Minimax Theory and its Applications

Volume 11 (2026), No. 5s, 180-197

Stone-Preserving Modality-Adaptive Preprocessing and Comparative Deep CNN Frameworkfor Early Detection of Kidney Stones in CT and Ultrasound Images

Dhanesh Kumar[1], Donavalli Haritha[2]

[1]Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India,
Email: krdhanesh06@gmail.com
[2]Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India
Email: haritha_donavalli@kluniversity.in

Abstract
Kidney stone disease affects a large and growing share of the population, and its early detection depends on recognizing small calculi in computed tomography (CT) and ultrasound images that are degraded by quantum noise and speckle. Deep convolutional neural networks (CNNs) report very high accuracy on public kidney CT datasets, but most studies treat preprocessing as a fixed step and do not examine whether denoising removes the very foci that indicate early disease. This study proposes a robust detection framework built around a Stone-Preserving Adaptive Enhancement (SPAE) pipeline: modality-specific denoising (non-local means for CT, speckle-reducing anisotropic diffusion for ultrasound) is combined with a multiscale Laplacian-of-Gaussian and top-hat saliency map that restores small bright foci suppressed by denoising, followed by kidney region-of-interest (ROI) extraction and contrast-limited adaptive histogram equalization, after which ResNet, DenseNet, EfficientNet and ConvNeXt backbones are compared under a common protocol. The pipeline was validated on synthetic CT and ultrasound kidney phantoms with known stones. Speckle reduction alone lowered small-stone (radius ≤ 2 pixels) recall at about one false positive per image from 80.3% without filtering to 63.1%, whereas SPAE restored it to 76.2%; in CT, SPAE raised small-stone recall from 82.0% with non-local means to 89.1% (120 test phantoms each). ROI extraction achieved mean Dice coefficients of 0.94 for CT and 0.88–0.90 for ultrasound and retained every stone within the cropped region. Noise-reduction quality remained close to the base denoisers (CT PSNR 33.4 versus 34.2 dB at noise SD 0.07). The CNN comparison on public CT and ultrasound datasets is specified with a leakage-controlled protocol, and the experiments are in progress.

Keywords— Kidney stone detection; Computed tomography; Ultrasound; Speckle reduction; Image enhancement; Region of interest; Convolutional neural networks; ConvNeXt; EfficientNet

Introduction

Kidney stone disease (nephrolithiasis) is a common and recurrent condition in which crystalline concretions form within the urinary tract, causing pain, obstruction, infection and, in the long term, loss of renal function [1]. Population surveys have reported a lifetime prevalence of about one in eleven adults in the United States, with a rising trend over recent decades [2]. Early detection of small stones allows conservative management and prevention before obstruction or complications occur, which makes reliable imaging-based detection clinically valuable.

Non-contrast CT is the reference standard for stone detection because of its high sensitivity, whereas ultrasound is radiation-free, inexpensive and widely available, and is preferred for initial assessment in many settings, in children and in pregnancy [3]. A large randomized trial showed that initial ultrasonography was associated with lower cumulative radiation exposure than CT without significant differences in serious adverse events [4]. Both modalities, however, face image-quality limitations that are most consequential for small stones: low-dose CT is affected by quantum noise, and ultrasound images are dominated by multiplicative speckle, both of which can mask calculi only a few pixels wide.

Deep learning has been widely applied to automatic kidney stone detection. A CNN trained on coronal CT images achieved 96.82% accuracy in distinguishing patients with and without stones [5], and a public CT dataset of 12,446 images of normal kidneys, cysts, stones and tumors has enabled many CNN and transformer studies, with the Swin transformer reaching 99.30% accuracy in its original evaluation [6]. Ensemble CNNs

have also been used to classify kidney abnormalities in ultrasound images [7]. Nevertheless, preprocessing is usually treated as a generic step of resizing, normalization and occasionally filtering, and its effect on small, early-stage stones is rarely measured.

Research Motivation

Denoising is a double-edged step for stone detection. Speckle-reducing anisotropic diffusion [8] and non-local means [9] suppress noise effectively but treat small isolated bright structures as noise to be smoothed, and contrast enhancement such as CLAHE [10] can amplify residual noise. A preprocessing pipeline designed for early detection must therefore reduce noise while explicitly preserving small hyperdense (CT) or hyperechoic (ultrasound) foci, and its benefit must be measured where it matters: in the detectability of small stones rather than in global image-quality indices alone.

Research Significance

This work is significant because it addresses a gap between image-processing practice and the clinical goal of early detection. By quantifying how standard denoisers affect small-stone conspicuity in CT and ultrasound and by introducing a stone-preserving alternative, it provides a principled basis for preprocessing choices. The common protocol for comparing CNN architectures on the enhanced, ROI-cropped images further supports evidence-based model selection for clinical decision support.

Problem Statement

Current deep learning approaches for kidney stone detection have four limitations. Preprocessing is not tailored to the noise physics of each modality; denoising is applied without verifying that small stones survive; whole images are classified although the kidney occupies only a fraction of the field of view, so irrelevant structures can drive decisions; and architectures are compared under different datasets, splits and preprocessing, so reported differences are difficult to interpret. In addition, very high accuracies on public datasets may partly reflect near-duplicate slices from the same patient appearing in both training and test sets.

Key Contribution

- A modality-adaptive enhancement and noise-removal pipeline, with non-local means for CT and speckle-reducing anisotropic diffusion for ultrasound.
- A Stone-Preserving Adaptive Enhancement (SPAE) step that uses multiscale Laplacian-of-Gaussian and top-hat saliency to restore small bright foci suppressed by denoising.
- An automatic kidney ROI extraction method based on multi-level Otsu thresholding and morphology, which crops the input to the renal region without losing stones.
- A phantom-based validation measuring PSNR, SSIM, contrast-to-noise ratio, ROI Dice and small-stone recall at a fixed false-positive rate.
- A leakage-controlled protocol for comparing ResNet, DenseNet, EfficientNet, ConvNeXt, VGG16 and MobileNetV3 on CT and ultrasound images.

Rest of the Paper

Section II reviews related work. Section III describes the proposed framework. Section IV reports the validation of the preprocessing pipeline and the experimental design for the CNN comparison. Section V concludes the paper.

Related Works

Yildirim et al. [5] developed an XResNet-50 model for automated kidney stone detection on coronal CT images from 433 subjects, reaching 96.82% accuracy, and used class activation maps to show that the network attended to the stone region. The study demonstrated that small stones could be detected by a CNN but also reported misses for very small calculi.

Islam et al. [6] collected and annotated the CT KIDNEY DATASET of 12,446 whole-abdomen and urogram images in four classes (normal, cyst, stone and tumor) and compared VGG16, ResNet50 and Inception v3 with the EANet, CCT and Swin vision transformers. The Swin transformer achieved the highest accuracy of 99.30%, and VGG16 reached 98.20%. Grad-CAM analysis indicated that VGG16 focused on relevant anatomy more consistently than ResNet50 and Inception v3.

Sudharson and Kokil [7] proposed an ensemble of pretrained deep networks for classifying kidney ultrasound images into normal, cyst, stone and tumor classes, and in later work [11] designed a computer-aided diagnosis

system that remained accurate when images were corrupted by speckle noise. Their results highlight that ultrasound-based detection is sensitive to image quality.

Elbedwehy et al. [12] combined neural networks with optimization techniques for kidney disease classification on the 12,446-image CT dataset and emphasized interpretability. Pimpalkar et al. [13] fine-tuned VGG16, ResNet50, AlexNet and InceptionV3 with hyperparameter optimization, Otsu binarization and watershed segmentation on CT images and reported accuracies up to 99.96% with wide neural network classifiers. Sharma et al. [14] fused a pretrained ResNet101 with a custom CNN and reported 100% test accuracy on the same dataset, and Khan et al. [15] proposed KTNET, a transformer with hierarchical feature blocks for kidney abnormality classification.

For ultrasound, Singla et al. [16] released the Open Kidney Ultrasound Data Set with expert annotations of kidney regions, providing a resource for kidney segmentation and ROI extraction in a modality where public data remain scarce.

Image enhancement for these modalities has a long history. Michailovich and Tannenbaum [17] reviewed despeckling of medical ultrasound images and showed that methods based on the multiplicative speckle model outperform generic filters, while Sagheer and George [18] surveyed medical image denoising algorithms across modalities and noted that most methods are evaluated with global fidelity measures rather than diagnostic tasks. For low-dose CT, Chen et al. [19] proposed RED-CNN, a residual encoder–decoder network that reduces noise while preserving structural detail. For ROI extraction, U-Net [20] has become the dominant segmentation architecture, but it requires pixel-level annotations that are rarely available for kidney stone datasets.

Overall, CNN and transformer models achieve very high accuracy on the public CT dataset, but the studies differ in splits, preprocessing and evaluation, and accuracies close to 100% raise the question of whether images from the same patient appear in both training and test sets. Modality-specific noise handling, its effect on small stones and kidney-focused ROI extraction have received little systematic attention.

Table 1. Comparison Table

Author	**Method**	**Data**	**Strengths**	**Limitations**
Yildirim et al. [5]	XResNet-50	Coronal CT, 433 subjects	Patient-level data; CAM	Misses very small stones
Islam et al. [6]	VGG16, ResNet50, Inception v3, Swin	CT, 12,446 images	Public dataset; Swin 99.30%	Split leakage not discussed
Sudharson & Kokil [7], [11]	Ensemble deep networks	Ultrasound	Robust to speckle noise	Limited public data
Elbedwehy et al. [12]	NN + optimization	CT, 12,446 images	Interpretability	Generic preprocessing
Pimpalkar et al. [13]	Fine-tuned CNNs, Otsu, watershed	CT, 4 classes	Up to 99.96% accuracy	Image-level evaluation
Sharma et al. [14]	ResNet101 + custom CNN	CT, 12,446 images	Feature fusion	100% test accuracy suggests leakage risk
Khan et al. [15]	KTNET transformer	CT	Hierarchical features	Not evaluated on ultrasound
Singla et al. [16]	Open kidney US data set	Ultrasound	Kidney ROI annotations	No stone labels

Author	Method	Data	Strengths	Limitations
Chen et al. [19]	RED-CNN low-dose CT denoising	CT	Structure-preserving denoising	Not evaluated for stone detectability

Table I summarizes the reviewed studies. The table shows that classification accuracy on public CT data is high, while preprocessing for small-stone preservation, ROI extraction and cross-modality comparison under a common protocol remain open problems, which motivates the proposed framework.

Proposed Methodology

The proposed framework processes CT slices and ultrasound frames with a modality-adaptive preprocessing pipeline, extracts the kidney ROI and classifies the enhanced crop with alternative CNN backbones under a common protocol. Fig. 1 gives an overview.

Data Collection

For CT, the public CT KIDNEY DATASET [6] is used, which contains 12,446 images collected from whole-abdomen and urogram examinations and labelled as normal (5,077), cyst (3,709), stone (1,377) or tumor (2,283). For ultrasound, a public kidney ultrasound collection with stone and normal labels is combined with the Open Kidney Ultrasound Data Set [16], whose expert kidney delineations are used to evaluate ROI extraction. Because the CT dataset is distributed as images without patient identifiers, near-duplicate slices are grouped by perceptual hashing before splitting, so that visually near-identical slices cannot appear in different partitions.

Near-duplicate grouping uses a 64-bit difference hash of each image, computed from the signs of horizontal intensity differences on a 9 × 8 down-sampled copy. Two images i and j are linked when the Hamming distance of their hashes is small, as in (1):

$$d_H(i,j) = \sum_{b=1}^{64} \mathbb{1}\left[h_{i,b} \neq h_{j,b}\right] \leq \tau_H \quad (1)$$

with $\tau H = 6$, and connected components of linked images form groups that are assigned to a single fold. This procedure approximates patient-level splitting when patient identifiers are unavailable.

Table II summarizes the data. Real datasets do not provide pixel-level stone annotations or noise-free references, so the preprocessing components are validated on synthetic phantoms whose ground truth is known, and the CNN comparison is carried out on the real images.

Table 2. Datasets Used in the Study

Dataset	Modality	Classes / annotations	Role
CT KIDNEY DATASET [6]	CT (abdomen, urogram)	Normal 5,077; cyst 3,709; stone 1,377; tumor 2,283	Stone vs. normal and 4-class CNN comparison
Public kidney US collection	Ultrasound	Stone / normal	Stone vs. normal CNN comparison
Open Kidney US Data Set [16]	Ultrasound	Expert kidney ROI masks	ROI-extraction evaluation
Synthetic kidney phantoms	CT and ultrasound	Known kidney masks and stones (radius 1–4 px)	Validation of denoising, SPAE and ROI

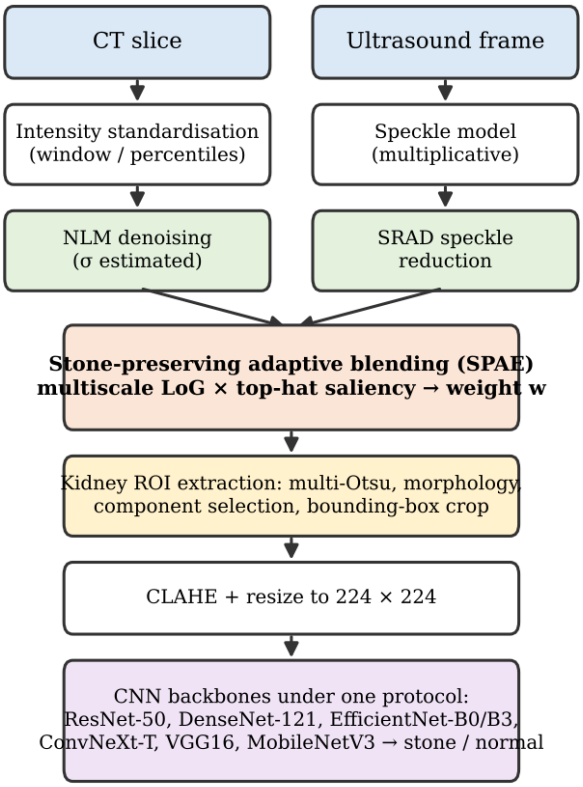

Fig. 1. Overview of the Proposed Stone-Preserving Detection Framework

Fig. 1. shows the two modality-specific branches, which differ in intensity standardization and denoising, and the shared stone-preserving blending, ROI extraction, contrast enhancement and CNN classification stages. The design keeps the classifier identical across modalities so that differences in performance can be attributed to the data and the preprocessing.

Modality-Specific Enhancement and Noise Removal

CT Intensity Standardization and Denoising

CT images are standardized to a soft-tissue window when Hounsfield units are available, and otherwise to the 1st–99th intensity percentiles of the body region. CT quantum noise is approximated as additive and signal-independent at the image level. Its standard deviation $\hat{\sigma}$ is estimated with the fast Laplacian-based estimator of Immerkær [21], and non-local means [9] is applied with filtering parameter $h = 0.8\hat{\sigma}$, as in (2):

$$\hat{u}(x) = \sum_{y\in\Omega} \frac{w(x,y)}{Z(x)} v(y), \quad w(x,y) = e^{-\frac{\|P(x)-P(y)\|_2^2}{h^2}} \qquad (2)$$

where v is the noisy image, P(x) the 5 × 5 patch around pixel x and Z(x) a normalizing constant.

Ultrasound Speckle Reduction

Ultrasound speckle is multiplicative, $v = u \cdot n$, where n follows approximately a Rayleigh distribution in amplitude images. Speckle-reducing anisotropic diffusion (SRAD) [8] evolves the image according to (3):

$$\frac{\partial I}{\partial t} = \operatorname{div}(c(q)\nabla I), \quad c(q) = \frac{1}{1+\frac{q^2-q_0^2}{q_0^2(1+q_0^2)}} \qquad (3)$$

where q is the instantaneous coefficient of variation, which is large at edges and small in homogeneous speckle, and q0 is the speckle scale, which decays over the 30 iterations. Diffusion is therefore inhibited at tissue boundaries but proceeds freely in speckled regions.

Stone-Preserving Adaptive Enhancement (SPAE)

Both denoisers can suppress stones that span only a few pixels, because such stones resemble isolated noise peaks. SPAE detects candidate stone foci with a multiscale blob measure and restores them from a lightly smoothed version of the input. The saliency map combines the scale-normalized negative Laplacian of Gaussian [22] with a white top-hat transform, as in (4):

$$S(x) = \left[\max_s\left(-s^2\nabla^2 G_s * v_p\right)\right]_+ \cdot TH\left(v_p\right) \quad (4)$$

where vp is the average of the denoised image and a Gaussian-smoothed (σ = 1) copy of the input, Gs a Gaussian of scale s ∈ {1, 2, 3, 4} pixels matching stone radii, and TH the white top-hat with a disk of radius 5, which keeps structures smaller than the disk and brighter than their surroundings. S is normalized to [0, 1] by its 99.9th percentile. The blending weight and the enhanced image are then given in (5):

$$w(x) = 1 - e^{-kS(x)},\ \tilde{t}(x) = w(x)(G_{0.5} * v)(x) + \left(1 - w(x)\right)\hat{u}(x) \qquad (5)$$

with k = 4. In homogeneous tissue w is close to 0 and the output equals the denoised image; at candidate stones w approaches 1 and the output preserves the original, only lightly smoothed, intensity of the focus. The saliency map also responds to bright edges such as the renal capsule, where blending restores sharpness that denoising may have reduced.

Region of Interest Extraction

The enhanced image is smoothed (σ = 2.5 for CT, 4 for ultrasound) and divided into four intensity classes by multi-level Otsu thresholding [23], [24], which selects thresholds t1<t2<t3 maximizing the between-class variance, as in (6):

$$(t_1^*, t_2^*, t_3^*) = \operatorname{argmax}_t \sum_{c=1}^{4} \omega_c(\mu_c - \mu_T)^2 \quad (6)$$

where ωc and μc are the probability and mean of class c and μT the global mean. Renal parenchyma belongs to the brightest tissue class, while stones are too small to form a class of their own. The mask of this class is opened morphologically, the connected component of plausible kidney size (between 1,500 pixels and a quarter of the image) closest to the image center is retained, and the hilum and sinus are filled by closing and hole filling. A bounding box with a margin of 8 pixels defines the crop, which is equalized with CLAHE [10] (clip limit 0.01, 32 × 32 tiles) and resized to 224 × 224 pixels.

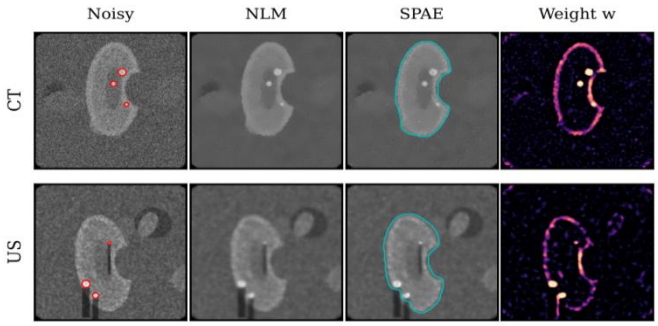

Fig. 2. Preprocessing Stages on Synthetic CT (Top) and Ultrasound (Bottom) Kidney Phantoms

Fig. 2. illustrates the pipeline. Red contours mark the true stones on the noisy images and cyan contours the extracted ROI on the SPAE output. Non-local means and SRAD remove most noise, but SRAD visibly flattens the small echogenic foci; the saliency weight concentrates on stones and on the bright renal boundary, and SPAE restores the stones while keeping the background smooth. The extracted ROI follows the kidney outline, so the crop passed to the CNN contains the kidney with its stones and excludes most surrounding structures.

CNN Architectures for Comparison

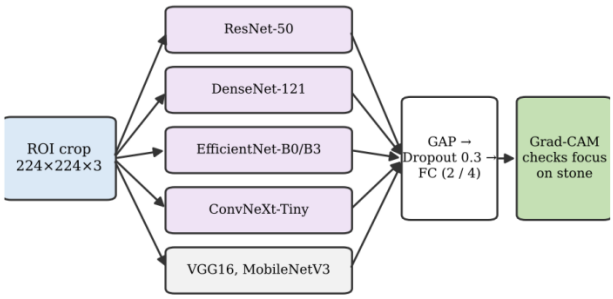

Fig. 3. Common Classification Head Used for All Compared Backbones

Fig. 3. shows the comparison design. ResNet-50 [25], DenseNet-121 [26], EfficientNet-B0 and B3 [27], ConvNeXt-Tiny [28], VGG16 [29] and MobileNetV3-Large [30] are initialized with ImageNet weights, and their classification layers are replaced by global average pooling, dropout of 0.3 and a fully connected layer with two (stone versus normal) or four outputs. Training minimizes class-weighted cross-entropy, as in (7):

$$\mathscr{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c}\alpha_c y_{ic}\log\hat{p}_{ic} \qquad (7)$$

where αc is inversely proportional to the frequency of class c, which compensates for the smaller number of stone images. Grad-CAM [31] is used to check that the networks attend to the stone region rather than to unrelated structures.

Results and Discussion

This section first validates the preprocessing pipeline on synthetic kidney phantoms, where noise-free references, kidney masks and stone positions are known exactly, and then specifies the CNN comparison on the public CT and ultrasound datasets, whose experiments are in progress. All preprocessing was implemented in Python 3.11 with NumPy, SciPy and scikit-image.

Phantom Design and Evaluation Protocol

Each 256 × 256 phantom contains a body outline, four randomly placed organs or bowel loops of variable intensity and a bean-shaped kidney with brighter cortex, darker medulla, renal sinus and hilar notch. One to three stones with radii of 1–4 pixels and intensities between 0.66 and 0.90 are placed within the parenchyma; in ultrasound phantoms each stone casts a posterior acoustic shadow. A Gaussian blur of σ = 0.7 models the system point-spread function and partial-volume effects, which lowers the contrast of the smallest stones as in early disease. CT noise is additive Gaussian with standard deviation 0.04, 0.07 or 0.10, and ultrasound speckle is multiplicative, derived from a spatially correlated Rayleigh field with strength 0.3, 0.5 or 0.7.

Image quality is measured by PSNR and SSIM [32] against the noise-free phantom and by the stone contrast-to-noise ratio, as in (8):

$$\mathrm{CNR} = \frac{\mu_{stone} - \mu_{kidney}}{\sigma_{kidney}} \qquad (8)$$

Detectability is measured with a simple, method-independent detector: local maxima of the white top-hat transform inside the kidney are candidate stones, the threshold is set on 30 calibration phantoms so that on average about one false positive occurs per image, and recall is computed on independent test phantoms, separately for small (radius ≤ 2 pixels) and larger stones. A candidate within the stone radius plus 2 pixels counts as a hit. ROI extraction is evaluated by the Dice coefficient with the true kidney mask, as in (9):

$$\mathrm{Dice} = \frac{2|A \cap K|}{|A| + |K|} \qquad (9)$$

Table 3. Preprocessing Parameters

Component	Setting

Component	Setting
CT denoising	Non-local means, 5 × 5 patches, search 13 × 13, h = 0.8σ̂ (Immerkær estimate)
Ultrasound denoising	SRAD, 30 iterations, Δt = 0.15, decaying speckle scale
SPAE saliency	LoG scales 1–4 px, top-hat disk radius 5, k = 4
ROI extraction	Gaussian σ = 2.5 (CT) / 4 (US), 4-class multi-Otsu, opening ×2, closing ×6
Contrast enhancement	CLAHE, clip limit 0.01, 32 × 32 tiles
Baselines	Gaussian σ = 1, median 3 × 3, bilateral (σ_c = 0.08, σ_s = 2)
Phantoms	30 calibration + 40 test per noise level; confirmation 30 + 120

Table III lists the settings. All parameters were fixed before the evaluation, and the same detector and threshold-calibration procedure were applied to every method so that differences in recall reflect the preprocessing alone.

Noise Removal and Image Quality

Table 4. Image Quality on CT Phantoms (40 Test Images per Level)

Method	PSNR (dB) σ=0.04 / 0.07 / 0.10	SSIM σ=0.07	CNR σ=0.07
No filtering	28.68 / 23.84 / 20.74	0.223	2.95
Gaussian	35.69 / 31.87 / 29.07	0.620	4.30
Median	35.91 / 31.30 / 28.30	0.587	4.12
Bilateral	35.39 / 29.55 / 24.64	0.460	3.86
Non-local means	38.86 / 34.22 / 31.13	0.725	5.42
SPAE (proposed)	38.43 / 33.40 / 29.86	0.688	4.92

Table 5. Image Quality on Ultrasound Phantoms (40 Test Images per Level)

Method	PSNR (dB) speckle 0.3 / 0.5 / 0.7	SSIM 0.5	CNR 0.5
No filtering	34.31 / 30.55 / 27.83	0.635	2.90
Gaussian	36.76 / 33.90 / 31.52	0.801	2.79
Median	37.59 / 33.52 / 30.69	0.791	2.93

Method	PSNR (dB) speckle 0.3 / 0.5 / 0.7	SSIM 0.5	CNR 0.5
Bilateral	36.61 / 35.37 / 33.48	0.889	3.12
SRAD	36.23 / 34.53 / 32.78	0.858	2.57
SPAE (proposed)	36.30 / 34.11 / 31.91	0.841	3.24

Tables IV and V show that, for CT, non-local means gave the highest PSNR at all noise levels (38.86, 34.22 and 31.13 dB) and SPAE remained within 0.4–1.3 dB of it, while bilateral filtering degraded strongly at high noise. For ultrasound, bilateral filtering and SRAD gave the highest PSNR and SSIM, and SPAE stayed within 0.9 dB of SRAD (and was 0.07 dB higher at the lowest speckle level). SPAE thus accepts a small loss in global image quality, which results from restoring slightly noisier intensities at salient locations, in exchange for preserving small structures. In ultrasound, SPAE gave the highest stone CNR (3.24 versus 2.57 for SRAD at speckle 0.5), whereas in CT non-local means already preserved stone contrast best (5.42 versus 4.92).

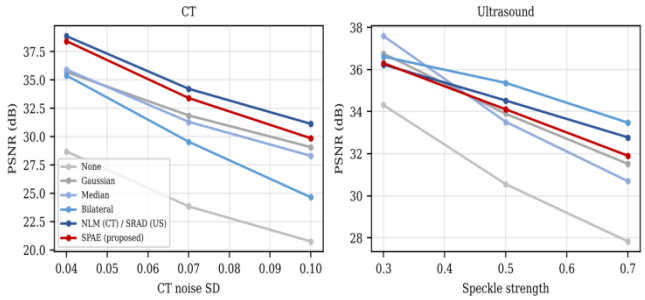

Fig. 4. PSNR of the Preprocessing Methods Across Noise Levels

Fig. 4. shows that all filters improved PSNR over the noisy input and that the ranking of methods was stable across noise levels. Global indices alone, however, would favor non-local means for CT and bilateral filtering for ultrasound, which motivates the detectability analysis below.

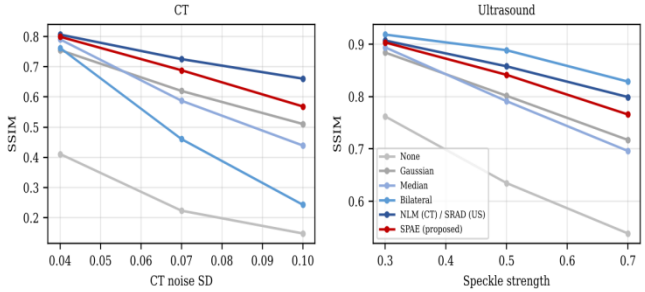

Fig. 5. SSIM of the Preprocessing Methods Across Noise Levels

Fig. 5. shows the corresponding SSIM values. The structural similarity of SPAE followed that of the base denoisers closely in both modalities, while bilateral filtering lost structural similarity rapidly with increasing CT noise. Together with Fig. 4, this shows that SPAE does not compromise the overall structure of the image while modifying it locally at candidate stones.

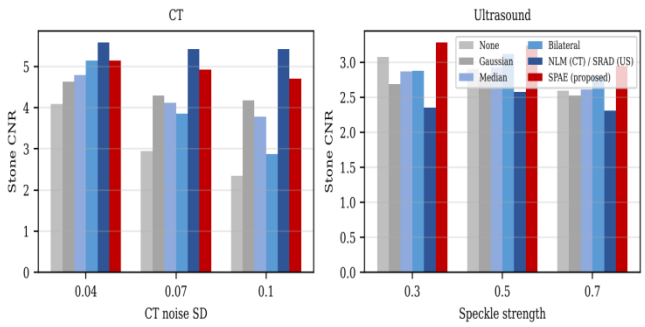

Fig. 6. Stone Contrast-to-Noise Ratio Across Methods and Noise Levels

Fig. 6. shows the stone contrast-to-noise ratio. In CT, non-local means achieved the highest CNR at all noise levels, with SPAE second, which is consistent with its strong edge- and detail-preserving behaviour. In ultrasound, SPAE achieved the highest CNR at every speckle level, while SRAD gave the lowest, confirming quantitatively that anisotropic diffusion reduces the conspicuity of small echogenic foci relative to their surroundings.

Small-Stone Detectability

Table 6. Stone Recall (%) at About One False Positive per Image

Modality / method	Small stones (r ≤ 2 px)	Larger stones (r = 3–4 px)	95% CI, small
CT, noise σ = 0.07 (120 test phantoms)			
No filtering	78.9	79.0	71.2–86.5
Non-local means	82.0	92.4	75.0–88.5
SPAE (proposed)	89.1	96.2	82.9–93.8
Ultrasound, speckle 0.5 (120 test phantoms)			
No filtering	80.3	81.3	72.3–87.1
SRAD	63.1	79.7	54.5–71.7
SPAE (proposed)	76.2	80.5	67.4–83.3

Table VI reports the confirmation experiment with 120 test phantoms per modality, containing 129 small CT stones and 122 small ultrasound stones. In ultrasound, SRAD reduced small-stone recall from 80.3% without filtering to 63.1%, confirming that speckle reduction erases small echogenic foci. SPAE restored recall to 76.2%, 13.1 percentage points above SRAD with non-overlapping confidence intervals, although not above the unfiltered image. In CT, SPAE achieved the highest small-stone recall (89.1% versus 82.0% for non-local means and 78.9% without filtering) and the highest recall for larger stones (96.2%); the confidence intervals of SPAE and non-local means overlap, so the CT advantage should be confirmed on real data. In the 40-image grid experiment, SPAE gave the best small-stone recall in ultrasound at every speckle level (97.8%, 92.9% and 71.4% versus 64.4%, 64.3% and 52.4% for SRAD), whereas in CT it matched non-local means at the lowest noise level, exceeded it at the middle level (91.3% versus 87.0%) and was slightly lower at the highest level (78.4% versus 81.1%).

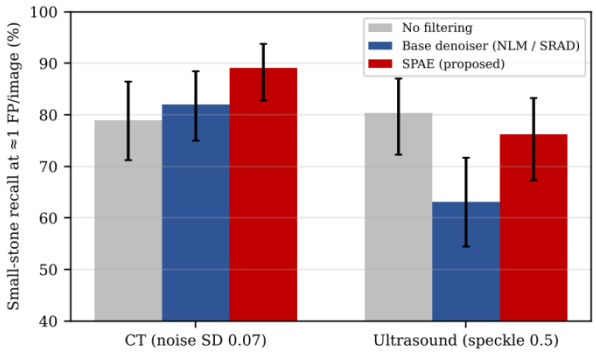

Fig. 7. Small-Stone Recall With 95% Bootstrap Confidence Intervals

Fig. 7. summarizes the detectability results. The key finding is that a denoiser can improve global image quality while harming the clinical task: SRAD increased ultrasound PSNR by about 4 dB yet lost about a fifth of the small stones that were detectable in the unfiltered image. Preprocessing for early stone detection should therefore be evaluated with task-based measures.

Table 7. Small / Larger Stone Recall (%) Across Noise Levels (40 Test Phantoms per Level)

Method	CT σ = 0.04	CT σ = 0.07	CT σ = 0.10	US 0.3	US 0.5	US 0.7
No filtering	91 / 98	76 / 83	54 / 64	93 / 89	86 / 85	67 / 80
Gaussian	91 / 100	91 / 97	86 / 87	89 / 86	79 / 87	60 / 77
Median	94 / 100	83 / 97	81 / 79	89 / 89	79 / 87	57 / 77
Bilateral	63 / 98	52 / 91	54 / 64	67 / 68	57 / 81	57 / 80
NLM (CT) / SRAD (US)	94 / 98	87 / 97	81 / 81	64 / 76	64 / 87	52 / 74
SPAE (proposed)	94 / 100	91 / 97	78 / 85	98 / 92	93 / 91	71 / 83

Table VII extends the comparison to all methods and noise levels. Two further observations emerge. Bilateral filtering, which ranked well on PSNR and SSIM in ultrasound, was the most harmful filter for small stones (52–67% recall), because its range kernel treats small bright foci as outliers.

Simple Gaussian smoothing was surprisingly competitive in CT at the highest noise level (86% small-stone recall), indicating that at very low dose the averaging of a few pixels can outweigh the edge preservation of more elaborate filters. In ultrasound, SPAE was the best method at every speckle level for both stone sizes.

ROI Extraction

ROI extraction on SPAE-enhanced phantoms reached mean Dice coefficients of 0.94–0.95 for CT across noise levels and 0.88–0.90 for ultrasound, where speckle and acoustic shadows blur the kidney boundary. In a separate check on 20 phantoms per modality, the bounding-box crop contained 100% of the stone pixels.

Restricting the CNN input to this crop removes on average about four fifths of the image area, including organs and bowel loops that can produce stone-like bright spots, and enlarges the kidney in the 224 × 224 network input, which increases the number of pixels covering each stone.

Table 8. ROI Extraction Dice Coefficient Across Noise Levels

Input to ROI extraction	CT 0.04	CT 0.07	CT 0.10	US 0.3	US 0.5	US 0.7
Raw image	0.943	0.940	0.945	0.906	0.879	0.895
Base denoiser	0.944	0.941	0.943	0.902	0.875	0.889
SPAE	0.944	0.942	0.945	0.902	0.876	0.893

Table VIII shows that ROI extraction was robust to the preprocessing choice and to the noise level, with Dice coefficients of 0.940–0.945 in CT and 0.875–0.906 in ultrasound. Because the ROI step relies on strong smoothing and on the size and position of the brightest tissue component, it is largely insensitive to fine-scale noise; the lower ultrasound values reflect boundary blurring by speckle and acoustic shadowing rather than gross failures.

Interaction With Contrast Enhancement and ROI Cropping

Table 9. Effect of Denoising Before CLAHE and of ROI Cropping (40 Phantoms per Modality)

Measure	CT ($\sigma = 0.07$)	Ultrasound (0.5)
Kidney background SD after CLAHE: raw / base / SPAE	0.202 / 0.109 / 0.116	0.188 / 0.169 / 0.159
Stone CNR after CLAHE: raw / base / SPAE	1.61 / 2.83 / 2.80	1.83 / 1.72 / 2.10
Stone pixels in 224 × 224 input: whole image / ROI crop	46.1 / 242.9	47.9 / 231.8

Table IX examines the last two pipeline stages. Applying CLAHE directly to noisy images nearly doubled the background variation within the kidney in CT (0.202 versus 0.109 after denoising), confirming that contrast enhancement must follow noise removal. After CLAHE, SPAE retained the stone CNR of non-local means in CT (2.80 versus 2.83) and gave the highest CNR in ultrasound (2.10 versus 1.72 for SRAD and 1.83 for the raw image). ROI cropping increased the number of input pixels covering stones about fivefold (from 46–48 to 232–243 pixels per image at 224 × 224), which gives the CNN considerably more spatial evidence for small stones than whole-image resizing.

Ablation and Parameter Sensitivity of SPAE

Table 10. Ablation of SPAE Components (60 Test Phantoms per Modality)

Configuration	US: PSNR (dB)	US: small recall (%)	CT: PSNR (dB)	CT: small recall (%)
No filtering	30.55	75.5	23.84	76.5
Base denoiser (SRAD / NLM)	34.52	56.6	34.22	89.7
SPAE, LoG saliency only	33.75	77.4	31.70	94.1
SPAE, top-hat saliency only	33.36	75.5	31.48	92.6

Configuration	US: PSNR (dB)	US: small recall (%)	CT: PSNR (dB)	CT: small recall (%)
SPAE, LoG × top-hat, k = 1	34.45	69.8	34.02	91.2
SPAE, LoG × top-hat, k = 2	34.33	75.5	33.79	92.6
SPAE, LoG × top-hat, k = 4 (default)	34.11	75.5	33.38	92.6
SPAE, LoG × top-hat, k = 8	33.80	75.5	32.75	92.6

Table X isolates the components of SPAE on a further 60 test phantoms per modality. Combining the LoG and top-hat responses multiplicatively gave the best balance: single-component saliency maps restored similar numbers of small stones but cost 0.4–0.8 dB more PSNR in ultrasound and about 1.7 dB more in CT, because each responds to more non-stone structures than their product. The blending strength k controls the trade-off; small-stone recall saturated at k = 2 in both modalities, while PSNR decreased steadily with larger k, so k = 2 would be a slightly better choice than the default k = 4 used in the main experiments.

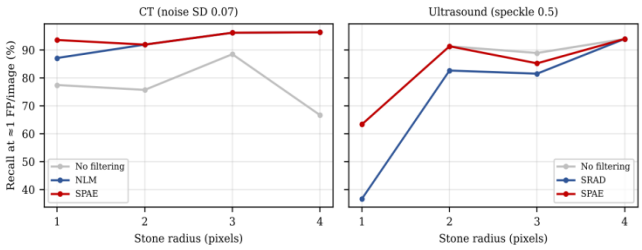

Fig. 8. Recall as a Function of Stone Radius

Fig. 8. shows recall by stone radius. The reduction in ultrasound loss resulting from SRAD was concentrated on the smallest stones (radius 1 pixel), where the recall decreased from 63% to 37%, and SPAE returned this back to 63%. For stones with radii of 3-4 pixels, there were no significant differences among any of the methods. In CT, both non-local means and SPAE increased recall for each stone radius over the original image, with SPAE having the largest increase in recall for radius-1 stones (94% vs. 87%). Therefore, the benefits of preserving stones through image enhancement are limited to those smallest and earliest stones, which have historically been the hardest to identify using clinical methods.

CNN Comparison Protocol

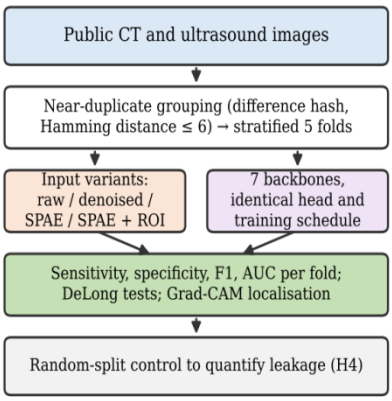

Fig. 9. Evaluation Protocol for the CNN Comparison

The evaluation process is outlined in Fig. 9. For every backbone, there will be training performed on four different types of input data: raw, denoised, SPAE-enhanced, and SPAE-enhanced with ROI cropping. This

allows the effect of each type of preprocessing to be evaluated for each of the architectures. In addition, an evaluation of the extent to which the reported accuracy of the model is affected by near-duplicate leakage will be conducted using a random split of images at the level of the image.

Table 11. Backbones Compared and Their Published Complexity

Backbone	Parameters (M)	GFLOPs (224^2)	Design feature
ResNet-50 [25]	25.6	4.1	Residual connections
DenseNet-121 [26]	8.0	2.9	Dense feature reuse
EfficientNet-B0 [27]	5.3	0.39	Compound scaling, MBConv
EfficientNet-B3 [27]	12	1.8	Larger compound-scaled model
MobileNetV3-L [30]	5.4	0.22	Mobile-optimized blocks
ConvNeXt-Tiny [28]	28.6	4.5	Modernized ConvNet design

Table XI shows the backbones along with their parameter counts and computational costs that were reported by the original authors. This selection of backbones includes classic, efficient, and modern designs, which will allow us to assess whether model size or architectural design is more important for identifying kidney stones after enhancing the images and cropping out the regions of interest. Each of these models has been fine-tuned using the AdamW optimizer [33] (at a learning rate of 1x10^-4; weight decay = 0.05; cosine scheduling; number of epochs = 30; batch size = 32; early stopping based upon validation loss). Additionally, all models used the same image augmentation techniques (rotation of +/-10 degrees; horizontal flipping; brightness adjustment of +/-10 percent), and five-fold cross-validation with nearly identical samples being grouped together within each fold.

Table 12. Training Configuration for the CNN Comparison

Item	Setting
Input	ROI crop, CLAHE, 224 × 224, 3 channels (replicated), ImageNet normalization
Optimizer	AdamW, learning rate 1×10^{-4}, weight decay 0.05, cosine decay
Schedule	30 epochs, batch 32, 3 warm-up epochs with frozen backbone, early stopping (patience 5)
Augmentation	Rotation ±10°, horizontal flip, brightness ±10%, small translation ±5%
Loss	Class-weighted cross-entropy (Eq. 7)
Validation	5-fold, stratified, near-duplicate groups kept within folds
Variants	Raw / denoised / SPAE / SPAE + ROI inputs for every backbone

Algorithm 1: SPAE_Preprocessing_and_Classification

Input: image v, modality m ∈ {CT, US}, trained classifier f

Output: stone probability and Grad-CAM map

1: if m = CT then standardize intensities; $\hat{\sigma} \leftarrow$ Immerkær(v); $\hat{u} \leftarrow$ NLM(v, h = 0.8$\hat{\sigma}$)

2: else $\hat{u} \leftarrow$ SRAD(v, 30 iterations)

3: $v_p \leftarrow (\hat{u} + G_1 * v) / 2$; $S \leftarrow [\max_s(-s^2\nabla^2 G_s * v_p)]_+ \cdot TH(v_p)$; normalize S

4: $w \leftarrow 1 - \exp(-k\,S)$; $\tilde{I} \leftarrow w\,(G_{0.5} * v) + (1 - w)\,\hat{u}$

5: $t^* \leftarrow$ multi-Otsu($G_\sigma * \tilde{I}$, 4 classes); M ← brightest class; open; select kidney component; close; fill holes

6: crop bounding box of M (+8 px); CLAHE; resize to 224 × 224

7: p ← f(crop); compute Grad-CAM; return p

Table 13. Hypotheses and Success Criteria for the CNN Comparison

Hypothesis	Experiment	Success criterion
H1: SPAE + ROI improves stone detection	Raw vs. denoised vs. SPAE vs. SPAE + ROI, each backbone	Higher stone sensitivity at equal specificity
H2: Gains are larger for ultrasound	Same comparison per modality	Larger improvement on ultrasound than CT
H3: Modern architectures help	ConvNeXt-T and EfficientNet vs. VGG16 and ResNet-50	Higher AUC; DeLong $p < 0.05$
H4: Leakage inflates accuracy	Random vs. duplicate-grouped splits	Report accuracy difference

Table XIII states the hypotheses. Sensitivity for the stone class, specificity, F1-score and AUC are reported, together with Grad-CAM localization relative to the stone region. H4 directly tests whether the near-perfect accuracies reported on the public CT dataset depend on the splitting procedure.

Detection performance in the CNN comparison is summarized by sensitivity, specificity and F1-score for the stone class, as in (10) and (11), and by the area under the ROC curve:

$$\mathrm{Sens} = \frac{\mathrm{TP}}{\mathrm{TP+FN}},\ \mathrm{Spec} = \frac{\mathrm{TN}}{\mathrm{TN+FP}} \qquad (10)$$

$$\mathrm{F1} = \frac{2 \cdot \mathrm{Prec} \cdot \mathrm{Sens}}{\mathrm{Prec+Sens}},\ \mathrm{Prec} = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \qquad (11)$$

Sensitivity is the primary endpoint because missed stones are clinically more consequential than false alarms in a screening setting; results are reported as mean and standard deviation over the five folds, and differences in AUC between backbones are tested with DeLong's test.

Table 14. Measured Runtime of Preprocessing Stages (256 × 256 Image, One CPU Core)

Stage	CT (ms)	Ultrasound (ms)
Base denoising (NLM / SRAD)	76.7	57.4
Stone saliency (LoG × top-hat)	24.8	23.4
Complete SPAE	89.4	73.6
ROI extraction	42.1	48.0
CLAHE	5.3	5.2

Table XIV reports runtimes measured on one core of a 2.3 GHz processor, averaged over 10 runs. The complete preprocessing chain takes about 0.14 s per CT image and 0.13 s per ultrasound frame, which is negligible compared with image acquisition and compatible with near-real-time use in ultrasound examinations; SPAE adds only about 13–16 ms to the base denoisers.

Fig. 10. shows accuracies reported in recent studies on CT data [5], [6], [13]. These results were obtained with different tasks, data and splits and are not directly comparable; values close to 100% on image-level splits of a dataset with many similar slices per patient are likely to overestimate performance on new patients. The proposed framework contributes a preprocessing stage whose benefit for small stones is measured explicitly and a comparison protocol designed to avoid such inflation.

Comparison with Existing Studies

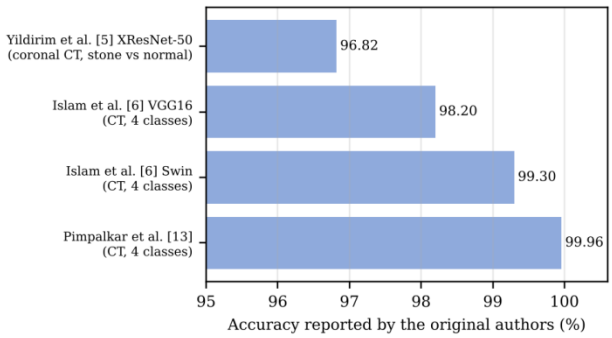

Fig. 10. Accuracies Reported by Recent Kidney Stone Detection Studies

Implementation and Reproducibility

All preprocessing components and the phantom generator are implemented with open-source libraries and fixed random seeds, so that every table and figure in this section can be regenerated. The CNN comparison is implemented in PyTorch with publicly available ImageNet weights; data splits, near-duplicate groups and trained weights will be released with the code to allow independent verification of the planned results.

Discussion

The phantom experiments show that preprocessing choices made for image quality can work against early detection. Speckle reduction, which is almost universally applied to ultrasound, removed about a fifth of the small stones that were detectable in the unfiltered image, and SPAE recovered most of this loss while keeping the background smooth. In CT, non-local means preserved small stones better than other filters, and the saliency-guided blending added a further, smaller improvement. The ROI stage complements enhancement by

concentrating the network input on the kidney, which both removes distractors and increases the effective resolution of small stones.

These findings have practical implications. Low-dose CT protocols, which are increasingly recommended for suspected stone disease to reduce radiation exposure, raise image noise and thus the importance of noise removal that does not erase small stones. Point-of-care ultrasound, often acquired by non-specialists with portable devices, produces images with strong speckle in which automated detection could support triage; the results suggest that conventional despeckling should not be applied blindly before automated stone detection. In both settings, task-based evaluation of preprocessing with known small targets, as performed here, offers a reproducible way to choose filters and parameters.

Threats to validity include the dependence of the detectability results on the simple detector and on the threshold calibration at one false positive per image; different operating points could change the ranking of methods. The phantom stone intensities and sizes were chosen to represent early, low-contrast stones, and larger or denser stones would be detected by all methods. The absence of individual patient information within the public CT data set that will be used for the CNN comparison restricts the near-duplicate grouping to approximating a split at the patient level. There are several limitations that must be addressed. The phantoms represent kidney anatomy, stones, noise, and speckle in simple ways; however, they do not simulate motion, beam hardening, ring-down or twinkling artifacts, vascular calcifications or phleboliths (all of which can appear to be stones). The detector used for the detectability analysis was purposely kept simple, therefore, absolute recall values cannot be interpreted as representing CNN performance. Additionally, the saliency map also identifies bright edges, potentially increasing false positive rates near the renal capsule. Finally, a CNN comparison using real patient data is ongoing and it has yet to be demonstrated if the improvements seen with phantoms translate into improved clinical image quality.

Conclusion and Future Work

The authors developed a reliable method to identify kidney stones early through both CT and ultrasound scans, using an adaptive enhancement pipeline based on preserving the stone's characteristics. This approach included denoising the images based upon their modality, restoring the small bright spots on each image using saliency-guided techniques, automatically identifying the regions of interest for the kidneys and enhancing their contrast, and then comparing several different types of convolutional neural network (CNN) models.On synthetic phantoms with known stones, SPAE kept noise-reduction quality close to non-local means and SRAD, restored small-stone recall in ultrasound from 63.1% after SRAD to 76.2%, and increased small-stone recall in CT from 82.0% to 89.1%, while ROI extraction achieved Dice coefficients of 0.88–0.95 and retained all stones in the crop.

Future work will complete the CNN comparison on the public CT and ultrasound datasets with duplicate-grouped cross-validation, evaluate the pipeline on clinical images with radiologist-annotated stones of known size, add stone segmentation and size estimation, and extend the framework to the multimodal fusion and explainability stages of this research programme.

References

[1] S. R. Khan et al., “Kidney stones,” Nature Reviews Disease Primers, vol. 2, art. 16008, 2016.
[2] C. D. Scales Jr., A. C. Smith, J. M. Hanley, and C. S. Saigal, “Prevalence of kidney stones in the United States,” European Urology, vol. 62, no. 1, pp. 160–165, 2012.
[3] A. Brisbane, M. R. Bailey, and M. D. Sorensen, “An overview of kidney stone imaging techniques,” Nature Reviews Urology, vol. 13, no. 11, pp. 654–662, 2016.
[4] R. Smith-Bindman et al., “Ultrasonography versus computed tomography for suspected nephrolithiasis,” New England Journal of Medicine, vol. 371, no. 12, pp. 1100–1110, 2014.
[5] K. Yildirim, P. G. Bozdag, M. Talo, O. Yildirim, M. Karabatak, and U. R. Acharya, “Deep learning model for automated kidney stone detection using coronal CT images,” Computers in Biology and Medicine, vol. 135, art. 104569, 2021.
[6] M. N. Islam, M. Hasan, M. K. Hossain, M. G. R. Alam, M. Z. Uddin, and A. Soylu, “Vision transformer and explainable transfer learning models for auto detection of kidney cyst, stone and tumor from CT-radiography,” Scientific Reports, vol. 12, art. 11440, 2022.
[7] K. Sudharson and P. Kokil, “An ensemble of deep neural networks for kidney ultrasound image classification,” Computer Methods and Programs in Biomedicine, vol. 197, art. 105709, 2020.
[8] Y. Yu and S. T. Acton, “Speckle reducing anisotropic diffusion,” IEEE Transactions on Image Processing, vol. 11, no. 11, pp. 1260–1270, 2002.
[9] A. Buades, B. Coll, and J.-M. Morel, “A non-local algorithm for image denoising,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2, 2005, pp. 60–65.
[10] K. Zuiderveld, “Contrast limited adaptive histogram equalization,” in Graphics Gems IV, P. S. Heckbert, Ed. San Diego, CA, USA: Academic Press, 1994, pp. 474–485.

[11] K. Sudharson and P. Kokil, “Computer-aided diagnosis system for the classification of multi-class kidney abnormalities in the noisy ultrasound images,” Computer Methods and Programs in Biomedicine, vol. 205, art. 106071, 2021.
[12] S. Elbedwehy, E. Hassan, A. Saber, and R. Elmonier, “Integrating neural networks with advanced optimization techniques for accurate kidney disease diagnosis,” Scientific Reports, vol. 14, 2024, doi: 10.1038/s41598-024-71410-6.
[13] A. Pimpalkar, D. K. J. B. Saini, N. Shelke, and A. Balodi, “Fine-tuned deep learning models for early detection and classification of kidney conditions in CT imaging,” Scientific Reports, vol. 15, art. 10741, 2025.
[14] K. Sharma, Z. Uddin, A. Wadal, and D. Gupta, “Hybrid deep learning framework for classification of kidney CT images: diagnosis of stones, cysts, and tumors,” arXiv:2502.04367, 2025.
[15] R. Khan, R. Mahum, U. Irshad, M. Shehab, and F. S. Butt, “Enhanced feature extraction for detection and classification of kidney abnormalities,” Current Medical Imaging, vol. 22, art. e15734056420229, 2026.
[16] R. Singla, C. Ringstrom, G. Hu, V. Lessoway, J. Reid, C. Nguan, and R. Rohling, “The Open Kidney Ultrasound Data Set,” in Simplifying Medical Ultrasound (ASMUS 2023), Lecture Notes in Computer Science, vol. 14337, 2023, pp. 155–164.
[17] O. V. Michailovich and A. Tannenbaum, “Despeckling of medical ultrasound images,” IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control, vol. 53, no. 1, pp. 64–78, 2006.
[18] S. V. M. Sagheer and S. N. George, “A review on medical image denoising algorithms,” Biomedical Signal Processing and Control, vol. 61, art. 102036, 2020.
[19] H. Chen, Y. Zhang, M. K. Kalra, F. Lin, Y. Chen, P. Liao, J. Zhou, and G. Wang, “Low-dose CT with a residual encoder-decoder convolutional neural network,” IEEE Transactions on Medical Imaging, vol. 36, no. 12, pp. 2524–2535, 2017.
[20] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: convolutional networks for biomedical image segmentation,” in Medical Image Computing and Computer-Assisted Intervention (MICCAI 2015), Lecture Notes in Computer Science, vol. 9351, 2015, pp. 234–241.
[21] J. Immerkær, “Fast noise variance estimation,” Computer Vision and Image Understanding, vol. 64, no. 2, pp. 300–302, 1996.
[22] T. Lindeberg, “Feature detection with automatic scale selection,” International Journal of Computer Vision, vol. 30, no. 2, pp. 79–116, 1998.
[23] N. Otsu, “A threshold selection method from gray-level histograms,” IEEE Transactions on Systems, Man, and Cybernetics, vol. 9, no. 1, pp. 62–66, 1979.
[24] P.-S. Liao, T.-S. Chen, and P.-C. Chung, “A fast algorithm for multilevel thresholding,” Journal of Information Science and Engineering, vol. 17, no. 5, pp. 713–727, 2001.
[25] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.
[26] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4700–4708.
[27] M. Tan and Q. V. Le, “EfficientNet: rethinking model scaling for convolutional neural networks,” in Proc. 36th International Conference on Machine Learning (ICML), 2019, pp. 6105–6114.
[28] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, “A ConvNet for the 2020s,” in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11976–11986.
[29] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in Proc. International Conference on Learning Representations (ICLR), 2015.
[30] A. Howard et al., “Searching for MobileNetV3,” in Proc. IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1314–1324.
[31] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: visual explanations from deep networks via gradient-based localization,” in Proc. IEEE International Conference on Computer Vision (ICCV), 2017, pp. 618–626.
[32] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600–612, 2004.
[33] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” in Proc. International Conference on Learning Representations (ICLR), 2019.