ~ 20% of renal masses turn out to be benign following resection, with the majority being < 4 cm in diameter, and identifying these patients beforehand could potentially spare them from unneeded surgery [3].

In routine clinical practice, a combination of qualitative and semi-quantitative evaluation is used to classify a renal mass as likely "benign" or "malignant." Visual assessments of tumor size, shape, texture, and enhancement are all important for determining the likelihood of cancer [4]. In contrast to enhancement being a quantifiable parameter measured in Hounsfield units on CT, tumor shape has typically been a qualitative assessment [5–9], making visual analysis of tumor contour subjective and susceptible to inter- or intra-observer interpretation variability.

In radiomics, quantitative tumor features such as size, shape, and texture are extracted from routine images. While most radiomics studies are heavily weighted towards applying texture analysis to distinguish various renal mass subtypes [7, 10–12], the literature on shape analysis is relatively sparse. Ding, et al implied that irregular morphology was an independent predictor of higher grade clear cell RCC [7], but their qualitative assessment of shape, based on two radiologists' ratings as either "round" or "non-round," highlights the inherent subjectivity in the visual analysis of tumor contour. While Shu et al have applied texture analysis as well as a small panel of quantitative shape metrics to clear cell RCC in predicting its Fuhrman grade [13], we are not aware of studies aimed at distinguishing benign from malignant renal masses by harnessing the wide array of shape and texture features long studied in the lung.

Previously developed radiomics panel in the literature demonstrated that malignant masses tend to be more lobulated and non-spheroidal and show different textures than benign masses, as well as the fact that shape and texture metrics were robust to manual segmentation [14, 15]. However, to date, the relative contributions of tumor shape versus texture to malignant behavior are still unclear.

In this study, using a larger cohort compared with 150 patients from prior studies, we assessed the potential in using both classes of metrics to noninvasively differentiate benign from malignant renal masses. We also studied the relative contributions of shape versus texture of a tumor on standard-of-care imaging to its malignant potential and the improvement in classification when combining the two, and we evaluated the necessity and contribution of shape metrics to the prediction model.

Patients and methods

This retrospective study complied with the Health Insurance Portability and Accountability Act (HIPAA), and the institutional review board granted approval with waiver of consent for inclusion.

Patients and tumors

Our patient population includes renal masses diagnosed on abdominal CT scans with pathologic diagnoses confirmed after resection at our institution. Patients were identified by retrospective query of a prospectively maintained surgical database of 1178 consecutive radical or partial nephrectomies between May 2007 and September 2018. Pathologic evaluation was performed by specialized genitourinary pathologists.

One hundred twelve patients were excluded due to the absence of evaluable preoperative imaging within a year before the nephrectomy. We excluded 49 patients with tumors arising outside the renal parenchyma, such as retroperitoneal liposarcoma, perirenal cyst, adrenal pseudocyst, and urothelial carcinoma. We also excluded 74 cases that lacked contrast-enhanced CT images, since tumor margins are difficult to accurately segment on non-contrast images. Two hundred eight cases were unable to be processed, largely due to digital imaging and communications in medicine (DICOM) incompatibilities from outside institutions that precluded processing in our radiomics pipeline.

Our final cohort contained 539 malignant and 196 benign masses. In a previously published study [14] evaluating the feasibility of shape analysis to differentiate between benign and malignant renal masses, we reported on 150 patients with renal masses that are also included here. This current study entails a larger patient cohort to include texture analysis and more renal mass subtypes, especially benign ones, to assess the integrated radiomics platform's role in renal mass evaluation.

Image acquisition

Preoperative CTs were obtained at our institution in 308/735 patients (42%), where a 64-detector row helical CT scanner (Brilliance, Philips Healthcare) was used to acquire images during patient breath-holding with these parameters: 120 kVp, variable tube current, slice thickness of 0.5 mm with reconstruction interval of 2 mm (Fig. 1). In total, 100–150 mL of nonionic intravenous contrast material (ISOVUE® 350; Bracco Imaging) dosed to weight was administered with a power injector at a rate of 4–5 mL/s.

The remaining patients had preoperative imaging performed at outside institutions prior to referral to our institution for surgical resection, but their CT examinations were uploaded onto our picture archiving and communication system (PACS) and thus available to be segmented and included in our cohort for processing in our radiomics pipeline.

Tumor segmentation

Using Synapse 3D software (Fujifilm), two senior radiologists-in-training (B.Q., M.N.G.) manually segmented renal tumors as three-dimensional regions of interest being blinded to pathologic diagnoses. Segmentation times varied from 20 to 40 min per case. Segmentations were then verified for accuracy by two radiologists (F.Y., V.D.) with 5 and 20 years of experience in abdominal imaging. Technical details beyond the scope of this journal regarding segmentation and data processing are elaborated further in Section A of Supplemental Materials. In general, the nephrographic phase provided the best delineation of the tumor and hence was used as the reference target for subsequent coregistration of other phases. If the nephrographic phase was not available for that case, the corticomedullary or excretory phase became the reference target instead. Images were coregistered by using the normalized mutual information cost function implemented in the Statistical Parametric Mapping software package (Wellcome Centre for Human Neuroimaging). Tessellated 3-D models of the tumor were created from segmented voxels using custom MATLAB (MathWorks) code.

Tumor shape and texture analyses

Shape analysis utilizes metrics to characterize the morphology, whereas texture analysis studies the variation of pixel intensity and their interrelationships. Shape

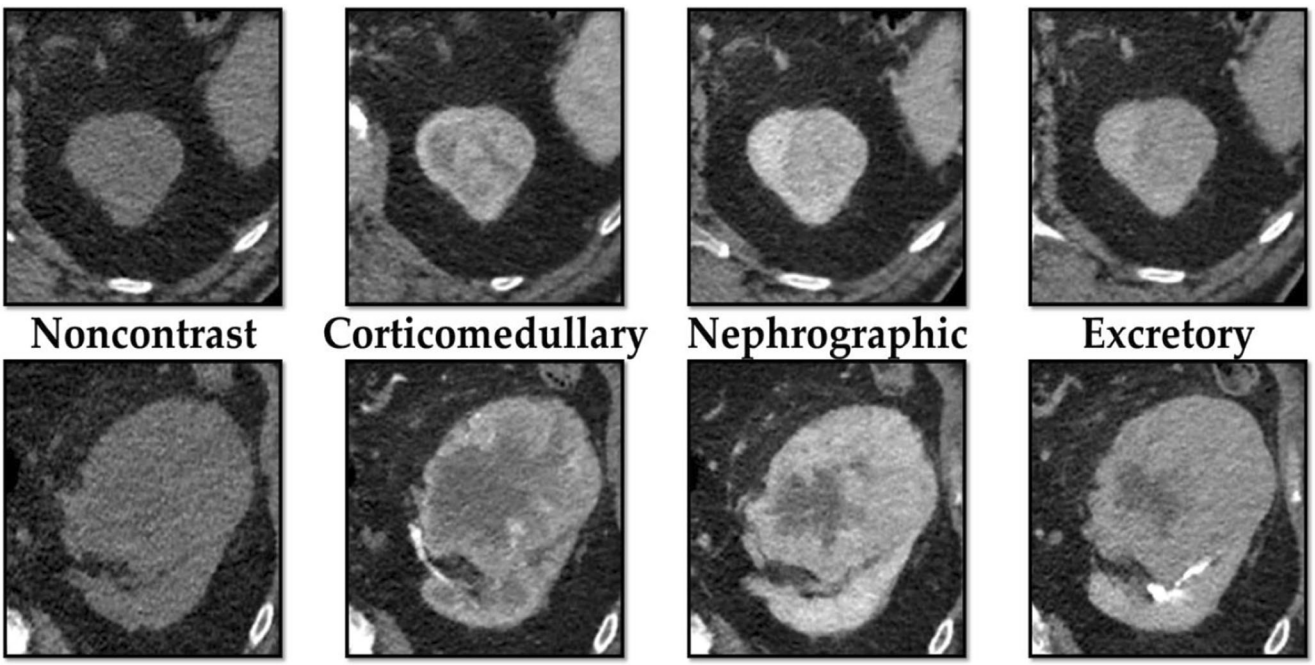

Fig. 1 Multiphase axial CT images show a 77-year-old male with a 3.7-cm left renal mass that proved to be a chromophobe renal cell carcinoma, stage pT3a (top); and a 69-year-old male with an 8.0-cm left renal mass that proved to be an oncocytoma (bottom). Non-contrast, corticomedullary (30 s), nephrographic (90 s), and excretory (5–7 min) phase images of the abdomen were obtained

analysis was performed for the whole tumor volume in the axial, coronal, and sagittal projections. Two-dimensional texture analysis was conducted in the orientation providing the largest tumor area in each phase in the axial, coronal, or sagittal projection. Three-dimensional texture analysis was conducted on the whole tumor volume. These techniques have been described in the literature [14–16] and are summarized in Fig. 2 and Fig. 3. More detailed descriptions and equations for each of the 33 shape and 760 texture features can be found in Sections B and C in Supplemental Materials and in Supplemental Figures 1–10.

Reliability assessment

We conducted a reliability analysis with 3 radiologists. Each radiologist segmented the margins independently for 15 subjects. Intraclass correlation (ICC) 2-way-mixed with absolute agreement was used to evaluate reliability.

Machine learning prediction rule development and statistical analysis

Random forest was used as the primary method for classifier development. We used average square error plots to select optimal numbers of decision trees and variables to try for each tree-building and leaf size. For bootstrapping at each tree, 60% of the original observations were used. Tenfold cross-validation was used to obtain robust classification performance. We used AUC to assess robust discrimination power based on predicted probability from each fold of testing data. The out-of-bag Gini index was used to rank the variable of importance. The gain in discriminatory power was assessed by comparing the full model (combined texture and shape classifier) vs. reduced model (e.g., texture only). *Z* test was used to compare AUCs.

For sensitivity analysis, 5 sub-cohorts of cases with respect to the acquisition phase were created. Sub-cohort I included cases that contain all four individual phases (non-contrast, corticomedullary, nephrographic, and excretory). Sub-cohorts II to IV include cases that contain a specific individual

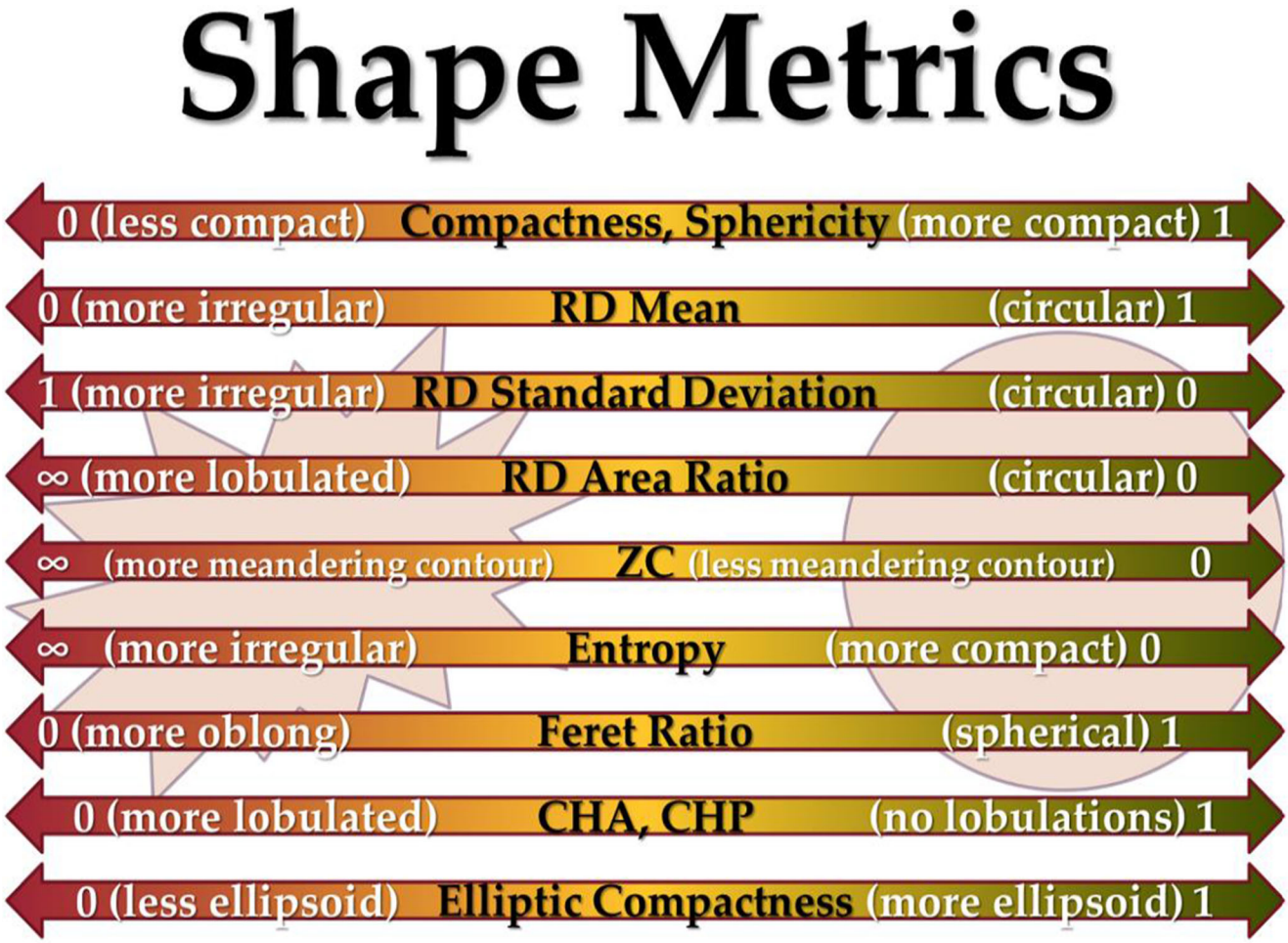

Fig. 2 Shape analysis using 33 shape metrics. Each shape feature focuses on certain characteristics of tumor morphology and factors into a quantitative calculation. RD, radial distance; ZC, zero-crossing count; CHA, convex hull area ratio; CHP, convex hull perimeter ratio

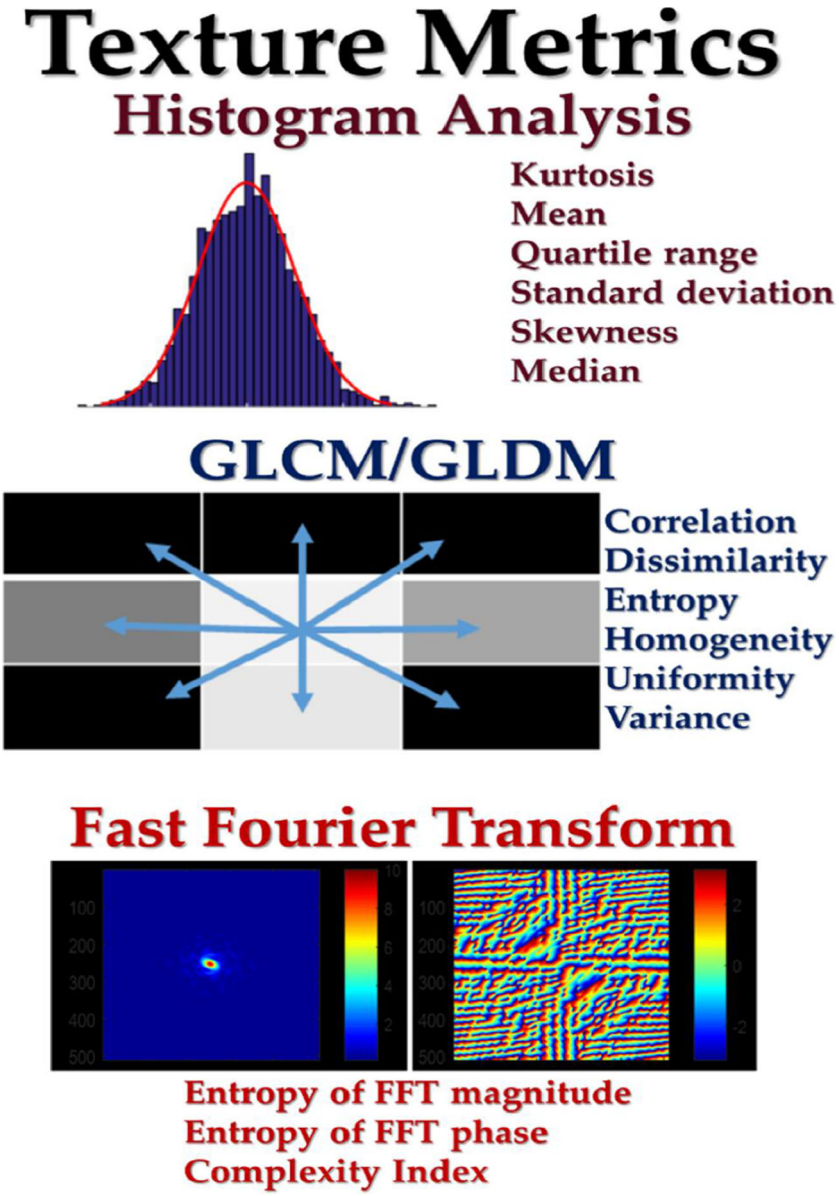

Fig. 3 Texture analysis using 760 texture metrics. Each texture feature involves the study of the variation of pixel image intensity. Different classes of texture metrics entail gray-level histogram analysis, gray-level co-occurrence matrix (GLCM) analysis, gray-level difference matrix (GLDM) analysis, and frequency analysis based on fast Fourier transform (FFT)

phase. Random forest classifiers were built and validated using these 5 sub-cohorts respectively. We repeated the above procedures using AdaBoost as well.

Many machine learning methods eliminate missing data such as data with missing phases within multiphase CT data. To apply those methods to our data with a high missing rate, we imputed missing data using the Markov Chain Monte Carlo (MCMC). This approach also helped compare the MCMC methodology based on built-in procedures of random forest and AdaBoost in dealing with data missingness using surrogate data. The average score across 10 imputed datasets was used for the final imputed data. Using the completed data after imputation, we developed and validated more classifiers: ElasticNet, random forest, AdaBoost, MARS, and NeuroNet. Model performance was validated through 1/3 independent testing data. To further test model robustness, we rebuilt the model using data from one institute and validated through other institutes. Supplemental Figure 14 illustrates all machine learning/validation approaches. SAS 9.4 was used for all data analyses.

Results

Patients, tumors, and scans

Five hundred thirty-nine of 735 (73%) patients with malignant renal tumors and 196/735 (27%) patients with benign renal tumors were included in our final patient cohort (Table 1).

Table 1 Distribution of tumor types within contrast-enhanced CT data

Tumor type	*n*	%	Tumor grade				Pathological stage			
			1	2	3	4	T1	T2	T3	T4
Malignant	539									
Clear cell RCC	407	75	27	241	124	12	306	24	74	3
Papillary RCC	73	14	4	42	20	2	55	8	10	0
Chromophobe RCC	42	8	N/A	N/A	N/A	N/A	25	6	10	0
Clear cell papillary RCC	10		6	3	1	0	10	0	0	0
Sarcomatoid RCC	4		0	0	0	3	0	1	2	0
Unclassified RCC	2		0	1	1	0	1	1	0	0
Collecting duct carcinoma	1		0	0	0	1	0	0	1	0
Benign	196		N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Oncocytomas	104	53	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Lipid-poor angiomyolipoma	59	30	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Cysts	13		N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Cystic nephromas	6		N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Metanephric adenomas	4		N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Pseudocysts	3		N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Dense fibrosis	2		N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Vascular malformations	2		N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Fibroma	1		N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Mixed epithelial and stromal tumor	1		N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Renal tubular hyperplasia	1		N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Note: chromophobe RCC although malignant is not sub-classified by Fuhrman grade. *RCC*, renal cell carcinoma; *N/A*, not applicable

Among clear cell RCCs, 135/401 (34%) were of high-grade (ISUP grades 3–4) and 77/401 (19%) were stage T3 or higher. Among stage T1 RCCs, 290/397 (73%) were T1a and 107/397 (27%) were T1b.

Table 2 tabulates the numbers of cases in each of the 5 sub-cohorts (I–V) with respect to the acquisition phases.

Table 2 Distribution of various acquisition phases within imaging studies

Sub-cohort	Cases that include:	*N* (total cases, 735)
I	All 4 phases	453 (62%)
II	Non-contrast phase	619 (84%)
III	Corticomedullary phase	520 (71%)
IV	Nephrographic phase	695 (95%)
V	Excretory phase	596 (81%)

453/735 (62%) cases in sub-cohort I had all 4 phases (non-contrast, corticomedullary, nephrographic, and excretory). The nephrographic phase was the most common contrast-enhanced phase, encompassing 695/735 (95%) cases in sub-cohort IV, followed by the excretory (596/735; 81%) and corticomedullary (520/735; 71%) phases in sub-cohorts V and III respectively. 619/735 (84%) patients had a non-contrast phase in sub-cohort II

In total, 495/735 (67%) patients were male. The mean age was 60.8 (range 17–93).

Reliability assessment

In the segmentation reliability assessment between 3 radiologists, 65% of features met ICC > 0.8. Thus, we conducted a sensitivity analysis with random forest using these robust features only. Supplemental Table 2 shows that using robust features only reached similar performance as using all features. Confusion matrices associated with the models presented in Table 2 have been tabulated in the Supplement as well (Supplement Table 4 and 5).

Top-performing metrics

The top 79/793 (10%) performing metrics within a given acquisition phase in differentiating renal masses are tabulated in the Supplemental Table.

Of the various metrics, the convex hull perimeter ratio (CHP) ranked consistently as a high-performing shape feature across all four phases, followed by elliptic compactness (EC). Of all four phases, shape metrics featured most prominently in the corticomedullary phase (sub-cohort III), with CHP and EC

Table 3 Discrimination power (AUC) of shape-only, texture-only, and combined shape and texture models in differentiating benign from malignant renal masses

Model	Sub-cohort I: all 4 phases	Sub-cohort II: non-contrast phase	Sub-cohort III: corticomedullary phase	Sub-cohort IV: nephrographic phase	Sub-cohort V: excretory phase
1 = shape only	0.64	0.65†	0.68†	0.64	0.66
2 = texture only	0.75 †	0.70†	0.70†	0.67	0.69
3 = combined	0.75	0.71	0.73	0.68	0.70

Values are mean gain or loss in AUC with 95% CI in parentheses. Model 1 = shape-only, 2 = texture-only, 3 = combined shape and texture

† Significant gain ($p \leq 0.05$)

both featuring as highly ranked within the top 10%. This has been repeated 10 and 6 times respectively during the tenfold cross-validation. In 4-phase sub-cohort I, CHP ranked among the top 67 metrics, appearing 3 times during the tenfold cross-validation. CHP also appeared 9 times in the non-contrast phase, 3 times in the nephrographic phase, and 7 times in the excretory phase during the tenfold cross-validation.

Comparison between shape-only, texture-only, and combined models

Table 3 is a tabulation of the segregation of various radiomics models obtained using shape only, texture only, and combined shape and texture metrics. In sub-cohort I entailing 4-phase studies ($n = 453$), an AUC of 0.64 in the independent testing subset was achieved by 33 shape metrics alone, whereas an AUC of 0.75 was achieved using 760 texture metrics (Fig. 4). Sensitivity analyses conducted in different individual phases with complete data also demonstrated similar results (Fig. 5 and Supplemental Figures 11–13), although the gaps between the AUCs for the isolated models were narrower. Shape-only models also attained comparable performance in the nephrographic and excretory phase sub-cohorts IV and V (AUCs 0.64 and 0.66 respectively, in comparison with 0.67 and 0.69 for texture-only models).

The texture-only model's performance slightly increased from 0.67–0.70 in the individual phase sub-cohorts II–V to 0.75 when all 4 phases are analyzed in sub-cohort I, whereas the shape-only model's performance was consistently in a similar range (0.64–0.68) regardless of whether all 4 phases or only individual phases were considered. This result is not surprising, given that the prediction rule from texture analysis is expected to improve in performance as additional data from multiple phases are included in the learning phase, whereas shape analysis is independent of the acquisition phase and its performance thus would not vary with phase(s).

Table 4 shows the gain in AUC by adding texture analysis to the shape-only model and vice versa. For sub-cohort I with all 4 phases, adding shape analysis to the combined model did not improve discrimination over the texture-only model (0.75 vs. 0.75, $p = 0.77$). However, within the corticomedullary phase sub-cohort III, even though texture metrics alone attained an AUC of 0.70, adding shape analysis to this sub-cohort significantly increased the AUC to 0.73 in the combined model ($p < 0.01$).

As expected, we saw comparable performance between AdaBoost and random forest, as they work well with missing data. Performance did not deviate significantly using random forest with the 4-phase complete data. Supplemental Table 3 tabulates the performance across different classifiers and testing scenarios (1/3 validation data and cross-institution validation).

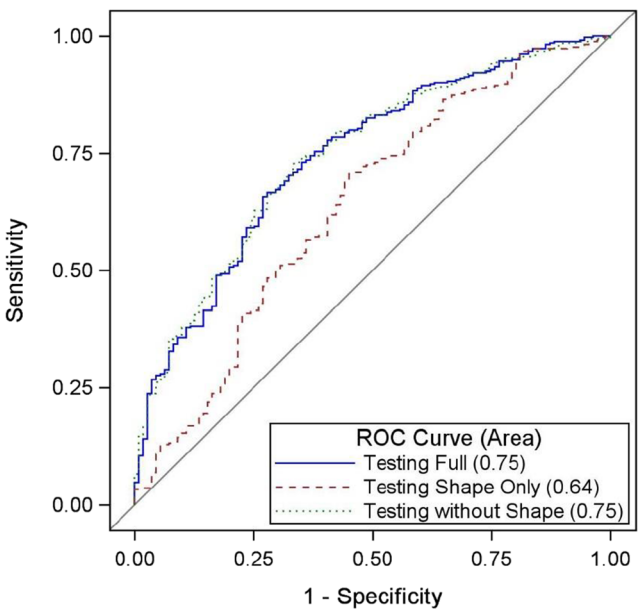

Fig. 4 Receiver operating characteristic (ROC) curves for the shape-only (dashed red), texture-only (dotted green), and combined (solid blue) radiomics models in the discrimination of benign and malignant renal masses using imaging data with all four phases available (sub-cohort I). AUC values are shown in the lower right corner

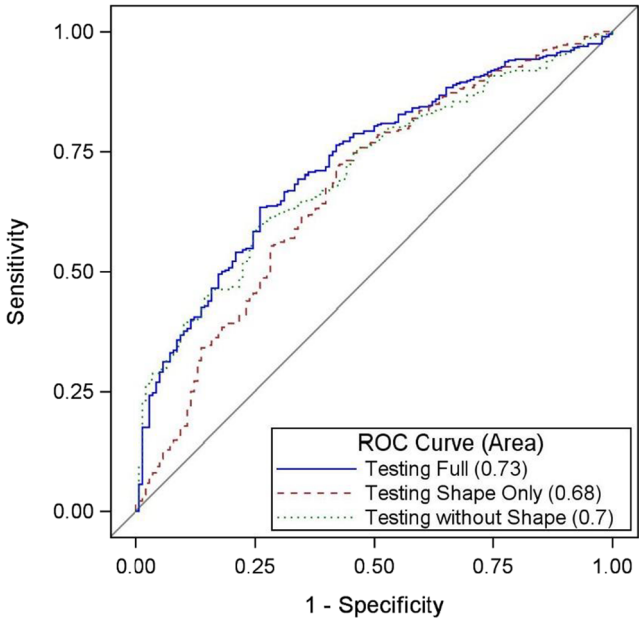

Fig. 5 Receiver operating characteristic (ROC) curves for the shape-only (dashed red), texture-only (dotted green), and combined (solid blue) radiomics models in the discrimination of benign and malignant renal masses using imaging data with a corticomedullary phase available (sub-cohort III). AUC values are shown in the lower right corner

Discussion

While many radiomics relationships have been explored, how much a renal tumor's shape correlates to its malignant behavior in comparison to texture is still unclear. Using a previously validated quantitative panel of metrics, we built different radiomics models based on 33 shape and 760 texture features from random forest classifiers. Our results show that for the task of distinguishing benign from malignant renal masses, shape metrics alone attain a reasonably high prediction performance and hold high variable importance in the combined radiomics model, while being independent of the acquisition phase (unlike texture).

Of all shape metrics, the convex hull perimeter ratio (CHP) was a consistently high-performing one regardless of phase, along with elliptic compactness (EC). This is concordant with our prior results identifying these two features as statistically significant between benign and malignant renal masses [14]. The CHP metric may be analogous to the fractional concavity feature described in the image biomarker standardization initiative [17], which was shown by Limkin et al [18] to be a reliable shape feature least affected by slice thickness or volume changes. As for EC (also termed anfractuosity or elliptic-normalized circumference), it appears that EC most resembles the volume/area density – minimum volume enclosing ellipsoid feature in the image biomarker standardization initiative [17].

While there is no general consensus which individual shape or texture feature is stable or robust in a systematic review [19], shape metrics have been shown to be more robust than texture metrics with respect to different imaging parameters [20] and respiratory motion patterns [21] in the setting of lung tumors. Given that texture features are less reliable than shape, and that tumor shape itself is independent of phase acquisition and hence more stable, we advocate for the inclusion of shape analysis alongside texture on future radiomics investigations and platforms, even though shape did not attain greater accuracies than texture in our study. When using combined features including both texture and shape, several shape features attained high-rank positions. We should respect the correlation between texture and shape especially with high-dimensional data, as this could result in competitive performance between the combined and texture-only models. However, when encountering missing or poor-quality data in real life, the shape can at least serve as a surrogate marker when texture feature is missing, of poor quality, or unstable between different scanners.

Using images of 118 RCC and 45 lipid-poor AMLs, Yang et al [22] achieved an AUC of 0.88 using a combination of radiomics and machine learning. Their high performance of random forest over other classifiers when considering all 4 phases is concordant with ours. Their higher AUC may reflect their application of Conditional Infomax Feature Extraction (CIFE) for feature selection. Foregoing a feature selection step risks including too many noisy, redundant, and irrelevant features, which may jeopardize

Table 4 Differences in benign-malignant discrimination power (AUC) between shape-only, texture-only, and combined shape and texture models

Comparison	Sub-cohort I: all 4 phases AUC	Sub-cohort II: non-contrast phase AUC	Sub-cohort III: corticomedullary phase AUC	Sub-cohort IV: nephrographic phase AUC	Sub-cohort V: excretory phase AUC
Texture contribution: gain of model 3 over model 1	0.11 (0.05 to 0.17, $p < 0.01$)†	0.06 (0.02 to 0.11, $p = 0.01$)†	0.05 (0 to 0.10, $p = 0.04$)†	0.04 (− 0.01 to 0.10, $p = 0.11$)	0.04 (− 0.01 to 0.10, $p = 0.12$)
Shape contribution: gain of model 3 over model 2	0.00 (− 0.01 to 0.01, $p = 0.77$)	0.01 (0 to 0.02, $p = 0.05$)†	0.03 (0.01 to 0.04, $p < 0.01$)†	0.01 (0 to 0.02, $p = 0.09$)	0.01 (0 to 0.02, $p = 0.06$)

Values are mean gain in AUC with 95% CI and p value in parentheses. Model 1 = shape-only, 2 = texture-only, 3 = combined shape and texture

† Significant gain ($p \leq 0.05$)

learning performance. However, a disadvantage is that if too few features are selected, some relevant features may be eliminated. Given our large sample size, we omitted the feature selection step and reported the performance based on all features. Kocak et al [23] using 47 cases showed that segmentation-based differences changed radiomics performance. Their contour-specific segmentation ($0.85 < AUC < 0.98$) outperformed their margin shrinkage method ($0.75 < AUC < 0.8$). Our study followed contour-specific segmentation and results from an inter-operator bias study on segmentation's effect on radiomics were comparable [14]. To evaluate texture analysis's role in predicting Fuhrman nuclear grade, Bektas et al [24] reported an AUC of 0.86 with tenfold cross-validation using a support vector machine (SVM). However, it was a single-institution study and does not account for shape metrics. Feng et al [25], also a single-institution study with 58 patients, showed that SVM classifier in combination with synthetic minority oversampling technique (SMOTE) and recursive feature elimination (RFE) achieved an AUC of 0.95 with fivefold cross-validation in discriminating between small AMLs without visible fat and RCCs. However, this study did not consider shape metrics.

Considering these studies, our AUC of 0.75 is comparable using multi-institutional data, a larger sample size, and a more comprehensive radiomics panel of both texture and shape in discriminating benign from malignant tumors. However, this prompts two questions: (1) whether radiomics analysis alone will generate a paradigm shift in renal mass evaluation, especially with a large cohort from multiple sites and differing protocols; (2) whether different tumor size categories need different algorithms. Our stratified sensitivity analysis based on tumor size (< 4 cm and > 4 cm) showed similar performance and variable of importance compared with the full cohort.

When presented with the challenge of high data dimensionality, machine learning such as a random forest is a prime method for dimensional reduction [26, 27]. Using other dimension reduction methods prior to random forest could be problematic due to data loss (e.g., feature filtered prior to learning) and poor stability [28] (e.g., a principal component derived from learning sample may not fit well in the testing sample). From a radiomics analysis standpoint, while we run the risk of overfitting, our model will assess all features [29]. In addition, given that one of radiomics' strengths is the numerous features that are computable from radiologic images, discarding features using statistical criteria without ascertaining their roles in the clinical question poses a greater risk for losing valuable information.

While percutaneous renal mass biopsies are accurate in 90% for diagnosis, it is an additional invasive procedure necessitating a small risk, expense, and time. Biopsies are also notoriously inaccurate for grading [30]. We plan to use this strategy to evaluate the grading of individual tumors as part of our future work.

There are a few limitations to this study. First, although we previously demonstrated that shape and texture metrics are robust to manual extraction without significant interobserver variability [14, 15], differences in imaging techniques may influence tumor segmentation as well as shape and texture analyses [19]. Second, while a large number of patients were scanned at our institution with the same four-phase imaging protocol, other patients in our study had imaging performed elsewhere with different imaging parameters and protocols prior to referral to our institution for resection, which introduces some variability in our quantitative metrics as discussed earlier. We adopted this methodology to closely reflect the real-life, standard-of-care heterogeneity of a patient population encountered in a tertiary care center. Third, randomly sampling a tertiary center population containing a greater proportion of malignant cases may cause extreme skewness and affect model performance. To achieve a more balanced distribution, instead of random sampling, we included all 196 (27%) benign cases. Fourth, in the absence of a reliable or widely accepted automated segmentation technique, tumors were manually segmented. In the future when automated segmentation is more robust, we will evaluate this cohort again, possibly with the use of deep learning methods.

Conclusion

Our results demonstrate that shape metrics alone, especially convex hull perimeter ratio and elliptic compactness, can attain a similar discriminatory power as texture metrics, signifying that shape analysis should not be overlooked in a future radiomics platform powered by machine learning. A framework as such should therefore utilize both shape and texture together rather than in isolation from each other.

Acknowledgments This study received funding from the Radiological Society of North America (RSNA) Research Fellow Grant (#RF1821) titled "The Shapely Renal Mass: Quantitative Contour Evaluation of Renal Cell Carcinoma."

Funding information Funding supported by the RSNA Research Fellow Grant #RF1821 titled "The Shapely Renal Mass: Quantitative Contour Evaluation of Renal Cell Carcinoma."

Compliance with ethical standards

Guarantor The scientific guarantor of this publication is Vinay A. Duddalwar, MD, FRCR.

Conflict of interest The authors of this manuscript declare relationships with the following companies:

Vinay Duddalwar, MD, FRCR, is a consultant for Intuitive Surgical and Radmetrix and sits on the advisory board of DeepTek.

Statistics and biometry Steven Y. Cen (biostatistician), one of the authors, kindly provided statistical support for this manuscript.

Informed consent Written informed consent was obtained from all subjects (patients) in this study.

Ethical approval Institutional Review Board approval was obtained.

Study subjects or cohorts overlap Some study subjects or cohorts have been previously reported in *Yap FY, Hwang DH, Cen SY, Varghese BA, Desai B, Quinn BD,* et al *Quantitative Contour Analysis as an Image-based Discriminator Between Benign and Malignant Renal Tumors. Urology 2018;**114**:121–7.* https://doi.org/10.1016/j.urology.2017.12.018. However, significantly larger sample size and more comprehensive analysis have been performed.

Methodology
- retrospective
- diagnostic study
- performed at one institution (multicenter data)

References

1. Shin T, Duddalwar VA, Ukimura O et al (2017) Does computed tomography still have limitations to distinguish benign from malignant renal tumors for radiologists? Urol Int 99:229–236. https://doi.org/10.1159/000460303
2. Choudhary S, Rajesh A, Mayer NJ, Mulcahy KA, Haroon A (2009) Renal oncocytoma: CT features cannot reliably distinguish oncocytoma from other renal neoplasms. Clin Radiol 64:517–522. https://doi.org/10.1016/j.crad.2008.12.011
3. Blute ML, Drewry A, Abel EJ (2015) Percutaneous biopsy for risk stratification of renal masses. Ther Adv Urol 7:265–274. https://doi.org/10.1177/1756287215585273
4. Gill IS, Aron M, Gervais DA, Jewett MAS (2010) Clinical practice. Small renal mass. N Engl J Med 362:624–634. https://doi.org/10.1056/NEJMcp0910041
5. Zhu YH, Wang X, Zhang J, Chen YH, Kong W, Huang YR (2014) Low enhancement on multiphase contrast-enhanced CT images: an independent predictor of the presence of high tumor grade of clear cell renal cell carcinoma. AJR Am J Roentgenol 203:W295–W300. https://doi.org/10.2214/AJR.13.12297
6. Davarpanah AH, Spektor M, Mathur M, Israel GM (2016) Homogeneous T1 hyperintense renal lesions with smooth borders: is contrast-enhanced MR imaging needed? Radiology 280:128–136. https://doi.org/10.1148/radiol.16151240
7. Ding J, Xing Z, Jiang Z et al (2018) CT-based radiomic model predicts high grade of clear cell renal cell carcinoma. Eur J Radiol 103:51–56. https://doi.org/10.1016/j.ejrad.2018.04.013
8. Lee-Felker SA, Felker ER, Tan N et al (2014) Qualitative and quantitative MDCT features for differentiating clear cell renal cell carcinoma from other solid renal cortical masses. AJR Am J Roentgenol 203:W516–W524. https://doi.org/10.2214/AJR.14.12460
9. Patel NS, Poder L, Wang ZJ et al (2009) The characterization of small hypoattenuating renal masses on contrast-enhanced CT. Clin Imaging 33:295–300. https://doi.org/10.1016/j.clinimag.2008.12.002
10. Coy H, Hsieh K, Wu W et al (2019) Deep learning and radiomics: the utility of Google TensorFlow™ inception in classifying clear cell renal cell carcinoma and oncocytoma on multiphasic CT. Abdom Radiol (NY) 44:2009–2020. https://doi.org/10.1007/s00261-019-01929-0
11. Li Z-C, Zhai G, Zhang J et al (2019) Differentiation of clear cell and non-clear cell renal cell carcinomas by all-relevant radiomics features from multiphase CT: a VHL mutation perspective. Eur Radiol 29:3996–4007. https://doi.org/10.1007/s00330-018-5872-6
12. He X, Zhang H, Zhang T et al (2019) Predictive models composed by radiomic features extracted from multi-detector computed tomography images for predicting low- and high- grade clear cell renal cell carcinoma: a STARD-compliant article. Medicine (Baltimore) 98:e13957. https://doi.org/10.1097/MD.0000000000013957
13. Shu J, Tang Y, Cui J et al (2018) Clear cell renal cell carcinoma: CT-based radiomics features for the prediction of Fuhrman grade. Eur J Radiol 109:8–12. https://doi.org/10.1016/j.ejrad.2018.10.005
14. Yap FY, Hwang DH, Cen SY et al (2018) Quantitative contour analysis as an image-based discriminator between benign and malignant renal tumors. Urology 114:121–127. https://doi.org/10.1016/j.urology.2017.12.018
15. Varghese BA, Chen F, Hwang DH et al (2018) Differentiation of predominantly solid enhancing lipid-poor renal cell masses by use of contrast-enhanced CT: evaluating the role of texture in tumor subtyping. AJR Am J Roentgenol 211:W288–W296. https://doi.org/10.2214/AJR.18.19551
16. Haralick RM, Shanmugam KS, Dinstein I (1973) Textural features for image classification. IEEE Trans Syst Man Cybern 6:610–621. https://doi.org/10.1109/TSMC.1973.4309314
17. Zwanenburg A, Leger S, Vallières M, et al (2016) Image biomarker standardisation initiative. arXiv preprint arXiv:1612.07003. [cs]
18. Limkin EJ, Reuzé S, Carré A et al (2019) The complexity of tumor shape, spiculatedness, correlates with tumor radiomic shape features. Sci Rep 9(1):1–12. https://doi.org/10.1038/s41598-019-40437-5
19. Traverso A, Wee L, Dekker A, Gillies R (2018) Repeatability and reproducibility of radiomic features: a systematic review. Int J Radiat Oncol Biol Phys 102(4):1143–1158. https://doi.org/10.1016/j.ijrobp.2018.05.053
20. Zhao B, Tan Y, Tsai W-Y et al (2016) Reproducibility of radiomics for deciphering tumor phenotype with imaging. Sci Rep 6(1):1–7. https://doi.org/10.1038/srep23428
21. Du Q, Baine M, Bavitz K et al (2019) Radiomic feature stability across 4D respiratory phases and its impact on lung tumor prognosis prediction. PLoS One 14(5). https://doi.org/10.1371/journal.pone.0216480
22. Yang R, Wu J, Sun L et al (2020) 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 30:1254–1263. https://doi.org/10.1007/s00330-019-06384-5
23. Kocak B, Ates E, Durmaz ES, Ulusan MB, Kilickesmez O (2019) Influence of segmentation margin on machine learning-based high-dimensional quantitative CT texture analysis: a reproducibility study on renal clear cell carcinomas. Eur Radiol 29:4765–4775. https://doi.org/10.1007/s00330-019-6003-8
24. Bektas CT, Kocak B, Yardimci AH et al (2019) Clear cell renal cell carcinoma: machine learning-based quantitative computed tomography texture analysis for prediction of Fuhrman nuclear grade. Eur Radiol 29:1153–1163. https://doi.org/10.1007/s00330-018-5698-2
25. Feng Z, Rong P, Cao P et al (2018) 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 28:1625–1633. https://doi.org/10.1007/s00330-017-5118-z

26. Díaz-Uriarte R, Alvarez de Andrés S (2006) Gene selection and classification of microarray data using random forest. BMC Bioinformatics 7(1):3. https://doi.org/10.1186/1471-2105-7-3
27. Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res:1157–1182
28. Sotiras A, Gaonkar B, Eavani H et al (2016) Chapter 10 - Machine learning as a means toward precision diagnostics and prognostics. In: Wu G, Shen D, Sabuncu MR (eds) Machine Learning and Medical Imaging. Academic Press, pp 299–334
29. Pattern Classification, 2nd Edition. In: Wiley.com. https://www.wiley.com/en-us/Pattern+Classification%2C+2nd+Edition-p-9780471056690. Accessed 3 Dec 2018
30. Bhindi B, Thompson RH, Lohse CM et al (2018) The probability of aggressive versus indolent histology based on renal tumor size: implications for surveillance and treatment. Eur Urol 74:489–497. https://doi.org/10.1016/j.eururo.2018.06.003