Assessing deep learning methods for the identification of kidney stones composition in endoscopic images

Andres Varelo-Silgado[1], Francisco Lopez-Tiro[1], Mauricio Mendez-Ruiz [1] Jonathan El Beze[2], Miguel Gonzalez[1], Vincent Estrade[3], Gilberto Ochoa-Ruiz[1], Christian Daul[4]

[1]Tecnologico de Monterrey, School of Engineering and Sciences, Mexico
[2]CHU Nancy, Service d'urologie de Brabois, Nancy, France
[3] Centre Hospitalier Universitaire Pellegrin, Bordeaux, France
[4] Centre de Recherche en Automatique de Nancy, Univ. de Lorraine, France
varleo95@gmail.com, gilberto.ochoa@tec.mx, christian.daul@univ-lorraine.fr

Abstract

Current analysis of kidney stones through morpho-constitutional assessments makes it is possible to establish treatments to reduce the recurrence of kidney stones formation, but this process is seen as time-consuming and prone to errors by expert urologists. Thus, many practitioners have advocated for the introduction of automated AI-based visual identification methods to be deployed during the endoscopic exploration and stone extraction process. Such CADx tools could have a tremendous impact in the urologists workflow, providing immediate insights of the stone composition, and thus allowing timely hygiene-dietary advice after the operation. In this paper, we investigate the applicability of deep learning-based computer vision techniques for automatically classifying kidney stones for real-time support systems, attaining an average classification precision of 97% using Inception v3 in a challenging dataset comprised of images of four types of stones acquired in vivo.

1. Introduction

Kidney stones with a diameter of more than a few millimeters cannot usually leave the urinary tract, causing severe pain. During a standard ureteroscopy, kidney stones are visualized using digital ureteroscopes and broken into fragments using a laser. These fragments are extracted from the urinary tract and their biochemical constitution is analyzed to understand the causes (i.e. lithogenesis) leading to the formation of the kidney stones and to prevent relapses with appropriate treatment (e.g., diet, drugs [4]). The class of the extracted kidney stones can be visually recognized by studying the textures, appearance, and colours of the surfaces and sections of the fragments using a microscope. Complementary information about the crystalline composition can then be determined using infrared-spectrophotometry [6].

However, in numerous hospitals, the result of such analyses [2] is usually available a couple of weeks after the procedure. A recent study [3] has shown that the results of such visual recognition from endoscopic images by an expert is strongly correlated with this analysis. A visual in vivo type recognition in endoscopic images could save precious time since the fragments can be pulverized and the SPIR analysis can be avoided. However, most urologists are not trained to perform this kidney stone type recognition efficiently and such a task is also strongly operator dependent.

Despite the inherent advantages and potential of AI-based automated and objective kidney stone recognition tools, only a few studies have been published in this domain. Both a classical approach (in [9] a Random Forest classifier exploits histograms of RGB colours and LBP encoding textures) and a deep learning method [10] have been investigated, but they obtained rather moderate classification results (a mean accuracy of 63% and 74% was obtained over four and five classes for [9] and [10], respectively). The authors in [1] clearly improved the classification results on five kidney stone types using the ResNet-101 architecture (the leave-one-out cross-validation led to recall values from 71% up to 94% according to the class). The main limitation of these previous works lies in the fact that the methods were tested on ex-vivo images obtained in very controlled acquisition conditions and without endoscopes.

In ureteroscopic in vivo data, the images are affected by blur, strong illumination changes between acquisitions, as well as by reflections, whereas the viewpoints are not easy to optimally adjust.

However, these works have shown the feasibility of automating kidney stone classification. The aim of this contribution is to assess whether or not CNN-based solutions can further improve the classification of kidney stones acquired with ureteroscopes.

2. Material and methods

2.1. Clinical image dataset

The employed dataset includes 177 kidney stone images which were acquired and annotated by an expert urologist, Prof. Vincent Estrade. The results of this visual classification were statistically confirmed by the concordance study in [3]. The dataset consists of 90 fragment surface images and 87 fragment cross-section images of the four kidney stone types with the highest incidence: whewellite (COM), weddellite (COD), acid uric (AU), and brushite (BRU). These clinical images were captured using either the URF-V or URF-V2 endoscopes from Olympus. Images of this dataset are shown in Fig. 1.

2.2. Patch extraction and data augmentation

As confirmed by the results of previous works [9, 10, 1, 7], image patches with a minimal size enable to capture enough texture and colour information for classification purposes. The use of image patches instead of the whole fragment surfaces and sections allows to increase the size of the training and test datasets. In order to avoid redundant information, the image areas including kidney stone fragments were scanned by square patches forming a regular grid whose neighbouring cells have a maximal overlap of twenty pixels. However, in previous works, the optimal size of these patches has not been studied.

The patch size was a hyper-parameter that was adjusted during the training of the ML models presented in Section 3. The best size value was obtained after several ablation studies using four patch areas (64x64, 128x128, 256x256, and 512x512 pixels), by monitoring the precision and loss curves for each patch size (see Fig. 2).

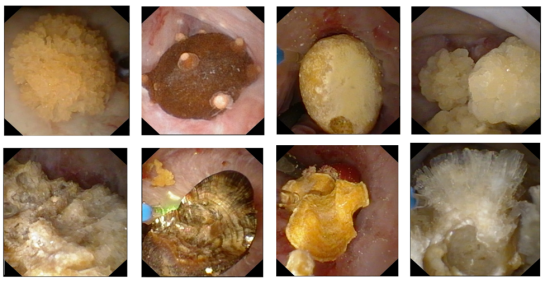

Figure 1. Examples of in vivo kidney stone images. From the left to the right: COM (whewellite), COD (weddellite), uric acid and brushite. Surface and section images are in the upper and lower line, respectively.

Table 1. Dataset metrics before and after sampling.

Stone type		Acquired images		Number of patches
View	Class	Number	Presence (%)	
Surface	COM	30	31.9	870
	COD	32	34.1	920
	Uric Acid	18	19.1	470
	Brushite	14	14.9	420
	Total	**94**	**100**	**2680**
Section	COM	27	31.0	820
	COD	28	32.2	780
	Uric Acid	18	20.7	460
	Brushite	14	16.1	410
	Total	**87**	**100**	**2470**

The best trade-off in terms of accuracy and recall was obtained with patches of 256x256 pixels. This patch size was used for the results given in Section 3. As shown in Table 1, 2680 surface and 2470 section patches were obtained. As the number of patches in unbalanced, we performed random weighted over-sampling with replacement.

Afterward, we performed data augmentation by applying different combinations of geometrical transformations to the original patches: flipping, affine transformations, and perspective distortions increasing the samples from 5,400 to 43,200 (10% were hold out for test purposes).

2.3. Feature extraction and classification

The aim of this paper is notably to compare the deep-learning methods to the best "classical" classification methods (i.e., non-DL based approaches). In [7], the feature vector (based on HSI colour energies and rotation invariant LBP histograms) was identified as leading to the highest separability of the kidney stones. Among the tested shallow ML methods, Random Forest and XGBoost obtained the highest precision and recall (see Table 2) for in vivo kidney-stone images. For these two classical ML methods, the results given in Section III were obtained by a hyper-

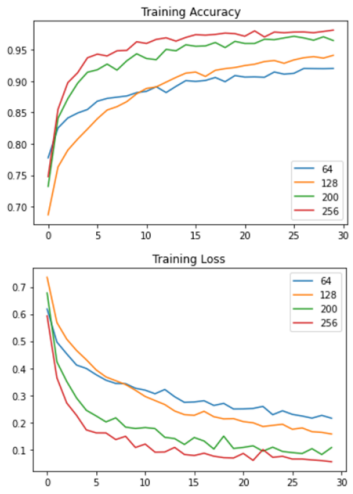

Figure 2. Ablation study with changing patch sizes.

Table 2. Weighted average metrics comparison for section and surface patches, as well as mixed patches.

Classifier	Surface		Section		Mixed	
	P	R	P	R	P	R
Random Forest	0.87	0.82	0.82	0.82	**0.91**	**0.91**
XGboost	0.93	0.93	0.89	0.89	**0.96**	**0.96**
AlexNet	0.93	0.95	0.83	0.82	0.92	0.92
VGG19	0.95	0.96	0.91	0.92	0.94	0.92
Inception	**0.98**	**0.97**	**0.94**	**0.96**	**0.97**	**0.98**

parameter tuning using 10 fold cross-validation (CV), averaging the results over 5 runs (for the non DL models we used leave-one-out CV due to the low number of samples).

DL architectures (AlexNet, VGG16, and Inception v3) were adapted for this contribution because theirs feature extractor backbones are optimal only with large datasets. The proposed method leverages the benefits of transfer learning by exploiting CNN backbones pre-trained with ImageNet. The fully connected (FC) layers of the original backbones are replaced by a custom FC layer of 25 channels, followed by a Batch Normalization, a ReLU activation function, another FC layer of 256 channels and a softmax layer with 4 class outputs. These two were randomly initialized, and connected to a softmax layer for predicting the patch class. During the training process, the weights in the convolution layers were fixed and only the FC layers were updated.

For all the reported experiments, we made use of Pytorch 1.7.0 and CUDA 10.1. The learning rates were obtained using the Pytorch Lightning 1.0.2 optimizer, yielding the following learning rate values: 0.0001 (AlexNet), 0.00005 (VGG16) and 0.0006 (Inception V3). We used the ADAM optimizer, a batch size of 64 and early stopping for all the experiments, whose results are discussed in the next section.

3. Results and discussion

We performed various experiments to assess the ability of the tested ML models to predict the kidney stone class, with surface and section patches separately, or mixed, as is it done in the morpho-constitutional analysis procedure [2]. To do so, all models were trained three times, with i) section patches, ii) surface patches, and iii) by mixing the two patches types. Precision (P) and recall (R) metrics are determined for each class individually.

The results obtained for Random Forest led to very similar performances as those reported in [7], showing that the class balancing compensates the increase of the number of classes. Moreover, still with respect to [7], an additional classifier was tested: XGBoost yielded significant results even comparable to those obtained with the DL-AlexNet model. In fact, only the Inception v3 model outperforms XGBoost. This model exhibits the highest average precision and recall for all classes and patch types.

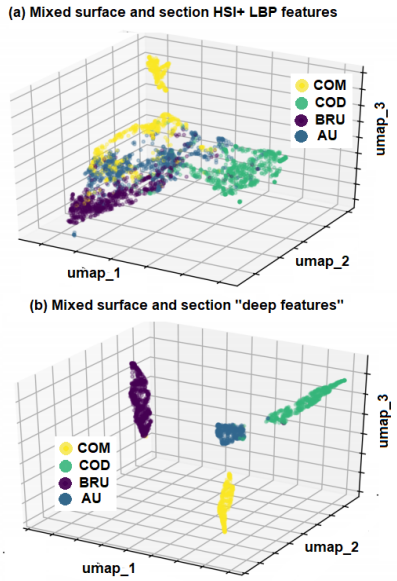

Figure 3. Feature visualization using the UMAP for (a) the HSI and LBP features and (b) the "deep features".

Table 2 shows that mixing surface and section information leads to the best results, whereas the three DL-based methods exhibit globally the best overall results when exploiting only surface images (confirming the results of the concordance study in [3]).

Furthermore, Figure 3(a) provides an UMAP visualization [8] which illustrates the class separability achieved using only the three most discriminant dimensions (umap1 to umap3) obtained after the dimensionality reduction of the HSI-LBP feature space. Fig. 3(b) shows that the same UMAP dimensionality reduction applied on the "deep features" produces tighter clusters and larger inter-class distances than in Fig.3(a).

4. Conclusions

In this work, we showed that is possible to train ML models for predicting kidney stone composition from digital images obtained from ureteroscopes. These results demonstrate that AI technology can be included in the urologists' workflow for identifying the causes (lithogenesis) of the kidney stone formation [5], because precious morphological information used for diagnosis can be extracted before proceeding to pulverizing the stone, speeding up preventive diagnosis measure. Furthermore, this method could be used to automatically adjust the laser settings during the ureteroscopy. However, as is previous works, our experiments shall include also other types of stones with mixed compositions to make it fully usable in clinical settings. Also, we made use of still images, which might limit the applicability of the proposed method in real interventions using video data which can be affected by body movements, surgical instruments, blood and debris.

References

[1] Kristian M Black, Hei Law, Ali Aldouhki, Jia Deng, and Khurshid R Ghani. Deep learning computer vision algorithm for detecting kidney stone composition. *BJU international*, 2020. 1, 2

[2] Michel Daudon, Arnaud Dessombz, Vincent Frochot, Emmanuel Letavernier, Jean-Philippe Haymann, Paul Jungers, and Dominique Bazin. Comprehensive morpho-constitutional analysis of urinary stones improves etiological diagnosis and therapeutic strategy of nephrolithiasis. *Comptes Rendus Chimie*, 19(11-12):1470–1491, 2016. 1, 3

[3] Vincent Estrade, Baudouin Denis de Senneville, Paul Meria, Christophe Almeras, Franck Bladou, Jean-Christophe Bernhard, Gregoire Robert, Olivier Traxer, and Michel Daudon. Toward improved endoscopic examination of urinary stones: a concordance study between endoscopic digital pictures vs. microscopy. *BJU international*, 2020. 1, 2, 3

[4] Justin I Friedlander, Jodi A Antonelli, and Margaret S Pearle. Diet: from food to stone. *World journal of urology*, 33(2):179–185, 2015. 1

[5] Victoria Jahrreiss, Julian Veser, Christian Seitz, and Mehmet Özsoy. Artificial intelligence: the future of urinary stone management? *Current opinion in urology*, 30(2):196–199, 2020. 3

[6] Aslam Khan. Prevalence, pathophysiological mechanisms and factors affecting urolithiasis. *International urology and nephrology*, 50(5):799–806, 2018. 1

[7] Adriana Martínez, Dinh-Hoan Trinh, Jonathan El Beze, Jacques Hubert, Pascal Eschwege, Vincent Estrade, Lina Aguilar, Christian Daul, and Gilberto Ochoa. Towards an automated classification method for ureteroscopic kidney stone images using ensemble learning. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1936–1939. IEEE, 2020. 2, 3

[8] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. 3

[9] Joan Serrat, Felipe Lumbreras, Francisco Blanco, Manuel Valiente, and Montserrat López-Mesas. mystone: A system for automatic kidney stone classification. *Expert Systems with Applications*, 89:41–51, 2017. 1, 2

[10] Alejandro Torrell Amado. Metric learning for kidney stone classification. 2018. 1, 2