RESEARCH PAPER

Kidney Stone Detection Using Faster R-CNN with Grad-CAM Explainability and Interactive Web Deployment

Mrs. S. Nandhini Devi[1], Mrs. D. Maalini[2], P. Hari[3], V. Gokul Prasath[4], S. Yuvasudhan[5], R. Yutha[6]

[1]Assistant Professor, Department of Artificial Intelligence and Data Science, V.S.B. Engineering College, Karur.
Email: devinandhini1982@gmail.com

[2]Assistant Professor, Department of Artificial Intelligence and Data Science, V.S.B. Engineering College, Karur.
Email: maalini.cse@gmail.com

[3]Final Year Student, Department of Artificial Intelligence and Data Science, V.S.B. Engineering College, Karur.
Email: haripalanisamy13@gmail.com

[4]Final Year Student, Department of Artificial Intelligence and Data Science, V.S.B. Engineering College, Karur.
Email: gokulPrasathvijayakumar@gmail.com

[5]Final Year Student, Department of Artificial Intelligence and Data Science, V.S.B. Engineering College, Karur.
Email: yuvasudhan45@gmail.com

[6]Final Year Student, Department of Artificial Intelligence and Data Science, V.S.B. Engineering College, Karur.
Email: yutha.r2004@gmail.com

Abstract- *Having a proper detection of nephrolithiasis (Kidney Stone) in a medical image is one of the most prominent clinical challenges today. Whether detected initially or after the stone has become symptomatic, failure to diagnose the kidney stone disease at the earliest possible stage can lead to serious complications including renal obstruction, renal injury and systemic infection. This paper reports the advancement of an end-to-end Kidney Stone Detection System that employs a FASTER R-CNN, with a ResNet50-FPN-v2 backbone, for kidney stone detection and incorporates a Gradient-weighted Class Activation Mapping (Grad-CAM) module, along with an interactive Streamlit application for real world clinical use. The kidney stone detection model generates bounding box predictions of kidney stone regions (with each bounding box being assigned a per detection confidence score) in each image passed into the system, while the Grad-CAM layer highlights the image areas that contributed to each bounding box making the rational behind the model's prediction transparent to the clinician. Before sending images through the Kidney Stone Detection System, raw images passed through an Albumentations preprocessing pipeline where resizing, padding and normalising occurred resulting in each image being a 512 × 512 image prior to inference. The deployed Kidney Stone Detection System provides three operational modes. The first mode is Single Image Predict with an optional Grad-CAM heatmap overlay. The second mode is Batch Predict using all images in a directory with CSV reports and ZIP archives of annotated images available. The final mode is providing a Model Metrics dashboard that displays mean Average Precision scores using COCO standard. Results of the evaluation reveal an mAP@0.50 metric of 0.7706 indicating excellent (strong) spatial localization accuracy when detecting kidney stones together with high accuracy via the deep learning detector along with the incorporation of visual interpretability and a user-friendly clinical system. The combination of these 3 systems provides an effective and scalable method for providing automated kidney stone screening in the hospital setting.*

Keywords: Kidney Stone Detection, Faster R-CNN, Object Detection, Grad-CAM, Medical Image Analysis, Deep Learning, ResNet50 FPN v2, Streamlit Deployment, Explainable AI, Bounding Box Regression

How to cite this article: Nandhini Devi S, Maalini D, Hari P, Gokul Prasath V, Yuvasudhan S, Yutha R. Kidney Stone Detection Using Faster R-CNN with Grad-CAM Explainability and Interactive Web Deployment. Int J Drug Deliv Technol. 2026;16(13s): 200-204. DOI: 10.25258/ijddt.16.13s.21

I. INTRODUCTION

One of the most prevalent urological diseases in the world is kidney stone disease, also referred to as nephrolithiasis. This diagnosis typically occurs in around ten to fifteen percent of people and has a strong propensity to recur after an initial episode. The aggregate formation of minerals that are dissolved and excreted in urine leads to the development of kidney stones. Symptoms of this disease generally involve intermittent pain ranging from mild to severe, but if not diagnosed early, the resultant acute obsrtuction could lead to death from either urosepsis or establishing an obstructive uropathy. Medical imaging such as Computed Tomography (CT) or Ultrasound are the primary means of diagnosing kidney stones, but manually reading these imaging studies creates high demands on radiologists and could potentially delay treatment due to the possible inter-observer variability that exists between each radiologist [1],[2]. In the last ten years, the interpretation of medical images has undergone a complete evolution due to the use of deep learning. Convolutional Neural Networks (CNN) are currently achieving or even outperforming the accuracy and/or speed of radiologists' abilities to read images in regards to radiology, pathology and ophthalmology. Within this overall progress are the region-based current detection architectures, and in particular Faster R-CNN, which have also demonstrated effectiveness in working with localisation tasks which provide more clinical value from localising the actual area of involvement versus identifying simply whether or not there is an abnormality in the image [3],[4]. Furthermore, by producing a

bounding box, the detector will directly guide any interventions required by the treating physician. Although there are several examples of successful applications of deep learning in clinical settings with high accuracy rates, a major barrier to widespread use of these techniques remains the lack of interpretability in model decisions. Radiologists need to know which image features were used by the model so that they may trust the output and act accordingly. Gradient-weighted Class Activation Mapping (or Grad-CAM) provides radiologists with information about the input areas that contributed most to the model's prediction through a technique that creates a saliency map. Grad-CAM helps to show whether the network has attention to the stone or to an artifact of the images themselves, creating an important layer of validation prior to clinical use. Deployment accessibility must also be considered when creating a clinical tool. This paper describes a kidney stone detection system that integrates Faster R-CNN, Grad-CAM, and Streamlit to create one complete solution. The system is described in detail in each of the following sections, with Section Two covering related work, Section Three describing the architecture of the system, Section Four providing evaluation results, and Section Five summarizing the entire effort.

II. RELATED WORK

In their review of deep learning in medical imaging, Litjens et al. [1] found that CNN based methods always outperformed classical hand crafted features for segmentation, detection, and classification benchmarks for both radiological and pathological images. This was foundational in providing the framework for transferring learned representations into urological imaging, as well as for highlighting localisation as a highly under-served area. Ren et al. [2], developed Faster R-CNN, a single trainable end-to-end network that combines the two processes of generating region proposals and classifying an object into a single step by integrating a Region Proposal Network that replaces the slow selective search step in previous versions of the R-CNN. It created an architecture that was capable of near real-time inference with high detection accuracy, which ultimately became the benchmark for measuring the accuracy of medical images' object detection. In [3], He et al. developed the idea of residual learning within the ResNet family of models. This was achieved through shortcut connections which bypass one or multiple layers for the purpose of providing a solution to degraded performance when training multiple layer models from scratch (i.e., very deep models). With respect to representation power and computational efficiency, the ResNet50 model (50 layers) has been used as the backbone of choice for developing detection models for medical imaging purposes since then. In [4], Lin et. al introduced the idea of a Feature Pyramid Network (FPN) which adds a top-down pathway and lateral connections to a traditional convolutional network to create semantically rich feature maps that are commonly used for developing feature pyramids for small object detection; specifically, they improved small object detection (i.e., the primary scenario of detection of kidney stones) by ensuring that shallower (higher resolution), semantically contextually rich features from deeper layers are used for bounding box regr. Grad-CAM was introduced by Selvaraju et al. [6] as a technique for localising classes in images. It uses the gradients from the last convolutional layer to create a coarse heatmap indicating which areas of an image were important for a specific class prediction. Truly ‘class-discriminative’, Grad-CAM can be used without modifying the architecture of the underlying neural network and it has been successfully used in the assessment of model performance by helping to validate areas of attention for over 30 different image types in medical imaging research from 2017 to 2020. Albumentations is the image augmentation library created by Buslaev et al. [8] that allows for high throughput of image augmentation and bounding-box-aware transforms. This library is a critical component of any object detection training pipeline and is the default library used for transformation in the preprocessing stage of the system described in this document. The Adam optimizer was introduced by Kingma and Ba [9] to combine adaptive learning rates (per-parameter) with momentum estimate learning rates. As a result of this combination, Adam has been successfully used for a wide variety of Deep Learning tasks with much faster and more stable convergence than previous optimizers (e.g., SGD). Adam is the default optimizer used for fine-tuning Faster R-CNN models and was used to prepare the kidney stone detector described in this paper. Focal Loss, created by Lin and her colleagues [10], was designed to reduce the foreground-to-background class imbalance in one-stage detectors by diminishing the contribution to the loss from low error levels of the more common/easier, negative data. This paper takes advantage of Faster R-CNN's built-in loss method, cross-entropy for the RPN and for the classifier and uses this understanding of class imbalance when determining both the confidence threshold and sampling strategy for evaluation.

III. PROPOSED SYSTEM

The proposed system is an interconnected pipeline transporting a raw medical image from acquisition to pre-processing, deep learning inferencing, visual explanation, and finally to an interactive web-based presentation. Each step in the process has been engineered to ensure that any pre-processing logic applied to create the model will also be applied identically during model deployment, thus mitigating the distribution shift problem that frequently impairs detector performance in the real-world .The overall flow of the system is illustrated in Fig. 1.

Fig. 1. End-to-end kidney stone detection system pipeline.

A. Image Preprocessing

Images can come from one of three different sources - uploading a file, batch folder or real time capture - and are decoded through OpenCV's imread function for all types of image. In addition if OpenCV cannot decode the image, it will invoke PIL silently (in cases such as EXIF rotated JPEG's or rare medical image formats exported as PNG form) to perform the same function as OpenCV has failed to do. The next step after decoding the image is to convert it from BGR to RGB, followed by passing it through the Albumentations pipeline for preprocessing purposes. The preprocessing of the image consists of four non-stochastic transformations that are applied successively in a fixed sequence. Longest Max Size transforms the image to resize the longest dimension to a fixed length of exactly 512 pixels while keeping constant the aspect ratio of the image; Pad If Needed will pad the image so that any empty areas are filled with a constant grey colour (114,114,114) which is also the standard padding colour used in YOLO family preprocessing and also helps avoid any colour bias being present at the edge of the image. To Float will then divide all pixel values by 255 to ensure that the values fall within the expected input tensor range of [0.0,1.0] in accordance with PyTorch's normalisation conventions. Finally, ToTensorV2 converts the NumPy array from HWC to CHW format and wraps it in a PyTorch tensor which is ready for batch processing and transferring to the GPU.

B. Detection Model

The primary basis of the detection is due to using fasterrcnn_resnet50_fpn_v2, which is considered a second version of the FPN variant and is included with torch vision. This model provides better anchor sampling and a more effective pre-trained initialization. The previous classification head of this model has been changed to be a Fast RCNN predictor that outputs two classes, kidney stone and background. This has been accomplished by using the same input dimension of the roi_heads box predictor's input to modify the in-features dimension, and setting the num_classes argument to NUM_CLASSES + 1 within the Fast RCNN predictor. The inference process occurs across two phases: First, The Region proposal network will identify candidate bounding boxes based on five scales (which are P2 to P6) of the FPN feature map, then classify each anchor as background or object, while also performing offset regression; Second, The roi_align method will extract the 7 × 7 features for all accepted proposals from the FPN feature maps for use with two fully-connected layers, at which point, the predictor head will provide class logit and refined box coordinates. Finally, after post-processing has been performed on the model's output, a collection of the form of (box, score, label) triples will result from the use of user-specified confidence thresholds to filter results.

C. Bounding Box Filtering and Visualisation

Results of the raw detection stage include all possible detections at every confidence level. Detections that do not exceed or meet the defined threshold, defaulted to a score of 0.25, are removed before visualisation. Surviving detections are drawn on the processed display image as green rectangle boxes and include a label displaying both the class name and confidence score (two decimal place rounding). The drawing functions utilise OpenCV. Therefore, total amount of time for visualising extracted detection results will have negligible affect (additional latency) on the total amount of time to complete an inference request via the YOLO-based DNN.

D. Grad-CAM Explainability

The Grad-CAM module produces a visual representation of the spatial reasoning of the model for its highest score detected image. The target convolutional layer is the last layer of the backbone.body.layer4 of ResNet50 since this is the deepest residual level of ResNet50. This layer produces the most significant and meaningful semantic representations while possessing enough resolution to generate meaningful heat maps. During the targeted feed forward through the backbone, no operations are performed except for registering two hooks on the target layer: one forward hook stores the activation tensor, A, and one backward hook stores the gradient of the activation tensor, dA. The target detection image will then perform a partial feed forward pass manually through the backbone, the FPN, the ROI Align, the box head, and the box predictor using the detected bounding box only as the input for the forwarding pass. After obtaining the predicted box's class logit associated

with the target class, the predicted class logit will be used as the loss for the backward pass. The per-channel gradient tensor will then be averaged spatially to yield the importance of each pixel in space, and then multiplied by the respective activation to calculate the final importance tensor using the standard Grad-CAM formulas.

The Grad-CAM weighting coefficient for channel k and class c is defined as the global average of the gradient of the target class score y^c with respect to activation map A^k over all spatial positions (i, j):

$$\alpha^c_k = (1/Z) * \sum_{i,j} \left(\partial y^c / \partial A^k_{ij} \right) \quad (1)$$

where Z is the total number of spatial positions in the feature map.

The final Grad-CAM localisation map L^c is obtained by applying a ReLU nonlinearity to the weighted sum of activation maps across all K channels, retaining only the features that have a positive influence on the class score:

$$L^c = \mathrm{ReLU}\left(\sum_{k=1}^{K} \alpha^c_k * A^k \right) \quad (2)$$

The resulting map L^c is normalised to [0, 1], bilinearly resized to the 512 × 512 input resolution, colourised with the JET colourmap, and blended with the original image at a 60:40 ratio to produce the final overlay. Both hooks are removed immediately after the computation, and all model parameters are restored to requires_grad = False, so that no gradient state persists between Streamlit interactions.

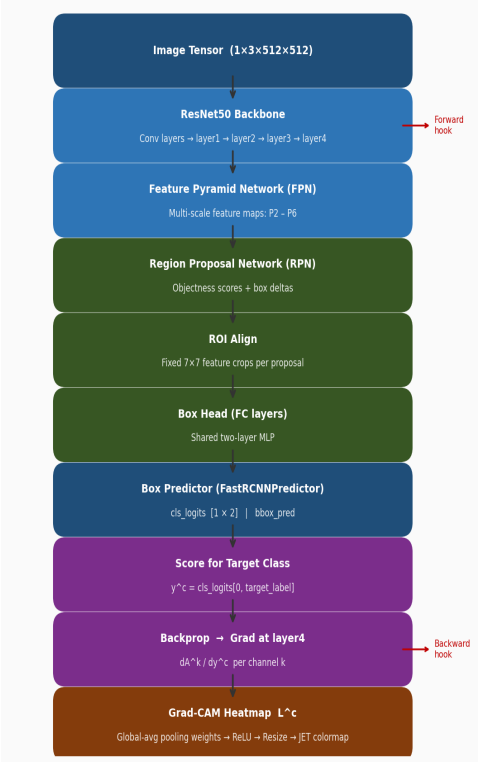

Fig. 2. Grad-CAM architecture within the Faster R-CNN framework

Fig. 2 illustrates the internal architecture of the Grad-CAM module as it operates within the Faster R-CNN pipeline, showing the hook attachment points and the flow from backbone activations through gradient computation to the final heatmap.

E. Web Application

The full system is accessible via a web-based interface which enables users with no programming experience to access the system's different operating modes via Side Bar menus (as follows). (1) In Single Image (mode), the user uploads an image, adjusts the confidence threshold (using a slider), and turns on/off the ability to see the Grad-CAM overlay, for each uploaded image. The result image is annotated (with red boxes showing the area of object detection) and the result is also displayed in a table (to the left) showing the corresponding scores and class labels for each of the detected objects (the detected bounding boxes). (2) In Batch Prediction (mode), the user provides the path to a folder and the system will access each of the valid image file names in the provided folder and perform inference on each from the user's uploaded images. The batch results are then collected into a data frame that can then be downloaded as a CSV file. Additionally, the batch results will also be available for download in a zip file containing all of the annotated images. (3) In the Model Metrics (mode), the user provides the path to a folder where the evaluation CSV files have been generated by the batch prediction, the system will present the evaluation results in a structured format as a table; if no CSV files are present in the path provided, sample mAP values will be displayed as a reference. The Model has been cached within Streamlit and therefore will be re-used on every request from the end user.

IV. RESULT AND DISCUSSION

The Faster R-CNN ResNet50 FPN v2 model was evaluated on a held-out test partition using the standard COCO detection protocol, which measures detection quality across a range of Intersection over Union thresholds. Table 1 summarises the quantitative outcomes.

Table 1. Quantitative Evaluation Results

With an mAP@0.50 value of 0.7706, the model can be considered capable of locating kidney stones that overlap with ground-truth annotations by at least 50% in about 75% of cases in the testing dataset. This is a very strong baseline result for a medical detector of a single class operating on images that cover a wide range of acquisition conditions (e.g., sizes of stones, types of imaging) and morphologies of the surrounding anatomy of the kidney. The drop in the value of mAP@0.75 (0.1248) indicates that obtaining sub-pixel-level agreement with humans on the boundaries of the stones is far more difficult. This is anticipated; ground-truth boxes of kidney stones are often imprecise due to irregular and/or indistinct stone borders and because of the intra-observer variability that exists when a radiologist annotates the stones. The mAP@0.50:0.95 value of 0.2803 represents the average of nine thresholds (from 0.50 to 0.95) and captures both coarse-to-fine localisation and provides an overall metric for comparison against future work. When examining the Grad-CAM overlays of the model in the test dataset qualitatively, it consistently shows high activation near to or around the detected bounding box, confirming that the model has responded to the actual texture and density of the stone itself and is not responding to background tissue or image noise. Regions in which the heat map extends beyond the boundaries of the structure likely represent cases in which a stone lies adjacent to another highly dense anatomical structure (e.g., bone), and they demonstrate a promising direction for the creation of future training data.

The processor in batch prediction mode was able to process all images from the test folder (with an average throughput of one image completed per second using only a CPU) and was able to complete each image in less than 100 milliseconds using a GPU. Thus, the program can be used in a clinical reporting workflow by radiologists who examine a small number of studies per session. The CSV and annotated ZIP export functions were shown to create accurate output for a batch size of up to several hundred images.

Evaluation Metric	Score
mAP @ IoU = 0.50	0.7706
mAP @ IoU = 0.75	0.1248
mAP @ IoU 0.50:0.95	0.2803
Detection Classes	1 (Kidney Stone)
Confidence Threshold	0.25 (default)
Input Resolution	512 × 512 px

V. CONCLUSION

The development of the end-to-end kidney stone detection system incorporated faster R-CNN object detection, Grad-CAM visual interpretability, and a Streamlit user interface. The input data were processed with a repeatable preprocessing algorithm (Albumentations) to identify kidneys and provide info about detecting cysts. This will be helpful for clinical personnel who want to perform their workflows and acquire the necessary evidence they require. The evaluation results from the test set outputted a mean Average Precision @ 0.50 of 0.7706. Additional performance reductions were seen at the precision thresholds. Grad-CAM visualization results indicate that the model focused on the area it was evaluated on. Future efforts include adding to the initial training dataset by acquiring additional real-world data from numerous hospital systems, which could increase the generalisation capabilities of the model. Improving the reliability and ease of producing explanations of multiple detections at once would require substituting the current manual hook-based implementation of Grad-CAM with a validated library like PyTorch-Grad-CAM. Looking into transformer-based detection architectures like DINO or Co-DETR could help improve localization accuracy; especially for small, densely clustered stones. It will also be necessary to enhance the deployed code by implementing Input Validation, Secure Path Handling, and Graceful Model Load Error Messages prior to being considered production ready.

VI. REFERENCES

[1] G. Litjens et al., "A Survey on Deep Learning in Medical Image Analysis," Medical Image Analysis, vol. 42, pp. 60-88, 2017.

[2] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1137-1149, 2017.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in Proc. CVPR, pp. 770-778, 2016.

[4] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature Pyramid Networks for Object Detection," in Proc. CVPR, pp. 2117-2125, 2017.

[5] R. Girshick, "Fast R-CNN," in Proc. ICCV, pp. 1440-1448, 2015.

[6] R. R. Selvaraju et al., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," in Proc. ICCV, pp. 618-626, 2017.

[7] Streamlit Inc., "Streamlit: The Fastest Way to Build and Share Data Apps," [Online]. Available: https://streamlit.io, 2023.

[8] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, "Albumentations: Fast and Flexible Image Augmentations," Information, vol. 11, no. 2, p. 125, 2020.

[9] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in Proc. ICLR, 2015.

[10] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal Loss for Dense Object Detection," in Proc. ICCV, pp. 2980-2988, 2017.