European Radiology (2019) 29:4776–4782
https://doi.org/10.1007/s00330-019-6004-7

IMAGING INFORMATICS AND ARTIFICIAL INTELLIGENCE

Differentiating kidney stones from phleboliths in unenhanced low-dose computed tomography using radiomics and machine learning

Thomas De Perrot[1] · Jeremy Hofmeister[1] · Simon Burgermeister[1] · Steve P. Martin[1] · Gregoire Feutry[1] · Jacques Klein[2] · Xavier Montet[1]

Received: 14 November 2018 / Revised: 21 December 2018 / Accepted: 11 January 2019 / Published online: 12 February 2019

Abstract

Objectives Distinguishing between kidney stones and phleboliths can constitute a diagnostic challenge in patients undergoing unenhanced low-dose CT (LDCT) for acute flank pain. We sought to investigate the accuracy of radiomics and a machine-learning classifier in differentiating between kidney stones and phleboliths on LDCT.

Methods Radiomics features were extracted following a semi-automatic segmentation of kidney stones and phleboliths for two independent consecutive cohorts of patients undergoing LDCT for acute flank pain.

Radiomics features from the first cohort of patients ($n = 369$) were ultimately used to train a machine-learning model designed to distinguish kidney stones ($n = 211$) from phleboliths ($n = 201$). Classification performance was assessed on the second independent cohort (i.e., testing set) (kidney stones $n = 24$; phleboliths $n = 23$) using positive and negative predictive values (PPV and NPV), area under the receiver operating curves (AUC), and permutation testing.

Results Our machine-learning classification model trained on radiomics features achieved an overall accuracy of 85.1% on the independent testing set, with an AUC of 0.902, PPV of 81.5%, and NPV of 90.0%. Classification accuracy was significantly better than chance on permutation testing ($p < 0.05$, permutation p value).

Conclusion Radiomics and machine learning enable accurate differentiation between kidney stones and phleboliths on LDCT in patients presenting with acute flank pain.

Key Points

- *Combining a machine-learning algorithm with radiomics features extracted for abdominopelvic calcification on LDCT offers a highly accurate method for discriminating phleboliths from kidney stones.*
- *Our radiomics and machine-learning model proved robust for CT acquisition and reconstruction protocol when tested in comparison with an external independent cohort of patients with acute flank pain.*
- *The high performance of the radiomics-based automatic classification model in differentiating phleboliths from kidney stones indicates its potential as a future diagnostic tool for equivocal abdominopelvic calcifications in the setting of suspected renal colic.*

Keywords Urinary tract · Lithiasis · Machine learning · Artificial intelligence

Abbreviations

AUC	Area under the curve
CV	Cross-validation
LDCT	Low-dose computed tomography
PCA	Principal component analysis
ROC	Receiver operating characteristics curve
VOI	Volume of interest

✉ Thomas De Perrot
thomas.deperrot@hcuge.ch

[1] Division of Radiology, Diagnostic Department, Geneva University Hospitals, Geneva, Switzerland

[2] Division of Urology, Department of Surgery, Geneva University Hospitals, Geneva, Switzerland

Introduction

Acute flank pain is a frequent motive for medical consultation in primary care and emergency medicine. It is often caused by renal colic, which results from kidney stones blocking the urine flow in the upper urinary tract, and has risen in incidence and prevalence over the last few decades [1]. The methods of diagnosing renal colic have

evolved over the last few years, from abdominal radiography coupled with ultrasound to unenhanced low-dose or even ultra-low-dose computed tomography (LDCT) [2]. The key for the diagnosis relies on the calcification's location in relation to the ureters: A calcification inside a ureter in the setting of acute flank pain typically indicates a kidney stone, while a pelvic calcification outside the ureter is mostly caused by a phlebolith. However, in daily practice, the identification of the ureter and calcification locations on LDCT can be challenging. In cases of diagnostic uncertainty following LDCT, patient management may require a complementary CT study with iodinated contrast enhancement or late CT acquisition and contrast excretion in order to identify the real position of the ureter. Correctly and accurately diagnosing kidney stones is thus of high clinical value, as persistent symptoms may require supplementary urological procedures, such as invasive ureteroscopy or extracorporeal shockwave lithotripsy, aimed at causing kidney stones to fragment [3].

Several studies have reported that kidney stones and phleboliths present different morphological structures, with distinct imaging characteristics observable using abdominal radiography or full-dose CT. Phleboliths have radiolucent centers and are calcified around their periphery [4], while kidney stones tend to instead present more densely organized microstructures due to their microcrystalline nature [5–7]. However, very little is known about the morphological structure of kidney stones and phleboliths on LDCT. Moreover, no study has investigated whether the textural or radiomics characteristics of these abdominopelvic calcifications might be of clinical interest, despite this field of quantitative image analysis and radiomics being a rapidly growing area, primarily in oncological applications [8–12]. In our study, we sought to assess the performance of a machine-learning classifier based on radiomics features designed to differentiate kidney stones from phleboliths on low-dose unenhanced CT scan.

Methods

Patient population

The study was approved by the local ethics committee for research on human subjects (CCER 2017-00922), which waived the need for written informed consent. A senior radiologist (with 18 years of experience) reviewed two independent patient cohorts who underwent LDCT for acute flank pain in two clinics of our institution. Data from the first cohort (i.e., training set) were employed to train our machine-learning classifier (see below), including LDCTs performed between August 2015 and March 2017 on 369 consecutive patients (mean age ± SD, 52 ± 15 years old; range, 19–94), namely 251 males and 118 females (Fig. 1). The second independent cohort (i.e., testing set) provided the data to test our machine-learning classifier in comparison with an independent dataset (see below), including LDCTs performed between January 2018 and August 2018 on 43 consecutive patients (mean age ± SD, 54 ± 17; range, 26–92), comprising 22 males and 21 females. Patients in the training and testing cohort were distinct to keep these datasets independent from each other.

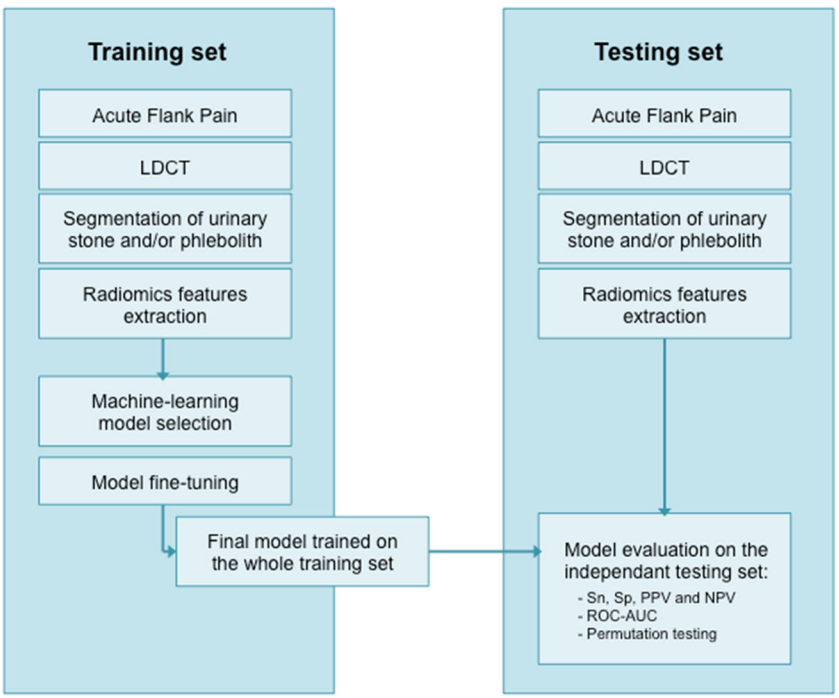

Fig. 1 The pipeline figure illustrates the successive study steps, from low-dose CT acquisition (LDCT) to the model performance assessment using an independent test set. The model selection and fine-tuning were based on the training set only, using cross-validation (see "Methods" section for details). LDCT, low-dose CT scan; Sn, sensitivity; Sp, specificity; PPV, positive predictive value; NPV, negative predictive value

Table 1 Radiomics features used in the study to characterize phlebolith or kidney stone textures

First-order features	GLCM features	Shape features	GLRLM features	GLSZM features	NGTDM features	GLDM features
10 percentile	Autocorrelation	Compactness 1	Gray level non-uniformity	Gray level variance	Coarseness	Small dependence emphasis
90 percentile	Average intensity	Compactness 2	Gray level non-uniformity normalized	High intensity emphasis	Contrast	Large dependence emphasis
Energy	Cluster prominence	Elongation	Gray level variance	High intensity large area emphasis	Busyness	Gray level non-uniformity
Entropy	Cluster shade	Flatness	High gray-level run emphasis	High intensity small area emphasis	Complexity	Dependence non-uniformity
Interquartile range	Cluster tendency	Maximum 2D diameter column	Long run emphasis	Intensity variability	Strength	Gray level variance
Kurtosis	Contrast	Maximum 2D diameter row	Long run high gray level emphasis	Intensity variability normalized		Dependence variance
Maximum	Correlation	Maximum 2D diameter slice	Long run low gray level emphasis	Large area emphasis		Dependence entropy
Mean absolute deviation	Difference average	Maximum 3D diameter	Low gray level run emphasis	Low intensity emphasis		Low gray level emphasis
Mean	Difference entropy	Roundness	Run entropy	Low intensity large area emphasis		High gray level emphasis
Median	Difference variance	Spherical disproportion	Run length non-uniformity	Low intensity small area emphasis		Small dependence low gray level emphasis
Minimum	Dissimilarity	Sphericity	Run length non-uniformity normalized	Size zone variability		Large dependence low gray level emphasis
Range	Energy	Surface area	Run percentage	Size zone variability normalized		Large dependence high gray level emphasis
Robust mean absolute deviation	Entropy	Surface volume ratio	Run variance	Small Area emphasis		
Root mean squared	Homogeneity1	Volume	Short run emphasis	Zone entropy		
Skewness	Homogeneity2		Short run high gray-level emphasis	Zone percentage		
Standard deviation	Inverse difference		Short run low gray-level emphasis	Zone variance		
Total energy	Inverse difference moment					
Uniformity	Inverse difference moment normalized					
Variance	Inverse difference normalized					
	Informal measure of correlation					
	Informal measure of correlation					
	Inverse variance					
	Maximum probability					
	Sum average					
	Sum entropy					
	Sum squares					
	Sum variance 1					
	Sum variance 2					

GLCM gray level co-occurrence matrix, *GLSZM* gray level size zone matrix, *GLRLM* gray level run length matrix, *NGTDM* neighboring gray tone difference matrix, *GLDM* gray level dependence matrix

Image acquisition

All LDCT examinations on the training set were performed using 64-slice multi-detector CT (GE Discovery 750 HD, GE Healthcare) applying the following protocol in helical mode: 0.7-s gantry rotation time, 120 kVp, 1.37:1 beam pitch, 40-mm table feed per gantry rotation, fixed *z*-axis tube current of 45 mA, and 64 × 0.625-mm detector configuration, as routinely used in clinical practice [2]. Images were reconstructed at 2.5 mm thickness with filtered back projection and blended with 40% adaptive statistical iterative reconstruction (ASIR-40) on a soft reconstruction kernel.

All LDCT examinations on the independent testing set were performed using 64-slice multi-detector CT (Somatom Definition Flash, Siemens Healthcare) following this protocol in helical mode: 0.5-s gantry rotation time, 120 kVp, 0.9 beam pitch, 34.5-mm table feed per gantry rotation, a fixed *z*-axis tube current of 40 mA, and 2 × 64 × 0.6-mm detector configuration, as routinely used in clinical practice [2]. Images were reconstructed at 2 mm thickness with advanced modeled iterative reconstruction (ADMIRE, strength 3) projection, on a soft reconstruction kernel.

Imaging analysis and reference standard

The same senior radiologist reviewed all the LDCT examinations, selecting one or two abdominopelvic calcifications on each LDCT according to intra- or extra-uretero-pelvic position. Calcifications positioned inside a well-defined and fat-surrounded ureter associated with urinary tract dilatation proximal to obstruction were diagnosed as kidney stones. Calcifications located inside the bladder were also considered. In contrast, those located distant from the ureters, ipsilateral to an upper urinary tract dilatation, and not in contact with a large vessel arterial wall were considered phleboliths. Soft-tissue calcifications prohibiting straightforward scan-based diagnosis were discarded. In the training set, we analyzed 412 abdominopelvic calcifications from 369 patients, finding 211 kidney stones and 201 phleboliths. In the independent testing set, we considered for analysis 47 abdominopelvic calcifications in 43 patients, finding 24 kidney stones and 23 phleboliths.

Kidney stone and phlebolith segmentation on LDCT

All abdominopelvic calcifications identified in the training and testing datasets were segmented by the same senior radiologist, using the semi-automatic Fast GrowCut algorithm implemented in 3D Slicer (Version 4.8).

Radiomics features extraction

Radiomics features were automatically extracted from the segmented abdominopelvic calcifications using PyRadiomics (Version 2.1.0) [13]. This enabled us to compute first-order features, shape features, gray level co-occurrence matrix (GLCM) features, gray level size zone matrix (GLSZM) features, gray level run length matrix (GLRLM), neighboring gray tone difference matrix (NGTDM), and gray level dependence matrix (GLDM) features from the volume segmented from the LDCT images (Table 1). Further features were added by applying filters to the native LDCT images (Table 2) prior to extracting the radiomics features. Using the PyRadiomics package on the native LDCT images and filtered images enabled the extraction of 147,029 radiomics features in total for each calcification, constituting a large number of input data for the classification procedure described below.

Machine-learning classification model

The machine-learning classification model was developed and optimized on the training set using cross-validation. It was further tested on the independent testing set only once, so as to avoid overfitting from repeated use of the testing set in selecting the model.

First, we sought to select the best machine-learning algorithm family enabling us to identify nephrolithiasis and phleboliths. To this end, we assessed the prediction accuracies of several machine classifiers on the training set using 10-fold stratified cross-validation (CV) (Table 3 for the assessed algorithms), with scikit-learn default parametrization applied to each classifier. The best classification model was selected based on the area under the receiver operating characteristic curve (ROC-AUC) obtained from this stratified CV strategy. Following this algorithm selection step, our best model was based on an AdaBoost classifier. It included a primary standardization step, followed by dimension reduction using principal component analysis (PCA) and a final classification step using an AdaBoost classifier.

Secondly, we aimed to fine-tune our model so as to identify the hyper-parameters enabling our model to achieve the best prediction accuracy on the training set. This hyper-parameter tuning was performed using a grid-search approach, with 5 × 3 (inner × outer loops) nested stratified CV [14]. The best

Table 2 Filters and wavelets used to obtain transformed images from the native LDCT acquisition before extracting iteratively the textural radiomics features

Wavelet family name (abbreviation)	Number of wavelets used
Biorthogonal (bior)	15
Coiflets (coif)	17
Daubechies (db)	38
Reverse biorthogonal (rbio)	15
Symlets (sym)	19
Haar (Haar)	1
Dmeyer (dmey)	1
Lagrangian of Gaussian (LOG)	4
Other filters	
Logarithmic (logarithm)	1

Table 3 Different machine-learning classifiers were assessed on the training set using tenfold stratified cross-validation (CV). The best classifier was selected based on its CV performance during this tenfold CV. For each classifier, preprocessing with standardization and dimension reduction using principal component analysis was fitted before classifying the algorithm with default scikit-learn parameters

Classifier	Accuracy mean ± std
AdaBoost	86.3 ± 3.1%
Support vector machine	83.2 ± 2.0%
Logistic regression	82.7 ± 2.5%
Stochastic gradient descent	81.0 ± 4.1%
Gaussian Naïve Bayes	76.6 ± 4.3%
K-nearest neighbors	71.4 ± 4.9%
Random forest	67.2 ± 2.3%

combination of parameter values was selected based on its ROC-AUC obtained from this nested stratified CV strategy using the training set (Table 3).

Finally, our model's capacity to identify nephrolithiasis was evaluated using the independent testing set, computing overall classification accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), ROC-AUC, and log-loss. Permutation testing was also performed on the testing set by randomly shuffling the label of the testing set *N*-times ($N = 10,000$) and computing prediction accuracy. All the classification steps were performed using scikit-learn package (Version 0.19.1) [15], an open-source python package for machine learning (installed through Anaconda version 4.5.10, with python version 3.6.6).

Results

Overall, 369 patients were included in the training set, each presenting with either kidney stones or phleboliths clearly identifiable on unenhanced LDCT. Our training set included the radiomics features of 211 kidney stones and 201 phleboliths. The independent testing set comprised 43 patients, with radiomics features of 24 kidney stones and 23 phleboliths.

The model was selected using a cross-validation strategy enabling us to identify the best machine-learning classifier (Table 3 for the results of each classifier family). This model was based on an AdaBoost classifier. Hyper-parameter tuning was then performed using grid-search nested CV and selected on the basis of its ROC-AUC. The best model was that which featured an AdaBoost classifier with 75 individual decision trees as base estimators and a learning rate set at 0.1, yielding a cross-validation AUC of 0.87.

The performance of our AdaBoost classifier on the testing set was evaluated as high, with an overall accuracy of 85.1%, sensitivity of 91.7%, specificity of 78.3%, positive predictive value of 81.5%, and negative predictive value of 90.0%, for a ROC-AUC at 0.902 and log-loss of 0.653 (Fig. 2). Permutation testing (with 10,000 permutations) demonstrated that our model performed significantly ($p < 0.001$, permutation p value) above chance level in classifying individual calcified pelvic structures in the independent testing set (Fig. 3). Figure 4 illustrates a clinical situation where such a classifier could be applied.

Discussion

We sought to investigate the accuracy of radiomics and machine learning to differentiate kidney stones from phleboliths in patients undergoing LDCT for acute flank pain. Using a machine-learning classifier trained on the radiomics features of abdominopelvic calcifications from a primary cohort of consecutive patients, we managed to accurately identify kidney stones and phleboliths in a second independent testing cohort.

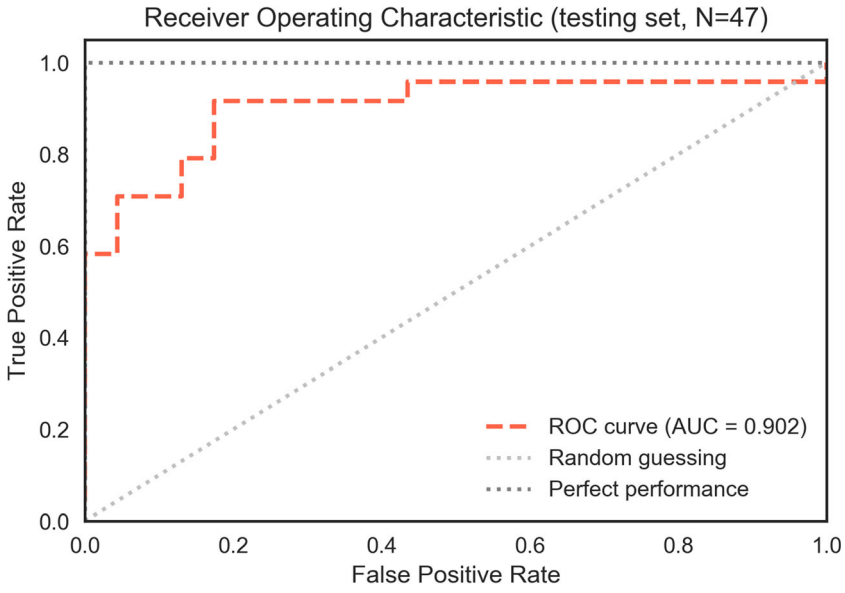

Fig. 2 After training an AdaBoost classifier on the training set, the model's area under the curve was computed using the independent testing set, demonstrating its high performance in classifying kidney stones versus phleboliths. ROC, receiver under the operating curve; AUC, area under the ROC

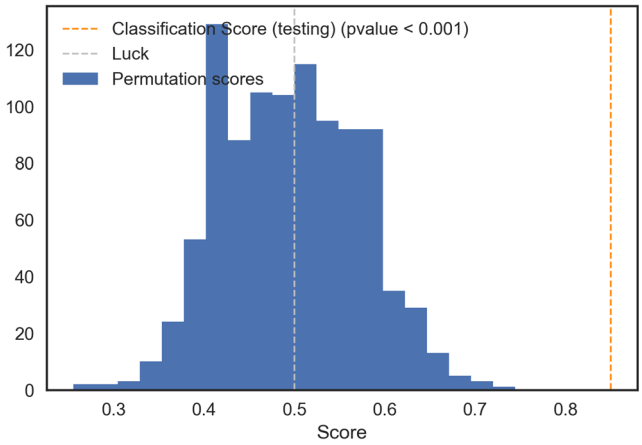

Fig. 3 After training an AdaBoost classifier on the training set, a permutation test was computed on the independent test set. The machine-learning model achieved significantly higher classification accuracy (vertical orange line) than on randomly shuffled labels on the test set (distribution in blue)

Differentiating kidney stones from phleboliths on LDCT review in cases of abdominopelvic calcifications can be challenging, and there are serious implications for clinical management such as additional diagnostic procedures for patients presenting with acute flank pain. Previous studies have revealed that the structure of phleboliths is characterized by a radiolucent center distinct from the phlebolith visible on plain abdominal radiography, yet this observation was not replicated in CT studies [4, 16]. Micro-CT studies, however, reported that kidney stones are not inert calcium formations, but present complex structural organizations [17, 18]. Moreover, macroscopic CT soft-tissue criteria (i.e., the rim and tail sign) have been applied with some success to differentiate kidney stones from phleboliths [19, 20]. However, as these criteria lack quantitative cutoffs, they may not be as easily reproducible on LDCT. We thus took advantage of the recent development of radiomics [21, 22] to show that sub-visual information about the texture of abdominopelvic calcifications coupled with a machine-learning classifier accurately differentiates kidney stones from phleboliths on LDCT. This finding could suggest that the internal microstructure and textural information of kidney stones and phleboliths differ on LDCT, as was previously reported using plain abdominal radiography and full-dose CT scan [4, 18].

One of the main advantages of a radiomics and machine-learning approach is its quantitative nature and reproducibility. Once an abdominopelvic calcification is segmented, a large number of quantitative radiomics features can be rapidly extracted and fed into a machine-learning classifier, which outputs the probability of the calcification being either a kidney stone or phlebolith with high accuracy. One of the downsides of such an approach, though, is its potential dependency upon a fixed CT acquisition protocol and the poor ability of the machine-learning model to generalize for LDCT scans acquired in different settings (e.g., distinct CT model or constructor, different acquisition parameters, or different reconstruction algorithms) [23, 24]. Our study demonstrated, however, that our machine-learning classifier might be robust enough in terms of acquisition settings while still yielding very high prediction accuracy due to it being trained and independently tested on two distinct CT scan models. We achieved very good sensitivity, specificity, PPV, and NPV on an independent testing cohort scanned using a Siemens

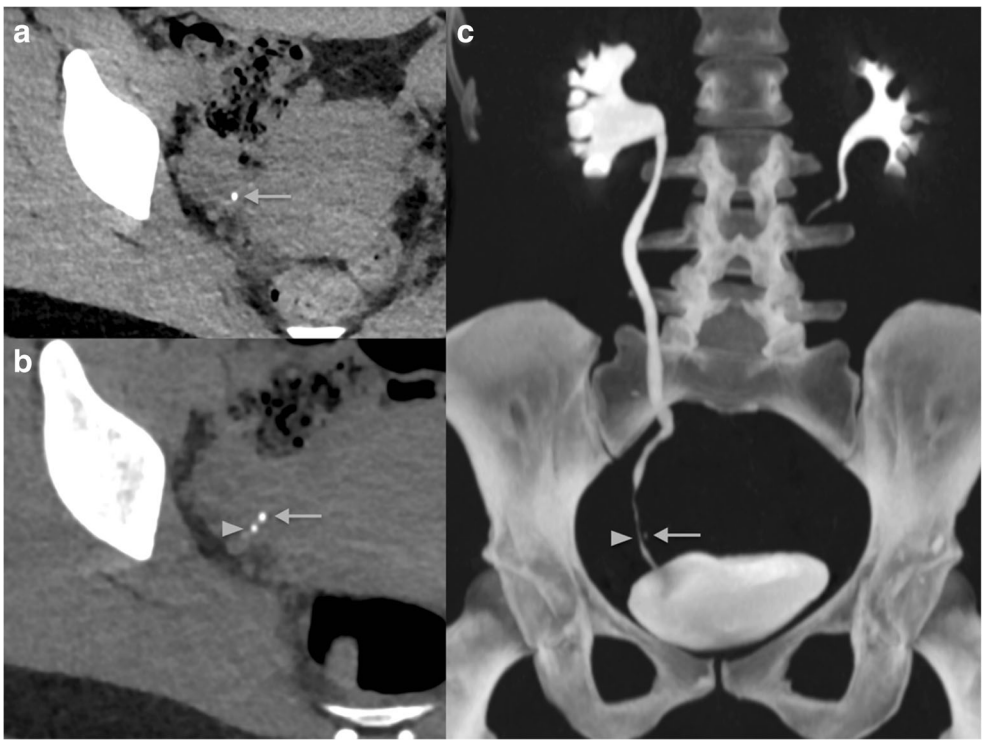

Fig. 4 Illustrative case of a young female patient (25 years old) with a calcified structure in the right pelvis (arrow) observed on LDCT (**a**), initially suspected as a kidney stone causing an upstream uretero-pelvic calyceal dilatation. The periureteral fatty planes being absent (**a**, **b**), the precise location of the calcification required a contrast injection in order to delineate the ureter (arrowhead). The late acquisition images in axial (**b**) and MIP coronal reconstruction (**c**) revealed the calcification to be a juxta-ureteral phlebolith. Using this example as a case external to the study population, the linear AdaBoost model classified the structure with the unenhanced images (**a**) as a phlebolith

Somatom Definition Flash, while the training cohort data were acquired on a GE Discovery 750 HD, with different acquisition parameters and reconstruction algorithms (notably different iterative reconstruction algorithms: ASIR-40 for GE and ADMIRE [strength 3] for Siemens).

All in all, we found that radiomics and machine learning offer accurate differentiation between kidney stones and phleboliths on LDCT, being robust for CT acquisition and reconstruction protocols in an external independent patient cohort with acute flank pain. With the rapid rise of machine learning and artificial intelligence, we believe that our findings have the potential to improve clinical management and reduce having to subject patients presenting with acute flank pain undergoing LDCT to additional procedures. Finally, our work contributes to the extension of radiomics beyond oncological imaging [25], with promising applications in this especially frequent motive for medical consultation in primary care and emergency medicine.

Compliance with ethical standards

Guarantor The scientific guarantor of this publication is Prof. Xavier Montet.

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 has significant statistical expertise.

Informed consent Written informed consent was waived by the Institutional Review Board.

Ethical approval Institutional Review Board approval was obtained.

Methodology
- retrospective
- experimental
- performed at one institution

References

1. Ziemba JB, Matlaga BR (2017) Epidemiology and economics of nephrolithiasis. Investig Clin Urol 58:299–306
2. Poletti PA, Platon A, Rutschmann OT, Schmidlin FR, Iselin CE, Becker CD (2007) Low-dose versus standard-dose CT protocol in patients with clinically suspected renal colic. AJR Am J Roentgenol 188:927–933
3. Luk AC, Cleaveland P, Olson L, Neilson D, Srirangam SJ (2017) Pelvic phlebolith: a trivial pursuit for the urologist? J Endourol 31:342–347
4. Traubici J, Neitlich JD, Smith RC (1999) Distinguishing pelvic phleboliths from distal ureteral stones on routine unenhanced helical CT: is there a radiolucent center? AJR Am J Roentgenol 172:13–17
5. Humphry GM (1896) Urinary calculi: their formation and structure. J Anat Physiol 30:296–311
6. Williams JC Jr, McAteer JA, Evan AP, Lingeman JE (2010) Micro-computed tomography for analysis of urinary calculi. Urol Res 38:477–484
7. Prien EL, Prien EL Jr (1968) Composition and structure of urinary stone. Am J Med 45:654–672
8. Summers RM (2016) Texture analysis in radiology: does the emperor have no clothes? Abdom Radiol (NY). https://doi.org/10.1007/s00261-016-0950-1
9. Parekh V, Jacobs MA (2016) Radiomics: a new application from established techniques. Expert Rev Precis Med Drug Dev 1:207–226
10. Larue RT, Defraene G, De Ruysscher D, Lambin P, van Elmpt W (2017) Quantitative radiomics studies for tissue characterization: a review of technology and methodological procedures. Br J Radiol 90:20160665
11. Aerts HJ, Velazquez ER, Leijenaar RT et al (2014) Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun 5:4006
12. Haralick RM, Shanmugam K, Dinstein I (1973) Textural features for image classification. IEEE Trans Syst Man Cybern SMC-3: 610–621
13. van Griethuysen JJM, Fedorov A, Parmar C et al (2017) Computational radiomics system to decode the radiographic phenotype. Cancer Res 77:e104–e107
14. Cawley GC, Talbot NLC (2010) On over-fitting in model selection and subsequent selection bias in performance evaluation. J Mach Learn Res 11:2079–2107
15. Pedregosa F, Varoquaux G, Gramfort A et al (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825–2830
16. Kim JC (2001) Central lucency of pelvic phleboliths: comparison of radiographs and noncontrast helical CT. Clin Imaging 25:122–125
17. Williams JC Jr, Lingeman JE, Coe FL, Worcester EM, Evan AP (2015) Micro-CT imaging of Randall's plaques. Urolithiasis 43(Suppl 1):13–17
18. Zarse CA, McAteer JA, Tann M et al (2004) Helical computed tomography accurately reports urinary stone composition using attenuation values: in vitro verification using high-resolution micro-computed tomography calibrated to fourier transform infrared microspectroscopy. Urology 63:828–833
19. Boridy IC, Nikolaidis P, Kawashima A, Goldman SM, Sandler CM (1999) Ureterolithiasis: value of the tail sign in differentiating phleboliths from ureteral calculi at nonenhanced helical CT. Radiology 211:619–621
20. Heneghan JP, Dalrymple NC, Verga M, Rosenfield AT, Smith RC (1997) Soft-tissue "rim" sign in the diagnosis of ureteral calculi with use of unenhanced helical CT. Radiology 202:709–711
21. Beig N, Patel J, Prasanna P et al (2018) Radiogenomic analysis of hypoxia pathway is predictive of overall survival in glioblastoma. Sci Rep 8(7)
22. Thawani R, McLane M, Beig N et al (2018) Radiomics and radiogenomics in lung cancer: a review for the clinician. Lung Cancer 115:34–41
23. Zhao B, Tan Y, Tsai WY et al (2016) Reproducibility of radiomics for deciphering tumor phenotype with imaging. Sci Rep 6:23428
24. Parmar C, Rios Velazquez E, Leijenaar R et al (2014) Robust Radiomics feature quantification using semiautomatic volumetric segmentation. PLoS One 9:e102107
25. Incoronato M, Aiello M, Infante T et al (2017) Radiogenomic analysis of oncological data: a technical survey. Int J Mol Sci 18