Automated Detection of Renal Calculi in Abdominal Computed Tomography Scans Using Deep Learning Techniques

Ian Herrera Vázquez [1] and Ismael Eliezer Pérez Ruiz[2]

[1]Universidad Modelo, Yucatán NJ 08544, MEX
15222025@modelo.edu.com

Abstract. The detection of kidney stones on computed tomography (CT) scans is crucial for preventing renal complications and facilitating timely treatment. However, traditional methods require extensive manual intervention, which can be laborious and error-prone. The growing use of artificial intelligence and deep learning techniques offers new ways to automate and enhance these diagnostic processes. In this context, the present study proposes an automated system that integrates: (1) a MobileNetV2 classifier with a high accuracy in slice orientation (axial/coronal) identification; (2) a YOLOv8n detector localizing kidneys with >95% accuracy; (3) adaptive stone segmentation via histogram thresholding. We use a public dataset comprising 2,757 abdominal CT images, including both kidney stone and non-stone cases. Due to the dataset's heterogeneity, including slices in axial/coronal planes, a MobileNetV2-based classifier was implemented to automatically distinguish between these two, facilitating image classification. YOLOv8 is then employed for localising the kidneys, enabling the renal anatomy to be segmented efficiently and automatically. Evaluation on test set showed that the system achieves 100% sensitivity, successfully detecting all true positive. However, the specificity reflects a higher rate of false positives, with a classification accuracy of 75.76% and an F1-Score of 84%. Real-time processing was achieved, with an average inference time of <1.5s per image. These results suggest that the proposed method could be integrated into clinical settings to accelerate and improve diagnostic accuracy. However, adjustments are still required to reduce false positives and validate the system on larger, more diverse datasets.

Keywords: Kidney Stone, Deep Learning, Image Processing.

1 Introduction

1.1 A Subsection Sample

Kidney stones (renal calculi) represent a pervasive and increasing global health burden. In the year 2019 alone, 115 million cases occurred worldwide with a prevalence rate ranging from 1% to 20%; it is estimated that 1 in 10 people in the United States has been affected by this disease [1]. These crystalline deposits form in the urinary tract,

often causing excruciating pain, urinary obstruction, and potential renal damage when left undiagnosed [2]. The clinical relevance for rapid and accurate detection stems from three critical factors: the need to prevent irreversible kidney injury, the necessity to guide appropriate surgical or medical interventions, and the substantial economic costs of emergency department visits for acute stone episodes – estimated at $5 billion annually in the United States alone [3].

In modern practice, non-contrast computed tomography (NCCT) is the undisputed gold standard for stone detection, offering unparalleled sensitivity in identifying calcifications [4]. This imaging modality surpasses traditional radiography and ultrasonography by reliably visualizing radiolucent stones and detecting secondary signs of obstruction such as perinephric stranding or hydronephrosis [5]. This diagnostic superiority comes with challenges: the deluge of imaging data in busy clinical settings often leads to radiologist fatigue, introducing concerning rates of missed diagnoses – particularly for smaller calculi – while interobserver variability remains problematic even among experienced specialists [6].

In this context, the automatic medical image analysis systems, driven by computer vision and deep learning techniques has emerged as a promising solution to improve efficiency and accuracy in computer-aided medical diagnosis (CAD) [7]. Particularly, the Convolutional Neural Networks (CNN's), has revolutionized the field through the capability of autonomously learning discriminative features across diverse imaging conditions [8].

Therefore, this work proposes a fully automated end-to-end pipeline that integrates three critical stages: (1) classification of the anatomical imaging plane, (2) precise localization of the kidneys, and (3) segmentation and quantification of renal calculi. This sequential integration mirrors the typical radiological diagnostic workflow while reducing the need for manual intervention or pre-processing [9]. This end-to-end pipeline harnesses two key architectures: the MobileNetV2, which is used for the effective and efficient slice plane classification, and the YOLOv8, which is used for the kidney localization.

2 Methodology

2.1 Dataset and Preprocessing

The development of this work begins with the use of the publicly available Kaggle's "CT KIDNEY DATASET: Normal-Cyst-Tumor and Stone" dataset, containing 12,446 anonymous CT across four diagnostic categories. In order to focus on kidney stone detection, cysts (n=3,709) and tumor (n=2,283) cases were excluded, as these represent distinct pathological entities with different imaging characteristics.

This yielded two relevant categories: Normal renal anatomy (n=5,077) and confirmed kidney stone cases (n=1,377). To address potential class imbalance [10], a stratified sampling approach was implemented. Additionally 15 images were identified as oblique planes through manual inspection and consequently excluded, resulting in a final balanced dataset of 1,362 normal cases to match the number of cases in the stone category.

A significant preprocessing challenge arose from the mixture of Axial and Coronal imaging planes within the dataset. This heterogeneity required the development of an automated classification system to distinguish between these two slice orientations. To address this, a transfer learning classifier based on MobileNetV2 was implemented to identify slice orientation and automate the dataset reorganization.

2.2 Architectures: MobileNetV2 and YOLOv8n

The selection of neural network architectures was based by considering clinical computational efficiency and task complexity. This section outlines the technical rationale behind selecting MobileNetV2 for plane classification and YOLOv8n for renal localization. Overall, the choice of MobileNetV2 and YOLOv8n responds to a strategy focused on maximizing the efficiency and portability of the system, while maintaining levels of accuracy suitable for application in an automated clinical environment [11].

To classify the type of CT slice, we implemented a convolutional neural network (CNN) based on the MobileNetV2 architecture, a lightweight CNN optimized for use on mobile or hardware-constrained devices. Its use of inverted residuals and depth wise separable convolutions reduce the number of parameters and operations while preserving accuracy.

Heavier alternatives, such as ResNet50, InceptionV3 and EfficientNet, were considered but discarded due to their higher computational demand and longer training times. These features are less desirable in this context without a substantial improvement in accuracy for this binary task [12].

For kidney localization, YOLOv8n, a real-time object detection architecture, was used for its minimal computational demand and suitability for low-resource settings [13]. Larger variants such as YOLOv8m or YOLOv8l were discarded, due to higher resource consumption, and alternative architectures such as Faster R-CNN or SSD imply higher latency and complexity in implementation [14]. The selected YOLO variant, version 8 model nano, is the lightest of the YOLOv8 family, designed to operate in low computational consumption environment, which aligns with the project objectives.

2.3 MobileNetV2-based Classifier Training

We implemented MobileNetV2 with pre-trained weights from ImageNet, retaining only its feature extraction backbone while removing the original classification layers.

To leverage transfer learning, we froze the base model weights, preserving its low- and mid-level features. This feature extractor was integrated into a custom sequential model to which we added a GlobalAveragePooling2D layer following the base model to reduce the spatial dimensions of the feature maps. Then, a dropout layer (0.2 rate) was included to introduce regularization to make the model learn more robust features [15].

The final layer was a single dense neuron with a sigmoid activation function, suitable for a binary classification (Axial versus Coronal). The model used the Adam optimizer, chosen for its adaptive learning capabilities, and the binary cross-entropy loss function, the standard for binary classification tasks [16]. The model contained approximately 2.26M, however, only 1,281 (exclusively to the classification head) of these were trained.

2.4 YOLOv8n-based Kidney Detector Training

The YOLOv8n model, with 72 layers and approximately 3M parameters, was initialized using pre-trained weights from the COCO dataset using Ultralytics default pipeline. Although no layers were explicitly frozen during training, the architecture inherently leveraged transfer learning by adapting its general-purpose object detection backbone to the specialized task of kidney localization in medical CT slices.

The model was trained using 780 annotated CT slices with at least one kidney instance, and 221 additional slices for validation. The only preprocessing applied was built-in resizing to 640x640px [17], with no external augmentation techniques.

Training was set for 50 epochs, with early stopping after 10 epochs. The training ended at epoch 31, when validation metrics plateaued, completing in 17 minutes and 24 seconds with a batch size of 16 using Google Colab's GPU. This setup achieved a mean Average Precision (mAP_{50}) of 0.992 on the validation set. On a separate test set (200 CT slices), the model demonstrated perfect recall by identifying all kidney instances while maintaining precise anatomical localization.

2.5 Stone Detection Algorithm

Following the localization of the kidney using the corresponding YOLOv8n model, the next step involved the extraction of the region of interest (ROI) to identify renal calculi. The bounding box generated by YOLOv8n was used to crop the kidney area from the original CT image. To minimize inclusion of adjacent structures, a fixed 10px margin was removed from each side of the bounding box before further processing.

The resulting ROI was converted to grayscale, facilitating processing with OpenCV. An adaptive thresholding strategy is implemented to isolate hyperdense regions indicative of stones. Specifically, we computed the pixel intensity mean and standard deviation within the grayscale ROI and defined a dynamic threshold as follows:

$$\mu + 2\sigma \times \alpha = umbral \tag{1}$$

Where μ is the mean intensity, σ is the standard deviation, and α is a scaling factor adjusted to enhance the sensitivity of the detection The scaling factor of 1.2 enhance sensitivity to regions with significantly higher intensity.

Pixels above this computed threshold were binarized using a hard threshold operation, effectively highlighting regions with significantly higher intensity values—characteristic of calcifications or stones in CT imaging [18].

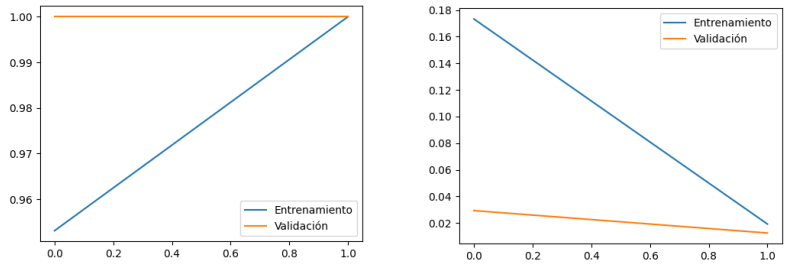
Post-thresholding, connected components were analyzed to count the number of individual stones, and their corresponding contours were overlaid CT image previously processed for visualization. This approach, proved effective for segmenting and quantifying stones within the kidney ROI

3 Results

3.1 Quantitative Evaluation of MobileNetV2 and YOLOv8n Architectures

The models demonstrated near-ideal learning trajectories during training [19] (see Fig. 1). The MobileNetV2 achieved perfect convergence by the second epoch, reaching 100% training accuracy (loss=0.0014) while maintaining optimal validation performance that persisted through subsequent epoch.

a)



b) c)

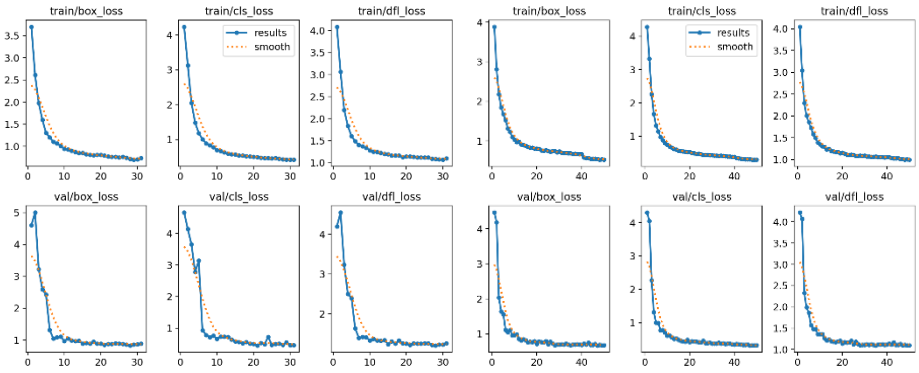


Fig. 1. Training and validation performance curves for the models. (a) MobileNetV2 Accuracy and loss plots; (b) YOLOv8n detector training metrics for kidney localization in Axial CT. (c) YOLOv8n detector training metrics for kidney localization in Coronal CT.

For the YOLOv8n detectors, the Axial-plane model demonstrated a marginally higher recall but slightly lower precision compared to the Coronal-plane variant. This reflects an accuracy gap of 0.5% in their respective test sets. Both YOLO models achieved balanced F1-scores (Axial=97.7, Coronal=97.44), with error rates below 5%. This indicates robust generalization, despite the inherently higher anatomical variability of the axial plane.

Table 1. Comparative performance metrics across models on the test set.

Model	Precision (%)	Recall (%)	Accuracy (%)	F1-Score (%)	Error (%)
MobileNetV2	100	100	100	100	0
YOLOv8n (Axial)	96.95	98.45	95.5	97.7	4.5
YOLOv8n (Coronal)	99.48	95.48	95.0	97.44	5

3.2 Quantitative Performance of the End-to-End Stone Detection Framework

The integrated pipeline demonstrated robust performance when evaluated on unseen CT images (see Fig. 2), achieving a perfect recall while maintaining a precision of 75.76%. This resulted in an overall accuracy of 84% for stone detection, with an error rate of 16%. The framework maintained this performance across both anatomical planes with no significant difference in detection rates between them.

The system showed particular strength in identifying true positive cases, as evidenced by the complete recall score, though some false positives were observed. The end-to-end processing time averaged under 1.35 seconds per image on standard GPU hardware, demonstrating clinical applicability.

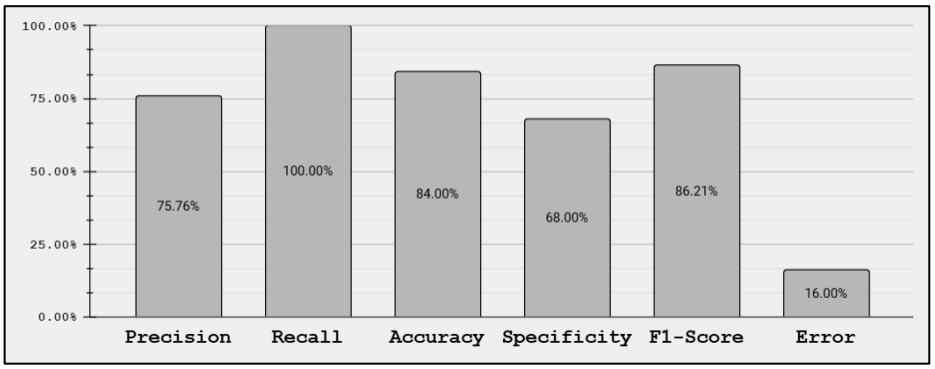


Fig. 2. Performance Metrics of the End-to-End Stone Detection Framework

To further illustrate the functionality of the proposed framework, Fig 3 presents a representative example of the entire pipeline in action. The left panel shows an axial CT image as input, while the right panel displays the final output after sequential processing: anatomical plane classification, kidney localization, and kidney stone detection.

Fig. 3. Visual Comparison of Original CT Input and Final Output Generated by the proposed End-to-End Framework.

4 Discussion

The proposed end-to-end framework demonstrated distinct performance characteristics across its components, reflecting their specialized roles in the diagnostic pipeline. MobileNetV2 achieved flawless performance in classifying slices. This reliability is critical for ensuring anatomically consistent processing in subsequent stages.

For kidney localization, YOLOv8n exhibited optimal (>95%) precision and recall, with marginally higher performance in coronal slices due to their more standardized

kidney morphology. The slight axial/coronal discrepancy aligns with known challenges in multi-planar renal imaging, where axial slices often exhibit greater anatomical variability [**20**].

The proposed end-to-end framework demonstrated perfect sensitivity in detecting renal calculi, ensuring no true positive cases were missed—a critical requirement for clinical deployment. However, the specificity revealed a conservative bias: false positives; This is mainly caused by the final stage, where hyperdense pixels (e.g., bone fragments or contrast) exceeded the adaptive threshold ($\mu + 2\sigma \times \alpha$) computed from the grayscale intensity of the kidney region.

Although effective at isolating stones, this method may misclassify residual bones, intense noise artefacts or isolated bright (specially in stone-free images). The lack of morphological or contextual validation in this step allows such false positives to occur.

While the inward 10-pixel margin reduced this effect, some residual bony structures at the kidney periphery remained challenging to exclude completely. This phenomenon explains the pipeline's 68% specificity and represents a key area for improvement. Future iterations could enhance specificity through anatomical context integration or advanced noise suppression techniques without compromising the system sensitivity.

5 Conclusion

This study presented a fully automated deep learning framework for renal calculi detection in abdominal computed tomography scans. The framework integrates anatomical plane classification using MobileNetV2, kidney localization via YOLOv8n detectors, and adaptive stone segmentation.

Our system achieved clinically critical 100% sensitivity in stone detection while maintaining reasonable precision (75.76%) and rapid processing times (<1.5s/image). The pipeline's perfect performance in kidney identification (95-99% accuracy across planes) and plane classification (100% accuracy) demonstrates the viability of deep learning for automating preliminary radiological assessments.

However, the overall accuracy and precision were affected by false positives, particularly in healthy kidneys, likely due to high-intensity tissue or bone fragments within the kidney region. This highlights a limitation of the framework, although it may be highly sensitive, false positives undermine its usefulness in confirming the disease [21].

These findings demonstrate that deep learning techniques can reliably automate kidney stone detection, offering potential support tools for radiologists. Nonetheless, future work should focus on improving specificity, possibly by incorporating more robust segmentation models such as U-Net architectures, refining thresholding strategies, and expanding the training dataset—especially for the detection components. Further validation with larger and more diverse clinical datasets is also recommended to enhance the generalizability and clinical applicability of the proposed framework.

References

1. [1] Zhang, L., Zhang, X., Pu, Y., Zhang, Y., Fan, J.: Global, Regional, and National Burden of Urolithiasis from 1990 to 2019: A Systematic Analysis for the Global Burden of Disease Study 2019. Clin. Epidemiol. 14, 971–983 (2022). https://doi.org/10.2147/CLEP.S370591
2. [2] Gottlieb, M., Long, B., Koyfman, A.: The evaluation and management of urolithiasis in the ED: A review of the literature. Am. J. Emerg. Med. 36(4), 699–706 (2018). https://doi.org/10.1016/j.ajem.2018.01.003
3. [3] Hyams, E., Matlaga, B.: Economic impact of urinary stones. Translational Andrology And Urology 3(3), 278–283 (2014). https://doi.org/10.3978/j.issn.2223-4683.2014.07.02
4. [4] Wollin, D.A., Gupta, R.T., Young, B., Cone, E., Kaplan, A., Marin, D., Patel, B.N., Mazurowski, M.A., Scales, C.D., Ferrandino, M.N., Preminger, G.M., Lipkin, M.E.: Abdominal radiography with digital tomosynthesis: an alternative to computed tomography for identification of urinary calculi? Urology 120, 56–61 (2018). https://doi.org/10.1016/j.urology.2018.06.041
5. [5] McCarthy, C.J., Baliyan, V., Kordbacheh, H., Sajjad, Z., Sahani, D., Kambadakone, A.: Radiology of renal stone disease. International Journal of Surgery 36, 638–646 (2016). https://doi.org/10.1016/j.ijsu.2016.10.045
6. [6] de Melo, J.A.C., Gelbcke, F.L., Amadigi, F.R., Huhn, A., da Silva, C., Ribeiro, G.: Psychological exhaustion of radiological nursing workers in nuclear medicine services. Rev. Bras. Enferm. 73(Suppl 1), e20200169 (2021). https://doi.org/10.1590/0034-7167-2020-0169
7. [7] Pinto-Coelho, L.: How Artificial Intelligence Is Shaping Medical Imaging Technology: A Survey of Innovations and Applications. Bioengineering (Basel) 10(12), 1435 (2023). https://doi.org/10.3390/bioengineering10121435
8. [8] Taye, M.M.: Theoretical understanding of convolutional neural network: concepts, architectures, applications, future directions. Computation 11(3), 52 (2023). https://doi.org/10.3390/computation11030052
9. [9] Cruz-Euán, V.H., Medina-Escobedo, M., Gutiérrez-Solís, A.L., Ávila-Nava, A., Ramírez-Jurado, A.A., González-Rocha, L.A., Lugo, R.: Concordancia de la ultrasonografía con urotomografía en el tamizaje y diagnóstico de urolitiasis en una población endémica. Rev. Mex. Urol. 79(4), e03 (2019). https://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S2007-40852019000400003
10. [10] Sánchez-Gutiérrez, M.E., González-Pérez, P.P.: Addressing the class imbalance in tabular datasets from a generative adversarial network approach in supervised machine learning. J. Algorithm. Comput. Technol. 17 (2023).
11. [11] Ultralytics: YOLOX: High Performance and Simplicity. https://docs.ultralytics.com/compare/yolov8-vs-yolox/#yolox-high-performance-and-simplicity, last accessed 2025/05/30.
12. [12] M. Tan and Q. V. Le, “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks,” arXiv preprint arXiv:1905.11946, 2019. https://arxiv.org/abs/1905.11946
13. [13] Yaseen, M.: What is YOLOv8: An in-depth exploration of the internal features of the next-generation object detector. arXiv preprint arXiv:2408.15857 (2024)
14. [14] More, G., Patil, O., More, O., More, M., Suryavanshi, S., Mali, M.: Comparison of Object Detection Algorithms CNN, YOLO and SSD. Int. J. Sci. R. Tech. 1(11), 137–144 (2024). https://doi.org/10.5281/zenodo.14186397
15. [15] Guerrero-Rangel, J.R.G., Sidorov, G., Maldonado-Sifuentes, C.E., Vargas-Santiago, M., Ortega-García, M.C., León-Velasco, D.A.: Natural Language Processing Approach Using a Neural Network Ensemble (CNN-HSNN) for Skin Cancer and Multi-Disease

Classification. Computación y Sistemas 28(3), 1243–1255 (2024). https://doi.org/10.13053/cys-28-3-5015
16. [16] Ruby, U., Yendapalli, V.: Binary cross entropy with deep learning technique for image classification. International Journal of Advanced Trends in Computer Science and Engineering, 9 (2020). https://doi.org/10.30534/ijatcse/2020/175942020
17. [17] Ultralytics: Model Evaluation Insights – Starting with a Higher Learning Rate. https://docs.ultralytics.com/es/guides/model-evaluation-insights/#starting-with-a-higher-learning-rate, last accessed 2025/05/30.
18. [18] Tajima, A., Bouisset, F., Ohashi, H., et al.: Advanced CT imaging for the assessment of calcific coronary artery disease and PCI planning. J. Soc. Cardiovasc. Angiogr. Interv. 3(3B), 101299 (2024). https://doi.org/10.1016/j.jscai.2024.101299
19. [19] Pushpa Singh, N. Singh, K.K. Singh, A. Singh: Diagnosing of disease using machine learning. In: Singh, K.K., Elhoseny, M., Singh, A., Elngar, A.A. (eds.) Machine Learning and the Internet of Medical Things in Healthcare, pp. 89–111. Academic Press, 2021. https://doi.org/10.1016/B978-0-12-821229-5.00003-3
20. [20] Takahashi, Y., Ohmoto-Sekine, T., Yoshida, M., Miyazaki, M.: Comparison of axial and coronal acquisitions by non-contrast-enhanced renal 3D MR angiography using flow-in time-spatial labeling inversion pulse. MAGMA 33(1), 95–102 (2020).
21. [21] Pascual Huerta, J.: Sensibilidad, especificidad y valores predictivos (Parte II). Revista Española de Podología 35(1), 69–70 (2024)