I.J. Information Technology and Computer Science, **2025, 5, 1-11**
Published Online on October 8, 2025 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2025.05.01

A ViT-based Model for Detecting Kidney Stones in Coronal CT Images

An Cong Tran*
Can Tho University, Can Tho, Vietnam
CTU-AIMED Leading Research Team, Can Tho, Vietnam
E-mail: tcan@ctu.edu.vn
ORCID iD: https://orcid.org/0000-0001-6184-909X
*Corresponding author

Huynh Vo-Thuy
Can Tho General Hospital, Can Tho, Vietnam
E-mail: vthuynh2310@gmail.com
ORCID iD: https://orcid.org/0009-0002-4416-9994

Received: 09 May 2025; Revised: 07 July 2025; Accepted: 21 August 2025; Published: 08 October 2025

Abstract: Detecting kidney stones in coronal CT images remains challenging due to the small size of stones, anatomical complexity, and noise from surrounding objects. To address these challenges, we propose a deep learning architecture that augments a Vision Transformer (ViT) with a pre-processing module. This module integrates CSPDarknet for efficient feature extraction, a Feature Pyramid Network (FPN), and Path Aggregation Network (PANet) for multi-scale context aggregation, along with convolutional layers for spatial refinement. Together, these trained components filter irrelevant background regions and highlight kidney-specific features before classification by ViT, thereby improving accuracy and efficiency. This design leverages ViT's global context modeling while mitigating its sensitivity to irrelevant regions and limited data. The proposed model was evaluated on two coronal CT datasets (one public and one private dataset) comprising 6,532 images under six experimental scenarios with varying training and testing conditions. It achieved 99.3% accuracy, 98.7% F1-score, and 99.4% mAP@0.5, higher than both YOLOv10 and the baseline ViT. The model contains 61.2 million parameters and has a computational cost of 37.3 GFLOPs, striking a balance between ViT (86.0M, 17.6 GFLOPs) and YOLOv10 (22.4M, 92.0GFLOPs). Despite having more parameters than YOLOv10, the model achieved a lower inference time than YOLOv10, approximately 0.06 seconds per image on an NVIDIA RTX 3060 GPU. These findings suggest the potential of our approach as a foundation for clinical decision-support tools, pending further validation on heterogeneous and challenging clinical datasets such as small (<2 mm) or low-contrast stones.

Index Terms: Kidney Stone, Coronal CT, Vision Transformer, CSPDarknet, FPN-PANet.

1. Introduction

Kidney stones are a common urological disorder that affects approximately 10% of the global population [1, 2]. The incidence rates are increasing due to factors such as high sodium diets, chronic dehydration, obesity, and metabolic syndromes. In the United States, the prevalence of kidney stones rose from 5.2% in the early 1990s to over 8.8% by the late 2000s [3]. This condition imposes a significant burden on healthcare systems, with annual costs exceeding $10 billion due to hospitalizations and lost productivity [4]. Early and accurate diagnosis is essential, as it allows for timely interventions such as lithotripsy or ureteroscopy, which can help prevent severe complications like renal failure and chronic kidney disease.

Among diagnostic tools, coronal computed tomography (Coronal CT) is widely regarded as the gold standard for diagnosing kidney stones, providing high spatial resolution to identify the location, size, and composition of the stones [5]. However, the manual analysis conducted by radiologists is labor-intensive and often requires 10 to 15 minutes per scan. It is also susceptible to inter-observer variability and subjective errors, particularly for small stones or those hidden by complex anatomical structures such as the renal pelvis or ureteral junctions [6, 7]. These challenges highlight the necessity for automated and reliable detection methods in medical imaging.

Deep learning has significantly enhanced kidney stone detection in medical imaging, with convolutional neural networks (CNNs) establishing crucial benchmarks. Ronneberger et al. introduced U-Net, a groundbreaking architecture for biomedical image segmentation, achieving 92% accuracy in MRI datasets during the MICCAI challenge [8]. This encoder-decoder model has been adapted for CT imaging, including Cicek et al.'s extension to 3D volumetric data, demonstrating strong kidney segmentation performance [9]. However, their model's accuracy fell to 75% in coronal planes, primarily due to geometric distortions and overlapping kidney structures.

Similarly, Tan and Le utilized EfficientNet-B0, a lightweight CNN, to classify medical anomalies, achieving an F1-score of 0.84 on the ImageNet dataset [10]. However, they encountered difficulties in detecting microcalcifications smaller than 3 mm due to subtle radiodensity profiles. This limitation aligns with broader critiques of CNNs, which often depend on local receptive fields, as noted by Litjens et al. [11]. This dependence restricts their capability to model global anatomical relationships. For instance, Roth et al. achieved 87% sensitivity for organ detection in axial CT scans using Faster R-CNN. However, this performance declined to 71% in coronal views due to structural interference from overlapping tissues [12].

Hybrid approaches combining object detection and classification have gained popularity in addressing existing limitations. Jocher et al. [13] introduced YOLOv5, a real-time object detection framework optimized for speed and precision, achieving a mean Average Precision (mAP) of 0.91 on the COCO dataset. Oktay et al. enhanced U-Net with attention mechanisms for multi-modal imaging, leading to an 8% increase in sensitivity (up to 85%) in cluttered environments [14]. In a related effort, Zhou et al. [15] integrated Mask R-CNN with Vision Transformer (ViT) for detecting prostate lesions in ultrasound images, reducing false positives by 20%. Valanarasu et al. combined CNNs with gated axial-attention mechanisms for detecting kidney stones in non-contrast CT scans, achieving a specificity of 80% [16]. Hatamizadeh et al. investigated a hybrid CNN-Transformer for multi-plane analysis, but the lack of coronal-specific ROI optimization allowed noise from adjacent anatomy to degrade performance [17].

Transformer-based models, especially Vision Transformers (ViTs), have emerged as promising alternatives in image analysis. These models utilize the self-attention mechanism to effectively capture long-range dependencies throughout entire images. For instance, Chen et al. developed TransUNet, a hybrid model that integrates ViT with U-Net [18]. This model achieved a Dice coefficient of 0.89 for multi-organ segmentation in abdominal CT datasets. In another advancement, Liu et al. introduced the Swin Transformer, a hierarchical variant of ViT, which outperformed ResNet-50 by 5–10% in natural image classification tasks, achieving an accuracy of 87.3% on ImageNet [19].

However, a significant challenge with Vision Transformers (ViTs) is their dependence on large training datasets, often requiring tens of thousands of images. This poses a particular problem in medical imaging, where annotated data is limited, as pointed out by Dosovitskiy et al. [20]. To address this issue, Hassani et al. introduced the Compact Convolutional Transformer (CCT) [21], which reduces parameters by approximately 30% while retaining competitive accuracy on smaller datasets, thus making it more appropriate for medical applications such as CT analysis.

The coronal orientation presents unique challenges, such as geometric distortions and overlapping anatomical structures, which require specialized solutions beyond those designed for axial or sagittal views. This study proposes a deep-learning model for kidney stone detection in coronal CT images that enhances the Vision Transformer (ViT) through a series of preprocessing layers to improve diagnostic precision. The model incorporates a background removal module featuring CSPDarknet for feature extraction, along with a Feature Pyramid Network (FPN) and a Path Aggregation Network (PANet) for multi-scale feature fusion, and convolutional layers for spatial refinement. Following these layers is a filtering step that isolates kidney regions, ensuring that ViT focuses on relevant features for stone classification. This integrated architecture capitalizes on ViT's strength in global feature modeling, tackling coronal-specific challenges and providing enhanced accuracy and efficiency amid the increasing prevalence of nephrolithiasis. The key contributions of our work include:

- An enhanced ViT-based architecture: The model enhances ViT by adding preprocessing layers, including CSPDarknet, FPN-PANet, and convolutional layers to improve spatial feature extraction and background suppression, enhancing classification performance for kidney stone detection in coronal CT images.
- Kidney stone dataset development: We developed a private coronal CT dataset consisting of 4,733 images, annotated by expert radiologists, capturing a variety of stone sizes and locations for comprehensive validation.
- Good performance and computational balance: The proposed model achieved an accuracy of 99.3%, an F1-score of 98.7%, and 99.4% of mAP@0.5, exceeding the performance of YOLOv10 and baseline ViT through comprehensive benchmarking. It maintains a favorable trade-off between complexity (61.2M parameters, 37.3 GFLOPs) and inference time (0.06 sec/image on RTX 3060), enabling practical use in clinical settings.

This paper is structured as follows: Section 2 describes the proposed model architecture, with a focus on the integration of CSPDarknet, FPN-PANet, and VIT. Section 3 outlines the datasets used and the experimental setup. Section 4 presents the results, including comparisons with baseline models. Section 5 discusses the clinical implications and limitations of the study, followed by the conclusion in Section 6.

2. Proposed Method and Implementation

The proposed methodology for kidney stone detection improves the Vision Transformer (ViT) framework by incorporating specialized layers before the ViT classification stage. This enhancement aims to increase classification accuracy and reduce computational complexity. The backbone of the model, CSPDarknet, serves as an efficient convolutional neural network (CNN) that extracts strong local features while minimizing input redundancy. The architecture also includes a Feature Pyramid Network (FPN) combined with a Path Aggregation Network (PANet), referred to as FPN-PANet. This combination facilitates multi-scale feature fusion, preserving both semantic abstraction and spatial details. Convolutional layers further refine these features spatially, helping to eliminate noisy backgrounds and isolate kidney-related regions for the ViT module. This layered design leverages the relationship between the spatial hierarchies of CNNs and the global attention mechanism of ViTs. By capturing both local and global structures, it enhances classification accuracy while reducing computational demands. Figure 1 illustrates the model architecture, which depicts this integrated pipeline.

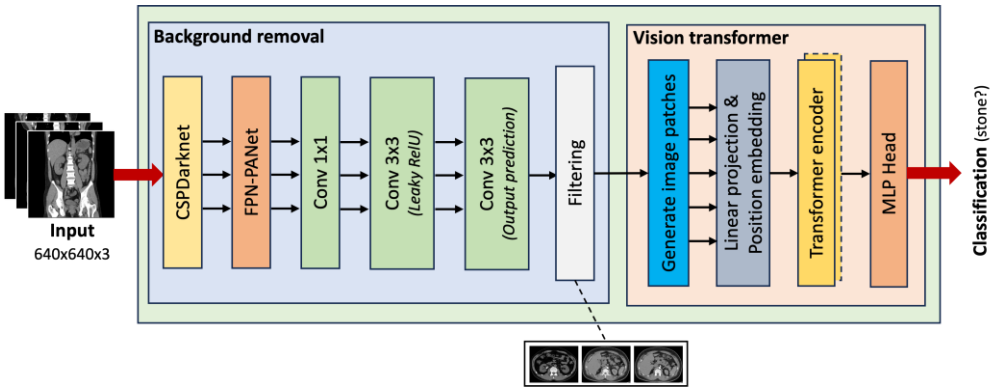


Fig.1. The proposed model architecture

2.1. Background Removal and Feature Extraction

The process begins with background removal using the CSPDarknet backbone, a lightweight feature extractor designed to optimize gradient flow and computational efficiency. CSPDarknet splits the input feature map into two parts, processing one part through a dense block while concatenating the other directly to the output, reducing redundancy. For an input image $I \in \mathbb{R}^{H\times W\times 3}$, CSPDarknet generates a feature map $F_{CSP} \in \mathbb{R}^{H'\times W'\times C}$ representing the reduced height, width, and channel dimensions, respectively.

Following CSPDarknet, a Feature Pyramid Network with a Path Aggregation Network (FPN-PANet) is employed to enhance multi-scale feature fusion. FPN constructs a top-down pathway to propagate high-level semantic features to lower levels, while PANet introduces a bottom-up path to reinforce fine-grained details. The FPN-PANet module takes feature maps at different scales $\{F_1, F_2, ..., F_n\}$, where $F_i \in \mathbb{R}^{H_i\times W_i\times C_i}$, and outputs fused features $\{P_1, P_2, ...P_n\}$. The fusion operation for a given level i in FPN can be expressed as:

$$P_i = \mathrm{Conv}_{1\times 1}(F_i) + \mathrm{Upsample}\left(\mathrm{Conv}_{1\times 1}(F_{i+1})\right) \tag{1}$$

PANet further refines these features by aggregating them through a bottom-up path:

$$Q_i = \mathrm{Conv}_{3\times 3}\left(P_i + \mathrm{MaxPool}\left(P_{i-1}\right)\right) \tag{2}$$

This dual-path mechanism ensures that the model concentrates on regions of interest, such as possible kidney stone locations, by maintaining both semantic and spatial information.

2.2. Convolutional Refinement

The fused feature maps Q_i are then processed through a series of convolutional layers to further refine the representations. First, a 1×1 convolution reduces the channel dimensionality, followed by a 3×3 convolution with Leaky ReLU activation to introduce non-linearity:

$$F_{\mathrm{conv}} = \mathrm{LeakyReLU}\left(\mathrm{Conv}_{3\times 3}\left(\mathrm{Conv}_{1\times 1}\left(Q_i\right)\right)\right) \tag{3}$$

Where $\mathrm{LeakyReLU}(x) = \max(0,x) + \alpha\min(0,x)$, with $\alpha = 0.1$. A final 3×3 convolution generates the output

prediction map, followed by a filtering step to isolate the most salient features. The filtering operation applies a threshold to suppress irrelevant regions, ensuring that only high-confidence features are passed to the Vision Transformer model.

2.3. Vision Transformer for Classification

The filtered feature map $F_{\text{filtered}} \in \mathbb{R}^{H'' \times W'' \times C'}$ is then processed by the Vision Transformer model. The feature map is divided into N non-overlapping patches of size $P \times P$, where $N = \frac{H'' \times W''}{P^2}$. Each patch is flattened into a vector $x_P \in \mathbb{R}^{\left(P^2 \cdot C'\right) \times D}$ and linearly projected to a fixed dimension D using a trainable matrix $E \in \mathbb{R}^{\left(P^2 \cdot C'\right) \times D}$:

$$z_P = x_P E \tag{4}$$

Positional embeddings $E_{pos} \in \mathbb{R}^{N \times D}$ are added to preserve spatial information:

$$z = \left[z_1, z_2, ..., z_n\right] + E_{pos} \tag{5}$$

The output of the transformer encoder is a sequence of embeddings $z_L \in \mathbb{R}^{N \times D}$, which is passed to a Multi-Layer Perceptron (MLP) head for classification. The MLP head consists of two linear layers with a GELU activation and dropout:

$$\text{MLP}(x) = \text{Linear}\left(\text{Dropout}\left(\text{GELU}\left(\text{Linear}(z)\right)\right)\right) \tag{6}$$

The final output is a probability distribution across the classes (e.g., the presence or absence of kidney stones), derived through a softmax function.

The integration of CSPDarknet and FPN-PANet reduces the computational burden by focusing on relevant features early in the model, while the Vision Transformer captures global contextual relationships. This approach achieves a balance between accuracy and efficiency, making it ideal for kidney stone detection in medical imaging.

3. Model Evaluation

In this section, we present the evaluation results of the proposed model, which include a description of the experimental dataset, the metrics utilized in the experiments, the experimental scenarios, the outcomes of the experiments, and an analysis of these results.

3.1. Experimental Datasets

We used two datasets to evaluate the proposed model. A summary is provided in Table 1.

- **Dataset 1**: The Coronal CT Kidney Stone Dataset was collected and annotated by a group of researchers from Firat University in Turkey [22]. The dataset includes data from 433 patients, comprising 278 diagnosed with kidney stones and 165 healthy individuals without them. In total, there are 1,799 images, with 790 being Coronal CT images featuring kidney stones and 1,009 images without them. This dataset is publicly available at https://github.com/yildirimozal/Kidney_stone_detection.
- **Dataset 2**: This is a private dataset collected by the authors from Can Tho City Hospital, Vietnam. It includes Coronal CT images of patients who visited the hospital for diagnosis and treatment between 2020 and 2022. The dataset consists of 4,733 images, including 2,222 coronal CT images of patients with kidney stones and 2,411 coronal CT images of individuals without kidney stones.

Both datasets followed the same CT acquisition protocol: 120 kV, auto tube current (ranging from 100 to 200 mA), a 128 mm detector, 203 mAs, and a slice thickness of 5 mm. Patients were between 18 and 80 years old. Individuals with double-J (pigtail) catheters, single kidneys, renal anomalies, or atrophic kidneys were excluded. All coronal CT images were independently annotated by a radiologist and a urologist based on the visual presence of kidney stones, without performing segmentation. Each image was labeled as either “stone” or “normal” at the image level.

To ensure unbiased evaluation, patient-level separation was strictly enforced between training, validation, and testing sets. However, neither dataset includes structured annotations for stone size or volume, which limits our ability to analyze performance based on stone sizes. We acknowledge this as a limitation and have noted it for future work.

Table 1. Experimental datasets

Dataset	Label	No of images	Size (MB)
1	Normal	1.009	340
	Stone	790	240
	Total	1.799	620
2	Normal	2.411	77
	Stone	2.322	74
	Total	4.733	151
Total		6.532	771

3.2. Experimental Scenarios

We designed six experimental scenarios to thoroughly evaluate the proposed model using the two datasets mentioned above. The experiments focus on two main objectives: (1) comparing the performance of individual models (YOLOv10 and the original ViT) against the proposed model, and (2) assessing the impact of data enrichment during training. Specifically, the scenarios are categorized into two groups:

- Group 1 (scenarios 1–3): Training on Dataset 1 and testing on one-third of Dataset 2.
- Group 2 (scenarios 4–6): Training on Dataset 1 combined with two-thirds of Dataset 2, followed by testing on the remaining one-third of Dataset 2.

Scenario 1 evaluates the kidney stone detection capability of YOLOv10. Scenario 2 focuses on classifying images as containing or not containing kidney stones using Vision Transformer (ViT). Scenario 3 employs our proposed model that improves the ViT model. Scenarios 4 to 6 repeat the procedures of Scenarios 1 to 3 but utilize an expanded training dataset (Dataset 1 + 2/3 of Dataset 2). This aims to assess the model's generalization ability on more diverse and larger datasets. To ensure consistency, all scenarios are tested on the same dataset (1/3 of Dataset 2). A summary of the experimental scenarios is presented in Table 2.

Table 2. Experimental scenarios

Dataset	Model	Training data	Testing data
1	YOLOv10	Dataset 1 *(1.799 images)*	1/3 Dataset 2 *(1.579 images)*
2	ViT		
3	Our model		
4	YOLOv10	Dataset 1 + 2/3 Dataset 2 *(4.953 images)*	
5	ViT		
6	Our model		

YOLOv10 was selected as our baseline due to its state-of-the-art performance in real-time object detection tasks, including applications in medical imaging. Its one-stage detection architecture makes it a strong candidate for comparison in terms of both detection accuracy and computational efficiency. Additionally, the original Vision Transformer (ViT) was chosen as a representative transformer-based model to evaluate the effect of our proposed architectural modifications, particularly under the same training and testing conditions. These two baselines allow for a direct comparison between CNN-based detectors and transformer-based architectures for kidney stone detection in CT images.

3.3. Experimental Configuration

The experimental setup for training and evaluating the proposed stone detection model was carefully designed to optimize performance on Coronal CT images. Hyperparameter values were determined through a systematic grid search and iterative experimentation to balance convergence speed, model stability, and generalization. The finalized hyperparameter settings used in the experiments are as follows:

- Optimizer: The Adam optimizer was selected for its adaptive learning rate properties, with an initial learning rate of 10^{-3} (0.001) for the Vision Transformer (ViT) model and 10^{-2} (0.01) for the YOLOv10 model. These values reflect ViT's sensitivity to smaller updates due to its transformer architecture and YOLOv10's robustness to larger steps during ROI detection training.
- Batch Size: A batch size of 16 was chosen and adjusted to fit the memory constraints of the NVIDIA RTX 3060 GPU (12 GB VRAM) used for training. This size ensured efficient gradient updates while maximizing GPU utilization.

- Loss Function: Categorical cross-entropy was employed as the loss function for both stages. For YOLOv10, this was adapted to a multi-task loss combining classification and bounding box regression (as per its default configuration), while for ViT, it directly optimized binary stone classification (stone/no stone).
- Number of Epochs: Training was conducted with an early stopping mechanism, halting if validation accuracy did not improve after 10 consecutive epochs, with a maximum cap of 100 epochs to prevent overfitting. This approach ensured efficient resource use while capturing optimal model performance.

In terms of hardware, all training and testing processes were performed on a single NVIDIA RTX 3060 GPU with 12 GB of VRAM, paired with an Intel Core i7-11700 CPU and 32 GB of RAM, running on a Linux-based system (Ubuntu 20.04). The software stack included PyTorch 1.10.0 for model implementation, CUDA 11.3 for GPU acceleration, and Python 3.8 for scripting, ensuring compatibility and reproducibility across experiments.

3.4. Evaluation Metrics

Three primary metrics were employed to comprehensively assess the proposed model's performance: Accuracy, F1-Score, and mean Average Precision (mAP). These metrics evaluate classification accuracy, detection robustness, and object localization precision, respectively, aligning with the dual objectives of kidney region detection (Stage 1) and stone identification (Stage 2).

- Accuracy: This metric measures the proportion of correct predictions relative to the total number of predictions. It is particularly relevant for ViT, where binary classification is the final output, though it is also computed globally across the models. It is calculated as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

Where:

TP (True Positive): Images with kidney stones are correctly classified as containing stones.
TN (True Negative): Images without kidney stones are correctly classified as stone-free.
FP (False Positive): Images without kidney stones are incorrectly classified as containing stones.
FN (False Negative): Images with kidney stones are incorrectly classified as stone-free

- F1-Score: The F1-Score balances precision and recall, making it ideal for evaluating performance on potentially imbalanced datasets (e.g., fewer stone-positive images). It is defined as:

$$\text{F1-Score} = \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

Where:

Precision: The ratio of correctly predicted stone instances to all predicted positives, calculated as:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{9}$$

Recall: The ratio of correctly predicted stone instances to all actual positives, calculated as:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{10}$$

- Mean Average Precision (mAP): mAP evaluates the model's ability to localize and classify kidney stones across multiple Intersections over Union (IoU) thresholds (e.g., 0.5 to 0.95). It is computed as:

$$mAP = \frac{1}{N} \sum_{i-1}^{N} AP_i \tag{11}$$

Where:

N: The number of classes (here, $N = 1$ for kidney stones, though kidney detection in Stage 1 could be treated separately).

AP_i: The Average Precision for the i-th class, derived from the precision-recall curve at varying IoU thresholds.

The proposed model does not perform object localization in the traditional sense (e.g., using bounding boxes or segmentation masks). Instead, the CSPDarknet and FPN-PANet modules are designed to enhance spatial features and suppress irrelevant background before the classification stage. The final prediction is a binary decision indicating

whether kidney stones are present in the image or not. Thus, the use of detection-oriented components serves to improve feature richness, but the overall task remains classification-focused. Consequently, the experiments do not use localization metrics like Intersection over Union (IoU) or Dice coefficient.

3.5. Evaluation Results

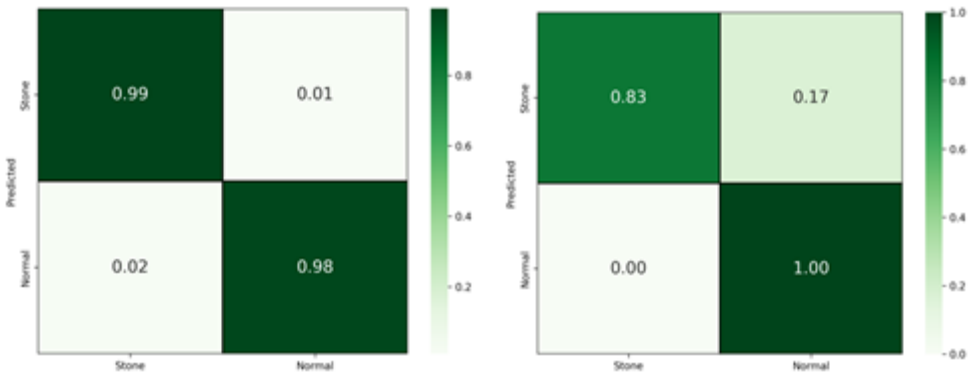
The experimental results, including accuracy-related metrics for the test scenarios on the test dataset, are presented in Table 3.

Table 3. Experimental results

Scenario	Accuracy	Precision	Recall	F1-score	mAP@0.5
1	0.986	0.987	0.983	0.985	0.985
2	0.924	0.833	1.000	0.909	0.922
3	0.983	0.964	1.000	0.982	0.981
4	0.988	0.980	0.992	0.986	0.990
5	0.904	1.000	0.832	0.908	0.922
6	0.993	1.000	0.974	0.987	0.994

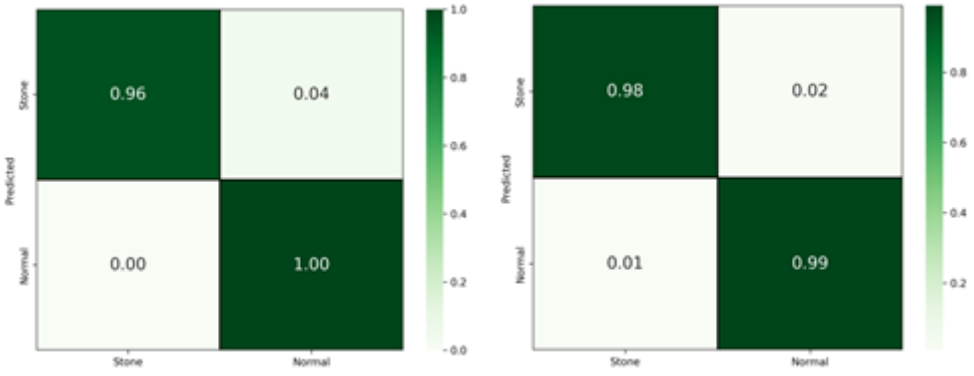
The experimental results highlight the differences in accuracy between individual models and the combined model, as well as the impact of expanding the training dataset. In scenario 1, the YOLOv10 model achieved an overall accuracy (accuracy) of 98.6% and a mAP@0.5 of 98.5%, reflecting its precise ability to detect kidney stones on the original dataset. However, when applying the original Vision Transformer (ViT) in scenario 2, although recall reached 100% (indicating that the model did not miss any cases of kidney stones), precision was only 83.3% due to a high false positive rate, resulting in an F1-score of 90.9%. This demonstrates the limitations of ViT in accurately distinguishing between kidney stones and similar structures when the input images contain background and a limited training dataset.

The proposed model in scenario 3 addressed these limitations, increasing the F1-score to 98.2% and mAP@0.5 to 98.1%. This improvement underscores the importance of the integrated layers, enabling ViT to focus on analyzing the micro-features of kidney stones without interference from complex backgrounds. When the training dataset was expanded (scenarios 4–6), the performance of both individual and combined models continued to improve. Specifically, Scenario 6 (our proposed model) achieved an accuracy of 99.3%, precision of 100%, and mAP@0.5 of 99.4%, demonstrating superior reliability and generalization capabilities.



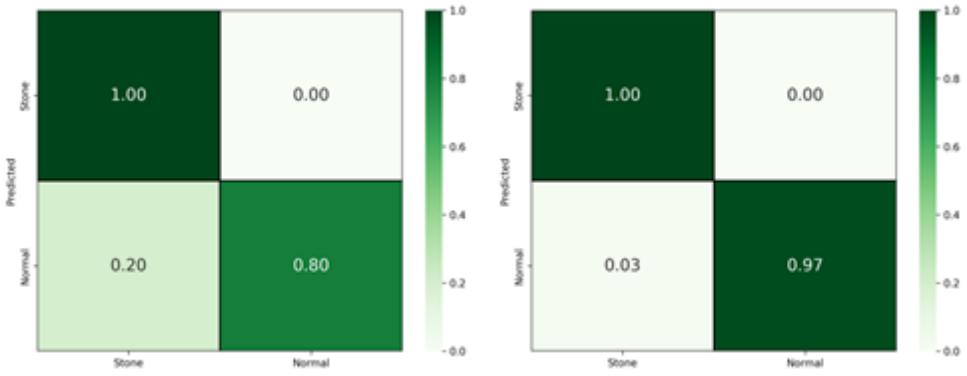
a) Scenario 1

b) Scenario 2



c) Scenario 3

d) Scenario 4



e) Scenario 5 f) Scenario 6

Fig.2. Confusion matrices on the test set

A comparison of the two groups of scenarios reveals that the addition of training data considerably enhanced the performance of YOLOv10: mAP@0.5 increased from 98.5% (scenario 1) to 99.0% (scenario 4), while recall improved from 98.3% to 99.2%. In contrast, for ViT (scenario 2 vs. 5), the expanded dataset achieved 100% precision, but recall decreased to 83.2%, highlighting a trade-off between accuracy and coverage. Nevertheless, when integrating the pre-processing layers (scenario 6), the model maintained equilibrium with an F1-score of 98.7% and a recall of 97.4%, demonstrating the effectiveness of the proposed approaches.

Figure 2 presents the confusion matrices for the six experimental scenarios, offering a comprehensive view of the model's performance in kidney stone detection across different configurations. Scenario 1 demonstrated high accuracy but had a relatively lower recall (98.3%), indicating that while the model detected most kidney stones, it missed a few, resulting in false negatives. Scenario 2 exhibited perfect recall (100%), but the low precision (83.3%) suggested an increase in false positives, highlighting a trade-off between sensitivity and specificity. In scenario 3, our proposed method achieved an accuracy of 98.3% with perfect recall and a precision of 96.4%. This indicates that our method effectively balanced sensitivity and specificity, identifying kidney stones with minimal false positives. Scenarios 4 to 6, which used both datasets 1 and 2, showed improved results, particularly scenario 6. This scenario achieved the highest accuracy (99.3%) and perfect precision (100%), with a recall of 97.4%. These results highlight the effectiveness of training with large datasets, as the model generalized well across diverse data and maintained high sensitivity and specificity.

Moreover, each scenario pair (1 vs. 3, and 4 vs. 6) used the same test data, allowing for controlled comparisons. Therefore, the performance improvements of the proposed model over YOLOv10 and ViT can be directly attributed to architectural enhancements. The consistent improvements across multiple metrics, along with reduced false positives in confusion matrices, indicate that these gains are unlikely to result from random variation.

In addition to model accuracy, we also analyzed the models' training and inference time (testing time). The training and inference times for the models are presented in Figure 3.

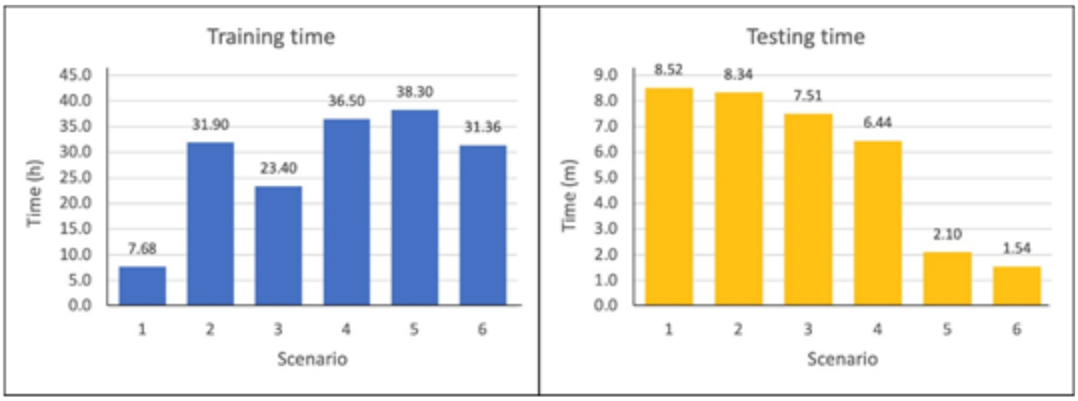


Fig.3. Model training and testing time

The experimental results indicate that the training time of the models varies significantly depending on the architecture used. The YOLOv10 model (scenarios 1 and 4) has the shortest training time, requiring only 7.68 hours when trained on the original dataset (scenario 1) and increasing to 36.5 hours when trained on the expanded dataset (scenario 4). This increase reflects the impact of dataset expansion while maintaining high computational efficiency due to the architecture's optimization for object detection tasks. In contrast, the ViT model (scenarios 2 and 5) requires significantly longer training times, ranging from 31.9 hours (scenario 2) to 38.3 hours (scenario 5). The primary reason is the self-attention mechanism, which consumes substantial computational resources and involves more parameters

than YOLOv10. Notably, when the dataset is expanded, the training time increases to 38.3 hours, highlighting the significant increase in computational complexity as the dataset grows.

The proposed model (scenarios 3 and 6) optimizes training time, requiring 23.4 hours (scenario 3) and 31.36 hours (scenario 6), respectively. This efficiency is achieved by removing the noisy background, which reduces the computational load on ViT while maintaining higher recognition performance. Testing time is critical in evaluating model performance, especially for applications requiring real-time processing. The YOLOv10 model (scenarios 1 and 4) demonstrates improved testing time when using the expanded dataset, decreasing from 8.52 minutes (scenario 1) to 6.44 minutes (scenario 4). This reflects the model's strong generalization ability, enabling it to maintain stable processing speed even as the complexity of the input data increases.

The ViT model (scenarios 2 and 5) exhibits higher testing times, approximately 8.34 minutes in scenario 2, due to its mechanism of processing the entire image rather than focusing on specific regions. However, when the dataset is expanded (scenario 5), the testing time significantly decreases to 2.1 minutes, indicating that the model has learned more generalized features, reducing computational demands during the inference phase. The proposed model (scenarios 3 and 6) achieves higher testing efficiency, with testing time decreasing from 7.51 minutes (scenario 3) to 1.54 minutes (scenario 6). The experimental results demonstrate that YOLOv10 has advantages in both training and testing speed, while ViT provides robust representation learning capabilities at the cost of significant computational resources. The proposed model strikes a balance between performance and processing time, optimizing both accuracy and computational efficiency.

Table 4. Computational comparison

Model	Params (M)	FLOPs (G)
YOLOv10	22.4	92.0
ViT (Baseline)	86.0	17.6
Our model	61.2	37.3

To evaluate efficiency, we also compare the number of parameters and FLOPs of our model with those of baseline methods. Table 4 provides a comprehensive computational comparison. Our proposed model has 61.2 million parameters and 37.3 GFLOPs, which lies between ViT (86 million parameters, 17.6 GFLOPs) and YOLOv10 (22.4 million parameters, 92 GFLOPs). Although ViT has fewer FLOPs, its transformer operations on the whole input image lead to longer inference times. Our model achieves a good balance by reducing the full-image processing cost of ViT through the incorporation of early-stage filtering. The balance of efficiency and accuracy underscores the model's suitability for real-time clinical deployment.

4. Conclusions

This study introduces a method for kidney stone detection in coronal CT images, leveraging a Vision Transformer (ViT) enhanced by CSPDarknet, FPN-PANet, and convolutional layers. The proposed model achieves superior performance, with an accuracy of 99.3%, an F1-score of 98.7%, and an mAP@0.5 of 99.4%, outperforming YOLOv10 and the original ViT. Incorporating Dataset 2 into training further improves mAP@0.5 by 1.3% and reduces inference time from 7.51 minutes to 1.54 minutes on the test dataset, demonstrating enhanced generalization and computational efficiency. As shown in Figure 3, the model maintains an average inference time of approximately 0.06 seconds per image, supporting near-real-time deployment in clinical settings such as PACS systems for rapid triage or second-opinion support. This efficiency stems from early-stage background filtering and streamlined ViT input, despite the complexity of multiple architectural components.

From a clinical perspective, the model's high precision and recall position it as a valuable diagnostic support tool for radiologists, potentially flagging kidney stones in coronal CT scans to reduce reading time and oversight. Its integration into PACS-based decision-support systems could enhance pre-screening workflows. However, limitations remain, including the need for improved performance on very small (<2mm) stones and low-contrast images, as well as the limited representation of global CT scanner heterogeneity and imaging protocols in the current datasets. Validation against radiologist performance is also essential to confirm clinical applicability. Real-world implementation would require further validation on multi-institutional data, seamless integration with hospital IT systems, and evaluation of user trust and interpretability.

Future research will focus on addressing existing limitations and enhancing the model's capabilities. To improve performance on small and low-contrast stones, we will investigate advanced image enhancement techniques and fine-tune the model using targeted datasets that contain challenging cases. To address scanner variability, we will expand the training data through multi-center collaborations to encompass diverse CT. We will also validate our model against radiologist performance through comparative studies to assess clinical equivalence and practical utility. Additionally, prospective clinical trials will evaluate diagnostic accuracy, radiologist confidence, and patient outcomes, while explainable AI techniques will be developed to enhance interpretability and user trust.

Statistical significance testing will validate performance gains, and we will benchmark our model against ViT-based object detection frameworks such as DETR and ViT-FRCNN to provide a comprehensive comparative analysis. Furthermore, we plan to extend the current classification-focused framework to support object localization or segmentation through annotated bounding boxes or masks. This would enable the model to produce spatial predictions and support evaluation using localization metrics such as Intersection over Union (IoU) or Dice coefficient, paving the way for robust real-world deployment.

References

[1] V. Romero, H. Akpinar, and D. G. Assimos, "Kidney stones: A global picture of prevalence, incidence, and associated risk factors," *Reviews in Urology*, vol. 12, no. 2-3, pp. e86, 2010.

[2] I. Sorokin, C. Mamoulakis, K. Miyazawa, A. Rodgers, J. Talati, and Y. Lotan, "Epidemiology of stone disease across the world," *World Journal of Urology*, vol. 35, no. 9, pp. 1301–1320, 2017.

[3] C. D. Scales Jr, A. C. Smith, J. M. Hanley, C. S. Saigal, and Urologic Diseases in America Project,, "Prevalence of kidney stones in the United States," *European urology*, vol. 62, no. 1, pp. 160–165, 2012.

[4] M. S. Pearle, E. A. Calhoun, G. C. Curhan, and Urologic Diseases of America Project, "Urologic diseases in America project: Urolithiasis," *The Journal of Urology*, vol. 173, no. 3, pp. 848–857, 2005.

[5] R. Eliahou, G. Hidas, M. Duvdevani, and J. Sosna, "Determination of renal stone composition with dual-energy computed tomography: An emerging application," *Seminars in Ultrasound, CT and MRI*, vol. 31, no. 4, pp. 315–320, Elsevier, 2010.

[6] M. W. Ciaschini, E. M. Remer, M. E. Baker, M. Lieber, and B. R. Herts, "Urinary calculi: radiation dose reduction of 50% and 75% at CT-Effect on sensitivity," *Radiology*,vol. 251, no. 1, pp. 105–111, 2009.

[7] R. P. Reimer, K. Klein, M. Rinneburger, D. Zopfs, S. Lennartz, J. Salem, A. Heidenreich, D. Maintz, S. Haneder, and N. Große Hokamp, "Manual kidney stone size measurements in computed tomography are most accurate using multiplanar image reformatations and bone window settings," *Scientific Reports*, vol. 11, no. 1, p. 16437, 2021.

[8] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Medical image computing and computer-assisted intervention–MICCAI 2015: 18th International Conference*, Munich, Germany, pp. 234-241, 2015.

[9] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 424–432, 2016.

[10] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Proc. International Conference on Machine Learning*, pp. 6105–6114, 2019.

[11] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[12] H. R. Roth, L. Lu, A. Farag, H.-C. Shin, J. Liu, E. B. Turkbey, and R. M. Summers, "DeepOrgan: Multi-level deep convolutional networks for automated pancreas segmentation," in *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 556–564, 2015.

[13] G. Jocher et al., "Ultralytics/yolov5: v6.0 - yolov5n 'Nano' models," *Roboflow integration, TensorFlow export, OpenCV DNN support*, vol. 10, 2021.

[14] O. Oktay et al., "Attention U-Net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[15] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: Redesigning skip connections to exploit multiscale features in medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 39, no. 6, pp. 1856–1867, 2021.

[16] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, and V. M. Patel, "Medical transformer: Gated axial-attention for medical image segmentation," in *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer International Publishing, 2021.

[17] A. Hatamizadeh et al., "Unetr: Transformers for 3d medical image segmentation," *in Proc. IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 574–584, 2022.

[18] J. Chen et al., "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[19] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10012–10022, 2021.

[20] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[21] A. Hassani et al., "Escaping the big data paradigm with compact transformers", *arXiv preprint arXiv:2104.05704*, 2021.

[22] K. Yildirim et al., "Deep learning model for automated kidney stone detection using coronal CT images," *Computers in biology and medicine*, vol. 135, p. 104569, 2021.

Authors' Profiles

An Cong Tran is a senior lecturer at the College of Information and Communication Technology, Can Tho University, Vietnam. He holds a Master's Degree (Hons) in Computer Science from the Asian Institute of Technology (AIT), Thailand, and a Ph.D. in Computer Science from Massey University, New Zealand. His current research interests encompass description logic learning, ontology learning, applications of blockchain in the public sector, and deep learning methods.

Huynh Vo-Thuy is an IT manager at Can Tho City General Hospital, Vietnam. She received a Master's Degree in Information Systems at Can Tho University, Vietnam. She is passionate about applying information systems to address real-world challenges, particularly in data management and decision support systems. With over 10 years of experience at Can Tho City General Hospital, she has played a key role in implementing IT solutions to enhance medical examination, treatment management, and patient data storage.

How to cite this paper: An Cong Tran, Huynh Vo-Thuy, "A ViT-based Model for Detecting Kidney Stones in Coronal CT Images", International Journal of Information Technology and Computer Science(IJITCS), Vol.17, No.5, pp.1-11, 2025. DOI:10.5815/ijitcs.2025.05.01