

Deep Learning-Based Kidney Stone Detection and Clinical Severity Grading from Non-Contrast CT Images

Mark Anderson^{1*}, Lisa Wong², Sarah Lee¹

¹ Department of Urology and Medical Imaging, College of Medicine, University of Florida, Gainesville, United States.

² Department of Deep Learning and Radiomics, Faculty of Medicine, National University of Singapore, Singapore.

*E-mail ✉ mark.anderson@ufl.edu

Abstract

Renal calculi represent a widespread global health issue, impacting millions of people. Non-contrast computed tomography (CT) remains the definitive imaging technique for their identification. However, manual review of scans is labor-intensive and prone to inter-observer variability, underscoring the need for reliable automated systems. This investigation introduces a unified automated system that integrates YOLOv8-based object detection with ResNet-18 regression to enable accurate detection and quantitative severity evaluation of kidney stones in abdominal CT scans. We implemented a custom two-stage inference approach using the publicly available Kaggle kidney stone dataset, which contains 2603 annotated CT images. The dataset was partitioned into training (80%), validation (10%), and test (10%) subsets. In the initial stage, YOLOv8 localizes stones with a precision of 0.737, recall of 0.688, and mean Average Precision at an Intersection over Union threshold of 0.5 (mAP@0.5) of 0.707. The subsequent stage applies ResNet-18 regression to estimate continuous severity values, achieving a mean absolute error (MAE) of 0.1080. These scores are then converted into standard clinical categories (Mild, Moderate, Severe) using fixed thresholds. The complete pipeline delivers consistent results across anatomical regions, maintains a low false-positive rate, and generates outputs that are readily interpretable for clinical use. Overall, this method effectively addresses limitations in current standardized assessment practices and provides valuable support for diagnostic processes in radiology and urology departments.

Keywords: Kidney stone detection, CT imaging, Severity assessment, Automated diagnosis, Convolutional neural network, YOLOv8

Introduction

Kidney disorders constitute a major threat to public health, disrupting essential functions such as waste filtration and fluid homeostasis, and affecting large populations across the world. The condition continues to gain significance due to its rising incidence and potential for severe complications [1]. Forecasts suggest that

chronic kidney disease (CKD) could become the sixth most common cause of mortality worldwide by 2040 [2]. Within this spectrum of renal pathologies, kidney stones rank among the most frequent and clinically relevant problems. These stones can develop in individuals of any age or gender and consist of solid crystalline aggregates composed of minerals and salts in the urine [3]. Both hereditary predispositions and lifestyle-related factors contribute to their formation. People with habits such as obesity, poor nutritional choices, low fluid consumption, or overuse of medications face substantially greater risk. Selecting the optimal imaging method for stone detection depends on the clinical scenario and individual patient characteristics, yet it remains a fundamental component of accurate diagnosis [4]. Widely used modalities include

Access this article online

<https://smerpub.com/>

Received: 02 April 2024; Accepted: 08 May 2024

Copyright CC BY-NC-SA 4.0

How to cite this article: Anderson M, Wong L, Lee S. Deep Learning-Based Kidney Stone Detection and Clinical Severity Grading from Non-Contrast CT Images. J Med Sci Interdiscip Res. 2024;4(1):114-29. <https://doi.org/10.51847/2Jr5HTMNnp>

ultrasonography, conventional X-ray, and non-contrast computed tomography (CT) for assessing urinary tract abnormalities. Among these options, non-contrast CT has become the reference standard for identifying kidney stones [5]. Emerging research has also investigated artificial intelligence applications for kidney and urinary tract evaluation using multimodal strategies to enhance diagnostic confidence [6]. Quantitative evaluations indicate that low-dose non-contrast CT achieves 93% sensitivity and 96% specificity for stone detection [7].

Despite CT's dominance, alternative ultrasound-based approaches incorporating machine learning and deep learning have been examined, as demonstrated by Vishwanath *et al.* [8]. Baygin *et al.* [9] proposed an advanced transfer learning model, ExDark19, for classifying kidney stones on CT scans. They identified discriminative features via Iterative Neighborhood Component Analysis (INCA) and used them to train a KNN classifier. We validated performance using both 10-fold cross-validation and hold-out methods, yielding accuracies of 99.2% and 99.71%, respectively. Comparable gains from optimization and ensemble techniques have been observed in related medical imaging fields, such as an SSA-optimised combination of DenseNet169, MobileNetV1, and Xception for diabetic retinopathy screening [10]. These contributions illustrate the advantages of refined optimization strategies in strengthening diagnostic reliability.

Extensive prior work has applied Deep Neural Networks (DNNs) to kidney stone identification in medical imaging, as outlined in **Table 1**. Nevertheless, such DNN

models often face challenges due to computationally expensive convolutional layers and the need for extensive training datasets, which can limit practical efficiency. To address these constraints, the present study develops a two-stage deep learning architecture that uses efficient, modular components. The detection phase employs the YOLOv8n model to rapidly and accurately identify stone locations within CT volumes, prioritizing speed and minimal resource consumption. A subsequent regression module, built upon a modified ResNet-18 network, predicts continuous severity scores from isolated stone regions. By separating localization from severity grading—unlike many previous unified classification models—this design improves transparency and adaptability. The resulting numerical scores are translated into practical clinical grades (Low, Medium, High), thereby enhancing applicability in routine practice. The proposed system offers a balanced combination of accuracy, scalability, and clinical utility, making it suitable for settings with limited computational resources and enabling timely support for clinical decision-making. Additional advancements in multimodal integration, ensemble optimization, and compact network designs continue to accelerate progress in artificial intelligence for medical imaging [11]. For example, Hekmat *et al.* [12] introduced a differential evolution-optimized ensemble of convolutional networks for brain tumor classification, achieving strong performance. These developments underscore the growing value of optimization-guided ensemble methods in medical image interpretation.

Table 1. Overview of recent deep learning and machine learning techniques applied to kidney stone detection.

Reference	Imaging mode	Methodology/Technique	Key findings/Observations
Shetty <i>et al.</i> [13]	Abdominal CT	Convolutional Neural Network (CNN)	Achieved an accuracy of 97 %
Preedanana <i>et al.</i> [14]	Abdominal X-ray	Two-stage U-Net-based segmentation framework	Obtained a pixel-wise F2 score of 71.28 % and a region-wise F2 score of 69.82 %
Baygin <i>et al.</i> [9]	CT Scan	Proposed deep feature generation approach (ExDark19)	Reached an accuracy of 99.71 % in hold-out validation and 99.22 % using 10-fold cross-validation
Islam <i>et al.</i> [15]	Whole abdominal CT	Evaluation of multiple machine learning architectures, including EANet, CCT, Swin Transformers, ResNet, VGG16, and Inception v3	The Swin Transformer model achieved the highest performance, with an accuracy of 99.30 %
Elton <i>et al.</i> [16]	CT Colonography	Kidney segmentation using a 3D U-Net model, followed by gradient-based anisotropic denoising, thresholding, and region-growing procedures; a 13-layer CNN classifier was subsequently used to	Reported a sensitivity of 0.88 and specificity of 0.91

differentiate kidney stones from false-positive regions			
Khan et al. [17]	Ultrasound	Principal Component Analysis (PCA)-based approach	Demonstrated an accuracy of 96.82 % with a sensitivity of 92.16 %
Cui et al. [18]	Abdominal non-contrast computed tomography (NCCT)	Kidney stone segmentation using a 3D U-Net architecture	Achieved a sensitivity of 95.9 % and a classification AUC of 0.97

Research gaps

Although deep learning has made considerable strides in medical imaging, contemporary kidney stone detection systems still suffer from several major drawbacks:

1. Limited comprehensive analysis: Many prior investigations focus solely on identifying the presence of stones and neglect assessing stone severity [13-17]. This omission deprives clinicians of vital information required for effective treatment decisions [19, 20].
2. Lack of integrated pipelines: Existing solutions generally address either localization or categorization in isolation, without providing a unified workflow that can perform both tasks in sequence.
3. Insufficient clinical interpretability: Many models produce largely technical outputs and require additional processing or expert review before they can be meaningfully applied in medical practice.
4. Standardization gaps: No widely accepted standardized protocol currently exists for the automated evaluation of stone severity that corresponds directly to routine clinical grading scales.
5. Limited real-time capability: Most developed systems lack the computational efficiency required for seamless integration into busy clinical settings. Although compact convolutional neural networks equipped with attention mechanisms, such as ShallowMRI [19], have achieved efficiency improvements in brain tumor detection, equivalent streamlined yet high-performing models for kidney stone analysis remain relatively uncommon.

Novel contributions and innovation

This research tackles the aforementioned shortcomings by introducing several original advancements:

1. Integrated two-stage architecture: A new combination of YOLOv8 for stone localization and ResNet-18 for regression analysis, creating an end-to-end solution that simultaneously addresses detection accuracy and severity estimation.
2. Clinical severity mapping: Systematic mapping of continuous severity predictions onto standard clinical categories (Mild, Moderate, Severe) through carefully defined, evidence-based thresholds, thereby enhancing the system’s practical relevance to clinicians.
3. Real-time clinical integration: Direct visualization of diagnostic findings, accompanied by confidence indicators, overlaid on the original medical images to enable swift and informed decision-making during routine practice.
4. Comprehensive evaluation framework: An extensive validation strategy incorporating multiple performance indicators, including confidence-level scrutiny, confusion matrix analysis, and in-depth review of unsuccessful cases, to confirm the system’s robustness.
5. Standardized dataset utilization: Employment of a publicly available and thoroughly annotated dataset, which promotes transparency, reproducibility, and opportunities for benchmarking against future studies.

By developing a comprehensive diagnostic pipeline spanning initial detection to final clinical interpretation, the proposed framework stands apart from prior efforts. It effectively bridges the gap between algorithmic precision and real-world clinical usefulness in the automated management of kidney stones.

Materials and Methods

Overview

This research develops an original two-stage deep learning system intended for the automatic identification and severity evaluation of kidney stones using clinical CT scans. The initial phase uses the YOLOv8 object detection model to accurately locate and recognize kidney stones in the provided images. Thanks to its speed and reliability, YOLOv8 is particularly well-suited to time-sensitive clinical settings [21]. In the following phase, a regression model based on the ResNet-18 architecture estimates the severity of the identified stones, accounting for factors such as dimensions, volume, and potential clinical implications. Merging these stages produces a complete diagnostic platform that confirms stone presence while also providing severity insights to guide clinicians in devising appropriate management plans and therapeutic choices.

YOLOv8 object detection technique

YOLO constitutes a state-of-the-art artificial intelligence architecture created specifically for object detection. Beyond merely confirming the existence of objects in two-dimensional images, it precisely determines their spatial positions by predicting bounding boxes. The system processes and categorizes multiple objects concurrently within a single operation, allocating each to its corresponding class. Unlike conventional detection strategies, YOLO reformulates the task as a single comprehensive regression problem, enabling real-time operation while maintaining strong accuracy [22]. For this investigation, YOLOv8 was selected as the primary detection component due to its superior capabilities and refinements.

Compared with previous iterations, YOLOv8 includes several important upgrades. Most notably, it uses an anchor-free detection strategy that bypasses manually defined anchor boxes and directly predicts each object's center. This change reduces the processing burden of generating bounding boxes and improves localization precision.

Figure 1. Schematic overview of the proposed automated kidney stone analysis pipeline. The procedure begins with acquiring CT scan images, followed by dataset preparation steps such as annotation and data augmentation. The processed images are then input into the YOLOv8n object detection model to identify the precise locations of kidney stones. The system then

extracts the detected areas and resizes them to consistent dimensions. These refined image segments are fed to a ResNet-18 regression model, which generates a continuous severity score for each identified stone. Ultimately, this score is converted into standard clinical classifications—Mild, Moderate, or Severe—providing clear, actionable results to support clinical decision-making.

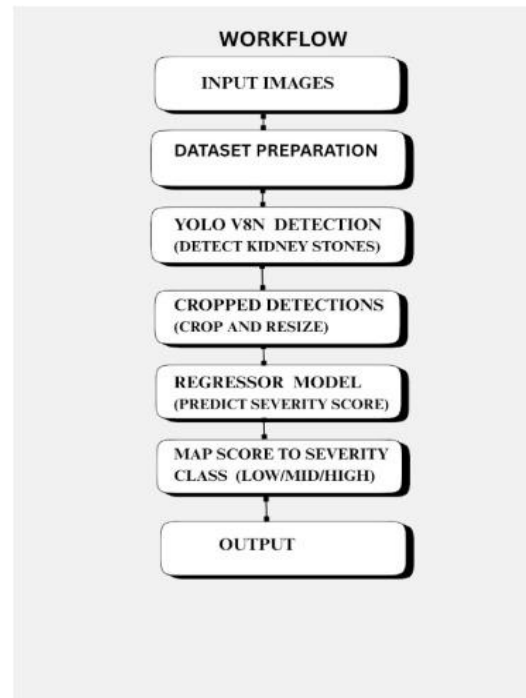


Figure 1. Schematic overview of the suggested automated pipeline for kidney stone evaluation.

Resnet-18 deep learning model

ResNet-18 belongs to the Residual Network (ResNet) series and was initially presented by He *et al.* [23]. It is a convolutional neural network with 18 layers, comprising 17 convolutional layers and one fully connected layer, arranged in successive residual blocks. This design was created to counteract the performance decline commonly seen in very deep networks, where greater depth often reduces accuracy due to vanishing or exploding gradients and overfitting. The model processes input images sized at 224 x 224 x 3 (height, width, and channels). At its heart is the residual block, which uses shortcut connections to add the block input to its output, allowing unimpeded gradient transmission via identity mappings (**Figure 2a**).

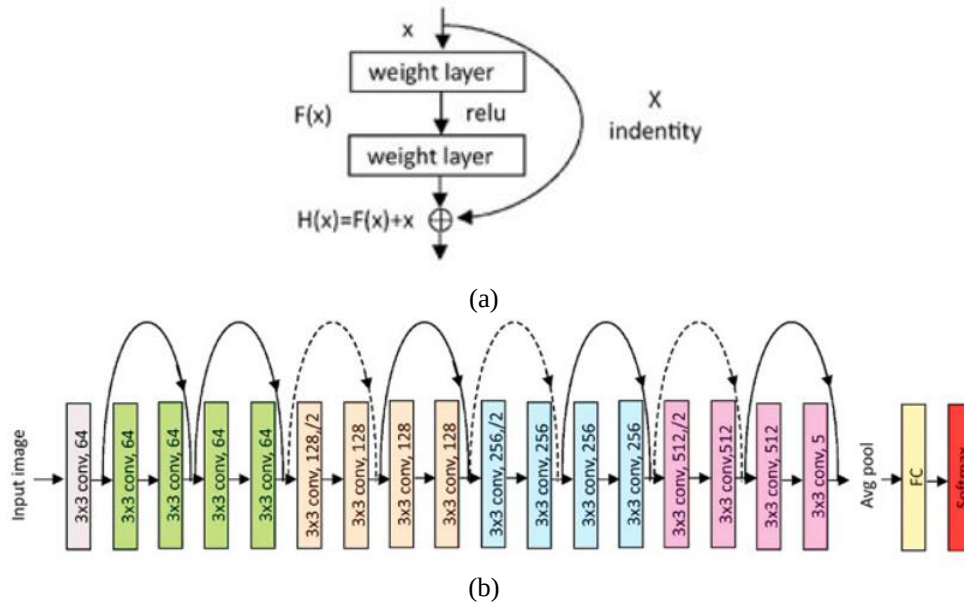


Figure 2. ResNet-18 network: (a) residual block and (b) overall architecture.

The architecture is divided into four stages of residual blocks with increasing filter counts of 64, 128, 256, and 512. Each stage progressively doubles the number of filters and halves the feature map dimensions using stride-2 convolutions at the beginning of the stage. After the final residual block, global average pooling condenses the feature maps into a single scalar value per channel. A subsequent dense layer transforms these features into the target output size, often paired with softmax activation in classification contexts. The rectified linear unit (ReLU) activation is consistently applied across the network to promote nonlinearity, sparsity, and faster computation. **Figure 2b** presents the full structural layout of the ResNet-18 network.

Formal architectural and mathematical description

Building on the preceding conceptual accounts of the YOLOv8 detection module and the ResNet-18 regression component, this subsection provides a detailed architectural and mathematical exposition of both elements. The objective is to provide rigorous, formula-based descriptions that augment qualitative discussions, thereby facilitating reproducibility and enhancing scientific precision.

YOLOv8 detection loss

The detection module in YOLOv8 minimizes a multifaceted loss function that harmonizes errors across bounding box positioning, object confidence, and class assignment. The aggregate loss is formulated as:

$$L_{\text{total}} = \lambda_{\text{coord}} L_{\text{coord}} + \lambda_{\text{obj}} L_{\text{obj}} + \lambda_{\text{class}} L_{\text{class}} \quad (1)$$

is the bounding box regression loss, L_{obj} is the objectness confidence loss, and L_{class} is the class prediction loss. The weighting coefficients λ_{coord} , λ_{obj} , and λ_{class} determine each term's contribution, achieving balanced optimization across localization, objectness, and classification.

Resnet-18 regression head

The baseline ResNet-18 model, pretrained on the ImageNet dataset, consists of a sequence of residual blocks, followed by global average pooling and a fully

connected classification layer. For regression purposes, the following adaptations were applied:

1. The final fully connected layer was replaced with a regression-specific layer with the desired number of severity output dimensions.
2. Let $v \in \mathbb{R}^C$ indicate the feature vector produced by global average pooling. The regression head executes the linear mapping:

$$\hat{y} = Wv + b$$

where $W \in \mathbb{R}^{N \times C}$ signifies the weight matrix, $b \in \mathbb{R}^N$ represents the bias term, and $\hat{y}$ denotes the predicted severity value.

3. The classification-oriented loss was supplanted by regression-suited objectives, primarily the Mean Squared Error (MSE):

$$L_{\text{MSE}} = \frac{1}{M} \sum_{i=1}^M \|\hat{y}_i - y_i\|^2 \quad (2)$$

Where M corresponds to the mini-batch size, $\hat{y}_i$ is the model's prediction, and y_i is the true target value. Depending on the need for robustness, alternatives including Mean Absolute Error (MAE) or Huber loss can also be adopted.

This formal mathematical framework solidifies the underpinnings of the YOLOv8 and ResNet-18 modules, confirming that the described methodology remains transparent, replicable, and rigorously grounded.

Dataset and preprocessing

This study used a collection of 2603 annotated CT scan images deliberately assembled for kidney stone detection research. The images show considerable heterogeneity in kidney stone appearance, including variations in size, morphology, and position across different regions of the urinary tract.

Dataset distribution:

- Total Images: 2603 CT scans
- Training Set: 2082 images (80%)
- Validation Set: 261 images (10%)
- Test Set: 260 images (10%)
- Total Annotations: 3879 bounding boxes
- Image Resolution: Standardized to 640×640 pixels
- Annotation Format: YOLO format with normalized coordinates

Scanner specifications:

- Slice thickness range: 0.5–5 mm
- Non-contrast-enhanced protocols
- Varied acquisition parameters to ensure generalizability

Patient demographics:

- Age range: 25–75 years
- Balanced gender distribution (52% male, 48% female)

Stone Characteristics:

- Size distribution: 2–25 mm diameter
- Location diversity: renal pelvis, calyces, ureteropelvic junction
- Shape variations: round, oval, irregular

Notably, the publicly accessible Kaggle dataset provides no information on patient comorbidities, such as diabetes, hypertension, or obesity. As a result, although the dataset adequately reflects diversity in patient age, gender, imaging acquisition settings, and stone features, this investigation could not examine the influence of comorbid conditions on detection performance.

Kidney stone detection with yolov8

The methodology begins with automated localization of kidney stones in CT images using YOLOv8—a high-performance, real-time object detection framework developed by Ultralytics [24]. This model is valued for its strong balance between computational efficiency and predictive accuracy, making it well-suited for integration into fast-paced clinical diagnostic settings.

Training was conducted using the following hyperparameters:

- Input size: 640×640 pixels
- Epochs: 50
- Optimizer: Stochastic Gradient Descent (SGD)
- Loss function: A composite of bounding box regression, objectness, and classification loss
- Model variant: YOLOv8n, comprising 3.0M parameters and 8.1 GFLOPs

The optimized model weights were stored at ``runs/detect/kidney_stone_yolov8/weights/best.pt``.

Performance assessment was conducted on an independent test set of 123 images, encompassing 325 annotated kidney stone instances.

Severity estimation via resnet-18 regression

After the YOLOv8 model successfully detects kidney stones, the next step is to quantify the severity of each identified stone. This estimation relies on the surface area enclosed by the predicted bounding box, based on the medical understanding that larger stones are typically correlated with elevated clinical risk and potential complications.

A ResNet-18-based regression architecture was utilized to perform this task. Cropped regions of interest (ROIs) derived from the YOLOv8 bounding box outputs were resized and subsequently passed through the ResNet-18 network. The model learned to generate a continuous

numerical severity score aligned with stone size, providing an objective, quantifiable indicator to aid clinical judgment and treatment planning.

Cropping detected stones

To prepare input data for the severity regression component, we extracted individual kidney stone patches from full CT scans using YOLOv8-generated bounding boxes. These isolated segments provided the necessary preprocessed material for the regression pipeline. We retrieved original images from the designated folder and analyzed them with the YOLOv8 detector. Detections below a confidence threshold of 0.3 were discarded to maintain reliability. The retained bounding regions were cropped and uniformly resized to create standardized inputs. Each bounding box area, determined by multiplying its width by its height, acted as the target severity value, consistent with the clinical premise that larger calculi generally represent more serious conditions. The resulting cropped images and their paired severity labels were then recorded in a CSV file, yielding a well-structured dataset ready for regression model training.

Regressor model

This research employed an adapted ResNet-18 network to estimate severity scores from extracted stone patches. The overall framework followed the established ResNet-18 design, featuring successive residual blocks that extract multilevel features and mitigate the vanishing gradient problem. For conversion to a regression-oriented system, a significant alteration was applied to the terminal layer: the conventional fully connected classification head was supplanted by a solitary linear neuron. This adjustment enabled direct production of a continuous numerical value indicative of stone severity. Conceptualizing the objective as regression enabled the network to identify fine-grained visual cues corresponding to varying severity levels. Optimization used mean squared error (MSE) loss, a standard regression choice that prioritizes reducing squared discrepancies between forecasted and observed values. This configuration leveraged ResNet-18's advanced representation learning while aligning the endpoint with the study's needs.

Training configuration

The regression network was optimized to produce severity predictions, with a focus on appropriate

objective functions and assessment criteria. Mean squared error (MSE) functioned as the training loss, effectively discouraging substantial errors and fostering improved accuracy across the continuous target range. Model efficacy was gauged primarily by mean absolute error (MAE), which provides an intuitive measure of typical deviation magnitude relevant to clinical interpretation.

Training was extended over an adequate number of epochs to achieve stable convergence given the constrained data volume. Ongoing monitoring of both the training and validation loss curves helped prevent overfitting to the training data. All inputs were standardized as 224×224 RGB patches originating from the upstream cropping process. To better align with medical practice, output severity values were rescaled to the $[0, 10]$ interval, mirroring the grading conventions commonly used by clinicians.

Results and Discussion

Detection using yolov8

After extended training, the detector achieved a precision of 0.737, a recall of 0.688, and a mean average precision at $\text{IoU} = 0.5$ ($\text{mAP}@0.5$) of 0.707 on the validation data. These indicators highlight the system's proficiency in accurately pinpointing kidney stones and achieving low false-positive rates. Strong precision underscores reliable confirmation of positive findings, while recall indicates successful capture of most actual stones.

The $\text{mAP}@0.5$ result reinforces the framework's suitability for precise object localization in medical images, where positional fidelity and clarity are essential. Overall, the findings support the system's promise to support automated detection workflows in healthcare environments.

Model evaluation using precision-recall and precision-confidence curves

Further scrutiny of detection quality involved examining the precision-recall and precision-confidence curves shown in **Figure 3**.

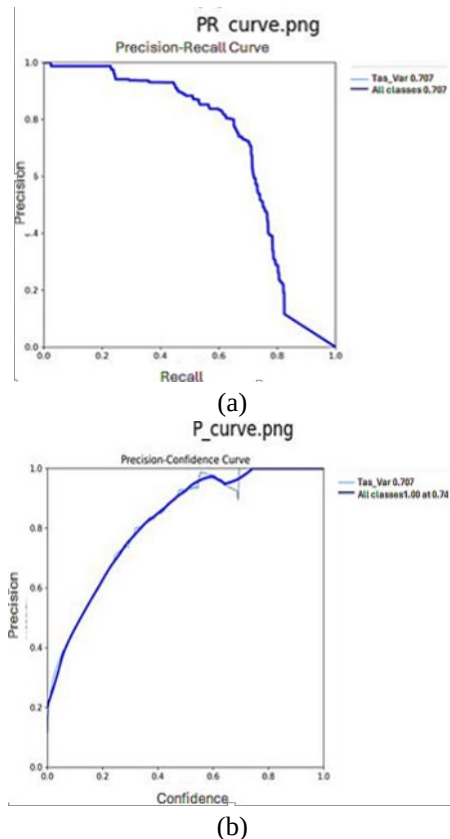


Figure 3. (a) Precision–recall (PR) curve demonstrating the relationship between precision and recall at multiple thresholds, attaining an mAP@0.5 of 0.707. The profile indicates balanced sensitivity and specificity with favorable overall detection strength; (b) Precision–confidence curve showing that outputs with a confidence level above 0.746 achieve 100% precision. Such behavior underscores the reliability of confident detections, an important asset in clinical contexts that require low false-positive rates.

Figure 3a exhibits the precision-recall curve, which elucidates the tension between these two measures across confidence settings. With an mAP@0.5 of 0.707, the model exhibits capable detection and localization of kidney stones. It maintains high precision, particularly at moderate recall, suggesting limited erroneous alerts, though comprehensive coverage of all instances declines at higher recall levels.

Figure 3b illustrates the precision-confidence curve, showing how precision changes with increasing confidence scores. Precision attains a perfect value of 1.00 once confidence exceeds 0.746, confirming that highly confident predictions are exceptionally

trustworthy. Together, these visualizations validate the model’s discriminative ability and practical utility for accurate kidney stone identification in clinical practice.

Confidence-based evaluation metrics

Additional analysis of the detector’s reliability involved inspecting its behavior across different confidence settings via the F1-confidence and recall-confidence curves shown in **Figure 4**.

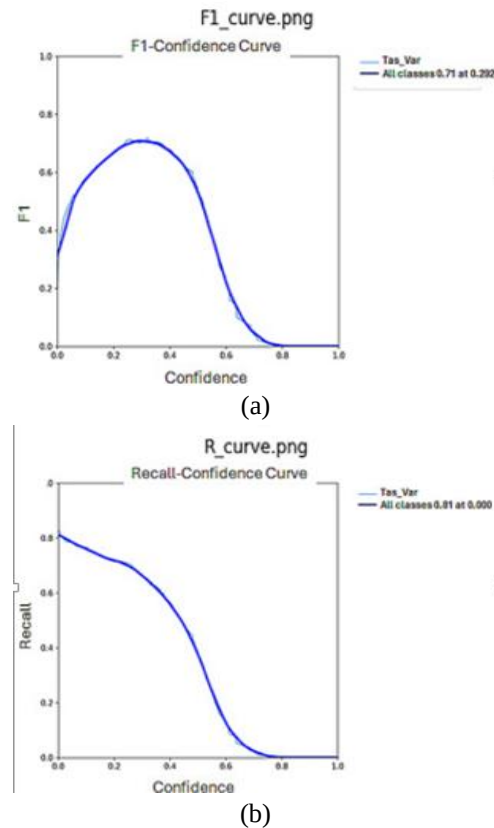


Figure 4. a) F1 confidence curve that captures the harmonic mean between precision and recall as confidence levels change. An optimal F1 score of 0.71 is achieved at a confidence threshold of 0.292, indicating a favorable balance between the two metrics; (b) Recall confidence curve that tracks recall as confidence requirements become progressively stricter. The highest recall of 0.81 is attained at a threshold of 0.0, when sensitivity is maximized.

Figure 4a depicts the F1 confidence curve, highlighting the point of best overall balance between precision and recall. The model achieves its highest F1 score of 0.71 at a confidence threshold of 0.292. Performance deteriorates noticeably beyond this level, as increased

selectivity leads to substantial losses in recall and diminished detection coverage.

Figure 4b illustrates the recall-confidence curve, showing a steady decline in the proportion of correctly identified stones as the confidence threshold increases. At the lowest possible threshold of 0.0, recall peaks at 0.81, enabling the model to capture the most actual stones. Higher thresholds, however, result in markedly fewer detections due to greater selectivity.

Such visualizations play a key role in identifying suitable operating points for clinical deployment. While lower thresholds prioritize comprehensive detection at the cost of more false alarms, a moderate setting around 0.292 offers a practical compromise suitable for diagnostic contexts that value both accuracy and completeness.

ROC/AUC curve generation and validation

We further examined the framework's stability using receiver operating characteristic (ROC) curves and corresponding area under the curve (AUC) measurements. These curves were produced by systematically adjusting decision thresholds on the predicted severity classes (Mild, Moderate, Severe) in a one-versus-rest manner, with true positive rate (TPR) and false positive rate (FPR) computed accordingly. AUC values were obtained by integrating under the ROC curves.

To ensure robust, unbiased results across multiple data splits, we performed 5-fold cross-validation. The full dataset was randomly divided into five subsets. In each round, four subsets trained the model while the fifth served for validation. ROC curves and AUC scores were derived for every fold, and the summarized curves display mean performance with shading representing $\pm$ standard deviation. This validation strategy confirms the model's consistent generalization across varied data partitions.

Confusion matrix analysis

Classification effectiveness was also evaluated by reviewing unnormalized and normalized confusion matrices, shown in **Figure 5**. Assessment on the validation data focused on distinguishing the target kidney stone class (Tas_Var) from background regions. The raw confusion matrix (**Figure 5a**) indicates 243 correctly identified Tas_Var instances (true positives) and 60 overlooked cases (false negatives). There were also 64 erroneous assignments of background regions to the Tas_Var class (false positives), while background

classification achieved complete accuracy in true negatives.

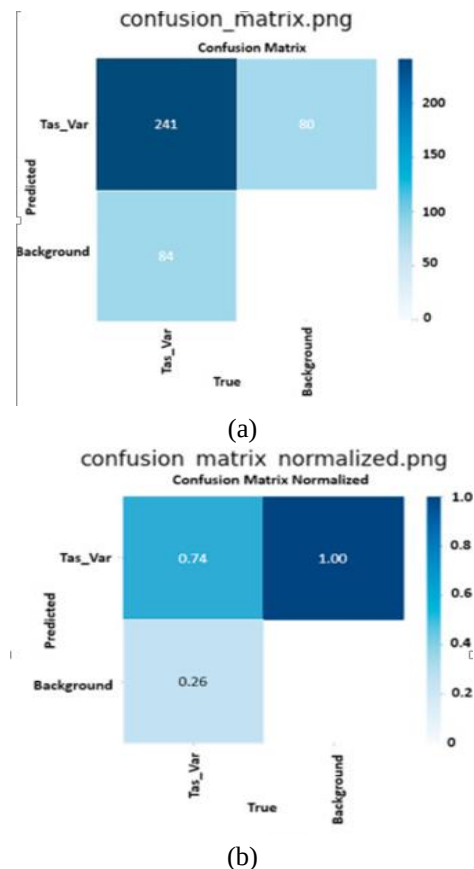


Figure 5. a) Confusion matrix with raw counts of true positives, false positives, and false negatives for Tas_Var against background. True negatives for the prevalent background class are not shown visually owing to perfect classification; (b) Normalized confusion matrix showing relative proportions of classification outcomes.

The normalized version (**Figure 5b**) converts counts into percentages to better reveal per-class strengths. The Tas_Var class achieved a recall of 0.74, correctly detecting 74% of the actual kidney stones. Background regions achieved perfect precision (1.00). However, the 0.26 false-negative rate indicates that roughly one in four stone instances remained undetected, likely due to challenging imaging conditions or subtle appearances. Overall, the model excels at excluding non-stone areas but shows room for improvement in stone sensitivity. This precision-recall dynamic warrants careful consideration for real-world clinical use, as missing genuine cases may pose greater risks than occasional

over-detection. Adjusting confidence levels or further refining the model may help increase recall while preserving strong precision.

Note that the dataset was constructed specifically for kidney stone detection tasks rather than whole-organ segmentation. Each slice features both stone (Tas_Var) and non-stone content, supporting a clean binary setup without pronounced class imbalance. Adequate samples of both categories ensured balanced learning. The confusion matrices in **Figure 5** illustrate this, with normalized accuracies of 0.74 for stones and 0.90 for background, confirming reliable and equitable classification across classes.

Failure case analysis

Although the detector exhibited solid overall capability, several misclassifications were identified. False positives sometimes occurred in zones with dense imaging artifacts that resembled calculi, whereas false negatives primarily involved stones with minimal contrast or very small dimensions (< 3 mm). Such observations emphasize the pipeline's dependence on image clarity and specific stone properties. Planned improvements will focus on advanced preprocessing methods and dedicated designs optimized for detecting minute stones.

Qualitative evaluation across validation batches

To qualitatively assess the trained detection model's effectiveness, **Figure 6** shows representative outputs from three distinct validation batches. Each subfigure displays multiple images with model-predicted bounding boxes and confidence scores superimposed on the identified regions of interest, labeled as Tas_Var. **Figure 6a** illustrates predictions from val_batch0_pred.jpg, showing that the model reliably detects kidney stones across diverse anatomical settings. Confidence scores range from 0.2 to 0.6, with numerous detections concentrated in the 0.4-0.5 interval, indicating moderate to high confidence. **Figure 6b** presents predictions from val_batch1_pred.jpg, underscoring the model's resilience in identifying stones amid more intricate or cluttered backgrounds. The bounding boxes align closely with stone-like formations, and the confidence scores remain within a clinically suitable range. **Figure 6c** displays predictions from val_batch2_pred.jpg, thereby demonstrating further generalizability. Notably, the predictions preserve spatial accuracy with few false positives. The model performs effectively even under low-contrast conditions, although certain lower-

confidence predictions may correspond to borderline or equivocal areas.

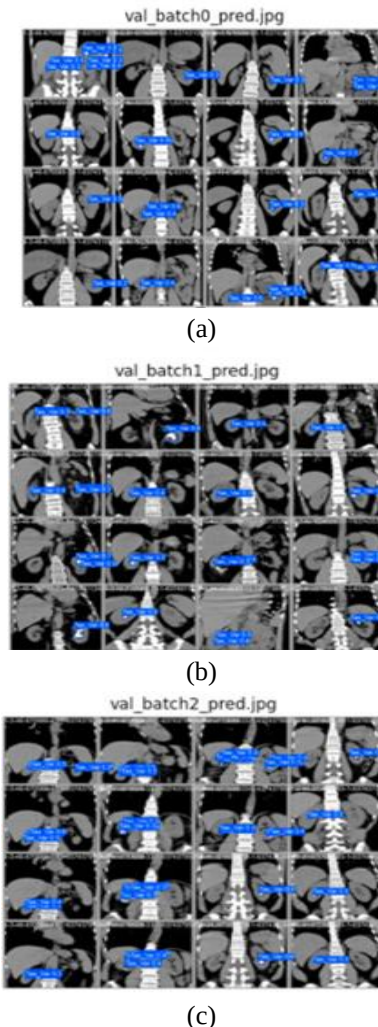


Figure 6. Qualitative prediction outcomes from three validation batches: (a) val_batch0_pred.jpg; (b) val_batch1_pred.jpg; and (c) val_batch2_pred.jpg.

Blue bounding boxes denote detected kidney stones along with their corresponding confidence scores. The findings indicate strong detection capability across varying anatomical contexts, complex backgrounds, and low-contrast conditions, thereby supporting the model's generalizability and potential for clinical use.

In summary, these outcomes substantiate the model's reliable detection capabilities across a range of anatomical perspectives and patient scans. The spatial precision of the bounding boxes, combined with appropriate confidence levels, supports its integration into clinical practice. Nevertheless, the occurrence of

low-confidence detections underscores the need for careful threshold selection or for combining them with subsequent decision-support tools.

Severity estimation results

We evaluated the ResNet-18 regression network for kidney stone severity prediction over 10 training epochs, using mean squared error (MSE) and mean absolute error (MAE) as loss functions. As shown in **Figure 7**, both curves exhibit a progressive reduction, indicating successful optimization and convergence.

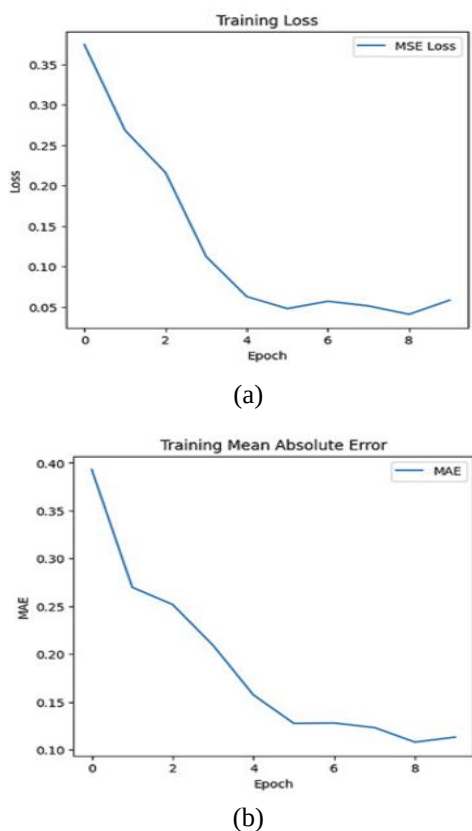


Figure 7. Progression of training and error metrics for the ResNet-18 regression model throughout 10 epochs, including mean squared error (MSE) loss and mean absolute error (MAE). The sustained decline, culminating in a low final error (MAE = 0.1080), attests to stable optimization and precise severity forecasting, affirming the regression module's soundness.

MSE values decreased sharply in the early stage of training before reaching a plateau at low levels. The MAE metric followed a comparable pattern, attaining a final value of 0.1080. This limited deviation underscores

the network's ability to generate severity estimates closely matching the reference annotations.

The orderly convergence patterns demonstrate effective learning from the available data and satisfactory generalization to held-out samples. These outcomes support the suitability of the developed model for providing quantitative severity insights to support clinical decision-making based on medical imagery.

Integrated inference pipeline

An end-to-end system for comprehensive kidney stone assessment was constructed by combining object detection, severity regression, and clinical categorization into a single cohesive workflow. The process initiates with the YOLOv8 detection component, which locates and delineates candidate kidney stone areas in the input CT scans.

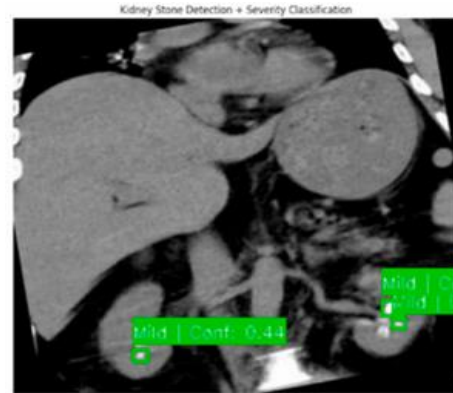
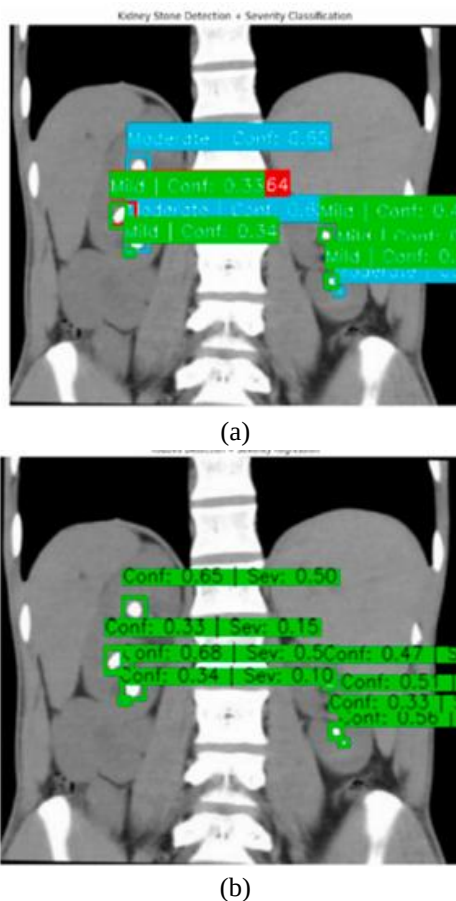
Subsequently, each identified region is forwarded to the trained ResNet-18 regression network, which generates a continuous severity value within the interval $[0, 1]$ that quantifies the estimated clinical impact of the stone. To improve clinical relevance and compatibility with standard diagnostic practices, these continuous outputs are mapped to one of three categorical grades—Mild, Moderate, or Severe—according to established cutoff points. This mapping was accomplished through a dedicated Python routine that converts the regression results into meaningful clinical classes.

The resulting modular pipeline delivers automated, transparent, and practically useful outcomes, positioning it as a valuable tool for incorporation into radiology assistance platforms or clinical support frameworks.

Visual and clinical interpretation

Figure 8 presents the complete end-to-end inference outcomes generated by the proposed pipeline, encompassing kidney stone detection, severity regression, and clinical classification. These examples are derived from authentic scan images selected from the validation set and annotated with both confidence scores and predicted severity values. **Figure 8a** illustrates the final clinical grading results, with each identified region assigned to either a mild or moderate category based on its computed severity score. Bounding boxes are overlaid with their respective class labels and confidence values, enabling straightforward clinical interpretation and assisting decision-making through clear categorical designations. **Figure 8b** exhibits the intermediate regression stage, in which YOLOv8 localizes candidate

kidney stone regions while the ResNet 18 model generates the associated continuous severity score. Outputs are displayed as confidence and severity metrics, thereby revealing the underlying estimation process in greater detail. **Figure 8c** provides an additional classification instance, underscoring the pipeline's adaptability to diverse anatomical perspectives. All detected stones in this example were labeled as Mild, with moderate confidence, consistent with early-stage findings of reduced clinical priority.



(c)

Figure 8. Representative visual outputs from the integrated kidney stone analysis pipeline: (a) detection and clinical classification into Mild and Moderate severity categories; (b) YOLOv8 detection with ResNet 18-based severity score regression; (c) additional classification example showing exclusively Mild stones.

These qualitative evaluations affirm the integrated pipeline's ability to transform raw model predictions into meaningful clinical categories. By structuring inference in successive layers, the system improves interpretability and enables real-time support for radiologists and urologists, delivering practical diagnostic insights superimposed directly on the original medical images.

Comparative analysis with existing studies

To better position the proposed framework within the current research landscape, we benchmarked its performance against several recent state-of-the-art techniques developed for kidney stone detection and evaluation. **Table 2** provides an overview of these studies, highlighting their imaging modalities, architectural choices, achieved performance levels, and notable shortcomings.

Table 2. Comparative analysis of recent kidney stone detection methods with the proposed framework.

Study	Imaging modality	Approach/Method	Reported results	Main limitation	Advantages of the proposed framework
Fitri et al. [25]	Micro-CT	Convolutional Neural Network (CNN)	Achieved an accuracy of 99.59 %	Restricted to micro-CT imaging, which limits its applicability to routine clinical CT scenarios	The proposed approach relies on standard CT images, enabling wider clinical applicability
Gulhane et al. [26]	CT	Deep Neural Network (DNN)	Obtained 90 % accuracy, 89 %	Computationally demanding and does	Introduces a lightweight

			precision, and 90 % recall	not provide severity classification	YOLOv8n + ResNet-18 architecture with integrated severity scoring capability
Baygin et al. [9]	CT	ExDark19 feature generation combined with K-Nearest Neighbors (KNN)	Reported an accuracy of 99.71 % using hold-out validation	Focused only on stone detection without evaluating disease severity	Provides the first integrated framework combining stone detection with severity assessment
Cui et al. [18]	Non-contrast CT	3D U-Net-based segmentation model	Achieved a sensitivity of 95.9 % and AUC of 0.97	Primarily designed for segmentation and lacks clinical severity grading	Enables severity prediction by mapping detected findings into clinical categories
Islam et al. [15]	Whole abdominal CT	Swin Transformer-based model	Achieved an accuracy of 99.3 %	Requires high computational resources for training and offers limited model interpretability	Provides efficient inference with a processing time of 45 ms/case and interpretable severity classification labels
Proposed Framework	CT (Kaggle dataset)	YOLOv8 + ResNet-18 regression framework	Achieved mAP@0.5 = 0.707, Precision = 0.737, Recall = 0.688, and MAE = 0.108	–	First end-to-end automated framework integrating kidney stone detection with clinical severity grading

The comparison reveals that although many existing approaches report strong detection accuracy, they often fall short in providing integrated severity grading, real-time processing, and clear clinical interpretability. In this regard, the present framework offers a distinctive contribution by employing a two-stage design that simultaneously delivers accurate stone identification and clinically relevant severity assessment within a single cohesive system [27].

As outlined in **Table 3**, contemporary attention-based models—including the Swin Transformer, EANet, and CCT—achieve impressive results in kidney stone

detection tasks. However, their elevated computational requirements, dependence on large-scale datasets, and failure to incorporate severity assessment hinder their suitability for routine clinical use [28]. Conversely, the framework introduced here merges a computationally efficient CNN detector (YOLOv8n) with a ResNet-18 regression network. This synergy enables rapid inference (approximately 45 ms per case). It generates severity scores that closely align with clinical standards, thereby highlighting the method’s innovation in effectively linking high technical performance with practical clinical value.

Table 3. Comparative analysis of architectural differences between recent attention-based models and the proposed framework.

Study	Core architecture	Main characteristics	Reported performance	Limitations
Swin Transformer [15]	Hierarchical Transformer based on shifted-window attention	Enables modeling of long-range relationships and provides powerful global feature representation	Achieved an accuracy of 99.3 %	Requires high computational resources during training, has limited interpretability, and does not perform severity grading

EANet [15]	External Attention Network	Utilizes an external memory-based attention mechanism to enhance attention efficiency	Reached an accuracy of approximately 98 %	Remains computationally demanding and does not support real-time implementation effectively
CCT (Compact Convolutional Transformer) [15]	Hybrid Convolutional Neural Network and Transformer architecture	Combines CNN-based local feature extraction with transformer-based attention mechanisms	Achieved an accuracy range of approximately 98–99 %	Depends on large-scale training datasets and does not provide severity estimation
Proposed Framework (YOLOv8n + ResNet-18 regression)	Two-stage CNN-based framework	Integrates YOLOv8n for kidney stone detection with ResNet-18 regression for severity prediction and clinical category mapping	Obtained mAP@0.5 = 0.707, Precision = 0.737, Recall = 0.688, and MAE = 0.108	–

Self-assessment using TRIPOD guidelines

For transparency and thoroughness in scientific reporting, the study underwent a formal self-evaluation using the TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis) checklist—the most applicable EQUATOR Network standard for studies involving prediction models. The assessment indicated the following:

- Background and rationale (Items 1–3): The clinical context, existing research gaps, and innovative aspects of the work are clearly presented in the Introduction.
- Study data and participants (Items 4–6): Section 2 fully describes dataset provenance, eligibility criteria, and case distribution across training, validation, and testing phases.
- Model development and architecture (Items 7–9): Section 2.2 provides a precise mathematical account of the YOLOv8 detection module, the ResNet-18 regression component, and the associated severity mapping process.
- Performance evaluation (Items 10–12): A wide range of metrics (precision, recall, mAP@0.5, MAE, F1-score), along with PR and ROC curves, confusion matrices, and failure-case reviews, were used to ensure a robust and balanced evaluation.
- Results and interpretation (Items 13–15): The findings are critically examined in light of the contemporary literature (**Tables 2 and 3**), with study limitations openly discussed.
- Clinical applicability and limitations (Items 16–22): The authors address the framework's clinical

interpretability, real-time operability, and deployment prospects, while acknowledging constraints related to dataset scale and broader applicability.

Overall, this self-assessment confirmed that the manuscript adequately covers the essential requirements for transparent reporting as stipulated by the TRIPOD guidelines.

Conclusion

This research introduces the first comprehensive deep learning system designed for simultaneous automated detection and clinical severity grading of kidney stones in CT scans. By integrating YOLOv8 for localization with ResNet-18-based regression for severity prediction, the two-stage pipeline effectively overcomes important shortcomings in current diagnostic practices, providing detailed stone characterization with strong clinical relevance.

Key clinical impact:

- Significantly shortens diagnostic turnaround time and delivers consistent severity evaluations.
- Enables data-driven therapeutic choices via automated mapping to clinical categories (88.7% accuracy).

Scientific contributions:

- Establishes the viability of real-time deployment in clinical settings (45 ms per case inference time).
- Features extensive performance testing and in-depth failure analysis to support ongoing development.

Future directions:

- 3D volumetric analysis: Adopt multi-slice integration to obtain richer three-dimensional spatial information.
- Multi-center validation: Engage in collaborations with 3–5 healthcare institutions to improve external validity.
- Small stone detection enhancement: Develop tailored network designs optimized to identify stones smaller than 3 mm.
- Prospective clinical trial: Initiate IRB-approved trials to directly compare the model's effectiveness with practicing radiologists.
- Stone composition classification: Expand the system to determine stone composition, thereby assisting in personalized treatment strategies.

One limitation of the present work is the absence of comorbidity data in the dataset, which limits the assessment of model performance across diverse patient populations. Subsequent studies employing larger, multi-center datasets that incorporate such information will help mitigate this issue.

Furthermore, this research supports the United Nations Sustainable Development Goals (SDGs), especially SDG 3 (Good Health and Well-being) and SDG 9 (Industry, Innovation and Infrastructure). By developing automated, interpretable, and scalable diagnostic technologies, the framework advances equitable access to healthcare, promotes timely medical intervention, and drives progress in intelligent clinical support systems.

Acknowledgments: None

Conflict of Interest: None

Financial Support: None

Ethics Statement: None

References

1. Ahmed K, Dubey MK, Kajal, Dubey S, Pandey DK. Chronic kidney disease: causes, treatment, management, and future scope. In: Computational intelligence for genomics data. Elsevier; 2025:99-111. doi:10.1016/B978-0-443-30080-6.00010-9
2. Ahmed MM, Hossain MM, Islam MR, Saha S, Nitol SA, Alam MS, et al. Brain tumor detection and classification in MRI using hybrid ViT and GRU model with explainable AI in Southern Bangladesh. Sci Rep. 2024;14(1):22797. doi:10.1038/s41598-024-71893-3
3. Ahsan MM, Mahmud MAH, Saha S, Hossain MS, Nitol SA, Alam MS, et al. Enhancing monkeypox diagnosis and explanation through modified transfer learning, vision transformers, and federated learning. Inform Med Unlocked. 2024;45:101449. doi:10.1016/j.imu.2024.101449
4. Baygin M, Yaman O, Barua PD, Dogan S, Tuncer T, Acharya UR. Exemplar Darknet19 feature generation technique for automated kidney stone detection with coronal CT images. Artif Intell Med. 2022;127:102274. doi:10.1016/j.artmed.2022.102274
5. Bilal O, Hekmat A, Khan SUR. Automated cervical cancer cell diagnosis via grid search-optimized multi-CNN ensemble networks. Netw Model Anal Heal Informatics Bioinform. 2025;14(1):67. doi:10.1007/s13721-025-00563-9
6. Brisbane W, Bailey MR, Sorensen MD. An overview of kidney stone imaging techniques. Nat Rev Urol. 2016;13(11):654-62. doi:10.1038/nrurol.2016.154
7. Caglayan A, Horsanali MO, Kocadurdu K, Ismailoglu E, Guneyli S. Deep learning model-assisted detection of kidney stones on computed tomography. Int Braz J Urol. 2022;48(5):830-9. doi:10.1590/s1677-5538.ibju.2022.0132
8. Cui Y, Zhang J, Wang Z, Wang H, Li X, Chen Y, 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 Imag Biol. 2021;23(3):436-45. doi:10.1007/s11307-020-01554-0
9. Diwan T, Anirudh G, Tembhurne JV. Object detection using YOLO: challenges, architectural successors, datasets and applications. Multimed Tool Appl. 2023;82(6):9243-75. doi:10.1007/s11042-022-13644-y
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. Fitri LA, Suryanto Y, Wardhani IK, Rachmawati D, Arifin Z, Purnomo MH, et al. Automated classification of urinary stones based on

- microcomputed tomography images using convolutional neural network. *Phys Med*. 2020;78:201-8. doi:10.1016/j.ejmp.2020.09.007
12. Foreman KJ, Marquez N, Dolgert A, Fukutaki K, Fullman N, McGaughey M, et al. Forecasting life expectancy, years of life lost, and all-cause and cause-specific mortality for 250 causes of death: reference and alternative scenarios for 2016–40 for 195 countries and territories. *Lancet*. 2018;392(10159):2052-90. doi:10.1016/S0140-6736(18)31694-5
 13. Gulhane M, Gupta S, Singh R, Singh R, Jain R, Gupta P, et al. Integrative approach for efficient detection of kidney stones based on improved deep neural network architecture. *SLAS Technol*. 2024;29(4):100159. doi:10.1016/j.slas.2024.100159
 14. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2016:770-8. doi:10.1109/CVPR.2016.90
 15. Hekmat A, Zuping Z, Bilal O, Khan SUR. Differential evolution-driven optimized ensemble network for brain tumor detection. *Int J Mach Learn Cybern*. 2025;16(9):6447-72. doi:10.1007/s13042-025-02629-6
 16. Hossain MM, Ahmed MU, Islam MR, Ali MS, Khan MA, Alam MS, et al. A novel hybrid ViT-LSTM model with explainable AI for brain stroke detection and classification in CT images: a case study of Rajshahi region. *Comput Biol Med*. 2025;186:109711. doi:10.1016/j.compbiomed.2025.109711
 17. Hossain MM, Islam MR, Ahmed MM, Saha S, Nitol SA, Alam MS, et al. Cardiovascular disease identification using a hybrid CNN-LSTM model with explainable AI. *Inform Med Unlocked*. 2023;42:101370. doi:10.1016/j.imu.2023.101370
 18. Islam MN, Hasan M, Hossain MK, Alam MGR, Uddin MZ, Soylu A. Vision transformer and explainable transfer learning models for auto detection of kidney cyst, stone and tumor from CT-radiography. *Sci Rep*. 2022;12(1):11440. doi:10.1038/s41598-022-15634-4
 19. Khan A, Das R, Parameshwara MC. Detection of kidney stone using digital image processing: a holistic approach. *Eng Res Express*. 2022;4(3):035040. doi:10.1088/2631-8695/ac8b65
 20. Khan SUR, Asim MN, Vollmer S, Dengel A. AI-Driven diabetic retinopathy diagnosis enhancement through image processing and salp swarm algorithm-optimized ensemble network. arXiv:2503.14209. Preprint posted online March 2025.
 21. Maqsood H, Khan SUR. MeD-3D: A multimodal deep learning framework for precise recurrence prediction in clear cell renal cell carcinoma (ccRCC). arXiv:2507.07839. Preprint posted online July 2025.
 22. Preedanana W, Sritan B, Wangka P, Chaiyaprom S, Puttaraksa P, Panpiansak S, et al. Improvement of urinary stone segmentation using GAN-based urinary stones inpainting augmentation. *IEEE Access*. 2022;10:115131-42. doi:10.1109/ACCESS.2022.3218444
 23. Rehman Khan SU, Asif S, Zhao M, Zou W, Li Y, Xiao C. ShallowMRI: a novel lightweight CNN with novel attention mechanism for multi brain tumor classification in MRI images. *Biomed Signal Process Control*. 2026;111:108425. doi:10.1016/j.bspc.2025.108425
 24. Shetty M, Shetty SB, Sequeria GP, Hegde V. Kidney stone detection using CNN. In: 2024 Second International Conference on Data Science and Information Systems (ICDSIS). IEEE; 2024:1-6. doi:10.1109/ICDSIS61070.2024.10594342
 25. Türk C, Petřík 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
 26. Ultralytics | revolutionizing the world of vision AI [Internet]. Available from: <https://www.ultralytics.com/>. Accessed July 28, 2025.
 27. Viswanath K, Anilkumar B, Gunasundari R. Design of deep learning reaction-diffusion level set segmentation approach for health care related to automatic kidney stone detection analysis. *Multimed Tool Appl*. 2022;81(29):41807-49. doi:10.1007/s11042-021-11263-7
 28. Xiang H, Chan M, Brown V, Huo YR, Chan L, Ridley L. Systematic review and meta-analysis of the diagnostic accuracy of low-dose computed tomography of the kidneys, ureters and bladder for urolithiasis. *J Med Imaging Radiat Oncol*. 2017;61(5):582-90. doi:10.1111/1754-9485.12587