2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

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

Francisco Lopez[1], Andres Varelo[1], Oscar Hinojosa[1], Mauricio Mendez[6], Dinh-Hoan Trinh[7], Yonathan ElBeze[4], Jacques Hubert[4,5], Vincent Estrade[3], Miguel Gonzalez[6], Gilberto Ochoa[6], Christian Daul[2]

Abstract— **Knowing the type (i.e., the biochemical composition) of kidney stones is crucial to prevent relapses with an appropriate treatment. During ureteroscopies, kidney stones are fragmented, extracted from the urinary tract, and their composition is determined using a morpho-constitutional analysis. This procedure is time-consuming (the morpho-constitutional analysis results are only available after several weeks) and tedious (the fragment extraction lasts up to an hour). Identifying the kidney stone type only with the in-vivo endoscopic images would allow for the dusting of the fragments and eneable early treatments, while the morpho-constitutional analysis is ready. Only few contributions dealing with the in vivo identification of kidney stones have been published. This paper discusses and compares five classification methods including deep convolutional neural networks (DCNN)-based approaches and traditional (non DCNN-based) ones. Even if the best method is a DCCN approach with a precision and recall of 98% and 97% over four classes, this contribution shows that an XGBoost classifier exploiting well-chosen feature vectors can closely approach the performances of DCNN classifiers for a medical application with a limited number of annotated data.**

Keywords: *Kidney stone recognition; endoscopy; deep learning; in vivo classification.*

I. 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 in order to understand the causes (i.e. lithogenesis) leading to the formation of the kidney stones and to prevent relapses with an appropriate treatment (e.g., diet, drugs [1]). 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 (FTIR) [2]. However, in numerous hospitals, the result of such analysis [3] is usually available only several weeks after the ureteroscopy, and removing the kidney stone fragments is a tedious task that can last up to an hour.

[1]IEEE Student Member, INAOE, UAG (Mexico), UP (Colombia)
[2]CRAN (Université de Lorraine and CNRS), F-54000 Nancy
[3]CHU Pellegrin place Amémlie Raba Léon F-33000 Bordeaux
[4]CHU Nancy, Service d'urologie de Brabois, F-54511 Nancy
[5]IADI-UL-INSERM (U1254), F-54511 Vandœuvre-lès-Nancy
[6]Tecnologico de Monterrey, Escuela de Ingeniería y Ciencias, México
[7]Viettel Cyberspace Center, Vietnam
Corresponding Authors:
gilberto.ochoa@tec.mx, christian.daul@univ-lorraine.fr

Furthermore, only very few, highly trained experts are able to recognize the type of a kidney stone by exclusively observing in vivo endoscopic images. A recent study [9] has shown that the results of such visual recognition from endoscopic images by an expert is strongly correlated with those of the morpho-constitutional analyses. A visual in vivo type recognition in endoscopic images could save precious time since the fragments can be pulverized, speeding up the process, without losing precious diagnostic information. However, most urologists are not trained to perform this kidney stone type recognition efficiently and such a task is also strongly operator dependent. In addition, it has been found that laser fragmentation can change the composition of kidney stones, which can bias the FTIR analyses [4].

Despite the inherent advantages of an automated and objective kidney stone recognition method, only few studies have been published in this domain (see Table I for an overview of the literature). Both classical approaches (in [5] a Random Forest classifier exploits histograms of RGB colours and local binary patterns -LBP- encoding rotation invariant textures) and deep learning methods [6] (based on a Siamese Convolutional Neural Network, CNN) 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 [5] and [6], respectively). The authors in [7] 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 on 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.

In a previous work [8], it was shown that by choosing an appropriate colour space (HSI) and by extracting colour and texture features exploited by classical classifiers such as a Random Forest tree or a KNN ensemble model, the results can be significantly improved (an accuracy of about 85% was obtained over 3 classes), even for patient data acquired with an endoscope in uncontrolled conditions. The aim of this contribution is to show how CNN-based solutions can further improve the classification of kidney stones acquired with ureteroscopes, even with a moderate number of images.

TABLE I

OVERVIEW OF THE DATA USED IN THE LITERATURE OF KIDNEY STONE CLASSIFICATION. SIX KIDNEY STONE TYPES WITH DIFFERENT COMPOSITIONS WERE TREATED: URIC ACID (UA), CALCIUM OXALATE MONOHYDRATE (COM), CALCIUM OXALATE DIHYDRATE (COD), STRUVITE (STR), CYSTINE (CYS) AND BRUSHITE (BRU). THE CLASSES AND THE ACQUISITION CONDITIONS ARE GIVEN FOR EACH CONTRIBUTION.

References	Kidney Stone Composition						Image Type		Acquisition Mode	
	UA	COM	COD	STR	CYS	BRU	Surface	Section	Ex Vivo	In Vivo
Serrat et al 2017 [5]	✓	✓	✓	✓	✓		✓	✓	✓	
Torrell et al 2018 [6]	✓	✓	✓	✓	✓	✓	✓		✓	
Black et al 2020 [7]	✓	✓		✓	✓	✓	✓	✓	✓	
Martinez et al 2020 [8]	✓	✓	✓				✓	✓		✓
This contribution	✓	✓	✓			✓	✓	✓		✓

The rest of this paper is structured as follows. Section II gives an overview of the acquired endoscopic image dataset and details the undertaken sampling and augmentation strategies to adapt the data to deep learning (DL) implementations. Section II also presents the model training approach. Section III compares the performances of the DL-strategies with those of conventional machine learning techniques. Finally, Section IV concludes the contribution and proposes perspectives.

II. MATERIAL AND METHODS

A. Clinical Image Dataset

The dataset includes 177 kidney stone images which were acquired and annotated by an expert, MD. Vincent Estrade (the data was collected following the ethical principles outlined in the Helsinki Declaration of 1975, as revised in 2000, with the consent of the patients). The results of this visual classification were statistically confirmed by the concordance study in [9]. The dataset consists of 90 fragment surface images and 87 fragment cross-section images of the four kidney stone types with the highest incidence (see the last row of Table II). These clinical images were captured using either the URF-V and URF-V2 endoscopes from Olympus, or with a BOA endoscope from Richard Wolf. Images of this dataset are shown in Fig. 1.

As in previous works [5-8], the classification is not performed on whole images, but using square patches localized on kidney stone surfaces or sections (surrounding tissues are not visible in the patches). In practice, kidney stone fragments have to be segmented prior to the classification. This process can be done either manually, or using automated methods, either in offline or online fashion.

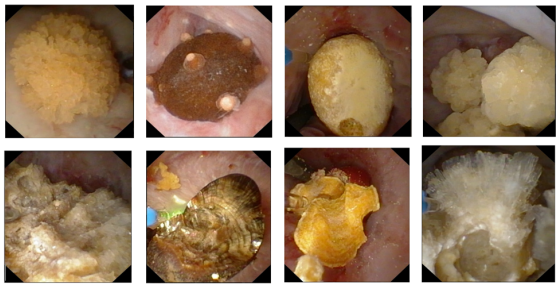

Fig. 1. Examples of in vivo kidney stone images. From the left to the right: COM (whewellite), COD (weddelite), uric acid and brushite. Surface and section images are in the upper and lower line, respectively.

B. Patch extraction and data augmentation

As confirmed by the results of previous works [5-8], 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 which was adjusted during the training of the machine learning models presented in Section II.C. The best size value was obtained after several ablation studies using four patch areas (64x64, 128x128, 256x256 and 512x512 pixels, respectively) and by monitoring the precision and loss curves for each patch size. 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 III. As shown in Table II, 2680 surface patches and 2470 section patches were obtained in all with this procedure.

As it can be observed in the last column of Table II, the resulting number of patches per class is imbalanced due to the changing fragment sizes and the number of images available per class. Two approaches were tested in order to balance the number of patches per class. In a "down-sampling approach", the number of patches of each class is reduced to that of the brushite class with lowest sample number (420 and 410 patches for the surface and section im-

Stone Type		Acquired Images		Number
View	Class	Number	Presence (%)	of patches
Surface	COM	30	31.9	870
	COD	32	34.1	920
	Uric Acid	18	19.1	470
	Brushite	14	14.9	420
	Total	**94**	**100.0**	**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**	**100.0**	**2470**

TABLE II

NUMBER OF ACQUIRED IMAGES AND OF THEIR (ALMOST) NON OVERLAPPING SQUARE PATCHES

ages, respectively). In this "down-sampling" process, patches were randomly removed from the COM, COD and uric acid classes. In contrast, in the "up-sampling approach", the number of images of the brushite, COD and AU classes is increased to match that of the COM class with the highest sample number (870 and 820 patches for surface and section fragments, respectively). New patches, which are not located on the initial grid of patches, are randomly extracted from the images to increase the sample number. Classification tests have shown that this "up-sampling approach" led to slightly better results in terms of accuracy, which means that even if redundant information is present, the increase of the patch number favours the accuracy of the classification. With the "up-sampling approach", 750 patches are available per class.

Then, the number of patches of each class was still increased since the performance of deep learning (DL) approaches relates to the amount of available training data. The data was augmented by applying different combinations of geometrical transformations to the original patches: patch flipping, affine transformations, and perspective distortions. The number of patches increased from 5,400 to 43,200 using this data augmentation (10% of the original patches were kept for test purposes). The patches were also "whitened" using the mean m_i and standard deviation σ_i of the colour values I_i in each channel ($I_i^w = (I_i – m_i)/\sigma_i$, with i = R, G, B). This dataset was split in three parts for the training, validation and test stages.

C. Feature extraction and classification

The aim of this paper is notably to compare the deep-leaning methods to the best "classical" classification methods (i.e., non-DL based approaches). In [8], 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 classical methods exploiting these features, Random Forest trees and XGBoost classifiers obtained the highest precision and recall (P and R, respectively in Table III) for in vivo kidney-stone images. For these two classical machine learning methods, the results given in Section III were obtained by a hyper-parameter tuning using a 10 fold cross-validation (CV) and by averaging the results over five runs (for the non DL models we used leave-one-out CV due to the low number of samples).

DL architectures of various levels of complexity (AlexNet, VGG16 and Inception v3) were adapted for this contribution. The performance of the feature extractor backbones of these models is optimal only when a large number of training images is available. 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. The two FC layers were randomly initialized and connected to a softmax layer for predicting the patch class. During the training process of the three DL models using the kidney stone patches, the weights in the convolution layers were fixed and only the weights in the FC layers were updated.

Classifier	Surface		Section		Mixed	
	P	R	P	R	P	R
R. 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	0.94	0.97
VGG19	0.95	0.96	0.91	0.92	0.95	0.96
Inception	**0.98**	**0.97**	**0.94**	**0.96**	**0.97**	**0.98**

TABLE III

WEIGHTED AVERAGE METRICS COMPARISON FOR SECTION AND SURFACE PATCHES, AS WELL AS MIXED PATCHES.

Classifier	COM		COD		UA		BRU	
	P	R	P	R	P	R	P	R
R. Forest	0.84	0.86	0.90	0.95	0.88	0.67	0.90	0.92
XGBoost	0.92	0.96	0.91	0.91	0.97	0.96	0.96	0.94
AlexNet	0.93	0.98	0.95	0.85	0.88	0.92	0.93	0.92
VGG19	0.97	0.97	0.92	0.93	0.93	0.83	0.94	0.92
Inception	**0.98**	**0.97**	**0.93**	**0.96**	**0.95**	**0.90**	**0.96**	**0.95**

TABLE IV

PRECISION (P) AND RECALL (R) VALUES OBTAINED OVER THE FOUR CLASSES WITH ALL CLASSIFIERS.

For all the reported experiments, we made use of Pytorch 1.7.0 and CUDA 10.1. The learning rates for each model 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.

III. RESULTS AND DISCUSSION

Various experiments were carried out for evaluating the machine learning methods described in section II.C using the patch data introduced in II.B. In particular, the ability of these models to predict the kidney stone class either based on surface and section patches taken separately, or by using simultaneously the two patch types, as is it classically done in the morpho-constitutional analysis procedure [3]. To do so, all models were trained three times, namely i) solely with the section patches, ii) only with the surface patches, and iii) by mixing the two patch types. Each section or surface patch was classified two times, i.e. by its dedicated model and by the model for mixed data. The well-known precision (P, see Tables III and IV) and recall (R) metrics are determined for each kidney stone class to assess the clinical applicability of the studied fragment classification methods.

The results obtained for the Random Forest classifier (see the first rows of Tables III and IV) led to very similar performances as those reported in [8], showing that the class balancing compensates the increase in the number of classes. Moreover, still with respect to [8], an additional classifier was tested: XGBoost which yielded significantly better results than the Random Forest classifier. These results are even comparable to those obtained with the DL-AlexNet model.

The results of the chosen deep learning (DL) models are summarized in the last three rows of Tables III and IV. It

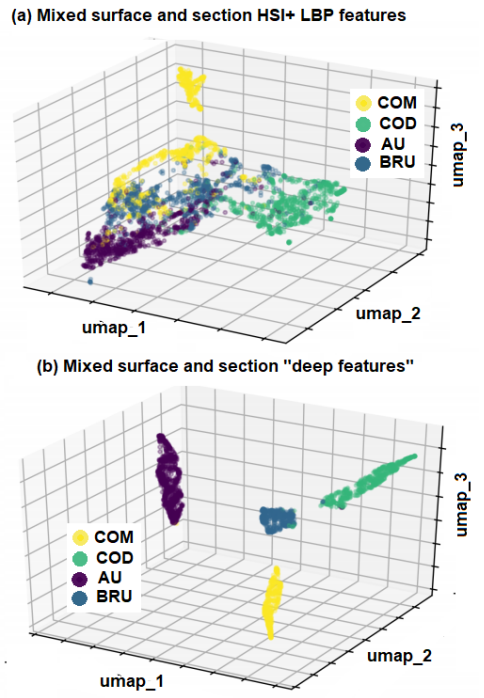

Fig. 2. Feature visualization using the UMAP dimensionality reduction technique for (a) the HSI and LBP features and (b) the "deep features".

is noticeable that only the Inception v3 model outperforms the best traditional machine learning method based on the XGBoost classifier. As seen in the last row of Table IV, this model exhibits the highest mean precision (P) and recall (R) for all classes and for all patch types (section patches, surface patches and mixed patch types).

It can be observed in Table III that for the classical models, the simultaneous use of surface and section patches lead 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 [9]). However, among the DL approaches, only Inception v3 has globally a better precision (P = 0.97) and recall (R = 0.98) than the XGBoost classifier (P = R = 0.96).

Figure 2(a) provides an UMAP visualisation [10] 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. In Fig. 2(b) it is noticeable that the same UMAP dimensionality reduction applied on the "deep features" produces tighter clusters and larger inter-class distances than in Fig. 2(a). This suggests that the superior performance of the DL models relates to the richness of the information extracted from the patches using efficient feature extraction backbones associated with transfer learning.

Inception v3 and XGBoost improve the precision and recall metrics by about 30% and 10% in comparison to the traditional approaches in [5] and [8], respectively. These two models outperform also the DL model tested in [7] on ex-vivo data, which exhibited a high average recall of 97%, but a lower average precision of 80%.

IV. CONCLUSION

In this work, it was shown that it is possible to effectively train DL models for predicting kidney stone types using a rather small dataset acquired during ureteroscopies. This study confirms that AI methods can be incorporated in the urologists' workflow for identifying the causes of the kidney stone formation [11]. Thus, the tedious fragment extraction phase can be avoided since pulverizing kidney stones becomes possible without losing the information necessary for diagnosis purposes. However, this work focused only on four classes of kidney stones. The proposed method should be improved to identify classes with a lower incidence and/or mixed composition. Additionally, all previous works including ours make use of "still" images. Classifying kidney stones using video data could further facilitate the urologist's work and represents a fertile area of research, but poses very complex challenges, as the regions of interest might be blurred due to the endoscope motion, affected by reflections, or occluded by surgical instruments, blood or debris.

ACKNOWLEDGMENTS

The authors wish to thank the AI Hub and the CIIOT at ITESM for their support for carrying the experiments reported in this paper in their NVIDIA's DGX computer.

We also thank the Verano de la Investigación Cientifica Delfin program for assisting Andres Varelo with a mobility grant. and CONACYT for the master scholarship for Oscar Hinojosa at UAG and Mauricio Mendez at Tec de Monterrey.

REFERENCES

[1] J.I. Friedlander, J.A. Antonelli and M.S. Pearle, Diet: from food to stone, World Journal of Urology, vol. 33, n. 1, p. 179 - 185, 2015

[2] A. Khan, Prevalence, pathophysiological mechanisms and factors affecting urolithiasis, International Urology and Nephrology, vol. 50, number 5, pages, 799 - 806, 2018

[3] 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, vol. 19, number 11, pages 1470-1491, 2016

[4] E.X. Keller, V. de Coninck, M. Audouin, S. Doizi, D. Bazin, M. Daudon and O. Traxer, Fragments and dust after Holmium laser lithotripsy with or without "Moses technology": How are they different?, Journal of Biophotonics, vol. 12, number 4, 2019

[5] J. Serrat, F. Lumbreras, F. Blanco, M. Valiente and M. Lopez-Mesas, MyStone: A system for automatic kidney stone classification, Expert Systems with Applications, no. 89, pp. 45–51, 2017.

[6] A. Torrell, Metric learning for kidney stone classification, BSc. thesis, Escola D'Enginyeria, Universitat Autònoma de Barcelona, 2018.

[7] K.M. Black, L. Hei, A. Aldoukhi, J. Deng and K.R. Ghani, Deep learning computer vision algorithm for detecting kidney stone composition, BJU International, vol. 125, number 6, pages 920-924, 2020

[8] A. Martinez, D.-H. Trinh, J. El Beze, J. Hubert, P. Eschwege, V. Estrade, L. Aguilar, C. Daul, G.Ochoa, Towards an automated classification method for ureteroscopic kidney stone images using ensemble learning, 42nd Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society (EMBC), 1936-1939, 2020

[9] V. Estrade, B. D. de Senneville, P. Meria, C. Almeras, F. Bladou, J.C. Bernhard, R. Gregoire, O. Traxer, M. Daudon, Toward improved endoscopic examination of urinary stones: a concordance study between endoscopic digital pictures vs. Microscopy, BJU International, 2020.

[10] L. McInnes, J. Healy, J. Melville, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, Online Pre-print: arXiv:1802.03426, 2018.

[11] Jahrreiss, J. Veser, C. Seitz, M. Ozsoy, Artificial intelligence: thefuture of urinary stone management?, Current Opinion in Urology,vol.30, number 2, 2020