Automatic detection of calcium phosphate deposit plugs at the terminal ends of kidney tubules

Katrina Fernandez[1,2], *Mark Korinek*[1], *Jon Camp*[1], *John Lieske*[3], *David Holmes*[1] ✉

[1]*Biomedical Imaging Resource, Mayo Clinic, Rochester, MN, USA*
[2]*University of Minnesota, Minneapolis, MN, USA*
[3]*Department of Nephrology & Hypertension, Mayo Clinic, Rochester, MN, USA*
✉ *E-mail: holmes.david3@mayo.edu*

Published in Healthcare Technology Letters; Received on 26th September 2019; Accepted on 2nd October 2019

Kidney stones are a common urologic condition with a high amount of recurrence. Recurrence depends on a multitude of factors the incidence of precursors to kidney stones, plugs, and plaques. One method of characterising the stone precursors is endoscopic assessment, though it is manual and time-consuming. Deep learning has become a popular technique for semantic segmentation because of the high accuracy that has been demonstrated. The present Letter examined the efficacy of deep learning to segment the renal papilla, plaque, and plugs. A U-Net model with ResNet-34 encoder was tested; the Letter examined dropout (to avoid overtraining) and two different loss functions (to address the class imbalance problem. The models were then trained in 1666 images and tested on 185 images. The Jaccard-cross-entropy loss function was more effective than the focal loss function. The model with the dropout rate 0.4 was found to be more effective due to its generalisability. The model was largely successful at delineating the papilla. The model was able to correctly detect the plaques and plugs; however, small plaques were challenging. Deep learning was found to be applicable for segmentation of an endoscopic image for the papilla, plaque, and plug, with room for improvement.

1. Introduction: Proper kidney function is required for blood filtering, maintaining homeostasis, and synthesising hormones. Unfortunately, the kidneys can be afflicted with diseases that alter its ability to function. One such affliction is kidney stone disease, in which hard deposits made of minerals and salts, usually calcium oxalate or calcium phosphate coalesce into granules. Kidney stone disease is associated with an increased risk of end-stage renal failure [1, 2], renal cell carcinoma [3], and other life-threatening conditions. Worldwide, there has been an increase in the number of people afflicted by this disease [4]. Also, once an individual has a first kidney stone episode, the likelihood of another kidney stone episode increases [5]. Owing to these phenomena, individuals must be aware that they are at risk to make lifestyle changes such as increasing fluid intake and decreasing salt intake [6–8]. In addition to lifestyle changes, several different treatment strategies have been developed and approved to reduce the likelihood of future stone development. [9]. Given that these treatments affect different biological pathways [10], the nephrology research community has been looking for biomarkers, which can inform treatment [9, 10]. One way to identify those at risk for recurring kidney stones is to assess the incidence of their precursors in the kidney.

There are two precursors or pathways for the formation of calcium-based kidney stones. One precursor is a plug, which is the result of crystal formation and retention within collecting ducts [11]. The other pathway is plaque, which is made up of sub-epithelial apatite crystals found on the surface of the papilla [11]. Identifying these precursors allows for removal before stone formation and identifying individuals at risk for a recurring kidney stone episode or linked conditions [12–14].

Historically, the primary characterisation of stone precursors has been chemical analysis and histologic analysis. However, there has been progress in the development of intact imaging for precursor characterisation [15]. Micro computed tomography (CT) is an effective tool; however, it requires the extraction of plugs and plaques. It cannot be used *in vivo*. Clinical CT does not provide the resolution required to identify plugs and plaques. Moreover, as demonstrated in [16, 17], clinical CT findings do not correlate with precursor burden. One method for detection and characterisation of stone precursors is video endoscopy [13]. In this procedure, the physician uses an endoscopic camera to interrogate each of the poles of a kidney. Images are digitally recorded and analysed offline. Data collection for video characterisation often occurs during the removal of existing stones (see [9] for guidance on minimally invasive removal of stones).

Currently, there are two approaches to assess stone precursors. The first approach is subjective, requiring the surgeon to review the images and score each pole based on the surgeon's observations [13, 18]. The second approach is quantitative characterisation [17]. This approach was validated against chemical and histologic analysis. Unfortunately, identifying stone precursors from the video images, via manual segmentation, is a tedious process. A trained analyst must scan through the video sequence to video segments, which correspond to the individual poles. Once determined, the analyst identifies a representative image containing the papilla and any precursor deposits. The analysis time varies from patient to patient but cases routinely require 60 min of analyst time to characterise a single subject. Owing to the effort and expertise required, it is prudent to identify alternate segmentation approaches that can be automated to decrease the amount of time to create relevant data and reduce human variability (Fig. 1).

Machine learning is one approach for automating video analysis process. It is often used to automate and decrease the amount of time it takes for tasks such as segmentation and classification. Specifically, in the medical field, it has been used for semantic segmentation tasks such as outlining nuclei [19], categorising, segmenting brain tumours [20] etc. Deep learning is a subset of machine learning that has demonstrated superior performance with reduced featuring engineering [21, 22]; as such, it does not require a large amount of data preprocessing and feature extraction. Instead, the architecture learns the features from the data.

There are various deep learning architectures. A commonly used architecture is the U-Net, a convolutional neural network. The U-Net's popularity is due to its ability to encode and decode. This functionality allows it to steadily decrease spatial dimension via pooling layers, the encoding architecture, and then recover the spatial dimension along with object details, the decoding architecture [23].

doi: 10.1049/htl.2019.0086

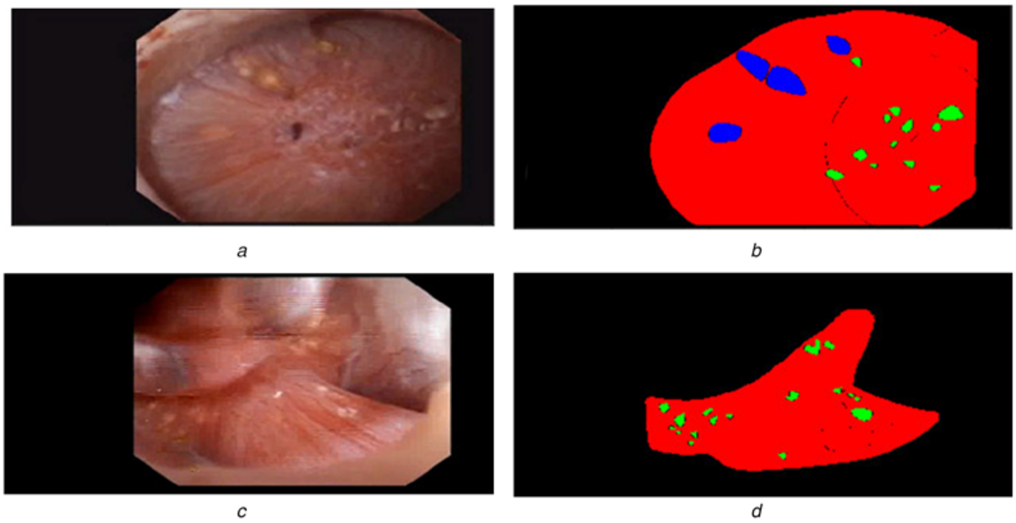

Fig. 1 *Representative endoscopic imaging and corresponding object maps*
a, c Frames from an endoscopic video
b, d Corresponding labelled imaged outlining the papilla, red, the plaque, green, and the plugs blue

The ResNet encoder is a network that uses a skip connection to avoid the vanishing/exploding gradient problem that occurs when the number of layers in a deep learning network is increased. This connection adds an input to its output after several weight layers [24]. Encoders vary in the number of parameters and hidden layers. They also vary in memory size to fit various computing restrictions.

Deep learning is not only common in fields such as cell biology and neurology; it has also been utilised in nephrology. Past studies have applied deep learning for stone detection [25, 26], renal cell carcinoma [27] etc. However, publications to date do not describe the use of deep learning algorithms for kidney plug and plaque detection.

Given the clinical demand for renal mapping of stone precursors, there is a need to reduce the time for segmentation and decrease human variability in the analysis pipeline. The goal of this Letter is to use previously developed deep learning architectures to improve the processing pipeline for plug and plaque characterisation. Although there are validated architectures for semantic segmentation of video data, none has been used in this domain. In this Letter, the use of deep learning models to identify the precursors of kidney stones and renal papilla is explored.

2. Method: Over 200 endoscopic videos were recorded from patients after they underwent kidney stone removal surgery. As per the institutional review board (IRB) protocol, clinical information was excluded from the image analysis to avoid any bias in the processing. Owing to the focus on the plugs and plaques, video images of the stone were not included in the analysis. These videos were then manually segmented for the papilla, plaque, and plug using Mayo Clinic N (Biomedical Imaging Resource, Rochester, MN), a visualisation and analysis software for medical imaging. A single trained technician processed all videos under the direct supervision of a board-certified urologic surgeon. For each analysis, the technician selected representative images for each pole. The image was loaded into ANALYZE and the papilla, plaques, and plugs were hand traced. The labelled image was stored as an Object Map and an ANALYZE file format for segmented and labelled data. Manual analysis is described in detail in [17].

During the data preparation phase, the manual segmentations were extracted from the video sequences along with the corresponding labelled images. Python 3.7.3 (Anaconda, Inc., Austin, TX) was used for all preprocessing. Each red, green, and blue image was padded and/or resized to 480×1024 px^2 and saved into an hdf5 file (to facilitate reading batches of images during the training). The labelled image contained the four different objects – background, papilla, plaque, and plug. Each object was assigned a numeric value and resized to 480×1024 px^2 to maintain correspondence with the image data.

The deep learning models were created using the segmentation models library, a Python library for image segmentation that is built on the Keras/TF libraries. (https://github.com/qubvel/segmentation_models). Several different deep learning architectures were explored. The final model selected was a U-Net. The U-Net model was implemented with a ResNet-34 encoder. The pre-trained ResNet-34 encoder had been previously trained on ImageNet, which is well-suited for video/photographic image analysis. The deep learning algorithm was trained on 1666 normalised images and tested on 185 normalised images with a batch size of 7 for 500 epochs. An activation layer of softmax, upsampling decoder architecture, and batch normalisation in between the convolution and activation layers were used. For the decoder blocks, the convolution layer filters used were 256, 128, 64, 32, and 16. One algorithm had a dropout rate of 0 and the other had a dropout rate of 0.4. The models were compiled with the ADAM optimiser. The metric used for these models is the intersection over union (IOU) score or the Jaccard index, where A and B are sample sets (1). The Jaccard index is a measurement that compares the similarity and variance between sample sets

$$\mathrm{IOU} = J(A,\ B) = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

Two loss functions were tested, primarily to address the class size imbalance between the normal tissue, papilla, plaques, and plugs. The first loss function used was the sum of the Jaccard loss and the categorical cross-entropy equation, where N is the number of training examples, $\hat{y}_j$ is the output, y is the true output, and J is the number of output nodes as in the equation below:

$$\begin{aligned} \mathrm{JCE} = (1 - J(A,\ B)) + \frac{-1}{N} \sum_{i=0}^{N} \sum_{j=0}^{J} y_j * \log(1 - \hat{y}_j) \\ + (1 - y_j) * \log(1 - \hat{y}_j) \end{aligned} \tag{2}$$

The second loss function used was a focal loss (3).

Focal loss attempts to take into account the probability of observing a given class during the training. In this model, m is the number

doi: 10.1049/htl.2019.0086

of classes, c is the class, and o is the observation

$$\text{focal loss} = \sum_{c=1}^{m} -\alpha * (1 - p_{o,c})^{\gamma} * y_{o,c} * \log(p_{o,c}) \qquad (3)$$

After training, the deep learning algorithms were evaluated on the training and test datasets to obtain their overall loss and IOU scores.

3. Results: The performances of the models with the Jaccard-cross-entropy (JCE) and focal loss are reported in Table 1. The JCE loss model was able to achieve excellent performance on the train dataset but moderate performance on the test dataset. This is highlighted by the test loss value, which is ten times larger than the train loss value and the 22% lower IOU score on the test dataset. This is a typical example of an overfitted model and represents low generalisability. The focal loss model also exhibited overfitting and low generalisability as the test loss value is 100 times larger than the train loss value and had more than a 24% lower IOU score on the test dataset.

When considering the dropout, the models differed in their loss value and Jaccard index for the datasets, as seen when evaluated on the training dataset, the non-dropout model had a lower JCE loss score (0.0575) and a higher IOU score (0.958) than the model with the dropout rate of 0.4 (0.174 and 0.900, respectively). However, the deep learning algorithm with the 0.4 dropout rate outperformed the regular modal for the test dataset.

Figs. 2 and 3 provide representative examples of successful and unsuccessful segmentations, respectively. In Fig. 2, the papilla (red) is delineated from the surrounding tissue using the dropout model.

Table 1 Metrics and loss scores of the deep learning algorithms on the training and test datasets

Type of loss	Dataset	Loss	IOU score
JCE	train	0.174	0.900
f	train	0.0679	0.709
JCE	test	1.11	0.679
f	test	15.5	0.469

JCE = sum of Jaccard loss and categorical cross-entropy.
F = focal loss.

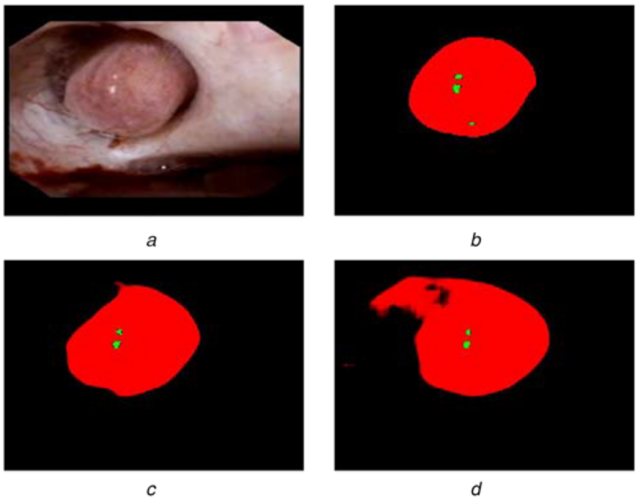

Fig. 2 *Successful segmentation*
a Input given to the deep learning algorithms after training to compare performance
b Ground truth of Fig. 2*a*
c Prediction made by the deep learning algorithm with the dropout rate of 0 based on Fig. 2*a*
d Prediction made by the deep learning algorithm with the dropout rate of 0.4 based on Fig. 2*a*

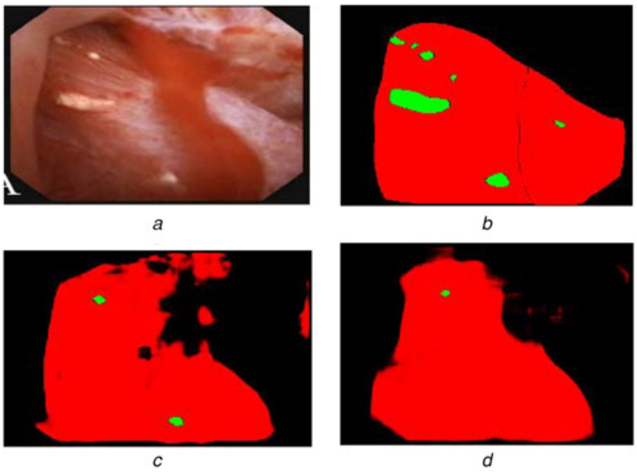

Fig. 3 *Unsuccessful segmentation*
a Input given to the deep learning algorithms after training to compare performance
b Ground truth of Fig. 3*a*
c Prediction made by the deep learning algorithm with the dropout rate of 0 based on Fig. 3*a*
d Prediction made by the deep learning algorithm with the dropout rate of 0.4 based on Fig. 3*a*

The papilla segmentation bleeds beyond the border of the papilla in the non-dropout model. In both cases, two plaques were identified. In Fig. 3, neither the non-dropout or dropout models were able to delineate the papilla clearly; however, both models appear to be approaching the correct segmentation. The dropout model was able to identify two of the plaques but missed several others. Both models were not able to identify the largest of the plaques.

4. Conclusion: The deep learning algorithm with the dropout rate of 0.4 performed better than the deep learning algorithm without dropout. The initial non-dropout model was able to achieve excellent performance on the training dataset but poor performance on the test dataset. This type of training response is a typical example of an overfitted model. The deep learning algorithm with the dropout of 0.4 exhibited better generalisability. Having a dropout rate reduced overfitting by randomly dropping neurones in the neural network. The generalisability of the model with the higher dropout rate was better because it learnt the defining features of the papilla, plugs, and plaques from the training dataset, rather than just ‘memorising’ individual images and instances of precursor formations.

The problem of class imbalance is significant in this data, as well. Focal loss was specifically engineered to address the class imbalance problem by attempting to account for the likelihood of seeing a class; however, this method did not perform as well as JCE. JCE is widely used in the community.

Early evaluation of the deep learning architecture is critical to the eventual success of the model. There are many established approaches. For this work, several different architectures were studied. Nearly, all of these architectures performed poorly at this task, except the U-Net architecture. This reinforces the general acceptance of the U-Net model for semantic segmentation.

Although it was determined that one particular implementation was better than the others (i.e. U-Net/ResNet32 with dropout and a JCE loss function), there are ample opportunities to improve the model. One such method would be through data augmentation. Our dataset of 200+ cases is the largest dataset of segmented video endoscopic data of the kidney; however, there is great variability in the images and presentation of the papilla. Data augmentation involves various techniques such as flips, translations, or rotations to alter the current data to artificially increase the amount of usable data.

Gamma correction or colour correction could also be used as a form of data augmentation to normalise the variation in illumination within the dataset [28]. Gamma correction is useful for datasets, where the data collection conditions are not always consistent such as if there are various cameras used, different perspectives of the object obtained, and variation in the amount of light or luminance of image capture. Data augmentation could lead to an increase in accuracy since the deep learning algorithm could learn more with increased exposure to data by allowing the model to learn more about features that define the papilla, plaque, and plug, rather than just overfitting to the training dataset [29]. However, it is important to proceed with caution as the augmented data may contribute to the overfitting or may result in training to unimportant features. This is one of the reasons that a pre-trained encoder was selected; by using a pre-trained model, some generic features will already be accounted for in the model.

Another method would be to decrease the amount of irrelevant data. The original segmentation data used to train the deep learning algorithm was created over 9 years, with multiple month gaps in between segmentations. Owing to the nature of data creation, there was probably human variability introduced to the segmentation data. Also, the accuracy of the manually segmented papilla, plaque, and plug object maps increased over the years. This might implicate that some of the older data may not represent what would be the accepted segmentation if reassessed or manual segmentation was repeated. These circumstances may have decreased the veracity of the object maps and may have negatively impacted the performance of the deep learning algorithms.

The deep learning algorithm could also be improved by exhaustive hyper-parameter tuning. Hyperopt (https://github.com/hyperopt/hyperopt) is one such package for hyper-parameter optimisation. Although dropout rate was studied in this Letter, the dropout rate could be increased to a higher rate to address overfitting further and increase generalisability. The algorithm could have also been improved by using a different encoder. As with other approaches to optimised model development, hyper-parameter optimisation should be addressed cautiously since it can lead to bias and overfitting as well [30].

Deep learning is a powerful technique for achieving automated image segmentation. Video endoscopy of the kidney is an application, which should be well-suited for deep learning because several of the deep learning architectures were originally designed and validated based on colour camera images. Unfortunately, deep learning requires large amounts of diverse but representative data to adequately train the model. If insufficient data is available, then the model underperforms and/or becomes overtrained. Owing to the optimistic findings of this Letter, future studies will focus on data augmentation to develop a sufficient database of images for training.

5. Acknowledgments: This work was supported by the National Institute of Health (NIH) (nuSURF-NIH Grant no. R25-DK101405) and the Mayo Clinic O'Brien Urology Research Centre (P50 DK083007).

6 References

[1] Mikawlrawng K., Kumar S., Vandana R.: 'Current scenario of urolithiasis and the use of medicinal plants as antiurolithiatic agents in Manipur (North East India): a review', *Int. J. Herb. Med.*, 2014, **2**, (1), pp. 1–12

[2] El-Zoghby Z.M., Lieske J.C., Foley R.N., ET AL.: 'Urolithiasis and the risk of ESRD', *Clin. J. Am. Soc. Nephrol.*, 2012, **7**, (9), pp. 1409–1415

[3] Cheungpasitporn W., Thongprayoon C., O'corragain O.A., ET AL.: 'The risk of kidney cancer in patients with kidney stones: a systematic review and meta-analysis', *QJM: Int. J. Med.*, 2014, **108**, (3), pp. 205–212

[4] Romero V., Akpinar H., Assimos D.G.: 'Kidney stones: a global picture of prevalence, incidence, and associated risk factors', *Rev. Urol.*, 2010, **12**, (2–3), pp. e86–e96

[5] Rule A.D., Lieske J.C., Li X., ET AL.: 'The ROKS nomogram for predicting a second symptomatic stone episode', *J. Am. Soc. Nephrol.*, 2014, **25**, (12), pp. 2878–2886

[6] Han H., Segal A.M., Seifter J.L., ET AL.: 'Nutritional management of kidney stones (nephrolithiasis)', *Clin. Nutr. Res.*, 2015, **4**, pp. 137, doi: 10.7762/cnr.2015.4.3.137

[7] Gul Z., Monga M.: 'Medical and dietary therapy for kidney stone prevention', *Korean J. Urol.*, 2014, **55**, pp. 775, doi:10.4111/kju.2014.55.12.775

[8] Assadi F., Moghtaderi M.: 'Preventive kidney stones: continue medical education', *Int. J. Prev. Med.*, 2017, **8**, p. 67, doi:10.4103/ijpvm.IJPVM_17_17

[9] Assimos D., Krambeck A., Miller N.L., ET AL.: 'Surgical management of stones: American urological association/endourological society guideline, Part I', *J. Urol.*, 2016, **196**, (4), pp. 1153–1160

[10] Coe F.L., Worcester E.M., Evan A.P.: 'Idiopathic hypercalciuria and formation of calcium renal stones', *Nat. Rev. Nephrol.*, 2016, **12**, (9), pp. 519–533

[11] Bird V.Y., Khan S.R.: 'How do stones form? Is unification of theories on stone formation possible?', *Arch. Esp. Urol.*, 2018

[12] Williams J.C., Borofsky M.S., Bledsoe S.B., ET AL.: 'Papillary ductal plugging is a mechanism for early stone retention in brushite stone disease', *J. Urol.*, 2018, **199**, (1), pp. 186–192

[13] Cohen A.J., Borofsky M.S., Anderson B.B., ET AL.: 'Endoscopic evidence that Randall's plaque is associated with surface erosion of the renal papilla', *J. Endourol.*, 2017, **31**, (1), pp. 85–90

[14] Borofsky M.S., Williams J.C., Dauw C.A., ET AL.: 'Association between Randall's plaque stone anchors and renal papillary pits', *J. Endourol.*, 2019, **33**, (4), pp. 337–342

[15] Borofsky M.S., Dauw C.A., Cohen A., ET AL.: 'Integration and utilization of modern technologies in nephrolithiasis research', *Nat. Rev. Urol.*, 2016, **13**, (9), pp. 549–557

[16] Krambeck A.E., Lieske J.C., Li X., ET AL.: 'Current computed tomography techniques can detect duct of Bellini plugging but not Randall's plaques', *Urology*, 2013, **82**, (2), pp. 301–306

[17] Linnes M.P., Krambeck A.E., Cornell L., ET AL.: 'Phenotypic characterization of kidney stone formers by endoscopic and histological quantification of intrarenal calcification', *Kidney Int.*, 2013, **84**, (4), pp. 818–825

[18] Pless M.S., Williams J.C., Andreassen K.H., ET AL.: 'Endoscopic observations as a tool to define underlying pathology in kidney stone formers', *World J. Urol.*, 2019, **37**, pp. 2207–2215

[19] Höfener H.: 'Deep learning nuclei detection: a simply approach can deliver state-of-the-art results', *Comput. Med. Imaging Graph.*, 2018, **70**, pp. 43–52

[20] Kermi A.: 'Deep convolutional neural networks using U-Net for automatic brain tumor segmentation in multimodal MRI volumes'. Int. MICCAI Brainlesion Workshop, 2018, vol. **32**, pp. 37–48

[21] Loussaief S., Abdelkrim A.: 'Deep learning vs. bag of features in machine learning for image classification'. Int. Conf. Advanced Systems Electrical Technologies, March 2018

[22] Clemente C.J., Jaafar F., Malik Y.: 'Is predicting software security bugs using deep learning better than the traditional machine learning algorithms?'. IEEE Int. Conf. QRS, July 2018

[23] Ronneberger O., Fischer P., Brox T.: 'U-Net: convolutional networks for biomedical image segmentation'. Medical Image Computing and Computer-Assisted Intervention Conf., May 2015

[24] He K., Zhang X., Ren S., ET AL.: 'Deep residual learning for image recognition', arXiv:1512.03385, 2015

[25] Kazemi Y., Mirroshandel S.A.: 'A novel method for predicting kidney stone type using ensemble learning', *Artif. Intell. Med.*, 2018, **84**, pp. 117–126

[26] Längkvist M., Jendeberg J., Thunberg P., ET AL.: 'Computer aided detection of ureteral stones in thin slice computed tomography volumes using convolutional neural networks', *Comput. Biol. Med.*, 2018, **97** pp. 153–160

[27] Han S., Hwang S.I., Lee H.J.: 'The classification of renal cancer in 3-phase CT images using a deep learning method', *J. Digit. Imaging*, 2019, **32**, pp. 638–643

[28] Bianco S., Cusano C., Napoletano P., ET AL.: 'Improving CNN-based texture classification by color balancing', *J. Imaging*, 2017, **3**, pp. 33

[29] Gaonkar B., Bui A., Brown M., ET AL.: 'Extreme augmentation: Can deep learning based medical image segmentation be trained using a single manually delineated scan?', 2018

[30] Skocik M., Collins J., Callahan-Flintoft C., ET AL.: 'I tried a bunch of things: the dangers of unexpected overfitting in classification', bioRxiv 078816, doi: https://doi.org/10.1101/078816