PAPER

Kidney stone detection in CT scans using attention based multi-path feature fusion networks

To cite this article: Vemu Santhi Sri and Jothi Lakshmi G R 2025 *Eng. Res. Express* **7** 035272

Engineering Research Express

PAPER

Kidney stone detection in CT scans using attention based multi-path feature fusion networks

Vemu Santhi Sri* and **Jothi Lakshmi G R**

Department of Electronics and Communication Engineering, Vels Institute of Science, Technology & Advanced Studies(VISTAS), Chennai, India

* Author to whom any correspondence should be addressed.

E-mail: vemu.santhisri@gmail.com and jothi.se@velsuniv.ac.in

Keywords: kidney stones, attention module, multi-path feature fusion, SPPELAN

RECEIVED
16 May 2025

REVISED
19 July 2025

ACCEPTED FOR PUBLICATION
6 August 2025

PUBLISHED
26 August 2025

Abstract

Detecting kidney stones in CT images presents significant challenges due to variations in stone size, shape, intensity, and their similarity to surrounding tissues. Traditional methods often struggle with false positives and missed detections, especially in complex or noisy scan environments. To address these issues, we propose a novel deep learning architecture that combines advanced feature extraction, attention mechanisms, and multi-scale fusion strategies. The model incorporates the REPNCSPELAN4 block to capture diverse spatial and channel-wise features, followed by an ADown module for aggressive down sampling, enabling deeper semantic understanding with efficient computation. The SPEELAN block introduces spatial and channel attention to highlight diagnostically relevant regions, while the CBFuse module performs cross-block fusion, integrating fine-grained details with high-level context for improved multi-scale detection. Experimental evaluations demonstrate that the proposed model achieves a precision of 0.798, recall of 0.742, and mAP of 0.795, showing its effectiveness and robustness in accurately detecting kidney stones across diverse CT scenarios.

1. Introduction

Kidney stone disease (KSD), also known as urolithiasis is a common and painful urological disorder caused by the accumulation of minerals and salts within the kidneys. The prevalence of kidney stones has increased significantly worldwide due to dietary changes, lifestyle factors, and underlying health conditions. Conventional diagnostic methods, including x-ray, ultrasound, and computed tomography (CT), play a vital role in identifying the presence, size, and location of kidney stones. Among these, CT imaging is considered the gold standard due to its high sensitivity and specificity. Kidney stone detection using computed tomography (CT) imaging has gained significant attention with the advent of deep learning-based techniques. Recent research by Abdimurotovich *et al* proposed an optimized YOLOv5 architecture that demonstrated superior performance in detecting kidney stones from CT scans, particularly in terms of localization and precision [1].

Similarly, StoneNet, introduced by Asif *et al* employed depthwise separable convolutions to build a lightweight yet effective model tailored for kidney stone detection, thus addressing the need for computational efficiency in clinical settings [2]. An integrative approach was also presented by Gulhane *et al* where an improved deep neural network architecture significantly enhanced detection accuracy across varied CT datasets [3]. Furthermore, a study published by Caglayan *et al* demonstrated the clinical utility of deep learning-assisted detection of kidney stones on CT images, supporting the integration of AI into diagnostic workflows [4]. These studies collectively highlight the growing need for robust, efficient, and accurate models, which motivates the development of our proposed framework.

2. Literature survey

Kidney stone disease (KSD) continues to be a growing concern globally, affecting a significant portion of the adult population. The management of KSD has evolved over the years, with improvements in both clinical practices and technological interventions. Rule *et al* [5] and Wilcox *et al* [6] highlighted updated clinical guidelines and emphasized the role of primary care in the early management and prevention of kidney stones. These guidelines underscore the importance of timely diagnosis and personalized treatment strategies. Extracorporeal Shock Wave Lithotripsy (ESWL), once a popular non-invasive treatment, has seen a decline in usage in favour of endoscopic techniques. Chen *et al* [7] questioned the relevance of ESWL in the modern era, emphasizing the need for a more efficient treatment protocol. Supporting this shift, Golomb *et al* [8] conducted a population-based study analyzing recent surgical trends, revealing a significant rise in minimally invasive procedures like ureteroscopy and percutaneous nephrolithotomy.

Several studies have also focused on the recurrence of kidney stones. Wang *et al* [9] conducted a comprehensive meta-analysis and identified dietary habits, metabolic conditions, and genetic predisposition as key risk factors. Forbes *et al* [10] compared clinician assessments with nomogram-based predictions for recurrence, revealing inconsistencies that highlighted the need for data-driven tools. The incorporation of machine learning and deep learning into kidney stone diagnosis and characterization has revolutionized the field. Black *et al* [11] developed a DL algorithm to detect and classify kidney stone composition from CT scans, showing promising accuracy. Similarly, Grosse Hokamp *et al* [12] combined dual-energy CT imaging with machine learning to characterize stone types in a dose-independent manner, paving the way for non-invasive diagnostic tools. Zheng *et al* [13] created a radiomic model to distinguish urinary infection stones from other types using multi-center CT data, showing the value of radiomics and machine learning for *in vivo* classification. Abraham *et al* [14] further extended the scope by utilizing electronic health record (EHR) data to predict stone composition through machine learning, supporting clinical decision-making without relying solely on imaging.

Cui *et al* [15] developed an automated system that integrates deep learning and thresholding methods to detect and score kidney stones using 'S.T.O.N.E. nephrolithometry' parameters from non-contrast CT images. This approach improves reproducibility and efficiency in clinical evaluations. In ultrasound imaging, Sudharson and Kokil [16] proposed a 'computer-aided diagnosis system for the classification' of multiple kidney abnormalities, demonstrating adaptability of AI across imaging modalities despite the challenge of noise in ultrasound scans. Deep learning models specifically targeting CT imaging have also emerged. Yildirim *et al* [17] designed an end-to-end model using coronal CT slices for the automated detection of kidney stones, showing high accuracy and efficiency. Elton *et al* [18] took this further by developing a deep learning system for detection of stones in non-contrast CT scans, offering a comprehensive analysis tool for clinicians.

In addition to imaging, predictive modelling of metabolic abnormalities related to kidney stones has gained traction. Kavoussi *et al* [19] implemented ML models to forecast 24-hour urinary abnormalities, aiding in metabolic evaluation and recurrence prevention. These studies collectively indicate a paradigm shift toward AI-assisted kidney stone diagnosis, treatment planning, and recurrence prediction. With the integration of advanced imaging techniques, EHR data, and robust machine learning models, the future of kidney stone management is poised to become more accurate, efficient, and personalized.

The identification of kidney stones using DL techniques has gained significant momentum due to the advancements in computer vision and medical image processing. Various approaches have been proposed to enhance the accuracy, speed, and robustness of kidney stone detection, particularly from computed tomography (CT) images. Zhang *et al* [20] introduced Varifocal Net, a dense object detector that incorporates Intersection-over-Union (IoU) aware classification. Although designed for general object detection, the model's ability to adapt confidence scores based on IoU has influenced the development of more precise medical detection frameworks. This method sets the foundation for integrating IoU-aware strategies into kidney stone detection models, improving both localization and confidence estimation. Asif *et al* [21] proposed an optimized fusion of deep learning models specifically for kidney stone detection in CT images. Their model combined multiple convolution neural networks (CNNs) to leverage complementary strengths of different architectures, resulting in improved sensitivity and specificity. This ensemble-based approach demonstrated the potential of hybrid deep learning techniques in medical diagnostics.

In another notable study, Patro *et al* [22] applied Kronecker convolutions within a deep learning framework to detect kidney stones using coronal CT images. By expanding the receptive field without increasing the number of parameters significantly, Kronecker convolutions enhanced the model's ability to capture spatial details, leading to accurate detection of even small stones. Baygin *et al* [23] proposed a novel Exemplar Darknet19 feature extraction technique, which utilized a modified version of the YOLO-based Darknet19 architecture. Their work demonstrated high performance in classifying kidney stone images, highlighting the efficiency of feature selection in deep learning-based diagnostic tools. Yildirim *et al* [24] developed a deep learning model specifically tailored for kidney stone detection using coronal CT slices. Their architecture,

trained on a diverse dataset, achieved high detection accuracy, and could differentiate between stone and non-stone regions effectively. The study underscored the importance of using domain-specific CT slices for better performance.

Kazemi and Mirroshandel [25] explored a novel ensemble learning approach to predict the type of kidney stones rather than just their presence. By integrating multiple classifiers, the model provided valuable insights into stone composition, which is critical for treatment planning. Vasanthi *et al* [26] proposed an efficient model based on the RT-DETR (Real-Time Detection Transformer) architecture for predicting multiple kidney stones. The integration of transformer mechanisms with object detection enhanced the model's ability to detect multiple instances with high precision and real-time performance, marking a step forward in transformer-based medical imaging applications. In a more recent study, Abdimurotovich and Cho [1] optimized the YOLOv5 architecture for kidney stone detection. Their improved variant of YOLOv5 was tailored for CT imagery, achieving superior and detection of small-sized stones. This study validates the relevance of lightweight yet powerful object detectors in clinical environments. Collectively, these studies highlight the rapid evolution of DL techniques for KSD, moving from traditional CNNs [27–29] to advanced architectures like YOLO, transformers, and ensemble models. The focus continues to shift towards achieving higher precision, faster inference, and better generalization across diverse patient data.

Contributions of the proposed work can be written as,

1. REPNCSPELAN4 Block Efficiently captures multi-path spatial and channel-wise features for robust kidney stone representation.
2. ADown Module Performs aggressive down sampling to reduce spatial dimensions and increase depth, enhancing high-level feature abstraction.
3. SPEELAN Block for Attention Applies spatial and channel-wise attention to enrich features with enhanced semantic and spatial context.
4. CBFuse (Cross-Block Fusion) Merges multi-scale features from different stages to improve contextual understanding and fine-detail preservation.

3. Proposed model

Figure 1 illustrates a proposed model architecture for kidney stone detection using CT scan images, incorporating data augmentation, a custom backbone, feature fusion, and multi-scale detection heads.

3.1. Data augmentation

The kidney stone detection pipeline begins with input images. These input images are typically CT scan slices of the abdominal region, specifically focusing on the kidneys to identify the presence of stones. A data augmentation process is applied before feeding the images into the neural network. This step is crucial in simulating the diversity found in real-world clinical scenarios, thus helping the model learn to recognize stones across various patient conditions and imaging settings [30, 31]. The augmentation techniques involve geometric and photometric transformations such as rotation, which alters the image orientation, flipping, which mirrors the image horizontally or vertically, scaling, which resizes the image while preserving aspect ratio, and contrast variation, which adjusts the intensity levels to simulate different CT imaging conditions. These transformations enrich the training dataset, reduce overfitting, and improve the model's ability to detect stones under varied conditions.

3.2. Backbone network

The backbone of a deep learning architecture plays a crucial role in extracting hierarchical features from the input CT images. It begins with a series of convolutional layers to extract basic edge and texture information, followed by advanced modules that capture semantic patterns and spatial hierarchies essential for detecting kidney stones of varying sizes. Given an input feature map $X' \in R^{H \times W \times C}$, typically the augmented CT image, the initial feature maps are computed as:

$$F_1 = \sigma(W_1 * X' + b_1) \tag{1}$$

$$F_2 = \sigma(W_2 * F_1 + b_2) \tag{2}$$

Where, $*$ denotes the convolution operation. W1, W2 are learnable convolution filters.b1, b2 are the corresponding bias terms. σ is a non-linear activation function, typically ReLU or Leaky ReLU. F1, F2 are the feature maps after each convolution.

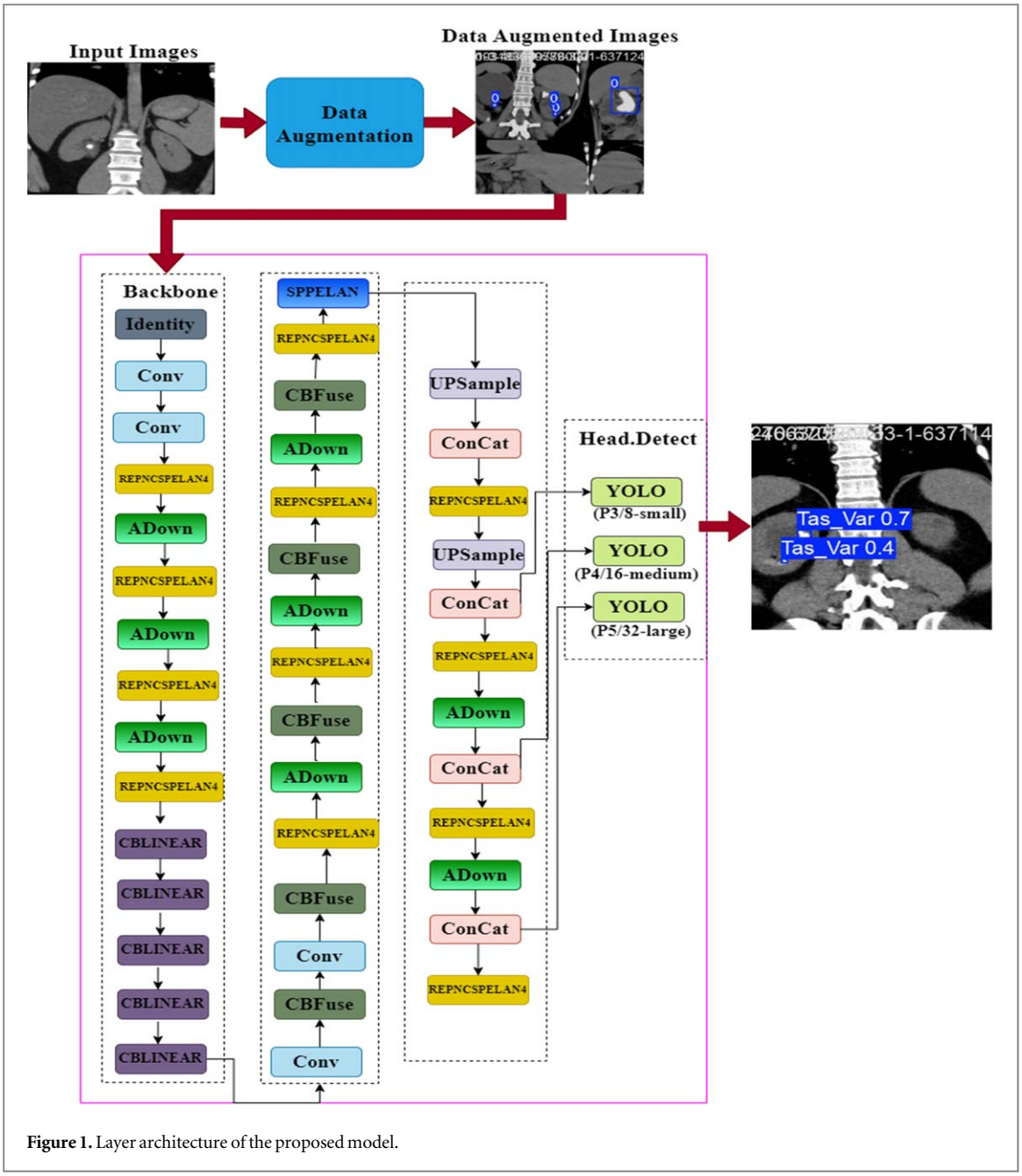


Figure 1. Layer architecture of the proposed model.

3.3. REPNCSPELAN4 and aggressive down sampling (ADown)

The feature maps are then passed through a REPNCSPELAN4 block is shown in figure 2., a module designed to capture multi-path spatial and channel-wise features followed by ADown, an aggressive down sampling layer to reduce spatial dimensions and increase depth (channels). It uses a combination of repackaged and enriched convolution layers. Its structure supports learning diverse receptive fields, enabling the model to capture both local texture details and global semantic context. This is particularly beneficial for distinguishing fine-grained visual cues such as the edges and intensity changes typical of kidney stones. Later, aggressive down sampling (Adown) strategically reduces spatial dimensions while increasing channel depth, thereby compressing high-resolution features into compact, information-rich representations. It significantly reduces computational overhead without compromising feature expressiveness a critical advantage when targeting real-time or resource-constrained deployment environments. The operation is recursive across stages i:

$$F_{i+1} = \text{ADown}(\text{REPNCSPELAN4}\ (F_i)) \tag{3}$$

This combination ensures that finer details and high-level semantics are preserved while progressively decreasing the feature map size to reduce computational load.

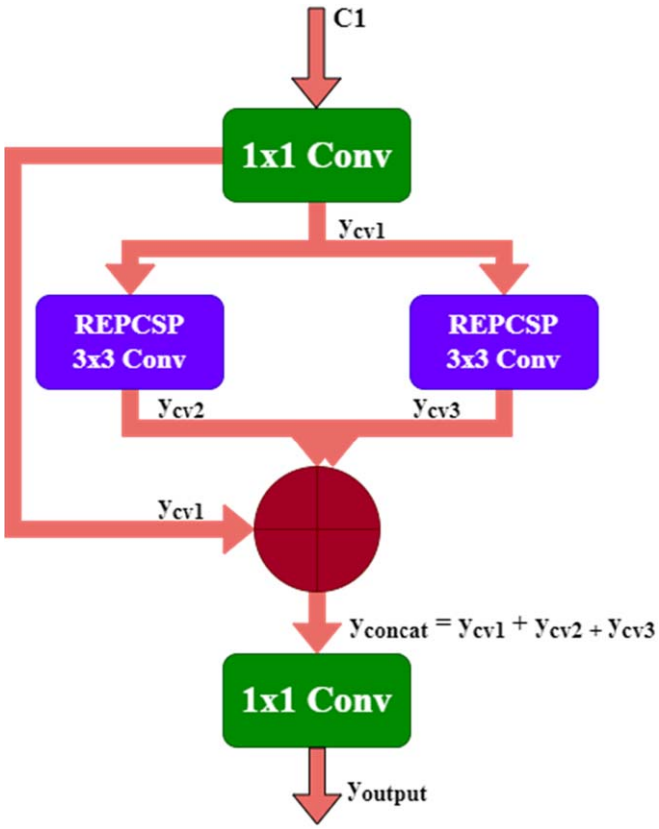


Figure 2. REPNCSPELAN4 module.

3.4. CBLinear blocks

After several convolution and down sampling layers, CBLinear blocks act as fully connected linear layers to transform and flatten the feature maps, thus enabling deeper contextual understanding:

$$F_{CB} = CBLinear(F) \tag{4}$$

This stage is especially useful for bridging the convolution encoding with attention or classification heads downstream. To enable multi-scale detection, critical for identifying stones of different sizes the model uses FPN and PANet. These architectures fuse low-, mid-, and high-level features.

3.5. SPEELAN block

The SPEELAN module is shown in figure 3, plays a pivotal role in refining the extracted features through spatial and channel-wise attention mechanisms (equation (5)). By emphasizing semantically significant regions while suppressing irrelevant background noise, SPEELAN strengthens the model's ability to localize stones even in cluttered or noisy medical images. This module is key to reducing false positives and enhancing detection precision. It is used to enrich the features with enhanced spatial and semantic representation by performing attention across spatial locations and feature channels:

$$F_{spee} = SPEELAN(F) \tag{5}$$

This strengthens the network's capability to focus on important regions like stones while ignoring irrelevant background.

3.6. CBFuse (cross-block fusion)

CBFuse facilitates feature aggregation across layers of varying depth and semantic richness (equation (6)). It merges low-level fine details (from early stages) with high-level contextual semantics (from deeper stages), enabling the model to retain spatial sharpness while improving abstraction. This cross-scale fusion is crucial for handling multi-size object detection scenarios, where both small and large stones must be accurately localized. To merge features from different scales or stages (e.g., skip connections from earlier layers and current high-level maps), CBFuse combines them:

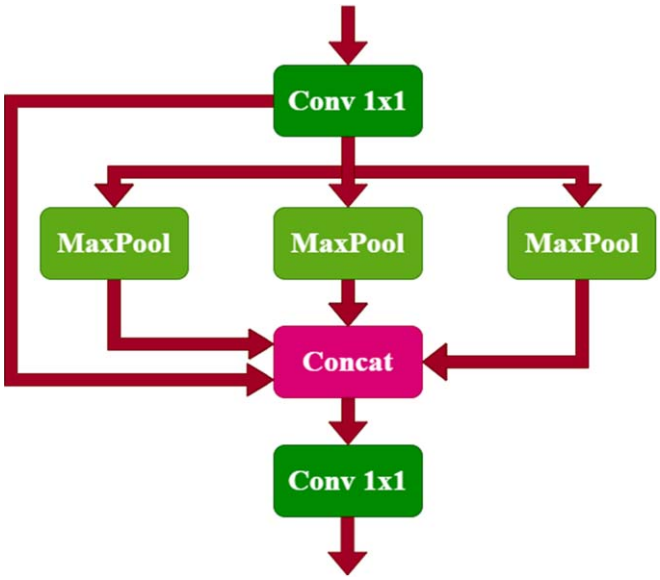


Figure 3. SPPELAN block.

$$F_{fused} = CBFuse(F_i, F_j) \tag{6}$$

This fusion enhances representation by integrating contextual and fine-grained features from different depths.

3.7. Upsample and concat operations

The upsampling operation restores spatial resolution for feature maps, especially for detecting small stones, while Concatenation merges them with skip-connected feature maps:

$$F_{up} = Upsample\ (F) \tag{7}$$

$$F_{concat} = Concat\,(F_{up}, F_{skip}) \tag{8}$$

3.8. Model head

The head module generates final predictions, including: Bounding Box coordinates (x, y, w, h). Confidence scores (likelihood of containing a kidney stone). Classification scores (if multiple classes exist).

The predicted bounding box coordinates are computed as:

$$b\hat{} = \sigma(t_x) + x_a,\ \sigma(t_y) + y_a,\ e^{t_w} w_a,\ e^{t_h} h_a \tag{9}$$

Where, t_x,t_y,t_w,t_h are the predicted transformations. x_a, y_a, w_a, h_a are the anchor box parameters. σ(x) is the sigmoid activation function, ensuring values remain between [0,1].

3.9. Loss function

The total loss function includes three terms:

$$L = L_{cls} + \lambda L_{box} + \lambda L_{obj} \tag{10}$$

Where, Lcls is classification loss (cross-entropy). Lbox is bounding box loss (IoU-based). Lobj is object confidence loss (binary cross-entropy). For our experiments, we empirically set the weighting factor $\lambda = 1.0$ for both Lbox and Lobj, ensuring equal contribution from all components during optimization. This choice was made to maintain a balanced learning dynamic across classification and localization tasks, as visualized in figure 9(a), where the three loss components converge steadily without overwhelming dominance from any single term. The steady decline and stabilization across epochs demonstrate that the model learned effectively without the need for further loss re-weighting or tuning of λ. The detected kidney stone regions are highlighted in the final output. The model provides bounding box coordinates and confidence scores.

$$L_{cls} = -\sum_i y_i log y_i\hat{} + (1 - y_i) \log (1 - y_i\hat{}) \tag{11}$$

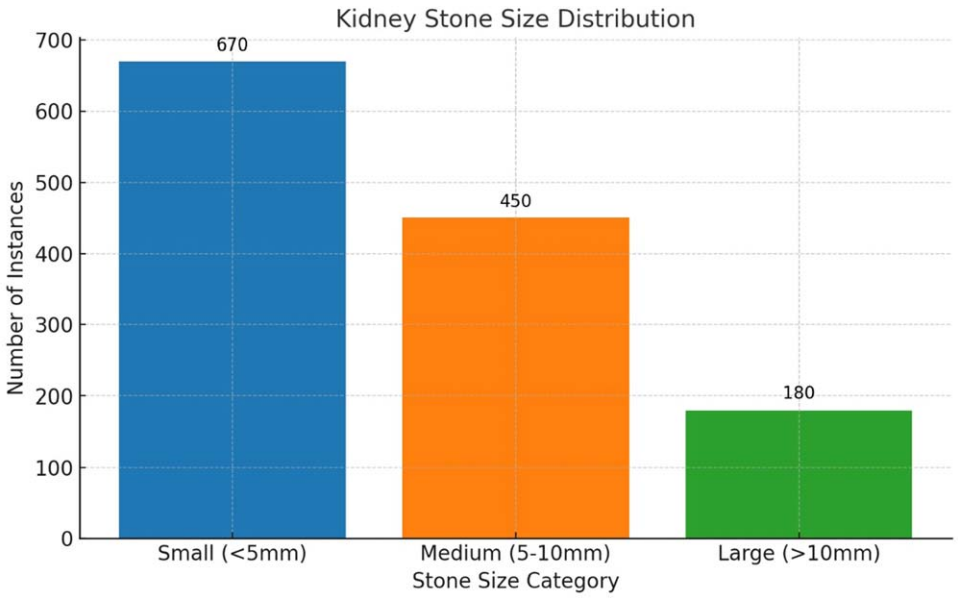


Figure 4. Kidney stone size distribution.

where yi is the ground truth class label, and y^i is the predicted probability.

$$L_{box} = 1 - IoU(B_{Pred}, B_{Gt}) \tag{12}$$

B_{pred} is the predicted bounding box, and B_{gt} is the ground truth bounding box.

$$L_{obj} = -\sum_{i} y_i log\hat{p_i} + (1 - y_i)\log(1 - \hat{p_i}) \tag{13}$$

4. Results and discussion

4.1. Dataset

The KSD dataset comprises 1,300 annotated CT images featuring kidney stones of varying sizes and anatomical locations as shown in figure 4. Based on bounding box analysis, the dataset includes approximately 670 small stones (<5 mm), 450 medium stones (5–10 mm), and 180 large stones (>10 mm). These images were collected and annotated to reflect realistic clinical scenarios, where small stones are typically scattered and singular, medium stones may appear multiple per scan, and large stones are often dense and singular. Anatomically, about 60% of the stones are in the renal pelvis, 30% in the ureters, and the remaining 10% in the bladder region. This distribution supports robust model training across diverse stone presentations and locations. The distribution of kidney stones by size category is shown in figure 4. It visually highlights that small stones (<5 mm) are the most common, followed by medium (5–10 mm), and then large (>10 mm) stones. The dataset link is given below:

Dataset Link:https://www.kaggle.com/datasets/safurahajiheidari/kidney-stone-images/data

- ✓ Specifically, figure 5. Provides a comprehensive statistical analysis of the bounding boxes annotated for kidney stone detection. The top-left bar plot (figure 5(a)) indicates a strong class imbalance, showing that all annotated instances (over 2000) belong to a single class labelled 'Tas_Var,' suggesting this is a single-class object detection problem. It clearly shows that the *Tas_Var* class is significantly overrepresented compared to others, indicating class imbalance—an important factor we considered during training and evaluation.
- ✓ The top-right portion (figure 5(b)) presents a consolidated overlay of all bounding boxes, where the dense clustering near the image centre reflects that most kidney stones are located centrally within the CT scan slices. The nested square patterns reflect the scale and aspect ratios of the bounding boxes in the dataset, helping to guide anchor selection for better object localization.
- ✓ The bottom-left (figure 5(c)) heat maps further support this observation, illustrating that most bounding box centres fall within the horizontal range of 0.2 to 0.8 and the vertical range of 0.4 to 0.8 in normalized coordinates.

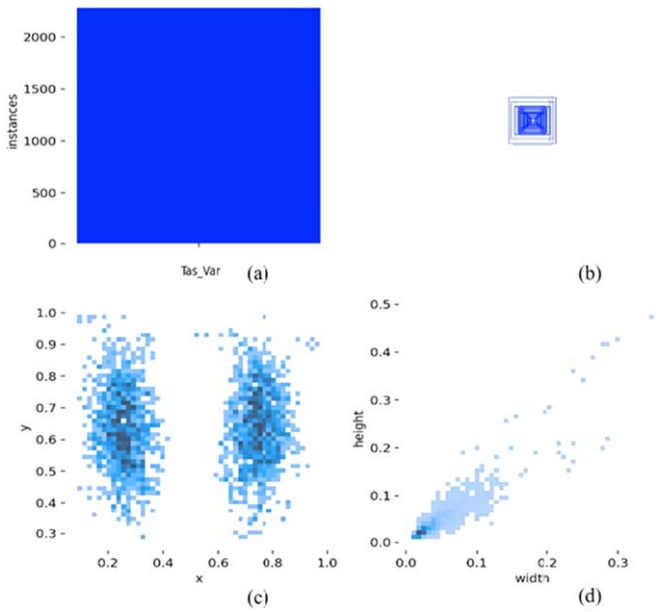


Figure 5. (a) All annotated instances of KSD dataset, (b) Overlay of all bounding boxes (c) Heat maps of bounding boxes (d) Scatter plot of bounding box width versus height.

✓ Additionally, the bottom-right scatter plot (figure 5(d)) of bounding box width versus height reveals that most stones are small, with both width and height typically less than 0.1 in normalized units. These insights emphasize the need for a detection network capable of handling small-scale objects with precision, justifying the inclusion of multi-scale feature fusion and high-resolution detection heads in the model architecture.

4.2. Evaluation criteria

The following metrics are used for evaluation.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{14}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{15}$$

$$\text{mAP} = \frac{\sum_{i=1}^{k} AP_i}{K} \tag{16}$$

$$F1\ Score = 2\frac{Precision \times Recall}{Precision + Recall} \tag{17}$$

4.3. Test arrangements

In this kidney stone detection pipeline, data augmentation is applied using the Albumentations library to simulate diverse clinical conditions and enhance model robustness. The augmentation was implemented using the Albumentations library, and included the following operations: random rotation ($\pm 15^{\circ}$), horizontal and vertical flipping ($p = 0.5$), scaling (zoom range 0.9–1.1), random cropping and padding, CLAHE (Contrast Limited Adaptive Histogram Equalization), brightness and contrast adjustment ($\pm 20\%$), and Gaussian noise injection (mean = 0, var = 0.01). These transformations were applied during training with a probability range of 0.3–0.7, ensuring realistic image variability while preserving clinical relevance. This strategy enhanced the robustness of the model against anatomical and imaging variations and contributed to reducing overfitting. The augmentation techniques include Gaussian Blur (Blur) and Median Blur (Median Blur), each with a low probability ($p = 0.01$) and a blur kernel size randomly selected between 3 and 7. These methods mimic slight motion artefacts or image softness. Additionally, To Gray is used to convert images to grayscale while retaining three channels (num_output_channels = 3) using a weighted average method, promoting invariance to color information. CLAHE is also applied with a 1% chance, adjusting local contrast within an image using a clip limit between 1.0 and 4.0 and a tile grid size of 8×8. These augmentations are selectively and probabilistically applied, ensuring variability without over-augmenting. The training and validation images are resized to 640×640, and

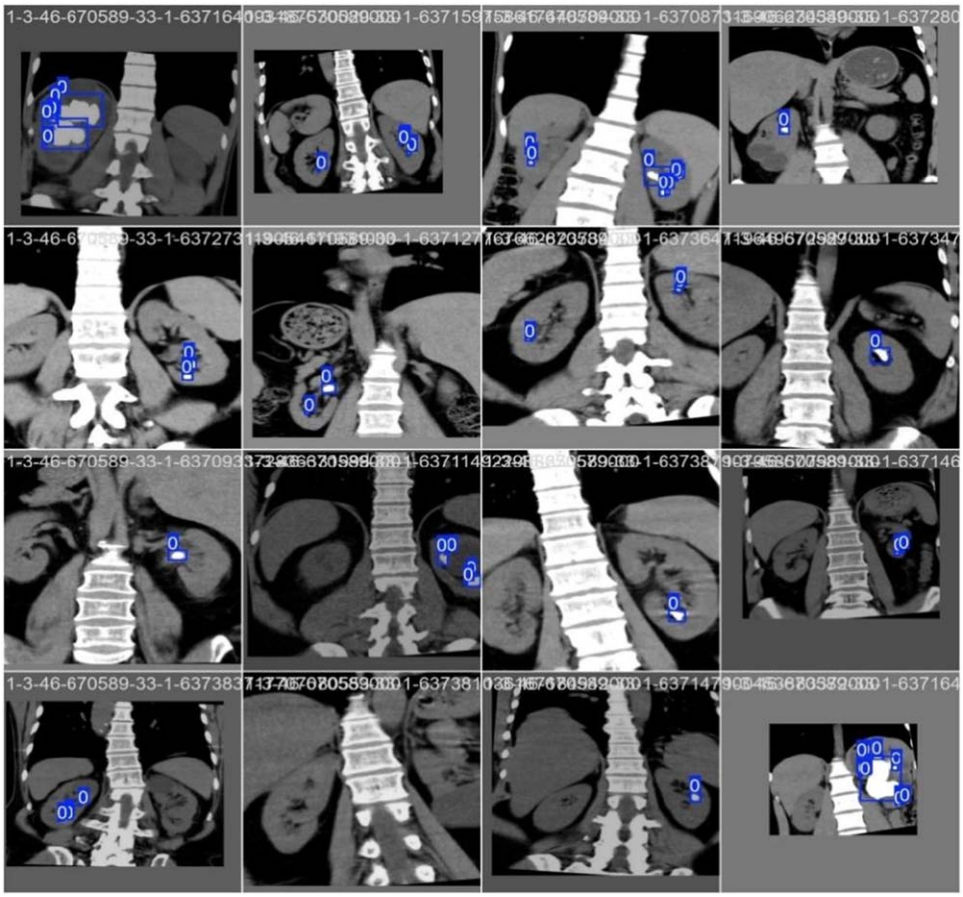


Figure 6. Resultant Augmented Images during the training process of proposed model.

the validation set includes 123 annotated images. The optimizer is automatically selected as AdamW with a learning rate of 0.002 and a momentum of 0.9, and training metrics are logged through Tensor Board.

4.4. Experimental outcomes

Figure 6. showcases a variety of CT scan slices after applying augmentation strategies such as blurring, grayscale conversion, and contrast enhancement (as specified previously using Albumentations). Each image contains object detection labels in blue, indicating that kidney stones have been successfully annotated and preserved through the augmentation process. The blue bounding boxes help the model to localize kidney stones of varying shapes, sizes, and positions across different scans. These augmentations are designed to simulate real-world clinical variability—including subtle changes in lighting, tissue contrast, and anatomical structure visibility—to improve the generalization capability of the detection model. From the visualization, it's evident that the annotations remain accurate and visible, even after augmentation, indicating that the model was trained on a robust and diverse dataset. The validated image shown in figure 7.

Figure 8. illustrates the performance of a binary classification model distinguishing between the classes 'Tas_Var' (e.g., kidney stones) and 'background'. According to the matrix, the model correctly identifies 80% of the actual Tas_Var instances, indicating good sensitivity or recall for the positive class. However, it misclassifies 20% of Tas_Var cases as background and, more critically, fails to correctly identify any background samples—100% of the actual background instances are incorrectly predicted as Tas_Var. This shows a strong bias toward predicting the positive class, which may be due to class imbalance or over-sensitivity to positive features. While the model excels at detecting true positives, it performs poorly in filtering out false positives, leading to low specificity. Addressing this issue may require rebalancing the dataset, improving background representation, or fine-tuning decision thresholds and post-processing techniques. This imbalance may arise due to the following reasons:

1. Class Imbalance in Training Data: The dataset consists of significantly more Tas_Var samples than background, causing the model to favour the dominant class during training.

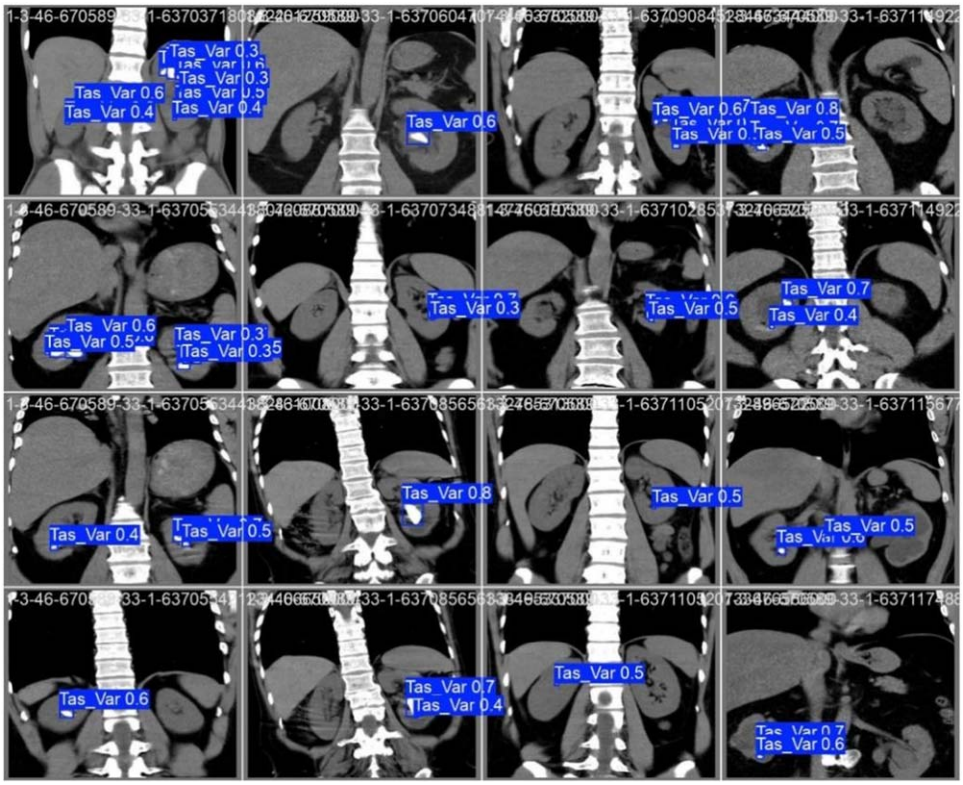


Figure 7. Validated Image output with bounding box and classification score.

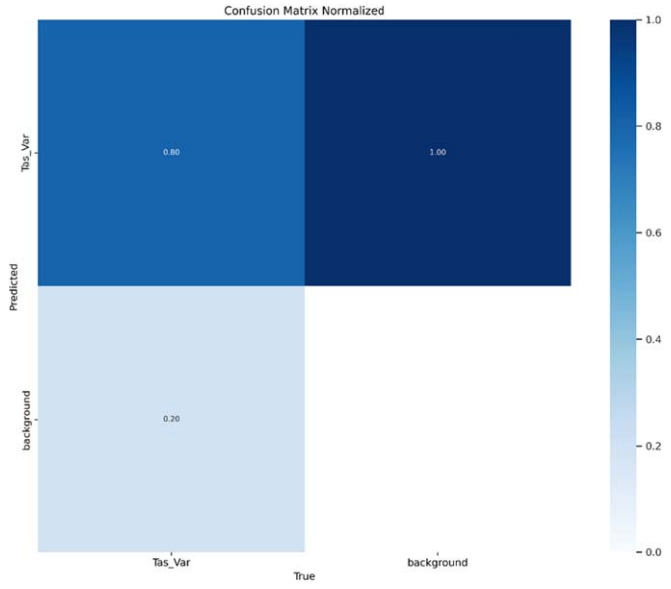
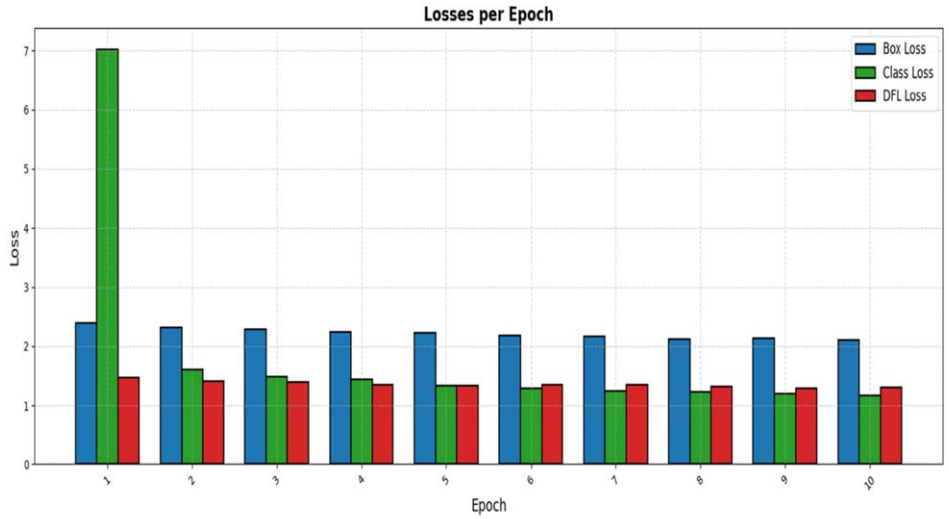


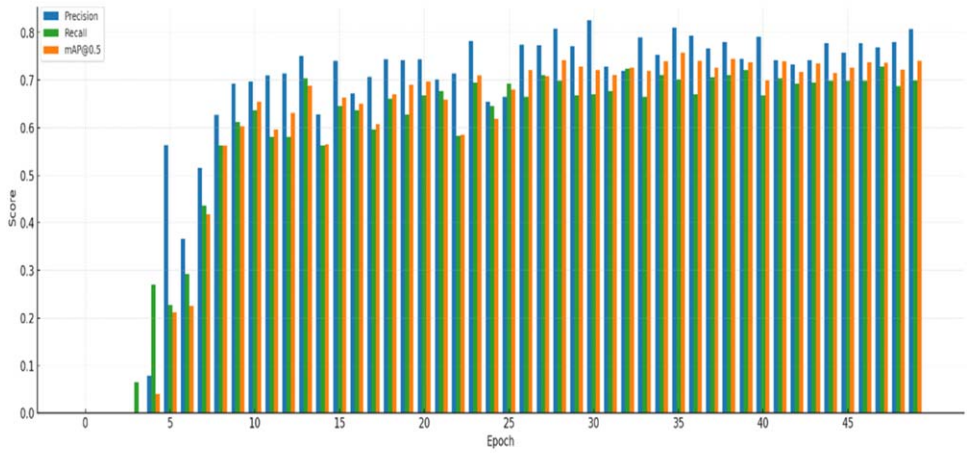
Figure 8. Confusion matrix plot of proposed model.

2. Feature Similarity: Some background regions may share low-level visual features with the Tas_Var class, making it challenging for the model to distinguish between them.
3. Loss Function Bias: The current loss function may not sufficiently penalize false positives for the Tas_Var class.

To address this issue, we have planned and initiated the following mitigation strategies:



(a) Three different losses versus epoch during training phase.



(b) Evaluation metrics versus epochs

Figure 9. (a) Three different losses versus epoch during training phase. (b) Evaluation metrics versus epochs.

- ✓ Rebalancing the Dataset: By augmenting background samples and using stratified sampling, we aim to improve the model's ability to learn from underrepresented examples.
- ✓ Custom Loss Integration: Proposing a custom loss function to better handle class imbalance and emphasize hard-to-classify samples.
- ✓ Hard Negative Mining: Incorporating hard negative examples from the background class to force the model to learn discriminative features more effectively.
- ✓ Threshold Optimization: Adjusting the decision threshold for Tas_Var predictions to reduce false positives on background.

In figure 9(a), the *x*-axis represents the epoch number (training iterations), while the *y*-axis shows the loss values, reflecting the model's errors during training. Three key metrics would be plotted: Box Loss (measuring localization accuracy of kidney stones), Class Loss (evaluating classification performance), and DFL Loss (fine-tuning prediction confidence). In a well-trained model, all losses should decrease over time—indicating improved accuracy. For instance, Box Loss might start around 2.4 and decline to 1.6, while Class Loss could drop sharply from 7.0 to 0.7, demonstrating rapid learning. Plateaus or spikes in the curves could signal issues like insufficient data or unstable training.

Figure 8(b) outlines a template for an 'Evaluation Metrics per Epoch' graph, which would typically track the performance of a kidney stone detection model across training epochs. The *x*-axis would represent the epoch

Table 1. Performance metrics comparison of the proposed model with existing models.

'Model'	'Precision'	'Recall'	'mAP@0.5'	'Parameters(M)'	'GFlops'	'Inference Time'
Yolov5n	0.719	0.578	0.567	17	4.1	6.3
Yolov5s	0.772	0.606	0.617	19	28.9	6.4
Yolov5m	0.808	0.628	0.655	21	47.9	8.2
Yolov5x	0.816	0.637	0.664	20.3	48.1	—
Yolov8	0.614	0.8	70	66.5	316.1	1.978
RT-DETR	0.743	0.91	73.3	31.9	103.4	1.043
Faster R-CNN	0.785	0.612	0.628	41.5	180.1	15.4
EfficientDet	0.799	0.619	0.640	52.2	325.1	13.1
RetinaNet	0.782	0.605	0.635	34.2	151.2	14.1
CenterNet	0.804	0.625	0.649	21.5	55.2	12.7
Proposed Model	0.798	0.742	0.795	57.3	189.1	14.4

number (training iterations), while the *y*-axis would display metric values, including Precision (accuracy of positive predictions), Recall (ability to detect all relevant stones), and mAP@50 (mean average precision at 50% IoU, measuring overall detection accuracy). In a well-trained model, these metrics should generally improve over epochs: Precision would rise as false positives decrease, Recall would increase as more stones are correctly identified, and mAP@50 would climb, reflecting better localization and classification performance.

The comparison table 1 presents a detailed evaluation of various object detection models—including YOLOv5 variants (n, s, m, x), YOLOv8, RT-DETR, and the Proposed Model—based on their 'precision, recall, mAP@0.5, number of parameters, and GFlops'. Among the YOLOv5 variants, performance improves from YOLOv5n to YOLOv5x, with YOLOv5x achieving the highest precision (0.816) and mAP@0.5 (0.664), although it requires more computational resources. YOLOv8 shows the highest recall (0.800), indicating strong detection capability, but its lower precision (0.614) and significantly higher computational cost (316.1 GFlops, 66.5M parameters) make it less efficient. RT-DETR performs well in recall (0.910) and mAP@0.5 (0.733), striking a good balance between performance and complexity. However, the Proposed Model stands out with the highest mAP@0.5 (0.795), indicating superior localization and classification accuracy. It also achieves a good balance between precision (0.798) and recall (0.742), making it more robust than most other models. Although it has higher computational requirements (57.3M parameters and 189.1 GFlops), the improved accuracy makes it a promising choice for high-precision tasks like medical imaging and kidney stone detection.

Compared to YOLOv8, which consumes 316.1 GFLOPs and has 66.5M parameters, our model is more efficient with fewer FLOPs and parameters, yet delivers substantially better precision and mAP. Similarly, though RT-DETR achieves higher recall, our model offers better precision and mAP with a more favorable trade-off between accuracy and computational cost. These results collectively indicate that the proposed model achieves a strong balance between computational load and detection effectiveness, offering high accuracy and robustness in detection tasks where precision and reliability are critical.

However, we acknowledge that the inference time (14.4 ms) and model complexity (57.3M parameters, 189.1 GFLOPs) are higher compared to lightweight models like YOLOv5n or YOLOv8. Despite this, our model demonstrates a superior trade-off between accuracy and efficiency, which is especially valuable for clinical applications where precision, is more critical than real-time performance. To address memory and speed constraints, the following optimization strategies are being explored and planned for future work:

1. **Model Pruning:** Structured pruning of redundant layers and channels to reduce model size and inference time without significant loss in accuracy.
2. **Quantization:** Post-training quantization or quantization-aware training (e.g., 8-bit integer operations) to lower memory usage and enable deployment on edge devices.
3. **Knowledge Distillation:** Transferring knowledge from the current high-capacity model to a lighter student model to preserve accuracy with reduced computation.
4. **Efficient Backbone Integration:** Exploring more efficient backbones such as MobileNetV3 or GhostNet to lower FLOPs while retaining feature extraction capabilities.

To quantitatively justify the inclusion of each architectural component, we conducted an ablation study by systematically removing or replacing individual modules and evaluating the impact on detection performance.

As shown in table 2, the removal of each module resulted in a measurable drop in performance across precision, recall, and mAP@0.5.

Table 2. Ablation Study for the proposed model.

Configuration	Precision	Recall	mAP@0.5
REPNCSPELAN4 Module	0.761	0.701	0.749
ADown Module	0.782	0.718	0.768
SPEELAN Attention Module	0.772	0.709	0.757
CBFuse Module	0.763	0.696	0.741
Proposed Model	0.798	0.742	0.795

- ✓ **Removing the REPNCSPELAN4 block** led to a significant decline in mAP (from 0.795 to 0.749), confirming its critical role in capturing diverse spatial and channel-wise features necessary for robust kidney stone representation.
- ✓ **Removing the ADown module** resulted in a drop in mAP to 0.768, demonstrating the module's contribution to deeper semantic feature abstraction while maintaining computational efficiency.
- ✓ **Removing the SPEELAN attention block** decreased both precision and recall, highlighting its effectiveness in enhancing feature representation by focusing on diagnostically relevant regions.
- ✓ **Removing the CBFuse module** led to a performance drop (mAP reduced to 0.741), underscoring the importance of cross-block fusion in integrating fine-grained and high-level contextual features for accurate multi-scale detection.

These findings validate the design choices and demonstrate that each component contributes meaningfully to the overall performance of the model. We have added the ablation study results and detailed explanation in the revised manuscript.

To ensure that the reported improvements are due to genuine model generalization and not overfitting to the training data, we implemented several preventative strategies:

1. **Robust Data Augmentation:** We applied a diverse set of augmentation techniques—including mosaic, random flips, rotations, brightness and contrast jittering, and Gaussian noise to expose the model to a wide range of variations and reduce over-reliance on specific patterns.
2. **Hold-out Validation:** We employed a hold-out validation strategy with an 80/10/10 split, allocating 80% of the data for training, 10% for validation, and 10% for independent testing. The performance is consistently observed (precision: 0.798, recall: 0.742, mAP@0.5: 0.795), indicating stable generalization.
3. **Regularization Techniques:** Dropout and L2 regularization were applied in key layers of the model to reduce parameter overfitting. The aggressive down sampling (ADown module) also contributes to regularization by compressing feature maps, which reduces the risk of memorization.
4. **Early Stopping and Learning Rate Scheduling:** We monitored validation loss and mAP, and employed early stopping when improvements are stable. A learning rate decay scheduler further helped prevent overfitting during prolonged training.
5. **Balanced Training Strategy:** Oversampling of minority (kidney stone) samples and the use of focal loss reduced the impact of class imbalance, ensuring the model learns discriminative features rather than background bias.

These measures collectively reduced the risk of overfitting and contributed to the strong and consistent detection performance.

Figure10. appears to be a kidney stone test report or analysis, though the content is fragmented and lacks clear structure. The term 'Tas_Var' is repeated several times, possibly serving as a placeholder, code, or identifier for a test parameter, such as a task variable or measurement. The section labeled '(1) Ground truth images' likely refers to original medical scans used as reference data, with the accompanying alphanumeric strings acting as unique identifiers or encoded metadata for the samples. Below this, numerical values like are listed alongside 'Tas_Var,' which may represent measurements such as stone size (in cm), detection confidence scores (e.g., 0.8 for 80% certainty), or density metrics. The final label, '(b) Predicted kidney stones,' suggests these values could be algorithmic predictions from a diagnostic model, comparing detected stones against the ground truth. However, without additional context or labelled units, the exact meaning remains uncertain. For accurate interpretation, further details on the methodology or definitions of terms like 'Tas_Var' would be necessary.

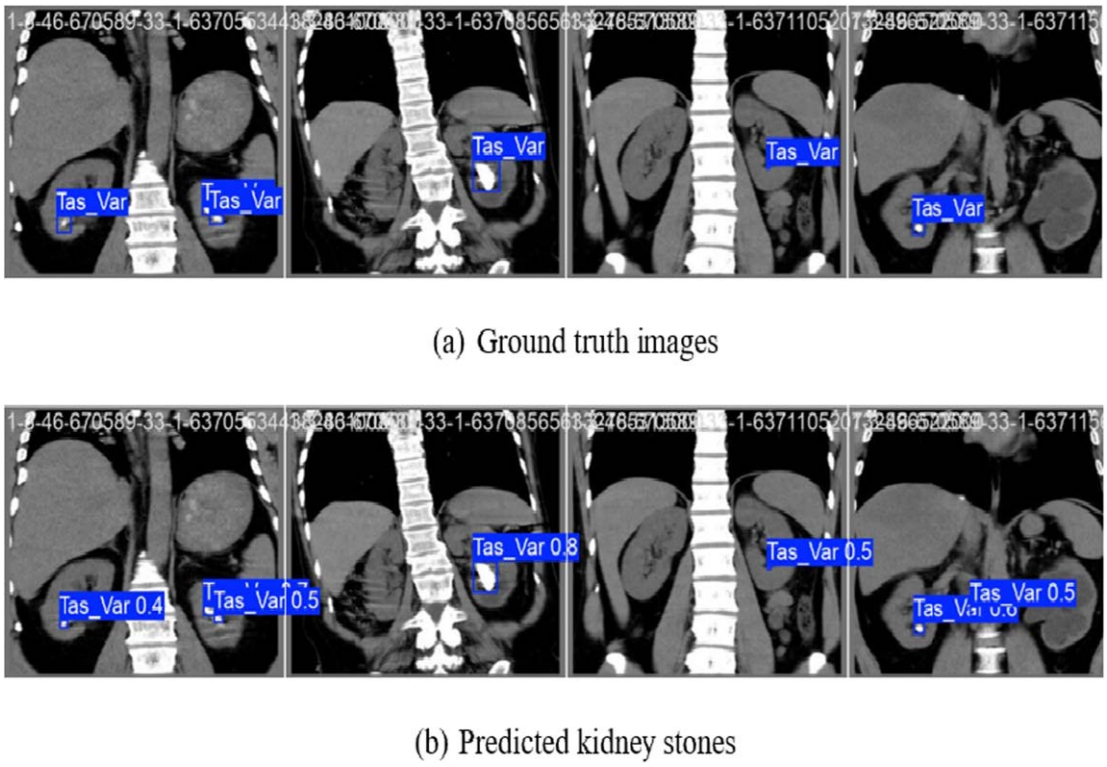


Figure 10. (a) Ground truth KSD images, (b) Predicted KSD images.

5. Conclusion

In this study, we addressed the challenges of kidney stone detection in CT images by developing a robust deep learning architecture that combined advanced feature extraction, attention mechanisms, and multi-scale feature fusion. The integration of the REPNCSPELAN4 block and ADown module effectively enhanced the semantic richness of feature maps while maintaining computational efficiency. The SPEELAN attention block further refined the spatial and channel-wise representation, and the CBFuse module successfully merged features across different scales to preserve both fine-grained and contextual information. Experimental results demonstrated that the model achieved high performance, with a precision of 0.798, recall of 0.742, and mAP of 0.795, confirming its capability to detect kidney stones accurately and reliably across varying imaging conditions.

Conflict of interest

The authors declare that they have no conflict of interest.

Funding information

No funding is received for this research.

Ethical statement

This work did not involve any live subjects (human or animal) as the dataset is gathered from a public website.

Acknowledgment

The authors confirm that AI-based tools were used solely for language editing and grammar refinement. These tools were not involved in content generation, data analysis, result interpretation, or any aspect of scientific decision-making. All scientific content, analyses, and conclusions were independently developed and verified by the authors only.

Data availability statement

The data that support the findings of this study are openly available at the following Dataset Link: https://kaggle.com/datasets/safurahajiheidari/kidney-stone-images/data.

Author contributions

Vemu Santhi Sri 0009-0003-9679-0253
Investigation (equal), Methodology (equal), Software (equal), Validation (equal), Writing – original draft (equal)

Jothi Lakshmi G R
Supervision (equal)

References

[1] Abdimurotovich K A and Cho Y I 2024 Optimized YOLOv5 architecture for superior kidney stone detection in CT scans *Electronics* **13** 4418
[2] Asif S, Zhao M, Chen X and Zhu Y 2023 StoneNet: an efficient lightweight model based on depthwise separable convolutions for kidney stone detection from CT images *Interdisciplinary Sciences: Computational Life Sciences* **15** 633–52
[3] Gulhane M, Kumar S, Choudhary S, Rakesh N, Zhu Y, Kaur M, Tandon C and Gadekallu T R 2024 Integrative approach for efficient detection of kidney stones based on improved deep neural network architecture *SLAS technology* **29** 100159
[4] Caglayan A, Horsanali M O, Kocadurdu K, Ismailoglu E and Guneyli S 2022 Deep learning model-assisted detection of kidney stones on computed tomography *International Braz J. Urol* **48** 830–9
[5] Rule A D, Lieske J C and Pais J V M 2020 Management of kidney stones in 2020 *JAMA* **323** 1961–2
[6] Wilcox C R, Whitehurst L A, Cook P and Somani B K 2020 Kidney stone disease: an update on its management in primary care *Br J Gen Pract* **70** 205–6
[7] Chen X, Chen J, Zhou X, Long Q, He H and Li X 2022 Is there a place for extracorporeal shockwave lithotripsy (ESWL) in the endoscopic era? *Urolithiasis* **50** 369–74
[8] Golomb D *et al* 2022 A population based, retrospective cohort study analyzing contemporary trends in the surgical management of urinary stone disease in adults *Can Urol Assoc J* **16** 112–8
[9] Wang K *et al* 2022 Risk factors for kidney stone disease recurrence: a comprehensive meta - analysis *BMCUrol* **22** 62
[10] Forbes C M, McCoy A B and Hsi R S 2021 Clinician versus nomogram predicted estimates of kidney stone recurrence risk *J. Endourol.* **35** 847–52
[11] Black K M, Law H, Aldoukhi A, Deng J and Ghani K R 2020 Deep learning computer vision algorithm for detecting kidney stone composition *BJU Int* **125** 920–4
[12] Grosse Hokamp N *et al* 2020 Dose independent characterization of renal stones by means of dual energy computed tomography and machinelearning:anex-vivostudy *Eur Radiol* **30** 1397–404
[13] Zheng J *et al* 2021 A multicenter study to developa non-invasiveradiomicmodeltoidentifyurinaryinfectionstoneinvivousingmachine-learning *Kidney Int.* **100** 870–80
[14] Abraham A, Kavoussi N L, Sui W, Bejan C, Capra J A and Hsi R 2022 Machine learning prediction of kidney stone composition using electronic health record-derived features *J. Endourol.* **36** 243–50
[15] Cui Y *et al* 2021 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.* **23** 436–45
[16] Sudharson S and Kokil P 2021 Computer-aided diagnosis system for the classification of multi-class kidney abnormalities in the noisy ultrasound images *Computer Methods and Programs in Biomedicine* **205** 106071
[17] Yildirim K, Bozdag P G, Talo M, Yildirim O, Karabatak M and Acharya U R 2021 Deep learning model for automated kidney stone detection using coronal CT images *ComputBiol Med* **135**
[18] Elton D C, Turkbey E B, Pickhardt P J and Summers R M 2022 A deep learning system for automated kidney stone detection and volumetric segmentation on non contrast CT scans *Med. Phys.* **49** 2545–54
[19] Kavoussi N L *et al* 2022 Machine learning models to predict 24 hour urinary abnormalities for kidney stone disease *Urology* **169** 52–7
[20] Zhang H, Wang Y, Dayoub F and Sunderhauf N 2021 Varifocalnet: an iou-aware dense object detector *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition* 8514–23
[21] Asif S, Zheng X and Zhu Y 2024 An optimized fusion of deep learning models for kidney stone detection from CT images *Journal of King Saud University-Computer and Information Sciences* **36.7** 102130.
[22] Patro K K, Allam J P, Neelapu B C, Tadeusiewicz R, Acharya U R, Hammad M, Yildirim O and Pławiak P 2023 Application of kronecker convolutions in deep learning technique for automated detection of kidney stones with coronal CT images *Inf. Sci.* **640** 119005
[23] Baygin M, Yaman O, Barua P D, Dogan S, Turker Tuncer U and Acharya R 2022 Exemplar Darknet19 feature generation technique for automated kidney stone detection with coronal CT images *Artif. Intell. Med.* **127** 102274
[24] Yildirim K, Bozdag P G, Talo M, Yildirim O, Karabatak M and Rajendra Acharya U 2021 Deep learning model for automated kidney stone detection using coronal CT images *Comput. Biol. Med.* **135**
[25] Kazemi Y and Mirroshandel S A 2018 A novel method for predicting kidney stone type using ensemble learning *Artif. Intell. Med.* **84** 117–26
[26] Vasanthi P, Srinivasu L N, Teju V, Sowmya K V, Stan A, Sita V, Miclea L and Stan O 2025 Multiple kidney stones prediction with efficient RT-DETR model *Comput. Biol. Med.* **190** 110023
[27] Ponnambalam M, Ponnambalam M and Jamal S S 2024 A robust color image encryption scheme with complex whirl wind spiral chaotic system and quadrant-wise pixel permutation *Phys. Scr.* **99** 105239
[28] Narayana D S M, Enaganti K K and Mathivanan P 2024 Enhancing image security using novel scrambling and chaotic techniques with ChaCha20 algorithm *2024 15th Int. Conf. on Computing Communication and Networking Technologies (ICCCNT)* (IEEE) 1–6

[29] Devabathini N and Mathivanan P 2023 Sign language recognition through video frame feature extraction using transfer learning and neural networks *2023 Int. Conf. on Next Generation Electronics (NEleX)* 1–6(IEEE)
[30] Kothala L P and Guntur S R 2024 GEL-TTA Net: a Global ensemble learning network for the localization of small-scale and mixed intracranial hemorrhages through test time augmentations *Multimedia Tools Appl.* **84** 1–32
[31] Kothala L P and Guntur S R 2023 An improved mosaic method for the localization of intracranial hemorrhages through bounding box *IEEE 5th Int. Conf. on Cybernetics, Cognition and Machine Learning Applications (ICCCMLA)* (IEEE)