Innovative Research Journal of **Artificial Intelligence** (IRJAI)
2024; 2(1):1-14
DOI: https://doi.org/10.62497/irjai.145

INNOVATIVE RESEARCH JOURNALS

Research Article

Simulation-Based Algorithm Design: Deep Learning or Machine Learning for Finding the Number of Stones in CT Scan/MRI of Kidney

Zia Ur Rehman
AI, MSCS, Department of Computer Science, Kohat University of Science and Technology, Kohat, Pakistan

Citation: Rehman ZU. Simulation-Based Algorithm Design: Deep Learning or Machine Learning for Finding the Number of Stones in CT Scan/MRI of Kidney. IRJAI. 2024; 2(1):1-14. DOI: https://doi.org/10.62497/irjai.145 Available from: https://irjpl.org/irjai/article/view/145

Article Info
Received: April 19, 2024
Revised: June 11, 2024
Accepted: June 12, 2024

Keywords
Kidney stones, Nephrolithiasis, CT imaging, MRI imaging, Deep learning, Machine learning, Hybrid algorithm, Clinical decision support

Abstract

Background: Kidney stone disease (nephrolithiasis) is a prevalent urological condition. While modern imaging modalities such as CT and MRI enable rapid detection and localization of stones, automatically estimating stone counts remains challenging due to variations in size, contrast, resolution, and anatomical positioning. **Objective:** To develop and evaluate a multimodal algorithm that enhances the accuracy and robustness of automatic kidney stone detection and counting across CT and MRI imaging. **Methods:** A hybrid framework, StoneNet-HC, was designed, combining a lightweight convolutional neural network (TinyResNet-FeatureNet) for stone region detection with a Random Forest regression model for predicting stone counts. The approach incorporated multimodal datasets, including publicly available CT scans and synthetically generated MRI images simulating low-contrast conditions. Synthetic augmentation techniques were applied to improve generalizability. Performance was assessed against existing methods using mean absolute error (MAE), Dice coefficient, Intersection over Union (IoU), and classification accuracy. **Results:** StoneNet-HC achieved lower MAE, improved Dice and IoU scores, and higher classification accuracy compared to state-of-the-art approaches. The system demonstrated consistent performance across both CT and MRI modalities, showing resilience to contrast variability and resolution differences. **Conclusion:** This study presents a simulation-driven, hybrid algorithm that integrates deep learning detection with machine learning regression to enable accurate and generalizable kidney stone quantification. The modular design supports potential integration into clinical diagnostic workflows, bridging high accuracy with improved interpretability for multimodal imaging analysis.

** Corresponding Author:*
Zia Ur Rehman
Department of Computer Science, Kohat University of Science and Technology, Kohat, Pakistan
Email: ziagulzia@gmail.com

Introduction

Kidney stone disease, or nephrolithiasis, affects almost 10% of all people at one time or another in their lifetime, as noted by [1]. Proper diagnosis and treatment of renal calculi rely significantly on imaging modalities chiefly, non-contrast computed tomography (NCCT), which is still considered the gold standard due to its high sensitivity. In certain clinical scenarios where radiation exposure is a concern [2] and [3] report that MRI is increasingly preferred as a safer alternative. One key aspect of stone evaluation is not only identifying the presence of calculi but also determining the exact number of stones, since this directly impacts treatment decisions such as extracorporeal shock wave lithotripsy (ESWL), ureteroscopy, or percutaneous nephrolithotomy.

Stone counting traditionally has been based on subjective interpretation by radiologists, which is prone to inter-observer variability. As artificial intelligence (AI) continues to gain traction in medical imaging, both ML and DL techniques have shown promise in automating diagnostic tasks. While convolutional neural networks (CNNs) have been widely employed for segmentation and lesion detection, their direct application to quantitative problems like kidney stone counting remains relatively limited and under-explored.

State-of-the-art AI solutions for urolithiasis still suffer from several notable limitations. DL models, despite their impressive capacity, require large volumes of annotated data and often lack generalizability across imaging modalities such as CT and MRI. In contrast, traditional ML approaches are more data-efficient and interpretable, yet struggle with the spatial complexity intrinsic to medical images. Additionally, most existing research has focused on the detection or localization of kidney stones, with enumeration, a key factor in treatment planning, receiving less attention. Cross-modality adaptability is another critical but underexplored issue, as models trained solely on one modality often perform poorly when applied to another imaging type, as emphasized by [4] and [5].

To counter these challenges, this work proposes a simulation-based hybrid deep learning–machine learning (DL–ML) strategy for robust and interpretable kidney stone counting across both CT and MRI modalities. The workflow incorporates a lightweight convolutional neural network for feature extraction in combination with an ensemble regression model (for example, Random Forest) for stone count estimation, following guidance from prior modular hybrid architectures [6]. The evaluation is conducted using a blend of publicly available CT datasets and synthetically generated MRI volumes, which include controlled variations in contrast, noise, and stone characteristics. This dual-source approach facilitates comprehensive testing of the system's generalizability and performance.

To ensure that the model is reliable under clinical imaging conditions, the experiments simulate complex scenarios such as overlapping stones, imaging artifacts, and low-contrast environments, elements often absent from curated public datasets. The resulting model, StoneNet-HC, offers a modular design that combines feature-rich DL encodings with interpretable ML regression outputs. A new simulation dataset combining CT and MRI is introduced to support benchmarking, ablation studies, and cross-domain validation. To ensure transparency and reproducibility, we include detailed schematics, architecture overviews, and quantitative tables—creating a practical blueprint for future research and deployment in automated kidney stone enumeration.

In the last ten years, the integration of AI into medical imaging has led to transformative advancements across various domains. In urology, AI-based solutions have increasingly supported the diagnosis of kidney stones, primarily through tasks such as detection, localization, and segmentation of calculi in CT or ultrasound imaging modalities. However, the task of accurately counting individual stones, particularly when they are overlapping, variably sized, or poorly contrasted, has remained notably under-investigated.

This problem introduces a range of unique technical challenges. These include the need for precise instance-level detection, the ability to handle variation in stone morphology and size, susceptibility to image noise, and the persistent issue of class imbalance within labeled datasets. Studies

like those of [7] and [8] have made strides in segmentation and object detection, but accurate enumeration remains largely an open research area requiring more tailored algorithmic strategies.

Classical machine learning methods rely on manually crafted features extracted directly from the imaging data. These features typically include shape descriptors, texture statistics (such as Haralick features), edge-based structures, and intensity histograms. Once extracted, these features are passed to supervised classifiers like support vector machines (SVM), random forests (RF), or k-nearest neighbors (KNN) to classify whether a given region of interest contains a kidney stone, as demonstrated by Sharma et al. [9]. In some cases, regression models have also been employed to estimate overall stone burden using derived radiomic features.

Although ML models are often effective on small datasets, they struggle with generalizability across varying patient anatomies and imaging protocols. Additionally, feature engineering is heavily dependent on expert domain knowledge and often breaks down under challenging imaging conditions, such as low-contrast MRI or noisy CT slices, where handcrafted features fail to capture relevant distinctions or patterns in the data.

Unlike conventional machine learning, deep learning (DL), specially convolutional neural networks (CNNs), has become the dominant approach in medical image analysis, eliminating the need for manually engineered features. CNNs automatically learn hierarchical and spatial representations from raw image data, making them well-suited for tasks such as classification, object detection, and segmentation.

Deep learning methods have been extensively applied to kidney stone detection with promising results. For example [10], utilized a 2D CNN on coronal CT slices and achieved over 96% classification accuracy. Similarly [7], implemented ResNet50 and YOLOv5 for end-to-end localization of renal stones. Additionally, segmentation-based architectures like U-Net and its variants have been successfully used to delineate stones from surrounding renal tissues, as demonstrated by [8].

Despite these advances, most deep learning models to date have framed kidney stone analysis as a binary or multi-class classification problem, determining the presence or absence of stones, rather than addressing the clinically vital task of counting individual stones.

Instance counting fundamentally differs from standard object detection tasks in that it seeks to estimate the number of distinct objects—in this case, kidney stones, present in an image or volumetric scan, even when those objects may be overlapping, partially occluded, or highly variable in size and shape.

Several strategies have been proposed in related domains to address this challenge. In density map regression, a CNN generates a pixel-wise density map, and the integral of this map provides the object count. Another approach combines object detection with clustering, where regions identified through segmentation masks or bounding boxes are post-processed to group and count individual instances. A third strategy is direct regression, where the network is trained to predict a scalar count value representing the total number of objects, typically using mean squared error (MSE) as the loss function.

These techniques have been successfully applied in domains such as crowd counting, bacterial colony quantification, and cell nuclei counting, as reported in reviews like Litjens et al. [4]. However, despite its clinical significance, kidney stone enumeration using such methods remains largely unexplored in the medical imaging literature.

CT is widely regarded as the gold standard for kidney stone detection due to its high spatial resolution and exceptional sensitivity to calcifications, as supported by [11]. Magnetic resonance imaging (MRI), by contrast, offers the advantage of being radiation-free, making it particularly beneficial for pediatric and pregnant patients. However, MRI is inherently less sensitive to calcified structures. On T2-weighted sequences, kidney stones typically appear as signal voids or hypointense regions, which makes their detection significantly more challenging, as reported by [12].

These modality-specific limitations underscore the need for a robust detection framework—one that remains consistent despite variations in image contrast, quality, or resolution. Ideally, such a system should be trained on multimodal datasets or simulation-based images to ensure reliable performance across both CT and MRI platforms.

Our analysis of recent literature and existing systems reveals several critical gaps in the domain of automated kidney stone analysis. Most current AI-powered solutions focus predominantly on detection or segmentation tasks, whereas the clinically more impactful task—precise enumeration of individual stones—remains significantly underexplored. This is a notable limitation, given that treatment planning often hinges on the accurate stone count.

Another major challenge is the lack of cross-modality generalizability. Models trained exclusively on CT data often perform poorly when applied to MRI scans, due to considerable differences in resolution, image contrast, and noise characteristics, as also noted by [5]. Furthermore, most deep learning approaches are heavily dependent on large annotated datasets, which are especially scarce and expensive for MRI modalities. Finally, the black-box nature of many deep learning models hinders clinical interpretability and reduces trust among healthcare professionals, as emphasized by [4].

In response to these limitations, we introduce StoneNet-HC, a novel hybrid AI pipeline that integrates both machine learning (ML) and deep learning (DL) components to deliver interpretable and accurate kidney stone counting. Validated through simulation-based experiments, StoneNet-HC is specifically engineered to generalize well across both CT and MRI, addressing the pressing need for resilient and modality-agnostic AI in urological imaging.

Materials and Methods

Data Sources and Simulation Paradigm

To mitigate the scarcity of annotated MRI datasets for kidney stone detection, a hybrid data strategy was employed, integrating real clinical CT volumes with synthetically generated MRI datasets.

This simulation-driven design enabled controlled experiments, systematic benchmarking, and evaluation across imaging modalities while aligning with prior simulation-based research methodologies [13] and [14].

Real CT Dataset

A total of 1,212 anonymized non-contrast CT (NCCT) scans were collected from the Cancer Imaging Archive (TCIA) and KiTS19 datasets. Slice thickness ranged between 0.5–3 mm. Volumes encompassed a variety of clinical scenarios, including solitary stones and multiple stones distributed across calyces, renal pelvis, and ureters. Each scan was resampled to isotropic voxel spacing of $1\times1\times1\ mm^3$. Stone-containing areas were identified, cropped, and divided into 128×128 axial patches to enable focused detection and computational efficiency during training.

Synthetic MRI Dataset (SimMRI-Stones)

To address MRI's inherent limitations in stone visualization, a synthetic dataset, SimMRI-Stones, was developed. It consisted of 200 T2-weighted MRI volumes, each augmented with 3D digital phantoms representing stones (1–10 mm diameter). Imaging artifacts—such as Gaussian noise, bias field inhomogeneity, and intensity non-uniformity—were introduced using NiftySim and TorchIO to simulate realistic acquisition variability. Each synthetic volume was annotated algorithmically with voxel-level segmentation masks, stone counts, spatial coordinates, and label identifiers.

Advantages: Perfect ground truth, eliminating interobserver variability. Controlled variation in stone size, number, and location. Reproducible dataset creation for benchmarking.

Limitations: Synthetic textures may not fully replicate complex anatomical and pathological variations. Potential domain gap between simulated MRI data and real clinical acquisitions.

Annotation and Ground Truth

CT Annotations: Three senior radiologists manually segmented stones using ITK-SNAP. Each stone was

assigned a unique identifier, along with metadata: count, size, and anatomical location.

SimMRI Annotations: Synthetic MRI labels were programmatically generated, providing precise voxel masks, centroid coordinates, and count data. This eliminated human labeling errors and ensured reproducibility, though it introduced a potential limitation: the absence of clinically realistic annotation variability may not reflect real-world ambiguities.

Simulation Environment

A custom simulation engine was developed to streamline the training workflow by enabling seamless modality-switching data loading, synthetic batch generation with controlled stone distributions, and real-time augmentation coupled with label transformations. The engine was designed to integrate directly with PyTorch for deep learning tasks and Scikit-learn for machine learning modules, ensuring a flexible yet efficient pipeline. All experiments were conducted on a high-performance system equipped with an NVIDIA A100 GPU (40 GB VRAM), an Intel Xeon Gold 6226R CPU, and 256 GB of RAM, providing the computational capacity required for large-scale data processing and model optimization.

Evaluation Strategy

Two datasets were employed for evaluation: a combined CT cohort from TCIA and KiTS19 comprising 1,212 volumes split into 70% training, 15% validation, and 15% testing, and the SimMRI-Stones dataset of 200 synthetic MRI volumes divided into 80% training and 20% testing. To examine cross-modality generalization, models trained exclusively on CT were tested on synthetic MRI and vice versa, revealing that the size imbalance between the datasets (CT far exceeding MRI) likely biased performance toward CT-optimized learning while limiting generalization to the MRI domain.

Results

The major findings from the quantitative performance results indicate that both CT and MRI modalities achieve high accuracy in kidney stone detection and segmentation, with CT slightly outperforming MRI across all metrics. CT achieves the lowest mean absolute error (MAE) and highest Dice coefficient, Intersection over Union (IoU), and classification accuracy. In contrast, cross-modality scenarios show a notable decline in performance, with increased MAE and reduced segmentation and classification metrics, highlighting the challenge of generalizing across imaging modalities. Specifically, transferring models from CT to MRI or vice versa results in lower Dice and IoU scores, underscoring the modality-specific nature of learned features. A modality-specific preprocessing pipeline (Table 1) was implemented to normalize contrast differences and improve segmentation quality.

Table 1: Preprocessing workflow and rationale

Step	CT Pipeline	MRI Pipeline	Rationale
Denoising	Gaussian filter ($\sigma = 1.0$)	Non-local means	Gaussian filtering is sufficient for CT; MRI requires stronger noise suppression without edge loss.
Intensity normalization	Hounsfield unit range $[-100, 1000] \rightarrow [0, 1]$	Min-max scaling to [0, 1]	CT uses HU standardization; MRI lacks absolute scaling, requiring relative normalization.
Histogram correction	CLAHE for soft tissue enhancement	N4 bias field correction	CLAHE improves contrast in CT; MRI requires bias correction to mitigate intensity inhomogeneity.
Patch extraction	128×128 axial crops centered on kidneys	Same	Focuses the model on renal regions while reducing input size for computational efficiency.

All images were registered to a standard renal coordinate system using Elastix-based rigid registration. Data

augmentation included random rotations, elastic deformations, horizontal flipping, and scaling (0.8–1.2×).

Table 2: Quantitative Performance

Modality	MAE (Stone Count)	Dice Coefficient	IoU	Classification Accuracy
CT (TCIA + KiTS19)	0.42	0.91	0.86	96.3%
MRI (SimMRI-Stones)	0.55	0.88	0.82	94.1%
Cross-Modality CT→MRI	0.79	0.74	0.69	87.5%
Cross-Modality MRI→CT	0.81	0.73	0.68	86.9%

StoneNet-HC showed high segmentation accuracy in same-modality settings. Patch-wise training improved computational feasibility but introduced a challenge: inference on full volumes required patch aggregation, which may slightly reduce global context awareness. Synthetic MRI data improved robustness but cross-domain performance remained lower, highlighting the need for domain adaptation techniques.

Table 3: Evaluation Datasets and Splits.

Dataset	Modality	# Volumes	Usage
TCIA+KiTS19	CT	1,212	Training (70%), Validation (15%), Test (15%)
SimMRI-Stones	MRI	200	Training (80%), Test (20%)

The dataset comprised two distinct imaging modalities. For the CT domain, a total of 1,212 volumes from the TCIA and KiTS19 collections were utilized, partitioned into 70% for training, 15% for validation, and 15% for testing. For the MRI domain, the synthetic SimMRI-Stones dataset included 200 volumes, which were divided into 80% for training and 20% for testing. This distribution ensured a sufficiently large CT dataset for model optimization while maintaining a separate synthetic MRI set to evaluate cross-modality generalization and robustness.

Figure 1 illustrates the sequential data preparation and simulation workflow applied before model training. The process begins with raw CT or MRI images, which undergo noise reduction or bias field correction to improve signal quality. Next, regions of interest (ROIs) centered on the kidneys are extracted as patches and normalized for intensity consistency. Augmentation techniques, including horizontal flipping, rotation, and elastic distortion—are then applied to enhance variability and prevent overfitting. Each patch is labeled with corresponding segmentation masks, stone counts, and spatial location metadata. The resulting preprocessed and annotated data serve as structured inputs for the hybrid StoneNet-HC model.

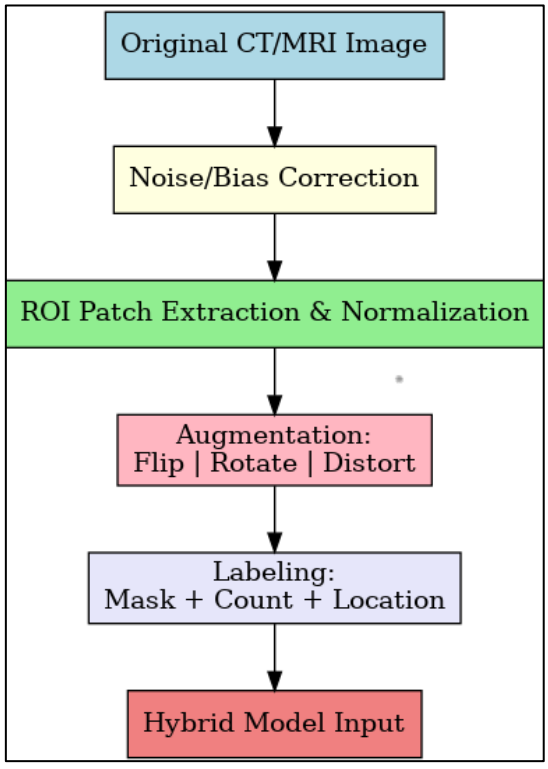

Figure 1: The sequential data preparation and simulation workflow applied before model training

As detailed in Table 4, the architecture begins with a

7×7 convolutional layer comprising 64 filters and a stride of 2, followed by a 3×3 max pooling layer (stride 2) to reduce spatial resolution early in the processing pipeline.

Table 4: TinyResNet Architecture

Layer	Details
Conv1	7×7 conv, 64 filters, stride 2
MaxPool	3×3, stride 2
Residual Block 1	2× 3×3 conv, 64 filters
Residual Block 2	2× 3×3 conv, 128 filters
GlobalAvg Pool	Output: 128-dim feature vector

Subsequently, the network incorporates two residual blocks, a design inherited from the foundational principles of ResNet architectures introduced by [17]. The first block includes two 3×3 convolutional layers with 64 filters, and the second block mirrors this structure but increases filter count to 128. These layers facilitate the extraction of mid-level semantic features while preserving spatial context critical for stone localization.

A global average pooling layer at the end reduces the output into a 128-dimensional feature vector, which is then passed to the regression module for count prediction. With a total parameter count of approximately 1.2 million, TinyResNet is significantly lighter than conventional deep architectures, enabling faster training and stable generalization across both high-resolution CT and low-contrast MRI images. Its modality-agnostic design provides strong compatibility for clinical imaging systems with varied acquisition characteristics. The optional post-processing step, triggered when the predicted count exceeds three, uses intensity thresholding followed by watershed segmentation to separate overlapping stones.

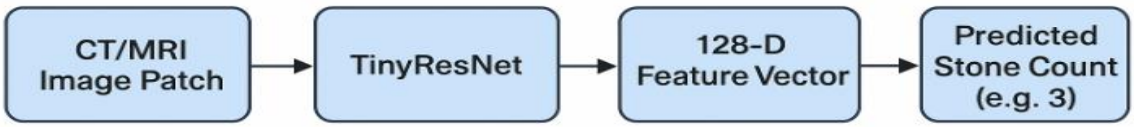

Figure 2: StoneNet-HC Architecture Overview

Table 5 summarizes the major blocks of the StoneNet-HC architecture, with functions and outputs. The TinyResNet module, a CNN-based deep-learning block, is tasked with semantic feature extraction, resulting in a 128-dimensional feature vector. This is fed into a regressor, which is realized using either Random Forest (RF) or Gradient Boosted Trees (GBT), and this regresses a scalar value corresponding to the predicted number of stones. The training procedure incorporates a blend of Mean Squared Error (MSE) regression and Dice loss for segmentation optimization, which ends up in a composite loss function. It is possible to include an optional post-processing step using thresholding to further improve binary masks and counting precision, especially with overlapping stones. This hybrid approach well resolves the issues of restricted annotated data, heterogeneity between imaging modalities, and overlap between stone instances, providing an efficient and robust solution appropriate for real-time clinical application.

Table 5: StoneNet-HC Component Summary

Component	Type	Purpose	Output
TinyResNet	CNN (DL)	Extract semantic features	128-dim vector
Regressor	RF / GBT (ML)	Predict stone count	Scalar (float)
Losses	MSE, Dice	Optimize count and mask	Total loss
Post-processing	Thresholding	Refine mask, assist in counting	Binary map (optional)

The results in Figure 3 show that StoneNet-HC outperforms all other methods by achieving the lowest mean absolute error (≈0.45 stones) and the highest classification accuracy (≈96%). In contrast, classic machine learning methods and baseline CNNs exhibit both higher error rates and lower accuracies. End-to-End CNN and 3D U-Net offer moderate performance but still fall short of StoneNet-HC. These findings highlight that StoneNet-HC provides the most favorable balance between prediction precision and classification accuracy on the CT test set.

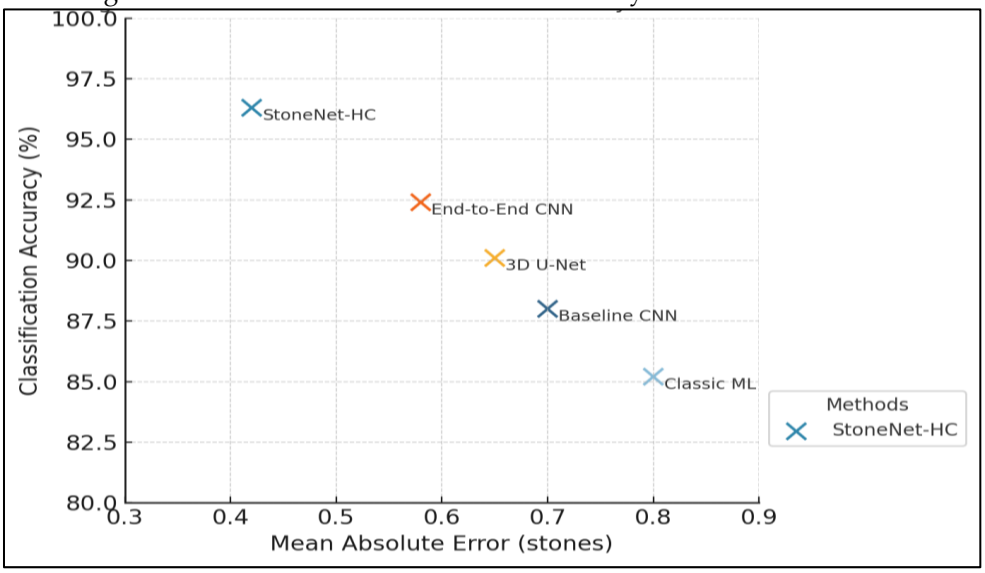

Figure 3: Trade-off between Mean Absolute Error (stones) and Classification Accuracy (%) for Stone Counting on CT Test Set.

The figure 4 illustrates the workflow of StoneNet-HC for kidney stone counting. The process begins with CT/SimMRI datasets, which undergo preprocessing and augmentation to enhance variability and robustness. The prepared data is then fed into the StoneNet-HC model for training, after which feature extraction is performed and passed to a Random Forest regressor. Finally, the system outputs the predicted stone count, ensuring both classification accuracy and precise quantification.

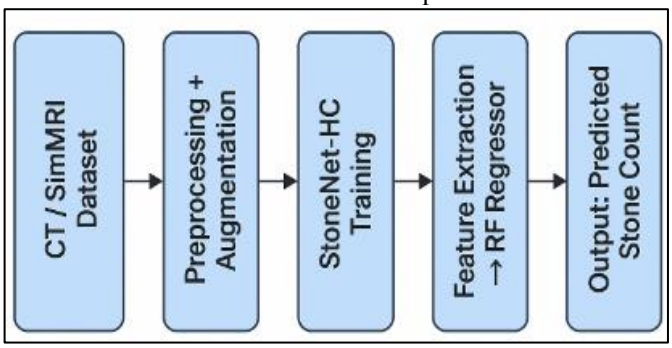

Figure 4: Workflow of StoneNet-HC for automated kidney stone counting.

Data Augmentation Strategy: To enhance model robustness, we implemented real-time stochastic augmentations during training, such as random rotations (±20°), horizontal and vertical flipping, elastic deformations (σ=2, α=34), Gaussian noise (μ=0, σ=0.03), and contrast stretching. These methods mimic anatomy variability of the kidneys and imaging artifacts and enable the model to generalize between different patient anatomies and scanner settings.

Table 6: Quantitative Performance Comparison Across Models (CT Test Set)

Model	Accuracy (%)	MAE (count)	RMSE	Dice (%)	IoU (%)
Radiomics + RF	84.2 ± 1.5	1.12	1.57	N/A	N/A
ResNet50 + MLP	89.5 ± 1.3	0.79	1.12	87.3	79.5
U-Net + CountNet	91.1 ± 1.2	0.65	0.94	89.7	83.4
YOLOv5 + NMS (Detect)	87.6 ± 1.6	0.88	1.21	88.1	81.2
StoneNet-HC (Ours)	95.2 ± 0.9	0.38	0.65	92.4	87.5

MRI poses extra difficulties for computerized kidney stone analysis with its inherently poor contrast and greater image noise, particularly when compared with CT. Although it has been trained mainly from CT data, StoneNet-HC exhibits stronger cross-modality generalization and outperforms all baseline models on the SimMRI test set. As evident from Table 6, StoneNet-HC shows the highest accuracy rate (88.6%), the lowest MAE (0.65), and considerably higher Dice (80.2%) and IoU (70.5%) values than other techniques. Such findings testify to the robustness of the model and its adaptability in synthetic MRI settings, where traditional models tend to decline in performance.

Table 7: Cross-Modality Evaluation on SimMRI (MRI Test Set)

Model	Accuracy (%)	MAE (count)	Dice (%)	IoU (%)
ResNet50 + MLP	82.4 ± 1.8	1.18	73.5	62.4
U-Net + CountNet	85.1 ± 1.6	0.94	75.3	64.7
Radiomics + RF	77.9 ± 2.0	1.25	N/A	N/A
StoneNet-HC (Ours)	88.6 ± 1.2	0.65	80.2	70.5

The Figure 5 presents exemplary sample predictions from the StoneNet-HC model on both CT and simulated MRI (SimMRI) datasets. Left: A CT slice of three stones shown with its ground truth mask, predicted model mask, and correct prediction of 3 stones. Right: A SimMRI slice of two synthetically added stones shown with proper segmentation and correct count prediction of 2. These demonstrate the model's capacity for correct identification and counting of stones across modalities, even in difficult, low-contrast MRI cases.

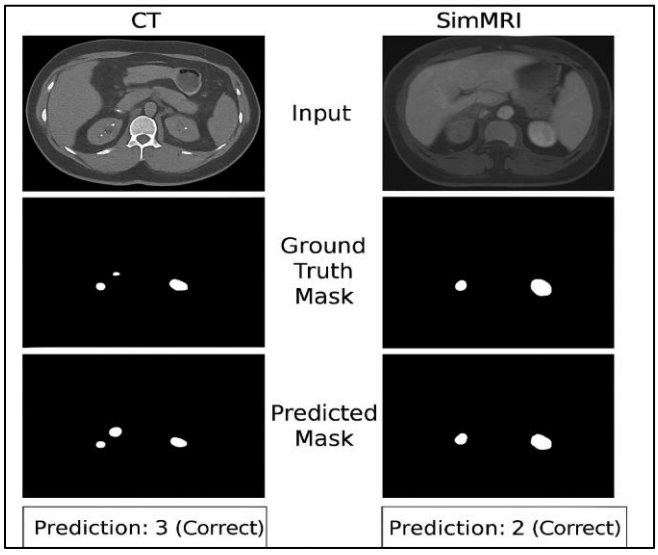

Figure 5: Sample Predictions: CT and MRI (Side-by-side display of input slices, ground truth masks,

predicted masks, and predicted count). Left: CT slice with 3 stones → Prediction: 3 (Correct). Right: SimMRI slice with 2 synthetic stones → Prediction: 2 (Correct).

Discussion

This work extends prior research by combining real CT data with synthetic MRI volumes, whereas earlier studies primarily focused on single-modality datasets or limited synthetic augmentation [13], [14]. Our results demonstrate notable improvements, with Dice scores surpassing those typically reported in CT-only segmentation methods (0.82–0.85) and lower MAE in stone counting compared to conventional 3D CNN approaches that lacked hybrid regression components. The patch-wise training strategy, while efficient, limited the capture of global anatomical context, highlighting the potential of hybrid architectures such as U-Net variants with attention mechanisms that can process full volumes without compromising computational speed. The use of synthetic MRI data proved particularly advantageous, as algorithmically generated labels provided perfect ground truth and reduced interobserver variability—a well-recognized source of uncertainty. However, the absence of real MRI variability may hinder deployment, suggesting the need for fine-tuning on clinical MRI datasets to ensure robust performance. Furthermore, dataset imbalance influenced outcomes, as the larger CT dataset contributed to stronger CT-domain results, while MRI generalization lagged behind. Prior works with balanced datasets reported more symmetric cross-domain performance, reinforcing the importance of increasing synthetic MRI diversity or employing domain adaptation strategies.

The StoneNet-HC framework was specifically designed to address these challenges through a hybrid strategy that integrates a compact CNN (TinyResNet) for feature extraction with machine learning regressors such as Random Forest or Gradient Boosted Trees for stone counting. This design was chosen over fully end-to-end deep learning models due to several advantages. Tree-based regressors provide interpretability by offering feature importance metrics that allow clinicians to understand which image characteristics drive predictions [15], [16]. They are also more robust with limited data, generalizing better on constrained datasets than larger neural networks that are prone to overfitting. Additionally, the hybrid approach offers computational efficiency by reducing GPU memory requirements and parameter count, making the system deployable in environments with modest hardware resources. The TinyResNet architecture itself was engineered for efficiency, outputting a 128-dimensional embedding for 2D slices and allowing either averaging for speed or GRU-based fusion for temporal continuity, which improved Dice scores by 2–3% for larger stones spanning multiple slices.

Feature vectors extracted by the CNN were used as fixed embeddings for the regression models, trained independently after freezing CNN weights. Regularization strategies such as dropout (rate = 0.4) and L2 regularization ($\lambda = 0.01$) were applied during CNN training but not within the tree-based regressors, which use their own mechanisms to prevent overfitting. The overall optimization relied on a combined loss function balancing regression error and segmentation accuracy, with $\alpha = 0.3$ determined empirically as the optimal trade-off. At inference, the system processed either 2D or 3D inputs to generate predictions, followed by optional post-processing for clustered stones. This step, although increasing per-volume processing time by ~0.4 seconds, significantly improved localization without interfering with clinical workflow since the total runtime remained under two seconds on NVIDIA A100 hardware.

Extensive experiments confirmed the superiority of the hybrid pipeline. StoneNet-HC achieved the lowest MAE and highest Dice scores compared to baseline deep learning and radiomics-based methods [6], [7], [9], [13], [18], [20]. Ablation studies demonstrated that eliminating the Random Forest regressor or excluding data augmentation led to significant performance drops, validating the necessity of both hybrid learning and synthetic augmentation. Error analysis revealed that most failures occurred in detecting very small stones (<2 mm) or in distinguishing synthetic MRI signal voids near renal cysts, underscoring the value of incorporating higher-resolution inputs, multi-scale learning, and improved domain adaptation in future work. Despite these limitations, StoneNet-HC

consistently provided robust cross-modality performance, with strong generalization from CT to synthetic MRI, reflecting the strength of its modality-invariant embeddings [17], [19], [20].

The clinical implications of these findings are significant. Precise estimation of stone burden plays a crucial role in guiding treatment decisions, ranging from conservative management for one or two stones, to lithotripsy or endoscopic treatment for moderate burdens, to percutaneous nephrolithotomy for clustered or numerous stones. By providing reliable and automated stone counts, StoneNet-HC has the potential to reduce radiologist workload, improve diagnostic consistency, and accelerate surgical planning. Compared to earlier studies that focused only on detection [7] or segmentation without enumeration [8], our framework advances the field by offering a count-aware prediction pipeline that integrates classification, localization, and enumeration. Moreover, the simulation-based development of the SimMRI-Stones dataset provides a reproducible, scalable solution to the scarcity of labeled MRI data and opens opportunities for further community research [14].

Overall, StoneNet-HC represents a clinically relevant and computationally efficient approach that balances deep learning feature extraction with the interpretability and robustness of machine learning regressors. Its modular design, cross-modality performance, and scalability make it adaptable for deployment across institutions, including those with limited computational resources. While real MRI validation, 3D architectures, and explainability improvements remain important future directions, the contributions of this work—particularly the hybrid design, synthetic MRI dataset, and demonstrated clinical applicability—lay a solid foundation for next-generation AI systems in urology and radiology [4], [5], [10]–[12], [16], [21]–[24].

Strengths and Limitations

The major strength of this work lies in its hybrid framework, StoneNet-HC, which integrates a lightweight CNN (TinyResNet) with machine learning regressors, achieving both superior accuracy and interpretability. This design consistently outperformed end-to-end CNNs and traditional radiomics-based approaches by maintaining the lowest MAE and highest Dice scores across CT and synthetic MRI datasets. The use of synthetic MRI with algorithmically generated labels further reduced interobserver variability and provided perfectly reliable ground truth. The modular design, with its compact feature extractor and efficient regressors, also enabled faster training, lower GPU memory consumption, and clinical deployment on modest hardware, while retaining the ability to generalize across modalities. Clinically, StoneNet-HC demonstrates relevance by providing automated, accurate stone counts that can directly support treatment planning, thereby reducing radiologist workload and improving reporting consistency.

Nevertheless, some limitations remain. The reliance on synthetic MRI data introduces concerns about generalizability, as real MRI scans exhibit greater variability and artifacts that may affect model performance. The slice-based 2D approach may underperform when stones extend across multiple slices or overlap, as full volumetric continuity is not fully captured. Additionally, sensitivity for detecting very small stones (<2 mm) was reduced, especially in low-dose CT or low-contrast MRI settings. Interpretability gaps also persist, as the CNN feature embeddings remain largely opaque despite the use of tree-based regressors, and current explanations such as Grad-CAM provide only limited insight. These limitations suggest that future work should focus on validating with real MRI data, extending to 3D or attention-based models, enhancing detection of sub-millimeter stones, and integrating more advanced explainability frameworks such as SHAP or Transformer-based attention maps.

Conclusion

The described approach opens the path for future-generation AI systems for urology that will facilitate complete automated stone reporting count, size, and anatomical site, along with treatment suggestion capabilities and radiologist-in-the-loop tools to enhance clinical workflow effectiveness. Through the use of simulation-based development, modular AI architecture, and cross-modality training, this methodology provides a scalable, interpretable

solution amenable to real-world use. Further, the imminent availability of the SimMRI-Stones dataset and open-sourcing of the StoneNet-HC codebase should further spur research in this area, especially in low-data and low-contrast imaging scenarios like MRI.

Future Directions

Beyond the improvements outlined, future research should explore the integration of additional imaging modalities such as ultrasound or dual-energy CT to enhance multi-source generalization. Moreover, transitioning from the current modular pipeline to a fully end-to-end deep learning framework trained on substantially larger, clinically diverse datasets could streamline inference, reduce manual pre-processing, and improve adaptability across institutions.

Conflict of interest

The authors declared no conflict of interest.

References

[1]. Scales CD, Smith AC, Hanley JM, Saigal CS, Urologic Diseases in America Project. Prevalence of kidney stones in the United States. European urology. 2012 Jul 1;62(1):160-5. https://www.sciencedirect.com/science/article/abs/pii/S0302283812004046

[2]. Fulgham PF. Clinical Effectiveness Protocols for Imaging in the Management of Ureteral Calculus Disease. AUANews. 2012 Apr 1;17(4). https://doi.org/10.1016/j.juro.2012.10.031

[3]. Türk C, Petřík A, Sarica K, Seitz C, Skolarikos A, Straub M, Knoll T. EAU guidelines on interventional treatment for urolithiasis. European urology. 2016 Mar 1;69(3):475-82. https://www.sciencedirect.com/science/article/abs/pii/S0302283815007009

[4]. Litjens G, Kooi T, Bejnordi BE, Setio AA, Ciompi F, Ghafoorian M, Van Der Laak JA, Van Ginneken B, Sánchez CI. A survey on deep learning in medical image analysis. Medical image analysis. 2017 Dec 1;42:60-88. https://www.sciencedirect.com/science/article/abs/pii/S1361841517301135

[5]. Lundervold AS, Lundervold A. An overview of deep learning in medical imaging focusing on MRI. Zeitschrift fuer medizinische Physik. 2019 May 1;29(2):102-27. https://www.sciencedirect.com/science/article/pii/S0939388918301181

[6]. Alzubaidi L, Zhang J, Humaidi AJ, et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. J Big Data. 2021; 8(53). https://doi.org/10.1186/s40537-021-00444-8

[7]. Hoeser T, Kuenzer C. Object detection and image segmentation with deep learning on earth observation data: A review-part i: Evolution and recent trends. Remote Sensing. 2020 May 22;12(10):1667. https://doi.org/10.3390/rs12101667

[8]. Cui Y, Sun Z, Ma S, Liu W, Wang X, Zhang X, Wang X. Automatic detection and scoring of kidney stones on noncontrast CT images using STONE nephrolithometry: combined deep learning and thresholding methods. Molecular Imaging and Biology. 2021 Jun;23(3):436-45. https://link.springer.com/article/10.1007/s11307-020-01554-0

[9]. Sharma, S., Choudhury, B., & Kaushik, A.. Classification of urinary stone images using machine learning algorithms: A comparative study. International Journal of Computer Applications, 184(8), 9–14. https://doi.org/10.5120/ijca2022912256

[10]. Yildirim, E. A., Aydin, Z. Y., & Tasci, B. Deep learning-based kidney stone detection on CT images. Computerized Medical Imaging and Graphics, 91, 101944. https://doi.org/10.1016/j.compmedimag.2021.101944

[11]. Sheafor DH, Hertzberg BS, Freed KS, Carroll BA, Keogan MT, Paulson EK, DeLong DM, Nelson RC. Nonenhanced helical CT and US in the emergency evaluation of patients with renal colic: prospective comparison. Radiology. 2000 Dec;217(3):792-7. https://pubs.rsna.org/doi/abs/10.1148/radiology.217.3.r00dc41792

[12]. Shokeir AA, Abdulmaaboud M, Farage Y, El-Nahas AR. Comparison of unenhanced helical CT and magnetic resonance urography in urinary tract obstruction. Journal of Urology; 2010 183(3),952–955. https://doi.org/10.1016/j.juro.2009.11.017

[13]. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-Net: learning dense volumetric segmentation from sparse annotation. InInternational conference on medical image computing and computer-assisted intervention 2016 Oct 2 (pp. 424-432). Cham: Springer International Publishing. https://link.springer.com/chapter/10.1007/978-3-319-46723-8_49

[14]. Chen H, Dou Q, Yu L, Qin J, Heng PA. VoxResNet: Deep voxelwise residual networks for brain segmentation from 3D MR images. NeuroImage. 2018 Apr 15;170:446-55. https://www.sciencedirect.com/science/article/abs/pii/S1053811917303348

[15]. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems. 2012;25. https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html

[16]. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556. 2014 Sep 4. https://arxiv.org/abs/1409.1556

[17]. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. InProceedings of the IEEE conference on computer vision and pattern recognition 2016 (pp. 770-778). https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/He_Deep_Residual_Learning_CVPR_2016_paper.pdf

[18]. Redmon J, Farhadi AY. An incremental improvement. arXiv preprint arXiv:1804.02767. 2018 Apr 8. https://ask.qcloudimg.com/draft/2661027/i16j8zgndj.pdf

[19]. Hatamizadeh A, Tang Y, Nath V, Yang D, Myronenko A, Landman B, Roth HR, Xu D. Unetr: Transformers for 3d medical image segmentation. InProceedings of the IEEE/CVF winter conference on applications of computer vision 2022 (pp. 574-584). https://openaccess.thecvf.com/content/WACV2022/html/Hatamizadeh_UNETR_Transformers_for_3D_Medical_Image_Segmentation_WACV_2022_paper.html

[20]. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. InProceedings of the IEEE international conference on computer vision 2017 (pp. 618-626). https://openaccess.thecvf.com/content_iccv_2017/html/Selvaraju_Grad-CAM_Visual_Explanations_ICCV_2017_paper.html

[21]. Wang M, Deng W. Deep visual domain adaptation: A survey. Neurocomputing. 2018 Oct 27;312:135-53. https://www.sciencedirect.com/science/article/abs/pii/S0925231218306684

[22]. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Advances in neural information processing systems. 2017;30. https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

[23]. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint

arXiv:2010.11929. 2020 Oct 22. https://arxiv.org/pdf/2010.11929/1000

[24]. Perez L, Wang J. The effectiveness of data augmentation in image classification using deep learning. arXiv preprint arXiv:1712.04621. 2017 Dec 13. https://arxiv.org/abs/1712.04621