Interpretable Deep Learning Classifier by Detection of Prototypical Parts on Kidney Stones Images

Daniel Flores-Araiza[1], Francisco Lopez-Tiro[1], Elias Villalvazo-Avila[1], Jonathan El-Beze[2], Jacques Hubert[2], Gilberto Ochoa-Ruiz[1], Cristian Daul[3]

[1]Tecnologico de Monterrey, School of Engineering and Sciences, Mexico
[2]CHU Nancy, Service d'urologie de Brabois, Nancy, France
[3]Centre de Recherche en Automatique de Nancy, Université de Lorraine, France

Abstract

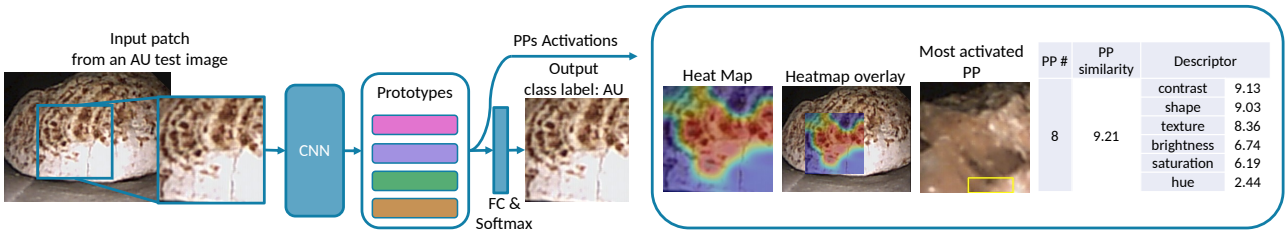
Identifying the type of kidney stones can allow urologists to determine their formation cause, improving the early prescription of appropriate treatments to diminish future relapses. However, currently, the associated ex-vivo diagnosis (known as morpho-constitutional analysis, MCA) is time-consuming, expensive and requires a great deal of experience, as it requires a visual analysis component that is highly operator dependant. Recently, machine learning methods have been developed for in-vivo endoscopic stone recognition. Shallow methods have demonstrated to be reliable and interpretable but exhibit low accuracy, while deep learning-based methods yield high accuracy but are not explainable. However, high stake decisions require understandable computer-aided diagnosis (CAD) to suggest a course of action based on reasonable evidence, rather than to merely prescribe one. Herein, we investigate means for learning part-prototypes (PPs) that enable interpretable models. Our proposal suggests a classification for a kidney stone patch image and provides explanations in a similar way as those used on the MCA method.

1. Introduction

Urolithiasis disease refers to the formation of kidney stones (KS). Several industrialized countries present a high incidence of kidney stone episodes (around 10% of the population is affected [21, 14]). The stone formation is a multifactorial process [4, 6], where the diet is one of the most important factors [10, 19] but several complementary factors can produce it (e.g., hereditary-family history, chronic diseases, and sedentary lifestyle). Early identification of the type of kidney stone aids the urologist to have an accurate diagnosis, enabling them to prescribe the appropriate treatment (e.g., diet adaptation or surgery), but most importantly, to reduce eventual relapses [10].

The kidney stone type of an ex-vivo sample (extracted during endoscopic surgery) can be identified using a two-step procedure, a method known as Morpho-Constitutional Analysis (MCA) [5, 6]. First, a microscopic morphological examination of the visual characteristics of the stone (e.g., size, form, color, texture, and appearance of the surface and section view) is strongly correlated with the molecular study [4] and it enables to preserve important diagnostic information. On the other hand, an Infrared-spectrophotometry analysis leads to a more precise identification of the crystalline composition of the stone [8]. Although the MCA efficiently establishes the type of ex-vivo kidney stones, it is very difficult to provide a reliable diagnosis during an endoscopic intervention (the results of the MCA may take days). Also, it is time-consuming and tedious (fragments extraction can take up to one hour of surgery) [4] and very difficult to train specialists on. Recently, it has been suggested that an endoscopic (intra-operative) stone recognition (ESR) CAD tool could help to obtain a faster diagnosis based solely on the video signal information provided by the endoscope, in tandem with the visual aid on the screen [9]. Also, the operations are quicker to perform and less traumatic, due to dusting can fragment and destroy the kidney stones inside the urinary tract.

Several methods have been proposed in recent years for performing automated ESR, based both on traditional and deep learning techniques with very encouraging results [16]. On the one hand, shallow Machine Learning (ML) models have shown that the efficient extraction of features (e.g., color and texture) in kidney stone views (surface and section) can have a significant impact on the classification (with accuracy $\geq 90\%$) on in-vivo endoscopic images (strongly correlated to the morphological analysis of kidney stones) [12, 11, 16]. However, in visualizations such as UMAP [13], the clusters are not tight enough, which could



(a) Model architecture

(b) Explanations

Figure 1: (a) Overall view of the proposal workflow, using ProtoPNet to obtain particular explanations for an input image. By use of PPs we provide explanations of the output classification, in tree different ways on (b), with a heatmap of the relevant parts on the input image, training images detected similar to the input, and measures of visual characteristics (descriptors) important for the activated PPs.

mean that they are not the best features that could be used in the classifier. On the other hand, Deep Learning (DL) based models [11, 9, 2, 16] have shown excellent results (high accuracy $\geq 95\%$) for extracting features relevant to the classifier (and tight clusters in UMAP). However, DL models lack interpretation of the features they extract, making these models not very useful in clinical settings.

As a matter of fact, current ML and DL models are unable to describe the inner workings that led to a given prediction beyond the class label. Therefore, these types of models cannot provide useful information to the specialist to understand how the input image was used to perform a diagnosis (i.e., classify the kidney stone type).

In order to pave the way for AI-based ESR using deep learning techniques, in this work, we leverage recent strides in explainability that seek to base image classifiers decisions on case-based reasoning to make them more interpretable [3, 15, 17, 1]. Additionally, we provide both visual explanations and quantitative information about visual characteristics deemed important by the network. It must be emphasized that our approach follows the reasoning processes of urologists in detecting morphological relevant features of each image (i.e., MCA). Overall, this work is aimed at facilitating human-machine collaboration in the context of CAD tools for urolithiasis prevention.

2. Materials and Methods

2.1. Kidney stone dataset

The ex-vivo dataset includes 305 kidney stone images acquired (two reusable digital flexible ureteroscopes from Karl Storz using video columns: Storz Image 1 Hub and Storz image1 S) and labeled manually by the urologist Jonathan El Beze[2] (for more details, see [7]). For this study, we make use of an ex-vivo image dataset divided in three subsets: 177 surface images, 128 section images and the third subset of 305 images (177 surface and 128 section images) of the six kidney stone types with the highest incidence: Acide Urique (AU), Brushite (BRU), and Cystine (CYS), Struvite (STR), Weddellite (WD), Whewellite (WW). Patches of this dataset are shown in Fig. 2.

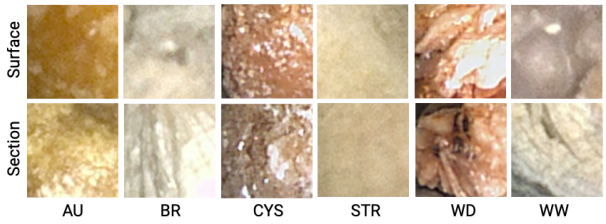


Figure 2: Examples of ex-vivo kidney stones generated patches.

However, the identification of kidney stones is not usually performed on whole images [18, 20, 2, 12]. Thus, patches of 200×200 pixels (minimal size enable to capture enough texture and color information) were cropped from the original images to increase the size of the training dataset (for more details, see [11]). A total of 2000 patches are available per class (AU, BRU, CYS, STR, WD, and WW) and view (surface, section, and mixed). The train and test relationship were 80% (38400 images) and 20% (9600 images), respectively. In order to limit the over-fitting produced by the small size of the available training dataset, data augmentation was performed. Additional patches were obtained by applying geometrical transformations (patch flipping, and perspective distortions), increasing the number of training patches from 38400 to 1152000 using data augmentation. The patches were also "whitened" using the mean m_i and standard deviation σ_i of the color values I_i in each channel ($I_i^w = (I_i - m_i\sigma_i)$, with $i = R, G, B$).

2.2. ProtoPNet plus descriptors

By use of a Prototypical Part Network (ProtoPNet), able to identify several parts of an image, where it thinks a part of the image looks like a learned prototypical part of some class (as it can be seen in Fig. 1a and 1b). This type of model makes its prediction on a weighted combination of the similarity scores between parts of the image and the learned part-prototypes (PPs). This capacity yields

Table 1: Weighted average metrics comparison for section, surface, and mixed patches. ProtoPNet with VGG19 with batch normalization as the backbone (PPN-VGG19bn). VGG 19-layer model, configuration 'E', with batch normalization (VGG19bn).

Model	Accuracy			Precision			Recall			F1 score		
	Surface	Section	Mixed	Surface	Section	Mixed	Surface	Section	Mixed	Surface	Section	Mixed
PPN-VGG19bn	0.98	0.99	0.97	0.98	0.99	0.97	0.98	0.99	0.97	0.98	0.99	0.97
VGG19bn	0.99	1.00	1.00	0.99	1.00	1.00	0.99	1.00	1.00	0.99	1.00	1.00
AlexNet	0.96	0.97	0.97	0.96	0.97	0.97	0.96	0.97	0.97	0.96	0.97	0.97

predictions that are relatively easy to understand by users, rendering it interpretable. We apply the methodology presented in previous works, that proposed and used ProtoPnet models [3, 17]. However, PPs may still depend on non-apparent characteristics from the input image, reason for us to quantify the sensitivity of PPs to a set of perturbations [15], which we call "descriptors". These descriptors indicate why the classification model deemed an image patch and PP (part-prototype) similar. It's worth noticing that PPs are vectors in latent space that should learn discriminative, prototypical parts of a class. Thus, high dimensional reduction projections, UMAP visualizations as an example, contain global information of the main characteristics of the whole ProtoPNet model.

2.3. Implementation details

The ProtoPNet architecture consists of a standard Convolutional Neural Network (CNN, e.g. ResNet), followed by a prototype layer and a fully-connected layer. The prototype layer consists of a pre-determined number of class-specific prototypes. Herein, we use 10 part-prototypes per class, the initialization and training procedure for our training also follows [3, 15], using a pre-trained VGG19 (with batch normalization) on ImageNet as CNN backbone.

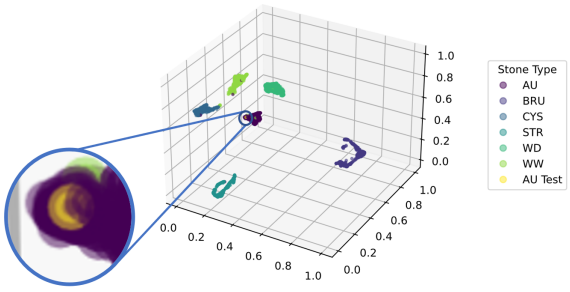


Figure 3: UMAP of the Part-Prototypes activations on section images. Our approach allows obtaining separate clusters of the output classes. On the zoomed example (blue circle) can be appreciated the projection of a new classification (yellow point) is surrounded by samples of the same class (purple points), indicating a high certainty on the correct classification of the test sample.

3. Results and Discussion

Evaluation metrics of our model, its backbone model, and AlexNet (as reference), are reported in Table 1. The performance of ProtoPNet is comparable with its corresponding uninterpretable backbone model ($\leq$ 3% difference). In contrast to black-box classifiers, our proposal provides explanations for input images, by showing the activation area of PPs, the corresponding representative image to each activated PPs, and their descriptors (Fig. 1b).

We plot PPs and their descriptors activations for input images on a UMAP visualization of the three most discriminant dimensions (umap1 to umap3). This UMAP allows seeing class separability for each output class of the ProtoPNet, as shown in Fig. 3. In this way, additional global insight is gained in the case of a new classification observed surrounded by samples of the same class, which provides confidence of a correct classification (also described in Fig. 3). However, it was observed that multiple PPs end up indicating the same training patch as their explanation, a behavior similar to mode collapse on Generative Adversarial Networks (GANs), which limits the variety of possible explanations provided for an output. The use of descriptors mitigates the cases for visually similar PPs by providing details on the characteristics most relevant for each PP [15].

4. Conclusion and Future work

We showed that by training of PPs and extracting their descriptors we convert an uninterpretable VGG19 into an interpretable model. This can facilitate the use of these models for ESR by a urologist. However, mode collapse of the learned PPs is a limitation on the current implementations of ProtoPNets. To prevent this, better initialization procedures and loss function adjustments will be explored. Finally, we found indications of better class separability by use of PPs and their descriptors, to the point UMAP visualizations could be used to provide global context of the certainty of the output classification of a particular image.

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.

References

[1] Alina Jade Barnett, Fides Regina Schwartz, Chaofan Tao, Chaofan Chen, Yinhao Ren, Joseph Y. Lo, and Cynthia Rudin. IAIA-BL: A case-based interpretable deep learning model for classification of mass lesions in digital mammography. *CoRR*, abs/2103.12308, 2021. 2

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

[3] Chaofan Chen, Oscar Li, Chaofan Tao, Alina Jade Barnett, Jonathan Su, and Cynthia Rudin. This Looks Like That: Deep Learning for Interpretable Image Recognition. *CoRR*, jun 2018. 2, 3

[4] Mariela Corrales, Steeve Doizi, Yazeed Barghouthy, Olivier Traxer, and Michel Daudon. Classification of stones according to michel daudon: a narrative review. *European Urology Focus*, 7(1):13–21, 2021. 1

[5] 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

[6] Michel Daudon, Paul Jungers, Dominique Bazin, and James C Williams. Recurrence rates of urinary calculi according to stone composition and morphology. *Urolithiasis*, 46(5):459–470, 2018. 1

[7] Jonathan El Beze, Charles Mazeaud, Christian Daul, Gilberto Ochoa-Ruiz, Michel Daudon, Pascal Eschwège, and Jacques Hubert. Evaluation and understanding of automated urolithiasis recognition methods. *BJU International*, 2022. 2

[8] Laurence Estepa and Michel Daudon. Contribution of fourier transform infrared spectroscopy to the identification of urinary stones and kidney crystal deposits. *Biospectroscopy*, 3(5):347–369, 1997. 1

[9] Vincent Estrade, Michel Daudon, Emmanuel Richard, Jean-Christophe Bernhard, Franck Bladou, Gregoire Robert, and Baudouin Denis de Senneville. Towards automatic recognition of pure & mixed stones using intraoperative endoscopic digital images. *BJU International*, abs/2105.10686, 2021. 1, 2

[10] 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

[11] Francisco Lopez, Andres Varelo, Oscar Hinojosa, Mauricio Mendez, Dinh-Hoan Trinh, Yonathan ElBeze, Jacques Hubert, Vincent Estrade, Miguel Gonzalez, Gilberto Ochoa, et al. Assessing deep learning methods for the identification of kidney stones in endoscopic images. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 2778–2781. IEEE, 2021. 1, 2

[12] 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. 1, 2

[13] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. 1

[14] Anmar Nassir, Hesham Saada, Taghreed Alnajjar, Jomanah Nasser, Waed Jameel, Soha Elmorsy, and Hattan Badr. The impact of stone composition on renal function. *Urology Annals*, 10(2):215, 2018. 1

[15] Meike Nauta, Annemarie Jutte, Jesper C. Provoost, and Christin Seifert. This looks like that, because ... explaining prototypes for interpretable image recognition. *CoRR*, abs/2011.02863, 2020. 2, 3

[16] Gilberto Ochoa-Ruiz, Vincent Estrade, Francisco Lopez, Daniel Flores-Araiza, Jonathan El Beze, Dinh-Hoan Trinh, Miguel Gonzalez-Mendoza, Pascal Eschwège, Jacques Hubert, and Christian Daul. On the in vivo recognition of kidney stones using machine learning. *arXiv preprint arXiv:2201.08865*, 2022. 1, 2

[17] Dawid Rymarczyk, Lukasz Struski, Jacek Tabor, and Bartosz Zielinski. Protopshare: Prototype sharing for interpretable image classification and similarity discovery. *CoRR*, abs/2011.14340, 2020. 2, 3

[18] 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. 2

[19] R Siener and A Hesse. Fluid intake and epidemiology of urolithiasis. *European journal of clinical nutrition*, 57(2):S47–S51, 2003. 1

[20] Alejandro Torrell Amado. Metric learning for kidney stone classification. 2018. 2

[21] Adie Viljoen, Rabia Chaudhry, and John Bycroft. Renal stones. *Annals of clinical biochemistry*, 56(1):15–27, 2019. 1