Comparing feature fusion strategies for
Deep Learning-based kidney stone identification

Elias VILLALVAZO-AVILA[1], Francisco LOPEZ-TIRO[1], Daniel FLORES-ARAIZA[1], Gilberto OCHOA-RUIZ[1],
Jonathan EL-BEZE[2], Jacques HUBERT[2], Christian DAUL[3]

[1]Escuela de Ingenieria y Ciencias, Tecnologico de Monterrey
Av. Eugenio Garza Sada 2501 Sur, Tecnológico, 64849 Monterrey, N.L., Mexico

[2]CHU Nancy, Service d'urologie de Brabois, F-54511 Nancy, France

[3]CRAN (UMR 7039), Université de Lorraine and CNRS,
2 avenue de la Forêt de Haye, 54518 Vandœuvre-lès-Nancy cedex, France

A01798097, A01799045, A01098051, gilberto.ochoa@tec.mx
j.elbeze@chru-nancy.fr, j.hubert@chru-nancy.fr, christian.daul@univ-lorraine.fr

Résumé – Cette contribution présente une méthode d'apprentissage profond pour l'extraction et la fusion d'informations d'images acquises sous différents points de vue dans le but de produire des caractéristiques plus discriminantes entre objets. Notre approche a été conçue pour mimer l'analyse morpho-constitutionnelle utilisée par les urologues pour classer visuellement des fragments de calculs rénaux à partir de leur surface et section. Des stratégies de fusion de caractéristiques profondes ont permis d'améliorer les performances des extracteurs (structures principales des réseaux) à vue unique de plus de 10 % en termes de précision de la classification des calculs rénaux.

Abstract – This contribution presents a deep-learning method for extracting and fusing image information acquired from different viewpoints with the aim to produce more discriminant object features. Our approach was specifically designed to mimic the morpho-constitutional analysis used by urologists to visually classify kidney stones by inspecting the sections and surfaces of their fragments. Deep feature fusion strategies improved the results of single view extraction backbone models by more than 10% in terms of precision of the kidney stones classification.

1 Introduction

Urolithiasis refers to the formation of kidney stones that cannot be expelled from the urinary tract. This is a medical condition that has been increasing over the last few years [1]. Urolithiasis is caused by multiple factors, where diet is the most important, but also genetic inheritance, water intake, and a sedentary lifestyle could promote the formation of kidney stones [2]. The Morpho-constitutional analysis (MCA) is the most important method for kidney stone characterisation. MCA is a combination of a visual examination under the microscope of the stone's texture, appearance, and color (surface and section views), and a biochemical analysis by Fourier Transform Infrared Spectroscopy (FTIR) [3]. If carried out properly, a timely treatment (diet adaptation, surgery) can be prescribed for each patient, reducing the risk of stone recurrence [3].

However, MCA has a major drawback : the results of this analysis are often available only after several weeks. Thus, urologists increasingly aim at visually identifying the morphology of kidney stones only with the help of the image displayed on the screen [4] during the removal process (Endoscopic Stone recognition (ESR)). However, this visual analysis requires a great deal of experience due to the high similarities between classes that only a limited number of specialists have.

Therefore, different Machine Learning (ML) approaches have been proposed [4, 5, 6, 9] for the classification of kidney stones, demonstrating that it is a problem that can be solved with traditional and deep learning techniques with very encouraging results. However, most of these models were trained on ex-vivo stones placed in controlled environments, whereas in reality, images may suffer from motion blur, reflections, illumination variations, as occurs in common practice during an endoscopic imaging session. Moreover, there is no ordered manner of mixing surface and section information for exploiting the visual information in a way that a specialist would do it. Besides, in most cases the amount of training data available is limited, thus these contributions use data augmentation techniques to increase the amount of input data, but some limitations have not been addressed. Nonetheless, these works [4, 5, 6] have demonstrated the potential of automatic ESR in an in-vivo dataset. Table 1 gathers the previous works which have addressed the problem of recognizing kidney stone types using only images. This table provides also an overview on the used data (urinary calculus classes and acquisition conditions).

Multi-View (MV) classification is an area of ML that combines features from different sources or feature subsets, known as views, to identify objects with higher accuracy, since diverse

TABLE 1 – 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), WW : Whewellite (Calcium Oxalate Monohydrate), WD : Weddellite (Calcium Oxalate Dihydrate), STR : Struvite, CYS : Cystine, and BRU : Brushite.

Reference/Feature	Kidney Stone Composition						Image Type		Acquisition
	AU	WW	WD	STR	CYS	BRU	Surface	Section	
Serrat et al. (2017) [9]	✓	✓	✓	✓			✓	✓	Ex-vivo
Torrel et al. (2018) [8]	✓	✓	✓	✓		✓	✓		Ex-vivo
Black et al. (2020) [7]	✓	✓		✓	✓	✓	✓	✓	Ex-vivo
Martinez et al. (2020) [6]	✓	✓	✓				✓	✓	In-vivo
Lopez et al. (2021) [10]	✓	✓	✓			✓	✓	✓	In-vivo
Estrade et al. (2021) [5]	✓	✓	✓				✓	✓	In-vivo
This contribution	✓	✓	✓	✓	✓	✓	✓	✓	Ex-vivo

TABLE 2 – Comparison of the precision obtained by the three main contributions for the most common kidney stone classes. The precision is given for each individual class and classifier. The taxonomy per stone class same as in Table 1. The average precision (weighted by the image number of each class) are also given for each kidney stone type. In the last table line, the precision values are given for surface images. For section images, the authors in [5], reported a precision of 0.95, 0.94 and 0.94 for AU, WW and WD, respectively. The corresponding weighted precision equals 0.94.

Reference	Precision Per Class						Weighted Precision	ML Method
	AU	WW	WD	STR	CYS	BRU		
Serrat et al. (2017) [9]	0.65	0.55	0.69	0.50	N/A	N/A	0.63	Random Forest
Torrel et al. (2018) [8]	0.76	0.67	0.80	071	N/A	0.72	0.74	Siamese CNN
Black et al. (2020) [7]	0.94	0.95	N/A	0.71	0.75	0.75	0.85	CNN – ResNet101
Martinez et al. (2020) [6]	0.91	0.94	0.92	N/A	N/A	N/A	0.92	Random Forest
Lopez et al. (2021) [10]	0.98	0.93	0.95	N/A	N/A	0.96	0.95	Inception
Estrade et al. (2021) [5]	0.99	0.90	0.93	N/A	N/A	N/A	0.94	ResNet152v2

characteristics are extracted, synthesized and combined [11].

This variant of learning can improve the performance by optimizing multiple functions, one per view, and in that way, information can be obtained from different perspectives of the same data inputs. Moreover, MV can also be applied to Convolutional Neural Networks (CNNs) to boost the performance in situations where a single image does not yield sufficiently discriminative information for accurate classification by combining useful information from different views, so more comprehensive representations may be learned yielding a more effective classifier. This MV approach stands in contrast to previous works for ML-based ESR, which made use of single models which mixed image patches to train either a shallow or a DL models, as summarized in Table 2. These previous works showed the potential gains that could be obtained through an automated mechanisms for ESR, but those methods either used ex-vivo images acquired using high-quality microscopes or digital cameras or did not take into account the fusion of section and surface information to perform a prediction.

In this work, we leverage recent strides in DL research that have sought to combine information from multiple views and we demonstrate that MV learning can be applied to ESR, with promising results. We show that such an approach can be beneficial for the identification of kidney stones by maximizing the amount of information that the model can use for classification. Through several experiments, we demonstrate that by combining the information of surface and section views in the same model, we can obtain a method that is more explainable and similar to what specialists do in clinical practice (i.e., MCA).

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]. To reproduce in-vivo conditions, the experimental setup used in this work consists of a small diameter tube where the inner walls were covered with a yellowish film to display the appearance of the urinary tract (for more details, see [12]). The ex-vivo dataset consists of three subsets : the first subset consist of 177 surface images, 128 section images for the second subset, and the third subset of 305 images (177 section + 128 surface) of the six kidney stone types with the highest incidence : Type Ia (Whewellite, WW), Type IIb (Weddellite, WD), Type IIIb (Acide Urique, AU), Type IVc (Struvite, STR), Type IVd (Brushite, BRU), and Type Va (Cystine, CYS). Images of this dataset are shown in Fig. 1.

Classification of kidney stones (i.e., MCA) is usually not performed on whole images [6, 7, 8, 9]. Therefore, in this work as in previous works, patches of 256×256 pixels were cropped from the original images to increase the size of the training dataset (for more details, see [10]). However, the number of resultant patches for each class is imbalanced (due to the changing fragment sizes, image resolution, and the number of images in the original dataset). In order to balance the number of patches per class, a random sampling approach was used.
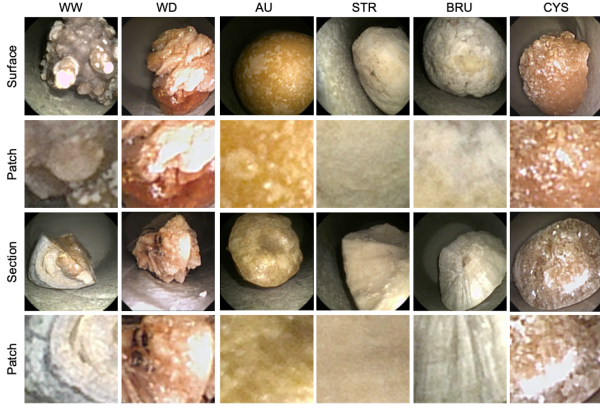
FIGURE 1 – Examples of ex-vivo kidney stones images. From left to right : WW, WD, AU, STR, BRU, and CYS. The surface view is in the first row, and their respective generated patches are in the second row. The section view is in the third row, and their respective generated patches are in the fourth row.

This step yielded a total of 1000 patches per class (WW, WD, AU, STR, BRU, and CYS) and view (surface, section, and mixed). This new dataset was then split into 19200 images (80%) for training and validation, and 4800 images (20%) for the test. In order to limit the over-fitting produced by the small size of the available training dataset, data augmentation was heavily performed. Additional patches were obtained by applying geometrical transformations (patch flipping, affine transformations, and perspective distortions). The number of patches increased from 19200 to 153600 using 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 color values I_i in each channel [10].

2.2 Methods

2.2.1 Pre-training Stage

Previous approaches have used Deep Learning architectures such as AlexNet, or VGG16 for assessing the kidney stone classification task [4, 10]. For this contribution, the previously-mentioned architectures are used for the creation of the MV models. This network was trained on the entire training data, mixing both surface and section patches, and served as a baseline or comparison for the multi-view implementations introduced in this paper. Once trained, the feature extraction layers of this single-view network are frozen to ensure that each branch from the multi-view model extracts the same features and that any variation in the performance will rely on the unfrozen layers (fusion and fully-connected layers).

2.2.2 Multi-View Model

The frozen layers are duplicated. In this way, one copy will process only images of the surface of the stone, while the other copy will process images of the section view. These frozen layers are connected to a fusion layer, which will be respon-
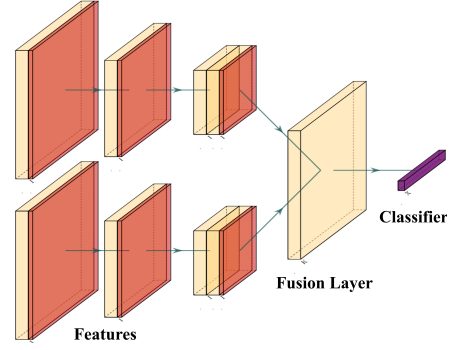

FIGURE 2 – Proposed Multi-View model. First part of the model corresponds to the duplicated feature extraction layers. One copy will process only surface images, the other will process section images. A fusion layer is added to combine information from the different views. The fused feature map is then fed to the classification layers

sible for mixing the information of the two views. In this work, the two late-fusion methods proposed in [11] are explored. The first method concatenates the feature vectors obtained from each view, and connects the resulting representation to a fully connected layer. As per the second method, feature vectors are stacked and max-pooling is applied across them. Lastly, the output of the fusion layer is connected to the remaining part of the MV model, which merely consists of the classifier. The proposed model is shown in Fig. 2. To make a direct comparison of the feasibility and performance of this architecture against previous works, we used the same hyper-parameters as [4]. Cross-entropy loss is used to compute the classification loss, and optimization is performed using Adam optimizer with a learning rate of $2e^{-4}$. Batch-size selected was 64 for both multi-view and single-view networks.

3 Results and Discussion

Several experiments were performed to assess the ability of MV models to predict the kidney stone class, combining information from surface and section views, as done during an ESR procedure [5]. Precision (P) and Recall (R) metrics are determined for each class individually. The results reported for both fusion techniques in the multi-view models show that combining information from different classifiers yields significant results compared to the single-view classifiers. For instance, for a single-view AlexNet network, the results obtained were 0.84 and 0.83 for precision and recall, respectively. In contrast, MV networks, independent of the fusion strategy, performed better compared to the other experiments. Table 3 shows the scores for all the models used for this work, and the UMAP visualizations on Fig. 3a, and Fig. 3b show how stone type clusters are distributed for both SV and MV networks. One disadvantage of using concatenation as fusion strategy is that the number of features of the first layers of the classifier increases considerably, limiting its implementation on systems with reduced memory.
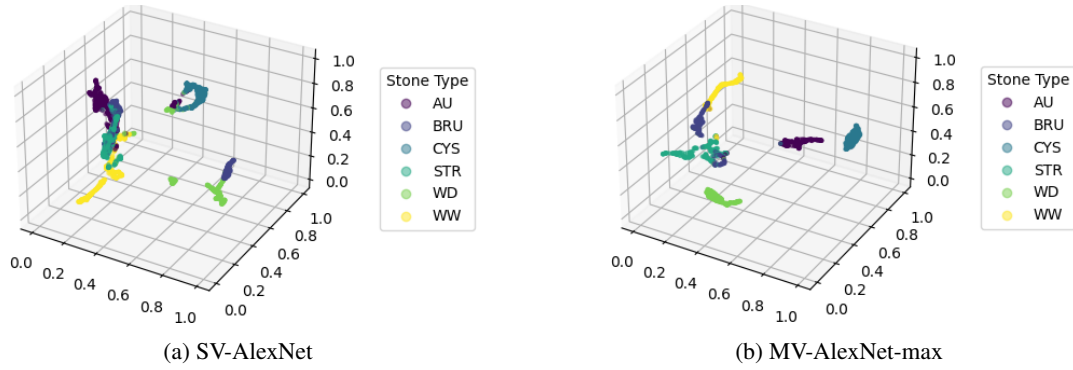
(a) SV-AlexNet   (b) MV-AlexNet-max

FIGURE 3 – UMAP visualizations showing the learned representations (i.e., feature maps) for (a) the Single-View (SV) AlexNet architecture with no fusion, and (b) in Multi-View AlexNet architecture with a max-pool fusion layer. As it can be observed in the figure, more separate clusters are obtained through the use of feature fusion layers

TABLE 3 – Weighted average metrics comparison for section, surface, and mixed patches. MV-AlexNet-max : Multi-View network with max-pool as fusion strategy. MV-VGG16-max : Multi-View network with max-pool as fusion strategy. MV-AlexNet-conc : Multi-View network with concatenation as fusion strategy. SV-AlexNet : Single-View AlexNet network. SV-VGG16 : Single-View VGG16 network.

Classifier	Surface		Section		Mixed	
	P	**R**	**P**	**R**	**P**	**R**
MV-AlexNet-max	–	–	–	–	**0.95**	**0.94**
MV-VGG16-max	–	–	–	–	**0.94**	**0.94**
MV-AlexNet-conc	–	–	–	–	**0.94**	**0.93**
SV-AlexNet	0.77	0.71	0.88	0.87	0.84	0.83
SV-VGG16	0.79	0.70	0.89	0.89	0.83	0.81

4 Conclusion and future work

We showed that by mixing information from different views, it is possible to train more accurate models for predicting kidney stone composition from images obtained from ureteroscopy. Thus, AI technology can be included in the current stone removal workflow, speeding up preventive diagnosis measures. However, we make use of a very-limited ex-vivo dataset in a simulated environment. We aim to solve this problem by applying metric learning in future work to tackle the amount of data that we require for training, as well as to increase inter-class separability.

Références

[1] A. Viljoen, R. Rabia et J. Bycroft. *Renal stones*. Annals of clinical biochemistry, 2019.

[2] J.I. Friedlander, J. Antonelli et M.S. Pearle. *Diet : from food to stone*. World journal of urology, 2015.

[3] M. Daudon, P. Jungers, B. Dominique et J.C. Williams. *Recurrence rates of urinary calculi according to stone composition and morphology*. Urolithiasis, 2018.

[4] G. Ochoa-Ruiz, V. Estrade, F. Lopez, D. Flores-Araiza, J. El-Beze, D-H. Trinh, G. Gonzalez-Mendoza, P. Eschwège, J. Hubert et C. Daul. *On the in vivo recognition of kidney stones using machine learning*. arXiv preprint arXiv:2201.08865, 2022.

[5] V. Estrade, M. Daudon, E. Richard, J.C. Bernhard, F. Bladou et G. Robert. *Towards automatic recognition of pure & mixed stones using intraoperative endoscopic digital images*. BJU International, 2021.

[6] A. Martínez, T. Dinh-Hoan, J. El Beze, J. Hubert, P. Eschege, V. Estrade, L. Aguilar, C. Daul et G. Ochoa. *Towards an automated classification method for ureteroscopic kidney stone images using ensemble learning*. 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), 2020.

[7] K.M. Black, H. Law, A. Aldoukhi, J. Deng et K.R Ghani. *Deep learning computer vision algorithm for detecting kidney stone composition*. Wiley Periodicals Inc., 2020.

[8] A. Torrell Amado. *Metric learning for kidney stone classification*. 2018.

[9] J. Serrat, F. Lumbreras, F. Blanco, M. Valiente et M. López-Mesas. *myStone : A system for automatic kidney stone classification*. Expert Systems with Applications. Elsevier, 2017.

[10] F. Lopez, A. Varelo, O. Hinojosa, M. Mendez, D.-H. Trinh, Y. ElBeze, J. Hubert, V. Estrade, M. Gonzalez et G. Ochoa. *Assessing deep learning methods for the identification of kidney stones in endoscopic images*. 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), 2021.

[11] D. Li et Y. Tian. *Multi-view metric learning for multi-instance image classification*. arXiv preprint arXiv:1610.06671, 2016.

[12] J. El-Beze, C. Mazeaud, C. Daul, G. Ochoa-Ruiz, M. Daudon, P. Eschwège et J. Hubert. *Evaluation and understanding of automated urolithiasis recognition methods*. BJU International, 2022.