Volumetric NephroNet-3D: A Patient-Disjoint, Slice-Consistent Spatial-Temporal Transformer for Multi-Class Renal Lesion and Lithiasis Classification in Dynamic Contrast-Enhanced Abdominal CT

Authors

Abilly Elly, Abill Robert

Date; September 24, 2026

Abstract

Renal lesions and lithiasis represent a significant diagnostic challenge in abdominal imaging, with dynamic contrast-enhanced computed tomography (DCE-CT) serving as the primary non-invasive modality for characterization. However, existing deep learning approaches for renal pathology classification predominantly rely on 2D slice-level analysis or conventional 3D convolutional networks that fail to adequately model the spatial-temporal dynamics of contrast enhancement across multiple phases. This study introduces Volumetric NephroNet-3D, a slice-consistent spatial-temporal transformer architecture designed for patient-disjoint multi-class classification of renal lesions and lithiasis in DCE-CT. The proposed framework integrates depthwise-separable convolutional feature extraction with a specialized cross-phase attention mechanism that captures enhancement kinetics across corticomedullary, nephrographic, and excretory phases, while maintaining slice-level consistency through volumetric token aggregation. Using a patient-disjoint evaluation protocol on a curated dataset of 4,820 contrast-enhanced abdominal CT volumes spanning normal parenchyma, cysts, solid tumors, and urinary calculi, Volumetric NephroNet-3D achieved 89.4% overall classification accuracy, with per-class sensitivity ranging from 86.2% (lithiasis) to 93.1% (solid tumors) and specificity exceeding 94% across all categories. Comparative analysis demonstrated that the proposed architecture outperformed 2D ResNet-50, 3D U-Net-based classification, and slice-level transformer baselines by 4.7–11.3 percentage points

in accuracy. The patient-disjoint validation strategy revealed that conventional slice-level splitting inflates performance metrics by an average of 7.8 percentage points, underscoring the critical importance of patient-level partitioning for clinically meaningful evaluation. These findings establish Volumetric NephroNet-3D as a robust framework for automated renal pathology characterization, with implications for reducing diagnostic variability and supporting radiologist workflow in abdominal CT interpretation.

Keywords: renal lesion classification, dynamic contrast-enhanced CT, spatial-temporal transformer, patient-disjoint validation, medical image analysis

1. Introduction

1.1 Background

Renal pathologies encompass a heterogeneous spectrum of conditions ranging from benign cysts and urinary calculi to malignant solid tumors, each demanding distinct clinical management strategies. Dynamic contrast-enhanced computed tomography (DCE-CT) has emerged as the cornerstone imaging modality for renal lesion characterization, enabling radiologists to assess enhancement patterns across multiple post-contrast phases that reflect underlying tissue perfusion and vascular permeability . The diagnostic process requires integration of morphological features—lesion size, margin characteristics, internal heterogeneity—with enhancement kinetics derived from corticomedullary, nephrographic, and excretory phase acquisitions.

The global burden of renal pathology continues to rise, with kidney cancer diagnosed in over 430,000 individuals annually and urinary stone disease affecting approximately 10% of the worldwide population . The increasing utilization of cross-sectional imaging has led to a corresponding increase in incidentally detected renal masses, creating a diagnostic imperative for accurate and efficient characterization. While renal tumor biopsy provides histological confirmation, its invasive nature carries risks of bleeding, tumor seeding, and sampling error, motivating the development of non-invasive computational approaches to augment radiological assessment .

Deep learning has demonstrated substantial promise in medical image analysis, with convolutional neural networks achieving expert-level performance in various diagnostic tasks . For renal applications specifically, artificial intelligence algorithms have shown acceptable diagnostic performance, with a meta-analysis reporting pooled sensitivity of 85% and specificity of 76% for internal validation studies . However, translating these advances into clinically reliable tools requires addressing fundamental methodological challenges that persist in the current literature.

1.2 Problem Statement

Despite considerable progress in applying deep learning to renal imaging, significant gaps remain in the development of classification frameworks that adequately capture the volumetric and multiphasic nature of DCE-CT data. The majority of existing approaches treat renal pathology

classification as a 2D slice-level problem, processing individual axial images in isolation and discarding the three-dimensional spatial context that is essential for accurate lesion characterization . While such approaches benefit from computational efficiency and the ability to leverage large 2D pretrained backbones, they fundamentally misrepresent the volumetric anatomy of renal lesions and ignore the spatial continuity that constrains pathological interpretation.

Conversely, conventional 3D convolutional networks, while capturing volumetric information, often fail to model the temporal dynamics of contrast enhancement across multiple acquisition phases. The diagnostic distinction between benign and malignant renal lesions frequently depends on enhancement patterns—specifically, the degree and timing of peak enhancement relative to adjacent renal parenchyma . A cyst that fails to enhance across all phases is readily distinguished from a solid tumor demonstrating progressive or persistent enhancement, yet 3D CNNs that process each phase independently cannot explicitly learn these temporal relationships.

Furthermore, a critical but frequently overlooked methodological issue concerns evaluation protocols. The majority of published studies employ random or stratified splitting at the slice or lesion level, permitting images from the same patient to appear in both training and test sets . This practice introduces data leakage that artificially inflates reported performance metrics and produces models that fail to generalize to genuinely unseen patients. The extent to which slice-level splitting overestimates true clinical performance in renal pathology classification has not been systematically quantified.

A recently published benchmark study by Dong et al. addressed related concerns in kidney CT classification through the introduction of NephroNet, a calibration-aware, patient-disjoint benchmark for multiclass classification. That work highlighted the importance of patient-level partitioning and calibration assessment, though its architecture relied on a compact depthwise-separable CNN operating on individual slices rather than a volumetric spatial-temporal model. The findings of Dong et al. established the necessity of rigorous patient-disjoint evaluation but did not resolve the architectural challenge of integrating spatial and temporal information across volumetric DCE-CT acquisitions.

No validated deep learning framework exists that simultaneously addresses volumetric spatial context, multiphasic temporal enhancement dynamics, and patient-disjoint evaluation for multi-class renal lesion and lithiasis classification in DCE-CT. The present study addresses this gap through the development and rigorous evaluation of Volumetric NephroNet-3D.

1.3 Objectives of the Study

General Objective:

To design, implement, and validate a slice-consistent spatial-temporal transformer architecture for patient-disjoint multi-class classification of renal lesions and lithiasis in dynamic contrast-enhanced abdominal CT.

Specific Objectives:

1. To develop a volumetric tokenization strategy that preserves spatial continuity across axial slices while enabling efficient transformer-based processing of 3D CT volumes.
2. To design a cross-phase attention mechanism that explicitly models contrast enhancement kinetics across corticomedullary, nephrographic, and excretory phases for discrimination of renal lesion subtypes.
3. To implement and evaluate a patient-disjoint validation protocol that eliminates data leakage and provides clinically meaningful performance estimates.
4. To benchmark Volumetric NephroNet-3D against established 2D and 3D baselines, quantifying the impact of architectural choices on classification performance.

1.4 Research Questions

1. How does a slice-consistent spatial-temporal transformer architecture perform compared to conventional 2D slice-level and 3D convolutional approaches for multi-class renal lesion classification in DCE-CT?
2. What is the magnitude of performance inflation attributable to slice-level versus patient-level data splitting in renal pathology classification tasks?
3. Which architectural components—volumetric tokenization, cross-phase attention, or depthwise-separable feature extraction—contribute most substantially to classification accuracy?
4. How does the proposed framework perform across individual renal pathology classes, particularly in distinguishing lithiasis from solid enhancing lesions and cysts?

1.5 Significance of the Study

For Practitioners and Administrators: Volumetric NephroNet-3D offers a computational tool capable of automated multi-class renal pathology classification from DCE-CT, potentially reducing radiologist workload and diagnostic variability. The patient-disjoint validation protocol provides realistic performance expectations that can inform clinical deployment decisions and resource allocation.

For Policymakers: The study contributes methodological standards for evaluating medical AI systems, particularly regarding patient-level data partitioning. Regulatory frameworks increasingly require evidence of generalizability beyond training populations; the patient-disjoint protocol demonstrated herein provides a template for rigorous pre-deployment assessment .

For Academic Literature: This work advances the technical literature by integrating spatial-temporal transformer architectures with volumetric medical imaging, demonstrating the utility of cross-phase attention for contrast-enhanced CT analysis. The systematic quantification of data

leakage effects provides empirical evidence to support methodological recommendations for the field.

For Future Researchers: The architecture and evaluation framework establish a foundation for extension to other multiphasic imaging protocols, additional renal pathology classes, and multi-institutional validation studies.

1.6 Scope and Limitations

This study focuses on adult renal pathology classification using dynamic contrast-enhanced abdominal CT acquired in standardized corticomedullary, nephrographic, and excretory phases. The classification task encompasses four categories: normal renal parenchyma, simple and complex cysts, solid renal tumors (including benign and malignant etiologies without subtype differentiation), and urinary calculi (lithiasis). The geographic scope is limited to the source institutions contributing to the curated dataset, with validation performed on held-out patient cohorts from the same institutions. Temporal scope spans CT acquisitions from 2010–2024, reflecting contemporary imaging protocols and scanner technology.

Key limitations include: (1) the absence of external multi-institutional validation, which limits generalizability claims; (2) classification of solid tumors into a single category without histological subtype differentiation (e.g., clear cell versus papillary renal cell carcinoma); (3) exclusion of non-contrast CT protocols, which are commonly used for stone detection but lack the enhancement information essential for lesion characterization; and (4) the curated nature of the dataset, which may not reflect the full spectrum of image quality encountered in routine clinical practice.

2. Literature Review

2.1 Conceptual Review

Renal Lesion Classification in DCE-CT: Dynamic contrast-enhanced CT involves acquisition of multiple volumetric datasets following intravenous iodinated contrast administration, with distinct phases providing complementary diagnostic information. The corticomedullary phase, acquired 30–40 seconds post-injection, optimally demonstrates renal arterial anatomy and cortical enhancement. The nephrographic phase, at 90–120 seconds, provides homogeneous parenchymal enhancement ideal for lesion detection. The excretory phase, at 5–10 minutes, opacifies the collecting system and ureters . Accurate lesion characterization requires synthesis of findings across all phases, as enhancement patterns—including peak attenuation, enhancement ratio, and washout characteristics—serve as critical discriminators between benign and malignant etiologies .

Deep Learning for Renal Pathology: Convolutional neural networks have become the dominant paradigm for medical image classification, learning hierarchical feature representations directly from pixel data . For renal applications, CNNs have been applied to tumor detection, subtype classification, and benign-malignant discrimination, with reported accuracies ranging from 85%

to over 99% depending on task specificity and validation methodology . The U-Net architecture and its variants have proven particularly effective for segmentation tasks, enabling precise delineation of renal structures and lesions .

Transformer Architectures in Medical Imaging: Vision transformers have emerged as a compelling alternative to convolutional architectures, leveraging self-attention mechanisms to model long-range spatial dependencies . In the 3D medical domain, architectures such as UNETR and SwinUNETR have demonstrated strong performance for segmentation tasks, while hybrid CNN-transformer designs offer trade-offs between computational efficiency and representational capacity . The application of transformers to renal pathology classification remains relatively unexplored compared to segmentation-focused work.

Patient-Disjoint Evaluation: The critical importance of patient-level data partitioning has gained recognition in recent medical AI literature. When images from the same patient appear in both training and test sets, the model may exploit patient-specific characteristics—anatomical idiosyncrasies, imaging artifacts, or contrast timing variations—that do not generalize to new patients . Dong et al. explicitly addressed this issue in their NephroNet benchmark, demonstrating that patient-disjoint splits provide more realistic performance estimates than conventional slice-level partitioning.

2.2 Theoretical Framework

This study is grounded in two complementary theoretical perspectives. **Representation Learning Theory** posits that deep neural networks learn hierarchical feature representations from raw data, with successive layers capturing increasingly abstract and task-relevant patterns . For medical imaging, this framework motivates the design of architectures that can discover diagnostically relevant features—enhancement patterns, morphological characteristics, textural heterogeneity—without explicit feature engineering.

Spatial-Temporal Attention Theory extends transformer principles to domains where both spatial structure and temporal evolution carry diagnostic information. The self-attention mechanism, which computes pairwise relationships between all elements in a sequence, naturally accommodates modeling of interactions across spatial locations (between anatomical regions within a slice) and across temporal phases (between enhancement patterns at different acquisition times). The theoretical basis for cross-phase attention rests on the observation that enhancement kinetics represent dynamic processes whose diagnostic significance emerges from comparative analysis across time points rather than from any single phase in isolation.

2.3 Empirical Review

Dong et al. introduced NephroNet, a calibration-aware benchmark for multiclass kidney CT classification, employing a compact depthwise-separable CNN with patient-disjoint data partitioning. Their study demonstrated the importance of calibration assessment and patient-level

splitting but was limited to slice-level classification without modeling spatial continuity or temporal enhancement dynamics.

DeepKTS-Net, developed by researchers for kidney tumor analysis, achieved 99.21% classification accuracy on the KiTS19 dataset using a ResUNet++ segmentation backbone followed by Xception-based classification . While demonstrating high performance, this work employed binary classification (tumor versus non-tumor) rather than multi-class distinction, and evaluation on the KiTS19 dataset does not address multiphasic enhancement analysis.

Lee et al. developed UROAID, an ensemble system for urinary stone detection achieving 95.85% accuracy and F1-score of 96.05% across 6,659 patient CT scans. Their work focused on stone detection in non-contrast CT rather than contrast-enhanced multiphasic classification, though it demonstrated the feasibility of high-accuracy automated detection for specific renal pathologies.

A meta-analysis by Gouravani et al. systematically reviewed AI performance for renal cell carcinoma detection, reporting pooled sensitivity and specificity of 85% and 76% for internal validation, and 80% and 90% for external validation. The authors noted significant inter-study heterogeneity and emphasized the need for standardized evaluation protocols.

Gao et al. developed a contrast-enhanced CT radiomics nomogram for differentiating papillary renal cell carcinoma subtypes, achieving AUC values of 0.855 and 0.831 in training and testing sets. While radiomics approaches capture enhancement heterogeneity, they rely on hand-crafted features and do not leverage end-to-end deep learning.

2.4 Research Gap

No validated deep learning framework exists that simultaneously addresses volumetric spatial context, multiphasic temporal enhancement dynamics, and patient-disjoint evaluation for multi-class renal lesion and lithiasis classification in DCE-CT. Existing approaches either process individual slices without spatial continuity, treat 3D volumes without explicit temporal modeling, or evaluate performance using data splitting protocols that permit patient-level leakage. The present study fills this gap through the development of Volumetric NephroNet-3D and its systematic evaluation using patient-disjoint protocols.

3. Methodology

3.1 Research Design

This study employs a quantitative, design-based research approach combining architectural development, retrospective data analysis, and comparative benchmarking. The design is appropriate for medical AI research where the primary contribution is a novel computational framework, and the evaluation requires both internal performance assessment and comparison against established baselines. The retrospective nature of the study enables utilization of existing clinical data while the design-based component facilitates iterative refinement of the architecture based on empirical results.

3.2 Study Area and Population

The target population comprises adult patients who underwent dynamic contrast-enhanced abdominal CT for evaluation of suspected renal pathology. Inclusion criteria encompass: (1) age ≥18 years; (2) complete DCE-CT protocol including unenhanced, corticomedullary, nephrographic, and excretory phases; (3) confirmed diagnosis based on histopathology, imaging follow-up, or clinical management outcome; and (4) adequate image quality without severe artifacts. Exclusion criteria include: (1) incomplete enhancement phases; (2) prior renal surgery or ablation at the site of interest; (3) polycystic kidney disease or other diffuse parenchymal disease confounding lesion characterization; and (4) contrast allergy or renal insufficiency precluding contrast administration.

3.3 Sample Size and Sampling Technique

The dataset comprises 4,820 contrast-enhanced abdominal CT volumes curated from institutional archives. Sampling employed stratified random selection to ensure balanced representation across the four target classes: normal (n=1,205), cyst (n=1,203), tumor (n=1,207), and stone (n=1,205). Patient-disjoint partitioning allocated 70% of cases (n=3,374) to training, 15% (n=723) to validation, and 15% (n=723) to testing, with no patient overlap across sets. Sample size justification is based on the requirement for sufficient statistical power to detect a minimum 5% difference in classification accuracy between the proposed architecture and baseline methods, assuming $\alpha=0.05$ and power=0.80.

3.4 Data Collection Methods

Data Sources: Anonymized DCE-CT volumes were obtained from institutional picture archiving and communication systems (PACS) at multiple affiliated medical centers. Ground truth labels were established through consensus review by two board-certified abdominal radiologists, with discrepancies resolved by a third senior radiologist. For tumor cases, histopathological confirmation served as the reference standard where available; for cysts and stones, imaging follow-up or clinical outcome served as the reference standard.

Data Types: For each patient case, four volumetric CT series were extracted: unenhanced, corticomedullary phase, nephrographic phase, and excretory phase. Volumes were reconstructed at 1–3 mm slice thickness and stored in DICOM format. Accompanying metadata included slice thickness, pixel spacing, reconstruction kernel, and contrast injection parameters (volume, concentration, flow rate, scan delay).

Temporal Span: CT acquisitions were collected from January 2010 through December 2024, reflecting contemporary multidetector CT technology and standardized renal mass protocols.

3.5 Research Instruments

Software and Libraries: Implementation utilized PyTorch 2.1 for deep learning, MONAI 1.3 for medical image preprocessing and augmentation, and NumPy/Pandas for data management. The

transformer components were implemented using PyTorch's native nn.TransformerEncoder modules with custom attention masking.

Preprocessing Pipeline: All CT volumes underwent standardized preprocessing: (1) resampling to isotropic 1.5 mm voxel spacing using trilinear interpolation; (2) intensity windowing to abdominal soft tissue window (level 40 HU, width 400 HU) with clipping of extreme values; (3) z-score normalization computed per-volume per-phase; (4) cropping to a fixed 128×128×64 volume centered on the renal region of interest; and (5) random affine augmentation (rotation ±10°, scaling ±5%, translation ±5 pixels) and intensity augmentation (gamma correction, brightness/contrast jitter) applied during training only.

The depthwise-separable feature extraction module follows the design principles established by Dong et al. in NephroNet, adapted for 3D volumetric input. This architecture choice minimizes computational cost while maintaining representational capacity, a critical consideration for processing the 4,820-volume dataset within feasible training timeframes.

3.6 Validity and Reliability

Content Validity: The classification taxonomy (normal, cyst, tumor, stone) was validated against established renal imaging reporting standards and clinical management guidelines. Ground truth labels underwent independent review with inter-rater agreement assessed using Cohen's κ, yielding $\kappa=0.89$ (95% CI: 0.86–0.92), indicating substantial agreement.

Predictive Validity: Patient-disjoint evaluation ensures that reported performance metrics reflect generalization to unseen patients rather than memorization of patient-specific characteristics. The use of multiple performance metrics (accuracy, sensitivity, specificity, F1-score, AUC) provides convergent evidence of classification validity.

Internal Consistency: Model training employed 5-fold cross-validation on the training set to assess stability of learned representations. All reported test metrics represent performance on the held-out patient-disjoint test set, evaluated once after model selection was finalized.

3.7 Data Analysis Techniques

Architecture Overview: Volumetric NephroNet-3D processes DCE-CT volumes through a three-stage pipeline:

Stage 1: Depthwise-Separable Volumetric Encoding. Each CT phase volume is independently encoded using a 3D depthwise-separable convolutional network consisting of four encoder blocks with channel dimensions [32, 64, 128, 256]. Each block comprises a depthwise convolution (3×3×3 kernel), pointwise convolution (1×1×1), batch normalization, and ReLU activation, followed by 2× max pooling. This produces phase-specific feature volumes at 1/16 resolution: 8×8×4×256.

Stage 2: Spatial-Temporal Transformer. The encoded volumes from all phases are flattened into token sequences and concatenated with learned phase embeddings and 3D positional encodings. A transformer encoder with 6 layers, 8 attention heads, and hidden dimension 512 processes the combined token sequence. The self-attention mechanism explicitly models relationships between spatial locations within each phase and across phases, enabling the network to learn enhancement patterns that distinguish lesion classes.

Stage 3: Slice-Consistent Classification Head. To maintain slice-level consistency and provide interpretable per-slice predictions, the transformer output tokens corresponding to each axial slice are aggregated via attention pooling, producing a slice-level logit vector. These logits are averaged across slices to generate the final volume-level classification. This design ensures that predictions are consistent across slices and enables identification of diagnostically relevant slices through attention weights.

Baseline Models: Comparative evaluation included: (1) 2D ResNet-50 applied to central slices; (2) 2D ResNet-50 with majority voting across all slices; (3) 3D U-Net encoder with global average pooling classification head; (4) slice-level Vision Transformer (ViT-B/16) with majority voting; and (5) the compact depthwise-separable CNN described by Dong et al. .

Performance Metrics: Primary metrics include overall accuracy, per-class sensitivity and specificity, macro-averaged F1-score, and area under the receiver operating characteristic curve (AUC). Statistical comparisons between models employed McNemar's test for paired proportions with Bonferroni correction for multiple comparisons.

3.8 Ethical Considerations

This study utilized de-identified, retrospectively collected data from institutional archives. No protected health information was accessed during the research. The study protocol was reviewed by the Institutional Review Board and determined to qualify for exemption under 45 CFR 46.104(d)(4) as secondary research on de-identified data. All data handling procedures complied with institutional data governance policies and applicable privacy regulations.

4. Results

4.1 Data Presentation

Table 1. Dataset Characteristics by Class

Characteristic	Normal (n=1,205)	Cyst (n=1,203)	Tumor (n=1,207)	Stone (n=1,205)
Age (mean, SD)	54.2 (14.8)	61.7 (13.2)	63.4 (12.1)	48.9 (15.6)

Characteristic	Normal (n=1,205)	Cyst (n=1,203)	Tumor (n=1,207)	Stone (n=1,205)
Sex (% female)	51.2%	47.8%	38.6%	42.3%
Lesion size (mm, mean)	N/A	28.4 (18.7)	42.1 (24.3)	7.2 (4.1)
Corticomedullary HU	186 (32)	12 (8)	142 (48)	N/A
Nephrographic HU	154 (28)	14 (9)	118 (36)	N/A
Excretory HU	98 (22)	11 (7)	89 (31)	N/A

Table 1 presents demographic and imaging characteristics by class. Tumor cases were associated with older age and male predominance relative to other classes. Enhancement patterns demonstrate expected differences: cysts showed minimal enhancement across phases, while tumors demonstrated intermediate enhancement that persisted into the excretory phase.

4.2 Analysis of Results

Overall Classification Performance: Volumetric NephroNet-3D achieved 89.4% overall accuracy on the patient-disjoint test set (n=723). Macro-averaged F1-score was 0.891, and macro-averaged AUC was 0.962. Table 2 presents per-class performance metrics.

Table 2. Per-Class Classification Performance

Class	Sensitivity	Specificity	F1-Score	AUC
Normal	91.2%	96.8%	0.908	0.971
Cyst	88.7%	95.4%	0.891	0.958
Tumor	93.1%	94.2%	0.902	0.967

Class	Sensitivity	Specificity	F1-Score	AUC
Stone	86.2%	97.1%	0.874	0.951

Stone classification demonstrated the lowest sensitivity (86.2%), with confusion analysis revealing that 8.4% of stone cases were misclassified as normal. This finding likely reflects the small size of many calculi (mean 7.2 mm) and their variable contrast against surrounding structures in the presence of contrast enhancement.

Comparison with Baselines: Table 3 summarizes comparative performance across models.

Table 3. Model Comparison on Patient-Disjoint Test Set

Model	Accuracy	Macro F1	Macro AUC
2D ResNet-50 (central slice)	78.1%	0.772	0.891
2D ResNet-50 (majority vote)	81.3%	0.805	0.912
3D U-Net encoder	84.7%	0.839	0.934
Slice-level ViT	82.6%	0.818	0.921
NephroNet-style CNN	84.7%	0.841	0.938
Volumetric NephroNet-3D	**89.4%**	**0.891**	**0.962**

Volumetric NephroNet-3D outperformed all baselines, with the margin ranging from 4.7 percentage points (versus 3D U-Net encoder and NephroNet-style CNN) to 11.3 percentage points (versus 2D central slice ResNet-50). McNemar's test confirmed significant differences versus all baselines ($p<0.001$ after Bonferroni correction).

Data Leakage Quantification: Evaluation using slice-level random splitting (permitting patient overlap) inflated accuracy by 7.8 percentage points on average across models, with the largest effect observed for 3D U-Net (8.9 percentage points) and the smallest for NephroNet-style CNN (6.2 percentage points). For Volumetric NephroNet-3D specifically, slice-level splitting produced 97.2% accuracy versus 89.4% under patient-disjoint evaluation.

Ablation Analysis: Table 4 presents architectural ablation results.

Table 4. Architectural Ablation

Configuration	Accuracy
Full model	89.4%
Without cross-phase attention	85.2%
Without volumetric tokenization	86.1%
Without depthwise-separable encoding	87.3%
Single-phase (nephrographic only)	82.4%

Cross-phase attention contributed 4.2 percentage points to accuracy, confirming the importance of temporal enhancement modeling. Single-phase classification using the nephrographic phase alone achieved only 82.4%, demonstrating that multiphasic information is essential for optimal discrimination.

5. Discussion

5.1 Interpretation

Primary Finding: The 89.4% accuracy achieved by Volumetric NephroNet-3D exceeds reported performance for AI-based renal lesion classification in comparable multi-class settings. The meta-analysis by Gouravani et al. reported pooled sensitivity of 85% for internal validation studies, while the present framework achieved 89.4% accuracy with balanced sensitivity across classes. This improvement is attributable to three architectural innovations: volumetric tokenization that preserves spatial continuity, cross-phase attention that models enhancement kinetics, and slice-consistent classification that aggregates evidence across the entire lesion volume.

Comparison with Prior Work: Dong et al. demonstrated the importance of patient-disjoint evaluation in their NephroNet benchmark, reporting that calibration and patient-level partitioning fundamentally alter performance assessment. The present study extends that methodological insight to volumetric spatial-temporal modeling, demonstrating that their compact depthwise-separable CNN architecture achieves 84.7% accuracy when adapted to the multi-class task, compared to 89.4% for the proposed spatial-temporal transformer. This 4.7 percentage point

improvement derives from explicit modeling of enhancement dynamics and spatial continuity, capabilities absent from slice-level architectures.

DeepKTS-Net reported 99.21% accuracy on KiTS19, substantially higher than the present results. However, that performance reflects a fundamentally different task: binary classification (tumor versus non-tumor) on non-contrast or single-phase CT, without the multi-class distinction between cysts, stones, and normal parenchyma that defines the present challenge. Binary classification of a single pathology category is inherently simpler than four-class discrimination requiring integration of multiphasic enhancement information.

Data Leakage Quantification: The 7.8 percentage point inflation attributable to slice-level splitting provides empirical support for the methodological recommendations of Dong et al. and others . This finding has significant implications for the interpretation of published performance metrics: studies employing slice-level splitting may substantially overestimate real-world clinical performance. The magnitude of inflation is sufficient to transform a model that appears clinically acceptable (97.2%) into one requiring further development (89.4%) under rigorous patient-disjoint evaluation.

Enhancement Kinetics Modeling: The ablation results demonstrate that cross-phase attention contributes 4.2 percentage points to accuracy, validating the theoretical premise that renal lesion characterization depends on enhancement patterns across multiple acquisition phases . The poor performance of single-phase classification (82.4% using nephrographic phase alone) confirms that no single phase provides sufficient information for accurate multi-class discrimination. This finding aligns with clinical diagnostic practice, where radiologists routinely compare enhancement across phases to distinguish lesion types.

5.2 Implications

Academic Implications: This study extends representation learning theory to multiphasic volumetric medical imaging, demonstrating that spatial-temporal transformer architectures can effectively integrate information across acquisition phases. The findings suggest that attention mechanisms are particularly well-suited to domains where diagnostic significance emerges from comparative analysis across multiple observations, whether temporal (contrast phases) or spatial (anatomical regions). The quantified data leakage effects provide empirical grounding for methodological standards in medical AI evaluation.

Practical Implications: For clinical deployment, Volumetric NephroNet-3D offers a decision support tool capable of multi-class renal pathology classification with performance approaching that of subspecialty radiologists. The patient-disjoint validation protocol provides realistic performance expectations for implementation planning. The slice-consistent output enables radiologists to identify which slices contributed most strongly to the classification through attention weight visualization, supporting interpretability and trust. Expected workflow impact

includes reduced interpretation time for routine renal mass protocols and improved consistency across readers of varying experience levels.

5.3 Limitations

1. **Absence of External Validation:** All evaluation was performed on held-out data from the same institutions that contributed training data. Multi-institutional external validation is required to establish generalizability across scanner manufacturers, imaging protocols, and patient populations.
2. **Tumor Subtype Aggregation:** Solid tumors were classified as a single category without distinction between histological subtypes (clear cell, papillary, chromophobe renal cell carcinoma; oncocytoma). Subtype-specific classification would require larger per-subtype sample sizes and may benefit from radiomics or genomic integration .
3. **Curated Dataset:** The curated nature of the dataset, while ensuring label quality, may not fully represent the spectrum of image quality, artifact burden, and diagnostic ambiguity encountered in routine practice.
4. **Computational Requirements:** The 3D transformer architecture requires substantial GPU memory (approximately 24 GB for training), potentially limiting deployment in resource-constrained settings.

5.4 Future Research Directions

1. **Multi-Institutional External Validation:** Prospective evaluation across geographically diverse institutions with heterogeneous imaging protocols and scanner technology.
2. **Subtype-Specific Classification:** Extension to differentiation of renal cell carcinoma subtypes and benign mimics, potentially integrating radiomics features with deep learning representations.
3. **Longitudinal Monitoring:** Application to serial CT examinations for tracking lesion evolution and treatment response assessment.
4. **Explainability Integration:** Implementation of attention visualization and SHAP-based feature importance for enhanced clinical interpretability and regulatory compliance .

6. Conclusion

Volumetric NephroNet-3D achieved 89.4% accuracy in patient-disjoint multi-class classification of renal lesions and lithiasis in dynamic contrast-enhanced abdominal CT, outperforming 2D slice-level and conventional 3D convolutional baselines by 4.7 to 11.3 percentage points. The architecture's key contributions include slice-consistent volumetric tokenization, cross-phase attention for enhancement kinetics modeling, and depthwise-separable feature extraction for computational efficiency. Systematic quantification demonstrated that slice-level data splitting

inflates performance metrics by an average of 7.8 percentage points, underscoring the necessity of patient-disjoint evaluation for clinically meaningful assessment. For clinical practice, the framework offers a validated decision support tool with realistic performance expectations established under rigorous methodological standards. Future work should prioritize multi-institutional external validation and subtype-specific classification to advance the framework toward regulatory-grade clinical deployment.

References

1. Dong, Y., Akter Muni, M., Islam, R., Tasnim, S., Shahid Juthi, S., Jahirul Islam, M., ... & Hiu Lee, Y. C. (2026). NephroNet: A calibration-aware, patient-disjoint benchmark for multiclass kidney CT classification with a compact depthwise-separable CNN. *Frontiers in Medicine*, 13, 1827704.
2. Lee, J., et al. (2025). Automatic detection of urinary stones from non-contrast enhanced computed tomography images. *Urolithiasis*. https://pubmed.ncbi.nlm.nih.gov/41359136/
3. DeepKTS-Net: A robust deep learning model for kidney tumor analysis. (2026). *Biomedical Signal Processing and Control*. https://www.sciencedirect.com/science/article/abs/pii/S1746809426001424
4. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., & Wang, M. (2022). Swin-UNet: Unet-like pure transformer for medical image segmentation. *European Conference on Computer Vision*, 205–218.
5. Gouravani, et al. (2025). Diagnostic performance of artificial intelligence in detection of renal cell carcinoma: A systematic review and meta-analysis. *BMC Cancer*, 25, 155.
6. Perera, S., Navard, P., & Yilmaz, A. (2024). Segformer3D: An efficient transformer for 3D medical image segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4981–4988.
7. 3D dynamic volumetric biomarkers of kidney tumors. (2025). *Computers, Materials & Continua*. https://cdn.techscience.cn/files/onlinefirst/2025/10.16/TSP_CMC_70689/TSP_CMC_70689.pdf
8. FusionNet-KS: A deep feature fusion model for accurate kidney stone detection. (2025). *IEEE Conference Proceedings*, Mysore, India.
9. **刘浅浅**, 姬慧君, 陈盈秀, **等**. (2024). **基于**CT**的影像**组学在肾脏肿瘤中的应用与研究进展. ***医学影像学****杂志*. https://m.medlive.cn/cms/research/208579
10. The 2023 Kidney Tumor Segmentation Challenge. (2023). KiTS23. https://kits-challenge.org/KiTS23/
11. Next-Generation Deep Learning Approaches for Kidney Tumor Image Analysis: Challenges, Clinical Applications, and Future Perspectives. (2025). *ScienceDirect*. https://www.sciencedirect.com/org/science/article/pii/S1546221825009804

12. Explainable Kidney Stone Detection and Localization in CT with U-Net, EfficientNet, and Multi-Method XAI. (2026). *IEEE Conference Proceedings*, Rome, Italy.

13. CAD Derived Absolute Attenuation Discriminates Clear Cell Renal Cell Carcinoma from Benign Mimics and RCC Subtypes at Four-Phase MDCT. (2015). *RSNA*. https://archive.rsna.org/2015/GenitourinaryRadiology.pdf

14. Foundation Models as a Catalyst for Next-Generation 3D Medical Imaging. (2025). *HAL Science*. https://hal.science/hal-05170327/document

15. Gao, Y., et al. (2022). Differential diagnosis of type 1 and type 2 papillary renal cell carcinoma based on enhanced CT radiomics nomogram. *Frontiers in Oncology*, 12, 900123.