STREAMLINE.
OPTIMIZE.
TRUST.

ADAPTIVO | LINACVIEW | DOSEVIEW | PIPSPRO | QA PILOT | IMSURE | STRUCTSURE AI QA

COMPLETE INTEGRATED QA

STANDARD IMAGING®

CLICK HERE TO LEARN MORE

A deep learning system for automated kidney stone detection and volumetric segmentation on non-contrast CT scans

Daniel C. Elton[1], Evrim B. Turkbey[1], Perry J. Pickhardt[2], Ronald M. Summers[1]

1. Radiology and Imaging Sciences, National Institutes of Health Clinical Center, Bethesda, MD 20892-1182, USA

2. School of Medicine and Public Health, University of Wisconsin, Madison, WI 53726, USA

email: Daniel C. Elton (delton@mgh.harvard.edu) or Ronald M. Summers (rms@nih.gov)

Abstract

Purpose: Early detection and size quantification of renal calculi are important for optimizing treatment and preventing severe kidney stone disease. Prior work has shown that volumetric measurements of kidney stones are more informative and reproducible than linear measurements. Deep learning based systems that use abdominal non-contrast CT scans may assist in detection and reduce workload by removing the need for manual stone volume measurement. Prior to this work no such system had been developed for use on noisy low-dose CT or tested on a large-scale external dataset.

Methods: We used a dataset of 91 CT colonography (CTC) scans with manually marked kidney stones combined with 89 CTC scans without kidney stones. To compare with a prior work half the data was used for training and half for testing. A set of CTC scans from 6,185 patients from a separate institution with patient-level labels were used as an external validation set. A 3D U-Net model was employed to segment the kidneys, followed by gradient-based anisotropic denoising, thresholding, and region growing. A 13 layer convolutional neural network classifier was then applied to distinguish kidney stones from false positive regions.

Results: The system achieved a sensitivity of 0.86 at 0.5 false positives per scan on a challenging test set of low-dose CT with many small stones, an improvement over an earlier work which obtained a sensitivity of 0.52. The stone volume measurements correlated well with manual measurements ($r^2 = 0.95$). For patient level classification the system achieved an area under the receiver operating characteristic (AU-ROC) of 0.95 on an external validation set (sensitivity = 0.88, specificity = 0.91 at the Youden point). A common cause of false positives were small atherosclerotic plaques in the renal sinus that simulated kidney stones.

Conclusions: Our deep learning based system showed improvements over a previously developed system that did not use deep learning, with even higher performance on an external validation set.

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/mp.15518

I. Introduction

People living in the United States have approximately a 9% lifetime risk of developing kidney stone disease, also known as urolithiasis or nephrolithiasis[1]. While X-ray, ultrasound, and CT may all be used to detect kidney stones, CT is by far the most common imaging modality employed owing to its high sensitivity and specificity (reported at 96% or higher)[2]. Emergency room visits due to kidney stones are very common[3] and nearly doubled from 1992 to 2009, when the number of visits reached around 1,000,000 per year in the United States[4]. Amongst patients who received a diagnosis of kidney stones after visiting the emergency department between 2007-2009, 71% had received a CT scan[4]The American College of Radiology's Data Science Institute has identified "Kidney stone detection on CT" as an important AI use-case[5]. AI systems may have particular utility when it comes to detecting kidney stones on scans performed for other reasons, such as CT colonography[6]. A 2010 study found that 8% of patients who underwent CT colonography had asymptomatic urolithiasis[6]. Ultra lose dose CT (ULD-CT) protocols have also been developed specifically for kidney stone screening and surveillance[7,8,9] and ULD-CT can be lower cost than renal ultrasound[8].

A 2005 study estimated that the total direct and indirect costs of nephrolithiasis was approximately $4.5 billion in 2000 in the United States and growing each year[10]. There is not much work on the cost effectiveness of early stage interventions, however Saigal et al estimate that a 75% effective intervention that that costs less than $300 per patient per year would be cost effective at reducing health care expenditures[10]. They note that a shift away from expensive medications towards low cost treatment modalities such as increased water intake and lemon juice[11] could increase cost effectiveness[10]. More work needs to be done but based on the work on screening cost-effectiveness reported so far[10,12] it is not unreasonable to conclude that earlier detection of kidney stones using CT, especially when done opportunistically, could reduce healthcare expenditures.

Accurate measurement of stone volume, which can only be done via CT, is important to determine the optimal course of treatment.[8] Stones which are small enough have a high likelihood of passing spontaneously on their own and therefore may not warrant costly treatment expenditures[13,14]. Selby et al found that stone volume but not stone diameter was a predictor of future symptomatic events (HR 1.35 per quartile)[15]. Statistical correlations

have been also been found between stone volume and the chance of spontaneous passage[13,14] and the risk of perioperative complications during percutaneous nephrolithotomy[16]. The rate at which stones grow over time may also inform treatment decisions as stones with a higher growth rate may be judged at greater risk of becoming symptomatic.

Detailed protocols for low-dose CT scanning for kidney stones have been described by Planz et al. where they argued these should be preferred over X-ray due to allow for volumetry.[8] While it was recently shown that low-dose high-noise scans do not significantly human diagnostic accuracy[17], no machine learning based system has yet been tested with ultra low-dose scans.

Despite the relevance of stone volume to making treatment decisions, accurate measurements of stone size and/or volume are not always performed owing to the added time required to make such measurements, creating a value proposition for automated measurement. A 2005 survey in the UK found that 15% of radiologists "guestimate" stone size rather than making a digital measurement[18]. It has been shown that stone volume measurement is a more reproducible measurement than stone diameter[8,19,20]. For instance, one study found an average of 26.3% inter-reader variation in stone diameter vs close to 0% inter-reader variation when using a semi-automated method for stone volume measurement[20]. One study found variation between radiologists depending on the windowing settings utilized while making the measurement, with a soft tissue window setting leading to overestimation of stone volume by an average of 57%[21]. It was been argued that stone volume provides a more objective means for stone surveillance[19]. Homayounieh et al. (2021) have demonstrated that stone radiomics features (most notably stone volume) are highly predictive of future hydronephrosis, future stone burden, and invasive treatment.[22]

Relatively few works have been published which tackle the challenge of computer-aided detection of kidney stones in CT. Lee et al. (2009) used texture and intensity based features to train an artificial neural network to distinguish kidney stones from vascular calcifications[23]. Liu et al. (2014) segmented the kidneys and then used total-variation flow denoising followed by the maximal stable extremal regions method to segment stones[24]. Features from the segmented stones were fed into an support vector machine (SVM) classifier to classify kidney stones vs false positives, achieving a sensitivity of 60.0% at an average of 2 false positives per scan[24]. Längkvist et al. (2018) performed thresholding at 250 HU and connected components

analysis followed by application of a convolutional neural network (CNN) to classify ureteral stones vs false positive detections[25]. A limitation of this work is that their training data only consisted of cases with very large and bright uretral stones and no stone free cases were used in either training or testing. The system achieved a 100% sensitivity but at an unacceptably high false positive rate of 2.7 false positives per scan. Parakh et al. (2019) developed a dual CNN system for detecting stones in the kidney, ureter, and bladder[26]. The system achieved a sensitivity of 0.873 and AUC of 0.954 for patient level stone detection[26]. A major limitation of their work is that their system does not report the number of stones detected or segment the stones to measure the stone volume. Most recently Cui et al. (2020) have utilized a dual stage 3D U-Net followed by simple thresholding and region growing to segment kidney stones. The first 3D U-Net segments the kidneys to allow a cropped box around each kidney to be generated. The cropped box is fed into a 2nd 3D U-Net which segments the renal sinus area, where most kidney stones are located. The system achieved a sensitivity of 96% with a false positive rate of 0.03 per patient for detecting stones $>$ 2mm in diameter (volume $>$ 4.18 mm^3)[27]. A major weakness of their system is their threshold based detection system which leads to many false positives on low dose CT scans.

To summarize or survey of existing work, no fully automated system has been demonstrated that obeys these necessary desiderata of providing fully automated and accurate stone detection and segmentation and the ability to work with noisy scans and small stones. Detecting small stones is of particular importance given that early treatment greatly reduces healthcare expenditures and patient's quality of life. The importance of stone volume measurement (as opposed to linear measurement) has been neglected in the recent works on automated kidney stone detection, even though it has been shown to be highly correlated with patient outcomes. Another major deficiency of the works reported so far is that none of the works test their system on an external dataset[25,26,27]. In this work we seek to meet what is actually needed for clinical use - a robust system validated on a large dataset capable of detecting and segmenting small stones on both noisy CT (the type of CT often used for kidney stone screening) and less noisy CT taken for CT colonography (for opportunistic detection).

II. Dataset preparation

The dataset used for initial training and testing is the same used by Liu et al.[24] (2014). The dataset, which we designate as NNMC-CTC, is a subset of a larger dataset of 1186 CT Colonography (CTC) scans from three institutions[28]. The images were selected as images containing kidney stones based on the extracolonic findings information, and the presence of stones was verified as described in[24]. Each patient was administered an oral contrast agent and scanned during a single breath hold using a four- or eight-channel CT scanner (General Electric Light Speed or Light Speed Ultra). CT scanning parameters included in-plane resolution of 0.55 - 0.75 mm, 1.25 – 2.5 mm section collimation, 1 mm reconstruction interval, and 120 kVp. Many of these images have a large amount of quantum noise. In prospective readings by experienced radiologists, 91 patients were reported to have renal calculi in the extracolonic findings. A radiologist with 6 years of experience with kidney stone detection (E.B.T) marked the coordinates for all the stones in these 91 cases and their findings were double checked by a second radiologist with 20 years of experience with kidney stone detection (R.M.S.). The renal calculi volumes were measured using a commercially available coronary artery calcium scoring tool (Vitrea Core fX v6, Vital Images, Minnetonka, MN). The settings employed for volume measurement were a lower threshold of 130 HU and a lower pixel threshold of 3 pixels, as recommended by Patel et al. to assess the volumes of calculi on non-contrast CT images[20]. In our literature survey we found that a lower threshold of 130 HU is the most common choice for measuring stone volume[15,24,29], with a minority of studies using either 200 HU[14] or 250 HU[25,30].

In addition to the scans of 91 patients with kidney stones, 89 patients without kidney stones were chosen from the remaining CTC images as negative examples. The dataset of 180 images was split evenly into training ($N = 90$) and test ($N = 90$) sets. The splitting was done to replicate the splitting used previously by Liu et al. as closely as possible[24]. There were 97 kidney stones in the training dataset and 77 in the test dataset. The distribution of stone sizes in the test dataset is shown in figure 1. The average stone size was 44.69 mm^3 and the range of stone sizes was 1 - 433 mm^3.

To validate our system on a newer dataset we used a set of 12,351 CTC images from the University of Wisconsin Medical Center.[31] The scans were helical CT (General Electric Discovery Series) taken with in-plane resolution of 0.45 - 0.75 mm, 0.75 or 1 mm recon-

struction interval, 1.25 mm slice thickness, and 120 kVp (a few scans had slice thicknesses of 1.5 or 3 mm). One scan was removed due to data corruption, 3 scans were removed for being decubitus or prone rather than supine, and 6,166 scans were removed due to lack of extracolonic findings information, leaving 6,185 scans. To isolate the scans with kidney stones we searched the extracolonic findings notes for at least one of the following keywords: "kidney stone", "kid stone", "nephrolithiasis", "renal stone", "renal calculi", "renal calculus", and "renal calc". Among the patients with extracolonic findings information, 841 had E-RADS scores reported for at least one extracolonic finding of a kidney stone. For patients having multiple scans with extracolonic findings, we took the first scan and ignored the rest. Altogether we found 755 patients (12%) with keywords indicating a kidney stone and 5381 patients assumed to be without stones. Some extracolonic findings were categorized based on the criteria laid out in the CT Colonography Reporting and Data System (C-RADS)[32]. In particular, categories E3 ("Likely Unimportant Finding, Incompletely Categorized") and E4 ("Potentially Important Finding") were of interest[7,32]. Among those, there were 14 kidney stones with Category E3 and 7 kidney stones with category E4. Retrospective analysis of all images employed was approved by the local Institutional Review Board and the informed consent requirement was waived.

III. Methods

Conceptually the system is similar to the one developed by Liu et al. in 2014[24] but with several important updates and simplifications. A diagram comparing both systems is shown in figure 2. To segment the kidneys we use a 3D U-Net[33] model developed previously[34] which was trained on an in-house dataset of 56 cases with ground truth segmentations. Details of the 3D U-Net architecture and training procedure can be found in previous work[34,35]. A connected components analysis is applied to the 3D U-Net segmentation to isolate the two largest objects which helps remove spurious segmentations outside the kidneys. We use the original scan thickness (1.0 mm for the NNMC-CTC dataset and 1.0-1.25 mm for the UW-CTC dataset, even though resampling to thicker slices would reduce image noise. This decision was informed by a number of studies from 2000 and onward which have shown that thicker slices lead to less accurate size measurements due to the partial volume effect[29,36,37,38]. In studies with ground truth stones embedded in phantoms, larger slice thicknesses lead to a

decrease in both the measured size and maximum intensity of stones[36,37]. Additionally, slice thicknesses greater than 3 mm can lead to small stones (<3 mm diameter) being missed[37,39]. On the basis of these findings, Kambadakone et al. recommend the use of thin slices (0.75 - 1.5 mm) for accurate kidney stone detection and size quantification[40]. To keep the radiation dose low with thinner slices, the tube current must be lowered, resulting in a noisier scan. However, as noted in the introduction, this does not hinder diagnosit accuracy[17].

To denoise the images two methods were tested head-to-head on a few training data cases - *denoise_tv_chambolle()* as implemented in the Python library Scikit-Image and *CurvatureAnisotropicDiffusionImageFilter()* as implemented in ITK. For further elaboration on the pros and cons of different denoising methods, see our prior work[24]. Denoising is performed iteratively until the number of connected components is less than 200 (this lowers the computational burden at later steps as the CNN must be applied to each connected component). Then threshold (130 HU) is set on the denoised image, followed by connected components analysis. Next, region growing of each connected component is performed on the original CT with a lower threshold of 130 HU to replicate the manual method for segmentation. We hypothesize that this low threshold helps compensate slightly for the effects of partial volume averaging and better replicates the recommended manual technique[41].

After region growing, any regions that are touching are joined together and a 24x24x24 voxel box centered around each region on the original CT is fed into a 13 layer CNN to classify kidney stones vs false positives. The CNN architecture is the same architecture developed by Perez et al. for lung nodule classification[42] and features batch normalization and dropout (dropout rate = 0.65) in each layer. The source code for the architecture is available online[43]. The CNN was trained on a set of boxes generated by running the first stage on the training scans. The boxes are reprocessed by clipping to -200 - 1000 HU and rescaling so the distribution of intensities is centered at zero with standard deviation of ± 1. The CNN was trained using data augmentation (random xyz jitter, random rotation, random flipping), the rectified-Adam optimizer[44], and a batch size of 8. Since false positives greatly outnumbered true positives, the true positives were reweighted in the training sample so there was a 50-50 mix during training. A validation set of 400 boxes was used to monitor the F1 score during training, and the training was stopped as soon as the validation F1 score plateaued. To ensure reproducibility, we have published the Python source code for our method on Github at https://github.com/rsummers11/CADLab.

IV. Results

Some initial hyperparameter studies and ablation studies were performed by hand on a few training cases. For instance, we tested three different thresholds for the max number of detections finding that it only made a difference in a few very noisy cases. Validation F1 score for several CNN architectures are shown in figure 3. Among the architectures studied, the CNN with box size 24x24x24 performed best in the validation set, so that architecture was used. We also tested several denoising methods. First we tested a variation flow denoising method very similar to the one utilized by Liu et al., as implemented in the *scikit-image* Python package (*denoise_tv_chambolle()*). We tested both 3D and 2D implementations, finding them to give very similar results visually on a few training cases. We found the method enhanced noise in a few situations and also artificially decreased the intensity of stones, often below the 130 HU threshold. We found that the anisotropic diffusion filtering method developed by Perona & Malik (1990)[45], as implemented in ITK (*GradientAnisotropicDiffusionImageFilter()*), did not enhance noise and only reduced the maximal stone brightness slightly, so we used that method instead. Thresholds of 100, 130, and 150 HU were tested. The differences were small but the 130 HU threshold yielded the best FROC curve in the test set.

The FROC curve and precision-recall curve on the test set is shown in figure 4. FROC curves vs stone volume size are shown in figure 5. Not surprisingly, large stones (ie volumes > 27 mm^3 / diameters > 3.7 mm) are much easier to detect, with a sensitivity of 0.91 at a false positive rate of < 0.05 per scan. A scatterplot comparing automated vs manual stone volume measurements is shown in figure 5 (Pearson's $r^2 = 0.95$). The relative average volume difference, defined as the average of $R = (V_{\text{predicted}} - V_{\text{true}})/V_{\text{true}}$, was 0.31 ± 0.92. This is an improvement over Liu et al., who obtained an average R of 1.15 ± 1.27[24].

The per-patient ROC curve for the CTC validation set is shown in figure 6. 6,163 patients had extracolonic findings in at least one scan. For patient-level detection the system achieved an AUC of 0.95 with a sensitivity of 0.88 and specificity of 0.91 at the Youden point. At the threshold corresponding to the Youden point, the detector found 6/7 of the E4 stones (85%) and 11/14 (78%) of the E3 stones. The single E4 stone that was missed was an improper label (image had a uretral stone, not a kidney stone). Using the sensitivity for the E3 and E4 stones specifically and the specificity for the detector overall we find AUC of 0.86

for the E3 stones and 0.91 for the E4 stones. These results make sense given that E4 are more severe stones.

We performed an analysis of a randomly drawn sample of false positive and false negative detections in the NNMC-CTC test set, obtained using a CNN classification threshold of 0.5 (see random examples shown in fig. 7). Out of 15 false positives, 9 (50%) were due to plaque in the renal sinus, 1 was due to a beam hardening artifact from oral contrast, 2 were due to image noise, 2 were due to likely missing labels, and 1 was due to a metal object near the kidneys. The two that were due to missing labels appeared to be very small stones and possibly uric acid stones. Out of 5 false negatives that were reviewed, they all were kidney stones of distinct size (eg $>$ 4-5 voxels) but less bright (ie max HU $<$ 400). These stones were detected in the first stages of the detector but misclassified by the CNN. It appears that it is difficult for the CNN to distinguish plaques in the renal arteries from kidney stones and that this issue was responsible for the majority of both false positives and false negatives. We also looked at a few of the false negatives that were responsible specifically for the plateau in sensitivity at 0.92 observed in the FROC and precision-recall curves (fig. 4). The plateau is largely due to very a few small stones which are lost in the denoising stage and therefore are never fed into the CNN.

We also conducted an analysis of false positives and false negatives in the UW-CTC dataset, again using a threshold of 0.5, which corresponded to an operating point with a sensitivity of 0.95 and a false positive rate of 0.16 (see figure 9). Out of 15 false positives surveyed, 7/15 (47%) were actual kidney stones of varying size, 5/15 (33%) were due to image noise, one was due to a rib bordering the kidney that was mistaken for a kidney stone, one was due to beam hardening from oral contrast, and one was due to a large calcified tumor. Out of 10 false negatives that we looked at, 9/10 were in very noisy images where we determined the problem was with the stones being lost during the denoising iterations. The 10th false negative case was the missed E4 case mentioned above - it was a large ureteral stone causing complete blockage of the left renal collecting system. The uretral stone was improperly classified in the extracolonic findings as a kidney stone.

V. Discussion

We have developed a fully automated system for kidney stone detection and volume quantification on CT. We adopted a framework similar to the one previously developed in our lab in 2014[24], but with important updates to the algorithms used in each step and a few simplifications. The system achieved an AUC of 0.95 for patient-level classification (fig. 6) on a large external validation set with 6,185 scans, with many ($\approx$ 50%) of the false positives corresponding to stones not reported in the extracolonic findings information. In the NNMC-CTC test set the improvement obtained over the 2014 work was substantial both in terms of sensitivity (0.86 vs 0.52 a false positive rate of 0.5/scan) and accuracy of volume measurement (fig 4). A comparison of the failure modes of the two systems indicate this improvement is mainly due to the use of a CNN rather than a feature-based SVM for false positive detection. This replicates a wealth of other studies which show that CNN based classifiers can perform much better than classifiers which use hand-crafted features.

Systems such as the one developed and validated in this work may offer clinical utility, particularly by providing an automated measure of stone volume. Stone volume measurements are time consuming to obtain manually but studies show volume is more reproducible than measures such as stone diameter[20].

The current system could likely be simplified further without loss in accuracy. For instance, the denoising algorithm could be removed and a patch-based 3D U-Net could be trained to segment kidney stones directly within the kidney region. Such an approach would be similar to the patch-based 3D U-Net that was recently demonstrated for segmenting small aortic plaques[46].

Two challenges encountered in this work were dealing with large amounts of image noise and distinguishing plaques from kidney stones. The issue of plaques causing false positives and false negatives has also been noted in another recent work[27]. Solving this problem was outside the scope of this work but might be tackled by assembling a joint training dataset to train a multiclass deep learning system to detect and segment both plaque and kidney stones. By training directly on the task of discriminating the two types of objects, such a system would likely have a lower chance of mistaking plaque for kidney stones or vice-versa.

In conclusion, in this work we showed that deep learning algorithm can detect renal

calculi in non-contrast CT scans with high sensitivity and specificity, including on high noise low dose scans. The system was demonstrated to generalize to both a hold-out test set and a large external dataset. Systems such as this which can detect kidney stones and provide accurate volume measurements on a wide range of CT scans may have substantial clinical utility.

Acknowledgements

This research was funded in part by the Intramural Research Program of the National Institutes of Health, Clinical Center. The research used the high-performance computing facilities of the NIH Biowulf cluster. The authors thank Drs. Perry Pickhardt, J. Richard Choi, and William Schindler for providing CT colonography scans.

Conflicts of interest

Potential financial interest: Author RMS receives royalties from iCAD, Philips, ScanMed, PingAn, and Translation Holdings and has received research support from Ping An (CRADA) and NVIDIA (GPU card donations). PJP is an adviser or consultant for Zebra Medical Vision and Bracco Diagnostics, and shareholder in Cellectar, Elucent, and SHINE.

Data Availability Statement

The data that support the findings of this study are available from corresponding author Dr. Ronald M. Summers upon reasonable request.

References

[1] C. D. Scales, A. C. Smith, J. M. Hanley, and C. S. Saigal, Prevalence of Kidney Stones in the United States, European Urology **62**, 160–165 (2012).

[2] I. Boulay, P. Holtz, W. D. Foley, B. White, and F. P. Begun, Ureteral calculi: diagnostic efficacy of helical CT and implications for treatment of patients., American Journal of Roentgenology **172**, 1485–1490 (1999).

[3] C. D. Scales, L. Lin, C. S. Saigal, C. J. Bennett, N. A. Ponce, C. M. Mangione, and M. S. L. and, Emergency Department Revisits for Patients with Kidney Stones in California, Academic Emergency Medicine **22**, 468–474 (2015).

[4] C.-W. Fwu, P. W. Eggers, P. L. Kimmel, J. W. Kusek, and Z. Kirkali, Emergency department visits, use of imaging, and drugs for urolithiasis have increased in the United States, Kidney International **83**, 479–486 (2013).

[5] L. B. Adair II, Kidney Stone Detection on CT, https://www.acrdsi.org/DSI-Services/Define-AI/Use-Cases/Kidney-Stone-Detection-on-CT (2020).

[6] C. J. Boyce, P. J. Pickhardt, E. M. Lawrence, D. H. Kim, and R. J. Bruce, Prevalence of Urolithiasis in Asymptomatic Adults: Objective Determination Using Low Dose Noncontrast Computerized Tomography, Journal of Urology **183**, 1017–1021 (2010).

[7] B. D. Pooler, M. G. Lubner, D. H. Kim, E. M. Ryckman, S. Sivalingam, J. Tang, S. Y. Nakada, G.-H. Chen, and P. J. Pickhardt, Prospective Trial of the Detection of Urolithiasis on Ultralow Dose (Sub mSv) Noncontrast Computerized Tomography: Direct Comparison against Routine Low Dose Reference Standard, Journal of Urology **192**, 1433–1439 (2014).

[8] V. B. Planz, N. M. Posielski, M. G. Lubner, K. Li, G.-H. Chen, S. Y. Nakada, and P. J. Pickhardt, Ultra-low-dose limited renal CT for volumetric stone surveillance: advantages over standard unenhanced CT, Abdominal Radiology **44**, 227–233 (2018).

[9] C. L. Moore, M. Bhargavan-Chatfield, M. M. Shaw, K. Weisenthal, and M. K. Kalra, Radiation Dose Reduction in Kidney Stone CT: A Randomized, Facility-Based Intervention, Journal of the American College of Radiology (2021).

[10] C. S. Saigal, G. Joyce, A. R. Timilsina, and the Urologic Diseases in America Project, Direct and indirect costs of nephrolithiasis in an employed population: Opportunity for disease management?, Kidney International **68**, 1808–1814 (2005).

[11] M. P. Kurtz and B. H. Eisner, Dietary therapy for patients with hypocitraturic nephrolithiasis, Nature Reviews Urology **8**, 146–152 (2011).

[12] E. S. Hyams and B. R. Matlaga, Economic impact of urinary stones, Translational Andrology and Urology **3** (2014).

[13] D. M. Coll, M. J. Varanelli, and R. C. Smith, Relationship of Spontaneous Passage of Ureteral Calculi to Stone Size and Location as Revealed by Unenhanced Helical CT, American Journal of Roentgenology **178**, 101–103 (2002).

[14] J. Jendeberg, H. Geijer, M. Alshamari, and M. Lidén, Prediction of spontaneous ureteral stone passage: Automated 3D-measurements perform equal to radiologists, and linear measurements equal to volumetric, European Radiology **28**, 2474–2483 (2018).

[15] M. G. Selby, T. J. Vrtiska, A. E. Krambeck, C. H. McCollough, H. E. Elsherbiny, E. J. Bergstralh, J. C. Lieske, and A. D. Rule, Quantification of Asymptomatic Kidney Stone Burden by Computed Tomography for Predicting Future Symptomatic Stone Events, Urology **85**, 45–50 (2015).

[16] S. Lai, B. Jiao, Z. Jiang, J. Liu, S. Seery, X. Chen, B. Jin, X. Ma, M. Liu, and J. Wang, Comparing different kidney stone scoring systems for predicting percutaneous nephrolithotomy outcomes: A multicenter retrospective cohort study, International Journal of Surgery **81**, 55–60 (2020).

[17] R. P. Reimer, J. Salem, M. Merkt, K. Sonnabend, S. Lennartz, D. Zopfs, A. Heidenreich, D. Maintz, S. Haneder, and N. G. Hokamp, Size and volume of kidney stones in computed tomography: Influence of acquisition techniques and image reconstruction parameters, European Journal of Radiology **132**, 109267 (2020).

[18] R. J. Kampa, K. R. Ghani, S. Wahed, U. Patel, and K. M. Anson, Size Matters: A Survey of How Urinary-Tract Stones are Measured in the UK, Journal of Endourology **19**, 856–860 (2005).

[19] S. R. Patel, S. Wells, J. Ruma, S. King, M. G. Lubner, S. Y. Nakada, and P. J. Pickhardt, Automated Volumetric Assessment by Noncontrast Computed Tomography in the Surveillance of Nephrolithiasis, Urology **80**, 27–31 (2012).

[20] S. R. Patel, P. Stanton, N. Zelinski, E. J. Borman, M. A. Pozniak, S. Y. Nakada, and P. J. Pickhardt, Automated Renal Stone Volume Measurement by Noncontrast Computerized Tomography is More Reproducible Than Manual Linear Size Measurement, Journal of Urology **186**, 2275–2279 (2011).

[21] A. Danilovic, B. A. Rocha, G. S. Marchini, O. Traxer, C. Batagello, F. C. Vicentini, F. C. M. Torricelli, M. Srougi, W. C. Nahas, and E. Mazzucchi, Computed tomography window affects kidney stones measurements, International Brazilian Journal of Urology **45**, 948–955 (2019).

[22] F. Homayounieh, R. D. Khera, B. C. Bizzo, S. Ebrahimian, A. Primak, B. Schmidt, S. Saini, and M. K. Kalra, Prediction of burden and management of renal calculi from whole kidney radiomics: a multicenter study, Abdominal Radiology **46**, 2097–2106 (2020).

[23] H. J. Lee, K. G. Kim, S. I. Hwang, S. H. Kim, S.-S. Byun, S. E. Lee, S. K. Hong, J. Y. Cho, and C. G. Seong, Differentiation of Urinary Stone and Vascular Calcifications on Non-contrast CT Images: An Initial Experience using Computer Aided Diagnosis, Journal of Digital Imaging **23**, 268–276 (2009).

[24] J. Liu, S. Wang, E. B. Turkbey, M. G. Linguraru, J. Yao, and R. M. Summers, Computer-aided detection of renal calculi from noncontrast CT images using TV-flow and MSER features, Medical Physics **42**, 144–153 (2014).

[25] M. Längkvist, J. Jendeberg, P. Thunberg, A. Loutfi, and M. Lidén, Computer aided detection of ureteral stones in thin slice computed tomography volumes using Convolutional Neural Networks, Computers in Biology and Medicine **97**, 153–160 (2018).

[26] A. Parakh, H. Lee, J. H. Lee, B. H. Eisner, D. V. Sahani, and S. Do, Urinary Stone Detection on CT Images Using Deep Convolutional Neural Networks: Evaluation of Model Performance and Generalization, Radiology: Artificial Intelligence **1**, e180066 (2019).

[27] Y. Cui, Z. Sun, S. Ma, W. Liu, X. Wang, X. Zhang, and X. Wang, Automatic Detection and Scoring of Kidney Stones on Noncontrast CT Images Using S.T.O.N.E. Nephrolithometry: Combined Deep Learning and Thresholding Methods, Molecular Imaging and Biology (2020).

[28] P. J. Pickhardt, J. R. Choi, I. Hwang, J. A. Butler, M. L. Puckett, H. A. Hildebrandt, R. K. Wong, P. A. Nugent, P. A. Mysliwiec, and W. R. Schindler, Computed Tomographic Virtual Colonoscopy to Screen for Colorectal Neoplasia in Asymptomatic Adults, New England Journal of Medicine **349**, 2191–2200 (2003).

[29] S. Demehri, M. K. Kalra, F. J. Rybicki, M. L. Steigner, M. J. Lang, E. A. Houseman, G. C. Curhan, and S. G. Silverman, Quantification of Urinary Stone Volume: Attenuation Threshold–based CT Method—A Technical Note, Radiology **258**, 915–922 (2011).

[30] J. B. Ziemba, P. Li, R. Gurnani, S. Kawamoto, E. K. Fishman, G. Fung, W. W. Ludwig, D. Stoianovici, and B. R. Matlaga, A User-Friendly Application to Automate CT Renal Stone Measurement, Journal of Endourology **32**, 685–691 (2018).

[31] P. J. Pickhardt, P. M. Graffy, R. Zea, S. J. Lee, J. Liu, V. Sandfort, and R. M. Summers, Automated CT biomarkers for opportunistic prediction of future cardiovascular events and mortality in an asymptomatic screening population: a retrospective cohort study, The Lancet Digital Health **2**, e192–e200 (2020).

[32] M. E. Zalis, M. A. Barish, J. R. Choi, A. H. Dachman, H. M. Fenlon, J. T. Ferrucci, S. N. Glick, A. Laghi, M. Macari, E. G. McFarland, M. M. Morrin, P. J. Pickhardt, J. Soto, and J. Yee, CT Colonography Reporting and Data System: A Consensus Proposal, Radiology **236**, 3–9 (2005).

[33] Özgün Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation, in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, pages 424–432, Springer International Publishing, 2016.

[34] V. Sandfort, K. Yan, P. J. Pickhardt, and R. M. Summers, Data augmentation using generative adversarial networks (CycleGAN) to improve generalizability in CT segmentation tasks, Scientific Reports **9** (2019).

[35] D. Elton, V. Sandfort, P. J. Pickhardt, and R. M. Summers, Accurately identifying vertebral levels in large datasets, in *Medical Imaging 2020: Computer-Aided Diagnosis*, edited by H. K. Hahn and M. A. Mazurowski, SPIE, 2020.

[36] K. C. Saw, J. A. McAteer, A. G. Monga, G. T. Chua, J. E. Lingeman, and J. C. Williams, Helical CT of Urinary Calculi, American Journal of Roentgenology **175**, 329–332 (2000).

[37] E. Ketelslegers and B. E. V. Beers, Urinary calculi: improved detection and characterization with thin-slice multidetector CT, European Radiology **16**, 161–165 (2005).

[38] R. Umbach, J.-K. Müller, G. Wendt-Nordahl, T. Knoll, and J. P. Jessen, In-vitro comparison of different slice thicknesses and kernel settings for measurement of urinary stone size by computed tomography, Urolithiasis **47**, 583–586 (2019).

[39] M. Memarsadeghi, G. Heinz-Peer, T. H. Helbich, C. Schaefer-Prokop, G. Kramer, M. Scharitzer, and M. Prokop, Unenhanced Multi–Detector Row CT in Patients Suspected of Having Urinary Stone Disease: Effect of Section Width on Diagnosis, Radiology **235**, 530–536 (2005).

[40] A. R. Kambadakone, B. H. Eisner, O. A. Catalano, and D. V. Sahani, New and Evolving Concepts in the Imaging and Management of Urolithiasis: Urologists' Perspective, RadioGraphics **30**, 603–623 (2010).

[41] E. W. Olcott, F. G. Sommer, and S. Napel, Accuracy of detection and measurement of renal calculi: in vitro comparison of three-dimensional spiral CT, radiography, and nephrotomography., Radiology **204**, 19–25 (1997).

[42] G. Perez and P. Arbelaez, Automated detection of lung nodules with three-dimensional convolutional neural networks, in *13th International Conference on Medical Information Processing and Analysis*, edited by J. Brieva, J. D. García, N. Lepore, and E. Romero, SPIE, 2017.

[43] LungCancerDiagnosis-pytorch, https://github.com/BCV-Uniandes/LungCancerDiagnosis-pytorch/blob/master/models/model_nod3.py, 2018, Accessed: 2021-08-07.

[44] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han, On the Variance of the Adaptive Learning Rate and Beyond, in *Proceedings of the 8th International Conference on Learning Representations (ICLR)*, 2020.

[45] P. Perona and J. Malik, Scale-space and edge detection using anisotropic diffusion, IEEE Transactions on Pattern Analysis and Machine Intelligence **12**, 629–639 (1990).

[46] R. M. Summers, D. C. Elton, S. Lee, Y. Zhu, J. Liu, M. Bagheri, V. Sandfort, P. C. Grayson, N. N. Mehta, P. A. Pinto, W. M. Linehan, A. A. Perez, P. M. Graffy, S. D. O'Connor, and P. J. Pickhardt, Atherosclerotic Plaque Burden on Abdominal CT: Automated Assessment With Deep Learning on Noncontrast and Contrast-enhanced Scans, Academic Radiology (2020).

Appendix

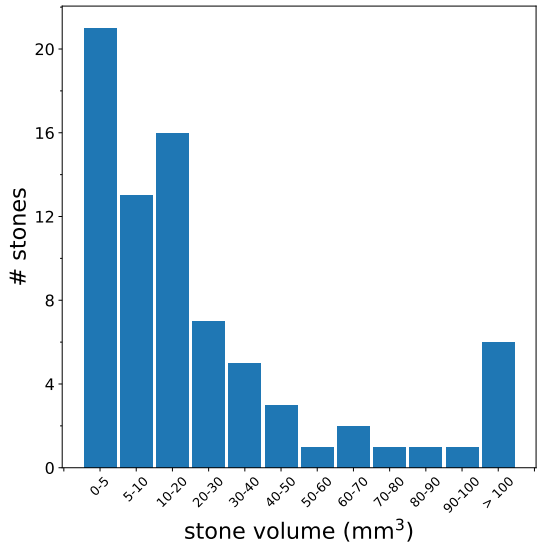

Figure 1: Stone volume distribution for the NNMC-CTC test set. The test set is challenging due to the combination of noisy images and 21 stones with volumes < 5 mm^3.

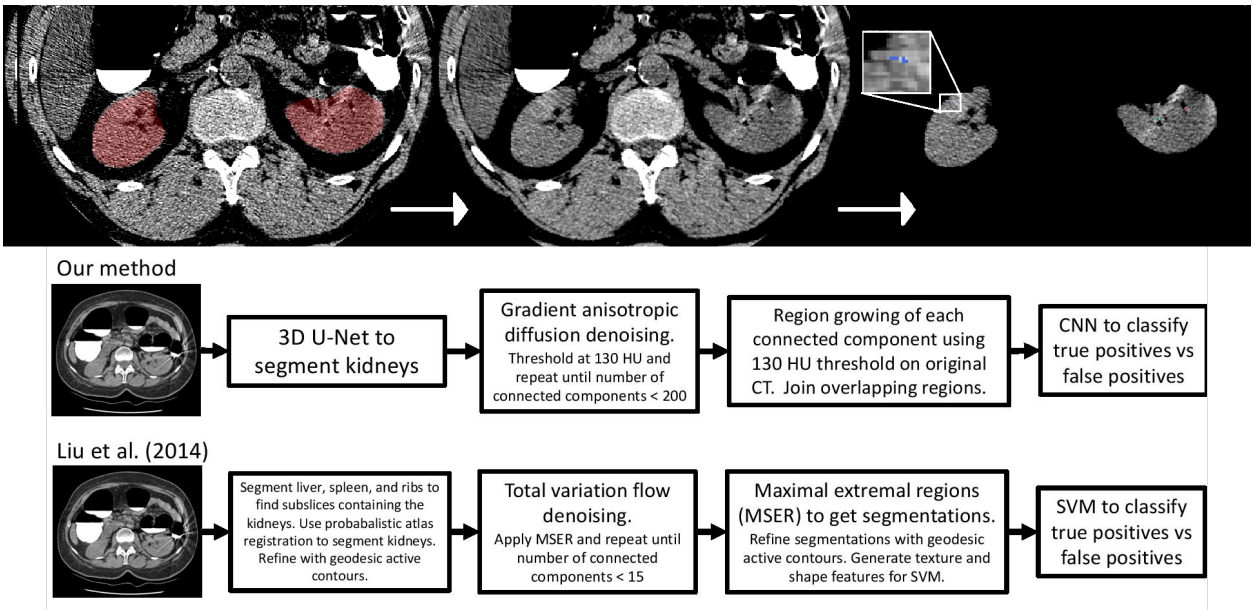

Figure 2: (top) Intermediate outputs from a randomly selected case with kidney stones. On the far left the kidney segmentation and CT is visualized with a window between -50 to 150 HU. The middle panel shows the CT after denoising. The right panel shows the detections. The blue detection (magnified) is classified as a false positive by the CNN whereas the green and red detections are classified as true positives. (bottom) The kidney stone detection system utilized in this work (top) vs the 2014 system previously developed in our lab (below)[24].

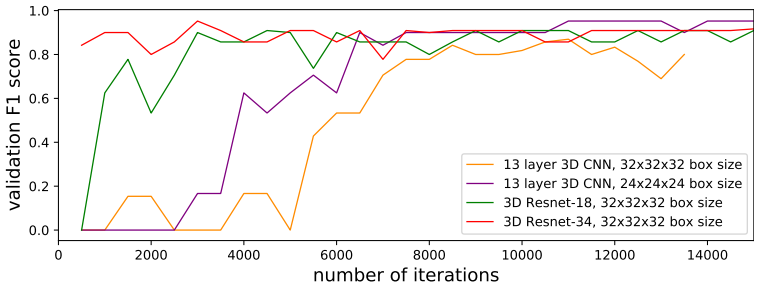

Figure 3: Validation F1 score during training for four models that were tested to classify kidney stones vs false positive detections. The 13 layer CNN with box size 24x24x24 was chosen since it had the highest validation F1 score.

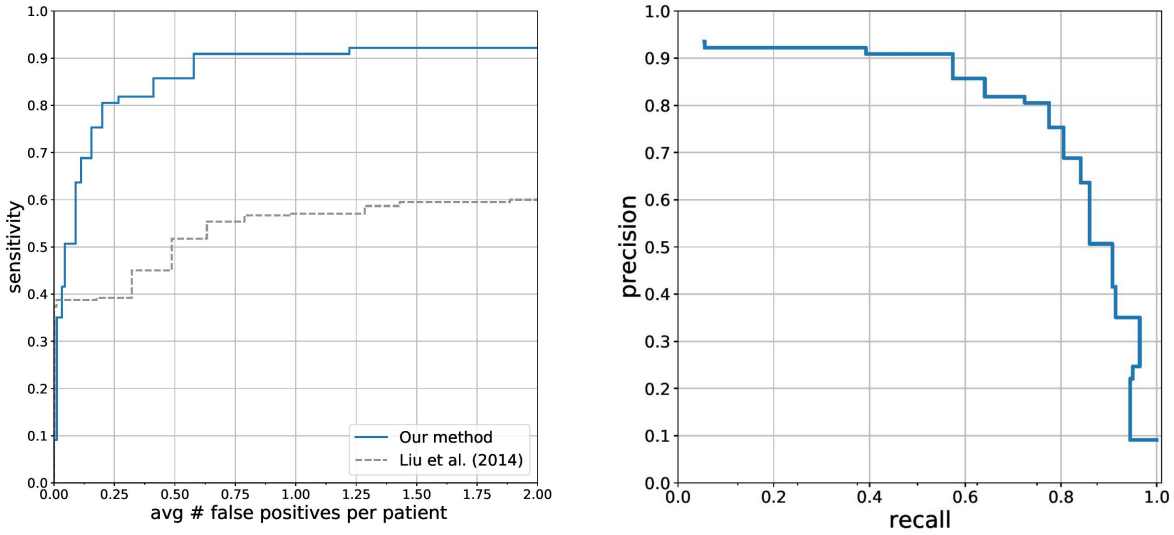

Figure 4: (left) FROC curve on the NNMC-CTC dataset. (right) Precision-recall curve.

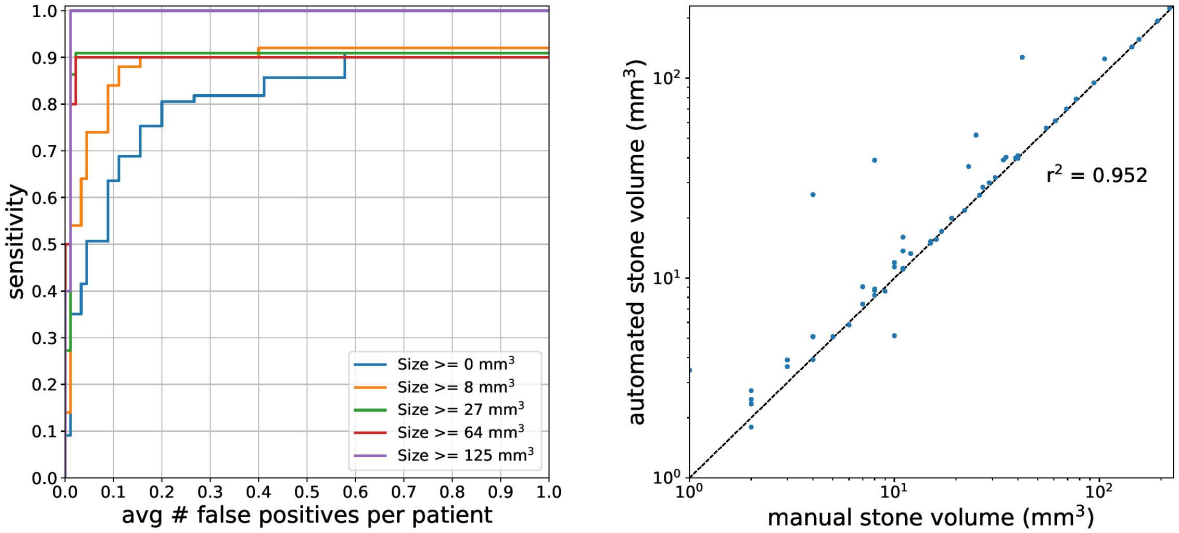

Figure 5: (left) FROC curve vs stone volume. (right) Automated vs manual stone volume measurements (with threshold = 130 HU).

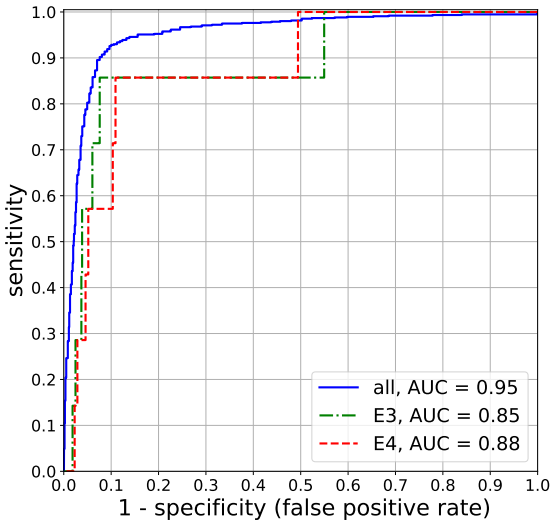

Figure 6: ROC curve for patient-level kidney stone detection on the UW-CTC dataset (N=6,136). Also shown are the ROC curves with sensitivity computed specifically for the kidney stones with E3 (N=15) or E4 (N=7) E-RADS scores.

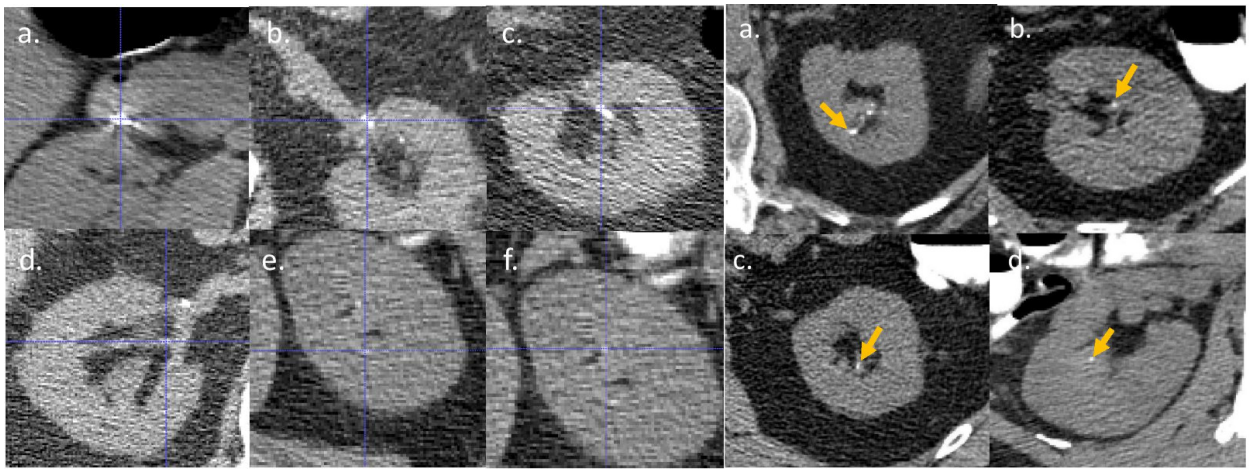

Figure 7: **(left) NMMC-CTC False positives:** (a.) metal object. (b.-d.) plaque. (e.-f.) image noise. **(right) NMMC-CTC False negatives:** (a.) False negative where there were 4 stones close together. One stone was lost after denoising. (b.-d.) These false negatives were all relatively small and less bright stones (max HU < 400).

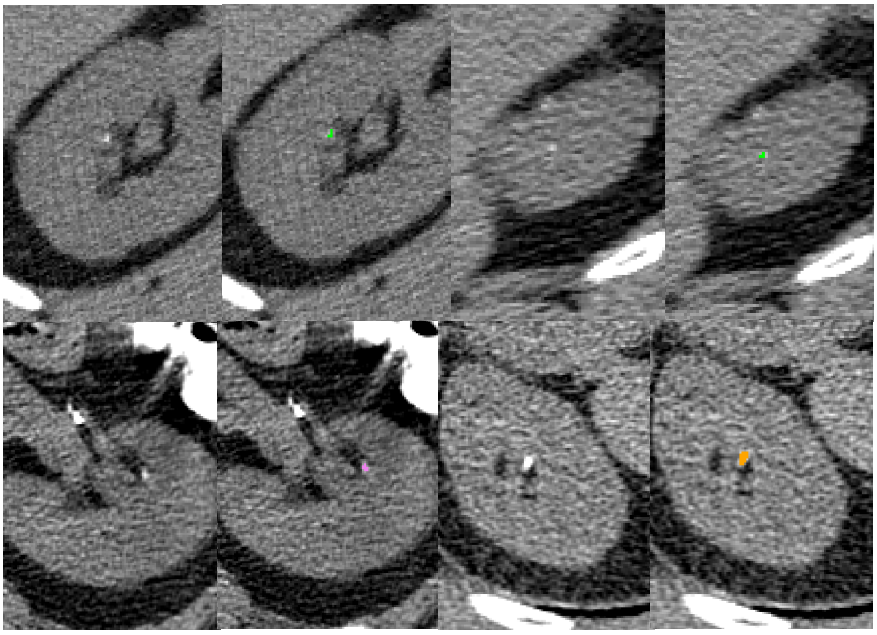

Figure 8: **NMMC-CTC True positives:** Four true positive examples are shown, with and without the segmentation overlay.

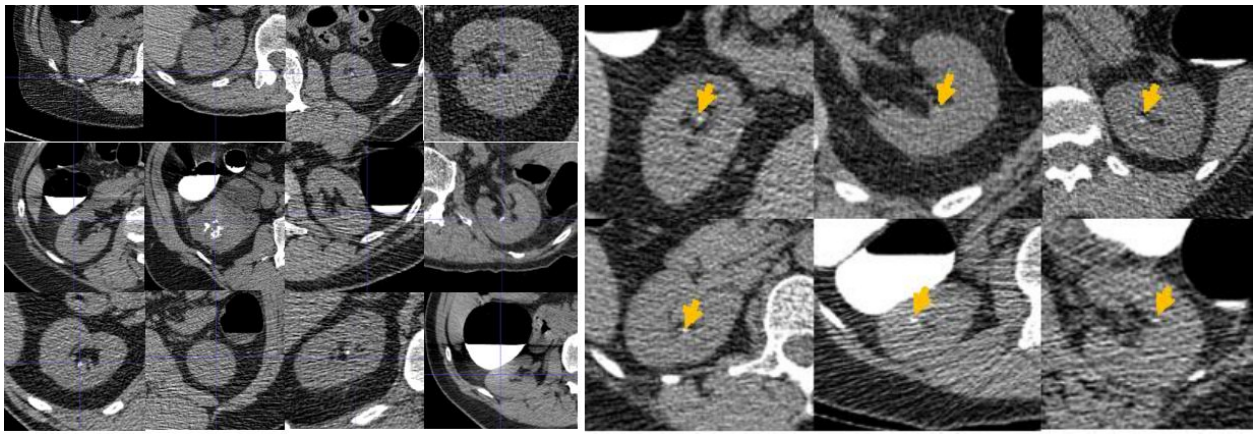

Figure 9: **(left) UW-CTC False positives.** A random sample is shown. The detections are shown with the crosshairs. Top row - noise/rib, noise, unlabeled stone, noise. Middle row - noise, calcified tumor, noise, and unlabeled stone. Bottom row - unlabeled stone, noise, unlabeled stone, and bright region from beam hardening. **(right) UW-CTC False negatives.** A random sample is shown and visualized with a soft tissue window between -160 to 240 HU.