European Radiology (2019) 29:4765–4775
https://doi.org/10.1007/s00330-019-6003-8

IMAGING INFORMATICS AND ARTIFICIAL INTELLIGENCE

Influence of segmentation margin on machine learning–based high-dimensional quantitative CT texture analysis: a reproducibility study on renal clear cell carcinomas

Burak Kocak[1] · Ece Ates[1] · Emine Sebnem Durmaz[2] · Melis Baykara Ulusan[1] · Ozgur Kilickesmez[1]

Received: 13 October 2018 / Revised: 19 December 2018 / Accepted: 11 January 2019 / Published online: 12 February 2019

Abstract

Objective To determine the possible influence of segmentation margin on each step (feature reproducibility, selection, and classification) of the machine learning (ML)-based high-dimensional quantitative computed tomography (CT) texture analysis (qCT-TA) of renal clear cell carcinomas (RcCCs).

Materials and methods For this retrospective study, 47 patients with RcCC were included from a public database. Two segmentations were obtained by two radiologists for each tumour: (*i*) contour-focused and (*ii*) margin shrinkage of 2 mm. Texture features were extracted from original, filtered, and transformed CT images. Feature selection was done using a correlation-based algorithm. The ML classifier was *k*-nearest neighbours. Classifications were performed with and without using *synthetic minority over-sampling technique*. Reference standard was nuclear grade (low versus high). Intraclass correlation coefficient (ICC), Pearson's correlation coefficient, Wilcoxon signed-ranks test, and McNemar's test were used in the analysis.

Results The segmentation with margin shrinkage of 2 mm (772 of 828; 93.2%) yielded more texture features with excellent reproducibility (ICC $\geq$ 0.9) than contour-focused segmentation (714 of 828; 86.2%), $p < 0.0001$. The feature selection algorithms resulted in different feature subsets for two segmentation datasets with only one common feature. All ML-based models based on contour-focused segmentation (area under the curve [AUC] range, 0.865–0.984) performed better than those with margin shrinkage of 2 mm (AUC range, 0.745–0.887), $p < 0.05$.

Conclusions Each step of the ML-based high-dimensional qCT-TA was susceptible to a slight change of 2 mm in segmentation margin. Despite yielding fewer features with excellent reproducibility, use of the contour-focused segmentation provided better classification performance for distinguishing nuclear grade.

Key Points

- *Each step of a machine learning (ML)-based high-dimensional quantitative computed tomography texture analysis (qCT-TA) is sensitive to even a slight change of 2 mm in segmentation margin.*
- *Despite yielding fewer texture features with excellent reproducibility, performing the segmentation focusing on the outermost boundary of the tumours provides better classification performance in ML-based qCT-TA of renal clear cell carcinomas for distinguishing nuclear grade.*
- *Findings of an ML-based high-dimensional qCT-TA may not be reproducible in clinical practice even using the same feature selection algorithm and ML classifier unless the possible influence of the segmentation margin is considered.*

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00330-019-6003-8) contains supplementary material, which is available to authorized users.

✉ Burak Kocak
drburakkocak@gmail.com

[1] Department of Radiology, Istanbul Training and Research Hospital, Istanbul, Turkey

[2] Department of Radiology, Cerrahpasa Medical Faculty, Istanbul University-Cerrahpasa, Istanbul, Turkey

Keywords Multidetector computed tomography · Radiomics · Clear cell renal cell carcinoma · Machine learning · Artificial intelligence

Abbreviations

AUC	Area under the curve
CE-CT	Contrast-enhanced computed tomography
CT	Computed tomography
ICC	Intraclass correlation coefficient
k-NN	*k*-nearest neighbours
LoG	Laplacian of Gaussian
NN	Nearest neighbours
qCT-TA	Quantitative computed tomography texture analysis
qTA	Quantitative texture analysis
RCC	Renal cell carcinoma
RcCC	Renal clear cell carcinoma
TCGA-KIRC	The Cancer Genome Atlas-Kidney Renal Clear Cell Carcinoma
WEKA	Waikato environment for knowledge analysis

Introduction

Quantitative texture analysis (qTA) is an image processing method for measuring certain pixel or voxel patterns that may not be perceptible with human eye [1, 2]. It produces numerous texture feature parameters, making the analysis high-dimensional. Although the field of high-dimensional qTA is still in its infancy, the literature suggests that it can be used in predicting tumour characterisation, stage, nuclear grade, response to treatment, and overall survival [1, 3, 4]. On the other hand, recent evidence suggests that the conclusions must be treated with caution since several texture features can vary significantly with slight changes in the images and acquisition parameters [5–7]. Even with the use of same acquisition protocols, the qTA might not guarantee similar texture feature values [5]. Therefore, reproducibility of texture features has been a critical challenge for acquiring consistent results in building predictive models to be used in clinical practice.

A machine learning (ML)-based high-dimensional qTA consists of a few consecutive steps including image preprocessing, segmentation, feature extraction, feature selection, and model development. Although each stage has a potential to be a source for reproducibility problems [7–15], the segmentation is thought to be the most critical and challenging component of radiomics studies [2]. Manual segmentation by experts is a widely used method and considered as a 'gold standard', but it also suffers from inter-observer variability. Despite covering the visible tumour contour seems to be intuitive to obtain all textural information from the lesion of interest, it might be complicated with textural details of the peritumoural or non-tumoural areas, which might have an influence on the reproducibility of the texture features, selected feature subsets, and in turn classification performance. To date, much work has been done for assessment of reproducibility of the qTA with relatively few features, mainly focusing on acquisition techniques and parameters [6, 7, 16]. Nonetheless, no attention has been given to the possible influence of the differences in segmentation margin of the tumours with relatively distinct contours like most renal tumours. In addition, to our best knowledge, no study has yet examined the possible influence of segmentation margin differences on reproducibility of high-dimensional quantitative computed tomography (CT) texture analysis (qCT-TA).

The present study was designed to determine the possible influence of slight differences in segmentation margin on high-dimensional CT texture feature reproducibility, feature selection, and ML-based classifications for distinguishing low and high nuclear grade renal clear cell carcinomas (RcCCs).

Materials and methods

Study design

No ethical approval was obtained for this experimental retrospective study because all patients' data included in this study are publicly and freely available for scientific purposes.

The patients included in this study were obtained from the Cancer Genome Atlas-Kidney Renal Clear Cell Carcinoma (TCGA-KIRC) [17, 18]. The results shown here are in whole or part based upon the data generated by the TCGA Research Network: http://cancergenome.nih.gov/. The authors acknowledge that they have previously used this public database (The Cancer Genome Atlas-Kidney Renal Clear Cell Carcinoma [TCGA-KIRC]) in different context [19].

To create a uniform imaging protocol, the inclusion criteria used in this study were as follows: (*i*) patients who had undergone pretreatment corticomedullary phase contrast-enhanced CT (CE-CT); (*ii*) CE-CT with a slice thickness of 5 mm; (*iii*) CE-CT with pixel spacing less than 1 mm; (*iv*) CE-CT with no overlap between slices; and (*v*) CE-CT performed with maximum tube voltage of 140 kV. Because the contrast delay time was not available in patients' imaging data, we used a previously reported quantitative method for inclusion of corticomedullary phase images [20]. We included the imaging studies if the difference in enhancement between cortex and medulla was equal or higher than 90 Hounsfield units [20].

The exclusion criteria were as follows: (*i*) patients with images that had significant image noise or significant artefacts; and (*ii*) patients with multiple tumours unless the target tumour was certain in the database. Multiple tumours were excluded because there was only one nuclear grade reported for multiple tumours in the public database.

In total, the TCGA-KIRC database includes 537 cases. However, only 267 patients' imaging data were available in http://www.cancerimagingarchive.net/. Among these 267 patients, only 47 patients with RcCC met our inclusion/exclusion criteria. Patients' demographics are presented in Table 1. The list of the included patients with codes as appeared in TCGA-KIRC database is presented in Supplementary Table 1.

The reference standard for classifications was publicly available nuclear grade of the RcCCs in TCGA-KIRC [17, 18]. The tumours were grouped as high (nuclear grades 3 and 4) and low nuclear grade (nuclear grades 1 and 2).

Table 1 Patients' demographics and characteristics

	Values
Mean age	59.7 years
Gender	
Female	24 (51.1%)
Male	23 (48.9%)
Mean tumour size*	74.7 mm
Nuclear grade	
Low	14 (29.8%)
High	33 (70.2%)
T stage	
T1a	11 (23.4%)
T1b	8 (17%)
T2a	2 (4.3%)
T2b	1 (2.1%)
T3a	19 (40.4%)
T3b	5 (10.6%)
T3c	1 (2.1%)
N stage	
N_x	30 (63.8%)
N_0	17 (36.2%)
M stage	
M_0	36 (76.6%)
M_1	11 (23.4%)
Stage groups	
Stage 1	19 (40.4%)
Stage 2	2 (4.3%)
Stage 3	15 (31.9%)
Stage 4	11 (23.4%)

T/N/M stages and stage groups refer to pathological evaluation

*maximum three-dimensional tumour diameter

To provide a better understanding to the readers, the flowchart in Fig. 1 summarises general methodological pipeline and the most critical technical steps used in this study.

Image processing

Details on the image processing are given in Online Supplement Part E1.

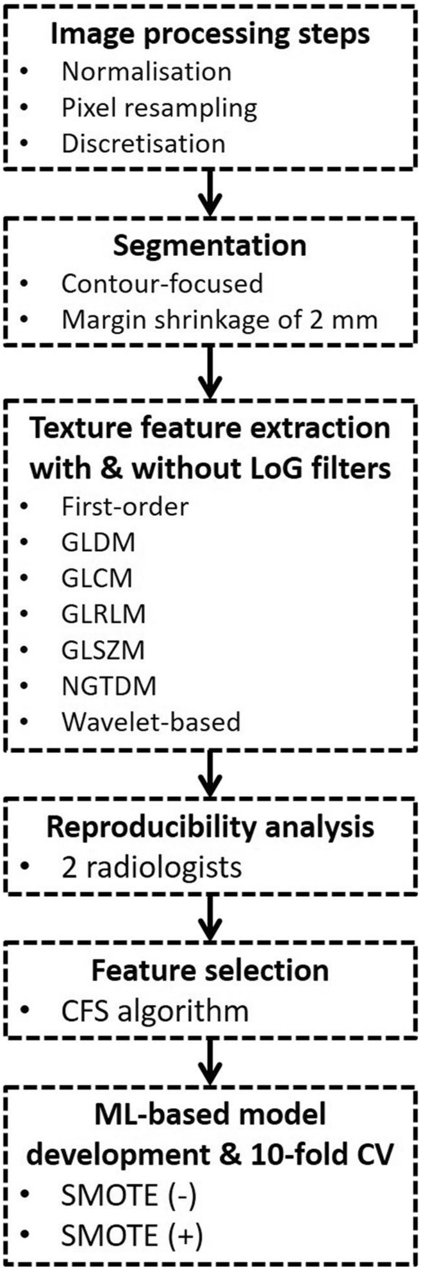

Fig. 1 The flowchart shows the methodologic pipeline of this study. LoG Laplacian of Gaussian, GLDM grey-level dependence matrix, GLCM grey-level co-occurrence matrix, GLRLM grey-level run-length matrix, GLSZM grey-level size zone matrix, NGTDM neighbouring grey-tone difference matrix, CFS correlation-based feature selection, ML machine learning, CV cross-validation, SMOTE *synthetic minority over-sampling technique*

Texture feature extraction

Texture features were extracted using PyRadiomics [21]. An axial image slice that represents the largest cross-sectional area of a tumour was selected. Two segmentations were obtained from each lesion with the following order: (*i*) one with contour-focused (focusing on visible outermost boundary of the tumours) and (*ii*) one with margin shrinkage of 2 mm from the lesion contour (Fig. 2). The contour-focused segmentation was created manually. On the other hand, the one with shrinkage was performed using margin shrinkage function of the software that creates the procedure equally in every direction.

Texture features were extracted from non-filtered (original), filtered, and wavelet transformed images. Laplacian of Gaussian (LoG) filter was used for image filtration with values of 2 mm, 4 mm, and 6 mm; where, 2 mm, 4 mm, and 6 mm represent fine, medium, and coarse patterns, respectively.

The total number of the features extracted was 828 per lesion. The extracted texture feature groups are given in Online Supplement Part E2.

Reproducibility analysis

To assess the reproducibility of the texture features, two radiologists independently segmented randomly selected 25 tumours. Both radiologists were blind to the nuclear grade of the lesions. Intraclass correlation coefficients (ICC) were calculated for each texture feature. The following scale of ICC was used for assessment of texture feature reproducibility: (*i*) ICC < 0.5, poor reproducibility; (*ii*) $0.75 > \text{ICC} \geq 0.5$, moderate reproducibility; (*iii*) $0.9 > \text{ICC} \geq 0.75$, good reproducibility; and (*iv*) ICC ≥ 0.9, excellent reproducibility [22]. The features with ICC ≥ 0.9 were included in the further feature selection steps. Paired proportions of the features with excellent reproducibility on two segmentation data (contour-focused versus margin shrinkage) were compared using McNemar's test. A two-tailed p value less than 0.05 indicated statistical significance.

Feature selection

The feature selection was performed using the Waikato Environment for Knowledge Analysis (WEKA) toolkit version 3.8.2 (The University of Waikato) [23, 24]. Further details about the feature selection are given in Online Supplement Part E3.

Possible collinearity of the selected features was assessed using Pearson's correlation coefficient (r). The r threshold for collinearity was 0.7 [25].

Machine learning classification

Machine learning (ML)-based classifications were performed using WEKA toolkit version 3.8.2. *K*-nearest neighbours (*k*-

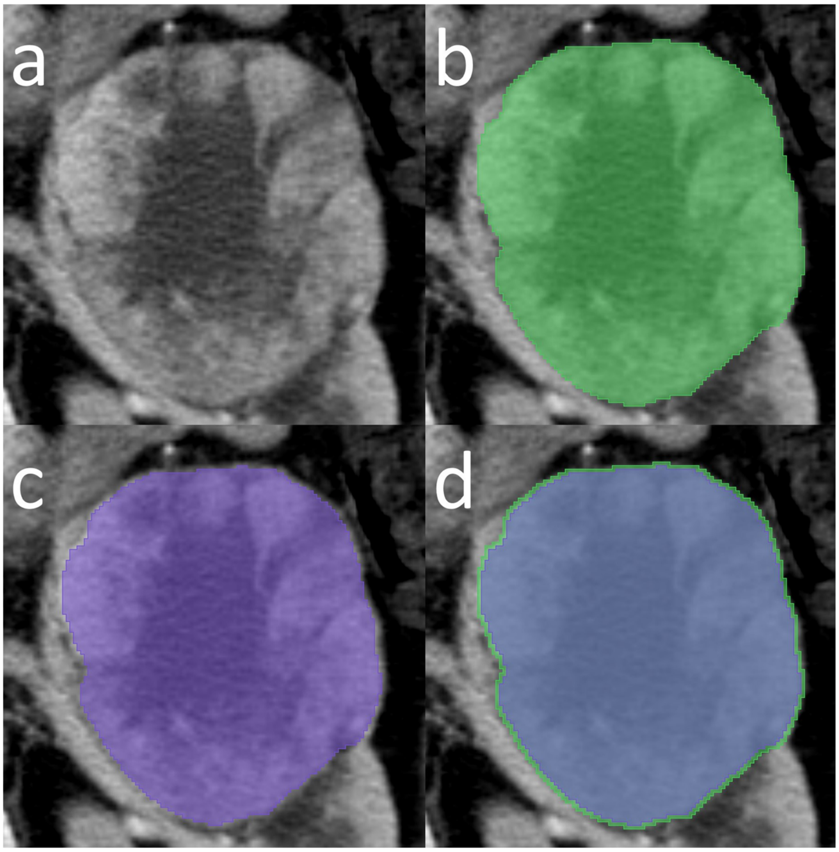

Fig. 2 Tumour segmentation technique. **a** A left-sided renal clear cell carcinoma. **b** Green-coloured area showing the contour-focused segmentation performed manually by focusing on the visible outer margin of the mass. **c** Purple-coloured area showing the segmentation with margin shrinkage of 2 mm performed by the software equally in every direction. **d** Overlay of both segmentations (contour-focused as green line; the one with margin shrinkage of 2 mm as blue area) presenting the equal difference in every direction

NN) algorithm was used for model development. Different k-values were used to avoid over- and under-fitting. All ML models created and validated with and without *synthetic minority over-sampling technique* (SMOTE) so as to avoid possible consequences of class imbalance [26]. Further details on the SMOTE are given in Online Supplement Part E4. Ten-fold cross-validation was adopted for the validation of the model. The performance of classifiers was mainly evaluated and compared by the area under the curve (AUC). Accuracy, sensitivity, specificity, precision, F-measure (weighted harmonic mean of precision and recall), and the Matthews correlation coefficient were also calculated. Comparisons of the 10-fold cross-validated predictive performance of k-NN classifiers were done using the Wilcoxon signed-ranks test [27]. A two-tailed p value less than 0.05 indicated statistical significance.

Results

Reproducibility analysis

Reproducibility analysis by two radiologists revealed that 86.2% of texture features (714 of 828) had excellent reproducibility (ICC ≥ 0.9) on contour-focused segmentation. The number of the features with good (0.9 > ICC ≥ 0.75), moderate (0.75 > ICC ≥ 0.5), and poor (ICC < 0.5) reproducibility was 36 (4.3%), 48 (5.8%), and 30 (3.6%), respectively. For contour-focused segmentation, Supplementary Table 2 shows the features without excellent reproducibility in detail.

On the other hand, using segmentation with margin shrinkage of 2 mm, 93.2% (772 of 828) of the texture features had excellent reproducibility (ICC ≥ 0.9). The number of the features with good (0.9 > ICC ≥ 0.75), moderate (0.75 > ICC ≥ 0.5), and poor (ICC < 0.5) reproducibility was 40 (4.8%), 9 (1.1%), and 7 (0.8%), respectively. For the segmentation with margin shrinkage, Supplementary Table 3 shows the features without excellent reproducibility in detail.

In summary, the number of the features with excellent (ICC ≥ 0.9) and good (0.9 > ICC ≥ 0.75) reproducibility increased from contour-focused segmentation to the segmentation with margin shrinkage of 2 mm.

Difference in proportions of the features with excellent reproducibility was statistically significant (difference [confidence interval], − 7% [− 8.9 to − 5.1%], $p < 0.0001$). The number and the proportions of the features based on two segmentation data are presented in Supplementary Table 4.

Feature selection

Using the correlation-based feature selection algorithm, the number of the features selected was 5 for contour-focused segmentation data and 4 for the segmentation with margin shrinkage of 2 mm. Only one texture feature appeared on both segmentation data. The remaining selected texture features were completely different. Table 2 shows the selected features in detail. The distribution of normalised texture feature values for the contour-focused segmentation and the one with margin shrinkage is presented in Fig. 3 and Fig. 4**,** respectively**.**

All of the selected texture features for two segmentation data were extracted from the images with LoG filter or wavelet transformation. For the segmentation with margin shrinkage of 2 mm, all of the four features were from wavelet images. On the other hand, for contour-focused segmentation, four out of five features were from wavelet images.

Regarding the texture feature classes like first-order, grey-level co-occurrence matrix and grey-level dependence matrix, all of the features were from the different classes in the qTA based on the contour-focused segmentation. On the other hand, there was a dominance of grey-level dependence matrix in the qTA based on the segmentation with margin shrinkage.

There was no significant collinearity between selected features for each segmentation type (Fig. 5).

Classification without SMOTE

Using the contour-focused segmentation, k-NN classifier with k-values of 3, 5, and 7 correctly classified 80.8%, 85.1%, and 82.9% of the RcCCs regarding nuclear grade with AUC values of 0.984, 0.870, and 0.865, respectively. On the other hand, using the segmentation with margin shrinkage of 2 mm, k-NN classifier with k-values of 3, 5, and 7 correctly classified 65.9%, 63.8%, and 70% of the RcCCs regarding nuclear grade with AUC values of 0.745, 0.786, and 0.797, respectively. Detailed performance metrics are presented in Table 3**.**

Classification with SMOTE

Using the contour-focused segmentation, k-NN classifier with k-values of 3, 5, and 7 correctly classified 89.3%, 89.3%, and 89.3% of the RcCCs regarding nuclear grade with AUC values of 0.949, 0.944, and 0.928, respectively. On the other hand, using the segmentation with margin shrinkage of 2 mm, k-NN classifier with k-values of 3, 5, and 7 correctly classified 77.2%, 77.2%, and 71.2% of the RcCCs regarding nuclear grade with AUC values of 0.887, 0.858, and 0.846, respectively. Detailed performance metrics are presented in Table 4**.**

Comparison of the models

Either with or without SMOTE, the AUC values in 10-fold cross-validation were statistically significantly different between all ML-based models created using the contour-focused segmentation and the one with the shrinkage of 2 mm, $p < 0.05$ (Fig. 6 and Table 5). The classification performance of the models with contour-focused outperformed the ones with the shrinkage of 2 mm.

Table 2 Selected feature subsets for each segmentation data

Segmentation type and features	Selected texture features			
	Image type	Feature class	Feature name	ICC
Contour-focused				
cTexF1	LoG filter of 6 mm	GLDM	Dependence non-uniformity normalised	0.997
*cTexF2**	Wavelet-HL	GLCM	Correlation	0.997
cTexF3	Wavelet-HL	First-order	Skewness	0.968
cTexF4	Wavelet-LL	GLRLM	Grey-level non-uniformity	0.998
cTexF5	Wavelet-LH	NGTDM	Complexity	1
Margin shrinkage of 2 mm				
*sTexF1**	Wavelet-HL	GLCM	Correlation	0.997
sTexF2	Wavelet-HH	GLDM	Grey-level non-uniformity	0.997
sTexF3	Wavelet-LH	GLDM	Small dependence high grey-level emphasis	0.995
sTexF4	Wavelet-LH	NGTDM	Contrast	0.972

*indicates the features appearing on both segmentation data sets

ICC intraclass correlation coefficient, *LoG* Laplacian of Gaussian, *GLDM* grey-level dependence matrix, *GLCM* grey-level co-occurrence matrix, *GLRLM* grey-level run length matrix, *NGTDM* neighbouring grey-tone difference matrix, *L* low frequency band, *H* high frequency band

Discussion

Overview

In this retrospective methodological study, we evaluated the influence of a slight difference of 2 mm in segmentation margin on ML-based high-dimensional qCT-TA for distinguishing low and high nuclear grade RcCCs. We found that the minor difference has an influence on each step of the ML-based qCT-TA including interobserver reproducibility of texture features, algorithm-based feature selection, and the ML-based classifications. Despite yielding fewer features with excellent reproducibility, the use of the contour-focused segmentation provided better classification performance in qCT-TA of RcCCs for distinguishing nuclear grade. Taken together, these results

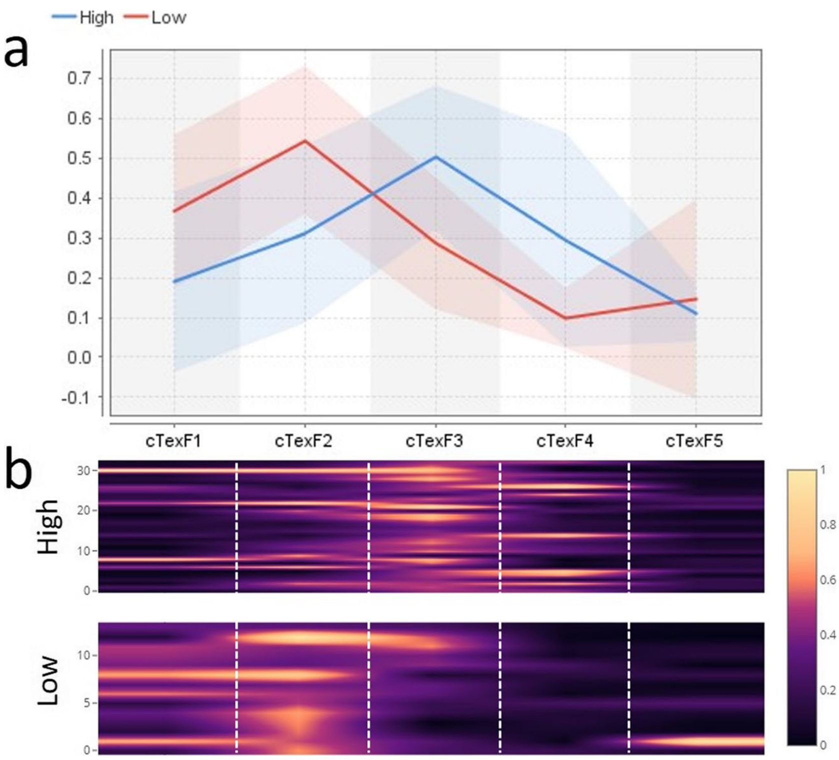

Fig. 3 Distribution of the selected texture features using contour-focused segmentation data. **a** Deviation plot shows mean (blue and red lines) of the normalised texture feature parameters with their corresponding one standard deviation (blue- and red-coloured areas). **b** Coloured and smoothed heat map shows the distribution and differences of normalised texture feature values by presenting each tumour's value. Please refer to Table 2 for the actual feature names. Low and High indicate the nuclear grade of the renal clear cell carcinomas

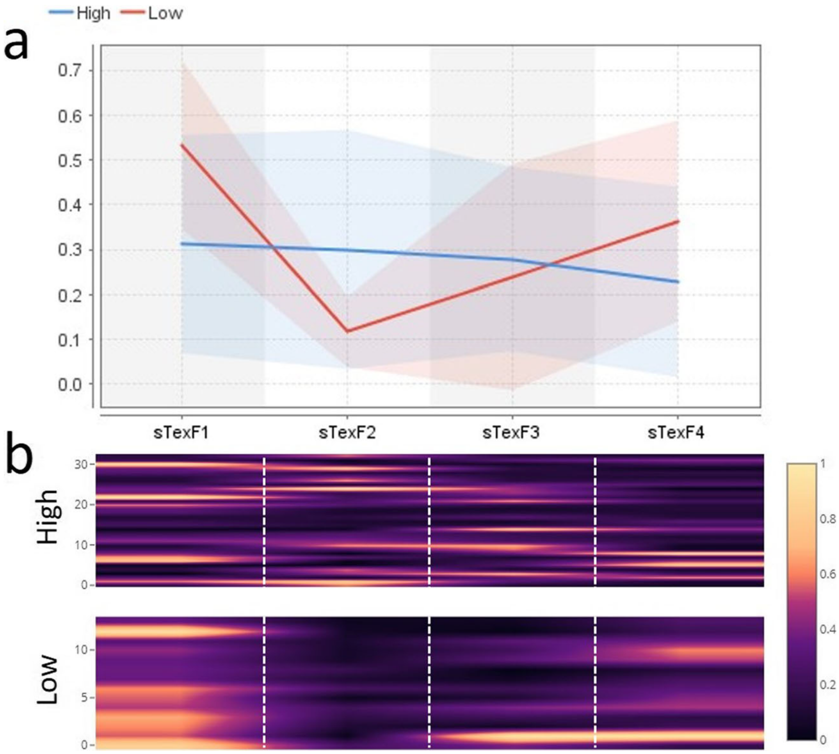

Fig. 4 Distribution of the selected texture features using the segmentation with margin shrinkage of 2 mm. **a** Deviation plot shows mean (blue and red lines) of the normalised texture feature parameters with their corresponding one standard deviation (blue- and red-coloured areas). **b** Coloured and smoothed heat map shows the distribution and differences of normalised texture feature values by presenting each tumour's value. Please refer to Table 2 for the actual feature names. Low and High indicate the nuclear grade of the renal clear cell carcinomas

suggest that findings of a high-dimensional qCT-TA may not be reproducible in clinical practice even using the same feature selection algorithm and ML classifier unless the segmentation margin (contour-focused versus margin shrinkage) is considered.

Practical implications

Although some researchers tend to segment renal cell carcinomas (RCCs) with a secure peripheral zone in order to avoid partial volume effect or volume averaging [28, 29], our findings

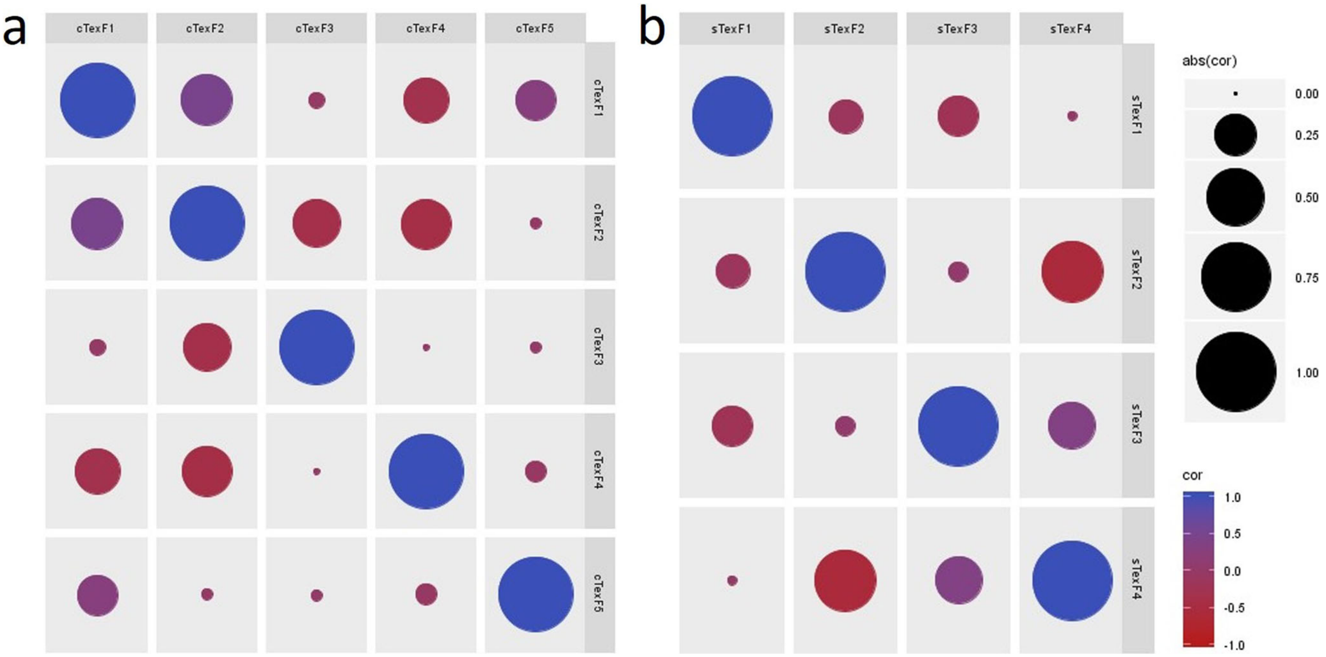

Fig. 5 The correlation matrix presenting the auto- and cross-correlation of the selected features for (**a**) contour-focused segmentation and (**b**) the one with margin shrinkage of 2 mm. No significant correlation is present among the features. Please refer to Table 2 for the actual feature names. abs absolute value, cor correlation

Table 3 Performance of the machine learning–based classifications without *synthetic minority over-sampling technique* (SMOTE)

Segmentation/classifier	Accuracy	Sensitivity	Specificity	Precision	F-measure	MCC	AUC	Confusion matrix		
								H	L	R
Contour-focused										
3-NN	80.8%	81.8%	78.6%	90%	0.857	0.575	0.984	27	6	High
		78.6%	81.8%	64.7%	0.710			3	11	Low
5-NN	85.1%	81.8%	92.9%	96.4%	0.885	0.696	0.870	27	6	High
		92.9%	81.8%	68.4%	0.929			1	13	Low
7-NN	82.9%	78.8%	92.9%	96.3%	0.867	0.663	0.865	26	7	High
		92.9%	78.8%	65%	0.765			1	13	Low
Margin shrinkage of 2 mm										
3-NN	65.9%	72.7%	50%	77.4%	0.750	0.219	0.745	24	9	High
		50%	72.7%	43.8%	0.467			7	7	Low
5-NN	63.8%	69.7%	50%	76.7%	0.730	0.187	0.786	23	10	High
		50%	69.7%	41.2%	0.452			7	7	Low
7-NN	70%	78.8%	50%	78.8%	0.788	0.288	0.797	26	7	High
		50%	78.8%	50%	0.500			7	7	Low

NN nearest neighbours, *MCC* Matthews correlation coefficient, *AUC* area under the curve, *H* and *High* high grade, *L* and *Low* low grade, *R* reference standard

suggest that using contour-focused segmentation rather than with shrinkage might improve the ML-based classification results by yielding better texture features despite having fewer reproducible features. These findings might be interpreted contradictory (fewer reproducible features versus high ML-based classification results). However, a higher number of reproducible features does not have to mean that the features have valuable information for the textural profile of the lesions or tumours. We think that these are completely different entities. Slight changes in segmentation margin might have a potential to change textural information and selected texture features gathered from the segmentation data. To obtain consistent results in ML-based qCT-TA of RcCCs, the concerns about segmentation margin presented in this study need to be considered carefully.

Our study may also have other potential implications because the nuclear grade of RcCC is considered one of the most

Table 4 Performance of the machine learning–based classifications with *synthetic minority over-sampling technique* (SMOTE)

Segmentation/classifier	Accuracy	Sensitivity	Specificity	Precision	F-measure	MCC	AUC	Confusion matrix		
								H	L	R
Contour-focused										
3-NN	89.3%	81.8%	97%	96.4%	0.885	0.797	0.949	27	6	High
		97%	81.8%	84.2%	0.901			1	32	Low
5-NN	89.3%	78.8%	100%	100%	0.881	0.806	0.944	26	7	High
		100%	78.8%	82.5%	0.904			0	33	Low
7-NN	89.3%	78.8%	100%	100%	0.881	0.806	0.928	26	7	High
		100%	78.8%	82.5%	0.904			0	33	Low
Margin shrinkage of 2 mm										
3-NN	77.2%	69.7%	84.8%	82.1%	0.754	0.552	0.887	23	10	High
		84.8%	69.7%	73.7%	0.789			5	28	Low
5-NN	77.2%	63.6%	90.9%	87.5%	0.737	0.567	0.858	21	12	High
		90.9%	63.6%	71.4%	0.800			3	30	Low
7-NN	71.2%	54.5%	87.9%	81.8%	0.655	0.450	0.846	18	15	High
		87.9%	54.5%	65.9%	0.753			4	29	Low

NN nearest neighbours, *MCC* Matthews correlation coefficient, *AUC* area under the curve, *H* and *High* high grade, *L* and *Low* low grade, *R* reference standard

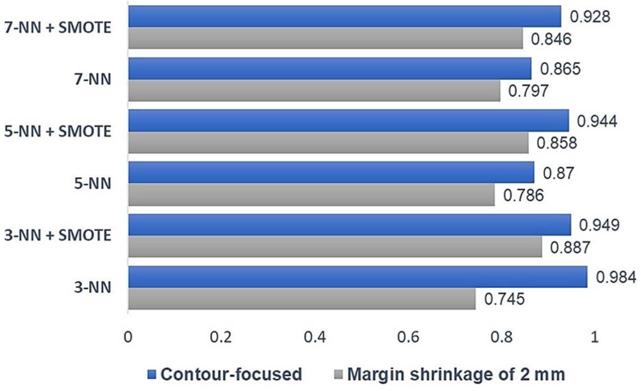

Fig. 6 The bar-chart shows the comparison of the predictive performance of the classifiers with and without using *synthetic minority over-sampling technique* (SMOTE) based on the two different segmentation data (contour focused versus margin shrinkage of 2 mm). The performance metric for comparison was the 10-fold cross-validated area under the curve value (AUC). The predictive performance of the classifiers using contour-focused segmentation data is better than those using the one with margin shrinkage. This comparison is also supported by the Wilcoxon signed-ranks test for all classifiers ($p < 0.05$). NN, nearest neighbours; SMOTE, *synthetic minority over-sampling technique*

important prognostic factors [30–32]. In a meta-analysis, the percutaneous biopsy, which is an invasive method and prone to significant sampling bias, showed a moderate concordance (87%) with nuclear grade [33]. In our study, non-invasive ML-based qCT-TA showed a comparable predictive performance with percutaneous biopsy [33]. Also, the qCT-TA might also be considered for active surveillance of small renal masses, which may allow repeated non-invasive assessment of the nuclear grade during follow-up [34, 35].

Generalisability issues, limitations, and future perspectives

Several generalisability issues and limitations to this experimental study need to be acknowledged. First, there were inherent downsides of a retrospective study design. For quantitative texture analysis, there is no absolute need for a prospective design [1]. Second, the number of our patient population was rather small, mainly due to our strict criteria for inclusion of the patients who had a corticomedullary phase CE-CT, which might lead a risk of overfitting regarding ML-based classification. However, we tried to minimise this potential bias by using a simple ML scheme (*k*-NN) along with different *k*-values. Third, our classes were relatively imbalanced. Therefore, we performed our analysis with and without SMOTE [26]. Fourth, even though three-dimensional qTA might be more representative for textural information [36], only the largest two-dimensional sections were used in this study. This was primarily due to our objective that is to be a guide for most of the future research because the majority of the clinical research on qTA of RCCs have been using single or a few slice-based segmentations. Fifth, the inclusion of patients from different centres might be seen as an important limitation yet it might be considered as a representation of the clinical practice. On the other hand, all of the image data sets in our study underwent a normalisation procedure to minimise inter-scanner variabilities and effects [37, 38]. In addition, all images also were rescaled and discretised because it has been shown that texture analysis has dependency to these preprocessing steps [6]. Sixth, using a slice thickness of 5 mm might be considered a limitation as in conventional analysis. In contrast, our group think that rather than thickness of the slices, the consistency of the thickness is much more important in qTA. Seventh, we did not use independent external datasets (single or multi-institutional) to validate the performance of the classifiers further. Eighth, we only included the corticomedullary phase images in the analysis because of its widespread use in most centres. Given the experimental nature of the study, we think this is not important at this stage. The primary purpose of this study was to draw the researcher's attention regarding the dependency of the method to the segmentation margin. Because it is not consistent in the dataset, we could not include the unenhanced or nephrographic phase in this study. Ninth, we did not perform separate group analysis for small and large lesions due to the small number of patients in total and groups, which would cause a risk of overfitting from the ML perspective. Tenth, we did not consider using semiautomatic and automatic segmentation techniques because manual segmentation is the most widely used technique. Nonetheless, future research would be done using these techniques with a comparative method. Eleventh, because we used a contrast-enhanced phase, we think that the location of a tumour should not be a concern for the analysis. However, based on our experience, it must be a major concern when using unenhanced CT images because in that case it is challenging to delineate the tumour contour even with the help of contrast-enhanced series. Twelfth, we only included RcCCs in the analysis. On the other hand, whether our findings might be extrapolated to the other RCCs should also be further studied.

Table 5 Statistical comparisons of the 10-fold cross-validated classification results based on two different segmentation data (contour-focused versus margin shrinkage of 2 mm) using area under the curve (AUC)

Classifier*	*Z***	*p* value**
3-NN	− 2.805	0.005
5-NN	− 2.347	0.019
7-NN	− 2.829	0.005
3-NN + SMOTE	− 2.809	0.005
5-NN + SMOTE	− 2.814	0.005
7-NN + SMOTE	− 2.552	0.011

*Each classifier used for two different segmentation data (contour-focused versus margin shrinkage of 2 mm)

**Statistical analysis was performed using Wilcoxon signed-ranks test

NN nearest neighbours, *SMOTE synthetic minority oversampling technique*, *Z* standardised test statistics

Thirteenth, tumour necrosis in RCCs is also known as an independent predictor of survival like nuclear grade. We think that necrosis may give important textural information and be responsible in tumour heterogeneity. Measuring only the solid portions of the tumours by excluding the necrotic components might also be considered for grading the tumours in the future.

Conclusions

Each step (interobserver texture feature reproducibility, feature selection, and classification) of the ML-based high-dimensional qCT-TA of RcCCs was susceptible to even a slight change of 2 mm in segmentation margin. Despite yielding fewer features with excellent reproducibility, use of the contour-focused segmentation provided better classification performance in qCT-TA of RcCCs for distinguishing nuclear grade. Taken together, findings of a high-dimensional qCT-TA may not be reproducible in clinical practice even using the same feature selection algorithm and ML classifier, unless the possible influence of the segmentation margin (contour-focused versus margin shrinkage) is considered.

Acknowledgements The abstract of this study has also been submitted to the European Congress of Radiology (ECR) 2019 (Control Number 19-P-339-ECR). The authors acknowledge that they have previously used this public database (The Cancer Genome Atlas-Kidney Renal Clear Cell Carcinoma [TCGA-KIRC]) in different context.

Funding The authors state that this work has not received any funding.

Compliance with ethical standards

Guarantor The scientific guarantor of this publication is Burak Kocak, MD.

Conflict of interest The authors of this manuscript declare no relationships with any companies, whose products or services may be related to the subject matter of the article.

Statistics and biometry One of the authors (Burak Kocak, MD) has significant statistical expertise.

Informed consent Written informed consent was not required for this study because all patients' data included in this study are publicly and freely available for scientific purposes (The Cancer Genome Atlas-Kidney Renal Clear Cell Carcinoma [TCGA-KIRC]).

Ethical approval Institutional Review Board approval was not required because all patients' data included in this study are publicly and freely available for scientific purposes (The Cancer Genome Atlas-Kidney Renal Clear Cell Carcinoma [TCGA-KIRC]).

Study subjects or cohorts overlap The authors acknowledge that they have previously used this public database (The Cancer Genome Atlas-Kidney Renal Clear Cell Carcinoma [TCGA-KIRC]) in different context.

Methodology
- retrospective
- experimental
- performed using a publicly and freely available database

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

References

1. Lubner MG, Smith AD, Sandrasegaran K, Sahani DV, Pickhardt PJ (2017) CT texture analysis: definitions, applications, biologic correlates, and challenges. RadioGraphics 37:1483–1503. https://doi.org/10.1148/rg.2017170056
2. Gillies RJ, Kinahan PE, Hricak H (2016) Radiomics: images are more than pictures, they are data. Radiology 278:563–577. https://doi.org/10.1148/radiol.2015151169
3. Kocak B, Yardimci AH, Bektas CT et al (2018) Textural differences between renal cell carcinoma subtypes: machine learning-based quantitative computed tomography texture analysis with independent external validation. Eur J Radiol 107:149–157. https://doi.org/10.1016/j.ejrad.2018.08.014
4. Bektas CT, Kocak B, Yardimci AH et al (2018) Clear cell renal cell carcinoma: machine learning-based quantitative computed tomography texture analysis for prediction of Fuhrman nuclear grade. Eur Radiol. https://doi.org/10.1007/s00330-018-5698-2
5. Berenguer R, Pastor-Juan MDR, Canales-Vázquez J et al (2018) Radiomics of CT features may be nonreproducible and redundant: influence of CT acquisition parameters. Radiology 288:407–415. https://doi.org/10.1148/radiol.2018172361
6. Shafiq-ul-Hassan M, Zhang GG, Latifi K et al (2017) Intrinsic dependencies of CT radiomic features on voxel size and number of gray levels. Med Phys 44:1050–1062. https://doi.org/10.1002/mp.12123
7. Mackin D, Fave X, Zhang L et al (2015) Measuring computed tomography scanner variability of radiomics features. Invest Radiol 50:757–765. https://doi.org/10.1097/RLI.0000000000000180
8. Balagurunathan Y, Gu Y, Wang H et al (2014) Reproducibility and prognosis of quantitative features extracted from CT images. Transl Oncol 7:72–87
9. Balagurunathan Y, Kumar V, Gu Y et al (2014) Test-retest reproducibility analysis of lung CT image features. J Digit Imaging 27:805–823. https://doi.org/10.1007/s10278-014-9716-x
10. Hunter LA, Krafft S, Stingo F et al (2013) High quality machine-robust image features: identification in nonsmall cell lung cancer computed tomography images. Med Phys 40:121916. https://doi.org/10.1118/1.4829514
11. Leijenaar RT, Carvalho S, Velazquez ER et al (2013) Stability of FDG-PET radiomics features: an integrated analysis of test-retest and inter-observer variability. Acta Oncol 52:1391–1397. https://doi.org/10.3109/0284186X.2013.812798
12. Echegaray S, Gevaert O, Shah R et al (2015) Core samples for radiomics features that are insensitive to tumor segmentation: method and pilot study using CT images of hepatocellular carcinoma. J Med Imaging (Bellingham) 2:041011. https://doi.org/10.1117/1.JMI.2.4.041011
13. Parmar C, Rios Velazquez E, Leijenaar R et al (2014) Robust radiomics feature quantification using semiautomatic volumetric segmentation. PLoS One 9:e102107. https://doi.org/10.1371/journal.pone.0102107
14. Fave X, Zhang L, Yang J et al (2016) Impact of image preprocessing on the volume dependence and prognostic potential of radiomics features in non-small cell lung cancer. Transl Cancer Res 5:349–363. https://doi.org/10.21037/8709

15. Hu P, Wang J, Zhong H et al (2016) Reproducibility with repeat CT in radiomics study for rectal cancer. Oncotarget 7:71440–71446. https://doi.org/10.18632/oncotarget.12199

16. Kim H, Park CM, Lee M et al (2016) Impact of reconstruction algorithms on CT radiomic features of pulmonary tumors: analysis of intra- and inter-reader variability and inter-reconstruction algorithm variability. PLoS One 11:e0164924. https://doi.org/10.1371/journal.pone.0164924

17. Akin O, Elnajjar P, Heller M, et al (2016) Radiology Data from The Cancer Genome Atlas Kidney Renal Clear Cell Carcinoma [TCGA-KIRC] collection. The Cancer Imaging Archive. https://wiki.cancerimagingarchive.net/display/Public/TCGA-KIRC#329b9f4f31cf4586831934c19c4f10f4. Accessed 4 May 2018

18. Clark K, Vendt B, Smith K et al (2013) The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. J Digit Imaging 26:1045–1057. https://doi.org/10.1007/s10278-013-9622-7

19. Kocak B, Durmaz ES, Ates E, Ulusan MB (2019) Radiogenomics in clear cell renal cell carcinoma: machine learning–based high-dimensional quantitative CT texture analysis in predicting PBRM1 mutation status. AJR Am J Roentgenol 1–9. https://doi.org/10.2214/AJR.18.20443

20. Cohan RH, Sherman LS, Korobkin M, Bass JC, Francis IR (1995) Renal masses: assessment of corticomedullary-phase and nephrographic-phase CT scans. Radiology 196:445–451. https://doi.org/10.1148/radiology.196.2.7617859

21. van Griethuysen JJM, Fedorov A, Parmar C et al (2017) Computational radiomics system to decode the radiographic phenotype. Cancer Res 77:e104–e107. https://doi.org/10.1158/0008-5472.CAN-17-0339

22. Koo TK, Li MY (2016) A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 15:155–163. https://doi.org/10.1016/j.jcm.2016.02.012

23. Kohavi R, John GH (1997) Wrappers for feature subset selection. Artif Intell 97:273–324. https://doi.org/10.1016/S0004-3702(97)00043-X

24. Bermejo P, Gamez JA, Puerta JM (2011) Improving incremental wrapper-based subset selection via replacement and early stopping. Intern J Pattern Recognit Artif Intell 25:605–625. https://doi.org/10.1142/S0218001411008804

25. Dormann CF, Elith J, Bacher S et al (2013) Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography (Cop) 36:27–46. https://doi.org/10.1111/j.1600-0587.2012.07348.x

26. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 16:321–357. https://doi.org/10.1613/JAIR.953

27. Demšar J (2006) Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res 7:1–30

28. Yan L, Liu Z, Wang G et al (2015) Angiomyolipoma with minimal fat. Acad Radiol 22:1115–1121. https://doi.org/10.1016/j.acra.2015.04.004

29. Hodgdon T, McInnes MD, Schieda N, Flood TA, Lamb L, Thornhill RE (2015) Can quantitative CT texture analysis be used to differentiate fat-poor renal angiomyolipoma from renal cell carcinoma on unenhanced CT images? Radiology 276:787–796. https://doi.org/10.1148/radiol.2015142215

30. Klatte T, Patard JJ, de Martino M et al (2008) Tumor size does not predict risk of metastatic disease or prognosis of small renal cell carcinomas. J Urol 179:1719–1726. https://doi.org/10.1016/j.juro.2008.01.018

31. Frank I, Blute ML, Cheville JC, Lohse CM, Weaver AL, Zincke H (2002) An outcome prediction model for patients with clear cell renal cell carcinoma treated with radical nephrectomy based on tumor stage, size, grade and necrosis: the SSIGN score. J Urol 168:2395–2400. https://doi.org/10.1097/01.ju.0000035885.91935.d5

32. Zisman A, Pantuck AJ, Dorey F et al (2002) Mathematical model to predict individual survival for patients with renal cell carcinoma. J Clin Oncol 20:1368–1374. https://doi.org/10.1200/JCO.2002.20.5.1368

33. Marconi L, Dabestani S, Lam TB et al (2016) Systematic review and meta-analysis of diagnostic accuracy of percutaneous renal tumour biopsy. Eur Urol 69:660–673. https://doi.org/10.1016/j.eururo.2015.07.072

34. Jewett MAS, Mattar K, Basiuk J et al (2011) Active surveillance of small renal masses: progression patterns of early stage kidney cancer. Eur Urol 60:39–44. https://doi.org/10.1016/j.eururo.2011.03.030

35. Abou Youssif T, Tanguay S (2009) Natural history and management of small renal masses. Curr Oncol 16(Suppl 1):S2–S7

36. Ng F, Kozarski R, Ganeshan B, Goh V (2013) Assessment of tumor heterogeneity by CT texture analysis: can the largest cross-sectional area be used as an alternative to whole tumor analysis? Eur J Radiol 82:342–348. https://doi.org/10.1016/j.ejrad.2012.10.023

37. Schieda N, Lim RS, Krishna S, McInnes MDF, Flood TA, Thornhill RE (2018) Diagnostic accuracy of unenhanced CT analysis to differentiate low-grade from high-grade chromophobe renal cell carcinoma. Am J Roentgenol:1–9. https://doi.org/10.2214/AJR.17.18874

38. Collewet G, Strzelecki M, Mariette F (2004) Influence of MRI acquisition protocols and image intensity normalization methods on texture classification. Magn Reson Imaging 22:81–91. https://doi.org/10.1016/j.mri.2003.09.001