# myStone: a system for automatic kidney stone classification

Joan Serrat*[1], Felipe Lumbreras[1], Francisco Blanco[2], Manuel Valiente[2], and Montserrat López-Mesas[2]

[1]Computer Vision Center and Dept. of Computer Science, Universitat Autònoma de Barcelona, 08193 Bellaterra, Spain.

[2]Centre Grup de Tècniques de Separació en Química, Unitat de Química Analítica, Departament de Química, Universitat Autònoma de Barcelona, 08193 Bellaterra, Spain.

July 20, 2017

*Corresponding author. Address: Computer Vision Center, Universitat Autònoma de Barcelona, Edifici O, 08193 Bellaterra, Spain. Tel.: +34 935811828; Fax: +34 935811670.
Email addresses: joans@cvc.uab.es (J. Serrat), felipe@cvc.uab.es (Felipe Lumbreras), frnblanco@gmail.com (Francisco Blanco), Manuel.Valiente@uab.cat (Manuel Valiente), Montserrat.Lopez.Mesas@uab.cat (Montserrat López-Mesas)

**Abstract**

Kidney stone formation is a common disease and the incidence rate is constantly increasing worldwide. It has been shown that the classification of kidney stones can lead to an important reduction of the recurrence rate. The classification of kidney stones by human experts on the basis of certain visual color and texture features is one of the most employed techniques. However, the knowledge of how to analyze kidney stones is not widespread, and the experts learn only after being trained on a large number of samples of the different classes. In this paper we describe a new device specifically designed for capturing images of expelled kidney stones, and a method to learn and apply the experts knowledge with regard to their classification. We show that with off the shelf components, a carefully selected set of features and an state of the art classifier it is possible to automate this difficult task to a good degree. We report results on a collection of 454 kidney stones, achieving an overall accuracy of 63% for a set of eight classes covering almost all of the kidney stones taxonomy. Moreover, for more than 80% of samples the real class is the first or the second most probable class according to the system, being then the patient recommendations for the two top classes similar. This is the first attempt towards the automatic visual classification of kidney stones, and based on the current results we foresee better accuracies with the increase of the dataset size.

Keywords: kidney stone; optical device; computer vision; image classification.

# 1 Introduction

Medicine and healthcare are among the most important fields where expert systems have found application (Wagner, 2017). For instance, computer aided detection and diagnosis systems can enhance the diagnostic capabilities of physicians or reduce the required time. Many of them are based on diverse imaging modalities like brain CT and MRIs, mammographies, chest X-ray and a long etc. Another type of expert systems are decision support systems, whose purpose is not as much to produce a diagnostic but to analyze data (eg. images) and present some kind of result so that decisions can be made more easily. At the core of such systems one can often find classifiers, which are typically trained on data samples to generate a discrete prediction (a class label) plus a confidence score or a probability for each class, when presented a new sample. This is the case of the work described in this paper, which deals with the problem of kidney stone classification.

Urinary lithiasis —the formation of kidney stones— shows a steady incidence increase in developed countries. Around 10% of population in developed countries suffer a stone episode at least once in his/her life (Romero et al., 2010; Scales et al., 2012). Emphasis should be made on the high prevalence affecting this disease. Some European follow-up studies have quantified the stone recurrence rate (repeated stone episodes for the same patient) at 40% in 5 years (Hesse et al., 2003; Andreassen et al., 2007). These dramatic numbers reflect not only a disturbing and painful disease but also a considerable burden for the national healthcare systems (Strohmaier, 2012).

Once the stone episode has passed, it is widely agreed that an adequate study of the causes of stone formation is required in order to decrease the high recurrence of this disease (Kok, 2012; Grases et al., 2002; Siener and Hesse, 2012). In fact, it has been pointed out that the correct treatment of stone patients can drop further stone formation as much as 46% (Nolde et al., 1993; Strohmaier, 2011). The urinary stone represents a solid description of the metabolic disturbances suffered by the patient, so it should be regarded as the starting point of an individualized treatment.

Aware of the importance of stone characterization, clinical scientists have used a variety of approaches for the description of urinary calculi. Already in the second half of the 19th century, the first classification of stones was devised, based on features such as color, hardness and shape. A number of techniques have been used from that starting point. IR spectroscopy, X-Ray diffraction and stereoscopic microscopy stand as the most extended analysis methodologies (Schubert, 2006). The rest of techniques (electron microscopy, Raman spectroscopy, hyperspectral imaging among them) have been mainly used for research purposes.

IR spectroscopy is easy to use and yields quantitative results, but the sample needs to be grinded and the distribution of components in the stone, related to its etiology, is inevitably lost. X-Ray diffraction presents the same drawback, as well as less available instrumentation. Hyperspectral imaging has proven to be a high-performance alternative. Like in IR spectroscopy, hyperspectral images are suitable to detect the presence of chemical components based on the spectral signature of the sample but, at the same time, to know their spatial distribution on the imaged surface of the stone. A few works have attempted to perform a pixel-wise classification according to some kidney stone classes with infrared and near infrared spectral imaging, like (Piqueras et al., 2011), (Blanco et al., 2012) and notably Blanco et al. (2015). However, their do nor perform a large scale study with hundreds of stones, like us. Additionally, the problem of this approach is that the equipment needed for its implementation is costly and not always available even in clinical laboratories.

In stereoscopic microscopy (Daudon et al., 1993; Cloutier et al., 2015; Grases et al., 2002) the external surface and a section of fragments are observed at a magnification of 10X to 40X. The color, 3D shape (e.g. lobulations, surface roughness and smoothness), size, shape and orientation of crystals, deposition layers, the existence of a core etc. provide cues to the expert with regard the stone class. Thus, stereoscopic microscopy relies on the expertise of the technician who performs the analysis. While very precise, it is time consuming, cannot be offered at a competitive price and generally the analysis is carried out in laboratories external to the hospital, so the waiting time for the results may become an issue.

This paper presents a system composed of a device and automatic classification method, specifically conceived for the fast and on-site analysis of urinary stones. The system is based on the principles of stereoscopic microscopy, so the sample is classified attending to the amount and distribution of mineralogical components, as they appear on images captured by a standard camera. To the best of our knowledge, this is the first attempt to automate the visual classification of kidney stones. Hence, we consider as the main contributions of this work :

1. The construction of a fully functional device, including hardware, user interface and classification software, for the visual recognition of renal calculi.

2. The first extensive dataset for this kind of samples, available upon request at (Lumbreras et al., 2017). It consists of 14,500 images of 908 stone fragments from 454 producers, recording separately both the external and internal side of each fragment, under visible and near infrared light sources.

3. The identification of discriminant color and texture features to train a state-of-the art classifier that attains a baseline accuracy of 63% for the top class and 83% for the top-2 classes, in spite of the large intraclass variability combined in some cases with considerable inter-class similarity.

4. We show that a boost in performance is possible with the use of the urinary pH level, obtaining 70% and 89% top-1 and top-2 accuracy, respectively. Moreover, confusions align with classes judged more similar by the annotator expert.

The organization of this paper is as follows. Section 2 presents a widespread taxonomy of kidney stones. They are characterized by the presence of certain chemical components that show up as color and textural features. Section 3 describes the device we have built to acquire images of kidney stone fragments. We follow a certain procedure to record a sample, which is a collection of images of a pair of fragments from one same patient. Accordingly, we have built a large dataset with samples of all the classes but the one less frequently found (section 4). On the images of the dataset we have computed a set of visual features related to color and texture which then are fed into a random forest classifier (section 5). In section 6 we combine the class probabilities given by this classifier with those obtained from the urinary pH level, a non-visual feature that helps to distinguish some classes. Section 7 reports the results for four variants of the classifier, depending on the scheme of classes used (8 main classes or 12 fine-grained classes) and the use or not of the urinary pH level as an additional feature. Finally, section 8 draws the main conclusions and avenues of future work.

## 2 Kidney stone taxonomy

There is a well known taxonomy proposed by M. Daudon in 1993 (Daudon et al., 1993). It is a hierarchical classification whose first level considers the main or two main chemical species in the stone (calcium oxalate, uric acid etc.). A second level is provided to account for different etiologies or pathologies that such broad classes do not discern well. Thus, urologists consider also what the minor components are and their spatial distribution. The later means, for instance, whether they are on the surface or the inside of the stone, forming a core, layers, radial structures, lobules or uniformly spread. Table 1 relates the two taxonomies.

This second level has a total of 21 classes, making it difficult to adapt to the clinical practice. Moreover, since our goal is to train a classifier, we will need as much samples as possible per class. Unfortunately, many second level classes (and even some of the first level) have a low natural frequency of occurrence, so they may not be well represented in a dataset. F. Grases and colleagues (Grases et al., 2002) simplified this classification scheme and we draw from them our first scheme of classes which are:

- calcium oxalate monohydrate (COM)
- calcium oxalate dihydrate (COD)
- mixed calcium oxalate and hydroxiapatite (CO−HAP)
- hydroxiapatite (HAP)
- struvite (STR)
- brushite (BRU)
- uric acid anhydrous and dihydrate (UA)
- mixed uric acid and calcium oxalate (UA+CO)
- cystine (CYS)

When collecting samples to build the dataset, we were able to gather just a few samples of cystine, one of the less frequent ones (around 1%). They were insufficient to properly train the classifier and hence we discarded it. In contrast, we realized that a few subclasses of COM and COD were relatively well populated (see table 1), thus giving us the chance to provide more specific classification results and consequently better patient recommendations. Hence we decided to expand the first scheme of classes to a second scheme. Along the way, we are going to borrow the class notation of (Grases et al., 2002) for which the classes of first scheme are denoted by numbers (2, 3 . . . 9) whereas the labels for the second are the former plus a suffix if necessary. Accordingly, COM was divided into:

- pure COM (2)
- COM with traces of hydroxiapatite and/or organic matter in the core (2b)
- COM with little amounts of COD in the core (2codt)

and COD was also split into:

- COD with presence of hydroxiapatite (3b)
- COD with COM, resulting from the transformation of COD, which is unstable, to COM (3t)
- COD, transformed to COM plus traces of hydroxiapatite (3bt)

The exhaustive visual description of each class is out of the scope of this paper. We refer the reader to the seminal paper (Daudon et al., 1993) and the more recent work (Cloutier et al., 2015), where they are listed and illustrated with examples. Nonetheless, we can get a glimpse of them in figures 1 and 2. They show 8 images of stone fragments for each main class, as acquired by our device. In each case, half the images are from the outer part of a fragment, half from an inner section. We can appreciate that some classes are quite similar in aspect (e.g. STR, HAP and BRU) and also the large intraclass variability common to all of them.

The first problem is a consequence of the lack of a clear frontier between classes because they may share the same components in different proportions. For instance, classes 2b, 3bt, CO−HAP and HAP form a continuum of calcium oxalate with increasing proportion of hydroxiapatite. The same phenomena occurs in the case of UA and UA+CO, and 2, 2codt and 3t. This makes kidney stone classification a tough problem for the human expert and obviously for the classification software.

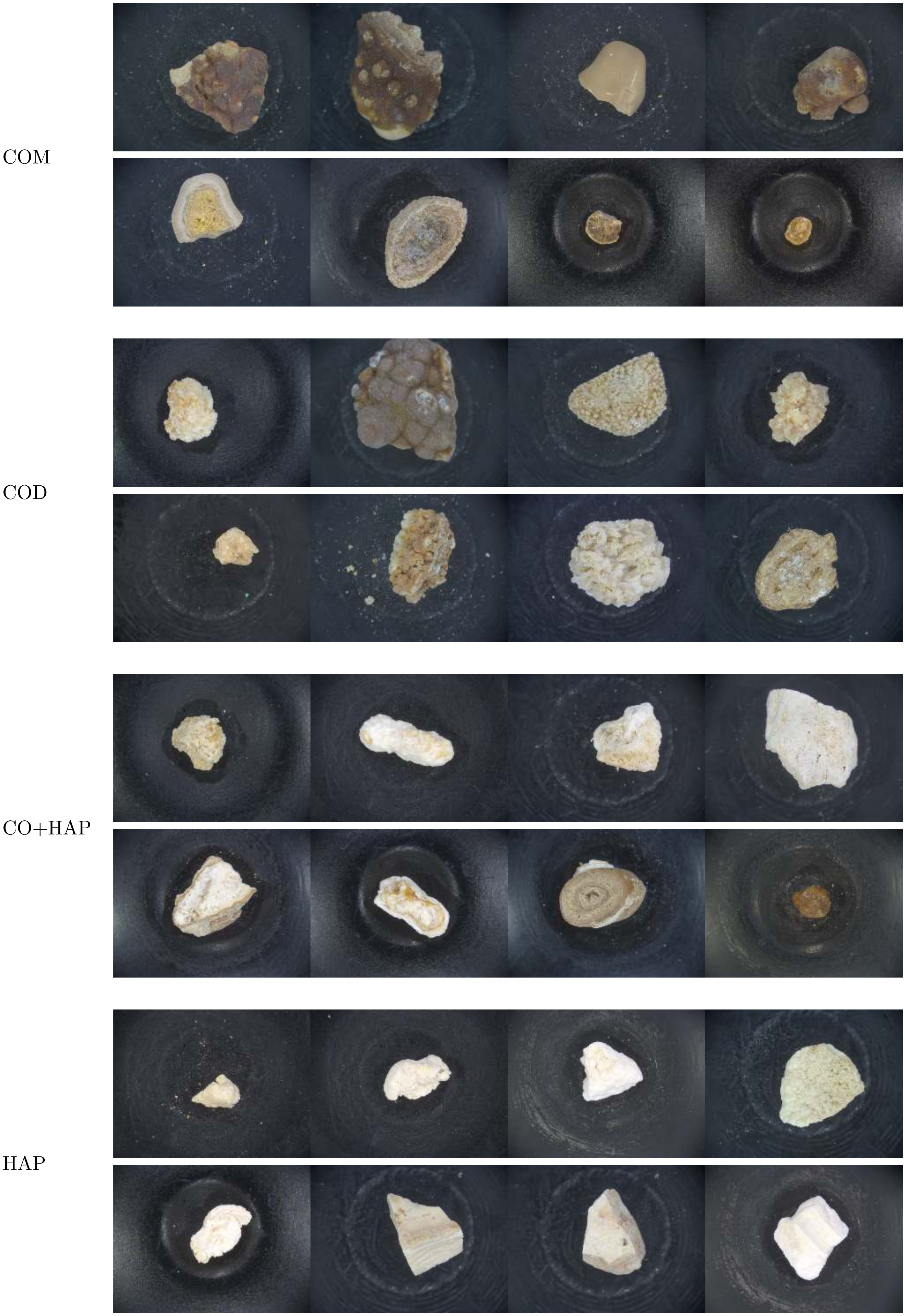


Figure 1: Sample images for the main stone classes acquired with white LEDs lighting. First row shows the external view of a fragment, second row the view of a section (core), most times of another fragment.

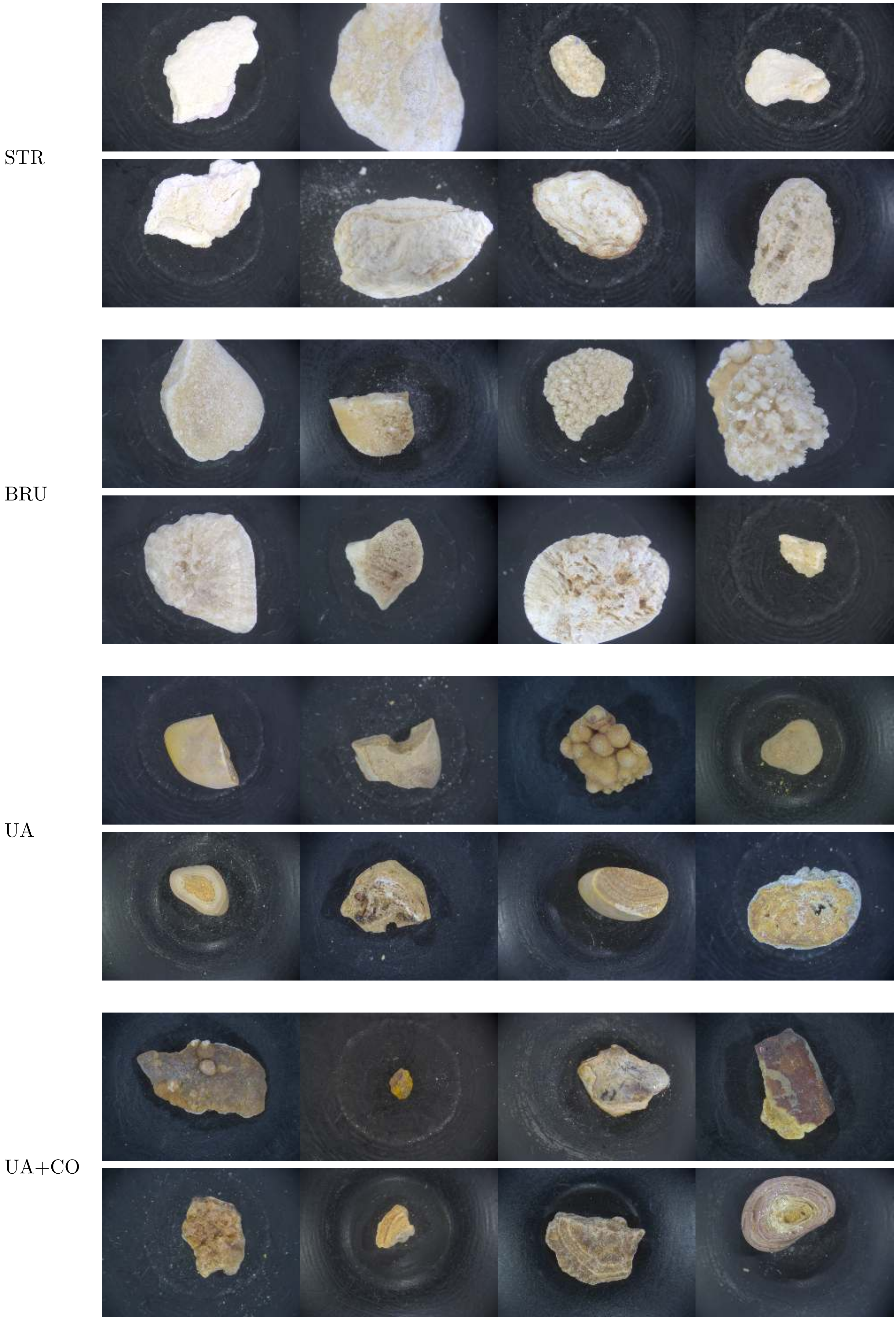


Figure 2: Continuation of figure 1.

## 3 Device

The device, specially developed for the analysis of kidney stones, assembles low-cost off-the-shelf components, being a simple, compact and robust piece of equipment. The final prototype consists of an enclosure that holds the camera, lens, focusing ring and a lighting board, plus a detached base acting as sample holder. Figure 3 shows an schematic view. All the structural components have been 3d printed in ABS plastic. Dimensions are 70 × 70 × 85 $\mathrm{mm}^3$, making it suitable to be placed on an desktop without disturbance. Figure 4 shows a picture of the first batch of prototypes where we can apreciate the relative size of the system.

The camera is a conventional RGB CMOS 5 megapixel small camera, model daA2500–14uc from Basler[1]. Its sensor spectral sensitivity allows us to acquire color images both in the visible and near infrared ranges (below 1000 nm) as we will explain. To this end, we got removed the cut-off IR coating and filter from the lens and camera, respectively. Image resolution is 2592 × 1944 pixels/channel. The camera is connected to the host computer with an USB3 cable through which our software obtains the pictures, sets the acquisition parameters, like the exposure time, and also sends two output signals that control the lighting board. In order to obtain similar colors across the several device exemplars we calibrate the gain parameter for each channel with a white background template.

The sample is located just about 60 mm below the sensor plane. At this distance a conventional lens with fixed aperture has a shallow depth of field and consequently a large part of the fragment surface is often out of focus. For this reason, our design includes a modified version of the conventional Evetar M12B1216IR lens, which is 12 mm focal length and F1.6 aperture (wide open). We contacted with the manufacturer[2] who provided us a F16 version of this lens, resulting in a larger depth of field whereby almost all the surface is in focus for almost all fragments. The loss of light by narrowing the aperture is not an issue because we control the time exposure and the light intensity. The focus is manually adjusted by rotating a wheel in contact to the lens.

The device is equipped with a specific lighting board. A ring of LEDs is arranged around the aperture for the camera lens, as shown in Figure 3. There are two sets of four white LEDs `w1`, `w2`, plus two other sets of four infrared LEDs `ir1`, `ir2`. The later emit light around 880 nm and 940 nm, respectively. The LEDs in each set are placed in cross shape and are switched on at the same time, to avoid self shadowing. Two on/off triggered signals control through a decoder the four groups.

[1] www.baslerweb.com

[2] www.leadingoptics.com

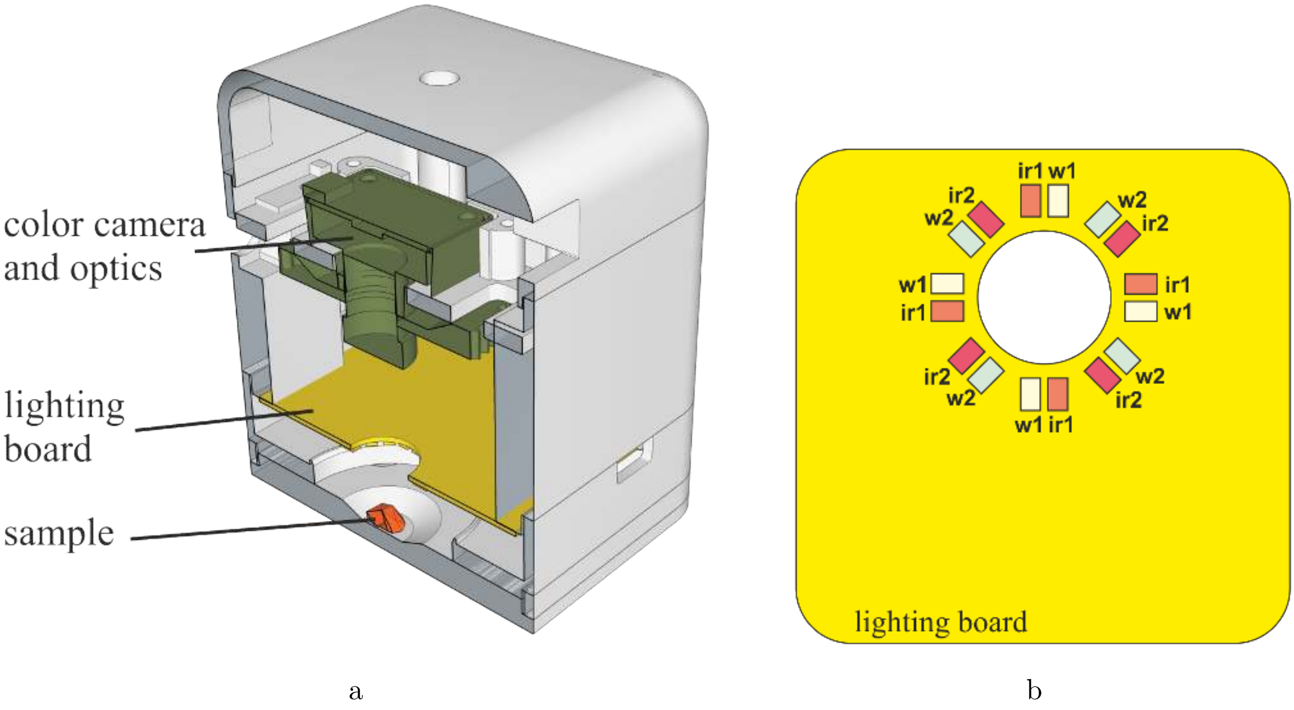


Figure 3: (a) Section of the device. (b) Lighting board.

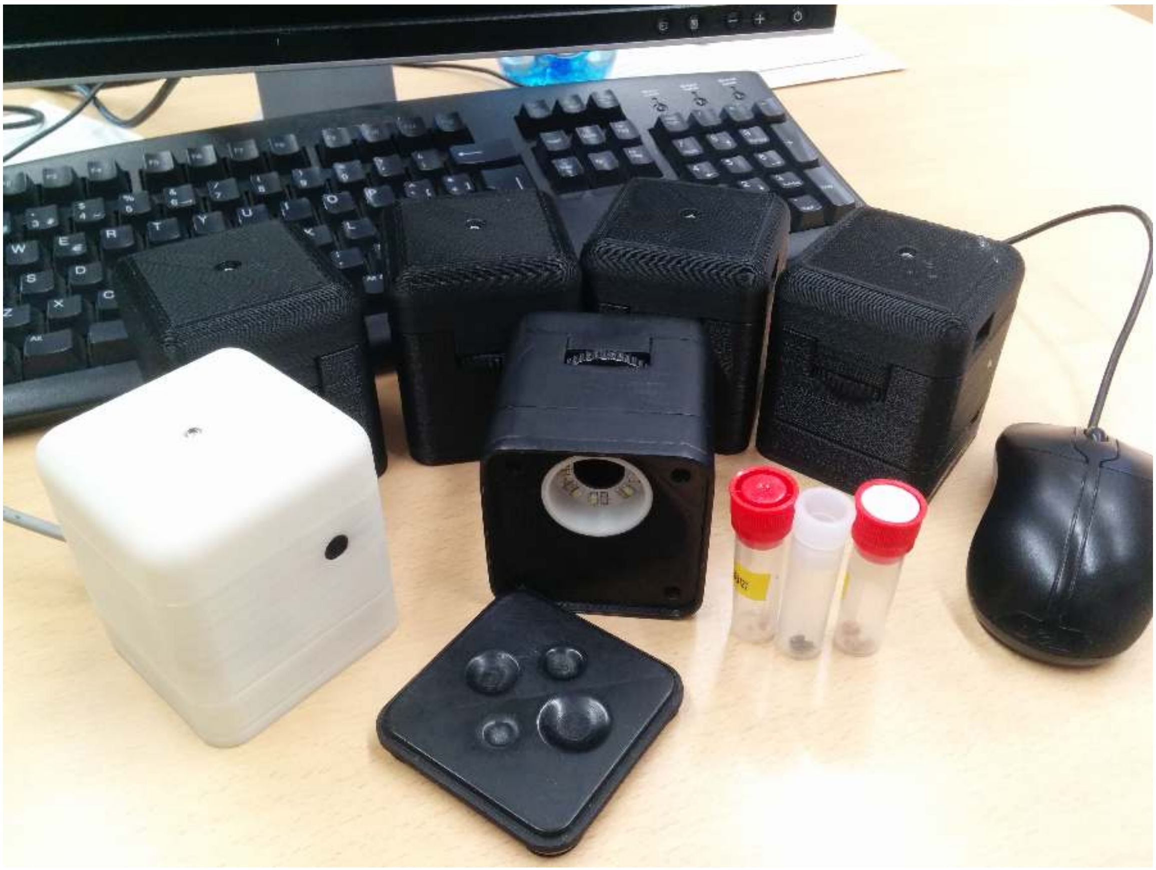


Figure 4: First set of prototype devices.

## 4 Dataset

The dataset of images is built upon a collection of 454 samples kindly provided by the urology department of the Hospital Universitary de Bellvitge (Barcelona, Spain) in the time span of several years. They cover all the main 9 classes but cystine, for which just 4 samples were available so we discarded this class as mentioned above. As for the rest, we tried to get all second scheme classes balanced and, at the same time, to record as much examples as possible to account for intraclass variability. The resulting percentages are shown in table 1. Note that HAP, BRU, STR and AU+CO have slightly more samples than their natural frequency.

One sample consists of two stone fragments from one same unique stone producer and episode, in order to assure independence. Stones are expelled either naturally or after extracorporeal shock wave lithotripsy. In the first case one of the stones is cut with the aid of a scalpel to expose its inner part. In the second, the operator has to select two fragments, each one with an inner and an outer surface. This is important because some classes exhibit characteristic traits in the inner part, like formation of nuclei or concentric layers, as mentioned. The classification of a kidney stone based just on its the external surface or just one fragment is faster, because it requires less time for fragment handling and image capture and processing. However, in preliminary experiments it proved to be less accurate than using both fragments, so we discarded this way.

For each side of each fragment we capture a series of 8 images by varying the exposure time —0.5 and 1 seconds— and light source —`w1`, `w2`, `ir1` and `ir2` LEDs. Hence, the dataset has a total of 14528 color images with a resolution of 2592 × 1944 pixels. Figure 5 shows some of them for two fragments.

In spite of this large figures, we are not going to use all the 32 images per sample in the classification. We recorded them in order to explore if they contained complementary information that could be advantageous for the segmentation and classification. As we will see, the infrared lighting `ir1` is quite suitable to obtain an almost perfect automatic segmentation of the stone region, but features computed over it did not result more discriminant or complementary to those computed over the white light illumination. In the end, to the effect of the stone analysis each sample is composed of 8 images: two fragments, two sides per fragment under white `w1` lighting at 1 second exposure time (for feature computation) and near infrared `ir1` at 0.5 seconds exposure time (for automatic segmentation), totaling 3632 images.

| Main | Class | | Dataset | | Natural |
|---|---|---|---|---|---|
| components | scheme 1 | scheme 2 | samples | percent | frequency |
| COM | 2 | 2 | 31 | 6.8 | 29.3 |
| | | 2b | 29 | 6.4 | |
| | | 2codt | 30 | 6.6 | |
| COD | 3 | 3b | 27 | 5.9 | 33.8 |
| | | 3t | 53 | 11.7 | |
| | | 3bt | 59 | 13.0 | |
| CO+HAP | 4 | 4 | 63 | 13.9 | 11.2 |
| HAP | 5 | 5 | 19 | 4.2 | 7.1 |
| STR | 6 | 6 | 38 | 8.4 | 4.1 |
| BRU | 7 | 7 | 14 | 3.1 | 0.6 |
| UA | 8 | 8 | 61 | 13.4 | 8.2 |
| UA+CO | 9 | 9 | 30 | 6.6 | 2.6 |
| CYS and others | | | | | 3.1 |
| Total | | | 454 | | |

Table 1: Percentage of kidney stone classes and subclasses in our dataset and naturally appearing according to (Grases et al., 2002). Even though figures vary depending on the world location of the study, COM, COD and UA calculi stand for the vast majority and their proportions are similar to those in the dataset.

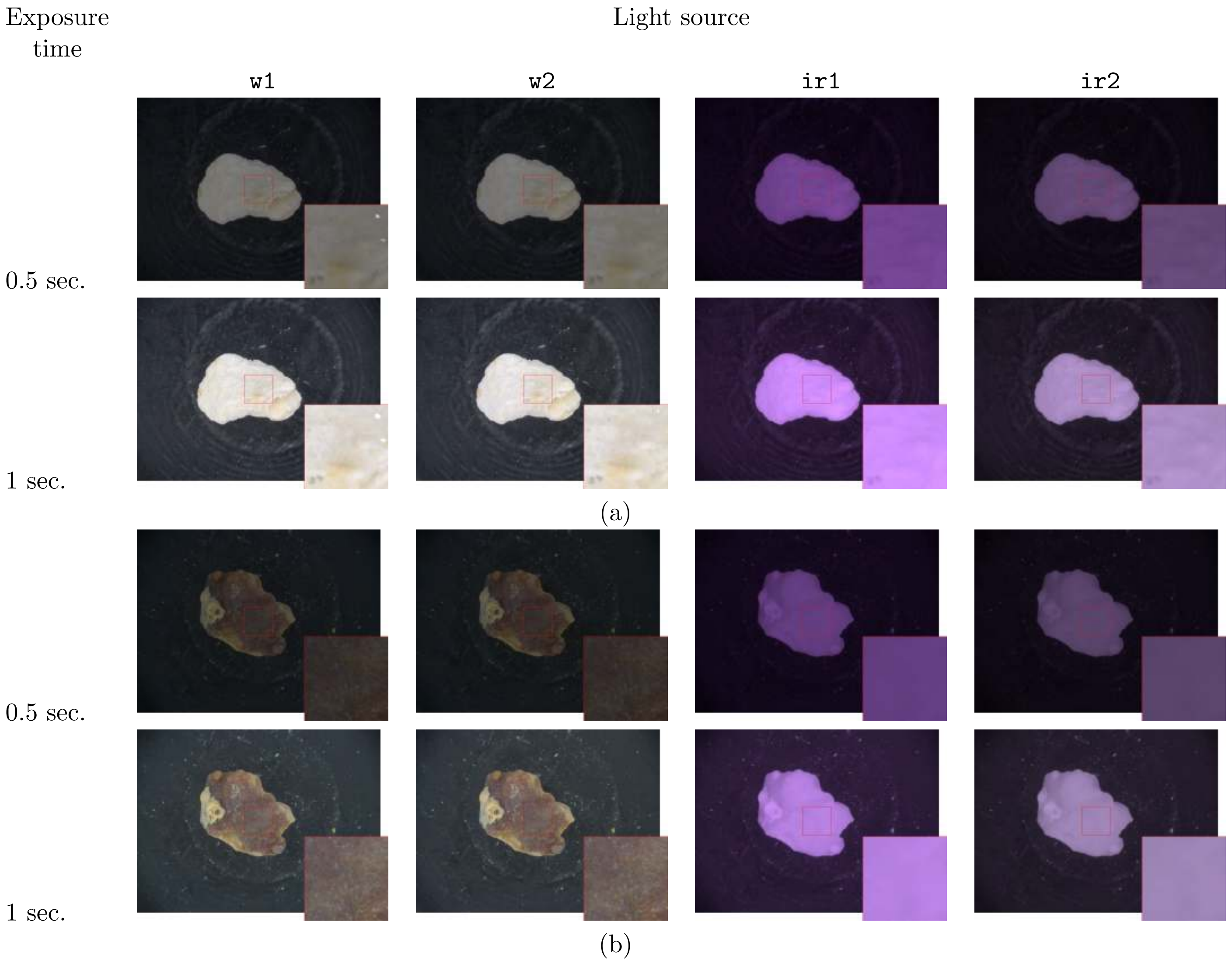


Figure 5: Set of 8 images taken for the external surface of an (a) struvite, (b) COM fragment. Typically struvite stones are the brightest and COM the darkest.

## 5 Classification

Our classification unit, a sample, is the set of four images of the external and internal view of the two fragments from one same stone or two stones produced by a subject. For each one of such images we compute a set of simple color and texture features, restricted to the region of interest (ROI) containing the stone fragment. Segmentation of this ROI is not trivial on the images recorded with white LEDs because of the wide range of brightness of the different classes and reflections of the background. But it turned out to be much easier on images acquired with infrared lighting (see for instance the two rightmost columns of figure 5). Actually, this is their main use, since features computed on them did not add much discriminative power to those computed on visible lighting images. Specifically, a simple thresholding on the image captured with `ir1` LEDs and exposure time equal to 0.5 seconds, followed by selection of the largest region, already yields a good segmentation in all cases. Given the clear bimodality of the histogram, the threshold value is determined by the classical Otsu method (Otsu, 1979; Sezgin and Sankur, 2004), which has proven quite reliable in our case.

We tried many combinations of feature types and their parameters, on all four light sources and exposure times, which we do not report here. In the end, the best results were achieved with the following features computed on images under white LEDs and 1 second exposure time:

- Rotational invariant local binary patterns (Ojala et al., 2002), computed with radius 2, 6 and 8 pixels, and 8, 8 and 10 sample points, respectively. They seem to represent well the different multiscale textures. Even though they are color textures, it was sufficient to compute them on a single channel (red).

- Color histogram, quantizing the color space into 16 bins per channel, in total thus a vector of dimension 4096.

- Gray level histogram, after converting from RGB to greyscale by simple channel averaging. In spite of already using the color histogram, we added this feature because it contributed to a slight increase in the total accuracy.

All these features, for all the four views of a sample (external and internal surfaces of two fragments), are concatenated into a single vector. To simplify and fasten the learning phase, we applied a simple dimensionality reduction technique. We removed those dimensions with variance below a threshold (e.g. bins corresponding to inexistent or very infrequent colors in the dataset). In order to determine its value and the sensitivity of the result to it, we fixed the features and computed the accuracy for a wide range of thresholds. Figure 6a shows that for a value of 1e-04 or below the global accuracy does not change significantly. The reduction is however important, keeping just about 700 features (figure 6b).

We obtained the best performance with a random forest classifier (Breiman, 2001) with 100 trees and 40% of features considered in each split. The quality of the splits was measured with the entropy criterion. Random forests obtained a slightly better performance than another classical method, Support Vector Machines, with the advantage of easier parameter optimization. We defer the discussion of the quantitative results to section 7, when we will have completed the presentation of all the classifier variants.

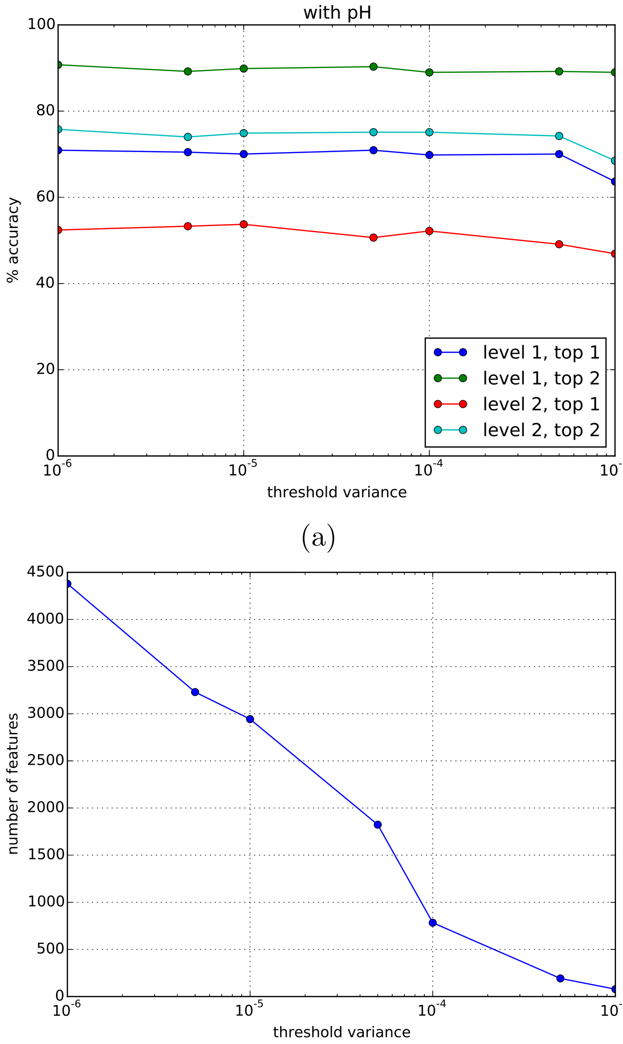
In cases of sufficient training data one divides the dataset into train, validation and test partitions, and then optimizes these parameters by repeatedly training and then checking the result on the validation set. However, our dataset is not that large: even though for each sample we have two fragments and many images, there few samples per class in comparison with other datasets, and specially for the two low frequency classes HAP and BRU. Hence we decided to train and test by the leave-one-out strategy so as to maximize the size of the training set : for every sample, train with all the rest and then test just with it. This means whenever we wanted to assess a new feature or change some parameter of the classifier or the feature computation, we had to train $N$ classifiers using the remaining $N-1$ samples, being $N = 454$ the size of the dataset.

This decision had an important implication with regard to the optimization of the many parameters of the features and classifier[3]. Unless training one leave-one-out classifier could be performed fast, searching for the optimum parameters would have been an extraordinary long task. Fortunately, our development environment was a CentOS Supermicro cluster with 28 nodes, each with 16 x86 64 GB cores. Thanks to this facility we were able to launch all the $N$ training jobs in parallel and get the result in a mere matter of minutes. We implemented the feature extraction and classification in Python on the Scikit-learn and Scikit-image (Pedregosa et al., 2011; van der Walt et al., 2014) libraries.

## 6 pH level as a feature

Major chemical components of a kidney stone start to precipitate below or above a certain pH level of the urine. Table 2 contains the ranges for each component according to several studies (Grases et al., 2002; Bichler

[3] For example, just for the LBP features we have to find the best number and size of the radius, the number of points on the circle in each case, from which channel to compute them, and whether we want standard, rotational invariance or non-uniform LBPs.



(a)

(b)

Figure 6: (a) accuracy versus threshold on variance of individual components in the feature vector, (b) number of selected features versus variance threshold.

et al., 2002; Spettel et al., 2013). We realize from this table that the pH level range, while not different for every class, clearly separates them into two groups. It thus may help to differentiate samples of visually similar classes like HAP and STR from BRU, and UA, UA+CO from calcium oxalate stones COM, COD, CO+HAP. In addition, it is a measure very easy and fast to obtain. Hence, we have integrated it into the classification pipeline, giving rise to a variant of the former classifier.

There are however two caveats. First, the values in table 2 are approximate and the pH level may vary with time for a subject, so we can not take them as sharp class borders. Second, we do not have the real pH for the samples in the dataset as we would like. Hence, we have figured out a reasonable pH level from the groundtruth class that we do know. In the intent of achieving a realistic simulation of the actual unknown pH value, given the groundtruth class of a sample, we draw a random value $p$ from a uniform distribution between the class limit in table 2 and a reasonable minimum/maximum urinary pH level that we have set to 4.5 and 7.0, respectively.

Now, we have to combine this new feature with those derived from images. The random forest classifier we employ produces not only the class label for the most probable class of a sample but also the probability of belonging to each one of the classes. Let $n$ be the number of classes and $P(C_i|\text{visual features}),\ i = 1 \ldots n$ these probabilities. A second vector of probabilities is obtained from the estimated pH value $p$ as follows,

$$P(C_i|\text{pH} = p) = \begin{cases} \sigma(p\,;\, l_i, s) & \text{if CO+HAP, HAP, STR} \\ 1 - \sigma(p\,;\, l_i, s) & \text{else} \end{cases} \tag{1}$$

being $l_i$ the class border for class $C_i$ found in table 2 and $\sigma(x\,;\, c, s) = (1 + e^{-(x-c)/s})^{-1}$ a scaled sigmoid function centered at $c$. Its purpose is to smooth the probabilities at class borders in a controlled way through the parameter $s$, as shown in figure 8. This contributes to avoid an excessive weight for the pH feature in the final decision. Like before with the threshold variance, figure 7 shows that the classifier accuracy is not much sensitive to a wide range of values of $s$ around the chosen value $s = 0.1$.

Considering the pH measure and the visual features as independent, the classifier output is

$$\arg\max_i P(C_i|\text{visual features})\, P(C_i|\text{pH} = p) \tag{2}$$

| class | pH level |
|---|---|
| COM | $< 6.3$ |
| COD | $< 6.3$ |
| CO+HAP | $> 6.0$ |
| HAP | $> 6.5$ |
| STR | $> 6.8$ |
| BRU | $< 6.6$ |
| UA | $< 5.5$ |
| UA+CO | $< 5.5$ |

Table 2: Ranges of urinary pH level per class.

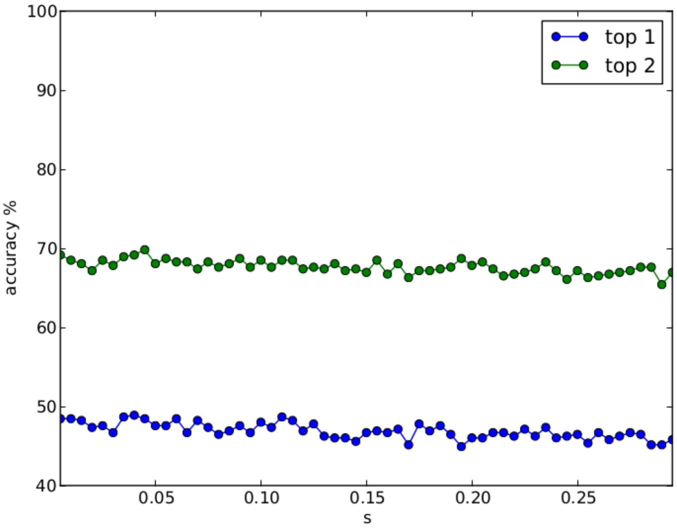


Figure 7: Accuracy of classification of detailed classes depending on the value of $s$.

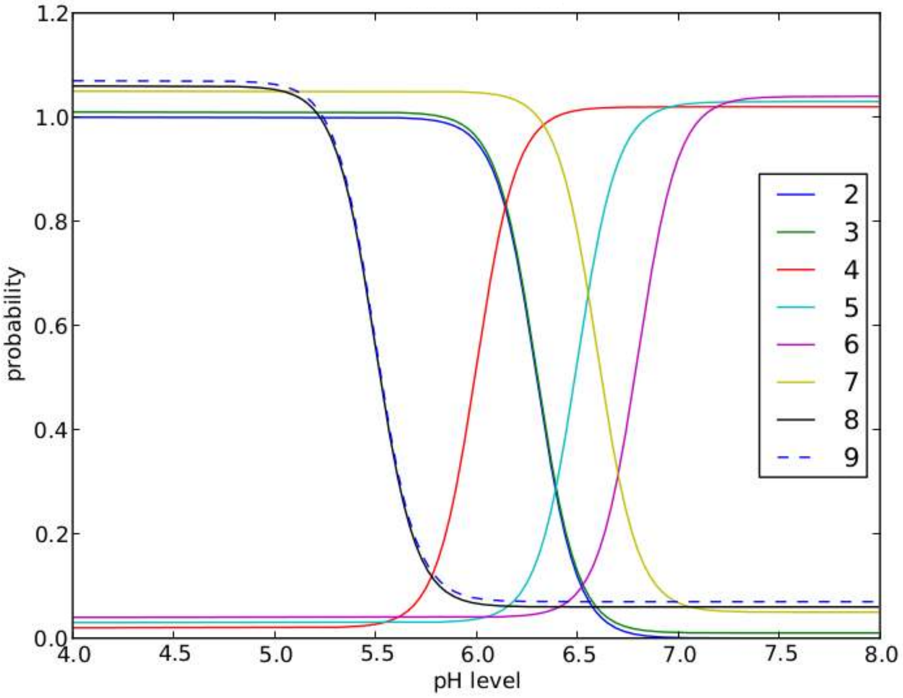


Figure 8: Class probabilities given pH level for $s = 0.1$. Curves have been vertically shifted for better visualization.

## 7 Results

We have obtained results for four variants of the random forest classifier based on the color and texture features described before. They are the combination of two factors:

- the two class schemes, with different number of classes (8 and 12)
- the use or not of the pH value

Table 3 shows their accuracy. Each figure is the mean of five runs of training and leave-one-out evaluation. Top-1 stands for the real class is equal to the predicted, whereas Top-2 means it is the most or second most probable according to the classifier. Standard deviations of Top-1 and Top-2 accuracies are less than 1.8 and 1.1, respectively. In all cases the threshold variance is set to 1*e*-04.

| Class scheme | Use of pH | Top-1 | Top-2 |
|---|---|---|---|
| 1 | No | 62.91 | 82.76 |
| | Yes | 70.01 | 88.95 |
| 2 | No | 43.90 | 66.92 |
| | Yes | 50.84 | 73.61 |

Table 3: Percent accuracy for four variants of the random forest classifier.

We note that the use of the attributed pH level consistently increases the accuracy around 7% in all cases. As already commented above, its effect is to disambiguate certain class confusions, specifically CO+HAP, HAP and STR from the rest. We confirm it through the confusion matrices in tables 4 and 5 for scheme 1 classes. The pH feature contributes mainly to better differentiate classes 3 and 4 (COD from CO+HAP). Also, it increases the precision of the minoritary class 5 (HAP). A similar trend is observed in the confusion matrices of scheme 2 classes shown in tables 6 and 7.

More importantly, there is a large drop in accuracy between class schemes 1 and 2, about 20%. Classification according to the 8 basic classes of scheme 1 was already a difficult problem because of large visual intra-class variations and also inter-class similarities. In effect, while in the medical literature (Daudon et al., 1993; Cloutier et al., 2015; Grases et al., 2002) one can find distinct descriptions of prototypical samples from such classes, reality is much more complex and varied. Sometimes even the human expert hesitates and resorts to complementary analysis (like infrared spectroscopy) to support his/her decision. The subdivision of COM and COD to form the second scheme, with the introduction

| | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | % recall |
|---|---|---|---|---|---|---|---|---|---|
| 2 | **72** | 14 | 1 | | | | 3 | | 80 |
| 3 | 20 | **107** | 9 | | | | 2 | 1 | 77 |
| 4 | 7 | 13 | **34** | 2 | 6 | | 1 | | 54 |
| 5 | 1 | | 6 | **2** | 10 | | | | 11 |
| 6 | | 4 | 11 | 3 | **20** | | | | 53 |
| 7 | | 6 | | | 4 | **0** | 4 | | 0 |
| 8 | 1 | 6 | 6 | | | | **43** | 5 | 70 |
| 9 | 8 | 5 | | | | | 13 | **4** | 13 |
| % precision | 66 | 69 | 51 | 29 | 50 | 0 | 65 | 40 | |

Table 4: Confusion matrix for scheme 1 classes, no pH

| | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | % recall |
|---|---|---|---|---|---|---|---|---|---|
| 2 | **73** | 16 | | | | | 1 | | 81 |
| 3 | 20 | **115** | 2 | | | | 2 | | 83 |
| 4 | 2 | 3 | **52** | | 4 | 2 | | | 83 |
| 5 | | | 7 | **4** | 8 | | | | 21 |
| 6 | | | 16 | 3 | **19** | | | | 50 |
| 7 | | 7 | 2 | | | **3** | 2 | | 21 |
| 8 | 1 | 8 | | | | | **47** | 5 | 77 |
| 9 | 8 | 5 | | | | | 13 | **4** | 13 |
| % precision | 70 | 75 | 66 | 57 | 61 | 60 | 72 | 44 | |

Table 5: Confusion matrix for scheme 1 classes using pH

of new four classes, only aggravates the problem. This is appreciated in the drop of recall for classes 2, 2b and 2codt for example, formerly 80 and 77% and now between 30 and 40%. The cause is to be found in the top-down nature of the random forest classifier, whereby at each node a certain feature and threshold value are selected to perform a good partition of training samples. The relative good news is that most of the confusions for the new classes are among themselves, that is, most of the errors in the group 2-2b-2codt are confined to the same group, and similarly for 3b-3t-3bt. This is just another manifestation of the former cause. Moreover, the rest of mistakes for each of these two groups are with the other group, keeping the pattern of the confusion matrix of scheme 1.

A last observation is that for class scheme 1, classes 5 and 7 have a very low recall, though the use of pH slightly increases it. The reason is that they have the lower number of samples, 4.2 and 3.1%, respectively.

|  | 2 | 2b | 2codt | 3b | 3t | 3bt | 4 | 5 | 6 | 7 | 8 | 9 | % recall |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | **13** | 9 | 4 |  | 4 | 1 |  |  |  |  |  | 1 | 41 |
| 2b | 11 | **9** | 2 |  | 2 | 2 |  |  |  |  | 3 |  | 31 |
| 2codt | 5 | 3 | **10** |  | 7 | 1 | 3 |  |  |  |  |  | 34 |
| 3b | 1 |  |  | **12** | 5 | 7 | 7 |  |  |  | 2 |  | 35 |
| 3t | 6 | 3 | 2 | 8 | **22** | 8 | 1 |  |  |  | 4 |  | 41 |
| 3bt | 2 | 4 | 1 | 2 | 16 | **19** | 6 |  |  |  | 1 |  | 37 |
| 4 | 2 | 3 | 1 | 2 | 3 | 5 | **36** | 3 | 6 |  |  | 2 | 57 |
| 5 |  |  |  |  |  | 1 | 6 | **3** | 9 |  |  |  | 16 |
| 6 |  |  |  | 1 |  |  | 11 | 3 | **21** | 1 | 1 |  | 55 |
| 7 |  |  |  | 1 | 2 | 1 | 1 |  | 3 | **3** | 3 |  | 21 |
| 8 |  |  |  | 1 | 1 | 1 | 6 |  | 1 |  | **45** | 6 | 74 |
| 9 | 5 | 1 | 1 |  | 1 | 1 |  |  |  |  | 16 | **5** | 17 |
| % precision | 29 | 28 | 48 | 44 | 35 | 40 | 47 | 33 | 52 | 75 | 60 | 36 |  |

Table 6: Confusion matrix for scheme 2 classes and no pH

|  | 2 | 2b | 2codt | 3b | 3t | 3bt | 4 | 5 | 6 | 7 | 8 | 9 | % recall |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | **13** | 9 | 4 |  | 4 | 1 |  |  |  |  |  | 1 | 41 |
| 2b | 11 | **9** | 2 |  | 2 | 2 |  |  |  |  | 3 |  | 31 |
| 2codt | 5 | 3 | **11** |  | 7 | 2 |  |  |  |  | 1 |  | 38 |
| 3b | 1 |  |  | **18** | 5 | 7 | 1 |  |  |  | 2 |  | 53 |
| 3t | 6 | 3 | 2 | 8 | **23** | 8 | 1 |  |  | 1 | 2 |  | 43 |
| 3bt | 2 | 4 | 2 | 6 | 16 | **21** |  |  |  |  |  |  | 41 |
| 4 | 1 |  |  |  | 2 | 3 | **54** |  | 3 |  |  |  | 86 |
| 5 |  |  |  |  |  |  | 8 | **4** | 7 |  |  |  | 21 |
| 6 |  |  |  |  |  |  | 14 | 3 | **21** |  |  |  | 55 |
| 7 |  |  |  | 2 | 2 |  | 2 |  |  | **6** | 2 |  | 43 |
| 8 |  |  |  | 1 | 1 | 1 |  |  |  |  | **52** | 6 | 85 |
| 9 | 5 | 1 | 1 |  | 1 | 1 |  |  |  |  | 16 | **5** | 17 |
| % precision | 30 | 31 | 50 | 51 | 37 | 46 | 68 | 57 | 68 | 86 | 67 | 42 |  |

Table 7: Confusion matrix for scheme 2 classes using pH

In addition to the global accuracy we have delved into the contribution of each feature. They have been selected on the basis of their success in other works on color and texture classification. Table 8 shows the accuracy of the selected features for the classifier two class schemes without pH. GLH stands for gray level histogram, CH is color histograms and LBPs local binary patterns. The number appended to CH is the quantization step for each channel. In all cases the threshold on the feature variance was set to $1e$-04 but for CH8 when it was $1e$-05 to compensate for the lesser number of selected features. It is noteworthy that the color histogram alone is just 3% below the top result, and contributes more than LBP which was computed only on the red channel. We believe the reason is that the color histogram is the only true color feature, and color turns out to be more discriminative than texture alone. However, a uniform coarse quantization like that provided by CH8 is not enough, since the color variability must be concentrated in a small region.

| | Class scheme | |
|---|---|---|
| features | 1 | 2 |
| GLH | 54.41 | 37.44 |
| CH8 | 55.75 | 37.89 |
| CH16 | 60.23 | 40.31 |
| LBP | 55.51 | 34.14 |
| CH8+LBP | 62.78 | 43.83 |
| GLH+CH16+LBP | 62.91 | 43.90 |

Table 8: Contribution to the global accuracy of each type of feature and combinations.

# 8 Conclusions

We have presented the first attempt to automate the visual classification of kidney stones. Kidney stone formation affects a substantial percentage of population, and also has a high rate of recurrence that can be reduced if stones are properly identified. To this end we have built an image capture device based on standard components. With it, we have collected a dataset of thousands of images of both the external surface and inner sections of kidney stone fragments, corresponding to a total of 454 samples/subjects. These samples belong to the main eight classes of kidney stones identified in the literature, with a frequency similar to their natural occurrence.

Subsequently, we have designed a classifier based on color and texture features, which are fed into a random forest classifier. With a leave-one-out training strategy to maximize the number of samples in the training set, we reach an accuracy of 63% for these eight classes. This is a low figure but we are dealing with a difficult problem mainly because of a large intra-class variability. In its defense, we report a top-2 accuracy of 83% and the analysis of the confusion matrix. This later shows that most errors happen between classes which are similar in composition (2-3, 3-4, 8-9) and, more importantly, in their associated recommendations for the patient, as contained for instance in (Grases et al., 2006; Friedlander et al., 2015). With regard practical applicability, at present we think of myStone not as a tool for fully automatic diagnostic —from which it is still far from—, but for decision support for the urologist, as it may provide not just a single answer but the computed probabilities for each class.

Nevertheless, we have one way to improve the accuracy by leveraging the relationship between classes and the urinary pH level, which increases the accuracy by about 7%. Admittedly, we simulate it from the class groundtruth, but the literature supports its discriminative nature and the simulated values are not ideal but drawn from a random distribution. On the contrary, splitting the most populated classes COM and COD into three subclasses each has resulted in a drop of their accuracy, though we reach a notable 67% top-2.

In comparison, (Blanco et al., 2015) attains a much higher global accuracy of 90.4% with the classes of the first scheme. Only that this figure refers to the accuracy of *pixel* classification, not sample (four images) or even a whole image. For this they did not have to design any color or texture feature like us but used the signature of an hyperspectral camera at each pixel, a vector of more than 100 features. Despite their dataset was only composed of 6 to 8 fragments per class, the fact that they got this result with a classifier as simple as a quadratic discriminant analysis is rather surprising. We interpret it as the extreme importance of selecting a good set of features.

Given these results, we wonder which alternative features and/or classifier could perform better on regular color camera images. The present trend, deep learning, is to learn the features themselves and the classifier altogether. Hence we have already done a number of experiments with convolutional neural networks for feature learning, followed by a few fully connected layers to output class probabilities. After trying several combinations of network architectures and layers plus the typical ingredients of regularization (dropout, weight regularization), we have just reached an accuracy short of 3% with regard the approach described here. We can not afford doing leave-one-out partitions and parallelization of the training phase as done here, because of the long time required to train a neural network and the huge number of folds. Instead, a conventional $k$-fold partition for some small $k$ needs be adopted.

As future work we have identified in this context the following four lines. First, we hypothesize that our dataset needs to grow considerably before we can improve the present results with deep learning techniques, which are quite demanding in supervised data. Thus, we intend to add more samples to the dataset, trying at the same time to balance the number of samples per class. Second, the network architecture for classifying our samples can not be the same as the standard deep convolutional networks proposed for single images: each sample consists of four images which are not pixelwise compatible and thus can not be merged into a multichannel single image. Some new type of multiview architecture must be designed. Third, we have to deal with the problem of unbalanced classes, for instance through weights or a non-uniform sampling scheme to build the minibatches. And finally, deep networks are known for being overconfident on their predictions. This means that some calibration procedure needs to be applied to the scores they provide so as to approach them to actual probabilities. This is relevant to ours because the output of the system is not only the most probable class but also the confidence or belief on each of them.

## Acknowledgments

This research has been supported by the Accio (Generalitat de Catalunya), project VALTEC13-1-0148 and the CaixaImpulse program (CI15-00013). We also thank NVIDIA Corporation for the donation of a GPU board.

## References

Andreassen, K., Poulsen, A., Olsen, P., Aabeck, J., and Osther, P. (2007). Classification of urolithiasis in Denmark: a national survey. *European Urology*, 2(1):126.

Bichler, K.-H., Eipper, E., Naber, K., Braun, V., Zimmermann, R., and Lahme, S. (2002). Urinary infection stones. *International Journal of Antimicrobial Agents*, 19(6):488–498.

Blanco F., Lpez-Mesas, M., Serranti, S., Bonifazi, G., Havel, J., Valiente M. (2012). Hyperspectral imaging based method for fast characterization of kidney stone types. *Journal of Biomedical Optics*, 17(7):076027-1-12.

Blanco, F., Lumbreras, F., Serrat, J., Siener, R., Serranti, S., Bonifazi, G., Lpez-Mesas, M., Valiente, M. (2015). Taking advantage of hyperspectral imaging classification of urinary stones against conventional infrared spectroscopy. *Journal of Biomedical Optics*, 19(12):126004-1-9.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Cloutier, J., Villa, L., Traxer, O., and Daudon, M. (2015). Kidney stone analysis: “Give me your stone and I will tell who you are”. *World Journal of Urology*, 33:157–169.

Daudon, M., Bader, C., and Jungers, P. (1993). Urinary calculi: review of classification methods and correlations with etiology. *Scanning Microscopy*, 7(3):1081–1104.

Friedlander, J. I., Antonelli, J. A., and Pearle, M. S. (2015). Diet: from food to stone. *World Journal of Urology*, 33(2):179–185.

Grases, F., Costa-Bauzá, A., and Prieto, R. M. (2006). Renal lithiasis and nutrition. *Nutrition Journal*, 23.

Grases, F., Costa-Bauzá, A., Ramis, M., Montesinos, V., and Conte, A. (2002). Simple classification of renal calculi closely related to their micromorphology and etiology. *Clinica Chimica Acta*, 322(12):29 – 36.

Hesse, A., Brndle, E., Wilbert, D., Khrmann, K.-U., and Alken, P. (2003). Study on the prevalence and incidence of urolithiasis in germany comparing the years 1979 vs. 2000. *European Urology*, 44(6):709 – 713.

Kok, D. J. (2012). Metaphylaxis, diet and lifestyle in stone disease. *Arab Journal of Urology*, 10(3):240 – 249. Stones / Endurology.

Lumbreras, F., Serrat, J., and Rotger, G. (2017). myStone web site. `http://www.cvc.uab.es/mystone` [Accessed March 2017].

Nolde, A., Hesse, A., Scharrel, O., and Vahlensieck, W. (1993). Modellprogramm zur Nachsorge bei rezidievierenden Harnsteinpatienten (Model program for follow-up of recurrent urinary stone formers). *Urologe B*, (33):148–154.

Ojala, T., Pietikäinen, M., and Mäenpää, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(7):971–987.

Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Piqueras, S., Duponchel, L., Tauler, R., De Juan, A. (2011). Resolution and segmentation of hyperspectral biomedical images by multivariate curve resolution-alternating least squares *Analytica Chimica Acta*, 705(1):182-192.

Romero, V., Akpinar, H., and Assimos, D. G. (2010). Kidney stones: A global picture of prevalence, incidence, and associated risk factors. *Reviews in Urology*, 12(2-3):89–96.

Scales, C., Smith, A., Hanley, J., and Saigal, C. (2012). Prevalence of kidney stones in the United States. *European Urology*, 62(1):160–165.

Schubert, G. (2006). Stone analysis. *Urological Research*, 34(2):146–150.

Sezgin, M. and Sankur, B. (2004). Survey over image thresholding techniques and quantitative performance evaluation. *Journal of Electronic Imaging*, 13(1):146–168.

Siener, R. and Hesse, A. (2012). *Comparative Costs of Various Treatment Strategies and Preventive Measures*, pages 897–901. Springer London, London.

Spettel, S., Shah, P., Sekhar, K., Herr, A., and White, M. D. (2013). Using hounsfield unit measurement and urine parameters to predict uric acid stones. *Urology*, 82(1):22–26.

Strohmaier, W. L. (2011). *Economic Implications of Medical and Surgical Management*, pages 245–250. Springer London, London.

Strohmaier, W. L. (2012). Economics of stone disease/treatment. *Arab Journal of Urology*, 10(3):273 – 278. Stones / Endourology.

van der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., Gouillart, E., Yu, T., and the Scikit-image contributors (2014). Scikit-image: image processing in Python. *PeerJ*, 2:e453.

Wagner, W. P. (2017). Trends in expert system development: A longitudinal content analysis of over thirty years of expert system case studies. *Expert Systems with Applications*, 76:85 – 96.