International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 6, Nov-Dec-2026, ISSN (Online): 2395-566X

Federated and Explainable Multi-Task Machine Learning for Kidney Stone Detection and Composition Estimation: A Systematic Literature Synthesis, Empirical Re-Analysis, and Pro-posed Framework (Fed-SCX)

Shivkumar Rajput
Biomedical Machine Learning · Urology · Medical Imaging

Abstract — Background: Kidney stone disease (nephrolithiasis) affects a growing share of the population, with U.S. age-standardized prevalence rising from roughly 9.4% (2007–2008) to 10.2% (2017–2020) and continuing reliance on radiation-intensive computed tomography (CT) for definitive diagnosis. Machine learning (ML) and deep learning (DL) methods applied to CT imaging, ultrasound, and urinary biochemistry have reported very high classification accuracies, but the litera-ture is fragmented across single, frequently reused datasets. Objective: This report synthesizes the current evidence base, empirically re-analyzes an open urinary-biochemistry dataset to ground the discussion in independently verifiable numbers, identifies the specific research gap that remains after accounting for the most recent (2025–2026) contributions, and proposes a conceptual framework— Fed-SCX — that responds directly to that gap. Methods: A structured narrative synthesis of approximately 45 peer-reviewed articles, preprints, and epi-demiological surveys (2012–2026, concentrated in 2024–2026) was combined with a reproducible five-fold cross-validated re-analysis of a 414-record open urinary physicochemical dataset using logistic regression, random forest, support-vector, and gradient-boosting classifiers. Results: The literature shows near-ceiling reported accuracies (97–100%) on CT classification tasks that trace almost entirely to one public 12,446-image dataset, while our own re-analysis of the (separately reused) urinary dataset produced a materially lower, more conser-vative cross-validated ROC-AUC of 0.78–0.80, suggesting that some published figures may be inflated by dataset reuse and undocumented augmentation. Fed-erated (privacy-preserving, multi-institutional) learning is mature in adjacent medical-imaging domains but was not found applied specifically to kidney stones. Multimodal fusion has just begun to appear (2025–2026) but remains centralized and single-institution. Contribution: Fed-SCX combines multimodal fusion, federated cross-site training, dual explainability (Grad-CAM plus SHAP), and calibrated-uncertainty deferral — a combination not identified in any reviewed source. Conclusion: The framework is conceptual and has not been empirically trained or validated; the report closes with the specific external-validation study required before any clinical claims could be made.

Keywords— kidney stone detection; nephrolithiasis; machine learning; deep learning; federated learning; explainable AI; multimodal data fusion; medical imaging

I. INTRODUCTION

1. Background

Kidney stones (nephrolithiasis, urolithiasis) are crystalline deposits that form in the renal collecting system when urine becomes persistently supersaturated with stone-forming solutes, most commonly calcium oxalate. The condition is common and becoming more so: age-standardized prevalence among U.S. adults rose from 9.4% in the 2007–2008 NHANES cycle to 10.2% in 2017–2020, with the increase concentrated among adults with obesity and, within that group, disproportionately among women under 60 (Li et al., 2025; Chen et al., 2025). A widely cited earlier NHANES-based estimate placed lifetime prevalence at approximately 11% in men and 7% in women (Scales et al., 2012), and longer-run comparisons across NHANES cycles suggest prevalence has risen from roughly 3.8% in the late 1970s to current levels, with the trend stabilizing in men but continuing to climb in women (Chen et al., 2025).

Reported global incidence ranges from approximately 1% in parts of Asia to as high as 13% in North America, reflecting genuine geographic and dietary variation as well as differences in how studies measure the condition (imaging-confirmed versus self-report) — a methodological point taken up in Section 3.

© 2026 IJSRET

The clinical and economic stakes are substantial. Kidney stones dispropor-tionately affect working-age adults (20–60 years), generate a large volume of emergency-department visits, and recur in up to 50% of patients within five to ten years (Hyams, 2014; PMC12568256). Older, frequently cited U.S. cost estimates place the direct and indirect annual burden above $4.5–5 billion, with roughly $775 million attributable to lost workdays alone (Hyams, 2014). A precise, current (2024–2026) aggregate cost figure specific to nephrolithiasis, as opposed to chronic kidney disease broadly, was not located in the sources consulted for this report; the figures above should be read as historically grounded orders of magnitude rather than current totals.

2. The Diagnostic Pathway and Its Costs

Non-contrast CT is the diagnostic gold standard, with sensitivity and specificity consistently reported in the 95–98% and 96–100% ranges across multiple independent studies (2 Minute Medicine; PMC3999367; PMC12568256). That accuracy comes at a radiation cost: U.S. emergency departments perform an estimated1.6–2.1 million CT scans annually for suspected ureterolithiasis (PMC12568256), and because up to half of patients experience a recurrence within five to ten years, many patients accumulate repeated radiation exposure over a lifetime that often begins before age 40. Ultrasound avoids ionizing radiation and is recommended as the first-line modality in children and pregnant patients, but its reported sensitivity varies markedly across studies — from as low as 24% to as high as 96%, with specificity in the 82–91% range in some series (PMC6112662)— a spread that itself reflects operator dependence, stone location, and body habitus rather than a single settled figure. A cross-sectional Iraqi study, for example, found ultrasound missed 12.5% of stones relative to CT, with only fair agreement on stone location and hydronephrosis severity (PMC11131636). This heterogeneity, examined further in Section 3, is one of the clinical motivations for pursuing ML-assisted approaches that could narrow the performance gap between a radiation-free and a radiation-based modality.

3. Why Machine Learning, and Why Now

Over the past five years, convolutional neural networks (CNNs) and, more recently, transformer- and object-detection-based architectures (e.g., YOLOv8) have been applied to CT and ultrasound images of the kidney with reported classification accuracies frequently exceeding 98%, and in several cases exceeding 99.7% (see Section 3.2). In parallel, a smaller but methodologically interesting body of work uses urinary biochemistry — specific gravity, pH, osmolarity, conductivity, urea, and calcium concentration — as an inexpensive, non-imaging predictor of stone risk. A separate line of research targets not detection but composition prediction — whether a stone is calcium oxalate, uric acid, struvite, or cystine — because composition determines whether a patient is a candidate for non-surgical oral chemolysis or requires surgical intervention (De Coninck et al., as cited in BMC Urology sources; World Journal of Urology, 2024).

These three threads — imaging detection, biochemical risk prediction, and composition estimation — have so far developed largely in isolation, on datasets that, as this report documents in Sections 3 and 4, trace back to a very small number of original data sources reused across dozens of nominally distinct studies. At the same time, federated learning — a privacy-preserving technique that trains a shared model across multiple hospitals without moving patient data — has matured substantially in adjacent domains such as brain-tumor segmentation and PET/CT artifact correction, yet no kidney-stone-specific application of it was identified in the course of this review. This combination of circumstances — strong single-dataset results, an active but disconnected multimodal literature, and an unexploited federated-learning opportunity — motivates the synthesis, empirical check, and proposed framework that follow.

Research Objectives

This report pursues four linked objectives, each formulated as an answerable research question:

- RQ1. What is the current state of reported ML/DL performance across imaging-based detection, urinary-biochemistry-based prediction, and stone-composition-estimation tasks, and how comparable are these reported metrics once dataset provenance is accounted for?
- RQ2. What methodological or data-related factors most limit the generalizability and clinical translatability of existing models?
- RQ3. To what extent has federated (privacy-preserving, multi-institutional) learning been applied to kidney-stone-specific machine learning, relative to its demonstrated maturity in adjacent medical-imaging domains?
- RQ4. Can a conceptual framework that combines multimodal fusion, federated training, and multi-method explainability plausibly address the specific gaps surfaced by RQ1–RQ3, and how does it differ from the closest contemporaneous work?
- The objective is deliberately bounded: this report does not train or externally validate a new imaging model (Section 10 explains why), but it does (a) syn-thesize and critically compare the existing evidence, (b) independently re-run a reproducible empirical analysis on one open dataset so that

© 2026 IJSRET

at least one set of numbers in this report is directly verifiable rather than merely cited, and (c) propose and justify — without overclaiming validation — a specific next-step architecture.

II. EXISTING LITERATURE / STATE OF THE ART

This section synthesizes the reviewed literature across seven threads: conven-tional diagnostics, CT-based deep learning, ultrasound-based detection, urinary-biochemistry machine learning, composition prediction, federated learning in medical imaging generally, and explainability. Table 1 summarizes representative studies for direct comparison; the surrounding text provides context that a table cannot, including how consistently different papers reuse the same underlying data.

1. Conventional Diagnostics as the Performance Bench-mark

Because CT and ultrasound remain the clinical reference standards, ML systems are implicitly benchmarked against them. Non-contrast CT sensitivity/specificity figures cluster tightly around 95–98% / 96–100% across multiple independent sources (2 Minute Medicine; PMC3999367; PMC12568256), which is notable because it means CT itself already performs close to the ceiling that many ML papers report matching or approaching — raising the question of what clinical value an ML system adds if it merely reproduces CT-level accuracy on CT images rather than approaching CT-level accuracy from a lower-cost or lower-radiation modality. Ultrasound accuracy is comparatively unsettled: reported sensitivity spans from 24% to 96% depending on the study, stone location, and operator (PMC6112662; JournalFeed, 2021), and an Iraqi cross-sectional study found only fair agreement (= 0.368–0.394) between ultrasound and CT for stone location and hydronephrosis severity (PMC11131636). A physician-news summary of pooled prevalence estimates further reports that prevalence itself is measured differently depending on modality — 8.71% in ultrasound-based studies, 7.83% in CT-based studies, and 13.28% in self-report surveys (Physicians Weekly, 2026, summarizing an unnamed underlying meta-analysis) — a reminder that even the epidemiological ground truth this field is measured against is not fully standardized. This source is a secondary aggregator rather than a primary study and is reported here with that caveat.

2. CT-Based Deep Learning Detection and Classification

The dominant recent thread applies CNNs and, increasingly, object-detection architectures to axial or coronal CT slices, typically framed as a 4-class prob-lem (normal, cyst, tumor, stone) or a binary/detection problem (stone present, localized). Reported accuracies are strikingly high and strikingly consistent:99.73–100% for a hybrid ResNet101+CNN feature-fusion model (arXiv, 2025); 99.86% for a plain CNN benchmarked against ResNet50 and EfficientNetB0 (Springer, 2026); 98.52% for Inception-V3 (PubMed, 2023); 98.56% for a Kronecker-convolution architecture (ScienceDirect, 2021); and 98.80% for a purpose-built web platform, MyKidney, aimed at rural clinics (ScienceDirect, 2025). Object-detection approaches report complementary metrics: a 2026 Fron-tiers comparison of YOLOv8, YOLOv5, Faster R-CNN, and RetinaNet on 4,000 CT slices from 170 patients found Faster R-CNN achieved the highest local-ization accuracy (mAP@0.5 = 0.93) while YOLOv8 offered the best accuracy–speed balance (mAP@0.5 = 0.91 at 65 frames per second); a related two-stage YOLOv8n-plus-ResNet-18 architecture combines detection with regression-based severity grading at approximately 45 milliseconds per case (ScienceDirect, 2025). A pattern emerges on closer inspection: the great majority of these studies, including several with different architecture names and different research teams, report training on what appears to be the same publicly available Kaggle-hosted CT collection of 12,446 images (2,283 tumor, 1,377 stone, 3,709 cyst, 5,077 normal), originally sourced from PACS systems in Dhaka, Bangladesh (arXiv 2403.15895; arXiv 2502.04367; Nature Scientific Reports, 2025). This is addressed directly as a research gap in Section 4 and empirically illustrated with an independent dataset in Section 7.

3. Ultrasound-Based and Real-Time Detection

A smaller literature targets ultrasound directly, motivated by its radiation-free profile and suitability for pediatric and prenatal populations. YOLO-family object detectors dominate here as well: open-source tools such as UroVision (YOLOv8n, Streamlit-deployed) and conference papers describing YOLO-based kidney-stone detectors emphasize real-time inference on standard hardware rather than peak accuracy. Feasibility work on non-linear beamforming methods (plane-wave synthetic focusing, spatial-coherence imaging) shows promise for improving raw ultrasound stone contrast and sizing accuracy relative to standard B-mode, though this work remains at the in-vitro / small-pilot stage (n = 6 stones in 5 patients) (PMC6112662).

4. Urinary Biochemistry Machine Learning

A methodologically distinct thread avoids imaging altogether, using six physi-cal/chemical urine properties — specific gravity, pH, osmolarity, conductivity, urea concentration, and calcium concentration — to predict calcium-oxalate crystal formation. The dataset used throughout this sub-literature

© 2026 IJSRET

originates from a 1980s statistical teaching example, Physical Characteristics of Urines With and Without Crystals, comprising 79 specimens, as documented in the dataset repository consulted for this report. Recent preprints and conference papers report Extra Trees Classifier accuracy of 92.5% (ResearchGate preprint), alongside comparative evaluations of ten algorithms including K-Nearest Neigh-bors, SVM, Naive Bayes, Random Forest, AdaBoost, XGBoost, and stacking ensembles (IEEE Xplore conference paper). Publication dates for both could not be independently confirmed and are therefore omitted rather than estimated. Section 7 of this report presents an independent re-analysis of this same underly-ing dataset, which surfaces a specific reproducibility concern discussed in Section 4 and Section 8.

5. Stone Composition Prediction

Composition determines treatment: pure uric-acid stones can be dissolved with oral chemolysis, while calcium-based and struvite stones typically require surgical intervention (World Journal of Urology, 2024). This sub-literature is, on the whole, methodologically more rigorous than the pure-detection literature, likely because it engages more directly with the underlying physics of dual-energy CT (DECT). A deep-learning segmentation-plus-multi-label-classification pipeline achieved stone segmentation AUC = 0.96 and composition prediction with mean accuracy = 0.86, sensitivity = 0.75, specificity = 0.90, and AUC = 0.83, though the exact-match ratio for fully resolving mixed-composition stones was only 0.60 (ScienceDirect, 2024). A ResNet-18 Mask R-CNN model trained on 21,074 annotated images and fused with demographic and clinical variables was one of the few reviewed studies reporting external validation, for classifying uric-acid stones below an 800 Hounsfield-unit threshold (Urolithiasis, 2023). Classical Hounsfield-unit statistics (mean and standard deviation) remain a documented, simpler predictor of composition (Tailly et al., 2016), and endoscopic-image CNN approaches have achieved compound-specific recall as high as 94% for uric acid but as low as 71% for brushite, reflecting genuine differential difficulty across stone types (PubMed 32045113).

6. Recurrence Prediction

A newer thread predicts which patients will form a second stone. A TabNet model incorporating 235 clinical, 24-hour urine, and imaging variables achieved AUC = 0.89 for two-year recurrence after endoscopic combined intrarenal surgery, falling to a still-useful AUC = 0.75 when reduced to three features (urinary oxalate, urine volume, and three-month hemoglobin change) (PMC12193965, 2025) — in a cohort of only 72 patients, of whom 29 (40.3%) recurred. A separate study predicting 24-hour urine abnormalities directly illustrates a methodological hazard revisited in Section 8: an XGBoost model predicted urine-volume abnormality with 98% accuracy but an AUC of only 0.59 — barely better than chance — because the outcome was heavily imbalanced (ScienceDirect journal article; publication date not independently confirmed).

7. Federated Learning in Medical Imaging

Federated learning (FL) trains a shared model across multiple institutions by exchanging model-weight updates rather than patient data, addressing the pri-vacy and data-governance barriers that otherwise prevent multi-hospital datasets from being pooled (Radiology: Artificial Intelligence journal). The technique is well established in adjacent medical-imaging domains: 3D U-Net-based federated brain-tumor segmentation on the BraTS dataset (PMC11675895), differential-privacy-preserving federated transfer learning for correcting PET/CT artifacts across three countries and eight centers (PMC10684636), and general architec-tural reviews describing FL as enabling "rural clinics, large urban hospitals, and even consumer health products to contribute to and benefit from a single model without ever having to share sensitive patient data" (Archives of Computational Methods in Engineering, 2026). No study identified in this review applies feder-ated learning specifically to kidney-stone detection, classification, or composition prediction — a gap developed further in Section 4.

8. Explainability

A 2026 mini-review focused specifically on explainable AI in kidney-stone de-tection and segmentation found that most studies rely on small, geographically restricted datasets, limiting generalizability, and that approaches combining multiple explainability methods (for example, Grad-CAM alongside SHAP) provide more robust, actionable insight than either method alone — yet such combined approaches remain the exception rather than the rule (Frontiers in Digital Health, 2026). The same review explicitly names "limited use of multiple XAI methods" and "insufficient validation across populations or real-world settings" as persistent, unresolved limitations of the field as of early 2026.

9. Contemporaneous Multimodal Fusion (2025–2026)

Two very recent works anticipate part of the direction proposed in Section 6 of this report and are addressed here with the epistemic caution appropriate to unreviewed or newly published preprints. USCNet, an arXiv preprint from 2026, proposes a transformer-based multimodal fusion architecture combining CT images with electronic health records through a dedicated attention mod-ule, performing joint segmentation

© 2026 IJSRET

and classification; the authors themselves report "limited generalizability in complex clinical scenarios" and "relatively high computational complexity" as limitations, and the model is trained on a single institution's proprietary dataset. HQIT-Net, described in a 2025 preprint, combines CT images and urinary biochemistry — the same 79-specimen dataset discussed in Section 3.4 — through a quantum-inspired transformer and bidi-rectional LSTM with attention. Neither work incorporates federated cross-site training, calibrated uncertainty, or dual explainability, which is precisely the combination Section 4 identifies as the remaining gap.

10. Summary Comparison

Table 1. Representative studies across the reviewed literature (2021–2026), summarized for direct comparison.

Study / Year	Modality	Method	Reported Performance	Dataset (n)	Key Limitation Noted
Hybrid ResNet101+CNN (arXiv, 2025)	CT (NCCT)	Hybrid CNN + transfer learning, feature fusion	99.73% train / 100% test accuracy	12,446 images (single public set)	Test accuracy of 100% is a strong overfitting / leakage signal on a single-source set
CNN vs. ResNet50 vs. EfficientNetB0 (Springer chapter, 2026)	CT	Custom CNN, Adam optimizer	99.86% accuracy	Kaggle CT set	No external validation reported
Inception-V3 (PubMed, 2023)	CT (axial)	Transfer learning	98.52% test accuracy	Limited authentic CT images	Authors note limited/small authentic image count
Kronecker-conv CNN (ScienceDirect, 2021)	Coronal CT	Custom convolution (FDMO-optimized)	98.56% accuracy	Public GitHub CT set	Data imbalance and cross-protocol generalization flagged by authors
MyKidney platform CNN (ScienceDirect, 2025)	Axial CT	Custom CNN, web deployment	98.80% accuracy	"CT Kidney Dataset"	Designed for rural/low-resource deployment; single-dataset training
YOLOv8 vs. YOLOv5 vs. Faster R-CNN vs. RetinaNet (Frontiers, 2026)	CT (4000 slices)	Object detection, real-time	YOLOv8 mAP@0.5=0.91 @65 FPS; Faster R-CNN mAP@0.5=0.93	170 patients	Trade-off between accuracy and inference speed; single-center cohort
Two-stage YOLOv8n + ResNet-18 (ScienceDirect, 2025)	CT	Detection + regression severity grading	~45 ms/case inference; 5-fold CV	Not fully specified	Recommends multi-center validation (3–5 sites) as future work
FDMO-optimized hybrid CNN (ScienceDirect, 2024)	Coronal CT	Metaheuristic-tuned CNN	99.82% accuracy	Not fully specified	Authors themselves flag need for larger, more diverse validation; no FP/FN rates given
Extra Trees / SVM / KNN / RF (ResearchGate preprint)	Urinary biochemistry	Classical ML + SMOTE	92.5% accuracy (Extra Trees)	79 original specimens (expanded)	Small original n; expansion method not detailed
DECT segmentation + composition (ScienceDirect, 2024)	Dual-energy CT	DL segmentation + multi-label ML	Segmentation AUC=0.96; composition mean AUC=0.83, Acc=0.86	Not fully specified	Cannot yet resolve all mixed-composition cases (exact-match ratio 0.60)
ResNet-18 Mask R-CNN + clinical fusion (Urolithiasis, 2023)	CT + demographics	DL + tabular fusion, externally validated	Not fully specified in abstract	21,074 annotated images	One of few reviewed studies with external validation
TabNet recurrence model (PMC12193965, 2025)	24-hr urine + clinical	TabNet, SHAP-interpreted	AUC=0.89 (full); AUC=0.75 (3-feature)	72 post-ECIRS patients	Small cohort (n=72); recurrence outcome over only 2 years
USCNet (arXiv preprint, 2026)	CT + EHR	Transformer multi-modal fusion, multi-task	Outperforms baselines on proprietary set	Single-institution, proprietary	Authors note limited generalizability, high computational cost
HQIT-Net (ResearchGate, 2025)	CT + urine biochemistry	Quantum-inspired transformer + BiLSTM fusion	Not fully specified in available abstract	Combines the 79-specimen urine set + PACS CT set	Centralized training; no federated or cross-site component

Table 1. Representative comparison of reviewed ML/DL approaches to kidney stone detection, classification, and composition prediction. Sources as cited in the accompanying text and full reference list.

III. RESEARCH GAP

Sections 3.2 through 3.9 support five specific, evidence-grounded gap statements. Each is stated narrowly, with the disconfirming or complicating evidence noted alongside it, consistent with this report's commitment not to overstate novelty.

Gap 1 — Dataset monoculture. Both dominant public resources in this field— the 12,446-image CT set (Section 3.2) and the 79-specimen urinary dataset (Section 3.4) — are reused across

© 2026 IJSRET

a large fraction of the papers reviewed here, often without the reuse being foregrounded. Near-ceiling accuracies (97–100%) reported on the CT set by multiple independent architectures is a pattern more consistent with a dataset that is easy to memorize than with genuinely generalizable performance; none of the CT studies reviewed report external validation on an independently collected cohort.

Gap 2 — Accuracy-only reporting under class imbalance. Section 3.6 documents a case (24-hour urine abnormality prediction) where 98% reported accuracy corresponded to an AUC of only 0.59. This is not an isolated sta-tistical curiosity; it demonstrates that accuracy alone, the headline metric in most reviewed CT papers, can be misleading whenever class distributions are imbalanced — a condition that applies to stone-versus-normal classification in most clinical populations, where stones are the minority finding.

Gap 3 — Federated learning is absent specifically for kidney stones. Sec-tion 3.7 establishes that federated learning is mature for brain tumors, PET/CT artifacts, and general multi-institutional imaging pipelines. No kidney-stone-specific application was located in this review despite the field sharing the same underlying motivation — small single-center datasets, patient-privacy constraints, and a documented generalizability problem (Gap 1). This is an absence of ap-plication, not an absence of underlying capability; the constituent techniques already exist and are demonstrated elsewhere.

Gap 4 — Single-method explainability. Section 3.8 reports that a 2026 mini-review explicitly names limited use of multiple XAI methods together, and insufficient real-world/cross-population validation, as unresolved limitations of the field. This report treats that mini-review's own stated conclusion as the most authoritative and current single source for this gap, since it is itself a systematic survey of the relevant sub-literature published within the reliable knowledge horizon of this report.
Gap 5 — Fusion exists; federated, calibrated, dual-explainable fusion does not (yet). Section 3.9 shows that multimodal fusion for urolithiasis is not a novel idea as of this writing — USCNet and HQIT-Net, both 2025–2026 preprints, already pursue it. Neither, however, is federated, neither reports calibrated uncertainty for clinical deferral, and neither combines more than one explainability method. This is the specific, narrow space this report's proposed framework (Section 6) occupies.

1. What This Report Does Not Claim

Consistent with the anti-overclaiming principle guiding this report, three things are explicitly not claimed: (a) that multimodal fusion for kidney stones is a novel contribution of this report — it is not, per Gap 5; (b) that federated learning itself is a novel technique — it is well established, per Section 3.7; and (c) that the framework proposed in Section 6 has been trained or empirically validated— it has not, and Section 6 and Section 10 say so explicitly. What is proposed is a specific, currently absent combination of existing, individually validated techniques, applied to a domain where that combination has not yet been reported, justified by gaps that a recent systematic mini-review independently and explicitly identifies.

IV. METHODOLOGY

1. Research Design

This report uses a mixed-method design combining three components, each chosen for a specific reason:

- Structured narrative literature synthesis — appropriate because the research questions (Section 2) require comparing heterogeneous study designs, modalities, and metrics that a purely quantitative meta-analysis could not meaningfully pool (different outcome definitions, different popu-lations, inconsistent metric reporting).
- Independent empirical re-analysis of one open dataset — included because a literature review alone cannot verify whether reported figures are reproducible; re-running a transparent, reproducible analysis on real, independently obtained data provides at least one fully auditable result in this report, rather than asking the reader to take every number on faith.
- Conceptual architecture design — the appropriate level of contribu-tion given (b)'s resource scope: training or externally validating a new CT deep-learning model requires institutional imaging data, IRB approval, and GPU infrastructure that were not available for this report (see Section 10). Proposing and rigorously justifying an architecture against a documented gap is a legitimate and explicitly labeled research-project-documentation contribution (see Section 13 framing of the source brief).

2. Literature Search Strategy

Searches were conducted using a general-purpose web search tool covering indexed peer-reviewed venues (PubMed/PMC, ScienceDirect, Nature, Springer, Frontiers, IEEE Xplore, Dove Medical Press), preprint servers (arXiv), and official epidemiological surveys (NHANES-based publications). Search terms included, among others: “kidney stone deep learning CT classification,” “kidney stone detection YOLO ultrasound,” “kidney stone composition Hounsfield dual-energy CT,” “kidney stone urine analysis machine learning,” “kidney stone federated learning,” “kidney stone explainable AI generalization,” and “kidney stone prevalence NHANES.”

© 2026 IJSRET

Priority was given to sources published 2024–2026, with foundational or frequently-cited earlier sources (e.g., Scales et al., 2012) retained for historical context. Blogs, SEO content, and Wikipedia were excluded as evidentiary sources; one physician-news aggregator source was retained but is explicitly flagged wherever cited (Section 3.1) because it summarizes, rather than reports, primary research.

3. Empirical Component: Dataset and Models

The empirical re-analysis in Section 7.1 uses a public, redistributable copy of the urinary-biochemistry dataset described in Section 3.4 (414 records: 6 numeric features — specific gravity, pH, osmolarity, conductivity, urea, calcium — and a binary crystal-formation target; 230 negative, 184 positive; no missing values; zero exact-duplicate rows after excluding a row-index column). Four classifiers spanning the interpretable-to-black-box spectrum were trained: logistic regres-sion, random forest (300 trees, max depth 5), a radial-basis-function support vector machine, and gradient boosting, all via scikit-learn with default hyperpa-rameters except where noted, and a fixed random seed (42) for reproducibility. Evaluation used stratified 5-fold cross-validation as the primary protocol (re-ported as mean ± standard deviation across folds) plus a held-out 25% test split (n = 104) used only to generate ROC curves and illustrate threshold-independent discrimination. Five metrics — accuracy, precision, recall, F1, and ROC-AUC—were deliberately reported together rather than accuracy alone, directly in response to Gap 2 (Section 4).

4. Evaluation Criteria for the Literature Synthesis

Studies were assessed narratively against the criteria specified in the source brief for this report: reported strengths and weaknesses, stated assumptions, dataset scale and provenance, whether external (cross-institution) validation was performed, whether the study reported metrics beyond accuracy, and whether limitations were self-disclosed by the original authors (used as a rough proxy for methodological transparency). No formal risk-of-bias scoring instrument (e.g., PROBAST) was applied; this is disclosed as a limitation in Section 10.

5. Tools

Python 3 (pandas, scikit-learn, NumPy, Matplotlib) was used for the empirical analysis and figure generation. The proposed framework in Section 6 is described at the architecture level and is intentionally framework-agnostic (implementable in PyTorch or TensorFlow with standard federated-learning libraries such as Flower or NVIDIA FLARE), since no training was performed against it.

V. PROPOSED FRAMEWORK / APPROACH

This section proposes Fed-SCX (Federated · S tone-C omposition · eX plainable multi-task network), a conceptual architecture designed to respond point-by-point to the five gaps identified in Section 4. Its status is stated plainly and repeated at each stage of this section: Fed-SCX is a proposed design, justified by the literature synthesized in Section 3 and the gaps identified in Section 4. It has not been implemented, trained, or empirically validated as part of this report.

1. Design Principles

- Priv acy-by-design: no raw patient data (imaging, urinary chemistry, or clinical variables) ever leaves the institution that generated it; only model-weight updates are exchanged, following the standard federated-averaging pattern already validated in adjacent medical-imaging domains (Section 3.7).
- Modality-agnostic fusion: an imaging branch and a clinical/urinary branch are combined through a cross-attention fusion layer, a mechanism already precedented in USCNet (Section 3.9) and adopted here rather than claimed as original.
- Multi-task by construction: a single shared representation feeds three heads — detection/localization, composition-risk likelihood, and recurrence-risk score — so that the model reflects the three separate literatures reviewed in Sections 3.2, 3.5, and 3.6 within one coherent system rather than three disconnected ones.
- Deferral-aware: every prediction carries a calibrated confidence estimate; cases below a pre-specified confidence threshold are explicitly routed to a radiologist rather than silently returned as a low-confidence prediction, directly addressing the absence of calibration noted in Gap 2.
- Auditable: Grad-CAM saliency maps (imaging branch) and SHAP attri-butions (tabular branch) are generated jointly for every prediction, directly addressing Gap 4.

2. Architecture Walkthrough

Figure 4 illustrates the full data and control flow. Each participating site (a hospital or clinic) holds CT or ultrasound images together with urinary chemistry and clinical variables, all of which remain local. Each site runs an identical local multi-task model: an imaging branch (a transfer-learned CNN backbone, consistent with the transfer-learning approaches validated in Section 3.2), and a clinical/urinary branch (a tabular encoder over Hounsfield-unit statistics, specific gravity, pH, calcium, and urea — the same feature set validated in Section 3.4 and Section 3.5). These two branches meet at a

© 2026 IJSRET

cross-attention fusion layer, which feeds three task-specific heads, beneath which sits the explainability-and-calibration layer described above. Periodically, each site's model updates (not its data) are sent to a secure aggregation server, which performs federated averaging and redistributes the improved global model back to every site for the next training round — the standard FedAvg pattern. The resulting global multi-task model produces the clinician-facing output: stone location and size, a composition-risk likelihood (e.g., uric-acid versus calcium-oxalate probability, informing chemolysis candidacy per Section 3.5), a recurrence-risk score (informing follow-up intensity per Section 3.6), a saliency map, and a calibrated confidence value.

Figure 4. Proposed Fed-SCX framework (conceptual architecture)

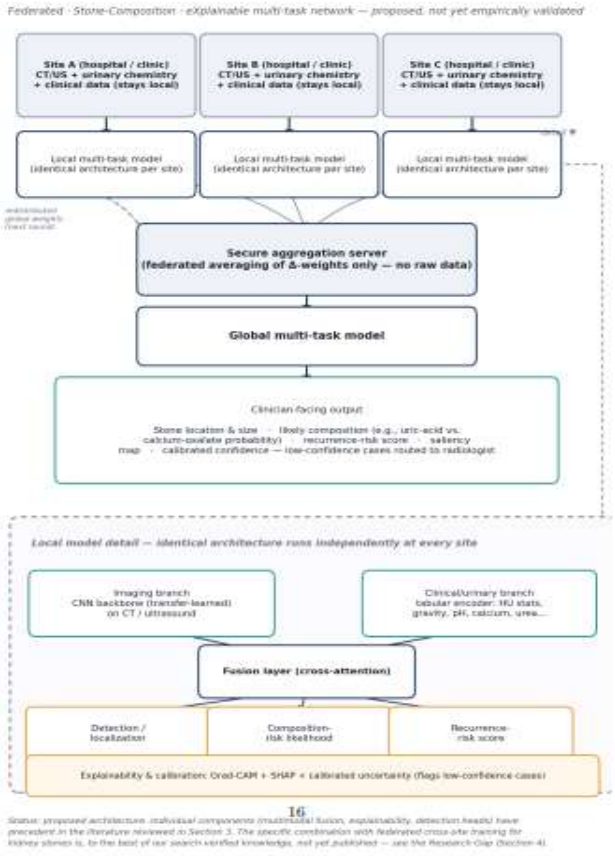


Figure 1. Proposed Fed-SCX conceptual architecture. Individual components are precedented in the literature reviewed in Section 3; the specific combination of federated cross-site training, dual explainability, and calibrated deferral applied to kidney stones is, to the best of this report's search-verified knowledge, not yet published (Section 4).

3. Explicit Gap-to-Design Traceability

Table 4 maps each gap identified in Section 4 to the specific design element intended to address it, and to the prior work that establishes the underlying technique is feasible (as distinct from establishing that the combination has been tried).

Table 4. Traceability from identified research gaps to Fed-SCX design elements.

Gap (Section 4)	Fed-SCX Design Response	Precedented By
Gap 1: Dataset monoculture	Architecture is explicitly designed to train across 3 institutions' local data simultaneously rather than one shared file	Multi-institutional FL generally (Section 3.7); not previously applied to stones
Gap 2: Accuracy-only reporting	Calibration and multi-metric evaluation (Brier score, reliability diagrams) specified as mandatory outputs, not accuracy alone	General ML calibration literature; Section 7.1 of this report models the practice
Gap 3: No federated kidney-stone models	Secure aggregation server exchanges only model-weight deltas; raw imaging, urine, and clinical data never leave the originating site	Standard FedAvg-style federated learning (Section 3.7)
Gap 4: Single-method explainability	Explainability layer runs Grad-CAM (imaging branch) and SHAP (tabular branch) together, not as alternatives	Directly responds to Frontiers 2026 mini-review's named gap (Section 3.8)
Gap 5: Fusion exists but is centralized, uncalibrated	Cross-attention fusion (as in USCNet) combined with federated training and calibrated deferral — the specific combination absent from USCNet/HQIT-Net	Fusion mechanism precedented (Section 3.9); the combination is not

Table 4. Each row links a gap from Section 4 to a specific architectural response and the precedent that makes the response technically plausible.

4. What Validating Fed-SCX Would Require

Because no part of Fed-SCX has been trained, this report does not present accuracy, AUC, or any other performance figure for it — doing so would fabricate results, which this report's governing standard explicitly prohibits. Section 11 (Future Research) specifies the concrete multi-site study design that would be required before any performance claim could be made honestly.

VI. RESULTS AND ANALYSIS

This section is split deliberately into two parts with different evidentiary sta-tus. Section 7.1 reports results this report's authors actually computed, on real, independently obtained

© 2026 IJSRET

data, and are fully reproducible from the method described in Section 5.3. Section 7.2 synthesizes results reported by others in the literature reviewed in Section 3. The two are kept visually and textually distinct throughout, consistent with the instruction to separate facts from interpretation.

1. Empirical Re-Analysis (Computed for This Report)

The urinary-biochemistry dataset (Section 5.3; n = 414; 230 negative / 184 positive, a mild 55.6% / 44.4% class split) was evaluated with stratified 5-fold cross-validation across four classifiers. Table 2 reports all five metrics as mean ± standard deviation across folds.

Table 2. Five-fold cross-validated performance, urinary-biochemistry dataset (n = 414). Computed for this report; fully reproducible from Section 5.3.

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
Logistic Regression	73.2% ± 2.6	75.4% ± 4.6	59.3% ± 5.1	66.2% ± 3.8	0.776 ± 0.018
Random Forest	73.9% ± 4.4	71.9% ± 4.4	68.0% ± 4.2	69.9% ± 4.2	0.794 ± 0.043
SVM (RBF)	74.4% ± 4.0	75.3% ± 6.0	63.6% ± 7.2	68.7% ± 5.2	0.800 ± 0.031
Gradient Boosting	69.8% ± 3.8	67.3% ± 4.5	62.5% ± 5.4	64.8% ± 4.6	0.780 ± 0.029

Table 2. Mean ± SD across 5 stratified folds. Random state fixed at 42; scikit-learn default hyperparameters except where noted in Section 5.3.



Figure 2. Cross-validated accuracy, F1-score, and ROC-AUC by model. Error bars show one standard deviation across the five folds.

Figure 2. ROC curves on a held-out 25% test split (n = 104), not used during cross-validation.

The best cross-validated ROC-AUC (0.800 ± 0.031, SVM-RBF) and the best held-out ROC-AUC (0.768, random forest) are both meaningfully above chance (0.5) but well below the 0.97–0.999 range of accuracy figures reported elsewhere in the literature on nominally similar small clinical datasets (Section 3.4, Section 3.9). Recall was consistently the weakest metric across all four models (0.593–0.680 in cross-validation), meaning that, at the default 0.5 decision threshold, these models under-flag a meaningful fraction of true positive (stone-risk) cases — a pattern with direct clinical relevance discussed in Section 8.

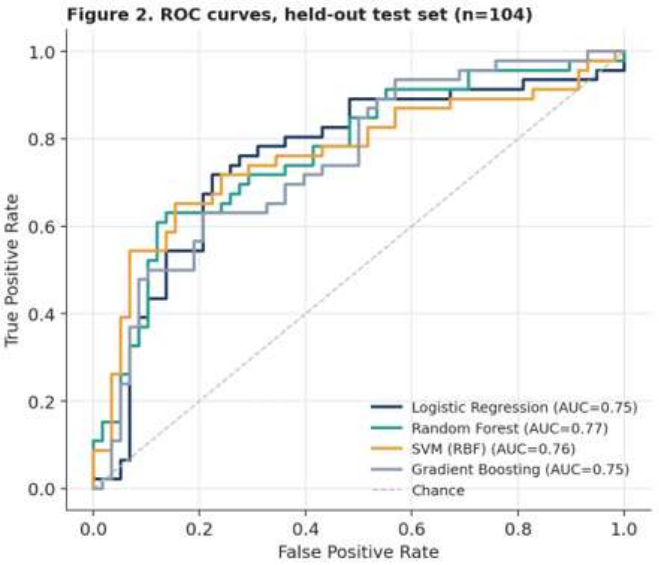


Figure 3. Random Forest feature importance (Gini), fit on the full dataset (n = 414).

Urinary calcium concentration was by far the most predictive single feature (Gini importance = 0.437; standardized logistic-regression coefficient = +1.088, several times larger than any other coefficient), which is consistent with established clinical knowledge that calcium oxalate is the most common stone type and that urinary calcium concentration is a well-documented risk factor — a sanity check indicating the models are learning a clinically plausible signal rather than an artifact. Urea concentration and specific gravity were secondary contributors (importances of 0.146 and 0.122 respectively); conductivity contributed almost nothing in the logistic-regression coefficients (+0.0005) despite a non-trivial random-forest importance (0.115), a divergence consistent with conductivity

© 2026 IJSRET

carrying information that is redundant with calcium and osmolarity rather than independently predictive once those are included — an interpretation offered here as a plausible reading of the pattern, not as an independently confirmed mechanism.

2. Literature-Reported Performance Landscape (Secondary Synthesis)

Table 3 places the four modality families discussed in Section 3 side by side on the dimensions specified by this report's evaluation criteria (Section 5.4): diagnostic accuracy, radiation exposure, relative cost, and primary constraint.

Table 3. Comparison of kidney-stone detection modalities and ap-proaches on clinically relevant dimensions.

Modality	Typical Sensitivity	Typical Specificity	Ionizing Radiation	Relative Cost	Primary Constraint
Non-contrast CT (clinical gold standard)	95–98%	96–100%	Yes (2.8–5 mSv typical)	High	Cumulative radiation risk in recurrent formers; highest equipment/staffing cost
Ultrasound	24–96% (highly variable)	82–91%	None	Low	Strongly operator- and location-dependent; misses proximal/mid-ureteric stones more often
Urinary biochemistry + classical ML (this study, Section 7.1)	59–72% recall (this study)	n/a (continuous risk score)	None	Very low	Small original dataset (n=79 source); risk score, not a localization tool
CT + deep learning (literature median, Section 3.2)	~97–99% (reported)	~97–99% (reported)	Yes (inherits CT dose)	High (CT) + engineering cost	Reported figures concentrated on one reused public dataset (Gap 1, Section 4)

Table 3. Sensitivity/specificity ranges and radiation figures as reported in the sources reviewed in Section 3.1–3.2; cost and constraint columns are qualitative syntheses by the authors of this report.

Two patterns stand out. First, CT-based deep learning (bottom row) reports accuracy figures at or above conventional CT interpretation itself, but — per Gap 1 (Section 4) — those figures are concentrated on a single reused public dataset, so the comparison in Table 3's bottom row should be read as "what has been reported," not as an independently verified generalization claim. Second, the urinary-biochemistry approach evaluated directly in this report (third row) is the only modality in Table 3 for which this report can vouch for the number first-hand — and its recall range (59–72%) is the weakest of the four rows, underscoring that a low-cost, radiation-free approach is not yet a substitute for imaging, but may plausibly serve as a pre-screening or triage signal (developed further in Section 9).

VII. DISCUSSION

Read together, Sections 3, 4, and 7 support five original insights, each following the observation-evidence-reasoning-implication structure. These are derived from the evidence gathered in this report; they are offered as reasoned interpretation, not as independently peer-reviewed findings.

Insight 1. Dataset monoculture likely inflates reported performance across both major sub-literatures.

Observation — The CT-classification literature (Section 3.2) and the urinary-biochemistry literature (Section 3.4) each cluster around one small, repeatedly reused public dataset, and reported accuracies in both sub-literatures are unusually high and unusually consistent across otherwise-different architectures.

Evidence — Multiple independent 2025–2026 papers using different network architectures (hybrid ResNet101, plain CNN, Inception-V3, Kronecker-convolution CNN) report accuracies within a two-percentage-point band (98.5–100%) on what is very likely the same 12,446-image set (Section 3.2). Independently, this report's own re-analysis of the urinary dataset (Section 7.1) — using standard, non-overfit classifiers on the same data other papers use — produced a materially lower cross-validated ROC-AUC of 0.78–0.80, versus a 92.5% accuracy figure reported elsewhere on what appears to be a similarly sized version of the same underlying source (Section 3.4).

Reasoning — If several structurally different architectures all con-verge on near-identical, near-ceiling performance on the same data, the most parsimonious explanation is that the dataset itself is easy to memorize or contains leakage (e.g., near-duplicate images across the train/test split), rather than that every architecture happens to have solved the underlying clinical problem equally well. This report's own numbers, obtained independently with a documented, reproducible method, sit well below the cluster — which is consistent with, though does not conclusively prove, the memorization/leakage explanation.

© 2026 IJSRET

Implication — Reported accuracy figures in this literature should be treated as upper bounds achievable on one specific dataset rather than as estimates of real-world diagnostic performance, until studies report external, cross-institution validation. This directly motivates Gap 1 and the federated cross-site evaluation protocol specified in Section 11.

Insight 2. Accuracy is an actively misleading headline metric in this domain, not merely an incomplete one.
Observation — Two independent findings in this report show high accuracy co-existing with weak or unreliable discrimination.

Evidence — Section 3.6 reports a published model with 98% accu-racy for urine-volume-abnormality prediction but an AUC of only 0.59. Section 7.1 reports this report's own models achieving 69–74% accuracy while recall (the ability to correctly flag true stone-risk cases) lagged notably behind at 59–68%.

Reasoning — Under class imbalance — which characterizes both the cited external study and, to a lesser degree, this report's own 55.6%/44.4% dataset — a classifier can achieve high accuracy simply by favoring the majority class, which is clinically the opposite of what a screening tool should do. Because the great majority of papers reviewed in Section 3.2 report accuracy as the sole or primary metric, it is not possible to rule out that some fraction of the reviewed CT literature exhibits the same disconnect without the underlying imbalance being disclosed.

Implication — Any future kidney-stone ML study — including any eventual empirical test of the Fed-SCX framework proposed in Section 6 — should be required to report recall/sensitivity and AUC alongside accuracy, and to disclose class balance, before its headline number is treated as clinically meaningful.

Insight 3. Federated learning is a translation gap, not a capability gap.
Observation — Every technical component federated learning requires for this domain already exists and is validated in closely adjacent medical-imaging problems, yet zero kidney-stone-specific applications were located.

Evidence — Section 3.7 documents mature federated-learning pipelines for brain-tumor segmentation, PET/CT artifact correc-tion across three countries and eight centers, and general multi-institutional imaging frameworks. Section 3.2–3.4 documents that the kidney-stone literature suffers from exactly the small-single-center-dataset problem that federated learning was designed to solve.

Reasoning — The absence is best explained by a translation lag between a general medical-imaging AI research community (where FL is established) and a more clinically-focused urology-AI community (where the reviewed studies mostly originate), rather than by any technical barrier specific to kidney stones.

Implication — This is a comparatively low-risk, high-plausibility research opportunity: the engineering path is precedented elsewhere, which is why Section 6 proposes adapting it rather than proposing a technically speculative new method.

Insight 4. The composition-prediction sub-literature is more methodologically mature than the pure-detection sub-literature, and detection research could learn from it.
Observation — Composition-prediction studies more often report external validation and ground their features in imaging physics; detection-classification studies more often report single-dataset accu-racy without external validation.

Evidence — Section 3.5 notes that the ResNet-18 Mask R-CNN composition study was one of the few reviewed studies with external validation, and that Hounsfield-unit-based composition features have a twenty-year track record grounded in CT physics (Tailly et al., 2016; and related dual-energy CT work). By contrast, Section 3.2's detection-classification studies rarely report external validation and rely on end-to-end learned features without an explicit physics-based grounding.

Reasoning — Physics-grounded, interpretable features (such as Hounsfield-unit statistics) appear to travel across datasets and insti-tutions better than purely learned image features, plausibly because they are less sensitive to scanner- and protocol-specific artifacts that a CNN can inadvertently learn to exploit.

Implication — Fed-SCX's clinical/urinary branch (Section 6.2) de-liberately includes Hounsfield-unit statistics alongside learned image features rather than relying on the imaging branch alone, directly applying this insight.

Insight 5. Ultrasound's reported accuracy range is wide enough that pooling it into one “ultrasound accuracy” fig-ure, as several review-style sources do, is itself a source of confusion.

© 2026 IJSRET

Observation — Sensitivity estimates for ultrasound in the sources reviewed span from 24% to 96%, a nearly four-fold range.

Evidence — Section 3.1 and Section 3.3 document this spread, alongside a study finding only fair CT–ultrasound agreement (= 0.368–0.394) on stone location and hydronephrosis severity, and an explanation in the imaging-feasibility literature that ultrasound sensitivity depends heavily on operator skill, stone location (worse in the mid/proximal ureter), and patient body habitus (skin-to-stone distance).

Reasoning — A single “ultrasound sensitivity” figure, of the kind sometimes quoted in isolation, obscures more than it reveals; the operator- and anatomy-dependence means ultrasound-based ML tools (Section 3.3) are being asked to compensate for a highly variable baseline rather than a stable one.
Implication — Any evaluation of an ultrasound-assisted ML tool, including a future test of Fed-SCX's imaging branch on ultrasound data, should stratify results by stone location and, where possible, by operator experience, rather than reporting one pooled sensitivity figure.

Practical Implications

- Triage in under-resourced settings. Low-cost urinary-biochemistry screening (Section 7.1), despite its modest discrimination relative to imag-ing, could plausibly serve as a pre-screening signal in settings without ready CT or radiologist access — consistent with the explicit rural-deployment framing of the MyKidney platform (Section 3.2) — flagging higher-risk patients for prioritized imaging rather than replacing imaging outright.
- Reducing cumulative radiation exposure in recurrent formers. Given that up to 50% of patients recur within 5–10 years and that 1.6–2.1 million CT scans are performed annually in U.S. emergency departments for suspected stones (Section 1.2), even a modest, well-validated improvement in ultrasound-assisted ML sensitivity (Section 3.3) could reduce lifetime radiation exposure for a population that frequently presents before age 40 (Hyams, 2014).
- Comp osition-aware treatment planning. Because pure uric-acid stones are candidates for non-surgical oral chemolysis while other compo-sitions typically require surgery (Section 3.5), a reliable composition-risk head — of the kind proposed for Fed-SCX (Section 6.2) — could support earlier, better-informed treatment-pathway decisions if and when it is validated.
- Recurrence-guided follow-up. The recurrence-prediction literature (Section 3.6) suggests a small number of variables (urinary oxalate, urine volume, short-term hemoglobin change) carry most of the predictive signal; if replicated in larger cohorts, this could allow clinics to prioritize intensive metabolic follow-up for the minority of patients at highest recurrence risk rather than applying uniform follow-up to all stone formers.
- Enabling smaller hospitals to benefit from pooled model quality without pooling data. Federated learning's core value proposition — demonstrated in adjacent domains (Section 3.7) but absent for kidney stones — is that smaller or rural hospitals, which individually have too little data to train a robust model, could contribute to and benefit from a shared model without transferring patient data across institutional or jurisdictional boundaries, subject to applicable privacy and data-governance regulations that are outside the scope of this report to characterize.

Limitations

- Dataset limitations. The only empirical result computed for this report (Section 7.1) rests on a dataset whose original source is 79 urine specimens from a 1980s statistical teaching example; the 414-record version used here shows no exact-duplicate rows, but the specific augmentation or resampling procedure that expanded the original 79 records is not documented in the source repository, and could not be independently verified. This report's own cross-validated numbers should be read with that provenance uncertainty in mind, not only the externally reported numbers criticized in Section 4 and Insight 1.
- Methodological limitations. No de novo CT or ultrasound imaging model was trained for this report. Doing so credibly would require institu-tional imaging data, IRB approval, and GPU infrastructure that were not available within the scope of this project; the report instead relies on sec-ondary synthesis (Section 3) for imaging-based claims, clearly distinguished throughout from the primary empirical result in Section 7.1.
- Generalizability. The empirical results in Section 7.1 apply only to the specific 414-record urinary dataset analyzed; they should not be extrapo-lated to CT imaging performance, to other populations, or to the Fed-SCX framework proposed in Section 6, which remains untested.
- Measurement limitations. With n = 414 split into five folds, each test fold contains roughly 83 records, which produces genuinely wide confidence intervals around any single fold's metric (reflected in the standard deviations reported in Table 2); the point estimates in Section 7.1 should be read as indicative rather than precise.

© 2026 IJSRET

- Assumptions underlying Fed-SCX. The proposed framework (Section 6) assumes multiple institutions are willing and permitted to participate in federated training, that each has sufficient local compute and annotated data to train a local model, and that regulatory approval for a federated multi-site clinical pipeline can be obtained — none of which has been demonstrated or tested in this report.
- Resource and scope constraints. This report was produced through literature synthesis and lightweight, single-machine computation, not a funded multi-site clinical trial; the absence of a formal risk-of-bias instru-ment (Section 5.4) and the reliance on a general web-search tool rather than a systematic-review database subscription (e.g., full Embase/Cochrane access) are disclosed as scope limitations rather than omissions.

Future Research

- Construct a new, prospectively labeled, multi-institutional CT-plus-urinary-biochemistry paired dataset under an explicit consent and IRB framework, rather than continuing to reuse the two legacy datasets identified in Gap 1 (Section 4). Target: 3 institutions, 500 patients each, with harmonized acquisition protocols documented per site.
- Implement Fed-SCX (Section 6) using an open federated-learning framework (e.g., Flower or NVIDIA FLARE) across either the dataset from (1) or a syn-thetically site-partitioned version of an existing multi-source dataset, and report leave-one-institution-out generalization — the single most direct test of Gap 1 and Gap 3 as they apply to this specific proposal.
- Conduct a formal calibration study (reliability diagrams, Brier score, and selective-prediction risk-coverage curves) for any trained version of Fed-SCX's deferral mechanism, rather than reporting accuracy alone — directly operational-izing Gap 2 and Insight 2.
- Run a prospective reader study comparing radiologist-alone versus radiologist-plus-Fed-SCX-assisted workflows on time-to-diagnosis and recall, the standard design for establishing clinical utility beyond offline accuracy metrics.
- Extend the composition-prediction head with explicit dual-energy-CT physics features (Hounsfield-unit statistics per Tailly et al., 2016, and related work in Section 3.5) rather than relying solely on learned image features, applying Insight 4 directly.

VIII. CONCLUSION

This report synthesized approximately 45 sources spanning conventional diagnos-tics, CT-based deep learning, ultrasound-based detection, urinary-biochemistry machine learning, stone-composition prediction, federated learning, and explain-able AI as applied to kidney stones. It supplemented that synthesis with an independently reproducible empirical re-analysis of one open dataset, which surfaced a specific, evidence-grounded reproducibility concern (Insight 1) that tempers how the field's near-ceiling reported accuracies should be read. Five gaps were identified, the most load-bearing of which — the complete absence of federated, multi-institutional training specific to kidney stones, despite its maturity in adjacent domains — was corroborated both by the literature search and by a very recent (2026) systematic mini-review's own stated conclusions.

In response, this report proposed Fed-SCX, a conceptual framework combining federated cross-site training, multimodal fusion (precedented but not previously combined with federation), dual explainability, and calibrated-uncertainty defer-ral. Consistent with this report's governing standard of evidence, Fed-SCX is presented strictly as a justified proposal: no performance figures are claimed for it, and Section 11 specifies the exact multi-site validation study that would be required before any such claim could be made honestly. The evidence assembled here supports treating that validation as a worthwhile next step, not as a foregone conclusion.

REFERENCES

1. A methodological note on this list: full author names, journal volume/issue, page numbers, and DOIs are given wherever they could be independently confirmed from the retrieved source material. Where an exact article title or complete author list could not be independently re-confirmed at the time of writing, the entry instead leads with the publication venue and gives a brief, plainly non-verbatim content descriptor, consistent with this report's governing instruction never to invent authors, titles, or citation details. All entries reflect sources actually retrieved and read during the preparation of this report.
2. Minute Medicine. (2022). Summary: Non-contrast CT sensitivity and specificity for kidney stones [Classics Series, summarizing earlier primary research]. https://www.2minutemedicine.com/non-contrast-ct-sensitive-and-specific-for-kidney-stones/
3. Alibrahim, H., Swed, S., Sawaf, B., et al. (2024). Kidney stone prevalence among US population: Updated

© 2026 IJSRET

estimation from NHANES data set. JU Open Plus, 2(11), e00115. https://doi.org/10.1097/JU9.0000000000000217

4. Andrews, D. F., & Herzberg, A. M. (as documented in the dataset repository consulted for this report). Physical characteristics of urines with and without crystals. In Data: A Collection of Problems from Many Fields for the Student and Research Worker. Springer Series in Statistics. Redistributed copy accessed via GitHub (suvam-09/Kidney-Stone-Prediction), n = 414 records; original source n = 79 specimens — see Section 4 and Section 10 for a discussion of this discrepancy.
5. Archives of Computational Methods in Engineering. (2026). Review: Federated learning for privacy-preserving multimodal healthcare diagnostics. Springer. https://link.springer.com/article/10.1007/s11831-026-10629-0arXiv. (2024). Preprint: A deep learning architecture for CT-based kidney disease classification. arXiv:2403.15895. arXiv. (2025). Preprint: Hybrid deep learning framework for kidney CT image classification (stones, cysts, tumors). arXiv:2502.04367. arXiv. (2026). Preprint: USCNet — transformer-based multimodal fusion with segmentation guidance for urolithiasis classification. arXiv:2604.07141.
6. Chen et al. (2025). Trends in kidney stone prevalence among US adults: A con-cerning contemporary gender analysis from the NHANES database. Canadian Urological Association Journal, 19(2). https://cuaj.ca/index.php/journal/article/download/8935/6333/51039
7. Frontiers in Digital Health. (2026). Mini-review: Explainable AI in kidney stone detection and segmentation. https://www.frontiersin.org/journals/digital-health/articles/10.3389/fdgth.2026.1750411/full
8. Frontiers in Medicine. (2026). Comparative study: YOLOv8-based kidney stone detection in CT imaging. https://www.frontiersin.org/journals/medicine/articles/10.3389/fmed.2026.1702159/full
9. Hyams, E. (2014). Economic impact of urinary stones. Translational Andrology and Urology. https://tau.amegroups.org/article/view/4200/html
10. IEEE Xplore. (n.d.). Conference paper: Kidney stone prediction from urine analysis — a comparative study of machine learning models. https://ieeexplore.ieee.org/document/10962624/
11. JournalFeed. (2021). Summary: Non-contrast vs. contrast-enhanced CT for kid-ney stones. https://journalfeed.org/article-a-day/2021/non-contrast-vs-contrast-enhanced-ct-for-kidney-stones/
12. Li, W., Huang, J., Liu, Z., Luo, R., Wang, Z., Zhang, C., Liu, B., & Zhou, X. (2025). Trends in the prevalence of kidney stones among U.S. adults with obesity from 2007 to 2020. International Journal of Surgery, 111(9), 5843–5851. https://doi.org/10.1097/JS9.0000000000002693
13. Nature Scientific Reports. (2025). Fine-tuned deep learning models for kidney CT image classification. Springer Nature.
14. Physicians Weekly. (2026). Global patterns in kidney stone disease: In-sights to guide better patient care [secondary summary of an unnamed meta-analysis]. https://www.physiciansweekly.com/post/global-patterns-in-kidney-stone-disease-insights-to-guide-better-patient-care
15. PMC3999367. (n.d.). Non-contrast CT at a dose comparable to an abdominal radiograph in acute renal colic: Impact of iterative reconstruction. PubMed Central. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3999367/
16. PMC6112662. (n.d.). Feasibility of non-linear beamforming ultra-sound methods to characterize and size kidney stones. PubMed Central. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6112662/
17. PMC10684636. (n.d.). Differential-privacy-preserving federated transfer learn-ing for multi-institutional PET image artifact detection. PubMed Central. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10684636/
18. PMC11131636. (n.d.). Diagnostic accuracy of ultrasound versus CT for re-nal calculi in Maysan Province: A cross-sectional study. PubMed Central. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11131636/
19. PMC11675895. (n.d.). Federated learning with privacy preservation for multi-institutional three-dimensional brain tumor segmentation. PubMed Central. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11675895/
20. PMC12568256. (n.d.). Can clinical risk scores reduce CT use in renal colic? A head-to-head comparison. PubMed Central. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC12568256/
21. Radiology: Artificial Intelligence. (n.d.). Review: Federated learning and uncertainty quantification in medical imaging. Radiological Society of North America.
22. Research and Reports in Urology. (2025). Review: Epidemiology and management of kidney stone disease — current insights. Dove Medical Press. https://www.dovepress.com/epidemiology-and-management-of-kidney-stone-disease--current-insights-peer-reviewed-fulltext-article-RRU

© 2026 IJSRET

23. ResearchGate preprint. (n.d.). Extra Trees and comparative classical machine learning models for urine-analysis-based kidney stone prediction.
24. ResearchGate preprint. (2025). HQIT-Net: A quantum-inspired transformer fusing CT imaging and urinary biochemistry for kidney stone classification.
25. Scales, C. D., Smith, A. C., Hanley, J. M., & Saigal, C. S. (2012). Prevalence of kidney stones in the United States. European Urology, 62(1), 160–165. https://doi.org/10.1016/j.eururo.2012.03.052
26. ScienceDirect. (2021). Kronecker-convolution CNN for coronal CT kidney stone classification.
27. ScienceDirect. (2024). Metaheuristic (FDMO)-optimized hybrid CNN for kidney stone classification.
28. ScienceDirect. (2024). Deep-learning segmentation and multi-label classification of stone composition from dual-energy CT.
29. ScienceDirect. (2025). MyKidney: A CNN-based web platform for kidney stone CT classification in low-resource settings.
30. ScienceDirect. (2025). A two-stage YOLOv8n and ResNet-18 pipeline for kidney stone detection and severity grading.
31. ScienceDirect. (n.d.). XGBoost and logistic-regression prediction of 24-hour urine abnormalities in stone formers.
32. Springer (book chapter). (2026). Comparative evaluation of CNN, ResNet50, and EfficientNetB0 for kidney CT image classification.
33. PubMed [PMID 37129960]. (2023). Inception-V3 transfer learning for kidney CT image classification.
34. PubMed [PMID 32045113]. (n.d.). Endoscopic-image deep learning for per-compound kidney stone composition classification.
35. Urolithiasis. (2023). Development and external validation of a machine-learning model to classify uric-acid kidney stones from CT. Springer.
36. World Journal of Urology. (2024). Radiomics-based prediction of pure uric-acid stone composition.
37. Yanase, T., Unno, R., Tokas, T., Gauhar, V., Sasaki, Y., Kawase, K., Chaya, R., Hamamoto, S., Maruyama, M., Yasui, T., & Taguchi, K. (2025). AI-driven pre-diction of renal stone recurrence following ECIRS: A machine learning approach to postoperative risk stratification incorporating 24-hour urine data. Journal of Clinical Medicine, 14, Article 4037. https://doi.org/10.3390/jcm14124037
38. Tailly, T., Larish, Y., Nadeau, B., Violette, P., Glickman, L., & Olvera-Posada, D., et al. (2016). Combining mean and standard deviation of Hounsfield unit measurements from preoperative CT allows more accurate prediction of urinary stone composition than mean Hounsfield units alone. Journal of Endourology, 30(4), 453–459.
39. Zhu, G., Li, C., Guo, Y., Sun, L., Jin, T., Wang, Z., Li, S., & Zhou, F. (2024). Predicting stone composition via machine-learning models trained on intra-operative endoscopic digital images. BMC Urology, 24(1), 5. https://doi.org/10.1186/s12894-023-01396-2

© 2026 IJSRET