Domain-Invariant and Calibration-Preserving Deep Learning for Multi-Center Kidney CT Classification: Evaluating NephroNet Architecture Across Heterogeneous Scanners and Contrast Phase Dynamics

Authors

Abilly Elly, Abill Robert

Date; September 24, 2026

Abstract

Computed tomography (CT) remains the primary imaging modality for renal pathology, yet deep learning models for kidney CT classification face persistent barriers to clinical translation, including slice-level data leakage, poor probability calibration, and limited generalizability across heterogeneous scanner platforms. This study presents a comprehensive evaluation of NephroNet, a compact depthwise-separable convolutional neural network with squeeze-and-excitation channel attention and spatial gating, within a rigorously controlled patient-disjoint benchmark for four-class kidney CT classification (Normal, Cyst, Tumor, Stone). Using a 12,446-image multi-center cohort from Dhaka, Bangladesh, we implemented group-stratified hold-out validation to prevent patient-series leakage, combined with a standardized training pipeline incorporating annealed MixUp/CutMix augmentation, class-weighted AdamW optimization, exponential moving average evaluation, test-time augmentation, and post-hoc temperature scaling ($T^* = 1.42$). NephroNet achieved an accuracy of 0.9997 (95% CI: 0.9984–1.0000), macro-AUC of 0.9969 (95% CI: 0.9953–0.9983), Brier score of 0.0007, and expected calibration error of 0.0021, surpassing 30 capacity-matched CNN and Vision Transformer baselines. These findings demonstrate that parameter-efficient architectural design combined with calibration-aware evaluation protocols can achieve near-ceiling discrimination while maintaining probability reliability, establishing a reproducible benchmark for domain-invariant renal imaging AI.

Keywords: kidney CT classification, deep learning, patient-disjoint evaluation, probability calibration, domain generalization

1. Introduction

1.1 Background

Renal pathologies, including cystic lesions, solid tumors, and urinary calculi, affect millions of individuals worldwide and impose substantial diagnostic and economic burdens on healthcare systems [1]. Computed tomography has emerged as the diagnostic modality of choice for renal imaging due to its high spatial resolution, rapid acquisition, and ability to characterize tissue attenuation across contrast phases [2]. The increasing volume of CT examinations has driven interest in automated deep learning systems capable of assisting radiologists in lesion detection and classification, with the potential to reduce interpretation time and improve diagnostic consistency across heterogeneous clinical settings [3].

Deep learning approaches to medical image classification have evolved substantially over the past decade, progressing from hand-crafted feature extraction to end-to-end convolutional neural networks (CNNs) and, more recently, Vision Transformers [4]. Within renal imaging specifically, numerous studies have reported high classification accuracy on publicly available datasets, particularly the Islam CT Kidney Dataset, which comprises 12,446 images across four diagnostic categories collected from picture archiving and communication systems (PACS) across multiple hospitals in Dhaka, Bangladesh [5]. However, the methodological rigor underlying these reported performance metrics has increasingly come under scrutiny, revealing systematic vulnerabilities that undermine the credibility of benchmark comparisons and limit clinical translation [6].

1.2 Problem Statement

Two persistent barriers undermine the validity of deep learning evaluations for kidney CT classification. First, patient-series leakage occurs when images from the same patient appear in both training and validation partitions, a particularly acute problem in CT datasets where each patient may contribute dozens of axial and coronal slices across multiple contrast phases [5][7]. When random image-level splitting is employed, the model effectively memorizes patient-specific anatomical features rather than learning generalizable pathological signatures, producing inflated accuracy estimates that collapse under external validation [8]. Despite widespread recognition of this issue, most published works on the Islam dataset have not implemented patient-disjoint splitting, rendering direct numerical comparisons methodologically inadmissible [5].

Second, probability calibration—the alignment between predicted confidence scores and actual correctness rates—has received insufficient attention in medical imaging deep learning [9]. Discrimination metrics such as accuracy and area under the receiver operating characteristic curve (AUC) evaluate ranking ability but provide no information about whether a model's predicted

probabilities are trustworthy at clinically meaningful decision thresholds [10]. This gap is particularly consequential in nephrology, where predicted probabilities may gate follow-up imaging, biopsy decisions, or surgical referral. A model with excellent AUC but poor calibration can systematically overestimate or underestimate malignancy risk, leading to inappropriate clinical actions [11]. Prior work on kidney CT classification has largely neglected calibration assessment, reporting only discrimination metrics and leaving the reliability of model outputs at deployment thresholds uncharacterized [5][12].

A third, less examined dimension of the problem concerns domain invariance across heterogeneous scanners and contrast protocols. The Islam dataset spans axial and coronal planes and both contrast-enhanced and non-contrast acquisitions, yet the extent to which model performance is preserved across these acquisition strata remains unknown [5]. Scanner vendor differences, reconstruction kernels, slice thickness, and contrast timing introduce distribution shifts that can degrade classification performance when models are deployed outside their training distribution [13]. Whether compact, parameter-efficient architectures can maintain calibration and discrimination across such heterogeneity without explicit domain adaptation mechanisms is an open question with direct implications for multi-center deployment.

1.3 Objectives of the Study

General objective: To rigorously evaluate the performance, calibration, and domain-invariance characteristics of the NephroNet architecture for four-class kidney CT classification under patient-disjoint, multi-center conditions.

Specific objectives:

1. To establish a patient-disjoint, group-stratified benchmark for kidney CT classification that eliminates slice-series leakage and enables methodologically admissible performance comparisons.
2. To evaluate whether the NephroNet architecture, combining depthwise-separable convolutions with squeeze-and-excitation channel attention and spatial gating, can achieve high discrimination accuracy at a tightly controlled parameter budget (≈1.46 M) relative to capacity-matched CNN and Vision Transformer baselines.
3. To quantify probability calibration quality before and after post-hoc temperature scaling, characterizing the reliability of model outputs at clinically relevant decision thresholds.
4. To assess the robustness of classification performance across contrast-phase and anatomical-plane strata within the multi-center cohort.

1.4 Research Questions

1. Does patient-disjoint splitting substantially degrade classification accuracy relative to previously reported random-split results on the Islam CT Kidney Dataset, and what does this reveal about the true generalizability of prior benchmarks?
2. How does NephroNet's parameter-efficient architecture compare to capacity-matched CNN and Vision Transformer baselines in terms of both discrimination (accuracy, macro-AUC) and calibration (Brier score, expected calibration error) on a patient-disjoint hold-out set?
3. To what extent does post-hoc temperature scaling improve probability calibration without degrading discrimination performance, and what temperature parameter is optimal for this task?
4. Do classification performance and calibration quality vary systematically across contrast-enhanced versus non-contrast acquisitions and across axial versus coronal planes?

1.5 Significance of the Study

For practitioners and radiologists: This study provides a rigorously validated benchmark that can serve as a reference point for evaluating future kidney CT classification models. The calibration-aware reporting protocol, including Brier score and expected calibration error with bootstrap confidence intervals, offers a template for assessing whether model outputs are trustworthy at the probability thresholds that inform clinical decisions [5][10].

For policymakers and health system administrators: The findings demonstrate that compact, parameter-efficient architectures can achieve near-ceiling accuracy on multi-center renal CT data, suggesting that deployment on resource-constrained hardware is feasible without sacrificing diagnostic performance. The emphasis on calibration addresses a key barrier to clinical adoption: the need for model outputs that clinicians can interpret as valid probability estimates [11].

For academic literature: This study contributes a methodological framework for patient-disjoint evaluation and calibration-aware benchmarking that addresses documented weaknesses in the existing kidney CT deep learning literature. The capacity-controlled comparison across 30 baselines establishes a reproducible standard for architectural comparison that isolates model design from training recipe variation [5][7].

For future researchers: The limitations identified in this study, particularly the absence of external validation across scanner vendors and geographic regions, define a clear research agenda for multi-institutional prospective studies and domain adaptation methods [5][13].

1.6 Scope and Limitations

This study analyzes a single, publicly available, de-identified dataset: the Islam CT Kidney Dataset (12,446 images; Normal 5,077, Cyst 3,709, Tumor 2,283, Stone 1,377), retrospectively compiled from PACS across multiple hospitals in Dhaka, Bangladesh [5]. The dataset spans axial and

coronal planes and contrast-enhanced and non-contrast acquisitions. This study does not include prospective data collection, external multi-institutional validation, or volumetric/multi-phase analysis. Key limitations include: (i) all data originate from a single geographic region, limiting generalizability across scanner vendors and patient demographics; (ii) the patient grouping heuristic relies on the Islam dataset filename convention, which may not be exhaustive; (iii) temperature scaling is applied post-hoc and may not be optimal across all operating thresholds; and (iv) challenging sub-classifications (e.g., fat-poor angiomyolipoma, early-stage transitional cell carcinoma) are not separately evaluated due to class taxonomy constraints [5][6].

2. Literature Review

2.1 Conceptual Review

Patient-Disjoint Evaluation. Patient-disjoint splitting refers to a validation protocol in which no patient contributes images to both training and validation partitions. This contrasts with random image-level splitting, which is prevalent in medical imaging literature but violates the independence assumption underlying generalization estimates when multiple images derive from the same patient or imaging series [5][7]. In CT datasets, where each study may generate dozens of axial and coronal slices, random splitting creates slice-series leakage: the model encounters near-duplicate anatomical content during training and validation, artificially inflating performance metrics [8].

Probability Calibration. A classification model is calibrated if its predicted probabilities correspond to observed frequencies: among samples assigned probability p for class k, approximately p should be correct [9]. Calibration is distinct from discrimination—a model can rank cases correctly (high AUC) while producing systematically overconfident or underconfident probabilities [10]. Calibration is quantified via metrics such as Brier score (mean squared error between predicted probabilities and one-hot labels) and expected calibration error (ECE; bin-weighted absolute difference between confidence and accuracy) [11]. Post-hoc temperature scaling, which divides logits by a learned temperature parameter before softmax, is a simple and effective calibration method [5][12].

Domain Invariance. Domain invariance refers to a model's ability to maintain performance across distribution shifts arising from differences in scanner hardware, acquisition protocols, reconstruction parameters, or patient populations [13]. In multi-center CT studies, scanner vendor differences, contrast timing, and slice thickness introduce domain shifts that can degrade model accuracy and calibration when models are applied outside their training distribution [14].

Depthwise-Separable Convolutions. Depthwise-separable convolutions factorize standard convolutions into a depthwise spatial convolution (one filter per input channel) followed by a pointwise 1×1 convolution, reducing parameters and computation by a factor of approximately 1/8

to 1/9 compared to standard convolutions [15]. This architectural choice enables compact models suitable for deployment in resource-constrained clinical environments.

2.2 Theoretical Framework

Signal Detection Theory provides the conceptual foundation for understanding the distinction between discrimination and calibration. In signal detection theory, the ability to distinguish signal (pathology present) from noise (pathology absent) is characterized by sensitivity (true positive rate) and specificity (true negative rate), which together define the receiver operating characteristic curve [4]. AUC, the area under this curve, measures discrimination: the probability that a randomly selected positive case receives a higher score than a randomly selected negative case. However, signal detection theory explicitly acknowledges that the criterion (threshold) used to convert continuous evidence into binary decisions affects the trade-off between sensitivity and specificity [11]. Well-calibrated probabilities provide a principled basis for threshold selection, because the optimal threshold depends on the relative costs of false positives and false negatives, which in clinical contexts may vary by pathology and patient population [10].

Domain Adaptation Theory addresses the problem of distribution shift between training and deployment environments. Covariate shift, where the marginal distribution of inputs changes while the conditional distribution of labels given inputs remains stable, is the primary challenge in multi-center medical imaging [13]. Scanner vendor differences, contrast phase variations, and reconstruction kernel differences constitute covariate shifts that can degrade model performance even when the underlying pathology-label relationship is unchanged [14]. Domain-invariant architectures aim to learn representations that are robust to such shifts without requiring target-domain labels, typically through architectural inductive biases (e.g., attention mechanisms that focus on pathology-relevant features) or explicit domain-adversarial training [13].

Capacity Control and Architectural Inductive Bias theory, articulated in recent medical imaging literature, holds that model performance depends not only on parameter count but on the alignment between architectural inductive biases and the structure of the data [5][7]. For medical imaging tasks where lesions exhibit distinctive morphological and attenuation characteristics, convolutional architectures with explicit channel and spatial attention may provide stronger inductive biases than pure attention-based models, particularly when training data are limited [4].

2.3 Empirical Review

Dong et al. [5] introduced NephroNet, a compact depthwise-separable CNN with squeeze-and-excitation channel attention and a spatial gate, evaluated under patient-disjoint conditions on the Islam CT Kidney Dataset. The study established the first patient-disjoint benchmark for this dataset, addressing the slice-series leakage problem that had compromised prior evaluations. NephroNet achieved accuracy of 0.9997 and macro-AUC of 0.9969, with Brier score of 0.0007 and ECE of 0.0021 after temperature scaling. The study's limitations include single-region data origin, reliance on filename-derived patient grouping, and absence of external validation.

Batool et al. [12] developed an explainable AI framework combining VGG16 with layer-wise relevance propagation (LRP) for kidney cancer classification on CT images. The model achieved 98.75% accuracy with LRP heatmaps localizing tumor margins in agreement with clinical criteria. However, this study employed random image-level splitting, likely introducing patient-series leakage, and did not assess probability calibration. The reported accuracy cannot be directly compared with patient-disjoint benchmarks.

Sabharwal [15] developed a NephroNet variant for renal cell carcinoma classification and synthetic image generation using histopathology data from Dartmouth-Hitchcock Medical Center. While conceptually related to the present NephroNet architecture, this work addresses a fundamentally different task (histological subtype classification from whole-slide images) and does not involve CT data, patient-disjoint evaluation, or calibration assessment.

Other studies on the Islam dataset have reported accuracies ranging from 95% to 99.9% using various CNN architectures including ResNet, EfficientNet, and Vision Transformers [5]. However, none of these works employed patient-disjoint splitting, and most neglected calibration metrics, rendering their performance estimates methodologically incomparable to the present benchmark.

2.4 Research Gap

No validated, calibration-aware, patient-disjoint benchmark exists that simultaneously (i) prevents slice-series leakage through rigorous patient-level grouping, (ii) reports both discrimination and calibration metrics with confidence intervals, and (iii) evaluates domain invariance across contrast phases and anatomical planes within a multi-center cohort. Prior work has either employed random splitting without acknowledging leakage, reported only discrimination metrics, or failed to characterize performance across acquisition strata [5][7][12]. This study fills this gap by establishing a reproducible benchmark and systematically evaluating the NephroNet architecture against 30 capacity-matched baselines under conditions that emulate real-world deployment constraints.

3. Methodology

3.1 Research Design

This study employs a quantitative, retrospective design based on secondary analysis of a publicly available, de-identified multi-center CT dataset. The design combines patient-disjoint hold-out validation with comprehensive discrimination and calibration assessment. This approach is appropriate because it emulates the deployment scenario of interest (classification of images from patients not seen during training) while enabling transparent, reproducible comparison across architectures under identical training conditions [5][7].

3.2 Study Area / Population

The study population comprises patients who underwent abdominal CT imaging at multiple hospitals in Dhaka, Bangladesh, and whose de-identified images were retrospectively compiled into the Islam CT Kidney Dataset [5]. The dataset includes four diagnostic categories: Normal (5,077 images), Cyst (3,709 images), Tumor (2,283 images), and Stone (1,377 images), for a total of 12,446 images. Images span axial and coronal planes and include both contrast-enhanced and non-contrast acquisitions. All patient identifiers and accompanying DICOM metadata were removed prior to public release, and images were converted to lossless JPG format.

3.3 Sample Size and Sampling Technique

The full dataset of 12,446 images was partitioned into training (9,956 images; 80%) and validation (2,490 images; 20%) subsets using group-stratified splitting. Group identifiers were derived from filename stems following the Islam dataset naming convention (e.g., Tumor_0012_003.jpg), where the group ID is formed as <class>_<patient_token>, ensuring that all slices and planes from the same patient share the same group regardless of series [5]. GroupShuffleSplit was used to partition groups, and StratifiedShuffleSplit was applied to group-majority labels to preserve class balance at the patient level. This sampling technique prevents patient-series leakage while maintaining representativeness across diagnostic categories.

3.4 Data Collection Methods

Data were extracted from the publicly available Islam CT Kidney Dataset. No new data were collected. The dataset comprises de-identified CT images obtained from PACS at multiple hospitals in Dhaka, with two-pass quality assurance performed by the original dataset curators (radiologist and technologist) [5]. Images were resized to 320 × 320 pixels and rescaled to [0, 1] intensity range. Class weights were computed as inverse frequencies: Normal 0.6129, Cyst 0.8389, Tumor 1.3631, Stone 2.2586, to address class imbalance during training [5].

3.5 Research Instruments

Software and Libraries. Experiments were implemented in TensorFlow-Keras with NumPy and OpenCV utilities for image processing [5]. The random seed was fixed at 1,234 for Python, NumPy, and TensorFlow to ensure reproducibility.

NephroNet Architecture. NephroNet comprises a depthwise-separable CNN with squeeze-and-excitation (SE) channel attention (reduction ratio r = 16) and a SpatialGate (7 × 7) [5]. The stem consists of two Conv-BN-GELU blocks (48 and 64 channels) with Gaussian noise regularization ($\sigma = 0.01$) followed by max pooling. Three depthwise-separable convolutional stages (128 → 320 → 640 channels) each incorporate SE and SpatialGate modules with max pooling. The head comprises global average pooling, a dense layer (256 units) with GELU activation, dropout ($p = 0.35$), and a 4-class softmax output. Total trainable parameters: approximately 1,461,587, controlled via a no-op ParamPad layer for capacity-fair comparisons [5][7].

Preprocessing and Augmentation. Training used the following augmentation schedule: minimal augmentation for epochs 0–3 (label smoothing $\varepsilon = 0.05$); MixUp ($\alpha = 0.3$) and CutMix (area-adjusted λ') for epochs 4–27 with label smoothing $\varepsilon = 0.03$; heavy augmentations disabled after epoch 28 (label smoothing $\varepsilon = 0.00$). Geometric augmentations included horizontal flip, rotation (±10°), translation (±10%), shear (6°), zoom (±20%), brightness (±10%), CLAHE, and random erasing [5].

3.6 Validity and Reliability

Content validity: The diagnostic categories (Normal, Cyst, Tumor, Stone) correspond to clinically salient renal pathologies routinely assessed on abdominal CT. The dataset was curated with two-pass quality assurance by a radiologist and technologist, ensuring diagnostic label accuracy [5].

Predictive validity: Patient-disjoint splitting ensures that performance estimates reflect generalization to unseen patients rather than memorization of patient-specific features. The 80/20 split with group stratification preserves class balance while maintaining patient independence.

Reliability: Fixed random seeds, standardized training recipes, and transparent reporting of all hyperparameters support reproducibility. Calibration metrics (Brier score, ECE) with bootstrap confidence intervals provide reliability estimates for probability outputs.

3.7 Data Analysis Techniques

Models Compared. NephroNet was compared against 30 CNN and Vision Transformer baselines, all constrained to approximately 1.46 M trainable parameters via ParamPad layers where necessary [5]. Baselines included EfficientNet variants, ResNet architectures, DenseNet, Vision Transformers (ViT), Swin Transformers, and hybrid CNN-Transformer models.

Performance Metrics. Discrimination was assessed via accuracy (with 600-resample bootstrap 95% CI) and per-class, micro, and macro ROC-AUC (with 400-resample bootstrap 95% CI). Calibration was assessed via Brier score and expected calibration error (15 bins) [5][11]. Confusion matrices were generated for per-class error analysis.

Cross-validation. A single hold-out validation set was employed rather than K-fold cross-validation, specifically to emulate real-world deployment where the model encounters patients not seen during training [5][7]. This design choice prioritizes external validity over variance reduction.

Training Protocol. All models were trained using AdamW (weight decay 10^{-4}, gradient norm clip 1.0) with a warmup-cosine learning rate schedule ($\eta_0 = 2.5 \times 10^{-4}$, $\eta_{min} = 10^{-6}$, 10% warmup). Training ran for 50 epochs with batch size 12. Online exponential moving average (EMA, $\beta = 0.999$) weights were used for evaluation only. Test-time augmentation (original + horizontal flip + rotations ±5°) averaged class probabilities across transformations. Post-hoc temperature scaling was optimized on the validation set using Adam [5].

3.8 Ethical Considerations

This study used a fully anonymized, publicly available dataset with no patient-identifying information. All patient identifiers and DICOM metadata were removed prior to public release by the original dataset curators. Because the research involved secondary analysis of de-identified public data with no prospective collection or intervention, formal Institutional Review Board (IRB) review was not required under applicable guidelines for secondary analysis of de-identified public data [5]. No protected health information (PHI) was accessed at any point during this study.

4. Results

4.1 Data Presentation

Table 1. Dataset Characteristics by Diagnostic Category

Category	Total Images	Training	Validation	Class Weight
Normal	5,077	4,061	1,016	0.6129
Cyst	3,709	2,967	742	0.8389
Tumor	2,283	1,826	457	1.3631
Stone	1,377	1,102	275	2.2586
Total	**12,446**	**9,956**	**2,490**	—

Table 1 presents the class distribution across training and validation partitions. The dataset exhibits moderate class imbalance, with Normal images comprising 40.8% of the total and Stone images comprising 11.1%. Inverse-frequency class weights were applied to compensate during training.

Table 2. Calibration Metrics Before and After Temperature Scaling

Setting	Brier Score	ECE (15 bins)
Before temperature scaling	0.0024	0.0063
After temperature scaling (T* = 1.42)	0.0007	0.0021

Table 2 presents calibration metrics on the patient-disjoint validation set (N = 2,490). Temperature scaling reduced the Brier score from 0.0024 to 0.0007 and the expected calibration error from 0.0063 to 0.0021, representing substantial improvements in probability reliability [5].

4.2 Analysis of Results

Discrimination Performance. NephroNet achieved accuracy of 0.9997 (95% CI: 0.9984–1.0000) on the patient-disjoint hold-out set. Macro-AUC was 0.9969 (95% CI: 0.9953–0.9983), with per-class AUC values of Cyst 0.9978, Normal 0.9971, Stone 0.9960, and Tumor 0.9967 [5]. The second-ranked model, EfficientNetV2-s, achieved accuracy of 0.9699 (95% CI: 0.9661–0.9735), with non-overlapping confidence intervals indicating a statistically significant performance gap ($\Delta = 0.0298$) [5][11].

Comparison Against Baselines. Across 30 capacity-matched CNN and Vision Transformer baselines trained under identical recipes, NephroNet achieved the highest accuracy and lowest validation loss. Notably, many larger and more complex architectures (e.g., Swin-Tiny, CaiT-XXS24-224) underperformed the compact NephroNet, suggesting that architectural inductive bias (dual attention with depthwise-separable convolutions) aligned well with the structure of this classification task [5][7].

Calibration Quality. After temperature scaling ($T^* = 1.42$), the expected calibration error of 0.0021 indicates that predicted probabilities deviate from observed frequencies by only 0.21 percentage points on average across 15 confidence bins [5][11]. This level of calibration is substantially better than typical medical imaging classifiers, which often exhibit ECE values in the range of 0.05 to 0.15 before recalibration [10].

Feature and Error Analysis. Confusion matrix analysis on a representative subset (approximately 472 images) revealed near-perfect diagonal dominance, with only 4 misclassifications across the full validation set (N = 2,490) [5]. The classification report showed per-class Precision/Recall/F1 scores all ≥ 0.9987, with Stone and Tumor achieving perfect scores (1.0/1.0/1.0) [5].

5. Discussion

5.1 Interpretation

The primary finding of this study is that a compact, parameter-efficient architecture (NephroNet, ≈1.46 M parameters) can achieve near-ceiling discrimination accuracy (0.9997) on patient-disjoint four-class kidney CT classification while maintaining excellent probability calibration (ECE = 0.0021 after temperature scaling). This answers Research Question 2 affirmatively: NephroNet outperformed all 30 capacity-matched baselines, demonstrating that architectural inductive bias—specifically, the combination of depthwise-separable convolutions with squeeze-and-excitation channel attention and spatial gating—provides advantages over larger, more complex architectures when parameter count is controlled.

Regarding Research Question 1, the patient-disjoint protocol did not substantially degrade performance relative to prior random-split results on the same dataset [5]. This finding is important because it suggests that the near-ceiling accuracy reported in the literature is not primarily an artifact of slice-series leakage, but rather reflects the visual separability of the four diagnostic categories in this dataset. The original images were extracted as regions of interest around the kidney, reducing background variability, and the four classes exhibit distinct attenuation characteristics (fluid density in cysts, calcific density in stones, soft-tissue mass in tumors, uniform parenchyma in normals) that are highly discriminable at the given resolution [5].

Regarding Research Question 3, post-hoc temperature scaling improved calibration substantially (Brier score 0.0024 → 0.0007; ECE 0.0063 → 0.0021) without degrading discrimination. The optimal temperature of 1.42 indicates that NephroNet's pre-calibration outputs were slightly overconfident, a common pattern in deep neural networks trained with cross-entropy loss [10][12]. The magnitude of improvement (ECE reduction of 0.0042) is clinically meaningful: at a decision threshold of 0.5, this translates to more reliable probability estimates for borderline cases.

Regarding Research Question 4, the dataset does not provide per-image metadata for contrast phase or anatomical plane in the publicly released version, precluding stratified analysis of calibration across these acquisition characteristics. This represents a limitation of the present study and a direction for future work with access to DICOM metadata.

5.2 Implications

Academic Implications. This study extends the theoretical understanding of capacity control and inductive bias in medical imaging by demonstrating that parameter-efficient architectures with well-designed attention mechanisms can match or exceed larger models on tasks with limited training data and relatively distinct visual classes [5][7]. The calibration-aware evaluation protocol, incorporating Brier score and ECE alongside discrimination metrics, provides a template that the broader medical imaging AI community can adopt to improve the clinical credibility of reported results.

Practical Implications. For radiology departments considering AI-assisted CT interpretation, the findings suggest that high-performance kidney CT classification can be deployed on resource-constrained hardware, as the NephroNet architecture requires only 1.46 M parameters—orders of magnitude smaller than typical CNN and transformer backbones [5]. However, the single-region data origin means that deployment across different scanner vendors and geographic populations requires external validation. The calibration protocol (temperature scaling on a held-out validation set) is computationally inexpensive and should be standard practice before clinical deployment. Clinicians should monitor ECE as a quality metric alongside accuracy, and recalibrate when local calibration degrades.

5.3 Limitations

1. **Single geographic origin:** All data originate from hospitals in Dhaka, Bangladesh, and generalizability to other geographic regions, scanner vendors, and patient populations remains unvalidated [5][13].
2. **Patient grouping heuristic:** The patient grouping relies on the Islam dataset filename convention, which, while manually spot-checked on a 200-filename sample, cannot be guaranteed exhaustive for all DICOM series configurations [5].
3. **Post-hoc calibration:** Temperature scaling is applied post-hoc and may not be optimal across all operating thresholds; differentiable calibration objectives may further improve reliability [5][11].
4. **Limited acquisition metadata:** The publicly released dataset does not include per-image contrast phase or anatomical plane labels, precluding stratified calibration analysis across these clinically relevant dimensions.
5. **Class taxonomy constraints:** The four-class taxonomy precludes evaluation of challenging sub-classifications such as fat-poor angiomyolipoma versus renal cell carcinoma or early-stage transitional cell carcinoma [5].

5.4 Future Research Directions

1. **External multi-institutional validation:** Prospective evaluation of NephroNet on CT cohorts from different geographic regions, scanner vendors (GE, Siemens, Philips, Canon), and patient populations to quantify domain shift and calibration degradation under real-world distribution shifts [13][14].
2. **Integration of volumetric and multi-phase information:** Extension to 3D volumetric analysis and multi-phase contrast fusion to improve discrimination in challenging cases and potentially enhance domain invariance [5].
3. **Differentiable calibration objectives:** Exploration of calibration-aware training losses (e.g., mL1-ACE, focal calibration) that optimize probability reliability during training rather than post-hoc [5][10].
4. **Extension to Bosniak risk stratification and stone composition prediction:** Application of the benchmark framework to clinically actionable sub-tasks with direct implications for management decisions [5].

6. Conclusion

This study established a patient-disjoint, calibration-aware benchmark for four-class kidney CT classification and demonstrated that the NephroNet architecture—a compact depthwise-separable CNN with dual attention mechanisms—achieves accuracy of 0.9997, macro-AUC of 0.9969, Brier score of 0.0007, and expected calibration error of 0.0021 on a multi-center cohort of 12,446 images. The primary contribution is a reproducible, leakage-free evaluation protocol that combines patient-level grouping, capacity-controlled baseline comparison, and comprehensive discrimination and calibration reporting. For radiologists and health system administrators, the practical takeaway is that high-performance renal CT classification is achievable with parameter-efficient architectures suitable for resource-constrained deployment, provided that probability calibration is explicitly optimized and monitored. Future work should prioritize external validation across scanner vendors and geographic regions, as near-ceiling performance on a single-region dataset does not guarantee clinical translation [5][13].

References

1. Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, *71*(3), 209–249.

2. Heilbrun, M. E., & Davenport, M. S. (2018). Renal imaging: A practical approach. *Radiologic Clinics of North America*, *56*(6), 849–863.

3. Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, *42*, 60–88.

4. Shen, D., Wu, G., & Suk, H.-I. (2017). Deep learning in medical image analysis. *Annual Review of Biomedical Engineering*, *19*, 221–248.

5. Dong, Y., Akter Muni, M., Islam, R., Tasnim, S., Shahid Juthi, S., Jahirul Islam, M., Fassi, M. A. L., Sharif Robbani, M., Zareen, S., & 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. https://doi.org/10.3389/fmed.2026.1827704

6. Varoquaux, G., & Cheplygina, V. (2022). Machine learning for medical imaging: Methodological failures and recommendations for the future. *npj Digital Medicine*, *5*(1), 48.

7. Roberts, M., Driggs, D., Thorpe, M., Gilbey, J., Yeung, M., Ursprung, S., Aviles-Rivero, A. I., Etmann, C., McCague, C., Beer, L., Weir-McCall, J. R., Teng, Z., Gkrania-Klotsas, E., Rudd, J. H. F., Sala, E., & Schönlieb, C.-B. (2021). Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nature Machine Intelligence*, *3*(3), 199–217.

8. Yagis, E., Atnafu, S. W., García Seco de Herrera, A., Marzi, C., Scheda, R., Giannelli, M., Tessa, C., Citi, L., & Diciotti, S. (2021). Effect of data leakage in brain MRI classification using 2D convolutional neural networks. *Scientific Reports*, *11*(1), 22544.

9. Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. *Proceedings of the 34th International Conference on Machine Learning*, *70*, 1321–1330.

10. Minderer, M., Djolonga, J., Romijnders, R., Hubis, F., Zhai, X., Houlsby, N., Tran, D., & Lucic, M. (2021). Revisiting the calibration of modern neural networks. *Advances in Neural Information Processing Systems*, *34*, 15682–15694.

11. Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J. V., Lakshminarayanan, B., & Snoek, J. (2019). Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. *Advances in Neural Information Processing Systems*, *32*, 13991–14002.

12. Batool, A., Ahmed, F., Naz, N. S., Altameem, A., Rehman, A. U., Adnan, K. M., & Almogren, A. (2025). An explainable deep learning framework for kidney cancer classification using VGG16 and layer-wise relevance propagation on CT images. *Diagnostics*, *15*(6), 731.

13. Guan, H., & Liu, M. (2021). Domain adaptation for medical image analysis: A survey. *IEEE Transactions on Biomedical Engineering*, *69*(3), 1173–1185.

14. Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., & Oermann, E. K. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS Medicine*, *15*(11), e1002683.

15. Sabharwal, Y. (2023). NephroNet: A novel program for identifying renal cell carcinoma and generating synthetic training images with convolutional neural networks and diffusion models. *International Science and Engineering Fair Abstracts*. https://abstracts.societyforscience.org/Home/PrintPdf?projectId=23989