RESEARCH ARTICLE

On the In Vivo Recognition of Kidney Stones Using Machine Learning

FRANCISCO LOPEZ-TIRO[1,2], (Member, IEEE), **VINCENT ESTRADE**[3], **JACQUES HUBERT**[4],
DANIEL FLORES-ARAIZA[1], **MIGUEL GONZALEZ-MENDOZA**[1],
GILBERTO OCHOA-RUIZ[1], **AND CHRISTIAN DAUL**[2], (Member, IEEE)

[1]School of Engineering and Sciences, Tecnologico de Monterrey, Monterrey, Nuevo Leon 64700, Mexico
[2]Centre de Recherche en Automatique de Nancy, UMR 7039, CNRS, Université de Lorraine, 54511 Vandœuvre-lès-Nancy, France
[3]CHU Pellegrin, 33000 Bordeaux, France
[4]Service d'Urologie de Brabois, CHU Nancy, 54512 Vandœuvre-lès-Nancy, France

Corresponding authors: Gilberto Ochoa-Ruiz (gilberto.ochoa@tec.mx) and Christian Daul (christian.daul@univ-lorraine.fr)

ABSTRACT Determining the type of kidney stones allows urologists to prescribe a treatment to avoid the recurrence of renal lithiasis. An automated in-vivo image-based classification method would be an important step towards an immediate identification of the kidney stone type required as a first phase of the diagnosis. In the literature, it was shown on ex-vivo data (i.e., in very controlled scene and image acquisition conditions) that an automated kidney stone classification is indeed feasible. This pilot study compares the kidney stone recognition performances of six shallow machine learning methods and three deep-learning architectures which were tested with in-vivo images of the four most frequent urinary calculus types acquired with an endoscope during standard ureteroscopies. This contribution details the construction of an in-vivo dataset of endoscopic images with four of the most recurrent classes in clinical practice. It also describes the design, implementation, and results of the classifiers (shallow machine learning and deep learning-based methods) of kidney stones. Even if the best results were obtained by the InceptionV3 architecture (weighted accuracy, precision, recall and F1-Score of $97 \pm 03\%$, $97 \pm 03\%$, $98 \pm 04\%$, and $98 \pm 03\%$, respectively), it is also shown that choosing an appropriate colour space and texture features allows a shallow machine learning method to approach closely the performances of the most promising deep-learning methods (the XGBoost classifier led to weighted accuracy, precision, recall and F1-Score values of $96 \pm 14\%$).

I. INTRODUCTION

Urinary lithiasis refers to the formation of crystalline accretions (kidney stones) from minerals dissolved in urine [1]. Kidney stones form themselves in the kidneys and migrate through the urinary tract (ureters, bladder, etc.). While small kidney stones evacuate naturally and imperceptibly, larger accretions (beyond a few millimeters) often cause severe pain (e.g., due to an obstructed ureter) and must be removed during an ureteroscopy (endoscopy of the upper urinary tract). Numerous developed countries [2], [3] exhibit a

The associate editor coordinating the review of this manuscript and approving it for publication was Jinhua Sheng.

high urinary lithiasis incidence since about 10% of their population is affected at least once by a kidney stone episode. The formation of kidney stones is favoured by various risk factors. Apart from reasons related to genetic inheritance, diet (eating too many fruits, vitamin C, vitamin B6, or animal proteins increases the risk of forming kidney stones), chronic diseases (e.g., diabetes) or an inappropriate lifestyle (e.g., a sedentary lifestyle that leads to a high body mass index) are some of the risk factors for urinary lithiasis. There is a direct relationship between these risk factors and the biochemical composition of the kidney stones [4], [5]. In developed countries, the stone recurrence rate approaches a very high value of 40% [6], [7].

TABLE 1. Simplified kidney stone classification: the frequency of appearance and a description of the crystal morphology are given for each stone type.

Type	Acronym	Occurrences	Crystal shape
Type I = Whewellite	WW	from 15 up to 35%	Dumbbell shape
Type II = Weddellite	WD	from 15 up to 35%	Bipyramidal
Type IIIb = Uric Acid dihydrate	UA	from 2 up to 13 %	Flat rhomboidal
Type IVc = Struvite	STR	from 20 up to 30%	Rectangular Prism
Type IVd = Brushite	BRU	from 5 up to 20%	Elongate narrow
Type V = Cystine	CYS	from 1 up to 3%	Hexagonal plates

TABLE 2. Overview of the kidney stone classes and acquisition conditions in the-state-of-the-art works, as well as for this contribution. Simplified taxonomy: UA: Uric Acid (Anhydrous and Dihydrate), COM: Calcium Oxalate Monohydrate (whewellite, WW), COD: Calcium Oxalate Dihydrate (weddellite, WD), STR: Struvite, CYS: Cystine, BRU: Brushite.

Reference	Kidney stone composition						Image type		Acquisition setup
	UA	WW	WD	STR	CYS	BRU	Surface	Section	
Serrat et al. [13]	✓	✓	✓	✓			✓	✓	Ex-vivo
Torrell et al. [14]	✓	✓	✓	✓		✓	✓		Ex-vivo
Black et al. [15]	✓	✓		✓	✓	✓	✓	✓	Ex-vivo
Martinez et al. [16]	✓	✓	✓				✓	✓	In-vivo
Estrade et al. [17]	✓	✓	✓				✓	✓	In-vivo
Mendez-Ruiz et al. [18]	✓	✓	✓	✓	✓	✓	✓	✓	Ex-vivo and In-vivo
Villalvazo-Avila et al. [19]	✓	✓	✓	✓	✓	✓	✓	✓	Ex-vivo
Flores-Araiza et al. [20]	✓	✓	✓	✓	✓	✓	✓	✓	Ex-vivo

Therefore, identifying the kidney stone types is crucial to avoid relapses [8], [9] through personalized treatments (diet adaptation, surgery, etc.) is considered of utmost importance by many practitioners [10]. For this purpose, several guidelines for visually recognizing some of the more common types of kidney stones (see Figure 1) have been proposed in recent years [11], [12] to be applied in the clinical practice.

The international morpho-constitutional classification of urinary stones includes seven groups denoted by roman numerals going from I to VII. Each group is associated with a specific crystalline type. Groups I to V designate whewellite, weddellite, uric acid dihydrate, calcium and non-calcium phosphates (i.e., brushite) and cystine, respectively. Group VI contains protein rich calculi which are very infrequent and group VII gathers all other kidney stone types. Each group is itself divided into several subgroups to differentiate morphologies and aetiologies for a given crystalline type (the subgroup names are designated by letters which complete the roman numbers). The most recent lithogenic events (i.e., cristal type) are located on the surface view, whereas less recent events are observable on the section view [21].

This contribution focuses on four of the six most recurrent subtypes in clinical practice [10]: whewellite (WW), weddellite (WD), uric acid (UA), and brushite (BRU). Table 1 lists the six most common subtypes of kidney stones, gives their corresponding acronym that will be used throughout the work, and provides the occurrence ranges in percentages of each subtype.

A. CONTEXT AND RECENT TRENDS IN URETEROSCOPY

Kidney stones [22] can be detected using various techniques, such as computed tomography [23], the ultrasound modality or X-ray imaging [24], [25]. Kidney stones can be destroyed using "extracorporeal shock wave lithotripsy" or ultrasound ureterolithotripsy [26]. However, the most widespread kidney stone diagnosis and removal technique is ureteroscopy, which is typically associated with laser lithotripsy [27].

Modern ureteroscopes are flexible endoscopes that illuminate the scene (the inner surfaces of hollow organs) with white light [29], [30]. The short focal length (approximately 6 mm) of the endoscope optics allows for the acquisition of high resolution images. The endoscope's distal tip is close to the observed surface, while the short focal length ensures a rather large angle of view. In the "chip on the tip" technology, the sensor (CCD) matrix is fixed on the distal tip and the electrical signals are transferred to the digital camera located on the proximal tip. A low-frequency (3)-5 Hz) laser is inserted in the endoscope's operating channel and is used to fragment and remove the kidney stones from the urinary tract (the fragmentation process is depicted on Fig. 2). The kidney stone fragments must remain large enough to allow for a morpho-constitutional analysis (MCA, top row of Figure 1) carried out with a microscope and for the subsequent infrared-spectrophotometry analysis [28]. The visual examination under the microscope aims to define the kidney stone surface and section in terms of textures, appearance of the crystals, colours, and morphological particularities to fully assess the possible causes of lithogenesis. On the other hand, infrared-spectrophotometry enables to identify the molecular and crystalline composition of the different areas (layers) of the kidney stone. This full morpho-constitutional analysis allows to identify the type (or class) of the kidney stone and to prescribe appropriate treatments (surgery, diets, etc.) for minimizing recurrences.

However, kidney stone identification using a morpho-constitutional analysis has two major drawbacks. First, in numerous hospitals, the results of this widespread analysis
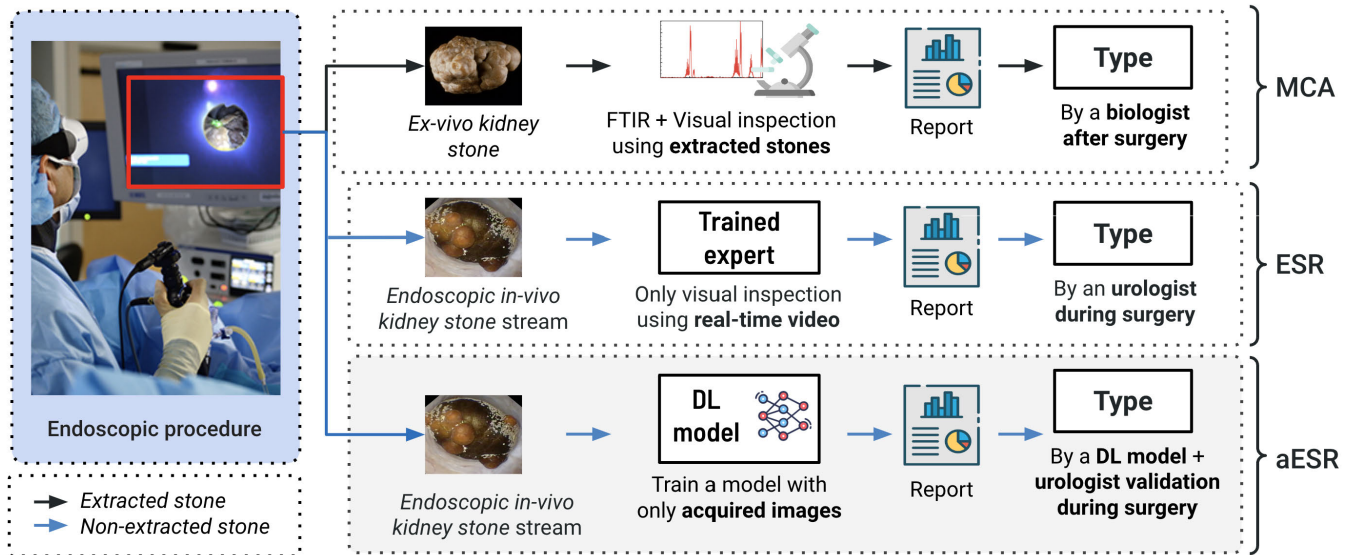
FIGURE 1. Morpho-Constitutional Analysis (MCA, [28]) proposed by Michel Daudon is the standard guideline for the identification of kidney stones (top row). MCA performs a visual inspection and is complemented by a biochemical analysis on extracted kidney stones (post-surgery). Endoscopic Stone Recognition (ESR, [10]) is a technique proposed by Vincent Estrade to perform in-vivo kidney stone identification during surgery using only the information displayed on the screen (middle row). Automatic Endoscopic Stone Recognition (aESR) is a method based on computer vision techniques and machine learning to classify in-vivo endoscope images (bottom row).
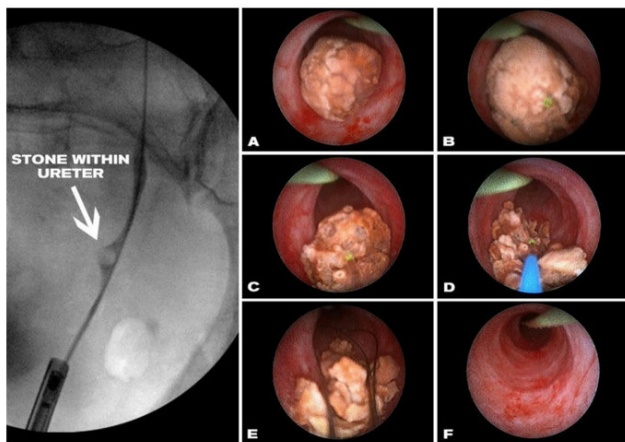
FIGURE 2. Classical ureteral stone removal process. Left: a stent is introduced in the ureter to guide the introduction of the endoscope (fluoroscopy image). A: Complete calculus visualized using a ureteroscope. B: Ureteral stone targeting using a laser (green dot). C and D: Ureteral stone fragmentation. E: Removal of stone fragments with a basket. F: Stone-free ureter.

are often available only after one or two months, even if in critical situations a therapeutic decision should immediately be taken [31], [32]. Second, removing fragments of kidney stones is a tedious and time-consuming task that can last from 30 up to 60 minutes [33], [34]. In this context, the increase in the imaging quality of endoscopes is leading more and more urologists seek to visually identify the morphology (or crystalline type) of kidney stones only with the help of images displayed on a screen. To do so, kidney stones are first fragmented in large parts, and both their external surfaces and their sections are visually analysed.

After their identification, the fragmented kidney stones are either collected or vaporized using a laser lithotripsy technique called "dusting". Fragmentation and dusting can be performed with the same laser, the difference being in the settings of the instrument. In the dusting mode, low energy (0.2-0.5 J) and high frequency (10-20 Hz) are used to vaporize the kidney stones [35].

A visual analysis performed by an urologist can be an appropriate first step for identifying the crystal type of a kidney stone. However, such an analysis requires a great deal of experience due to the high inter-class similarities and intra-class variations of the stones. This visual analysis, which is designated by the term "endoscopic stone recognition" (ESR, [17], middle row of Figure 1), can currently only be achieved by a limited number of specialists, whereas the management of urolithiasis diseases is part of the daily life of every urologist in the world. Moreover, even for experienced specialists, the classification remains often operator-dependent [36], [37]. Therefore, the implementation of automated and reproducible classification methods in this context would make it possible to take full advantage of the dusting technique.

B. PREVIOUS ATTEMPTS OF KIDNEY STONE CLASSIFICATION

Different approaches have been proposed to deal with the classification of kidney stones. These works exploit various image modalities such as hyper-spectral imaging [38], non-contrast computer tomography [23], [39] or multimodal data [40]. Some works exploiting CT images have focused their efforts and resources on automatically detecting the presence of kidney stones and obtaining information about

the calculi [41], [42], [43]. However a second procedure is required after the CT scan to fragment the kidney stones [44]. Thus, although promising results were obtained by these works, data modalities like CT are not used in a standard clinical situation since only endoscopic images are visualized in urology departments. Table 2 gathers the previous works which focus on automatic endoscopic stone recognition (aESR, bottom row of Figure 1) using only images. This table provides also an overview on the used data (urinary calculus classes and acquisition conditions).

The first work [13] dedicated to the classification of kidney stone images used ex-vivo data. In this contribution, collected kidney stones were placed in a closed enclosure that included light sources and cameras whose positions (distances and viewing angles) are optimized to capture large surface and section parts of fragmented kidney stones of known type. Images were acquired using a RGB CMOS 5 megapixel camera by switching between white and infrared light sources. Texture and color information was encoded in feature vectors to describe the kidney stones seen in high-resolution images. These feature vectors consist of RGB colour histograms and texture histograms of local and rotation invariant binary patterns. A Random Forest (RF) classifier was used to recognize the kidney stone type. Various ablation tests were performed to determine the best feature vector (e.g., the most informative features) and the best RF model configuration. Although the average accuracy of this method was rather moderate (see Table 3), the results showed that texture and color information can potentially be discriminant enough to automate the classification of various types of kidney stones.

In a continuation of this precursor work, the classification results were improved in [14] by exploiting a deep learning technique. However, the choice of the neural network was dictated by the limited size of the dataset. Despite this drawback, the encouraging results (an average accuracy of 74%, see Table 3) showed the potential of convolutional neural network (CNN) based methods to tackle this problem.

Another deep learning approach was also used in a recent study [15] to perform an ex-vivo classification. High resolution images were acquired for sixty-three kidney stones of various bio-chemical compositions (see the five stone classes given in Table 3 for [15]) provided by a laboratory performing morpho-constitutional analyses. At least one image was acquired for both the surface and the section of each kidney stone fragment, so that two or more images were available for all samples. As in other previous works, patches (including only kidney stone parts) were extracted from the images and fed to the machine learning algorithm.

In this work, a deep CNN, namely ResNet-101 which was pre-trained with ImageNet to prevent overfitting was used for feature extraction and classification of each patch of the dataset. The obtained model was assessed using the leave-one-out cross validation method, with the primary monitored outcome being the model recall to account for the reduced size of their dataset. While the average precision over five

classes is the best for this contribution, the classification metrics remain rather low for some classes, see the struvite (STR), cystine (CYS), and the brushite (BRU) classes in Table 3.

A more recent effort aiming to improve the classification of kidney stones images using traditional features (color histograms and LBP features for texture descriptors) and shallow methods (a Random Forest classifier) was presented Martinez et al. [16]. Their results showed that a well chosen color space can have a significant impact on the classification accuracy of 3 classes, attaining a 92% weighted average accuracy for kidney stones using features from surface and sections patches, as shown in the fourth row of Table 3.

In [17] has been investigated the applicability of deep CNNs to predict the morphology and composition of both pure and mixed stones. The authors made use of a dataset consisting of 347 images of the surface and section of kidney stones acquired with an Olympus URF-V flexible ureteroscope. However, it has to be noticed that, in contrast to most works in the recent literature, the authors also investigated the performance of machine learning models trained with a database of images including kidney stones with pure crystalline composition and urinary calculi with several layers with different crystalline composition. In order to train and test their models, the authors used image patches with a size of 256×256 pixels and trained two multi-class classification models based on ResNet-152-V2 (one for surface images and another for section images). For kidney stones with a pure crystalline composition, the authors made use of the same dataset as the one used in [16] (see the fifth row of Table 2). The authors in [17] expanded the training dataset using data augmentation to make their models more general.

A cross-validation was repeated ten times with randomly chosen image combinations for the training and testing steps. The full process was also repeated with different random initialization seeds for the deep CNN algorithm. Average standard test metrics were reported for each step (i.e., precision, area under the ROC curve (AUROC), specificity, sensitivity, see Table 4). Compared to other works in the literature [16], the authors in [17] did not investigate the effect of mixing sections and surface features during the network training, which has shown to improve the overall kidney stones recognition task.

The best sensitivity was obtained for the type IIIb (uric acid) using surface images (98% of the IIIb kidney stone images were correctly predicted). The most frequently encountered morphology was the Ia type (pure whewellite composition). It was correctly predicted in 91% and 94% of the cases using surface and section images, respectively.

In a more recent work [18], a novel metric learning approach was proposed for the classification of endoscopic kidney stone images. The authors proposed a method to generalize the classification of kidney stones in two different domains. Their approach exploited two datasets, namely the in-vivo dataset used in [45] and the ex-vivo dataset

TABLE 3. Comparison of the precision obtained by the state-of-the-art-contributions for the most common kidney stone classes. The precision is given for each individual class and classifier. The taxonomy of the stone classes is the same as in Table 2. The average precision (weighted by the image number of each class) are also given for each kidney stone type.

Reference	Precision Per Class						Weighted Precision	ML Method
	AU	WW	WD	STR	CYS	BRU		
Serrat et al. [13]	0.65	0.55	0.69	0.50	N/A	N/A	0.63	Random Forest
Torrell et al. [14]	0.76	0.67	0.80	071	N/A	0.72	0.74	Siamese CNN
Black et al. [15]	0.94	0.95	N/A	0.71	0.75	0.75	0.85	CNN – ResNet101
Martinez et al. [16]	0.91	0.94	0.92	N/A	N/A	N/A	0.92	Random Forest
Estrade et al. [17]	0.99	0.90	0.93	N/A	N/A	N/A	0.94	ResNet152v2
Mendez-Ruiz et al. [18]	N/A	N/A	N/A	N/A	N/A	N/A	0.88	ResNet50
Villalvazo-Avila et al. [19]	N/A	N/A	N/A	N/A	N/A	N/A	0.96	ResNet50
Flores-Araiza et al. [20]	N/A	N/A	N/A	N/A	N/A	N/A	0.85	DenseNet201

TABLE 4. Matched results between endoscopic and microscopic studies. "N/A" stands for not applicable and refers to data which are insufficient for a statistical use. This simplified class representation was adapted from [17].

Component	Stone type	% Correct Matches	Number of occurrences	AUC	Precision (PPV)	Recall (Sensitivity)
Whewellite	Ia or Ib	85%	205	0.87	0.88	0.85
Weddellite	IIa or IIb	85%	178	0.87	0.86	0.85
Uric Acid	IIIa or IIIb	91%	64	0.95	0.94	0.91
Carbapatite	IVa	81%	176	0.86	0.88	0.81
Struvite	IVc	50%	10	0.74	0.45	0.50
Brushite	IVd	65%	23	0.82	0.83	0.65
Cystine	Va	100%	7	N/A	N/A	N/A

described in [46]. The Meta-Learning scheme was based on ResNet50 and was implemented in two steps, i.e., Meta-training and Meta-testing. In the first step, Meta-training is used to sequentially learn the weights of different large datasets (ImageNet, CUB, Crop-Disease, EUROSAT and ISIC) containing thousands of images and classes for the training process. These images have distributions which are not similar to that kidney stone images. Subsequently, in a second step (Meta-testing), the weights obtained from the Meta-training step are fine-tuned and tested with only a few in-vivo or ex-vivo endoscopic images. As can be intuited, Meta-testing aims to fine-tune with the minimum of available endoscopic images and test with the rest of the dataset. The result obtained on complete endoscopic images for the section view is 88% (see the sixth row of Table 3). This work demonstrates that learning weights in a proper way, together with down-sampled learning, is effective in generalizing the classification of whole images in different domains (in-vivo, ex-vivo) with acceptable accuracy on individual views. Although the results are promising, the effect can only be measured on individual views. In addition, this work do not give results for classes taken individually and do not allow for a comparison with other related works.

In [16] it was shown that by mixing color and texture features, more discriminative embeddings could be obtained to train a random forest classifier. This idea was further explored in [19] where the authors implemented a novel multi-view approach based on ResNet50 to extract and combine in an organized way the information from both kidney stone views (surface and section) in order to make mixed embeddings to represent relevant information useful for the classification stage. The dataset of ex-vivo endoscopic

images evaluated in [19] consists of six of the most common classes in clinical practice (types Ia, IIa, IIIa, Iva, IVc, IVd, and Va). The authors in [19] trained two independent branches (one for each view) with a deep network based on ResNet50. For each view, the feature extraction layer is frozen. To improve the features of each model, attention layers were added at different levels of the network. Finally, the output embedding of each model is merged into a single feature embedding for both views. The results obtained by this approach suggest that combining information from different views (surface and section) helps to improve the results compared to single view classifiers. Although the results are promising for the surface, section and mixed views (83.2%, 90.4% and 91.2% of accuracy, respectively), the paper lacks information on the performances for each class taken separately (see the seventh row of Table 3).

AI explainability models [17], [20] have also been proposed in the frame of kidney stone identification. These models provide explanations that match clinical practice. An in-vivo endoscopic image dataset was used in [17] to classify pure kidney stones and mixed compositions (overall precision of 83% and 81% for surface and section views, respectively). The authors in [20] reached a precision of 80.2%, 87.6% and 85.2% for surface, section and mixed (surface + section) views. These results were obtained with a DenseNet201 model tested on the ex-vivo endoscopic dataset described in [46]. Grad-CAM representations [47] have been used to highlight regions of interest on which the model has focused to determine the type of kidney stone. Although this approach has clinical potential, there is still a need to standardize the process to achieve reliable explanations [48].

Two other recent works [49], [50] dealt with the automatic segmentation of kidney stones in video endoscopy. The segmentation task reported in these works consists on spatially localizing three regions, namely the kidney stones themselves, but also the surrounding tissue and the instruments used for the calculi fragmentation (such as the laser fiber). The work in [50], reports promising results since mean Dice Index and Jaccard Index values of up to 79.52% were obtained for the segmentation of the kidney stones and the laser fibers. On the other hand, in the work of [49], the authors report an accuracy of up to 99.56% for the segmentation of the kidney stones and other objects (surrounding tissue, laser fiber or other instruments). Despite the outstanding results obtained by these models, they do not integrate the ability to determine the different types of kidney stones during surgery. However, in future work, these segmentation models could be used to focus the feature extraction on region of interests (i.e., with kidney stone fragments) to improve the calculi recognition.

All these preliminary and promising results described above explain why the medical community in urology is convinced of the interest of kidney stone recognition methods based on artificial intelligence [51], [52] and of the importance of incorporating computer-aided diagnosis tools in their workflow [53]. This work is an extension of a preliminary study [16] which still improved the classification results using a RF classifier. However, the high precision of this last work was obtained for 3 classes only (see Table 3).

C. OBJECTIVES AND STRUCTURE OF THE PAPER

Except in [16] and [17], the kidney stone images used in the existing methods described in Section *"Previous attempts of kidney stone classification"* were acquired in ex-vivo under controlled acquisition conditions (well defined acquisition viewpoints, large and contrasted kidney stone surfaces, high resolution images, and diffuse illumination without reflections on the surfaces). These contributions gave an indication about the feasibility of in vivo kidney stone classification in optimal conditions and could be used to automate and speed-up the morpho-constitutional analysis, even if the described algorithms were probably not conceived with this aim.

In vivo images acquired with flexible uteroscopes are far of being captured in optimal conditions. On the one hand, it is difficult to control the endoscope's position and thus optimal acquisition angles and distances cannot be warranted. Large kidney stone surfaces and sections are difficult to be systematically captured. Moreover, in the vast majority of urology centres, ureteroscopes are equipped by cameras with HD sensor matrices (1024×720 pixels) whose resolution is clearly smaller than 5 mega-pixels used in previous works. Compared to the images used in existing the literature, the acquired kidney stone fragments are much smaller and with lower resolution. On the other hand, images can suffer

from numerous artefacts. For instance, motion blur due to high and non-constant endoscope displacement speeds and defocusing/refocusing due to changing distal tip/kidney stone distances affect globally the quality of numerous images. Moreover, specular reflections are also often visible due to the crystalline nature of the kidney stones and floating objects may occlude some regions of interest. Even if numerous efforts have been made in endoscopy to detect and segment artefacts in various applications (as in urology, gastroscopy or colonoscopy, see [54], [55]), images that are automatically selected in endoscopic videos are not often of optimal quality.

The aim of this contribution is to demonstrate the feasibility of the classification of kidney stones acquired in vivo during ureteroscopic procedures. The first results of a previous contribution [16] indicate that shallow (or classical) machine learning algorithms can improve the results of the literature, even when in vivo images are used. One goal of this contribution is to delve deeper into the feature selection and classifier tuning to improve the recognition results using shallow machine learning methods. This paper also investigates the advantages of using deep-learning methods for the classification of in vivo data and compare their performances to those of shallow machine learning approaches. The paper is organized as follows. Section *"Materials and Methods"* describes the construction of three steps, namely the collection of in vivo images, the optimal patch extraction from the images, and the tested class balancing methods. The first part of Section *"Design of a set of kidney stone recognition methods"* describes the main aspects of shallow machine learning methods (feature extraction, optimal feature selection, and the classifier tuning) for an automated kidney stone recognition. The second part of Section *"Design of a set of kidney stone recognition methods"* deals with the presentation of several deep-learning approaches for kidney stone classification. Section *"Results and Discussion"* details the results obtained by the different solutions described in Section *"Design of a set of kidney stone recognition methods"* and compares them with the results obtained in the literature. A conclusion and perspectives are given in Section *"Conclusions"*.

II. MATERIALS AND METHODS
A. DATABASE CONSTRUCTION

The images used in this contribution (see Table 5) were acquired by an urologist (Dr. Vincent Estrade) who is among the few experts in France able to visually recognize kidney stone types using only in vivo images displayed on a screen during an ureteroscopy. Additionally to this expertise, the annotation of the images used in this work was statistically confirmed in [56] by a concordance study exploiting the morpho-constitutional analysis of extracted kidney stone fragments (the morpho-constitutional analysis confirmed the visual classification made by the urologist), using microscopy and a Fourier transform infrared spectroscopy (FTIR) analysis, which were exploited as follows.

TABLE 5. Number of acquired images and of their (almost) non overlapping square patches. The whewellite (Type Ia), wedellite (Type IIb), uric acid (Type IIIb) and brushite (Type IVb) classes include 57 (30 surface and 27 section) images, 60 (32 + 28) images, 36 (18 + 18) images and 28 (14 + 14) images, respectively.

Image Type		Acquired Images		Number of patches
View	Class	Number	Occurrence (%)	
Surface	WW (Type Ia)	30	31.9	870
	WD (Type IIb)	32	34.1	920
	UA (Type IIIb)	18	19.1	470
	BRU (Type IVd)	14	14.9	420
	Total	**94**	**100.0**	**2680**
Section	WW (Type Ia)	27	31.0	820
	WD (Type IIb)	28	32.2	780
	UA (Type IIIb)	18	20.7	460
	BRU (Type IVd)	14	16.1	410
	Total	**87**	**100.0**	**2470**
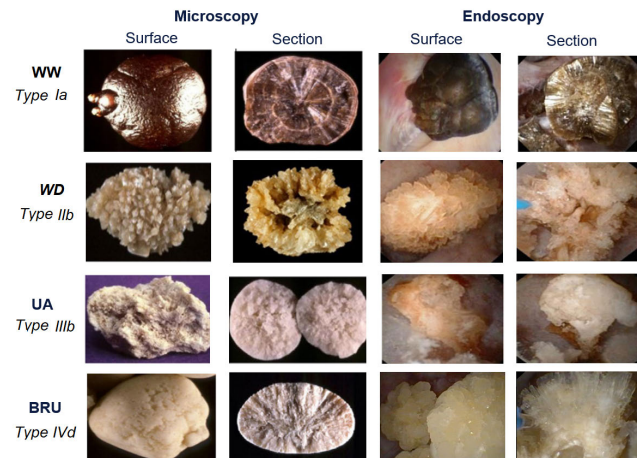
FIGURE 3. Ex-vivo surface (first column) and section (second column) images acquired with a microscope, and in vivo surface (third column) and section (fourth column) images captured with an endoscope. It has to be noticed that the kidney stones are not the same for the two image modalities. When moving from the first to the last line one have successively following classes: WW (Type Ia), WD (Type IIb), uric acid (Type IIIb) and brushite (Type IVd).

The ESR started with a visual observation of the kidney stone surfaces. Then, the kidney stones were split in two fragments using a laser. The Holmium-Yag laser parameters were set as follows: the laser pulse frequency, energy and power were adjusted at 5 Hz, 1.2–1.4 J and 6–7 W, respectively. The pulse sequence length was short and the fibre diameter was either 230 or 270 μm. A second visual observation of the fragment section was then performed. An additional fragmentation session was carried out when needed to allow for the analysis all types of pure and mixed stones.

The goal of the concordance study in [17] was to assess the efficiency of the ESR process. To do so, the fragmented kidney stones were analysed by a biologist (a MD with 40 years of experience) which performed a morphological analysis (visual inspection under a microscope), and a FTIR analysis done with a spectrometer. In a standard morpho-constitutional analysis, the surface, the section and the nucleus of each kidney stone are inspected. In the concor-

dance study in [17], the visual recognition of kidney stone types was considered as "confirmed" when a non-significant difference between the endoscopic and microscopic images (p-value > 0.05) was obtained. The area under the ROC curve (AUROC), the sensitivity, the specificity, the positive predictive value (PPV) and the negative predictive value (NPV) were calculated for a dataset with a confirmed concordance. The authors of [17] regrouped the following pairs of kidney stone morphologies that have similar aetiologies: Ia and Ib, IIa and IIb, IIIa and IIIb, Va and Vb, and VIa and VIb. These six morphological types cover 95% of the most common pure stones that urologists encounter in their daily practice.

Indeed, despite the variety of compositions and morphologies observed in urinary calculi, about 90% are composed of a limited number of crystalline species and morphological characteristics that are easily recognised through an endoscopic examination [10], [12]. The study in [17] showed a high concordance between endoscopic and microscopic typing of the kidney stones. More precisely, the concordance of the results was observed in 86.1% of the type I kidney stones, in 85% of the type II kidney stones, in 91% of the type III kidney stones and in 79% of type IV kidney stones made of calcium phosphate. Table 4 gives more details about this concordance.

A urologist (Vincent Estrade, 20 years of experience) intra-operatively and prospectively collected endoscopic digital images and videos of stones used for this study between January 2018 and November 2020 in a single centre (CHU Pellegrin, Bordeaux). A flexible digital ureterorenoscope (Olympus URF-V CCD sensor) was employed. The study adhered to all local regulations and data protection agency recommendations (National Commission on Data Privacy requirements). Patients were informed that their data would be used anonymously.

1) IMAGE DATASET
The dataset includes 181 kidney stone images which were acquired with four ureteroscope models, namely two from the Olympus company (the URF-V and URF-V2 endoscopes), and two other models from the Richard Wolf company (two different BOA models). The use of different endoscopic devices increases the variability of the image quality due to the acquisition conditions (changing illumination, uncontrolled viewpoints, etc.). Kidney stone surface and section images are shown for each of the four classes in the last two columns of Fig. 3. The relative image count of the images of the four classes (see Table 5) is in accordance with the typical kidney stone type occurrences observed in clinical situation (see Table 1). The classes consist of 57 (31%), 60 (33%), 36 (19.9%) and 28 (13.3%) images for the WW, WD, acid uric and brushite kidney stone types, respectively.

Experts who visually analyze kidney stone types do not observe globally a complete image, but rather interpret the image content by successively exploiting texture and color information of several image regions. The experts interpret in this way both the microscope images (see the
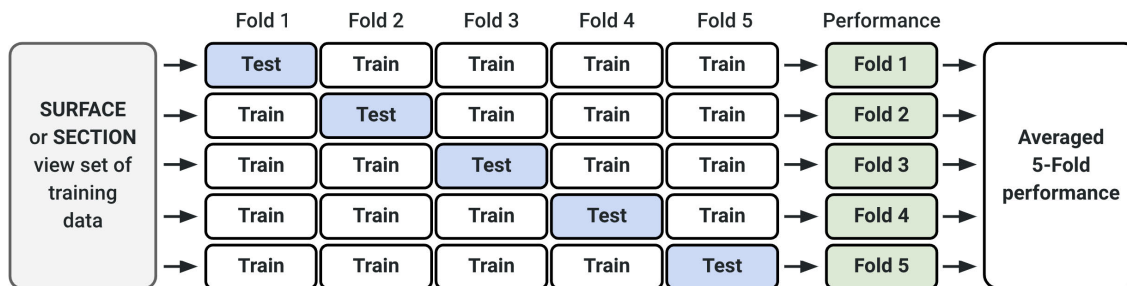
FIGURE 4. The data partition corresponds to a k-fold cross validation strategy (k = 5). The dataset (complete kidney stone images) for each view (surface or section) is distributed in 5 folds (where one fold is used for test and four for the training). The model performance represents the average of the five folds..

two columns on the left of Fig. 3) and the in vivo images (see the two columns on the right of Fig. 3) observed during a morpho-constitutional analysis and an ureteroscopic procedure, respectively. For this reason, the classification of urinary calculi is not performed on whole images in [13] and [16], but using square patches localized on kidney stone surfaces or sections delineated by edges (surrounding tissues should not be visible in the patches) [45]. In this contribution, stone fragment edges are automatically segmented in the images using a learning-based active contour method [57].

The way to organize and divide the dataset for subsequent training and testing was following a k-fold strategy (see Fig. 4 for additional details). The steps to train each model on a given fold partition are described below:

1) *Set the minimum number of kidney stone images per fold.* For each view (surface or section) with four classes each (WD, WW, UA, BRU), 5-folds were performed. The minimum number of images is given by the number of samples per class divided by the number of folds.
2) *Extract patches for each fold.* For each fold, square patches are extracted for each complete kidney stone image as described in Section II-A2.
3) *Class balancing for shallow machine learning methods.* For each fold, and its patches generated in step 2), color and texture features are extracted as described in Section II-B2, and class balancing strategies are subsequently applied to the produced vectors. Methods for balancing color and texture vectors are described in Section II-A3.
4) *Class balancing for DL-based methods.* For DL-based models, patches are generated following steps 1) and 2). Subsequently, under/over/sampling is applied to the minimum number of patches of the classes. With the patches resulting from under/over/sampling, they are used to train the DL models. To improve the network learning process, geometric data augmentation techniques are applied to the patches, as described in Section II-A were used.
5) *Training and testing of each fold.* Each fold was trained with 80% of the samples and tested with the remaining

20%. The performance results of each training and testing for each of the folds are averaged. The final result corresponds to the average of 5 folds. To avoid data leakage, a random selection of the samples without repetition was used.

The processes described in steps 1 to 5 are described in detail throughout Sections II-A2, II-A3, and II-A .
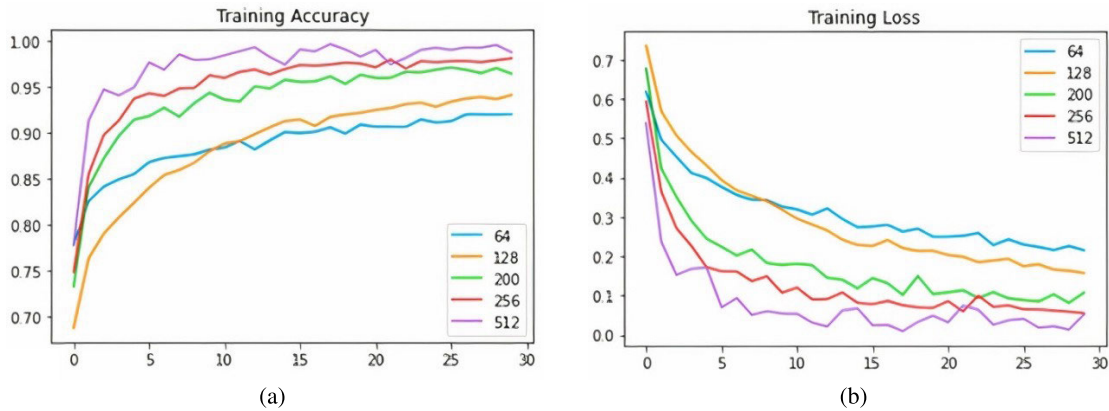
2) PATCH EXTRACTION

The use of patches instead of whole images is not only conform to the medical practice (both with the morpho-constitutional analysis and the visual urinary calculus identification by an expert), but it allows also to construct a larger training and test dataset [13], [16]. The image areas delineated by the segmented contours were overlapped by a regular grid whose square cells correspond to the patches. However, in order to maximize the amount of patches, the grid is not completely included inside the contour limits, to avoid unused kidney stone areas. The grids extend beyond the boundaries of the contours and the cells (patches) at the periphery of the segmented fragments also contain epithelial (kidney or ureter) tissue instead of urinary stones.

However, in some images, patches may contain parts of instruments used to fragment or extract kidney stones. These patches are identifiable since the instruments are easy to segment in the blue channel of the images (contrary to the instruments, epithelial tissues and kidney stones are characterized by color values for which the red and green channels carry stronger signals than the blue channel).

Thus, three precautions have to be taken during the patch extraction. First, the extracted patches have a maximal border overlap of twenty pixels to limit redundant information. Second, patches including a very high number of "non-kidney stone" pixels are 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 including instruments). Third, the patches should have an optimal size to capture local texture and color information (such a size adjustment was not presented in previous works [13], [16]).

FIGURE 5. Illustration of the classification efficiency dependence with the patch size. (a) Accuracy obtained for each patch size with a random forest tree and all 40 feature components (see Section *"Handcrafted feature extraction and selection"* for the handcrafted feature description). (b) Loss curves for the different patch sizes.

The side length of the square patches was a hyper-parameter which was adjusted during the training of the machine learning models presented in Section *"Design of a set of kidney stone recognition methods"*. The best size value was obtained after several ablation studies using five patch areas (64×64, 128×128, 200×200, 256×256 and 512×512 pixels, respectively) and by monitoring the training precision and loss curves for each patch size, as shown in Fig. 5 for a RF classifier. As noticeable in this figure, the accuracy and the loss values respectively increase and decrease significantly when the patch area becomes larger.

Increasing the patch size beyond 512×512 pixels does not improve the performances of the classifiers. With a patch size of 512×512 pixels, the accuracy of the shallow machine learning models is slightly better than with patches of 256×256 pixels. However, with a patch area of 512×512 pixels, the training of the deep learning models lead to over-fitting which is observable in the validation set. Consequently, the same 256×256 patch size was used for all the results reported in Sections *"Design of a set of kidney stone recognition methods"* and *"Results and discussion"*. With this procedure, 2680 and 2470 patches were respectively obtained for each k-fold, for the surfaces and the sections of the complete kidney stone fragment database. The last column of Table 5 gives the number of patches for each class.

3) CLASS BALANCING STRATEGIES
The sizes and the viewpoints of the fragmented kidney stone surfaces and sections are very variable (i.e., the number of useful patches extracted from the images depend on the number of pixels which effectively correspond to urinary calculi). Moreover, the number of images per class is statistically dependant on the urinary calculi type (see Table 1). These two facts explain why the number of patches per class is imbalanced, as noticeable in Table 5. Two strategies were tested to balance the classes.

- **Over-sampling approach**. For this strategy, the class with the highest patch number is taken as reference for the over-sampling. The WW class (Type Ia or Ib) is this reference class, for which 870 and 820 patches were extracted from the images of the surface and section fragments, respectively. The patch number of the brushite, WD and AU classes is increased by randomly extracting additional patches (still with 256×256 pixels) which do not correspond to the cells of the initial (regular) grid of patches. After this over-sampling, the four classes consist of 870 surface and 820 section patches.

- **Under-sampling approach**. This strategy follows a similar principle as the over-sampling since the number of patches of three classes (WW, WD and UA) is randomly reduced so that all urinary calculi classes include the same patch amount as the class with the smallest number of patches (the BRU class consists of 420 surface and 410 section patches).

These class balancing experiments were carried out with the Scikit Learn Imbalanced library [58]. Classification tests have shown that the "up-sampling approach" led to slightly better classification accuracy, which means that the presence of redundant information (partially superimposed patches) is completely compensated by the increase in the number of patches.

B. DESIGN OF A SET OF KIDNEY STONE RECOGNITION METHODS
This section describes the training and validation of shallow and deep learning-based models.[1] The training of the shallow machine learning models follows the traditional pipeline of Fig. 6, which is thoroughly discussed in Section *"Shallow machine learning methods"*. For the deep learning-based methods a different approach is followed by training three well-known models using an end-to-end approach and transfer learning to compensate for the relatively small size of the available dataset, as depicted in Fig. 9. The different

---

[1]Codes are available at https://github.com/CV-INSIDE/lopez2023vivo
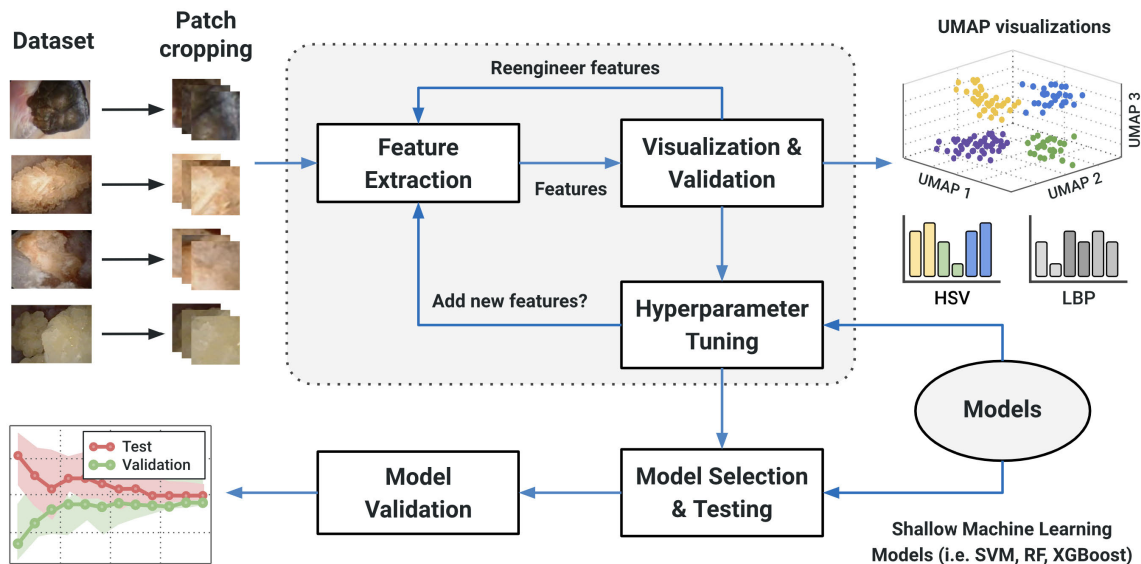
FIGURE 6. Illustration of the training of the six shallow machine learning models: feature extraction and selection, hyper-parameter tuning, and model validation.

choices in terms of deep-learning model design, training of the networks and their validation are detailed in Section *"Deep-learning based classification methods"*.

1) SHALLOW MACHINE LEARNING METHODS

An overview of the training and validation of the shallow models is shown on Fig. 6, which can be summarized as follows. In Section *"Database construction"* it was detailed how the relatively small number of patches was increased for training both classical and deep learning models. Section *"Handcrafted feature extraction and selection"* starts with an argued feature extraction process which encodes handcrafted color and texture information in vectors. The classification accuracy using different feature combinations was monitored to highlight the discriminating capacity of the features associated with a reference classifier. In [16] it was shown that random forest (RF) trees are appropriate to test the classification efficiency according to different feature combinations. These tests were performed using a k-fold cross validation approach and the ability of the features to form separate clusters in the feature space is visualized using UMAP [59]. After the determination of the most discriminant feature vectors, various well known shallow models were trained using an iterative tuning of their hyper-parameters based on a k-fold cross validation.

2) HANDCRAFTED FEATURE EXTRACTION AND SELECTION

Previous works [13], [14], [16] have shown that color and texture features are appropriate to describe the kidney stone surfaces and sections. This section goes more deeply in the justification of various choices made for the feature extraction (size of the local window for the texture encoding,

appropriateness of the chosen color space, and color feature type) for training shallow machine learning models.

a: COLOR FEATURES

Color spaces can be sorted in three families according to their general advantages [60]. The family of tri-stimulus spaces based on a set of primaries gathers all *RGB* color spaces (including the *XYZ* space). These colour spaces are widespread for technical reasons since cameras and screens acquire and display *RGB* values, respectively. Numerous classification algorithms use *RGB* color values which exhibit a major drawback: two visually different colors may be separated by a small distance in the color space coordinate system, and vice versa. For this reason, a second family of colors spaces (e.g. the *Lab* and *Luv* color spaces) were designed so that small or large numerical distances correspond to small and large color perception differences, respectively. These color spaces, in which the *L* component stands for the color intensity and the two other components are chromaticity values, are optimized to measure color differences. However, they do not reproduce the color perception of the human brain which separates the hue and intensity information (the brain perceives more subtle hue differences than intensity changes). The third family of color spaces (e.g., *HSI*, *HSV*, etc.) is based on three components: hue (the tint information), the saturation (the "amount of grey" in the color) and the brightness (the colour intensity).These color spaces allow for a closer simulation of the colour perception by the human brain since a change in only the light intensity does not affect the hue values. In the field of endoscopy, where the intensity of illumination can significantly change from one image to another, it is important to have an information (here the hue) that is independent of the image acquisition conditions. Thus,

the *HSV* space was used to extract the color information from the ureteroscopic images. The brightness, saturation and hue values are defined by (1), (2) and (3), respectively, where $MIN = min(R, G, B)$ and H is in degrees.

$$V = max(R, G, B) \tag{1}$$

$$S = \begin{cases} 1 - MIN/V & \text{if } V \neq 0 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

$$H = \begin{cases} 60(G - B)/(V - MIN) & \text{if } V = R \\ 120 + 60(B - R)/(V - MIN) & \text{if } V = G \\ 240 + 60(R - G)/(V - MIN)) & \text{if } V = B \\ \text{Undefined} & \text{if } R = G = B \end{cases} \tag{3}$$

After the *RGB* to *HSV* color conversion, the energies $e_{I_c}(x, y)$ were pixel-wise determined in each channel I_c ($c = H$, S or V) using (4) in order to capture local color changes:

$$\begin{cases} e_{I_c}(x, y) = \sqrt{g_x (I_c(x, y))^2 + g_y ((I_c(x, y))^2} \\ g_x(I_c(x, y)) = I_c(x + 1, y) - I_c(x - 1, y) \\ g_y(I_c(x, y)) = I_c(x, y + 1) - I_c(x, y - 1) \end{cases} \tag{4}$$

where $g_x(I_c(x, y))$ and $g_x(I_c(x, y))$ are gradient components along the x and y image axes. Representing the occurrences of these local energies using histograms lead to a global color description at the patch level. The energy values computed for each I_c channel were used to build a ten-bin histogram (30 bins in all for the three channels).

b: TEXTURE FEATURES

Haralick features determined with co-occurrence matrices are popular texture descriptors which are appropriate for large regions or entire images. The kidney stone fragment textures are strongly changing according to their location in the images, so that rotation invariant local binary pattern (LBP) values were preferred to capture texture information. These patterns were stored in histograms representing statistical information about local textures. Similarly to the optimal patch size search, classification tests were performed to find the best side size of the LBP windows. These window sizes (i.e., 5×5, 7×7 and 9×9) were hyper-parameters in the classification scheme depicted in Fig. 4. Using the RF classifier, the most discriminant texture features were obtained for a window area of 5×5 pixels. The LBP values are computed using grey level patches (the grey level values are given by the intensity channel of the *HSV* space presented in Section *"Color Features"*) and are used to determine 10 bin histograms.

c: TEXTURE AND COLOR FEATURE VECTORS

The complete feature vector (extracted either from surface or section patches) consists of a 40-bin histogram encoding hue energies (*eH*), saturation energies (*eS*), intensity energies (*eV*) and texture (*LBP*) information. The RF classifier was used to test the discrimination capacity of the features taken individually (*eH*, *eS*, *eV* or LBP taken separately, see Table 6), partially combined in different ways (*eH* + *eS* + *eV* or *LBP* + *eH*) or all jointly used (*LBP* + *eH* + *eS* + *eV*). The 80 components of the feature vectors extracted from the images of the mixed dataset result from the concatenation of the 40 histogram bins of a surface patch and of that of a section patch of the same kidney stone fragment.

UMAP visualizations, as given in Fig. 7, are very helpful for understating the results obtained with the RF classifier exploiting the different feature combinations given in Table 6. Several interesting conclusions can be drawn from the plots in Fig. 7. Fig. 7(a) shows that, for the surface patches, the vector including all *HSV* and *LBP* feature histograms provides a high separability of the classes. This visual representation of the class separability is in accordance with the high consistent performance across all metrics (accuracy 87±17%, precision of 87±17%, recall 86±16% and F1-Score 86±17%) obtained by the RF model (Table 6). As shown in Fig. 7(b), using only section images leads to a poorer class separability. This result is in accordance with the visual classification by human experts reported in the concordance study given in [17] and is also confirmed in Table 6 by the lower performance values close to 83±19% in all selected metrics. The hue component carries significant information leading to discernible clusters (Fig. 7(c)) and, combining hue and texture descriptors *LBP* + *eH* (for both sections and surfaces) helps the classifier to attain a general performance of 88±27% for accuracy, precision, recall, and F1-Score (Table 6). Finally, the clusters become more discernible when using the 80-component *LBP* + *eHSV* feature vectors including both surface and section information. The visualisation in Fig. 7(d) is again in accordance with the general performance of 91±15% obtained by the RF classifier (see Table 6) across all metrics.

It is noticeable in Table 6 that, in contrast with the results reported in the previous works (see also Table 3), the LBP features associated with a RF tree yield relatively high accuracy, recall precision, and F1-Score for both surface and section images. It is also noticeable that the hue channel taken separately provides significant discriminant capabilities. When using all texture and color features together one reach a general performance of 87±17% and 83±19% for surface and section images, respectively.

Furthermore, it is also noticeable that, in comparison to the preliminary work in [16] which also uses a RF classifier, the increase on the number of classes (i.e., the introduction of brushite) was not done in detriment of the overall performance (93% in [16] and 91±15% in Table 6) obtained when using the 80-component feature vectors. Fig. 8 shows a confusion matrix (including representative patches for each class) that partially explain why shallow features are somewhat ineffective: the data exhibit high intra-class variations, such that the aspect of the images of two classes can be close in terms of color and texture.

TABLE 6. Random Forest performance using different feature descriptors for section, surface and mixed patches. All results are presented as a percentage (average % ± standard deviation %). The metrics evaluated are Accuracy, Precision, Recall, and F1-Score..

| Descriptor	Surface images				Section images				Mixed images			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
eH	78±22	79±21	76±22	77±21	64±28	65±28	64±28	64±28	76±22	76±22	76±22	76±22
eS	57±22	59±22	57±21	57±21	61±20	61±20	61±20	60±20	64±18	64±18	64±18	64±18
eV	62±19	57±20	66±19	61±20	59±20	60±21	59±21	59±21	63±19	63±19	63±19	62±19
eHSV	76±23	81±22	77±22	70±22	71±25	71±25	72±25	71±24	83±29	83±29	83±29	82±29
LBP	85±28	85±28	85±28	85±27	72±34	73±35	72±34	71±34	87±27	87±26	87±27	88±27
LBP+eH	86±27	87±27	86±27	86±27	75±32	75±33	75±33	75±33	88±27	88±26	88±27	88±26
LBP+eS	82±20	81±29	82±29	81±29	76±32	77±32	77±32	76±32	89±26	89±25	88±25	90±26
LBP+eV	77±21	79±22	75±22	77±21	77±22	78±21	76±21	77±22	71±28	71±28	71±27	71±28
LBP+eHSV	**87±17**	**87±17**	**86±17**	**86±17**	**83±19**	**83±19**	**82±19**	**83±18**	**91±15**	**91±15**	**91±15**	**91±15**
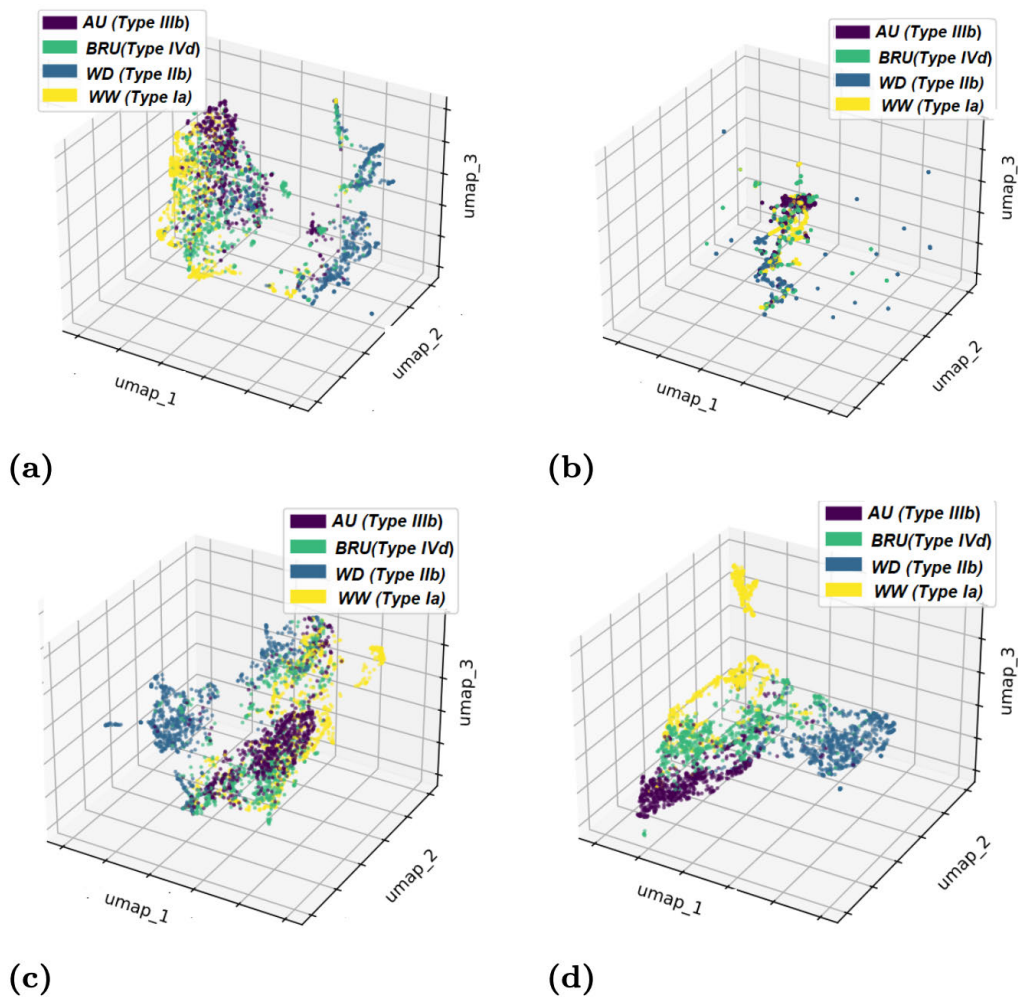
FIGURE 7. UMAP visualisation for the representation of the class separability according to different feature combinations and the separate or simultaneous use of surface and section patches. This UMAP representation is achieved using only the three most discriminant dimensions (umap1 to umap3) obtained after a dimensionality reduction of the *HSV -LBP* feature space. (a) Vector representation including all handcrafted *eH*, *eS*, *eV* and *LBP* features extracted from surface images. (b) Same representation as in (a), but for section images. (c) UMAP feature space representation obtained when using only the colour information (the 2 × 30 component vector of *eH*, *eS*, *eV* values) extracted both from section and surface images. (d) Same as in (c) but with all features (*eH*, *eS*, *eV* and *LBP*) extracted from both patch types and encoded in 80-component vectors.

3) TUNING OF THE CLASSIFIERS EXPLOITING HANDCRAFTED FEATURES

As depicted in Fig. 6, after having identified the best combination of handcrafted features, a set of six state-of-the-art machine learning models was trained (see Table 7). The hyper-parameter tuning consisted of a combination of a grid and random search using the Scikit-Learn software [58]. This search of optimal hyper-parameters was performed for three
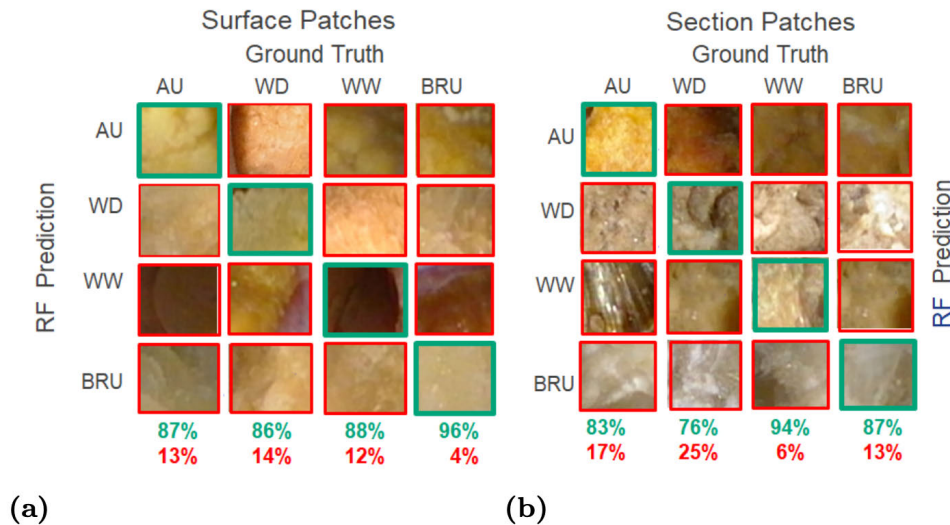
FIGURE 8. Confusion matrix for the four classes of the dataset. The results were obtained with a RF classifier (the percentages in green indicate the rate of correct recognition). Classification errors (percentages in red) are due to the intra-class variability and inter-class similarities.

datasets, namely i) by using only the surface patches from which vectors gathering all 40 color and LBP components were extracted, ii) by exploiting the same vectors extracted only from the section patches and iii) by exploiting the 80 component vectors obtained with a surface and section patch. The last dataset is the most representative of the clinical practice since the aspect of both the surfaces and the sections are taken into account by when human operators visually identify kidney stones. The two other datasets are useful to assess the contribution of the surface and section information taken separately. A stratified k-fold cross validation approach was used in order to maximize the number of data in the testing phase and to mitigate biases. The accuracy, precision, recall and F1-Score were measured for each of the three dataset configurations.

The six chosen models represent a relatively large pool of shallow machine-learning methods. The best hyper-parameter values for the combined surface and section patches descriptors are given below:

- **SVM.** The best model was obtained by setting the C parameter value at 1.16, and by using a sigmoid kernel with its defaults values for the coeff0 ($= 0.0$) and gamma ($=$ scale) hyper-parameters.
- **AdaBoost.** A set of decision tree classifiers was used for the model, the number of estimators and the maximum depth being equal to 100 and 12, respectively. The best LR was set to 0.1.
- **Bagging.** The bagging model has the same parameter values as AdaBoost, but the Random Forest three was employed as the base estimator (the number of estimators equals 160).
- **Multi-Layer Perceptron (MLP).** The MLP model had the following hyper-parameter settings: it consists of three layers, and 200 neurons were used in the hidden

later. Tt was trained for 200 epochs using the L-BFGS solver.
- **Random Forest (RF).** The RF model with the best hyper-parameter settings consists of 1800 estimators, a minimum split value of 5, a minimum of samples per leaf of 2, and a max depth of 50. Bootstrap was not used.
- **XGBoost.** The best hyperparemeter settings of the XGBoost model are the following. The base score value was set to 0.5. gbtree 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.

a: DEEP-LEARNING BASED CLASSIFICATION METHODS
The efficiency of deep-learning models lies on their ability to automatically extract highly discriminating features [61]. Section *"Training of the chosen deep-learning models"* presents the deep learning models that were designed for the kidney stone classification. Then Section *"Visualization of deep-feature data"* highlights the discrimination ability of the extracted deep-features.

b: TRAINING OF THE CHOSEN DEEP-LEARNING MODELS
Contrary to shallow machine learning solutions, the features extracted from images by deep-learning models do not correspond to a predefined physical information (e.g., relating to colors or textures), but depend on weight values linking the input data to class probabilities. After the model training, it is difficult to physically interpret the deep-features since they depend on numerous weights of the convolution layers. However, the appropriateness of deep-features to discriminate instances of different classes can be analysed after the learning phase using visualization tools like UMAP [59] or explainability techniques such as GradCAM [47]. The approach depicted in Fig. 9 was used
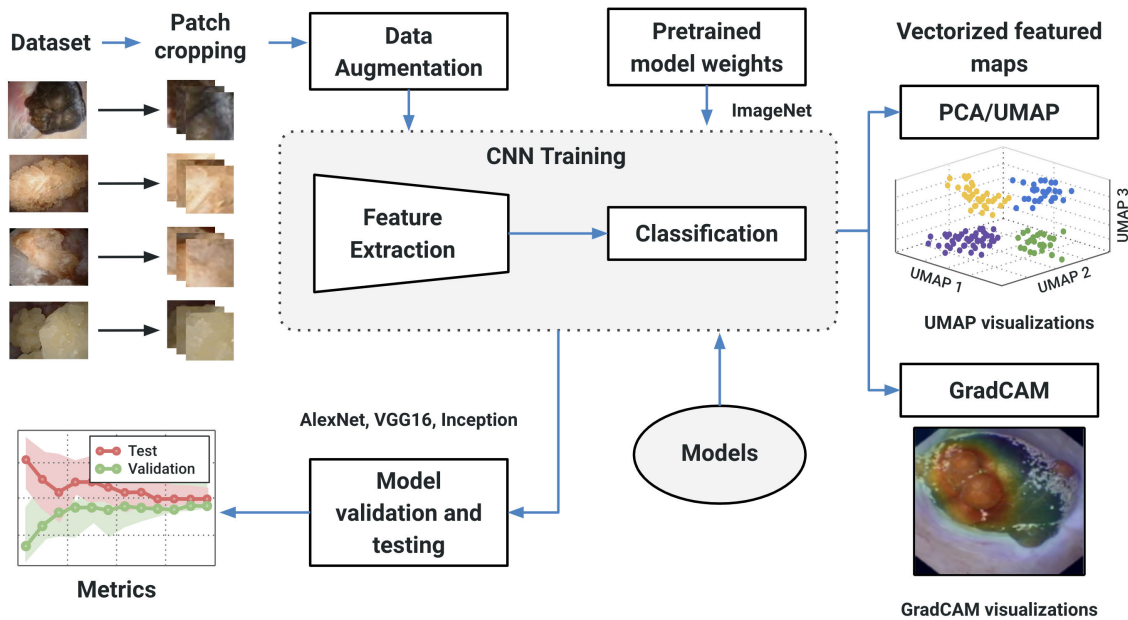
FIGURE 9. Process followed for training the deep learning models: data augmentation and training using pre-trained feature extraction backbones, model selection and validation, and feature visualization.

to exploit various DL-models. Three DCNNs were first pretrained using transfer learning, while the data (the patches whose generation was discussed in Section *"Class balancing strategies"*) used for the final training were augmented.

CNN models with different extraction backbones and architectures (AlexNet, VGG16 and InceptionV3) were tested in this contribution. Due to the moderate amount of available data, the convolution layers of the DL-models were pre-trained using ImageNet. Moreover, to adapt the DL-models to the kidney stone recognition task, the fully connected (FC) layer of the feature extraction backbones 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 4 class outputs for yielding the class prediction. The fully connected layers weights were randomly initialized. During the training of the three models, 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.

Extensive data augmentation was performed in order to limit the overfitting induced by the small size of the training dataset. Additional patches were obtained by applying vertical and horizontal flips, perspective distortions, and four affine transformations on the patches extracted from the images. With this data augmentation, the number of samples in the training set passed from 5,400 to 43,200 (10% of the samples were hold out for test purposes). Further, the patch values were "whitened" using (5) in which the mean m_i and standard deviation σ_i of the colour values $P_i(x, y)$ are

determined in each color channel:

$$P_i^w(x, y) = \frac{P_i(x, y) - m_i}{\sigma_i}, \quad \text{with} \quad i = R, G, B. \quad (5)$$

Each DL-model was trained three times, i.e. with different datasets (only section patches, only surface patches and both patch types combined). Thus, different parameter tuples were computed for each dataset which were classically split in three parts, namely training, validation and test sets. All the experimental studies reported in this paper made use of Pytorch 1.7.0 and CUDA 10.1. The hyperparameters such as the learning rates were automatically adjusted for each architecture using the optimizer provided by Pytorch (Lightning 1.0.2). LR values of 0.0001, 0.00005, and 0.0006 were obtained for AlexNet, VGG16, and InceptionV3, respectively. The ADAM optimizer, a batch size of 64, and early stopping were employed in all tests.

c: VISUALIZATION OF DEEP-FEATURE DATA
The result description in Section *"Deep Learning using transfer learning"* is based on two visualization tools which enabled us to better understand the ability of DL-models to recognize kidney stone types using the deep-features extracted from the three datasets (surface or/and section patches).

- **UMAP** (Uniform Manifold Approximation and Projection for Dimension Reduction, [59]) is used to construct a high dimensional (deep-learning) feature graph for each dataset type. The UMAP algorithm then reduces the dimensionality of the feature space by optimizing a low-dimensional graph so that both graphs are as structurally similar as possible. In this contribution,

the deep-learning features are represented in a three dimensional space whose dimensions umap1, umap2 and umap3 give component values obtained after a non-linear dimension reduction. These 3D representations illustrate the class separability of the deep-features.

- **GradCAM** (gradient weighted class activation mapping, [47]) Additionally, the features in the low-dimensional space are used for creating heat maps. This visualization technique uses the class-specific gradient information flowing into the final convolutional layer of a CNN to produce a coarse localization map of the important patch regions which triggered the classifier output. GradCAM representations allow for a better understanding of some of the errors made both by shallow and deep learning-based models. These class activation mappings can be used for determining the information (i.e., for finding the important color or texture features, or locating the image areas including the most important information) which favour a successful kidney stone recognition.

III. RESULTS

Various experiments were carried out for evaluating both shallow machine learning-based and deep learning-based methods. In all, nine models were trained three times, namely i) solely with the section patches, ii) only with the surface patches, and iii) by mixing the two patches types. Each section or surface patch was classified by its dedicated model and by the model for mixed data.

A. EXPERIMENTS USING SHALLOW MACHINE LEARNING METHODS

This section compares the results of six shallow machine learning models which were tuned with the validation pipeline depicted in Fig. 6. All results are given for the complete 40 component *HSV/LBP* feature vectors when either only surface or only section information is used for the recognition, and for the 80-component vector when both data types are simultaneously exploited for the classification.

When analysing Table 7 it becomes clear that the first results given in the literature for ex-vivo data were significantly improved by carefully choosing handcrafted features. Indeed, the authors in [13] obtained a precision of 63% (see Table 3) over four classes using a RF model exploiting *RGB* color values and non optimized *LBP* window sizes. As seen in Table 7, the general performance across all metrics 87±17%, 83±19%, and 91±15% for surface patches, section patches and both patch types, respectively) is greatly improved when a RF model exploits *HSV* color features and an appropriate *LBP* window size. The results reported in Table 7 for various shallow machine learning based methods outperform even the first results obtained with deep-learning approaches: in [14] the authors obtained a precision of 74% over four classes using a Siamese CNN

solution, while in [15] a ResNet 101 led to a precision from 71% up to 94% according to the class.

It is also noticeable in Table 7 that the performance criteria (accuracy, precision, recall and F1-Score) exhibit different values for surface and section patches taken separately. On the one hand, surface patches are globally more discriminant than section patches, and two classifiers lead globally to the best (XGBoost) and second best (MLP) results for the accuracy, precision, recall and F1-Score values for surface data. On the other hand, XGBoost enables the best classification for section patches, whereas no second best classifier really emerges for this patch type (for the section patches, the second best criterion values were obtained by three different classifiers). This difference in an automated exploitation of surface and section data confirms the visual analysis of the expert who noticed in [17] that the classification of surface images is easier than that of section images, as the former present more texture information due to the crystallization process than the latter. However, it can be noticed that when using both surface and sections patches two classifier obtain systematically the best (XGBoost) and second best results (Random Forest). These two shallow machine learning methods are able to reduce the effects of high intra-class variability and low inter-class differences when exploiting simultaneously surface and section data (the 80-component feature vector). The performances of the XGBoost and RF classifiers are compared in Section *"Deep Learning using transfer learning"* with those of the tested deep-learning methods.

B. DEEP LEARNING USING TRANSFER LEARNING

Tables 8 and 9 gather complementary results of both the two best shallow machine learning methods and the three tested deep-learning methods. Table 8 gives results for individual classes including the surface and section patches, whereas Table 9 provides class weighted performance criterion values for the surface patches, the section patches, and the patches of both types.

It is noticeable that all deep learning models (InceptionV3, AlexNet, VGG16) outperform the best shallow machine learning method based on the XGBoost classifier. As seen in the last row of Table 8, this DL-model exhibits the highest mean accuracy, precision, recall and F1-Score values for all classes. This result is also confirmed in Table 9 for all patch datasets (section patches, surface patches and mixed patch types). It can also be observed in Table 9 that, similarly to the shallow machine learning methods, the simultaneous use of surface and section patches leads to the best results, whereas the three DL-based methods exhibit globally superior results when exploiting only surface images. This last observation confirms the visual classification results of the concordance study in [17]. However, among the DL-approaches, InceptionV3 has globally a higher accuracy (97±03), precision (97±03), recall (98±04), and F1-Score (98±03) than the XGBoost classifier values

TABLE 7. Comparison of the performance of six shallow machine learning models according to the data type (section or/and surface kidney stone patches). The numbers in bold and italics represent the best and the second best results, respectively.

Classifier	Surface				Section				Mixed			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
SVM	85±28	83±28	86±28	84±28	81±37	76±38	*86±37*	80±38	78±20	79±20	77±20	78±20
AdaBoost	84±28	83±28	86±28	84±28	*83±29*	81±28	85±29	*83±29*	81±29	81±20	81±20	81±20
Bagging	76±23	76±23	76±23	76±23	77±30	77±30	77±30	77±30	75±21	75±21	76±21	75±21
MLP	*88±20*	86±29	*91±20*	88±29	72±23	80±23	64±24	71±24	85±22	84±23	86±23	85±22
R. Forest	87±17	*87±17*	86±17	86±17	*83±19*	*83±19*	82±19	83±18	*91±15*	*91±15*	*91±15*	*91±15*
XGBoost	**93±17**	**93±16**	**93±17**	**93±16**	**89±17**	**89±17**	**89±17**	**89±17**	**96±14**	**96±14**	**96±14**	**96±14**

TABLE 8. Precision and recall values obtained for the four classes with five classifiers. The best and second best criterion values are given for each class in bold and italics, respectively.

Method	WW (Ia)				WD (IIb)			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
Random Forest	85±21	84±21	86±22	85±21	92±22	90±22	*95±22*	92±22
XGBoost	94±18	92±18	96±18	93±18	91±19	91±18	91±18	91±19
AlexNet	96±07	93±07	**98±07**	95±07	90±09	**95±09**	85±08	91±09
VGG16	**97±06**	*97±07*	97±07	**97±07**	*93±08*	92±08	*93±08*	92±08
InceptionV3	**97±06**	**98±06**	97±05	97±06	**95±09**	93±09	**96±09**	**94±09**

Method	AU (IIIb)				BRU (IVd)			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
Random Forest	79±23	88±23	67±23	81±23	91±20	90±20	92±20	91±20
XGBoost	**96±19**	**97±10**	**96±19**	**96±19**	*94±18*	**96±19**	*94±19*	92±19
AlexNet	90±10	88±10	*92±10*	*91±10*	92±10	93±09	92±09	92±10
VGG16	88±10	93±10	83±11	90±11	93±09	94±09	92±08	*93±09*
InceptionV3	*92±09*	*95±08*	90±08	90±08	**95±08**	*96±08*	**95±09**	**95±08**

TABLE 9. Weighted average precision, recall and F1-Score comparison for the dataset including only surface patches, the dataset consisting only of section patches, and for the complete dataset with both patch types. The best and second best criterion values are given for each class in bold and italics, respectively.

Method	Surface				Section				Mixed			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
R. Forest	87±17	87±17	86±17	86±17	83±19	83±19	82±19	83±18	91±15	91±15	91±15	91±15
XGBoost	93±17	93±16	93±17	93±16	89±17	89±17	89±17	89±17	96±14	96±14	96±14	96±14
AlexNet	94±08	93±08	*95±08*	94±08	83±07	83±07	82±07	83±07	96±04	94±04	*97±04*	*97±05*
VGG16	94±08	*94±09*	94±08	*94±09*	*92±09*	*91±09*	*92±08*	*92±09*	95±05	95±05	96±05	96±05
InceptionV3	**95±08**	**95±08**	**95±08**	**95±08**	**94±08**	**94±08**	**94±08**	**95±08**	**97±03**	**97±03**	**98±04**	**98±03**

accuracy (96±14), precision (96±14), recall (96±14), and F1-Score (96±14) for both surface and section views.

The precision and recall values in bold (best result) and italics (second best result) given for each class in Table 8 were most often obtained by one of the three deep-learning models. Similarly, the deep-learning methods most often delivered the two best criteria values for the datasets including only one patch type (either surface patches or section patches in Table 9). However, when considering globally the results (weighted criterion values for all patch types exploited together, see Table 9), only the InceptionV3 model has slightly better results than the XGBoost model.

IV. DISCUSSION
The visualization of the three most discriminant UMAP components in Fig. 10 explains why AlexNet exhibits rather moderate kidney stone recognition performances. As it can be observed in Fig. 10(a), the feature extraction backbone of AlexNet determines deep-features that form compact and separated clusters for surface patches. This result is in

agreement with the high recall values obtained for surface images in the concordance study in [17]. In contrast, the deep feature clusters produced by AlexNet for section patches are close and elongated. For instance, Fig. 10(b) shows that the clusters corresponding to the weddellite and brushite classes are very close in the reduced UMAP feature space (the two clusters are even touching themselves in some reduced feature space places). The WD and WW clusters are also very close to each other, but without touching themselves. This explains why in Table 9 the precision and recall values are rather low for section images, and high for surface images when using AlexNet. Fig. 10(c) shows the feature clusters when training AlexNet on both surface and section patches. In this figure, the inter-cluster distances and compactness are visually similar to those of Fig. 10(a). This observation is conform with the fact that the accuracy, precision, recall and F1-Score values of the surface dataset (94 ± 08%, 93 ± 08%, 95 ± 08%, and 94 ± 08% respectively, see Table 9) are close to those of the mixed dataset (96 ± 04%, 94 ± 04%, 97 ± 04%, and
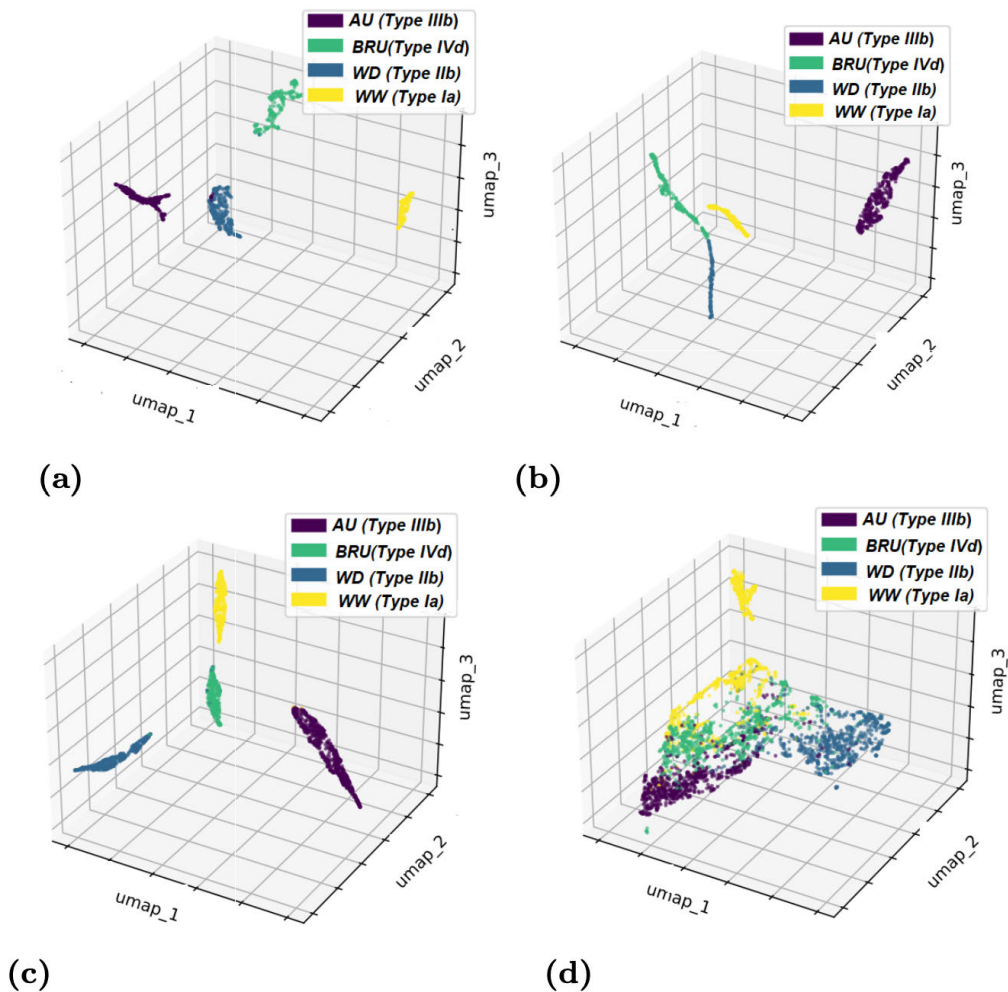
FIGURE 10. UMAP visualization given for the AlexNet model. The values of the umap1, umap2, and umap3 components were obtained after a dimensionality reduction of the initial feature space. (a) Cluster representation of the four kidney stones classes when training the AlexNet model only with surface patches. The initial feature space (before dimensionality reduction) consisted only of deep-features extracted by the AlexNet convolutional layers. (b) Same as in (a), but by training the AlexNet model only with section patches. (c) UMAP visualisation for the feature extraction performed by training AlexNet with both section and surface patches. Although being elongated, the more distant clusters indicate improved classification performances in comparison to the clusters in (a) and (b). (d) The clusters obtained after a dimensionality reduction of the handcrafted HSV/LBP feature space (same as in Fig. 7) are shown to allow for a comparison with the deep-feature clusters in (c). This comparison highlights the feature extraction improvement attained by a simple model such as AlexNet over traditional handcrafted features (here the three most discriminating components obtained for the 80-component vectors for mixed section and surface patches).

$97 \pm 05\%$). While the section dataset taken separately leads to rather moderate recognition results, it improves slightly the classification results when it is associated to surface patches. As seen in Table 9, the observation made for AlexNet can be extended to the two other tested deep-learning methods: surface images are in general almost sufficient for obtaining a high classification performance, but mixing surface and section images increases the recall values of about 1% in average (compare the recall values of AlexNet, VGG16 and InceptionV3 with and without section patches). For the sake of completeness, Fig. 10 provides also a comparison between the dimensionality reduced deep-feature space (see Fig. 10(c)) and the reduced handcrafted HSV-LBP feature

space determined with the surface and section datasets (see Fig. 10(d)).

When comparing these two 3D spaces, one could conclude that the class separability offered by deep-features leads to a more efficient classification than the handcrafted features, the inter-class distances and cluster compactness being visually the highest in Fig. 10(c). However, XGBoost is able to exploit the less promising handcrafted features to achieve a classification with a performance slighly better than that of the VGG16 and AlexNet networks (see the F1-Score for mixed data in Table 9).

The performances of the AlexNet feature extraction backbone are further discussed with the aid of the GradCAM
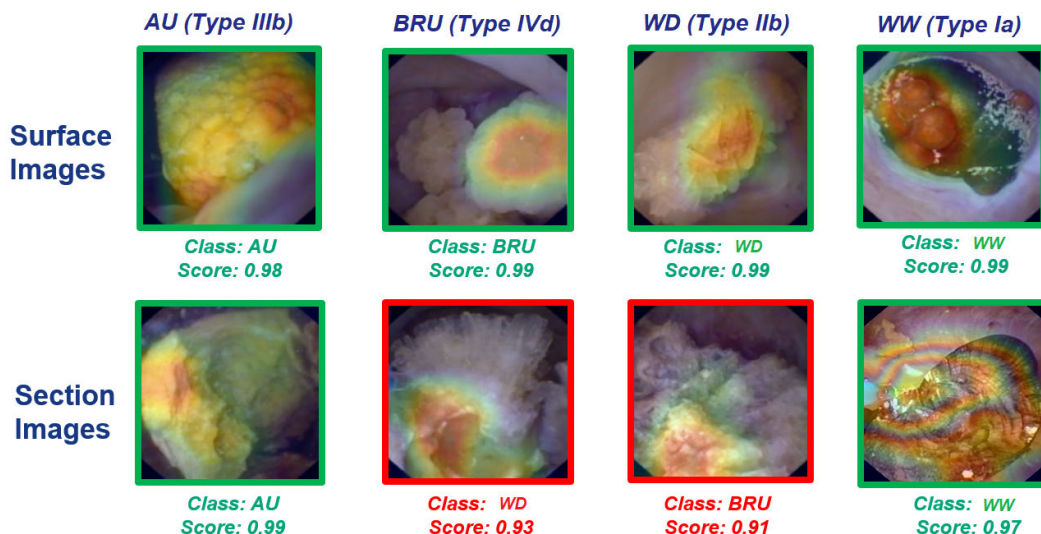
FIGURE 11. Grad-CAM visualizations illustrating the performances of the AlexNet model. Complete kidney stone images are superimposed by their corresponding gradient-weighted class activation maps. The first and second rows correspond to the uric acid, brushite, whewellite and and wedellite images acquired for the surface and section images, respectively. The green image frames indicate a correct classification by the GradCAM activation, while the red frames refer to as a very high activation score which led to a classification error (the WW and BRU section images were misclassified).

network dissection and visualization method. In this visualization technique, the images are overlapped by heatmaps in which color codes indicate the importance of particular image regions during the prediction making of deep learning models (from the red to the blue colors, the importance of the pixels decreases, while regions with grey-level values have no impact on the decision). Fig. 11 shows the activation maps for a surface (first row) and a section (second row) image of each class. The GradCAM visualisations were determined with models trained and tested on entire images of a unique patch type (either surface or section data). From the surface images of first row, it can be inferred that the feature extraction backbone focuses in well-defined regions with significant presence of both color and texture information to produce highly discriminant features enabling a classification with high confidence (the green frames indicate a correct classification and the class score approaches 1). Trained clinicians also focus their attention on local and significant colour and texture images to recognized kidney stones either in microscopic images (during a morpho-constitutional analysis [5]) or in endoscopic data (during ureteroscopies [12]).

However, it must be noted that GradCAM can produce relatively active heatmaps for images of a given class, even if the visualisation tool is fed with images from another class, as other components of the softmax layer vector might contribute to non-negligible class score values (for instance, the class score for the WD can be 0.75, whereas a value of 0.20 can be obtained for the brushite class and 0.025 for the other two classes). In such a situation, the precision, recall and F1-Score values tend to be weak and the activation maps are sparser. For this reason, heatmaps should always be analysed

by jointly considering the classification quality criteria. This is shown in the second row of Fig. 11 with GradCAMs of section images. Even if the activation maps indicate a significant presence of colours and texture information, correct classification scores (close to 1) were only obtained for the uric acid and whewellite (COM) samples (0.99 and 0.97, respectively). In contrast, AlexNet misclassified the brushite and whedellite (WD) samples (assigning them the wrong label with a high confidence score of 0.93 and 0.91 scores, respectively). These results are in accordance with the UMAP visualizations in Fig. 10(b), where the clusters for COD and brushite classes are very close to each other. The InceptionV3 deep-learning model exhibits the best overall performance (see Tables 8 and 9) since it produce compact and distant clusters for all datasets (surface, section or mixed patch types). The cluster separability of the InceptionV3 backbone is illustrated in Fig. 12 by the UMAP visualisation.

The performance of the proposed models is compared to that reported in the state-of-the-art (see Table 3 and [13], [14], [15], [16], [17], [18], [19], [20]).

As noticeable in the last line of Table 7, promising results (93±17%, 89±17%, and 96±14% for surface, section, and mixed views, respectively) were obtained for in-vivo endoscopic images of four of the most common classes encountered in clinical practice. These performance is higher than that of the other models based on shallow features. It is noticeable that these high performance was obtained on in-vivo data, while previous work in [13] (see Table 3) reached only a 63% overall precision for ex-vivo images acquired for four kidney stone sub-types. The results in the last line in Table 7 also surpass that of previous works tested on in-vivo
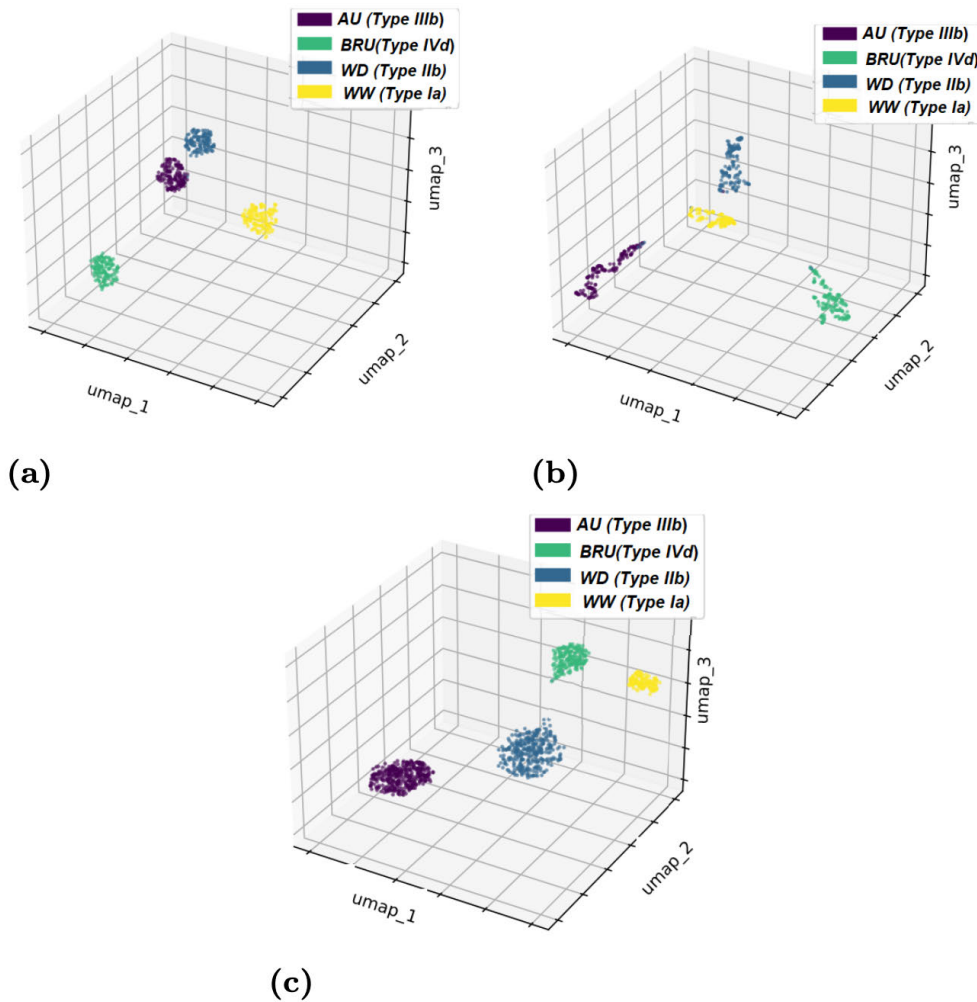
FIGURE 12. UMAP visualisation given for the InceptionV3 model. (a) Representation of the four kidney stones classes when training the InceptionV3 model only with surface patches. (b) Same cluster representation as in (a), but by training the model only with section patches. (c) Reduced feature space obtained by mixing both section and surface patches: the cluster separation is increased and leads to improved classification performances for the four classes.

data. For instance, the work in [16] (see again Table 3) led to a global precision of 79% and 89% over three classes for surface and section views respectively.

As noticeable in Table 9, the proposed DL-based models obtained outstanding results: a general performance of $95 \pm 08\%$, $94 \pm 08\%$, and $97 \pm 03\%$ was respectively reached for surface, section, and mixed views of in-vivo images of four classes. The proposed DL-model led to a higher precision than most of the models that classify kidney stones using ex-vivo images (average precision of 74% for up to 5 classes in [14] and average precision of 84% for five classes in Black et al. [15], see Table 3) The proposed model also outperformed other methods tested on in-vivo images (average precision of 88% for up to 6 classes in [18]). The work in [19] achieved a high precision of 88%, 84%, and 96% for surface, section, and mixed views of six classes. However, the dataset used by these authors consisted of ex-vivo images

acquired in more controlled acquisition conditions than for in-vivo data. The precision obtained by the proposed model (98%, 94%, and 97% for surface, section, and mixed views) was measured on in-vivo data including the ureteroscopic scene variability.

Finally, the results of the most recent models [17], [20] report a lower performance than that reached in this contribution (see Table 3). The work in [17] led to a precision of 83% and 81% for surface and section views respectively, on a set of in-vivo endoscopic images of four sub-types. Similarly, the work of [20], reached a precision of 80.2%, 87.6%, 85.2% surface, section, and mixed respectively, for ex-vivo endoscopic images of 6 classes.

As shown throughout this work, different techniques of shallow machine learning and deep learning have been applied to the classification of kidney stones in endoscopic images. Despite the promising results obtained so far, there

are still several limitations that should be considered in future work to develop robust IA systems for clinical practice. Following the major limitations that persist in the classification of kidney stones can be identified:

A. LACK OF LARGE IMAGE DATABASES

It is rather difficult to obtain large datasets, especially for kidney stone sub-types with few occurrences. In addition, it is complicated to obtain the same number of samples per class, since different types of kidney stones appear in different percentages (e.g., WD and WW types represent 76% of the total number of kidney stones). This work explored an approach to obtain a larger dataset through patch generation, which results in a balanced dataset for training and testing purposes. However, it is desired in clinical practice to perform classification of kidney stones on endoscopic images using complete images [62]. In future work, knowledge transfer based models such as Transfer Learning could be implemented to initialize models without the need to train them from scratch. Transfer learning methods are able to learn under a few samples scheme (e.g., Few Shot Learning) [63].

B. DOMAIN SHIFT

Currently, domain shift represents one of the biggest challenges in DL-based models [64]. The inability to use one model on different test sets (e.g., images from different endoscopes) is a concern. Models capable of adapting to domain shifts (from different sources) and robust to adverse conditions during an ureteroscopy (e.g., illumination changes or scene blur) are required. In order to deal with domain change, it would be interesting to explore techniques such as Meta Learning that are robust to slight changes in the scene and that have generalization capabilities (regardless of the hardware characteristics of the acquisition instrument) [65].

C. MODEL RETRAINING

An important limitation in DL-based models is the lack of learning novelty over time [66]. Current models require learning a task in a single episode (static models). However, in the clinical area, tasks may change (such as learning a new domain or learning a class never seen before) and require continuous adaptation/updating to learn new concepts. Therefore, approaches such as Continual Learning and its different paradigms in tasks learning new classes or new domains could be employed to have AI-based recommender systems in the clinical area [67].

D. LACK OF RELEVANCE IN THE SAMPLES

As mentioned above, the lack of data is one of the main limitations in the classification of kidney stones. Although current methods for data augmentation or extraction of relevant features are useful, instances are equally important to the learning task. Therefore, novel methods are required that allow selecting relevant samples for the learning task, especially in a context with few labels (new classes in the corpus or infrequent samples). Therefore, emerging methods such as Multi Instance Learning [68] represent an alternative in a weakly-supervised scheme, to determine relevant samples in a data set. In the context of kidney stone classification, this could mean boosting in the form of learning to a greater extent from the most relevant samples, and giving a lower weight to those that are not relevant.

Although there are a number of limitations of the current DL models, it can also be seen that there are different state-of-the-art techniques that could be explored to improve the existing methods for classification of renal calculi. In addition, new approaches are emerging every day that could contribute to the kidney stone identification task.

V. CONCLUSION

In this pilot study, it was shown that it is possible to train machine learning models (both shallow and deep learning-based) for recognizing the type of kidney stones using only digital images acquired with endoscopes during standard ureteroscopies. The results presented in this contribution show that AI methods are potentially a precise solution to help urologists to recognize the morphology (i.e, the crystal type) of four kidney stone types (subgroups Ia, IIb, IIIb and IVd). Additional works need to confirm this ability to identify the morphology on a larger number of pure and multilayered kidney stone classes. Until now, the FTIR analysis remains essential to complete the morphological analysis with the determination of the biochemical composition of the stone. A solution to move towards a complete diagnosis during the endoscopy would be to equip ureteroscopy operating rooms with spectrometers which can collect IR signals in hollow organs using an optical fiber passing through the endoscope's operating channel. This contribution represents an important first step towards the immediate determination of an appropriate treatment avoiding recurrence in terms of kidney stone formation, while making the vaporization of kidney stones more systematic.

As thoroughly discussed, the kidney stones have various visual aspects that have been used to propose taxonomies (based on color, texture and morphological descriptions) for aiding the urologist in their visual classification. Compared to the work by Serrat et al. [13], it was shown that a careful feature extraction and reduction can led to an efficient kidney stone type recognition using shallow machine learning methods. High performances were obtained when classifying the most common classes of urinary calculi (Ia, IIb, IIIb, IVd). For these four kidney stone types, the XGBoost method led to an precision of $93 \pm 17\%$, $89 \pm 17\%$ and $96 \pm 14\%$ when using only surface patches, solely section patches, and both patch types, respectively. Furthermore, this contribution is an extension of a previous preliminary study which only focused on shallow machine learning methods applied on a smaller class number. It was shown that some of the most common deep learning architectures (AlexNet, VGG16 and InceptionV3) can be effectively trained for obtaining solutions with comparable or higher

performances than those obtained by Black et al. [15]. Some tested CNN-models (e.g., InceptionV3) are with a lower complexity than ResNet-101 and Restnet-152, but reach a slightly better precision due to their improved information density capabilities [69]. In this contribution, the weighted average precision obtained for InceptionV3 equals (95 ± 08%, 94 ± 08%, 97 ± 03% for surface, section and mixed surface/section images, respectively), while high recall values were reached for all four used classes. However, the main difference between this study and previous works lies on the demonstration of the feasibility of classification methods making use of images acquired using flexible endosocopes under uncontrolled acquisition conditions (in previous studies such as [15], the results were obtained in ex-vivo under ideal acquisition conditions).

In comparison to the most recent work that investigated the use of deep learning techniques for classifying the morphology of different types of kidney stones also acquired in in-vivo [17], the main contribution of this work lies in the thorough comparison of both shallow and deep learning architectures. This comparison also focused on the under-standing of the features enabling a precise classification, as well as of the limitations of some methods. By comparing deep learning models of various levels of complexity, it was graphically possible to confirm the results of the concordance study by Estrade et al. [17]. For instance, it was verified with various classification experiments and with the UMAP visualizations that both shallow and deep learning models can reach a high accuracy when classifying UA (Type IIIb kidney stones), but are less effective when images the of weddellite (Type IIb stones) and brushite (Type IVd) classes need to be distinguished. To a lesser extent, whewellite images are also more complicate to be separated from the two previous classes. These observations explain the lower recall values for these three classes, both in the concordance study in [17] and in this contribution. Furthermore, the UMAP visualizations have been integrated into an interactive visualization tool [70] that enables the exploration of more complex models and databases (i.e, include more pure kidney stone classes or mixed stones for instance). It is also noticeable that the average precision obtained in [17] for surface and section images taken individually (94% in both cases, see Table 3) are lower than those obtained with an InceptionV3 architecture for surface images (98%, see Table 9) and mixed surface/section data (97%).

The results presented in this paper show the potential and interest of AI methods to automate the determination of the causes (lithogenesis) of the kidney stone formation. Nonetheless, additional tests should include other types of kidney stones with mixed composition to make an automated recognition procedure fully usable in clinical settings. Other kidney stone types with a unique biochemical composition (as struvite and cystine) should also extend the database. Additionally, most works in the literature (including this one) make use of still images, which might limit the applicability of the computer vision systems proposed so far (the video sequences which are displayed on screens might be affected by motion blur and blood or debris can hide kidney stone parts).

Other solutions that might be of interest to improve the classification results can be based on few shot learning approaches for object recognition and instance segmentation, the size of the available dataset being relatively small. Also, when more stones types are included in the dataset, the proposed models might benefit from online or active learning techniques for adapting to new settings (for instance, kidney stones from people from countries with very different weather, an aspect that has not been studied so far). Furthermore, training deep learning models using images in other color spaces (as the *HSV* or *HSI* color spaces) is another promising area of research, as the obtained results can be more robust and smaller the deep-learning networks could be deployed, speeding up the inference time [71].

Finally, "automated medical report generation" [72] can explain the decision taken by a model and favor translational medicine. Indeed, such techniques associate the visual features extracted from the images with text information to generate reports based on a learned vocabulary. Such reports justify the decision taken by the network with the terminology employed by the biologists who visually analyze the images during a MCA.

COMPLIANCE WITH ETHICAL APPROVAL
The images were captured in medical procedures following the ethical principles outlined in the Helsinki Declaration of 1975, as revised in 2000, with the consent of the patients.

REFERENCES
[1] J. Cloutier, L. Villa, O. Traxer, and M. Daudon, "Kidney stone analysis: 'Give me your stone, I will tell you who you are!'" *World J. Urol.*, vol. 33, no. 2, pp. 157–169, 2015.
[2] G. P. Kasidas, C. T. Samuell, and T. B. Weir, "Renal stone analysis: Why and how?" *Ann. Clin. Biochem., Int. J. Lab. Med.*, vol. 41, no. 2, pp. 91–97, Mar. 2004.
[3] P. M. Hall, "Nephrolithiasis: Treatment, causes, and prevention," *Cleveland Clinic J. Med.*, vol. 76, no. 10, pp. 583–591, Oct. 2009.
[4] S. F. R. D. Silva, D. C. D. Matos, S. L. D. Silva, E. D. F. Daher, H. D. H. Campos, and C. A. B. D. Silva, "Chemical and morphological analysis of kidney stones: A double-blind comparative study," *Acta Cirurgica Brasileira*, vol. 25, no. 5, pp. 444–448, Oct. 2010.
[5] M. Daudon and P. Jungers, "Stone composition and morphology: A window on etiology," in *Urolithiasis*. Berlin, Germany: Springer, 2012, pp. 113–140.
[6] C. D. Scales, A. C. Smith, J. M. Hanley, and C. S. Saigal, "Prevalence of kidney stones in the United States," *Eur. Urol.*, vol. 62, no. 1, pp. 160–165, Jul. 2012.
[7] A. Viljoen, R. Chaudhry, and J. Bycroft, "Renal stones," *Ann. Clin. Biochem.*, vol. 56, no. 1, pp. 15–27, 2019.

[8] G. Kartha, J. C. Calle, G. S. Marchini, and M. Monga, "Impact of stone disease: Chronic kidney disease and quality of life," *Urologic Clinics*, vol. 40, no. 1, pp. 135–147, 2013.

[9] J. I. Friedlander, J. A. Antonelli, and M. S. Pearle, "Diet: From food to stone," *World J. Urol.*, vol. 33, no. 2, pp. 179–185, Feb. 2015.

[10] V. Estrade, K. Bensalah, J.-P. Bringer, E. Chabannes, X. Carpentier, P. Conort, E. Denis, B. Doré, J. R. Gautier, H. Hadjadj, J. Hubet, A. Hoznek, E. Lechevallier, P. Meria, P. Mozer, C. Saussine, L. Yonneau, and O. Traxer, "Place de l'urétérorénoscopie souple en première intention pour le traitement des calculs du rein. Résultats de l'enquête de pratique du comité lithiase de l'AFU réalisée en 2011," *Progrès en Urologie*, vol. 23, no. 1, pp. 22–28, Jan. 2013.

[11] M. Daudon and P. Jungers, "Clinical value of crystalluria and quantitative morphoconstitutional analysis of urinary calculi," *Nephron Physiol.*, vol. 98, no. 2, pp. 31–36, Oct. 2004.

[12] V. Estrade, M. Daudon, O. Traxer, and P. Méria, "Pourquoi l'urologue doit savoir reconnaître un calcul et comment faire? Les bases de la reconnaissance endoscopique," *Progrès en Urologie-FMC*, vol. 27, no. 2, pp. 26–35, Jun. 2017.

[13] J. Serrat, F. Lumbreras, F. Blanco, M. Valiente, and M. López-Mesas, "MyStone: A system for automatic kidney stone classification," *Exp. Syst. Appl.*, vol. 89, pp. 41–51, Dec. 2017.

[14] A. T. Amado, "Metric learning for kidney stone classification," Universitat Autònoma de Barcelona, Titulació, Grau en Enginyeria Informàtica, Barcelona, Spain, Tech. Rep., 2018.

[15] K. M. Black, H. Law, A. Aldoukhi, J. Deng, and K. R. Ghani, "Deep learning computer vision algorithm for detecting kidney stone composition," *BJU Int.*, vol. 125, no. 6, pp. 920–924, Jun. 2020.

[16] A. Martínez, D.-H. Trinh, J. El Beze, J. Hubert, P. Eschwege, V. Estrade, L. Aguilar, C. Daul, and G. Ochoa, "Towards an automated classification method for ureteroscopic kidney stone images using ensemble learning," in *Proc. 42nd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2020, pp. 1936–1939.

[17] V. Estrade, M. Daudon, E. Richard, J. Bernhard, F. Bladou, G. Robert, and B. D. de Senneville, "Towards automatic recognition of pure and mixed stones using intra-operative endoscopic digital images," *BJU Int.*, vol. 129, no. 2, pp. 234–242, Feb. 2022.

[18] M. Mendez-Ruiz, F. Lopez-Tiro, D. Flores-Araiza, J. El-Beze, G. Ochoa-Ruiz, M. Gonzalez-Mendoza, J. Hubert, A. Mendez-Vazquez, and C. Daul, "On the generalization capabilities of FSL methods through domain adaptation: A case study in endoscopic kidney stone image classification," in *Proc. Mexican Int. Conf. Artif. Intell.* Cham, Switzerland: Springer, 2022, pp. 249–263.

[19] E. Villalvazo-Avila, F. Lopez-Tiro, J. El-Beze, J. Hubert, M. Gonzalez-Mendoza, G. Ochoa-Ruiz, and C. Daul, "Improved kidney stone recognition through attention and multi-view feature fusion strategies," 2022, *arXiv:2211.02967*.

[20] D. Flores-Araiza, F. Lopez-Tiro, J. El-Beze, J. Hubert, M. Gonzalez-Mendoza, G. Ochoa-Ruiz, and C. Daul, "Deep prototypical-parts ease morphological kidney stone identification and are competitively robust to photometric perturbations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 295–304.

[21] M. Corrales, S. Doizi, Y. Barghouthy, O. Traxer, and M. Daudon, "Classification of stones according to Michel daudon: A narrative review," *Eur. Urol. Focus*, vol. 7, no. 1, pp. 13–21, Jan. 2021.

[22] A. Basiri, M. Taheri, and F. Taheri, "What is the state of the stone analysis techniques in urolithiasis?" *Urol. J.*, vol. 9, no. 2, pp. 445–454, 2012.

[23] T. Kawahara, H. Miyamoto, H. Ito, H. Terao, M. Kakizoe, Y. Kato, H. Ishiguro, H. Uemura, M. Yao, and J. Matsuzaki, "Predicting the mineral composition of ureteral stone using non-contrast computed tomography," *Urolithiasis*, vol. 44, no. 3, pp. 231–239, Jun. 2016.

[24] G. Hidas, R. Eliahou, M. Duvdevani, P. Coulon, L. Lemaitre, O. N. Gofrit, D. Pode, and J. Sosna, "Determination of renal stone composition with dual-energy CT: In vivo analysis and comparison with X-ray diffraction," *Radiology*, vol. 257, no. 2, pp. 394–401, Nov. 2010.

[25] C. J. McCarthy, V. Baliyan, H. Kordbacheh, Z. Sajjad, D. Sahani, and A. Kambadakone, "Radiology of renal stone disease," *Int. J. Surg.*, vol. 36, pp. 638–646, Dec. 2016.

[26] V. K. Singh and P. K. Rai, "Kidney stone analysis techniques and the role of major and trace elements on their pathogenesis: A review," *Biophysical Rev.*, vol. 6, nos. 3–4, pp. 291–310, Dec. 2014.

[27] H. Alenezi and J. D. Denstedt, "Flexible ureteroscopy: Technological advancements, current indications and outcomes in the treatment of urolithiasis," *Asian J. Urol.*, vol. 2, no. 3, pp. 133–141, Jul. 2015.

[28] M. Daudon, A. Dessombz, V. Frochot, E. Letavernier, J.-P. Haymann, P. Jungers, and D. Bazin, "Comprehensive morpho-constitutional analysis of urinary stones improves etiological diagnosis and therapeutic strategy of nephrolithiasis," *Comp. Rendus Chimie*, vol. 19, nos. 11–12, pp. 1470–1491, Nov. 2016.

[29] R. Castañeda-Argáiz, J. Cloutier, L. Villa, and O. Traxer, "Evolution of endourology and flexible ureterorenoscopy, can they be useful to urologists to clarify stone composition and morphology?" *Comp. Rendus Chimie*, vol. 19, nos. 11–12, pp. 1590–1596, Nov. 2016.

[30] P. C. De Groen, "History of the endoscope [scanning our past]," *Proc. IEEE*, vol. 105, no. 10, pp. 1987–1995, Oct. 2017.

[31] E. Brain, R. M. Geraghty, C. E. Lovegrove, B. Yang, and B. K. Somani, "Natural history of post-treatment kidney stone fragments: A systematic review and meta-analysis," *J. Urol.*, vol. 206, no. 3, pp. 526–538, Sep. 2021.

[32] J. C. Williams, G. Gambaro, A. Rodgers, J. Asplin, O. Bonny, A. Costa-Bauzá, P. M. Ferraro, G. Fogazzi, D. G. Fuster, and D. S. Goldfarb, "Urine and stone analysis for the investigation of the renal stone former: A consensus conference," *Urolithiasis*, vol. 49, no. 1, pp. 1–16, Feb. 2021.

[33] M. Taratkin, E. Laukhtina, N. Singla, A. Tarasov, T. Alekseeva, M. Enikeev, and D. Enikeev, "How lasers ablate stones: In vitro study of laser lithotripsy in different environments," *J. Endourol.*, vol. 35, no. 6, pp. 931–936, 2021.

[34] E. Ventimiglia, F. Pauchard, F. Quadrini, S. Sindhubodee, H. Kamkoum, A. Jiménez Godínez, S. Doizi, and O. Traxer, "High- and low-power laser lithotripsy achieves similar results: A systematic review and meta-analysis of available clinical series," *J. Endourol.*, vol. 35, no. 8, pp. 1146–1152, Aug. 2021.

[35] V. De Coninck, E. X. Keller, and O. Traxer, "Metabolic evaluation: Who, when and how often," *Current Opinion Urol.*, vol. 29, no. 1, pp. 52–64, 2019.

[36] R. Siener, N. Buchholz, M. Daudon, B. Hess, T. Knoll, P. J. Osther, J. Reis-Santos, K. Sarica, O. Traxer, and A. Trinchieri, "Quality assessment of urinary stone analysis: Results of a multicenter study of laboratories in Europe," *PLoS ONE*, vol. 11, no. 6, Jun. 2016, Art. no. e0156606.

[37] G. Sampogna, D. Basic, P. Geavlete, J. A. Galán Llopis, J. Reis Santos, I. Saltirov, K. Sarica, S. Stavridis, A. Skolarikos, and A. Trinchieri, "Identificación endoscópica de la composición de los cálculos urinarios: Un estudio del Southeastern group for lithiasis research (SEGUR 2)," *Actas Urológicas Españolas*, vol. 45, no. 2, pp. 154–159, Mar. 2021.

[38] F. Blanco, M. López-Mesas, S. Serranti, G. Bonifazi, J. Havel, and M. Valiente, "Hyperspectral imaging based method for fast characterization of kidney stone types," *J. Biomed. Opt.*, vol. 17, no. 7, Jul. 2012, Art. no. 0760271.

[39] R. Grosjean, M. Daudon, M. F. Chammas, M. Claudon, P. Eschwege, J. Felblinger, and J. Hubert, "Pitfalls in urinary stone identification using CT attenuation values: Are we getting the same information on different scanner models?" *Eur. J. Radiol.*, vol. 82, no. 8, pp. 1201–1206, Aug. 2013.

[40] Y. Kazemi and S. A. Mirroshandel, "A novel method for predicting kidney stone type using ensemble learning," *Artif. Intell. Med.*, vol. 84, pp. 117–126, Jan. 2018.

[41] S. Qadri, "Role of machine vision for identification of kidney stones using multi features analysis," *Lahore Garrison Univ. Res. J. Comput. Sci. Inf. Technol.*, vol. 5, no. 3, pp. 1–14, Sep. 2021.

[42] C. S. Kolli, M. P. Raghunath, S. Meenakshi, K. Maheswari, C. F. Britto, and S. Kushwaha, "Efficient development of supervised learning algorithm for kidney stone prediction," in *Proc. Int. Conf. Inventive Comput. Technol. (ICICT)*, Jul. 2022, pp. 1373–1379.

[43] M. Sheikhi, S. Sina, M. Karimipourfard, and F. K. Shoushtari, "A study toward automatic identification of renal stone composition in single-energy or ultra-low-dose CT scan using deep neural networks," *Iranian J. Radiol.*, vol. 20, no. 2, pp. 1–13, May 2023.

[44] N. E. York, M. Zheng, H. M. Elmansy, M. E. Rivera, A. E. Krambeck, and J. E. Lingeman, "Stone-free outcomes of flexible ureteroscopy for renal calculi utilizing computed tomography imaging," *Urology*, vol. 124, pp. 52–56, Feb. 2019.

[45] F. Lopez, A. Varelo, O. Hinojosa, M. Mendez, D.-H. Trinh, Y. ElBeze, J. Hubert, V. Estrade, M. Gonzalez, G. Ochoa, and C. Daul, "Assessing deep learning methods for the identification of kidney stones in endoscopic images," in *Proc. 43rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Nov. 2021, pp. 2778–2781.

[46] J. El Beze, C. Mazeaud, C. Daul, G. Ochoa-Ruiz, M. Daudon, P. Eschwège, and J. Hubert, "Evaluation and understanding of automated urinary stone recognition methods," *BJU Int.*, vol. 130, no. 6, pp. 786–798, Dec. 2022.

[47] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[48] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Mach. Intell.*, vol. 1, no. 5, pp. 206–215, May 2019.

[49] W. Zhu, R. Zhou, Y. Yuan, C. Timothy, R. Jain, and J. Luo, "SegPrompt: Using segmentation map as a better prompt to finetune deep models for kidney stone classification," 2023, *arXiv:2303.08303*.

[50] S. Gupta, S. Ali, L. Goldsmith, B. Turney, and J. Rittscher, "Multi-class motion-based semantic segmentation for ureteroscopy and laser lithotripsy," *Computerized Med. Imag. Graph.*, vol. 101, Oct. 2022, Art. no. 102112.

[51] L. A. Fitri, F. Haryanto, H. Arimura, C. YunHao, K. Ninomiya, R. Nakano, M. Haekal, Y. Warty, and U. Fauzi, "Automated classification of urinary stones based on microcomputed tomography images using convolutional neural network," *Phys. Medica*, vol. 78, pp. 201–208, Oct. 2020.

[52] M. Mediouni, D. R. Schlatterer, H. Madry, M. Cucchiarini, and B. Rai, "A review of translational medicine. The future paradigm: How can we connect the orthopedic dots better?" *Current Med. Res. Opinion*, vol. 34, no. 7, pp. 1217–1229, Jul. 2018.

[53] V. Jahrreiss, J. Veser, C. Seitz, and M. Özsoy, "Artificial intelligence: The future of urinary stone management?" *Current Opinion Urol.*, vol. 30, no. 2, pp. 196–199, 2020.

[54] D. H. Trinh, C. Daul, W. Blondel, and D. Lamarque, "Mosaicing of images with few textures and strong illumination changes: Application to gastroscopic scenes," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 1263–1267.

[55] S. Ali, F. Zhou, B. Braden, A. Bailey, S. Yang, G. Cheng, P. Zhang, X. Li, M. Kayser, and R. D. Soberanis-Mukul, "An objective comparison of detection and segmentation algorithms for artefacts in clinical endoscopy," *Sci. Rep.*, vol. 10, no. 1, pp. 1–15, Feb. 2020.

[56] V. Estrade, B. Denis de Senneville, P. Meria, C. Almeras, F. Bladou, J. Bernhard, G. Robert, O. Traxer, and M. Daudon, "Toward improved endoscopic examination of urinary stones: A concordance study between endoscopic digital pictures vs microscopy," *BJU Int.*, vol. 128, no. 3, pp. 319–330, Sep. 2021.

[57] A. Hatamizadeh, D. Sengupta, and D. Terzopoulos, "End-to-end trainable deep active contour models for automated image segmentation: Delineating buildings in aerial imagery," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 730–746.

[58] G. Lemaitre, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 559–563, 2017.

[59] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform manifold approximation and projection for dimension reduction," 2018, *arXiv:1802.03426*.

[60] C. Daul, R. Rösch, and B. Claus, "Building a color classification system for textured and hue homogeneous surfaces: System calibration and algorithm," *Mach. Vis. Appl.*, vol. 12, no. 3, pp. 137–148, Oct. 2000.

[61] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sànchez, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017.

[62] H. E. Kim, A. Cosa-Linan, N. Santhanam, M. Jannesari, M. E. Maros, and T. Ganslandt, "Transfer learning for medical image classification: A literature review," *BMC Med. Imag.*, vol. 22, no. 1, p. 69, Dec. 2022.

[63] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM Comput. Surv.*, vol. 53, no. 3, pp. 1–34, May 2021.

[64] H. Guan and M. Liu, "Domain adaptation for medical image analysis: A survey," *IEEE Trans. Biomed. Eng.*, vol. 69, no. 3, pp. 1173–1185, Mar. 2022.

[65] T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey, "Meta-learning in neural networks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5149–5169, Sep. 2022.

[66] M. De Lange, R. Aljundi, M. Masana, S. Parisot, X. Jia, A. Leonardis, G. Slabaugh, and T. Tuytelaars, "A continual learning survey: Defying forgetting in classification tasks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3366–3385, Jul. 2022.

[67] M. Masana, X. Liu, B. Twardowski, M. Menta, A. D. Bagdanov, and J. van de Weijer, "Class-incremental learning: Survey and performance evaluation on image classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 5, pp. 5513–5533, May 2023.

[68] M. Ilse, J. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *Proc. Mach. Learn. Res.*, vol. 80, J. Dy and A. Krause, Eds. Stockholm Sweden, Jul. 2018, pp. 2127–2136. [Online]. Available: https://proceedings.mlr.press/v80/ilse18a.html

[69] A. Canziani, A. Paszke, and E. Culurciello, "An analysis of deep neural network models for practical applications," 2016, *arXiv:1605.07678*.

[70] *Visualization Tool for Pattern Recognition*. Accessed: Apr. 7, 2023. [Online]. Available: https://github.com/ivangarcia88/visualization_tool_for_pattern_recognition

[71] R. M. Anwer, F. S. Khan, and J. Laaksonen, "Compact deep color features for remote sensing scene classification," *Neural Process. Lett.*, vol. 53, no. 2, pp. 1523–1544, Apr. 2021.

[72] L. Gong, M. Wang, L. Shu, J. He, B. Qin, J. Xu, W. Su, D. Dong, H. Hu, J. Tian, and P. Zhou, "Automatic captioning of early gastric cancer using magnification endoscopy with narrow-band imaging," *Gastrointestinal Endoscopy*, vol. 96, no. 6, pp. 929–942, Dec. 2022.

FRANCISCO LOPEZ-TIRO (Member, IEEE) was born in Puebla, Mexico, in 1993. He received the M.Sc. degree in biomedical sciences from Instituto Nacional de Astrofísica, Óptica y Electrónica, Puebla, in 2019. He is currently pursuing the Ph.D. degree in computer science with Tecnólogico de Monterrey, Guadalajara, Mexico, and Université de Lorraine, Nancy, France. Since 2022, he has been a Lecturer with the School of Engineering and Sciences, Tecnologico de Monterrey. His research interest includes the development of AI methods for medical applications.

VINCENT ESTRADE was an Urologist with the University Hospital Bordeaux, in 2017, and the General Hospital of Angoulême, in 2003. Since 2015, he has been working on the development of endoscopic stone recognition for a better etiological approach to lithiasis by urologists and artificial intelligence. Recently, he has participated in the redaction of new French guidelines in lithiasis, particularly for the ESR Chapter. Since 2009, he has been a member of the French Lithiasis Committee.

JACQUES HUBERT trained as an Urologist with the University Hospital of Nancy, in 1984. He was appointed as a Professor with the University of Lorraine, in 1996. He has been a Teacher–Researcher with the IADI-UHP-INSERM Team (U 1254) of Pr. J. Felblinger, since 2004. He was the Head of the University Department of Urology, from 2008 to 2019. He was a world pioneer in robotic surgery when he started this activity in 2000 and developed pedagogical research in this field. Since 1994, he has been involved in the diagnosis and treatment of urinary stones. He is a member of the French Lithiasis Committee. He has also worked on imaging in urology with a synthesis book on the subject, in 2003.

DANIEL FLORES-ARAIZA was born in Oaxaca, Mexico, in 1990. He received the B.Eng. degree in mechatronic engineering with a concentration in applied robotics and the M.S. degree with a specialization in intelligent systems from Tecnologico de Monterrey, Puebla, Puebla, Mexico, in 2013 and 2019, respectively. He is currently pursuing the Ph.D. degree in computer science with Tecnologico de Monterrey, Guadalajara, Jalisco, Mexico. His research interests include obtaining insights from medical imaging classifiers through explainable methods, developing inherently interpretable medical imaging classifiers, and improving explanations for the outputs of classifiers with causal discovery and causal inference techniques.

GILBERTO OCHOA-RUIZ received the European Erasmus Mundus master's degree in computer vision and robotics from Heriot-Watt University at Edinburgh, Edinburgh, U.K., and the Ph.D. degree in electronic imaging and computer vision from Université de Bourgogne, France. He is currently an Associate Professor–Researcher in artificial intelligence and computer vision with the Advanced AI Group, Tecnologico de Monterrey. His main research interests include computer vision and image processing for applications in medical imaging and video analysis and the training and deployment of optimized deep learning models in constrained devices for edge computing applications.

MIGUEL GONZALEZ-MENDOZA received the Ph.D. degree in artificial intelligence from INSA-LAAS, Toulouse, France, in 2004. He is currently a Professor–Researcher with Tecnologico de Monterrey, Mexico. He is the author of more than 120 articles in JCR, congresses, and book chapters. His research interests include machine learning, data management, and computer vision applications, in which he has advised 13 Ph.D. theses and 30 master's theses. He is a member of the National System of Researchers (SNI) Level 2. He is a Regular Member of the Mexican Academy of Sciences and a Regular Member of the Mexican Academy of Computing. He has served as the general chair of national and international AI congresses and workshops. He leads several Mexican and European Union research projects (as the local chair). He is a Former President of the Mexican Society on Artificial Intelligence. He is a Young Scientist of the World Economic Forum for New Champions.

CHRISTIAN DAUL (Member, IEEE) received the Ph.D. degree in computer vision from Université Louis Pasteur, Strasbourg, France, in 1994. From 1990 to 1995, he was with Laboratoire des Sciences de l'Image, de l'Informatique et de la Télédétection (current laboratory name: ICube, UMR 7357 CNRS/Université de Strasbourg) before joining the Institute of Applied Mathematics of Kaiserslautern (ITWM, "Fraunhofer Institut für Techno- und Wirtschaftsmathematik," Germany), where he was a member of the Image Processing Group. Since October 1999, he has been with Centre de Recherche en Automatique de Nancy (CRAN), UMR 7039, CNRS, and Université de Lorraine (UL), Vandœuvre-lès-Nancy, France, where he is working in the area of endoscopic imaging (gastroscopy, colonoscopy, cystoscopy, and ureteroscopy). He is currently a University Professor with Ecole Européenne d'Ingénieurs en Génie des Matériaux, UL, where he is teaching in the fields of signal processing and computer science. His main research interests include image segmentation and classification, data registration, 3D data reconstruction, and 2D/3D image mosaicing.

● ● ●