


Original Article

Evaluation and understanding of automated urinary stone recognition methods

Jonathan El Beze^{1,2} , Charles Mazeaud^{1,2}, Christian Daul³, Gilberto Ochoa-Ruiz⁴, Michel Daudon⁵, Pascal Eschwège^{1,2,3} and Jacques Hubert^{1,2,6}

¹Department of Urology, CHU Nancy – Brabois, ²Université de Lorraine, ³CRAN UMR 7039, Université de Lorraine and CNRS, Nancy, ⁵Unit of Functional Explorations, INSERM UMRS 1155, Hospital Tenon, APHP, Paris, France, ⁴Tecnologico de Monterrey, Escuela de ingeniería y Ciencias, Mexico and ⁶IADI-UL-Inserm (U1254)

Objective

To assess the potential of automated machine-learning methods for recognizing urinary stones in endoscopy.

Materials and Methods

Surface and section images of 123 urinary calculi (109 *ex vivo* and 14 *in vivo* stones) were acquired using ureteroscopes. The stones were more than 85% ‘pure’. Six classes of urolithiasis were represented: Groups I (calcium oxalate monohydrate, whewellite), II (calcium oxalate dihydrate, weddellite), III (uric acid), IV (brushite and struvite stones), and V (cystine). The automated stone recognition methods that were developed for this study followed two types of approach: shallow classification methods and deep-learning-based methods. Their sensitivity, specificity and positive predictive value (PPV) were evaluated by simultaneously using stone surface and section images to classify them into one of the main morphological groups (subgroups were not considered in this study).

Results

Using shallow methods (based on texture and colour criteria), relatively high sensitivity, specificity and PPV for the six classes were attained: 91%, 90% and 89%, respectively, for whewellite; 99%, 98% and 99% for weddellite; 88%, 89% and 88% for uric acid; 91%, 89% and 90% for struvite; 99%, 99% and 99% for cystine; and 94%, 98% and 99% for brushite. Using deep-learning methods, the sensitivity, specificity and PPV for each of the classes were as follows: 99%, 98% and 97% for whewellite; 98%, 98% and 98% for weddellite; 97%, 98% and 98% for uric acid; 97%, 97% and 96% for struvite; 99%, 99% and 99% for cystine; and 94%, 97% and 98% for brushite.

Conclusion

Endoscopic stone recognition is challenging, and few urologists have sufficient expertise to achieve a diagnosis performance comparable to morpho-constitutional analysis. This work is a proof of concept that artificial intelligence could be a solution, with promising results achieved for pure stones. Further studies on a larger panel of stones (pure and mixed) are needed to further develop these methods.

Keywords

automated kidney stone recognition, ureteroscopy, morphoconstitutional analysis, urolithiasis, deep learning

Introduction

Urolithiasis is a frequent and recurrent pathology. Its management is medico-surgical. The medical aspect of the management of lithiasis disease is based on overall assessment and morpho-constitutional analysis of the stone [1]. The objective is to identify the aetiology and avoid recurrence.

The morpho-constitutional examination consists of two steps [2–4]: (1) morphological analysis of the surface and section of the extracted kidney stone and (2) infrared spectroscopy to determine the composition of the stone.

Diagnostic agreement of the morpho-constitutional analysis reaches 95% when the whole stone is sent for analysis, but it only reaches up to 60% for global infrared analysis of the powder of a stone or of small fragments. However, only 29.6% of urinary stones arrive whole in the laboratory [5].

As a result of the evolution of surgical techniques and lasers in recent decades, ureteroscopy with possible laser fragmentation of the stone has become a popular therapeutic option. However, given that the therapeutic objective of the intervention is to obtain a stone-free result [6], operating techniques may involve fragmentation of the stone, in

'popcorn' mode, into smaller and smaller fragments and / or spraying of the stone, in 'dusting' mode, into increasingly fine powder. This leads to a low rate of whole stones recovered for optimal laboratory analysis. In addition, it has been shown that spraying in dusting mode can modify the chemical composition of certain urinary stones and impair their recognition by spectrophotometry, thus changing their relation to aetiology [7,8].

These facts explain the growing interest in the intra-operative recognition of urinary lithiasis. Indeed, the urologist is the first to 'see' urolithiasis before and after its fragmentation, if this is performed. Intra-operative description of the stone can complement the morpho-constitutional analysis. Thanks to the work of Estrade et al. [5,9], this approach has enabled the validation of didactic boards of confirmed endoscopic images to aid in the peri-endoscopic recognition of certain types of urinary lithiasis. However, this method requires training and expertise that is difficult for all urologists to acquire. The objective of our study therefore was to assess the potential of an automated method for real-time recognition of urinary stones using machine-learning methods, utilizing images seen during ureteroscopy.

Materials and Methods

Morpho-constitutional Analysis

In the laboratory, a stone is examined according to morpho-constitutional analysis that comprises two steps: (1) a morphological step, in which the surface and the section of the stone are analysed under microscope with regard to colour, texture and shape and (2) a constitutional step, in which infrared spectrophotometric analysis is performed of the crystal component(s) of the urinary calculus.

A morpho-constitutional classification system was proposed by Daudon et al. [2,10]. The current classification divides urinary stones into seven types and 22 subtypes (Table 1).

Groups I and II include calcium oxalate monohydrate (Group I = whewellite) and calcium oxalate dihydrate (Group II = weddellite) stones, Group III includes uric acid and urates, Group IV includes calcium and/or magnesium phosphates, and Group V includes cystine. Group VI includes protein stones. Group VII contains miscellaneous types of stones.

Acquisition Equipment

In this study, we used 109 human urinary stones (Table 2a) from a historical series obtained from the CRISTAL laboratory, Paris France. These calculi are part of a study on the analysis of the densities of urinary stones evaluated on CT [11]. The stones are fully anonymized and were kept for study in compliance with the regulations in place during their endoscopic or surgical extraction.

The morpho-constitutional analysis of these stones has been previously established. They contain at least 85% of a single component. Thus, they are considered 'pure'. We acquired images of these stones using two reusable digital flexible ureteroscopes from Karl Storz® using video carts: Storz Image 1 Hub and Storz image1 S.

To reproduce *in vivo* conditions, namely, a closed environment such as urinary excretory cavities, we used a tube with a small diameter whose inner walls were covered with a yellowish film to simulate the appearance of the walls of the urinary tract. The aim was for the light, light reflection and distance to approximate the conditions found the *in vivo* environment (Fig. 1A).

For each stone, we acquired images from different points of view (far, near and/or from different angles). For the fragmented stones, we took images of the surface and the section. For unfragmented (whole) calculi, only surface images were acquired. Figure 1B shows some examples of the images taken of these stones.

For the images of the cystine stone shown (acquired with Storz® Image 1 Hub), the resolution was lower, the light reflection being more important than for the other images (acquired with Storz® Image 1 S). However, the difference was ultimately minimal and the automated recognition results remained high, as shown below.

For each type of urinary stone, we obtained between 25 (brushite) and 62 (whewellite [calcium oxalate monohydrate]) surface images, and between 20 (brushite) and 50 (uric acid) section images (Table 2b).

A limitation of the *ex vivo* database lies in the differing number of images available for each class. Because the brushite class was under-represented in terms of images in comparison to the other classes, the database was extended with images of brushite stones (14 images of section and 14 of surface) from a previous *in vivo* study [12].

Only parts of the images were used for classification purposes. Indeed, after eliminating obvious artefacts (e.g., instruments visible in the images), each image was divided into 'patches' of 256 × 256 pixels each. The use of patches instead of whole images is not only consistent with medical practice for morpho-constitutional analysis, but also allows the construction of a larger training and test dataset. Thus, the amount of information extracted for training the automated machine-learning models is increased, avoiding the well-known problem of overfitting.

The optimal patch size was determined in a previous study as that which offered the best accuracy with the minimal loss of information during the automated stone recognition [12]. The patch extraction process was performed randomly after the automated segmentation and image pre-processing process

Table 1 Morpho-constitutional analysis [2–4,7,10].

Type	Main crystal component	Main morphological stone characteristics		Common causes	
		Surface	Section		
I	Ia	Whewellite (COM)	Mammillary surface. Frequent umbilication and Randall's plaque. Colour: brown	Concentric layers with radiating organization starting from a nucleus (often Randall's plaque) Colour: brown	Insufficient water intake (low diuresis) and dietary hyperoxaluria (high consumption of oxalate-rich foods (e.g., dark chocolate and spinach) and hydroxyproline-rich foods and low calcium intake (increased oxalate absorption by the gut) Moderate hyperoxaluria with stasis Primary hyperoxaluria type I Malformative uropathy, stasis and confined multiple stones Enteric hyperoxaluria: inflammatory bowel disease (Crohn's disease and/or extensive ileal resections), bariatric surgery, and chronic pancreatitis Hypercalciuria
	Ib		Mammillary, rough surface. No umbilication. Colour: brown to dark brown	Unorganized. Colour: brown to dark brown	
	Ic		Budding surface. Colour: cream to pale yellow-brown	Finely granular and poorly organized. Colour: cream to pale yellow-brown	
	Id		Smooth surface. Colour: homogenous, beige, or pale brown.	Compact section showing thin concentric layers without radiations. Colour: beige, or pale brown.	
	Ie		Locally budding, mamillary or rough surface. Colour: often heterogeneous, pale yellow-brown to brown.	Locally unorganized section or radiating structure. Colour: often heterogeneous, pale yellow-brown to brown	
II	Ila	Weddellite (COD)	Spiculated surface showing aggregated bipyramidal crystals with sharp angles and edges. Colour: pale yellow-brown	Loose radial crystallization. Colour: pale yellow-brown	Hypercalciuria ± hyperoxaluria ± hypocitraturia Hypercalciuria, stasis and confined multiple stones
	Ilb	Weddellite ± whewellite	Smooth, long bipyramidal crystals, resembling small desert roses Colour: pale yellow-brown	Compact unorganized crystallization. Colour: pale brown-yellow.	
	Ilc	Weddellite	Rough Colour: grey-beige to pale brown	Unorganized core with diffuse concentric structure in periphery Colour: grey-beige to pale brown	
III	IIIa	Uric acid anhydrous	Homogeneous smooth surface. Colour: typically orange	Concentric layers with a radiating organization around a well-defined nucleus. Colour: typically orange	Stasis conditions with low urine pH (e.g., prostatic adenoma hyperplasia) Metabolic syndrome, diabetes Hyperuricosuria and alkaline urine (therapeutic alkalization or UTI) Hyperuricosuria and chronic diarrhoea
	IIIb	Uric acid dihydrate (± anhydrous)	Embossed, rough and porous. Colour: heterogeneous, beige to brown-orange	Poorly organized, porous. Colour: orange	
	IIIc	Urate salts, including ammonium hydrogen urate	Homogeneous or slightly heterogeneous, rough, and locally porous surface Colour: homogenous beige to greyish	Unorganized porous section Colour: whitish to greyish	
	III d	Ammonium hydrogen urate	Heterogeneous, embossed, rough and porous surface. Colour: greyish to brown	Alternated layers, thick and brownish or thin and greyish, locally porous Colour: greyish	
IV	IVa1	Carbapatite	Rough and homogenous. Colour: whitish to beige	Poorly organized or diffuse concentric layers. Colour: whitish to beige	Hypercalciuria, UTI

Table 1 (continued)

Type	Main crystal component	Main morphological stone characteristics		Common causes
		Surface	Section	
IVa2	Carbapatite	Embossed and varnished surface with small cracks. Glazed appearance. Colour: homogeneous, pale brown-yellow to pale brown	Alternated layers, thick brown-yellow and thin beige. Often multiple nuclei (from collecting duct origin)	Distal renal tubular acidosis
IVb	Carbapatite + struvite	Heterogeneous, both embossed and rough. Colour: heterogeneous, cream to dark brown	Alternate thick whitish and thin brown-yellow layers	UTI, hypercalciuria. Aetiology depends on minor components identified in the stone
IVc	Struvite	Aggregates of large crystals with blunt angles and edges. Colour: whitish	Diffuse, loose radial crystallization. Colour: whitish	UTI by urease-splitting bacteria
IVd	Brushite	Finely rough or dappled surface. Colour: whitish to beige	Radial crystallization with locally concentric layers. Colour: whitish to beige	Hypercalciuria, Primary hyperparathyroidism, phosphate leak
V	Va	Cystine	Homogeneous, rough surface. Waxy aspect. Colour: yellowish	Cystinuria
	Vb		Homogeneous smooth or finely rough surface. Colour: whitish to pale beige	Cystinuria + inadequate therapy
VI	Vla	Proteins	Matrix soft calculi, homogeneous surface. Colour: cream to pale brown	Chronic pyelonephritis
	Vlb	Proteins and drugs or metabolic compounds	Heterogeneous, irregularly rough surface. Locally scaled. Colour: dark brown to black. Other components often present in these stones may alter the structure and the colour	Proteinuria, drugs, clots
	Vlc	Proteins and whewellite	Homogeneous, smooth surface with clefts and scales. Colour: dark brown	End-stage renal failure and excessive calcium + vitamin D supplementation
VII	Miscellaneous	Various morphologies and colours according to the stone composition (infrequent purines and drugs)	Variable organization and colour according to the stone composition xanthine stones: xanthine oxidase deficiency; dihydroxyadenine stones: adenine phosphoribosyl transferase defect; and drug-containing stones: phenazopyridine, oxypurinol, silica, and calcite stones	

COD, calcium oxalate dihydrate; COM, calcium oxalate monohydrate.

described above, thereby mitigating any operator bias towards specific areas of the urinary calculus and preserving as much variance in the samples as possible.

Also, to avoid any bias, patches were extracted in such a way that they have a maximum border overlap of 20 pixels to

limit redundant information. Moreover, patches including a high number of 'non-stone' pixels were not included in the dataset (an experimentally set threshold value of 10% was used to discard inappropriate patches located close to the fragment periphery or those including instruments). Thus,

Table 2 Number of used stones, images and patches extracted depending on their type.

Morpho-constitutional classification	Group I	Group II	Group III	Group IV	Group IV	Group V	Total
Crystalyn component	Whewellite (COM)	Weddelite (COD)	Uric acid	Brushite (calcium phosphate dihydrate)	Struvite (magnesium ammonium phosphate hexahydrate)	Cystine	
(a) Type and number of stones used <i>ex vivo</i>							
Number	28	5	25	8	17	25	109
Number according to morpho-constitutional classification	Ia: 21 Ib: 4 Ic: 2 Id: 1	IIa: 2 IIb: 3	IIIa: 16 IIIb: 9	IVd: 8	IVb: 1 IVc: 16	Va: 24 Vb: 1	109
	Class	Images number	Presence (%)	Patches number			
(b) Number of images and patches							
Surface	COM	62	22.30	5614			
	COD	43	14.45	2642			
	Brushite	25*	9.00	2095			
	Uric acid	58	20.85	2185			
	Cystine	47	16.90	2058			
	Struvite	43	14.45	4237			
	Total	278	100.0	18831			
Section	COM	25	11.70	2260			
	COD	47	21.95	2355			
	Brushite	20*	9.35	2668			
	Uric acid	50	23.35	2837			
	Cystine	48	22.45	2695			
	Struvite	24	11.20	2048			
	Total	214	100.00	14863			
*To balance the different classes, 14 section images and 14 surface images were captured in <i>in vivo</i> procedures for the brushite type and were added to the database. COD, calcium oxalate dihydrate; COM, calcium oxalate monohydrate.							

more than 35 000 patches were obtained to construct the final dataset. Of these samples, 80% were used to train the machine-learning algorithm and for validation purposes, while 20% of the samples were used for the test dataset.

Recognition Methods (Stone Classification)

Two types of recognition approaches were studied: a shallow method and a deep-learning method. Shallow methods rely on the extraction of handcrafted statistical characteristics or feature vectors used to encode colour and texture information characteristics such as those the human eye could perceive.

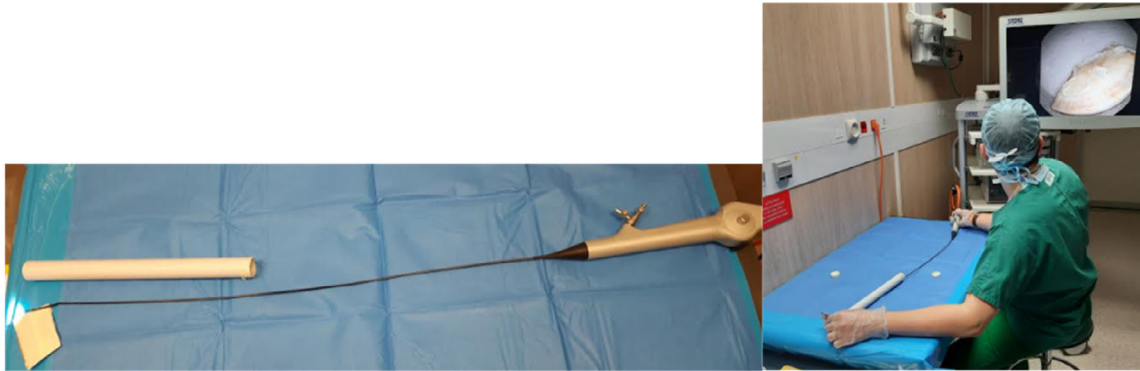
After a careful feature sensitivity analysis [13], a vector whose components encode hue, saturation, value of lightness (HSV) and local binary pattern (LBP) features (representing local textures) have been identified as those that best discriminate urinary stones. The colour information was represented in the HSI space, where H, S and I stand for hue (the tint of the colour), saturation (the amount of grey in the colour) and intensity (the brightness of the colour), respectively. With this colour space the advantage is that when only the colour intensity varies (as is the case when the endoscope’s viewpoint changes) the hue values remain constant (i.e., the

unchanged colour aspect favours the stone classification). LBPs encode textures in a binary code.

It must be emphasized that we have investigated the performance of other machine-learning methods in previous works dealing with *in vivo* data [12,13]. Our main goal in comparing shallow and deep-learning methods in these works was to show that effective feature extraction techniques can attain very competitive results compared to previous state-of-the-art solutions. (Serrat *et al.* [17] followed a similar approach, but their feature extraction phase was suboptimal). Thus, in a previous study [12], we demonstrated that a much better classification can be reached using a random forest classifier due to an improved feature selection process. The rationale for using and only focusing on XGBoost in this study was that it has been reliably demonstrated to be superior to most of the machine-learning methods in the literature. In fact, it is considered one the best off-the-shelf machine-learning methods when working on small datasets, a scenario where deep-learning methods will not perform as well. The improvements obtained by XGBoost compared to the second-best shallow machine-learning method (random forest) are of approximately 5%, increasing from 91% to 96% of PPV in our previous study [12] and from 87% to 91% with the dataset used in the present study, and by more than

Fig. 1 Image acquisition technique and examples of acquired images. **(A)** Device (left) and installation (right) during image acquisition. **(B)** Examples of calculi images used. COM, calcium oxalate monohydrate; COD, calcium oxalate dihydrate.

(A)



(B)

Calcium Oxalate		Brushite
Monohydrate (COM) = Whewellite	Dihydrate (COD) = Weddellite	
Uric acid	Cystine	Struvite

10% compared with methods based on Support Vector Machines.

The second method assessed in this study corresponds to a deep-learning approach exploiting features automatically learned during the training process of convolutional neural networks (CNNs; in contrast to the predefined feature vectors used for traditional machine-learning methods). In such an approach, a CNN is trained to learn and select the best features to be extracted from the images to maximize classification performance, without guiding or influencing the recognition by *a priori* knowledge.

The amount of information extracted is colossal, far exceeding that of shallow methods, and understanding the physical meaning of the features is not always possible. Their ability to establish relationships between these characteristics explains the strength of deep-learning methods, even if some of the information extracted is not significant for the recognition. For the reasons described above, we have made use of modern visualization techniques to compare the performance of the two methods; these visualizations can also help in understanding the decisions made by the algorithms when classifying a kidney stone given only visual data.

An important contribution of this work is that it compares shallow methods that are inherently 'interpretable' with deep-learning algorithms whose performance might be superior at the expense of reduced explainability of the results. In the future, we will explore some recent advances in the field of Explainable AI (XAI, [22]) to contrast our results with the morpho-constitutional analysis introduced by Daudon *et al.* [2].

Qualitative Assessment

It was possible to represent the class separability of stones in a three-dimensional space as explained below (Fig. 2). Each image was divided into patches. For each patch, a feature vector was extracted, either by a shallow method or by a deep-learning method. All the extracted features could then be reduced to a set of three 'main' features, shown as umap1, umap2 and umap3 in Fig. 2. This representation method is known as 'uniform manifold approximation and projection for dimension reduction' (UMAP) [14]. These main components were calculated from more complex features extracted from the image patches; the original features were vectors of very high dimension (40 elements for shallow models and 1024 elements for deep models) so UMAP tries to find a projection that best represents all the information of a patch in a compact way using three values: umap1, umap2 and umap3.

A point with coordinates (umap1-umap2-umap3) in the three-dimensional space represents a patch. Each type (*i.e.*, class) of urinary calculus was represented by a given colour

and the dots corresponding to the patches of each class form 'clouds' or clusters. The more compact the clouds were, and the more distant each cloud was from other clouds, the more discriminating the features extracted from the images are. On the other hand, the more the clouds overlapped in three-dimensional space, the lower the ability of the classification to correctly identify stones. After learning the classes visualized (Fig. 2), a new image (or patch) of urinary calculus, was treated as follows by the classification method. The features that make it possible to identify the urinary calculus were first extracted from the patches. Then, these characteristics were used to identify the class with the greatest resemblance. In other words, the closest point cloud, or one that included the point whose characteristics were extracted, was identified. This identification made it possible to recognize the stone type. In addition, for deep learning, the image class was recognized by a CNN.

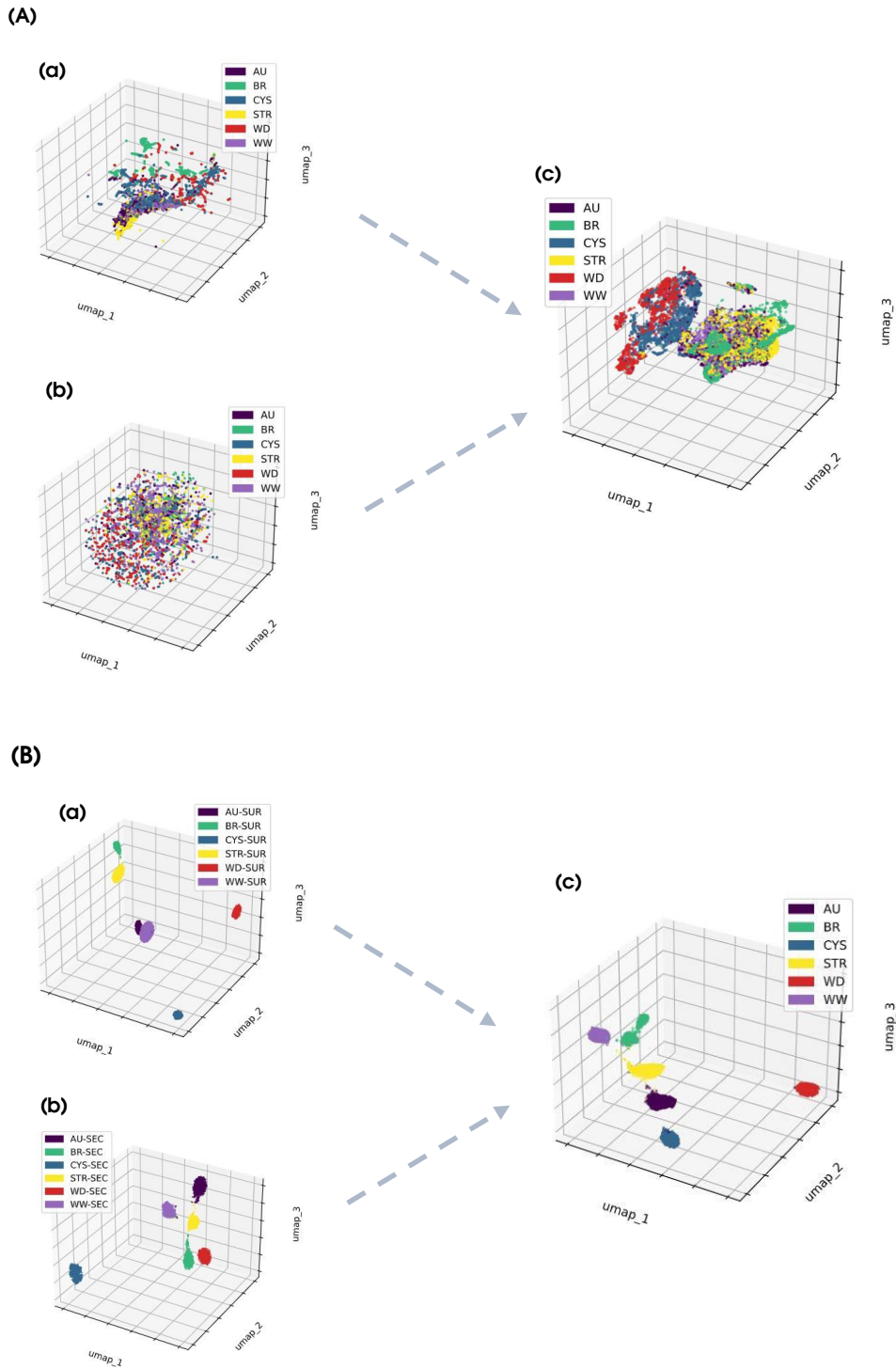
For the shallow approaches, the characteristic extraction method was the same regardless of the classifier (or the method of using the scatterplot to recognize a stone) as it represented a separate step. The classifier was used to name the identification method and used these data to the hyperplane that best separated the points of Fig. 2A. In this study, the classifier used for the shallow methods was XGBoost [15]. Readers interested in the technical details of the training and operation of the XGBoost classifier should refer to Appendix A of this paper.

For deep-learning methods, the feature extraction method was also specific to each type of neural network (referred to as neural network architecture). These methods were therefore distinguished both by the way they extract information from images or patches and by the way they used point clouds for classification. In this study, we present the results of a deep-learning method called 'Inception v3' [16]. The transfer-learning method and the inception v3 network architecture used in this study are described in detail in Appendix B.

Quantitative Assessment

Four complementary quality criteria were used to assess the performance of the classification methods: (1) sensitivity, which relates to the ability of the classifier not to miss a stone in a given class; (2) specificity, which refers to the ability to avoid over-detection in a given class; (3) positive predictive value (PPV), which represents the probability that a stone belongs to the recognized class; and (4) area under curve (AUC). The AUC values were determined using receiver-operating characteristic curves, which show the relationship between true-positive identification rates and false-positive identification rates. AUC values were individually determined for each class and for a given classification method. They

Fig. 2 Scatterplot using feature extractions. Each cloud represents a type of stone after classification by 'shallow methods' **(A)** and by the 'deep-learning method' **(B)**. For each method, the results are shown for surface patches **(a)**, section patches **(b)** and the effect of mixing both of them **(c)**. The more distant each cloud from other clouds, the more discriminating the features extracted from the images were. Using the deep-learning method **(B)**, clouds are more distant from each other, offering higher discrimination than that achieved using the shallow method. AU, uric acid (dark purple); BR, brushite (green); CYS, cystine (blue); STR, struvite (yellow); WD, weddellite (calcium oxalate dihydrate [COD]; red); WW whewellite (calcium oxalate monohydrate [COM]; light purple); Sur, surface; Sec, section. Hsl, hue, saturation intensity colour model; LBP, local binary patterns. **(a)** Scatterplot for individual surface patches. **(b)** Scatterplot for individual section patches. **(c)** Scatterplot obtained using both types of patches.



were also globally computed for all classes (weighted AUC across the classes) and a given classification method.

Evaluating a model based on both sensitivity and specificity is appropriate for most datasets because these measures consider all entries in the confusion matrix. Sensitivity relates to true-positive and false-negative rates, while specificity is determined using false positives and true negatives. The combined use of sensitivity and specificity therefore leads to a holistic measure in which both true positives and true negatives are considered. The AUC values given globally over all classes and for each individual class highlight respectively the discriminatory capabilities of the classifiers between a set of classes (weighted AUC over the classes), and the recognition power inside individual classes (individual class AUC).

Results

Qualitative Assessment

Shallow Methods

Figure 2A shows the different classes of urinary stones according to the UMAP method explained above, which uses texture features (LBP) and colour features (HSV) extracted in a controlled way and which has a physical meaning.

For individual surface or section data, point clouds intertwine and overlap. Automatic discrimination by a classifier is therefore complicated. By simultaneously using the surface and section images, point clouds corresponding to the classes 'weddellite' and 'cystine' become more discriminative (i.e., in the subplots in Fig. 2Ac and Bc, the two classes or clusters move apart in the three-dimensional space). The point clouds of the other classes remain intertwined and superimposed, offering little discrimination.

Deep Learning

Figure 2B represents the different classes of urinary stones whose cloud points were obtained using a feature extraction with the Inception v3 deep-learning method.

For individual surface or section images, each type of stone has its own more or less compact point cloud, these clusters being also more or less distant from each other in three-dimensional space. For surface patches (see Fig. 2Ba), however, the point clouds of uric acid and whewellite are touching each other, which does not facilitate their separation during classification. Combined use of surface and section images (Fig. 2Bc) leads to more compact and spaced point clouds, offering better discrimination (see, in particular, the classes of uric acid and whewellite that are no longer attached).

Quantitative Assessment

Table 3 shows the results of the classification of stones obtained with the shallow algorithm (XGBoost) and the deep-learning (Inception v3) methods, using both surface and section patches.

With the shallow method, the sensitivity, specificity and PPV are: 91%, 90% and 89%, respectively, for whewellite; 99%, 98% and 99% for weddellite; 88%, 89% and 88% for uric acid; 91%, 89% and 90% for struvite; 99%, 99% and 99% for cystine; and 94%, 98% and 99% for brushite. As shown in Table 3 for the XGBoost classifier, the individual AUC values range in interval (90% to 99%), while the corresponding weighted AUC value is 93%.

For the deep-learning approach, the sensitivity, specificity and PPV are: 99%, 98% and 97%, respectively for whewellite; 98%, 98% and 98% for weddellite; 97%, 98% and 98% for uric acid; 97%, 97% and 96% for struvite; 99%, 99% and 99% for cystine; and 94%, 97% and 98% for brushite. For the Inception V3 classifier, the individual AUC values range from 96% to 99%, with the weighted AUC value being 98%.

Classification tests can be performed at the github link provided in this paper.¹

Discussion

The peri-operative morphological recognition of stones is highly challenging. Estrade *et al.* and CLAFU [5,9] (*Comité Lithiase de l'Association Française d'Urologie* [Lithiasis

Table 3 Results obtained with shallow methods and deep-learning methods of classification using combined section and surface images, by urinary stone type.

Stone type	Classic methods				Deep learning			
	XGBoost				Inception V3			
	Sensitivity	Specificity	PPV	AUC	Sensitivity	Specificity	PPV	AUC
Whewellite	91%	90%	89%	92%	99%	98%	97%	98%
Weddellite	99%	98%	99%	98%	98%	98%	98%	97%
Uric acid	88%	89%	88%	89%	97%	98%	98%	97%
Struvite	91%	89%	90%	90%	97%	97%	96%	96%
Cystine	99%	99%	99%	99%	99%	99%	99%	99%
Brushite	94%	98%	99%	95%	94%	97%	98%	97%

AUC, Area under curve; PPV, positive predictive value.

Committee of the French Association of Urology)) have established didactic boards of confirmed endoscopic images for pure (I to VI) and mixed stones (IIb + Ia), (IIb + IVa1)c, (IIb + IVa1)i, IIIab + Ia.

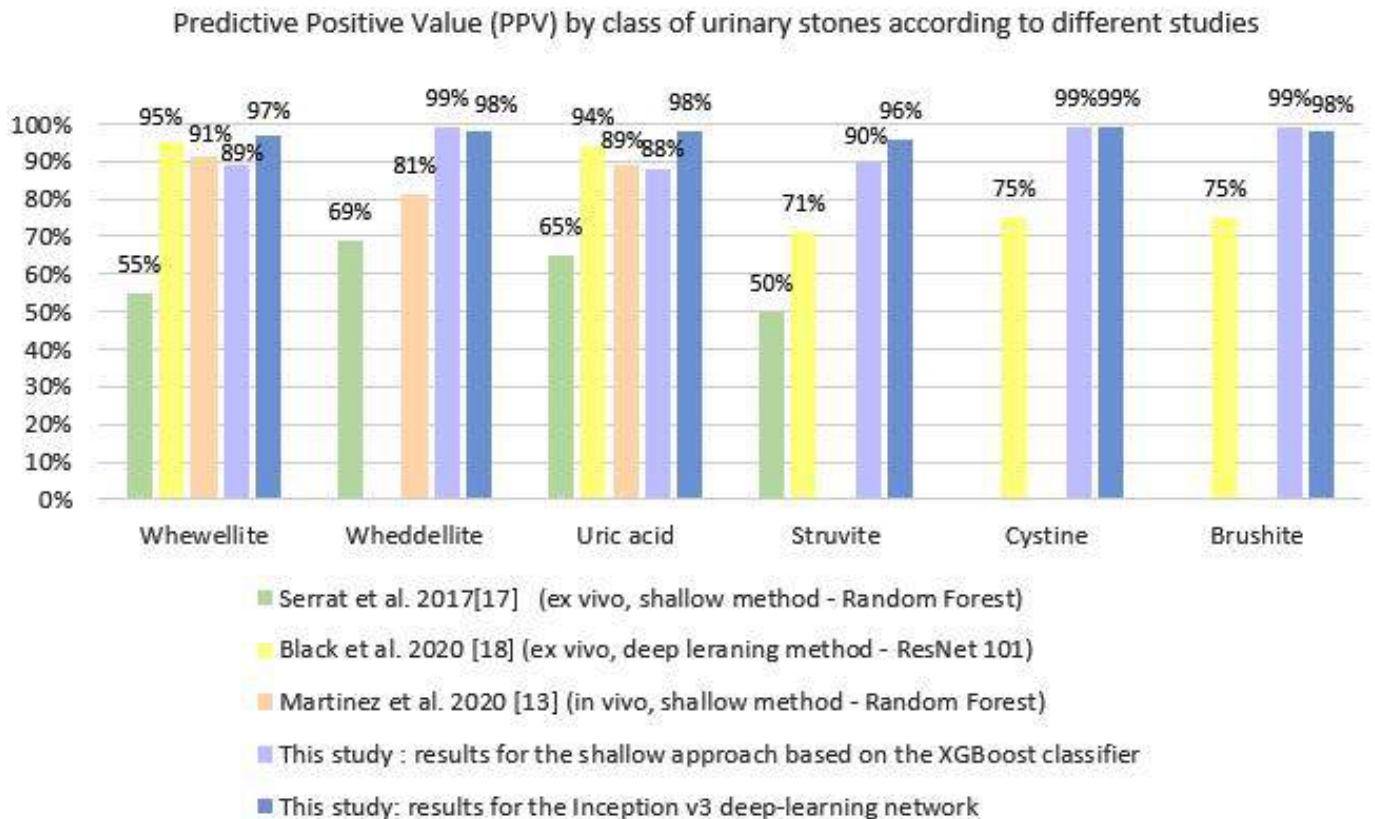
Furthermore, Estrade et al. published a comparison between *in vivo* endoscopic analysis of urinary stones by a urologist and *ex vivo* morphological analysis of these stones by a biologist according to the morpho-constitutional classification of Daudon et al. [2]. Good agreement between the results for the two recognition methods was established for the calculi of whewellite (Ia or Ib), weddellite (IIa or IIb), uric acid (IIIa or IIIb), carapatite-struvite association (IVb), and brushite stones (IVd). Endoscopic stone recognition is therefore visually possible when performed by a trained urologist.

In this study, we have shown that endoscopic stone recognition by artificial intelligence is possible for images acquired in *ex vivo* conditions. The proposed recognition methods use the morphological (colour and texture) aspect of the surface and the section, which is similar to the way in which morphological analysis is carried out by the biologist. Furthermore, the results are optimized by combining the surface view and the section view.

Figure 3 shows the results of other studies that have evaluated the automated recognition of urinary stones according to the morphology of the stone surface and section. Serrat et al. [17] studied an identification method based on 453 calculi from four classes, with *ex vivo* acquired images, analysed using a shallow classification method known as random forest. Black et al. [18] examined 63 stones of five classes of urinary calculi *ex vivo*, images of which were processed by a deep-learning method (ResNet101). Martinez et al. [13] used *in vivo* images of 125 calculi for three types of stone. The results obtained with the shallow classification method used in our study are consistent with these previous studies (Fig. 3), and sometimes even better, especially for struvite, cystine and brushite. Using the proposed deep-learning approach, precision is increased, especially for whewellite for which the PPV increases from 89% (shallow method) to 97% (deep learning). For brushite and weddellite, the shallow method offered a slightly better result than deep learning (99% and 98%, respectively).

Furthermore, Serrat et al. [17] and Black et al. [18] studied *ex vivo* images of urinary stones using devices allowing highly controlled acquisition conditions and conventional cameras not associated with an endoscope. In this study, images were obtained using a digital flexible ureteroscope, manipulated by

Fig. 3 Review of the literature.



human hand. Although motion artefacts were avoided to the maximum, they were still possible and impacted the quality of the images. However, these images remained exploitable. Two different ureteroscopes were used. While the image quality was not equal, the data were usable for automated recognition.

In addition, we divided the images into patches whose size was previously studied to optimize the ratio of information gain to amount of kidney stone data [13].

Also worthy of consideration is whether or not the proposed feature extraction and machine-learning methods would be generalizable to different acquisition conditions or to images acquired with different instruments and to cope with the increased number of classes in this study. In fact, in previous work [12,13] we have demonstrated that the results hold for images despite the use of different ureteroscopes in *in vivo* conditions, even with varying image resolutions, albeit only for three classes of kidney stones. In order to study the robustness of our results, we carried out extensive experiments in which we blurred the acquired images with different filters and trained the same models used in this study with these 'colour-distorted' images (these image changes simulate the differences between endoscopes to an extreme extent); the deep-learning models are indeed capable of extracting discriminating features even with large levels of blurring, at the cost of a slight loss in terms of PPV, sensitivity and specificity (a 3% loss on average for deep-learning models).

Deep-learning methods can extract more discriminating features and find more complex relationships among these features, yielding superior classifications results. Therefore, as with any deep-learning-based approach, in order to cope with different conditions, the models must be trained with images acquired under different operational conditions, which we consider a promising future avenue of research.

Nonetheless, this study has several limitations. The images were acquired in *ex vivo* conditions (the lighting conditions were not the same as those in the urinary tract). In *in vivo* conditions stones appear brighter. However, the main advantage of using *ex vivo* data is that the stone type is known since for each acquired fragment a morpho-constitutional analysis was available. The results obtained show the ability to recognize stone type (subtypes were not considered in this study). Moreover, as shown in a preliminary study [12], the same classification algorithms can be trained for *in vivo* data, the ground truth for the training being given by an expert who performed a visual classification based on images visualized and acquired during ureteroscopies. This expert (Vincent Estrade MD [7]) is among the few urologists able to perform a visual classification approaching the results given by a morpho-constitutional analysis. A similar high performance in terms

of kidney stone recognition (comparable criterion values as in Table 3) was also achieved by the algorithms described in this paper for these *in vivo* data. However, it is worth noting that, in this paper, the ground truth is provided by the widespread morpho-constitutional analysis, which remains undeniably the reference method. This justifies performing the study on *ex vivo* data to validate the classifications methods (the idea was to validate the classification methods on medically recognized ground truths, and then to adapt them later to *in vivo* conditions). The use of many patches extracted from few urinary stones can induce a favourable bias in the recognition of these urolithiasis and increase the corresponding PPV. It should also be noted that, for the brushite type, *in vivo* and *ex vivo* images were used in combination to train and test the classifiers since only few *ex vivo* data were available (the patch number of this class was increased using *in vivo* data). Mixing both image types increases the intra-class variability of the brushite kidney stones and makes the situation more complicated for the classification algorithm. The fact that for the brushite class the recognition performances remained very high (sensitivity 94%, specificity 98% and PPV 99% for the XGBoost approach and sensitivity 94%, specificity 97% and PPV 98% for the deep-learning Inception v3 algorithm) is an indication that the algorithm is robust and can deal with high intra-class variability. It is also notable that the AUC values given in Table 3 for individual classes are systematically high for both classification approaches. This result confirms that the capabilities of the classifiers to discriminate between the set of the six classes tested is high.

Acquisition of the images took place in air and not in liquid medium. The kidney stones have been kept for several years under dry conditions that may affect their morphological characteristics. For some calculi, these artefacts were clearly identifiable and were eliminated from image patches. Indeed, it is possible that the appearance of a stone will change after exposure to air. This study aims to understand the method and feasibility of an automated recognition system. To apply it in real endoscopy, it will be necessary to create a new database with images acquired *in vivo* with ureteroscopes. The urinary stones studied are considered pure, that is, they consist of at least 85% of a single constituent. However, it has been shown that almost half of the stones have mixed morphologies and consist of several different crystal components [3,4].

We used digital flexible reusable ureteroscopes in this study, the quality of whose images is better than that of other ureteroscopes commonly used, such as single-use flexible ureteroscopes, non-digital ureteroscopes, or even rigid ureteroscopes.

Furthermore, deep learning, by its nature, is uncontrolled. We do not yet understand what type of data it selects, or how it selects them.

In conclusion, automated endoscopic recognition of urinary stones would allow for reliable stone recognition during ureteroscopies independently of the urologist's level of expertise.

This study represents an important first step towards peri-operative stone recognition, which would make it possible to quickly offer patients targeted therapies or adequate metabolic investigations according to the type of urinary stone to prevent recurrence.

The laser energy required to spray stones may vary according to their composition [19]. We could imagine a system allowing first for automated stone recognition and then for the parameters of the laser to be adapted to target these to the recognized class in order to optimize spraying of the stone.

This preliminary study shows that automatic recognition of urinary stones can be used with ureteroscopes with good results. The deep-learning recognition method seems more accurate than shallow methods. The deep-learning method is more efficient as the database is important. Further *ex vivo* and *in vivo* studies on all types of calculi (pure and mixed) and in large numbers are necessary before a reliable model for a rapid and peri-operative automated recognition can be established.

In the future, we expect to see the deployment of a system that can recognize a stone during surgery, but our work remains for the moment a proof of concept. The aim is that the urologist takes pictures of the stone during the endoscopy, then the stone type is recognized in a few milliseconds directly intra-operatively via software installed on the endoscopy trolley. This could allow the urologist to recommend lifestyle and dietary measures or even medication as soon as the patient leaves the operating room according to the type of stone (e.g., to alkalinize the urine in case of a uric acid stone, or to extend antibiotic therapy for struvite stones). To facilitate accurate classification, the urologist can acquire the calculi from different points of view (for visual classification the kidney stone is also observed from various viewpoints). Indeed, several acquisitions of the same stone must lead to an identical recognition result (type). The clinician has an important role to play because the urologist must learn how to guide the image acquisition so that the algorithm is placed in realistic conditions, with images that can be used to recognize a stone.

Finally, it should be remembered that the morphological analysis of kidney stones is only one of the elements of the analysis of lithiasis diseases. Indeed, many other elements must be considered such as the patient's clinical data (weight, history, treatments taken, etc.), biological data (metabolic evaluation including a blood and urine analysis), and finally the analysis obtained from spectrophotometry of the stone.

ACKNOWLEDGEMENTS

We thank V. Estrade for the brushite images and the Karl Storz company for providing the flexible ureteroscopes used in this study during image acquisition and database construction.

Disclosures of Interest

None declared.

Endnote

¹<https://github.com/ML-INSIDE/KS-recognition>

References

- Haymann J-P, Daudon M, Normand M et al. First-line screening guidelines for renal stone disease patients: A CLAFU update. [Article in French]. *Prog Urol* 2014; 24: 9–12.
- Daudon M, Bader CA, Jungers P. Urinary calculi: Review of classification methods and correlations with aetiology. *Scanning Microsc* 1993; 7: 1081–104. discussion 1104–1106
- Corrales M, Doizi S, Barghouthy Y, Traxer O, Daudon M. Classification of stones according to Michel Daudon: A narrative review. *Eur Urol Focus* 2021; 7: 13–21.
- Cloutier J, Villa L, Traxer O, Daudon M. Kidney stone analysis: "Give me your stone, I will tell you who you are!". *World J Urol* 2015; 33: 157–69.
- Estrade V, Daudon M, Traxer O, Meria P. Why should urologist recognize urinary stone and how? The basis of endoscopic recognition. [Article in French]. *Prog Urol* 2017; 27: F26–35.
- Assimos D, Krambeck A, Miller NL et al. Surgical Management of Stones: American urological association/endourological society guideline. *PART I J Urol* 2016; 196: 1153–60.
- Estrade V, Denis de Senneville B, Meria P et al. Towards improved endoscopic examination of urinary stones: A concordance study between endoscopic digital pictures vs microscopy. *BJU Int* 2021; 128: 319–30.
- Keller EX, De Coninck V, Doizi S, Daudon M, Traxer O. Thulium fibre laser: Ready to dust all urinary stone composition types? *World J Urol* 2021; 39: 1693–1698.
- Estrade V, Daudon M, Molimard B, Traxer O. Endoscopic recognition of kidney stones: Validation of the first intraoperative images. [Article in French]. *Prog Urol* 2016; 26: 685
- Castiglione V, Jouret F, Bruyère O et al. Epidemiology of urolithiasis in Belgium on the basis of a morpho-constitutional classification. [Article in French]. *Nephrol Ther* 2015; 11: 42–9.
- Grosjean R, Sauer B, Guerra RM et al. Characterization of human renal stones with MDCT: Advantage of dual energy and limitations due to respiratory motion. *AJR Am J Roentgenol* 2008; 190: 720–8.
- Lopez F, Varela A, Hinojosa O, Mendez M, Trinh D-H, ElBeze J, et al.. Assessing deep learning methods for the identification of kidney stones in endoscopic images. 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), 2021, pp. 2778–2781. <https://doi.org/10.1109/EMBC46164.2021.9630211>
- Martinez A, Trinh D-H, El Beze J, Hubert J, Eschwege P, Estrade V, et al. Towards an automated classification method for ureteroscopic kidney stone images using ensemble learning. 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), 2020, pp. 1936–39. <https://doi.org/10.1109/EMBC44109.2020.9176121>

- 14 McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv180203426 Cs Stat* 2020
- 15 Chen T, Guestrin C. *XGBoost: A Scalable Tree Boosting System*. Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. New York, NY, USA: Association for Computing Machinery, 2016: 785–94. <https://doi.org/10.1145/2939672.2939785>
- 16 Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the Inception Architecture for Computer Vision. 2016 IEEE Conf. Comput. Vis. Pattern Recognit, 2016: 2818–26.
- 17 Serrat J, Lumberras F, Blanco F, Valiente M, López-Mesas M. myStone: A system for automatic kidney stone classification. *Expert Syst Appl* 2017; 89: 41–51.
- 18 Black KM, Law H, Aldoukhi A, Deng J, Ghani KR. Deep learning computer vision algorithm for detecting kidney stone composition: CNNs to predict kidney stone composition. *BJU Int* 2020; 125: 920–4.
- 19 Vassar GJ, Teichman JM, Glickman RD. Holmium:YAG lithotripsy efficiency varies with energy density. *J Urol* 1998; 160: 471–6.
- 20 Pedregosa F, Varoquaux G, Gramfort A et al. Scikit-learn: Machine learning in python. *J Mach Learn Res* 2011; 12: 2825–30.
- 21 Piccialli F, Somma VD, Giampaolo F, Cuomo S, Fortino G. A survey on deep learning in medicine: Why, how and when? *Inf Fusion* 2021; 66: 111–37.
- 22 M. Nauta, A. Jutte, J. Provoost, C. Seifert. This Looks Like That, Because ... Explaining Prototypes for Interpretable Image Recognition. *Arxiv:arXiv:2011.02863*

Appendix A

Training of the XG-Boost classifier and classification test

In order to find the best configuration of the XGBoost model, we performed hyper-parameter tuning and cross validation with 10 splits; we used a combination of a grid and random search using the Scikit-Learn software [20]. For training the model, we followed a stratified k-fold cross-validation approach in order to maximize the amount of data in the testing phase and to mitigate biases. Therefore, we trained and validated the model with 80% of the samples and tested it with 20% of the patches. The best hyper-parameter settings for the XGBoost model using the combined patches were as follows. The base score value was set to 0.5. gmtree was used as booster. The learning rate was set to 0.1, while gamma value was 0. A maximum depth of 3 and 100 estimators were used for the three sets of data. All the tests after the hyper-parameter tuning and validation have been carried out on the hold-out test data (20% of the original patches dataset). The data for training, validation and test data have been split randomly.

Appendix B

Training and validation of the Inception v3 architecture model using transfer learning

The choice of model is not arbitrary: it represents a good balance between algorithm complexity and accuracy performance when compared to other models in the literature.

For training this deep-learning model, we adopted a transfer-learning strategy, due to the relatively moderate amount of available training data. The rationale for using transfer-learning is that some models can be reused when pre-trained with natural images in order to take advantage of some useful patterns (i.e., features) already learned by the model. This strategy further avoids model overfitting to this new domain if combined with data augmentation strategies [21].

For our experiments, we proceeded as follows: the convolution layers of the Inception v3 model were pre-trained using ImageNet. Then, to specialize the model to our task (urinary calculus recognition), the fully connected (FC) layer of the feature extraction backbone was replaced by a custom FC layer consisting of 256 channels. The outputs of this layer are then concatenated with a batch normalization module, followed by a ReLU activation function, another 256 channel FC layer and ends with a softmax layer with six class outputs for yielding the class prediction. The fully connected layers weights were randomly initialized. During the training of the deep-learning model, the weights in the convolutional layers (obtained during the pre-training with ImageNet) were maintained constant, and only the weights in the FC layers were updated. As with the previous model, we used 80% of the patches for training and validation, and 20% of the samples (never seen during training) for testing the model. During training, data augmentation strategies were heavily applied for increasing the number of available samples; the applied transformation consisted of vertical and horizontal flips, perspective distortions, and four affine transformations on the original patches in our dataset.

All the experimental studies reported in this paper made use of Pytorch 1.7.0 and CUDA 10.1. The hyper-parameters such as the learning rates were automatically adjusted using the optimizer provided by Pytorch (Lightning 1.0.2). The learning rates used for training Inception v3 was 0.0006, using ADAM as optimizer and a batch size of 64. All the tests after the validation of the models have been carried out on the hold-out test data. The data for training, validation and test data have been split randomly.

Correspondence: Jonathan El Beze, Department of Urology, CHU Nancy - Brabois, France.

e-mail: yon.elbeze@gmail.com

Abbreviations: AUC, area under the curve; CNN, convolutional neural network; HSV, hue, saturation, value of lightness; LBP, local binary patterns; PPV, predictive positive value; UMAP, uniform manifold approximation and projection.