JOURNAL OF ENDOUROLOGY
Volume 38, Number 8, August 2024
© Mary Ann Liebert, Inc.
Pp. 748–754
DOI: 10.1089/end.2023.0740

Development of UroSAM: A Machine Learning Model to Automatically Identify Kidney Stone Composition from Endoscopic Video

Jixuan Leng, BS,[1] Junfei Liu, BS, BA,[1] Galen Cheng, MD,[2] Haohan Wang, PhD,[3] Scott Quarrier, MD, MPH,[2] Jiebo Luo, PhD,[1] and Rajat Jain, MD[2]

Abstract

Introduction: Chemical composition analysis is important in prevention counseling for kidney stone disease. Advances in laser technology have made dusting techniques more prevalent, but this offers no consistent way to collect enough material to send for chemical analysis, leading many to forgo this test. We developed a novel machine learning (ML) model to effectively assess stone composition based on intraoperative endoscopic video data.
Methods: Two endourologists performed ureteroscopy for kidney stones ≥ 10 mm. Representative videos were recorded intraoperatively. Individual frames were extracted from the videos, and the stone was outlined by human tracing. An ML model, UroSAM, was built and trained to automatically identify kidney stones in the images and predict the majority stone composition as follows: calcium oxalate monohydrate (COM), dihydrate (COD), calcium phosphate (CAP), or uric acid (UA). UroSAM was built on top of the publicly available Segment Anything Model (SAM) and incorporated a U-Net convolutional neural network (CNN).
Discussion: A total of 78 ureteroscopy videos were collected; 50 were used for the model after exclusions (32 COM, 8 COD, 8 CAP, 2 UA). The ML model segmented the images with 94.77% precision. Dice coefficient (0.9135) and Intersection over Union (0.8496) confirmed good segmentation performance of the ML model. A video-wise evaluation demonstrated 60% correct classification of stone composition. Subgroup analysis showed correct classification in 84.4% of COM videos. A *post hoc* adaptive threshold technique was used to mitigate biasing of the model toward COM because of data imbalance; this improved the overall correct classification to 62% while improving the classification of COD, CAP, and UA videos.
Conclusions: This study demonstrates the effective development of UroSAM, an ML model that precisely identifies kidney stones from natural endoscopic video data. More high-quality video data will improve the performance of the model in classifying the majority stone composition.

Keywords: ureteroscopy, metabolic stone, image guided therapy

Introduction

Kidney stone disease (KSD) affects 9% of the US population and results in billions of dollars of annual cost to society. The recurrent nature of KSD results in multiple emergency room visits, hospital admissions, and surgical procedures. Therefore, long-term prevention is important for KSD patients. Prevention involves obtaining a composition analysis of the kidney stone removed at the time of surgery and a panel of blood and urine tests.[1–4] Both AUA and EAU guidelines recommend that urologists *should* send stone material for composition analysis.[1,5] Previous literature suggests that stone composition meaningfully informs prevention counseling and can change pharmacological management.[6–8] Therefore, at a kidney stone center that manages both acute and chronic phases of KSD, stone composition is important to obtain.

[1]University of Rochester, Rochester, New York, USA.
[2]University of Rochester Medical Center, Rochester, New York, USA.
[3]University of Illinois Urbana-Champaign, Champaign, Illinois, USA.

Ureteroscopy with laser lithotripsy is the most commonly performed procedure for KSD. Historically, the standard has been to fragment the stone into small pieces, remove them using a basket, and send the fragments for chemical analysis. However, this clinical practice is evolving because of the advancement of laser technology. The appearance of adjustable pulse width, higher treatment frequencies, and separated cavitation bubbles have greatly improved the efficacy and efficiency of laser lithotripsy.[9,10] With these changes, stone "dusting" has evolved as an alternative to the traditional fragmentation and basketing method. Dusting entails fragmenting the stone into small enough pieces that can pass naturally. In older lasers, getting fragments small enough to pass was laborious and inefficient, limiting widespread use of dusting, but recent studies show that the acceptance of dusting as a standard of care is increasing.[9–13] In addition, it is important to recognize that using a stone extraction basket and stone analysis laboratory testing add significant expense baskets that can cost between $250–300, whereas composition testing ranges from $25–35. Therefore, any innovation that allows urologists to forgo basketing and stone analysis would be beneficial as health care costs rise.

Computer vision algorithms have been used on still images and clinical data to predict stone composition[14] and classify abnormalities in vision-based gastroscopy.[15] Deep learning in the form of convolutional neural networks (CNNs) has been used to predict chemical composition of stones both in laboratory and *in vivo* settings.[16,17] The process of separating an object from the background in an image is known as segmentation.[18] A publicly available segmentation model called Segment Anything Model (SAM) was used in this study. SAM allows natural images to be segmented but requires user input through selection of specific parts of the image and bounding boxes to select which portion of the image to focus on.[19] Computer vision models such as U-Net allow for automated kidney stone segmentation. U-Net was specifically created for segmentation of biomedical images with smaller datasets.[17,20] In our study, we created a machine learning model that can segment and characterize the majority stone composition from ureteroscopy videos. We used U-Net for initial raw mask generation, a heuristic postprocessing approach to extract prompts, and adapted SAM for final stone mask generation. Such a model would allow the urologist to forgo collecting stone fragments for laboratory chemical composition analysis, saving time and cost.

Materials and Methods

The institutional review board approved the study. Subjects undergoing ureteroscopy by fellowship-trained endourologists for stones ≥ 10 mm between July 2022 and November 2023 were identified through retrospective chart review. Boston Scientific LithoVue and Karl Storz Flex-X ureteroscopes were used based on surgeon preference. The stones were primarily dusted, but stone fragments were sent for chemical composition analysis to an external laboratory (Arup Laboratories, Salt Lake City, UT). Demographic and clinical data were recorded, and endoscopic video files were collected. We extracted individual frames from ureteroscopy videos that showed the stone with clarity and focus.

We took different approaches to process and prepare datasets for the two distinct task requirements for the UroSAM model as follows: stone segmentation and stone majority composition classification.

Stone segmentation

Trained research assistants traced images to identify the stone; this tracing was considered the "ground truth" (Fig. 1 top row). Figure 2 shows the schematics of the UroSAM model. The necessity for human-labeled prompts in SAM[19] is time and resource intensive. Therefore, the U-Net CNN model was used because of its superiority in handling complex imaging scenarios.[21] To improve the U-Net output, rather than using raw images, our approach involves deploying U-Net segmentation on numeric representations of images called image embeddings generated by the SAM vision encoder (Fig. 3a).

As illustrated by Figure 3b, the raw masks generated by U-Net may include undesired background noise which is problematic when precise segmentation is critical in the performance of subsequent tasks (i.e., composition classification). To address this, a heuristic method for mask postprocessing was developed. First, disconnected regions in the raw mask were removed if they were <40% of the largest region (Fig. 3b, c). Second, 60 points were identified along the mask edge, 30 inside and 30 outside, providing a distinct boundary between the kidney stone and its surroundings. Third, the center of the mask and two random internal points are included as foreground points (Fig. 3c). This refined mask is used as a prompt for the SAM, generating the final output (Fig. 3d) which is compared to the ground truth.

To assess the segmentation performance of UroSAM, a fivefold cross-validation was performed with the segmentation dataset, including 80% training images and 20% testing images. A number of metrics were calculated, including intersection over union (IOU), mean IOU (MIOU), precision (positive predictive value), recall (sensitivity), and Dice coefficient. IOU is a measure of the predicted segmentation and its overlap with the ground truth tracing. MIOU averages the segmentation of focusing on the stone and correctly identifying the background from the stone. Precision is the ratio of true positives (correctly identified stone pixels) to the total number of pixels predicted to be positive. Recall is the ratio of true positives over the total number of relevant pixels (true positives and false negatives). The Dice coefficient measures the similarity between the ground truth and predicted masks in image segmentation, indicating the accuracy of the segmentation with a value closer to 1 signifying better performance. F1 is a statistical measure that combines the metrics of precision and recall through a harmonic mean such that a value closer to 1 indicates improved segmentation. UroSAM performance was compared with a "vanilla" SAM model (i.e., using bounding boxes extracted from human-labeled ground truth masks as prompts for training and testing), as well as a U-Net model deployed on SAM image embeddings (i.e., no heuristic postprocessing and no SAM modules).

Composition classification

The majority stone compositions from the laboratory analysis were labeled as the "ground truth" for each video. We recorded the endourologists' prediction on the majority stone

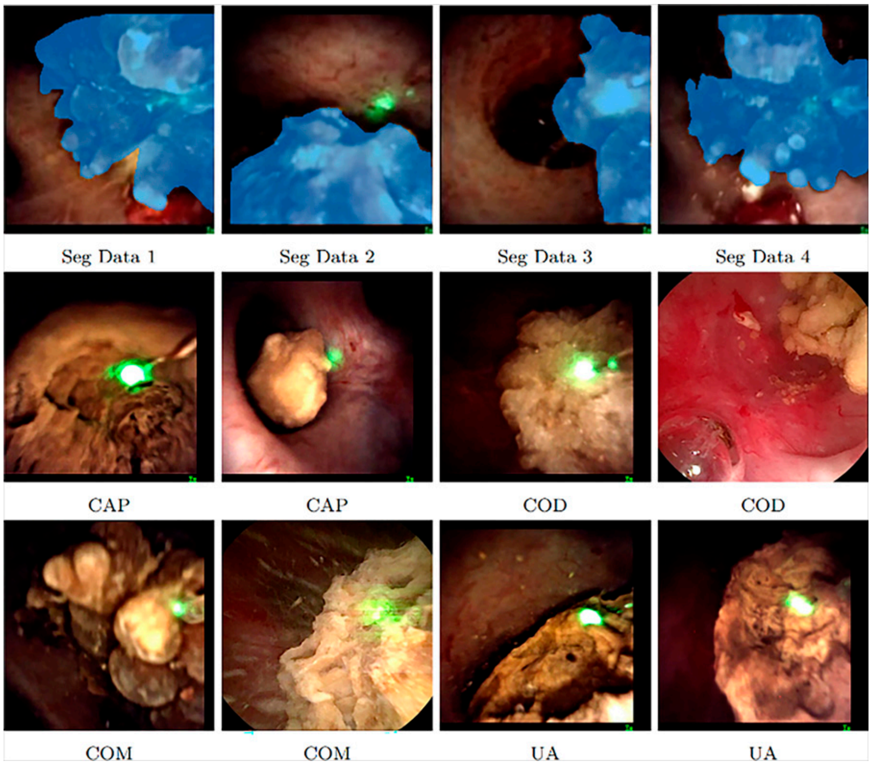
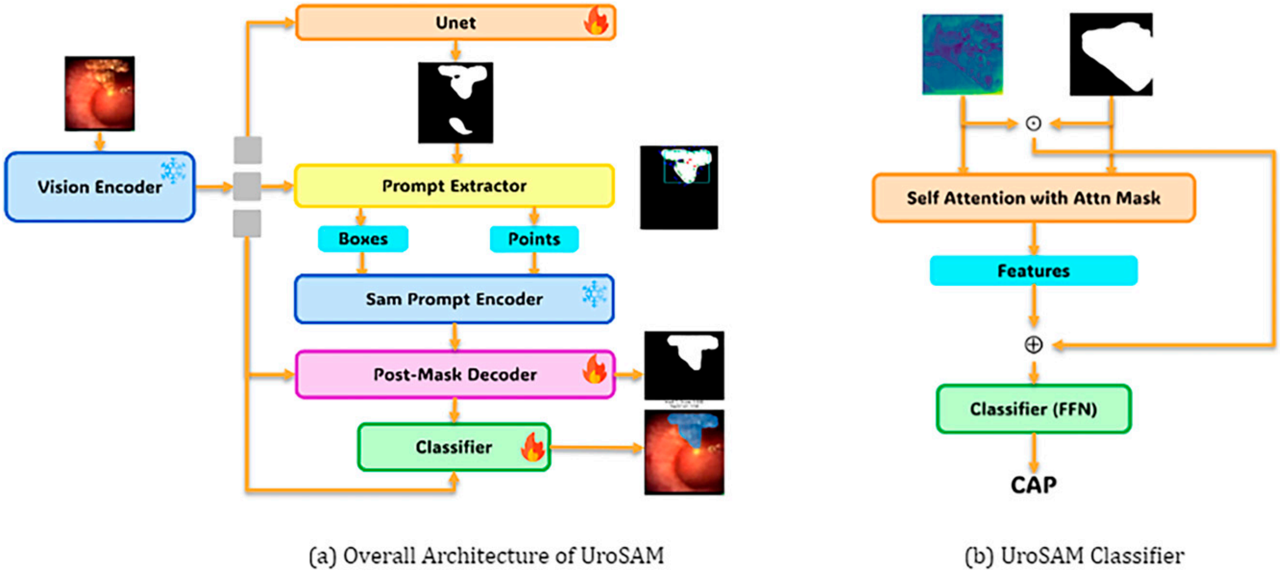


FIG. 1. Segmentation (top row) and visualization of different majority stone types.

composition based on the endoscopic video. In creating the composition classification model, we followed more stringent criteria in preprocessing the dataset to ensure the model's quality. Therefore, low resolution videos or blurry images were excluded. Examples of stone segmentation and stone classification are visualized in Figure 1.

Once we obtained the model-generated binary masks that indicate the foreground as 1 and 0 otherwise, a basic approach for classification would be to segment the images according to these masks, extracting features from the segmented regions, and then proceeding to classification. However, this would be time consuming. Therefore, we reused



(a) Overall Architecture of UroSAM

(b) UroSAM Classifier

FIG. 2. Overall architecture of UroSAM for kidney stone segmentation and classification (left), where "fire" indicates trainable modules and "ice" denotes freezing modules. The detailed architecture of the UroSAM classifier (right) integrates a single-layer self-attention with binary masks as attention masks and a residual connection, complemented by a feedforward network with one hidden layer.

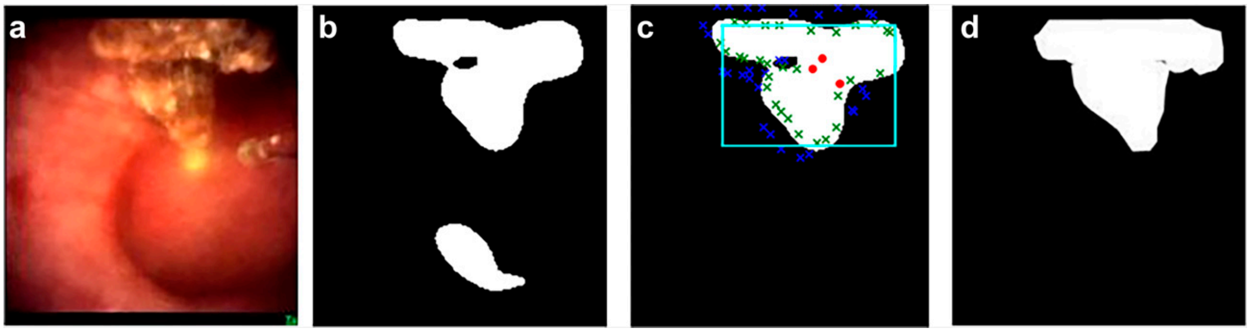


FIG. 3. Stages of Heuristic Postprocessing: From left to right: **(a)** Original Image (image embedding generated by SAM video encoder), **(b)** Initial U-Net mask output, **(c)** Prompt Extraction on U-Net masks, and **(d)** Final SAM output. SAM, Segment Anything Model.

the previously computed image embeddings (features) using the SAM vision encoder during the segmentation tasks (Fig. 2a), allowing us to optimize computational time and resources. To this end, we used a classifier that incorporated a self-attention layer and a residual connection (Fig. 2b). The binary masks were utilized within the self-attention mechanism to direct the model's focus on the foreground regions (stones). Furthermore, we performed element-wise multiplication between the image embeddings and the binary masks, which helped to diminish the background features in the embeddings. The modified embeddings were then combined with the output of the self-attention layer. Finally, these combined embeddings are fed into a feed-forward neural network with one hidden layer to carry out the classification task.

This classification dataset was then used to perform a video-wise analysis by employing a voting method. For each frame within a video, UroSAM assigned a prediction for the majority stone composition. To determine the overall classification for the video, we counted how often each class label occurred across all frames within a video. The confidence of a class label was then calculated by dividing the number of times that label occurred by the total number of frames in the video. The dataset was limited by having only 2 uric acid videos, therefore only two-fold cross-validation was performed. For the accuracy calculation, we combined the correct predictions from both folds and then divided this total by the overall number of videos in the dataset.

Results

Seventy-eight videos were recorded, and 1677 images were extracted and traced for segmentation tasks. Table 1 shows the clinical data for the cohort. Only 5 videos were of ureteral stones, while the rest were kidney stones. For the classification task, 26 videos were excluded because of low resolution and 2 were excluded because of missing laboratory composition data, leaving 50 videos for the final analysis.

Segmentation model performance

In the segmentation task, UroSAM exhibited strong performance across a range of metrics, including Intersection over Union (IOU), Mean IOU (MIOU), Dice/F1 score, Precision, and Recall. The model's precision was notably high, with an average precision of 0.9477 ± 0.0051 (Table 2). This performance suggests that the model reliably identified true positives while minimizing false positives. In addition, the Dice/F1 scores exceeded 0.9, indicating a balanced achievement in both precision and recall. Finally, both IOU metrics reflect the model's effectiveness in accurately segmenting the regions of interest.

We also evaluated the performance of the UroSAM model against two baseline configurations: a vanilla SAM model which utilizes bounding boxes extracted from ground truth masks as prompts, serving as an upper bound in performance, and a U-Net model using SAM image embeddings (Table 2).

Classification model performance

On average, UroSAM correctly predicted 60% of the majority stone composition. However, we observed that for videos of COD and CAP stones, our model often struggled to distinguish between COM and the true majority stone composition. This is likely because of 64% of the videos

Table 1. Fivefold Performance of UroSAM on Segmentation Data

Fold	*IOU*	*MIOU*	*Dice/F1*	*Precision*	*Recall*
Fold 1	0.8423	0.867	0.9085	0.9545	0.8788
Fold 2	0.8575	0.8762	0.9187	0.9464	0.9036
Fold 3	0.8502	0.8718	0.9151	0.9489	0.8932
Fold 4	0.8456	0.8681	0.9113	0.9404	0.8951
Fold 5	0.8522	0.8730	0.9140	0.9485	0.8900
Average ± Standard Deviation	0.8496 ± 0.0059	0.8712 ± 0.0038	0.9135 ± 0.0039	0.9477 ± 0.0051	0.8921 ± 0.009

IOU, Intersection over Union; MIOU, Mean IOU.

Table 2. Performance of Different Models on Segmentation Data Fold 1

Model	*IOU*	*MIOU*	*Dice/F1*	*Precision*	*Recall*
UroSAM (ours)	0.8423	0.867	0.9085	0.9545	0.8788
Vanilla SAM	0.8866	0.9009	0.9384	0.9577	0.9236
U-Net	0.854	0.8766	0.9156	0.9388	0.9066

IOU, Intersection over Union; MIOU, Mean IOU.

being COM class. To address this issue, we conducted a *post hoc* analysis using a threshold-based method for predictions. Specifically, for predictions categorized as COM, we applied a 60% confidence threshold. If the model's confidence level for COM exceeded this threshold, the prediction was accepted as COM. Otherwise, the second-highest prediction was accepted as the final classification. This approach was designed to mitigate the model's bias toward the over-represented COM class and improve the overall accuracy of our classification system. When the threshold was implemented, the model achieved an average accuracy of 62%. In addition, the introduction of the threshold method improved accuracy for all classes except COM, underscoring its effectiveness in enhancing the model's overall discriminatory capability for all classes.

Another *post hoc* analysis was performed by combining COM and COD majority stones together as CO, since prevention counseling is similar for both types of stones. The UroSAM model classification accuracy increased significantly from 60% (62% with adaptive threshold) to 78% (Table 3). In comparison, the endourologists predicted as much as 72% of the time (Table 3).

Discussion

In this study, we demonstrate the development of UroSAM, a machine learning model with Meta™ Segment Anything Model as the core foundation, complemented by U-Net and heuristic postprocessing to segment an endoscopic image to accurately identify kidney stones. Furthermore, we evaluated the model's ability to predict the majority stone composition. This to our knowledge is the first study to combine SAM and U-Net in complementary roles—this innovative approach enables accuracy and optimization in the analysis of ureteroscopy video images for stone prevention while also being relatively efficient in the use of computational resources.

Previous attempts to predict stone composition from images of kidney stones have used photographs or *ex vivo* models by placing stones in a cylinder.[16,22] As we used natural endoscopic images for UroSAM, ML model performance from these studies is not comparable with the present study. Other studies have used endoscopic images but employed different ML models. Setia and colleagues compared the performance of U-Net, U-Net++, and DenseNet, demonstrating a Dice coefficient of 0.84 for the U-Net++.[17] As the datasets were different, the results cannot be compared directly, but UroSAM performed well in segmentation, with high precision, recall, and Dice/F1. Similarly, Oh and associates applied a ResNet CNN to individual frames from endoscopic images. They combined COM, COD, and CAP in one group, with 88.2% correct classification. In our study's *post hoc* analysis, we elected to separate CAP from COM and COD as CAP stones require a significantly different prevention approach. We demonstrated that by combining COM and COD, the classification accuracy increased from 60% to 78% (62% to 78% with adaptive thresholding). The endourologist prediction was 72% which was comparable to that of the ML algorithm.

There are several limitations to this study. First, a retrospective design naturally introduces selection bias. Second, there were limited data for stones in the ureter, limiting the generalizability of the ML model. Third, a majority of the included ureteroscopy videos were COM, 32/50 (64%), which can cause "overfitting," a

Table 3. Clinical and Machine Learning Data for Each Stone Type. Calcium Oxalate Represents a Combination of COM and COD

*Majority stone composition (*n *= number of videos)*	*Age*	*BMI*	*Stone burden (mm)*	*Stone density (HU)*	*Model without threshold correct (number/%)*	*Model with threshold correct (number/%)*	*Endourologist correct (number/%)*
COM (*n* = 32)	59.8	32.4	17.4	1117.2	27/84.4%	23/71.9%	20/62.5%
COD (*n* = 8)	55.6	27.4	18.5	1134.9	1/12.5%	3/37.5%	7/87.5%
CAP (*n* = 8)	52.3	34.2	21.2	971.8	2/25.0%	4/50.0%	2/25.0%
UA (*n* = 2)	69.0	43.0	43.0	609.5	0/0%	1/50.0%	1/50%
CO (*n* = 40)	59.1	31.5	17.6	1120.4	37/92.5%	34/85.0%	33/82.5%
Total (*n* = 50)					30/60%	31/62%	30/60%
Total (COM and COD combined as CO [*n* = 50])					39/78%	39/78%	36/72%

COM, Calcium Oxalate Monohydrate; CO, Calcium Oxalate; COD, Calcium Oxalate Dihydrate; CAP, Calcium Phosphate; UA, Uric Acid.

phenomenon known in the ML literature with unbalanced datasets. This was apparent as COM had the highest correct percentage of 71.875%, whereas the other stones were 37.5%, 50%, and 50% for COD, CAP, and UA, respectively. This finding suggests that more video data are needed to improve accuracy, especially for the less prevalent stone types. Fourth, although vanilla SAM demonstrates high performance, its reliance on human-labeled prompts for inference renders it impractical for real-world applications where resources and time are limited. These comparisons underline the strengths of our model, particularly in achieving high precision, a key metric for classification tasks. Finally, the variable intracalculus architecture of kidney stones is well known.[23] Therefore, it is possible that the ground truth, that is, the chemical stone composition, does not accurately reflect the true stone composition. To solve this limitation, we would need to extract whole stones or large fragments, as can be done during percutaneous nephrolithotomy, and train the model on dusting of these larger fragments in a porcine model.

Future studies would collect data prospectively, examine the ability of an ML model in identifying minority stone composition which can influence medical management, and include real-time interpretation of ureteroscopy videos intraoperatively. In addition, clinical preprocedural factors, including demographics, laboratories, and previous stone type, could be included in the ML model to enhance its predictive capability.

Conclusions

Our study demonstrates the effective development of UroSAM, a machine learning model for identifying kidney stones and their compositions from endoscopic video data during ureteroscopy. UroSAM combines an adapted SAM, serving as the foundation, with a U-Net and heuristic post-processing for prompt extraction. Further work is required to improve the overall accuracy and generalizability of the model.

Authors' Contributions

J.L. and J.L. performed data annotation and developed the machine learning algorithm under the supervision of J.L. and H.W. J.L. and J.L. analyzed data under the supervision of J.L. and R.J. G.C. led the writing of the draft, with contributions from J.L., J.L., J.L., and R.J. R.J. and S.Q. were the endourologists involved in the study. R.J. and J.L. developed the concept and supervised all the aspects of the study. All authors were involved with editing and reviewing the article.

Disclosure Statement

No competing financial interests exist.

Funding Information

No funding was received for this article.

References

1. Pearle MS, Goldfarb DS, Assimos DG, et al. Medical management of kidney stones: AUA guideline. J Urol 2014;192(2):316–324; doi: 10.1016/j.juro.2014.05.006
2. Pak CYC, Poindexter JR, Adams-Huet B, et al. Predictive value of kidney stone composition in the detection of metabolic abnormalities. Am J Med 2003;115(1):26–32; doi: 10.1016/S0002-9343(03)00201-8
3. Kourambas J, Aslan P, Teh CL, et al. Role of stone analysis in metabolic evaluation and medical treatment of nephrolithiasis. J Endourol 2001;15(2):181–186; doi: 10.1089/089277901750134548
4. Coe FL, Wise H, Parks JH, et al. Proportional reduction of urine supersaturation during nephrolithiasis treatment. J Urol 2001;166(4):1247–1251.
5. Skolarikos A, Straub M, Knoll T, et al. Metabolic evaluation and recurrence prevention for urinary stone patients: EAU guidelines. Eur Urol 2015;67(4):750–763; doi: 10.1016/j.eururo.2014.10.029
6. Steely A, Worcester E, Prochaska M. Contrasting response of urine stone risk to medical treatment in calcium oxalate versus calcium phosphate stone formers. Kidney 360 2024; 5(2):228–236.
7. Guimerà J, Martínez A, Tubau V, et al. Prevalence of distal renal tubular acidosis in patients with calcium phosphate stones. World J Urol 2020;38(3):789–794.
8. Williams JC, Gambaro G, Rodgers A, et al. Urine and stone analysis for the investigation of the renal stone former: A consensus conference. Urolithiasis 2021;49(1):1–16; doi: 10.1007/s00240-020-01217-3
9. Fried NM. Recent advances in infrared laser lithotripsy [Invited]. Biomed Opt Express 2018;9(9):4552–4568; doi: 10.1364/BOE.9.004552
10. Stern KL, Monga M. The moses holmium system—Time is money. Can J Urol 2018;25(3):9313–9316.
11. Liao N, Tan S, Yang S, et al. A study comparing dusting to basketing for renal stones ≤ 2 cm during flexible ureteroscopy. Int Braz J Urol 2023;49(2):194–201; doi: 10.1590/s1677-5538.ibju.2022.0382
12. El-Nahas AR, Almousawi S, Alqattan Y, et al. Dusting versus fragmentation for renal stones during flexible ureteroscopy. Arab J Urol 2019;17(2):138–142; doi: 10.1080/2090598X.2019.1601002
13. Humphreys MR, Shah OD, Monga M, et al. Dusting versus basketing during ureteroscopy–which technique is more efficacious? A prospective multicenter trial from the EDGE Research Consortium. J Urol 2018;199(5):1272–1276; doi: 10.1016/j.juro.2017.11.126
14. Abraham A, Kavoussi NL, Sui W, et al. Machine learning prediction of kidney stone composition using electronic health record-derived features. J Endourol 2022;36(2):243–250; doi: 10.1089/end.2021.0211
15. Cong Y, Wang S, Liu J, et al. Deep sparse feature selection for computer aided endoscopy diagnosis. Pattern Recognit 2015;48(3):907–917; doi: 10.1016/j.patcog.2014.09.010
16. Black KM, Law H, Aldoukhi A, et al. Deep learning computer vision algorithm for detecting kidney stone composition. BJU Int 2020;125(6):920–924; doi: 10.1111/bju.15035

17. Setia SA, Stoebner ZA, Floyd C, et al. Computer vision enabled segmentation of kidney stones during ureteroscopy and laser lithotripsy. J Endourol 2023;37(4):495–501; doi: 10.1089/end.2022.0511
18. Snyder WQH. Segmentation. In: Fundamentals of Computer Vision. Cambridge University Press; 2017; pp. 149–199; doi: 10.1017/9781316882641.013
19. Kirillov A, Mintun E, Ravi N, et al. Segment anything. ArXiv 2023. preprint, abs/2304.02643,
20. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. arXiv 2015.
21. Ahmadi M, Farhadi Nia M, Asgarian S, Danesh K, Irankhah E, Lonbar AG, Sharifi A. Comparative analysis of segment anything model and u-net for breast tumor detection in ultrasound and mammography images. arXiv Preprint arXiv 2023 and 2306.12510.
22. El Beze J, Mazeaud C, Daul C, et al. Evaluation and understanding of automated urinary stone recognition methods. BJU Int 2022;130(6):786–798; doi: 10.1111/bju.15767
23. Williams JC, Jr., McAteer JA, Evan AP, Lingeman JE. Microcomputed tomography for analysis of urinary calculi. Urol Res 2010;38(6):477–484; doi: 10.1007/s00240-010-0326-x

Address correspondence to:
Rajat Jain, MD
University of Rochester Medical Center
601 Elmwood Avenue, Rochester
NY 14642
USA

E-mail: rkjain1@gmail.com

Abbreviations Used

ML	=	machine learning
KSD	=	kidney stone disease
CNN	=	convolutional neural network
COM	=	calcium oxalate monohydrate
COD	=	calcium oxalate dihydrate
CAP	=	calcium phosphate
UA	=	uric acid
SAM	=	segment anything model
IOU	=	intersection over union
MIOU	=	mean intersection over union