

Kidney Stone Segmentation from Computed Tomography Images Using Computer Vision Techniques

Rifat Burak Ergül

`rifat-ergul@hotmail.com`

Mus State Hospital

Poyraz Doğan

Istanbul University

Muhammed Edip Görür

Istanbul University

Mücahit Bayram

Istanbul Health and Technology University

Yakuphan Bilmez

Istanbul Health and Technology University

İsmail Taha Gürlek

Demre State Hospital

Mert Emre Erden

Istanbul University

Tzevat Tefik

Istanbul University

Research Article

Keywords: Kidney stone, computed tomography, deep learning, artificial intelligence, segmentation

Posted Date: March 8th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-8937650/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Objective

Urolithiasis is a common condition in urological practice, and accurate imaging is essential for proper diagnosis and management. Computed tomography (CT) is widely accepted as the gold standard for evaluating kidney stones due to its high sensitivity in assessing stone location, size, and morphology. However, manual interpretation of CT images is time-consuming and subject to observer-dependent variability. This study aimed to develop a deep learning–based model for automatic kidney stone segmentation and size estimation from CT images and to integrate this model into a web-based clinical decision support system.

Methods

An open-access dataset consisting of 3,584 CT images and corresponding kidney stone segmentation masks was used. The dataset was divided into training (70%), validation (20%), and test (10%) subsets. Image preprocessing and model development were performed using Python and the TensorFlow framework. A U-Net ++ architecture was employed for kidney stone segmentation. Model training was conducted in the Google Colab environment. Stone size estimation was performed using the equivalent circular diameter method based on the segmented stone area. Model performance was evaluated using Precision, Recall, Precision–Recall Area Under the Curve (PR-AUC), Intersection over Union (IoU), and Dice/F1-score metrics. The trained model was additionally deployed via a web-based interface to demonstrate potential clinical applicability.

Results

On the test dataset, the proposed model achieved a PR-AUC of 90.58%. The Dice/F1-score and IoU values were 82.10% and 69.63%, respectively. Precision and Recall were 76.66% and 88.36%. Qualitative evaluation demonstrated accurate localization and segmentation of kidney stones, including small calculi.

Conclusion

This study presents an effective deep learning–based approach for automated kidney stone segmentation and size estimation from CT images. The favorable quantitative results and consistent visual performance suggest that the proposed system may support clinical decision-making by reducing workload and improving diagnostic consistency in routine urological practice.

Introduction

Kidney stone disease is one of the most common urological disorders worldwide and represents a significant public health burden due to its high prevalence and recurrence rate. The disease may present with acute renal colic, hematuria, urinary tract infections, and, in advanced cases, obstructive uropathy leading to loss of renal function. Owing to its frequent acute presentation, urolithiasis accounts for a substantial proportion of emergency department admissions, making rapid and accurate diagnosis essential for effective clinical management [1–3]. Non-contrast computed tomography (NCCT) is currently regarded as the gold standard imaging modality for the evaluation of kidney stones, owing to its high sensitivity and specificity in detecting calculi, as well as its ability to accurately assess stone size, location, density, and anatomical relationships [4, 5]. Despite these advantages, manual interpretation of NCCT images remains time-consuming and prone to inter-observer variability. Factors such as increased workload, limited availability of experienced radiologists, and time pressure, particularly in emergency settings, may adversely affect diagnostic accuracy. [6]. Moreover, small stone size, low-contrast appearance, and complex anatomical variations can further contribute to missed or misinterpreted findings in routine clinical practice [7].

These limitations have prompted growing interest in automated and reliable tools to support radiological assessment. In recent years, artificial intelligence (AI), particularly deep learning–based approaches using convolutional neural networks (CNNs), has demonstrated remarkable performance in medical image analysis tasks, including classification, detection, and segmentation [8, 9]. Several studies have applied AI-based techniques to kidney stone detection and segmentation on NCCT images, reporting encouraging diagnostic performance and potential reductions in clinical workload [10, 11]. However, many previously published studies are limited by small sample sizes, single-center datasets, class imbalance, and insufficient consideration of clinical integration and real-time applicability [12, 13]. In addition, issues related to model explainability and deployment within clinical decision support systems have often been inadequately addressed. More recent investigations suggest that deep learning models trained on large, diverse datasets and designed with practical clinical workflows in mind may yield more robust and generalizable results [14].

Therefore, the present study aimed to develop a deep learning–based model for the automatic segmentation and size estimation of kidney stones from NCCT images using a large open-access dataset. Furthermore, the trained model was integrated into a web-based clinical decision support system to facilitate practical deployment and to support diagnostic decision-making in routine urological practice.

Materials and Methods

Dataset and Study Design

In this study, a deep learning–based approach was developed for the automatic segmentation of kidney stones from computed tomography (CT) images. The dataset consisted of CT images paired with corresponding binary segmentation masks indicating stone regions. Data organization and processing

were performed in accordance with recommended best practices for medical image analysis to ensure reproducibility and methodological transparency [8].

Input data included common raster image formats (PNG, JPG, TIFF) and DICOM (.dcm) files. For DICOM images, pixel intensities were normalized using VOI-LUT transformation and subsequently converted into a three-channel representation compatible with the model input. This preprocessing step aimed to reduce intensity variability arising from different CT acquisition protocols [15].

To minimize data leakage and improve patient-level generalizability, a group-based data splitting strategy was applied whenever patient-related identifiers could be inferred. This ensured that images originating from the same patient were not distributed across training, validation, and test subsets simultaneously. The dataset was divided into training (70%), validation (20%), and test (10%) sets, following commonly accepted practices in medical artificial intelligence studies [12].

Model Architecture

Kidney stone segmentation was performed using a U-Net ++ architecture. Compared to the classical U-Net, U-Net ++ employs nested skip connections that enable more effective multi-scale feature aggregation, which is particularly advantageous for segmenting small and low-contrast targets [16].

An ImageNet-pretrained EfficientNet-B3 backbone was used as the network's encoder. EfficientNet architectures are widely adopted in medical imaging due to their favorable balance between parameter efficiency and strong feature extraction capability [17]. Dropout was applied in the decoder pathway to reduce overfitting. The overall model architecture is illustrated in Fig. 1.

Preprocessing, Data Augmentation, and Mixing Strategies

All images were resized to match the model input dimensions while preserving the original aspect ratio using a letterbox/padding approach. Images were then normalized using ImageNet mean and standard deviation values before model ingestion.

To enhance robustness against variations in acquisition conditions and anatomical appearance, strong data augmentation was applied during training. Augmentation techniques included random brightness and contrast adjustment, noise injection, gamma correction, perspective transformation, elastic deformation, and horizontal and vertical flipping. Such augmentation strategies have been shown to improve generalization performance in medical image segmentation tasks substantially [18].

To further mitigate class imbalance and reduce overfitting, probabilistic MixUp and CutMix data-mixing strategies were applied during training. These approaches increase sample diversity and encourage smoother decision boundaries, contributing to more stable and generalizable learning. [18]

Training Configuration and Overfitting Prevention

Model training was conducted using the PyTorch framework, with AdamW employed as the optimization algorithm. To control overfitting, weight decay, gradient clipping, and early stopping were applied. Learning rate scheduling was performed using CosineAnnealingWarmRestarts, and Stochastic Weight Averaging (SWA) was applied during the later stages of training to improve generalization by averaging model weights across epochs [19].

Automatic Mixed Precision (AMP) was utilized throughout training to improve computational efficiency and reduce memory consumption. During validation, the probability threshold used to binarize the segmentation output was not fixed; instead, a threshold sweep was performed on the validation set to identify the value that maximized the Dice (F1) score. This strategy is particularly appropriate for segmentation problems with imbalanced pixel distributions [20].

Testing and Evaluation Metrics

Model performance was evaluated exclusively on the held-out test set. During inference, predictions were enhanced using test-time augmentation (TTA), where horizontally and vertically flipped versions of each image were processed, and the resulting probability maps were averaged. This approach improves prediction stability and robustness [21].

Segmentation performance was assessed using pixel-level Dice (F1-score) and Intersection over Union (IoU) metrics, as well as ROC and precision–recall (PR) curve–based area under the curve (AUC) values. These metrics are widely accepted for evaluating segmentation quality, particularly for small target structures such as kidney stones [22].

To generate clinically interpretable outputs, a subset of positive test cases was visualized by overlaying predicted masks, contours, bounding boxes, and equivalent circular diameter measurements onto the original CT images.

Web-Based Application

The developed segmentation model was deployed as a web-based clinical decision support tool (www.kidney-stone-segmentation.com) to facilitate practical use scenarios. The application allows users to upload CT images and receive real-time model outputs through an interactive interface. Uploaded images are processed using the same preprocessing pipeline as in training, and the model outputs probability maps and binary segmentation masks. Results are displayed visually with stone contours and measurement annotations. To ensure safe clinical use, the interface includes a disclaimer stating that the system does not replace physician judgment. An example screenshot of the web interface is shown in Fig. 2.

Results

The quantitative performance of the proposed kidney stone segmentation model is summarized in Table 1, which reports the precision, recall, PR-AUC, Dice (F1), and IoU values obtained on the independent test set. The model achieved 76.66% precision and 88.36% recall, indicating high sensitivity for kidney stone detection. The PR-AUC of 90.58% demonstrates strong discriminative performance under class-imbalanced conditions, as shown in the precision–recall curve in Fig. 3. In terms of spatial overlap, the model yielded a Dice (F1) score of 82.10% and an IoU of 69.63%, reflecting accurate pixel-wise agreement between predicted masks and ground truth annotations. Representative qualitative examples of kidney stone segmentations overlaid on CT images are presented in Fig. 4, further supporting the quantitative results.

Table 1
Performance metrics of the model

Precision (%)	Recall (%)	PR-AUC (%)	Dice (%)	IoU (%)
76,66	88,36	90,58	82,10	69,63

Discussions

In this study, a deep learning–based framework was developed for the automatic segmentation of kidney stones from non-contrast computed tomography (NCCT) images, and its performance was evaluated using both pixel-level overlap metrics and precision–recall–based measures. The results demonstrate that the proposed model achieves a favorable balance between sensitivity and segmentation accuracy, supporting its potential as a clinically relevant decision-support tool. In particular, the high PR-AUC value highlights the model’s robustness to substantial class imbalance, a common characteristic of kidney stone segmentation tasks in which stone pixels represent only a small fraction of the image [8].

Non-contrast CT remains the imaging modality of choice for evaluating nephrolithiasis due to its superior sensitivity and detailed anatomical depiction [23]. However, accurately interpreting NCCT images can be challenging in routine clinical practice, especially in high-workload environments and emergency settings. Small stones, low contrast relative to surrounding tissues, and complex anatomical variations may increase the likelihood of missed or inconsistent detections [6]. In this context, automated segmentation methods can provide consistent and objective delineation of stone regions, potentially reducing inter-observer variability and supporting radiologists during image interpretation.

Several deep learning–based approaches for kidney stone detection and segmentation have been reported in recent years. Prior studies have explored convolutional neural network–based frameworks for automated stone detection, volumetric segmentation, and stone burden assessment on CT images, reporting encouraging diagnostic performance [10, 11, 24]. More recent work has investigated advanced

architectures, including hybrid and attention-based models, to improve robustness and segmentation accuracy further [25, 26]. Compared with these approaches, the present study demonstrates competitive performance, particularly in precision–recall–based evaluation, which is increasingly recognized as a more informative metric than ROC-based measures in highly imbalanced medical imaging problems [27].

Beyond algorithmic performance, the clinical utility of artificial intelligence systems depends on their robustness, generalizability, and ease of integration into existing workflows. Previous methodological reviews emphasize that strong internal validation results alone are insufficient for clinical adoption, and that external validation and real-world testing are critical next steps [28, 29]. In the present study, integrating the segmentation model into a web-based interface represents an initial step toward practical deployment, enabling real-time visualization of segmentation outputs and facilitating interaction with end users in a clinical context.

Explainability and transparency are also increasingly recognized as key determinants of clinician trust in artificial intelligence–based systems. Although the current work focuses primarily on segmentation performance, future incorporation of explainable AI techniques, such as saliency maps or gradient-based visualization methods, may further enhance interpretability and support clinical acceptance [29, 30]. In addition, recent methodological studies highlight that combining overlap-based metrics with precision–recall–oriented evaluation provides a more comprehensive assessment of segmentation performance, particularly for small target structures such as kidney stones [27].

Several limitations of this study should be acknowledged. First, the model was evaluated on an internal test dataset derived from publicly available imaging sources, which may not fully capture the variability encountered in routine clinical practice. Second, the current framework focuses on binary stone segmentation and does not explicitly address clinically relevant attributes such as stone composition, anatomical location, or volumetric burden. Finally, although a web-based deployment was implemented, prospective and multicenter validation studies are required to assess the model’s real-world impact and generalizability.

Conclusion

In this study, a deep learning–based approach was developed for the automatic segmentation of kidney stones from non-contrast computed tomography (NCCT) images. It demonstrated robust performance on the internal test dataset, with reliable pixel-level delineation and strong discriminative capability under class-imbalanced conditions. The deployment of the model as a web-based application (www.kidney-stone-segmentation.com) represents an important step toward clinical translation by enabling real-time visualization of segmentation outputs through an accessible interface. Although external validation and prospective multicenter studies are required to confirm its generalizability and clinical impact, the results of this study suggest that artificial intelligence–driven segmentation systems have the potential to support radiological assessment and enhance urological diagnostic workflows.

Declarations

Ethics

Ethical approval was not required for this study.

Footnotes

Authorship Contributions

Concept and design of the study: R.B.E., P.D., T.T.; Data collection: M.E.G., M.E.E., İ.T.G.; Data preprocessing: R.B.E., İ.T.G.; AI model development and training: P.D., M.B., Y.B.; Statistical analysis and model evaluation: R.B.E., M.E.E.; Software development and coding: P.D., M.E.G.; Manuscript writing: P.D., M.E.G., M.B., Y.B.; Review and editing: R.B.E., İ.T.G., M.E.E., T.T.; Project supervision: T.T..

Conflict of Interest: No conflict of interest was declared by the authors.

Financial Disclosure: The authors declared that this study received no financial support

Data Availability

The dataset used in this study is publicly available from an open-access repository. Processed data and model outputs are available from the corresponding author upon reasonable request.

References

1. Finger M, Finger E, Bellucci A, Malieckal DA. Medical management for the prevention of kidney stones. *Postgrad Med J*. 2023;99(1169):112–8. doi: 10.1136/postgradmedj-2021-140971.
2. Scales CD, Jr., Smith AC, Hanley JM, Saigal CS, Urologic Diseases in America P. Prevalence of kidney stones in the United States. *Eur Urol*. 2012;62(1):160–5. doi: 10.1016/j.eururo.2012.03.052.
3. Sorokin I, Mamoulakis C, Miyazawa K, Rodgers A, Talati J, Lotan Y. Epidemiology of stone disease across the world. *World J Urol*. 2017;35(9):1301–20. doi: 10.1007/s00345-017-2008-6.
4. Turk C, Petrik A, Sarica K, Seitz C, Skolarikos A, Straub M, et al. EAU Guidelines on Diagnosis and Conservative Management of Urolithiasis. *Eur Urol*. 2016;69(3):468–74. doi: 10.1016/j.eururo.2015.07.040.
5. Waite S, Scott J, Gale B, Fuchs T, Kolla S, Reede D. Interpretive Error in Radiology. *AJR Am J Roentgenol*. 2017;208(4):739–49. doi: 10.2214/AJR.16.16963.
6. Brady AP. Error and discrepancy in radiology: inevitable or avoidable? *Insights Imaging*. 2017;8(1):171–82. doi: 10.1007/s13244-016-0534-1.
7. Moore CL, Daniels B, Ghita M, Gunabushanam G, Luty S, Molinaro AM, et al. Accuracy of reduced-dose computed tomography for ureteral stones in emergency department patients. *Ann Emerg Med*.

- 2015;65(2):189–98 e2. doi: 10.1016/j.annemergmed.2014.09.008.
8. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. *Med Image Anal.* 2017;42:60–88. doi: 10.1016/j.media.2017.07.005.
 9. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. *Nat Med.* 2019;25(1):24–9. doi: 10.1038/s41591-018-0316-z.
 10. Elton DC, Turkbey EB, Pickhardt PJ, Summers RM. A deep learning system for automated kidney stone detection and volumetric segmentation on noncontrast CT scans. *Med Phys.* 2022;49(4):2545–54. doi: 10.1002/mp.15518.
 11. Cui Y, Sun Z, Ma S, Liu W, Wang X, Zhang X, et al. Automatic Detection and Scoring of Kidney Stones on Noncontrast CT Images Using S.T.O.N.E. Nephrolithometry: Combined Deep Learning and Thresholding Methods. *Mol Imaging Biol.* 2021;23(3):436–45. doi: 10.1007/s11307-020-01554-0.
 12. Park SH, Han K. Methodologic Guide for Evaluating Clinical Performance and Effect of Artificial Intelligence Technology for Medical Diagnosis and Prediction. *Radiology.* 2018;286(3):800–9. doi: 10.1148/radiol.2017171920.
 13. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.* 2019;17(1):195. doi: 10.1186/s12916-019-1426-2.
 14. Mazurowski MA, Buda M, Saha A, Bashir MR. Deep learning in radiology: An overview of the concepts and a survey of the state of the art with focus on MRI. *J Magn Reson Imaging.* 2019;49(4):939–54. doi: 10.1002/jmri.26534.
 15. Dapamede T, Li F, Khosravi B, Purkayastha S, Trivedi H, Gichoya J. DICOM LUT is a Key Step in Medical Image Preprocessing Towards AI Generalizability. *J Imaging Inform Med.* 2025;38(5):3040–8. doi: 10.1007/s10278-025-01418-5.
 16. Zhou Z, Siddiquee MMR, Tajbakhsh N, Liang J. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. *Deep Learn Med Image Anal Multimodal Learn Clin Decis Support (2018).* 2018;11045:3–11. doi: 10.1007/978-3-030-00889-5_1.
 17. Keishing V, Ahn ES, Khashim Z, Loudermilk AD, Erickson BJ. Towards Automated Craniosynostosis Diagnosis Using EfficientNet-Based Artificial Intelligence Models: A Two-Class and Multi-Class Approach. *J Imaging Inform Med.* 2025. doi: 10.1007/s10278-025-01694-1.
 18. Zeng W. Image data augmentation techniques based on deep learning: A survey. *Math Biosci Eng.* 2024;21(6):6190–224. doi: 10.3934/mbe.2024272.
 19. Kunkel KJ, Anwaruddin S. Papillary Muscle Rupture Due to Delayed STEMI Presentation in a Patient Self-Isolating for Presumed COVID-19. *JACC Case Rep.* 2020;2(10):1633–6. doi: 10.1016/j.jaccas.2020.06.036.
 20. Youden WJ. Index for rating diagnostic tests. *Cancer.* 1950;3(1):32–5. doi: 10.1002/1097-0142(1950)3:1<32::aid-cnrcr2820030106>3.0.co;2-3.
 21. Rylaarsdam LE, Johnnecheck GN, Looyenga BD, Louters LL. GLUT1 is associated with sphingolipid-organized, cholesterol-independent domains in L929 mouse fibroblast cells. *Biochimie.* 2019;162:88–96. doi: 10.1016/j.biochi.2019.04.010.

22. Rodriguez-Espinosa N, Gonzalez-Colaco Harmand M, Miranda Saavedra F, Galvan Gonzalez MA, Plasencia Nunez M, Aldea Perona AM, et al. [Prescription of acetylcholinesterase inhibitors and memantine in the Canary Islands, comparison with the Spanish population.]. *Rev Esp Salud Publica*. 2021;95.
23. Smith RC, Rosenfield AT, Choe KA, Essenmacher KR, Verga M, Glickman MG, et al. Acute flank pain: comparison of non-contrast-enhanced CT and intravenous urography. *Radiology*. 1995;194(3):789–94. doi: 10.1148/radiology.194.3.7862980.
24. Frenken MWE, Thijssen KMJ, Vlemminx MWC, van den Heuvel ER, Westerhuis M, Oei SG. Clinical evaluation of electrohysterography as method of monitoring uterine contractions during labor: A propensity score matched study. *Eur J Obstet Gynecol Reprod Biol*. 2021;259:178–84. doi: 10.1016/j.ejogrb.2021.02.029.
25. Wu N, Gao H, Xu Q, Zhang Z. Characterization and Whole-Genome Analysis of a Zearalenone-Degrading *Stappia* sp. WLB 29. *Curr Microbiol*. 2022;79(6):179. doi: 10.1007/s00284-022-02874-w.
26. Hada AM, Potara M, Astilean S, Cordaro A, Neri G, Malanga M, et al. Linezolid nanoAntibiotics and SERS-nanoTags based on polymeric cyclodextrin bimetallic core-shell nanoarchitectures. *Carbohydr Polym*. 2022;293:119736. doi: 10.1016/j.carbpol.2022.119736.
27. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One*. 2015;10(3):e0118432. doi: 10.1371/journal.pone.0118432.
28. Miller H, Johns L. Interoperability of Electronic Health Records: A Physician-Driven Redesign. *Manag Care*. 2018;27(1):37–40.
29. Shi X, Du R, Zhang J, Lei Y, Guo H. Evaluation of the anti-cancer potential of *Cedrus deodara* total lignans by inducing apoptosis of A549 cells. *BMC Complement Altern Med*. 2019;19(1):281. doi: 10.1186/s12906-019-2682-6.
30. Wei J, Wang L, Yang X. Game analysis on the evolution of COVID-19 epidemic under the prevention and control measures of the government. *PLoS One*. 2020;15(10):e0240961. doi: 10.1371/journal.pone.0240961.

Figures

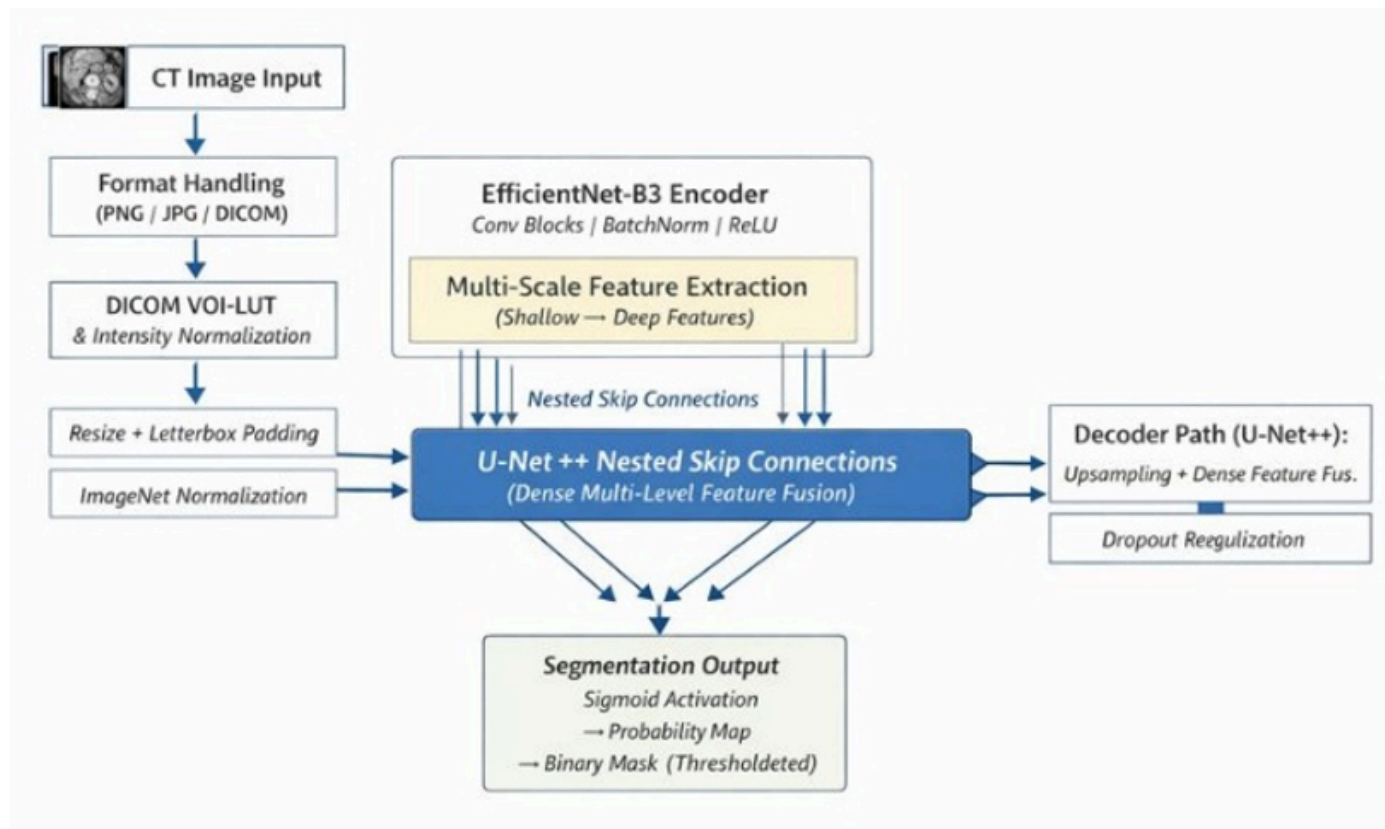


Figure 1

U-Net++-based kidney stone segmentation model with an EfficientNet-B3 encoder.

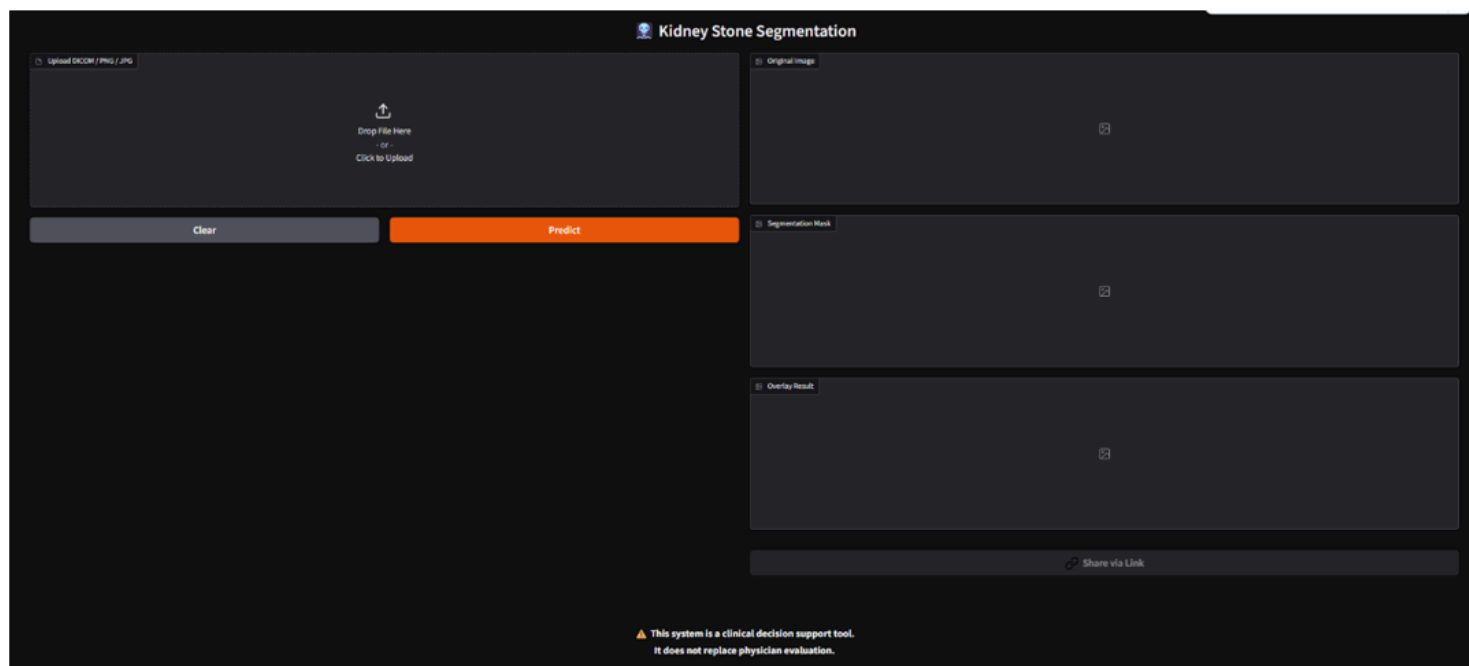


Figure 2

Screenshot of the web-based kidney stone segmentation interface (www.kidney-stone-segmentation.com).

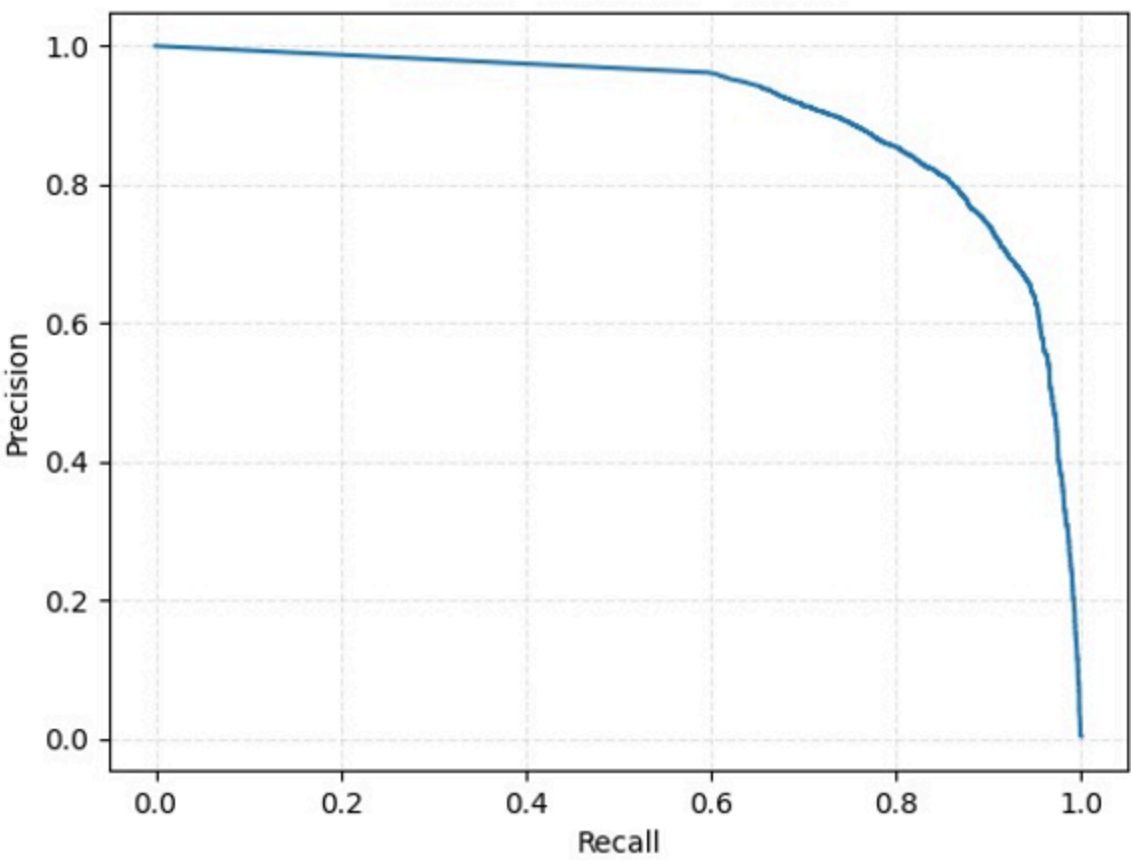


Figure 3

PR-AUC curve

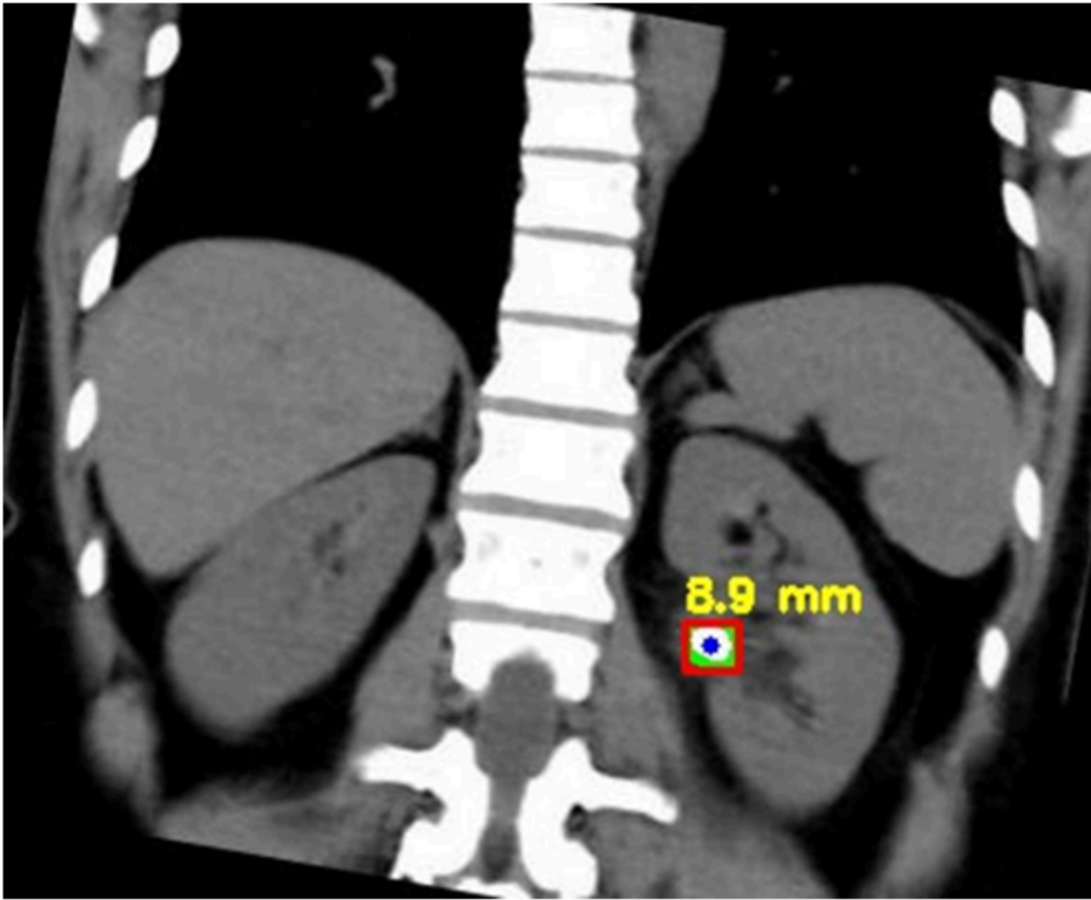


Figure 4

Overlaid CT image