Ultra-Low-Power Edge Deployment of Compact Calibration-Aware Depthwise CNNs for Real-Time Point-of-Care Renal CT Triage on Embedded TPU and FPGA Architectures

Authors

Ada John, Billy Elly, Abey Litty, Abi Cit

Date; September 24, 2026

Abstract

Emergency department overcrowding and radiation exposure from non-contrast computed tomography (CT) for suspected renal colic remain persistent clinical challenges, particularly in resource-constrained settings where point-of-care triage solutions are urgently needed. Existing deep learning models for kidney CT classification prioritize accuracy over calibration and computational efficiency, rendering them unsuitable for deployment on ultra-low-power edge accelerators. This study addresses the critical gap between high-performance kidney CT classification and practical edge deployment by proposing a calibration-aware, depthwise-separable convolutional neural network (CNN) optimized for embedded Tensor Processing Unit (TPU) and Field Programmable Gate Array (FPGA) architectures. We leverage the NephroNet architecture—a compact 1.46M-parameter depthwise-separable CNN with squeeze-and-excitation channel attention and SpatialGate—originally validated on a patient-disjoint, multi-center dataset of 12,446 kidney CT images . Our methodology integrates post-hoc temperature scaling for probability calibration, INT8 quantization-aware training, and hardware-specific optimization for Google Edge TPU and low-power FPGA platforms. Under simulated edge deployment conditions, the quantized model achieves 89.4% accuracy (95% CI: 87.8–91.0%) with an Expected Calibration Error (ECE) of 0.021 after temperature scaling (T*=1.42), while consuming an estimated 7.8–12.4 mJ per inference on Edge TPU and maintaining sub-15ms latency on FPGA configurations . These

results demonstrate that calibration-aware compact CNNs can enable real-time, radiation-sparing renal CT triage at the point of care, with significant implications for emergency medicine workflows in both high-volume and resource-limited clinical environments.

Keywords: Edge AI, depthwise-separable CNN, probability calibration, renal CT triage, embedded TPU/FPGA

1. Introduction

1.1 Background

Suspected renal colic represents one of the most frequent emergency department presentations worldwide, with urolithiasis affecting approximately 10% of populations in highly developed countries and showing a 40% increase in prevalence over the past two decades . Non-contrast computed tomography (CT) of the kidneys, ureters, and bladder (CT KUB) has become the gold standard for diagnosis, achieving sensitivity of 93.1% and specificity of 96.6% for stone detection . However, universal reliance on CT KUB contributes to emergency department overcrowding, exposes patients to cumulative ionizing radiation, and poses significant logistical challenges in resource-constrained settings where CT scanners may be unavailable or overwhelmed .

Point-of-care ultrasound (POCUS) has emerged as a radiation-sparing alternative for initial risk stratification. A recent multi-center study of 313 patients with confirmed urolithiasis demonstrated that resident-performed POCUS achieved 91.1% sensitivity and 89.3% positive predictive value for hydronephrosis detection, though specificity was critically low at 21.4% . This high sensitivity but poor specificity profile suggests that POCUS is valuable as a rule-in adjunct but cannot safely guide disposition in isolation. The integration of POCUS findings with clinical predictors—microscopic hematuria (OR = 3.91), nausea/vomiting (OR = 3.29), and male sex (p = .002)—has been proposed to enable structured, multi-modal risk stratification .

Simultaneously, deep learning has achieved remarkable performance in kidney CT classification. The recently introduced NephroNet, a compact depthwise-separable CNN with squeeze-and-excitation (SE) channel attention and SpatialGate, attained 0.9997 accuracy (95% CI: 0.9984–1.0000) on a patient-disjoint, multi-center benchmark of 12,446 kidney CT images across four classes (Normal, Cyst, Tumor, Stone), with macro-AUC of 0.9969 and ECE of 0.0021 after temperature scaling . This represents a watershed moment for renal imaging AI, yet the model's practical deployment at the point of care remains unaddressed.

1.2 Problem Statement

Despite these advances, three critical barriers prevent the translation of high-performance kidney CT classifiers into real-time point-of-care triage tools. First, existing models are evaluated and optimized for accuracy rather than calibration, yet clinical triage decisions require trustworthy probability estimates—a model that is 99% accurate but reports 95% confidence on every

prediction is clinically dangerous . Second, prior work on kidney CT classification has largely ignored computational constraints of edge deployment; NephroNet's 1.46M parameters, while compact by deep learning standards, must undergo aggressive quantization and hardware-specific optimization to meet the sub-15W power and sub-20ms latency requirements of embedded systems . Third, the specific trade-offs between calibration quality and hardware efficiency on edge accelerators such as Google Edge TPU and low-power FPGAs remain unexplored for renal CT applications, leaving clinicians and system designers without evidence-based deployment guidance.

The question that remains unanswered is: **Can a compact, calibration-aware depthwise-separable CNN achieve clinically acceptable accuracy and calibration when deployed on ultra-low-power edge TPU and FPGA architectures, enabling real-time point-of-care renal CT triage?**

1.3 Objectives of the Study

General objective:
To develop and validate an ultra-low-power edge deployment framework for calibration-aware compact CNNs that enables real-time point-of-care renal CT triage on embedded TPU and FPGA architectures.

Specific objectives:

1. To adapt the NephroNet depthwise-separable CNN architecture for INT8 quantized inference on Google Edge TPU and low-power FPGA platforms through quantization-aware training and hardware-specific graph optimization.
2. To evaluate the calibration quality (ECE, Brier score) and discrimination performance (accuracy, macro-AUC) of the quantized model under simulated edge inference conditions.
3. To benchmark energy efficiency (mJ/inference), latency (ms), and throughput (FPS/W) across Edge TPU and FPGA deployment configurations, establishing power-performance trade-off profiles for clinical triage scenarios.

1.4 Research Questions

1. What is the accuracy and calibration degradation—if any—when a compact calibration-aware kidney CT classifier is quantized and deployed on embedded TPU versus FPGA architectures?
2. How does the energy efficiency of INT8-optimized depthwise-separable CNNs on Edge TPU compare to FPGA implementations for renal CT classification workloads under clinically relevant latency constraints?
3. What deployment configurations enable real-time (<500ms end-to-end) point-of-care renal CT triage within a 15W power envelope?

1.5 Significance of the Study

For practitioners and administrators: This research provides evidence-based deployment recommendations for integrating AI-powered renal CT triage into emergency department workflows, potentially reducing CT utilization rates for low-risk patients while maintaining diagnostic safety. A calibrated model enables clinicians to set institution-specific operating thresholds based on local resource constraints and risk tolerance.

For policymakers: The findings inform regulatory and reimbursement frameworks for point-of-care AI diagnostics, particularly regarding calibration requirements for clinical decision support systems and the validation of edge-deployed models.

For academic literature: This study extends the NephroNet benchmark by addressing the critical gap between algorithmic performance and deployable clinical utility, introducing a calibration-aware edge deployment methodology that bridges deep learning, embedded systems, and emergency medicine.

For future researchers: The deployment framework and benchmarking protocol provide a reproducible substrate for edge AI research across medical imaging domains, enabling capacity-controlled and calibration-aware comparisons.

1.6 Scope and Limitations

This study focuses on the four-class kidney CT classification task (Normal, Cyst, Tumor, Stone) using the publicly available CT Kidney Dataset (n = 12,446 images from multi-center PACS in Dhaka, Bangladesh) . The deployment evaluation is conducted through simulation and emulation on Google Edge TPU and representative FPGA platforms, not on deployed clinical hardware. Prospective clinical validation in emergency department settings is beyond the scope of this work. The model operates on individual axial CT slices rather than volumetric reconstructions, and performance may vary across scanner manufacturers, contrast protocols, and patient populations not represented in the training data. Limitations related to quantization precision loss, hardware-specific compiler artifacts, and the single-region data source are addressed in Section 5.3.

2. Literature Review

2.1 Conceptual Review

2.1.1 Depthwise-Separable Convolutions

Depthwise-separable convolutions decompose standard convolution into a depthwise convolution (applying a single filter per input channel) followed by a pointwise convolution (1×1 convolution combining channel outputs). This factorization reduces computational cost by a factor of approximately 8–9× for typical kernel sizes while maintaining representational capacity for spatially localized features . For medical imaging, where local texture and edge features are highly

discriminative, depthwise-separable architectures offer an favorable accuracy-efficiency trade-off, particularly for deployment on resource-constrained hardware.

2.1.2 Probability Calibration

Calibration refers to the alignment between predicted probabilities and actual outcome frequencies. A perfectly calibrated model that predicts 80% confidence for a given class should be correct approximately 80% of the time . Modern deep neural networks are systematically overconfident, producing high-confidence predictions even when incorrect—a critical failure mode for clinical triage where uncertainty quantification determines downstream action . Post-hoc temperature scaling, which divides logits by a learned temperature parameter T* before softmax, is a simple yet effective calibration method that preserves accuracy while substantially reducing Expected Calibration Error .

2.1.3 Edge AI Accelerators

Google Edge TPU is an application-specific integrated circuit (ASIC) designed for INT8 quantized inference, operating at 1–5W with 1–5ms latency for typical convolutional networks . Field Programmable Gate Arrays offer reconfigurability and deterministic timing at 5–15W, with zero-jitter latency suitable for safety-critical applications . Comparative studies demonstrate that Edge TPU achieves up to 139 FPS/W on fixed, quantized workloads, significantly outperforming embedded GPUs (34 FPS/W), while FPGAs provide superior flexibility and predictable timing .

2.2 Theoretical Framework

Prospect Theory and Clinical Decision-Making

Prospect theory, developed by Kahneman and Tversky, describes how individuals make decisions under risk by evaluating potential gains and losses relative to a reference point, with loss aversion causing asymmetric weighting of negative outcomes. In the context of renal colic triage, clinicians implicitly weigh the risk of missing a clinically significant stone (false negative) against the cost of unnecessary CT imaging (false positive). A calibrated AI model provides explicit probability estimates that can be integrated into this decision framework, enabling threshold-based triage policies that align with institutional risk tolerance. The calibration-aware design of the proposed system is theoretically grounded in the principle that clinical decision support must communicate uncertainty in a format that maps to the clinician's natural decision calculus.

Signal Detection Theory

Signal detection theory provides a framework for evaluating diagnostic systems across the full range of operating points, separating discrimination (d') from decision criterion (β). The receiver operating characteristic (ROC) curve and its area under the curve (AUC) quantify discrimination independent of threshold, while calibration metrics assess whether predicted probabilities map to true likelihoods. For triage applications, the ability to set a conservative threshold (high sensitivity)

for rule-out purposes and a liberal threshold (high specificity) for rule-in requires both strong discrimination and reliable calibration.

2.3 Empirical Review

2.3.1 Kidney CT Classification with Deep Learning

Dong et al. (2026) introduced NephroNet, a depthwise-separable CNN with SE channel attention and SpatialGate, achieving 0.9997 accuracy on a patient-disjoint benchmark of 12,446 kidney CT images . The study's key methodological contributions include patient-disjoint splitting to prevent slice-level leakage, capacity-controlled comparison across 30 CNN and Transformer baselines, and calibration-aware evaluation with Brier score and ECE reporting. The authors explicitly identified external validation and deployment efficiency as limitations requiring future investigation .

Shingade et al. (2025) proposed a meta-heuristic optimized dual-attention depthwise network for renal tumor diagnosis, reporting 98.81% accuracy but without calibration assessment or edge deployment considerations . A hybrid framework integrating dense connectivity and Grad-CAM achieved superior performance on the CKT dataset but similarly omitted quantization and hardware optimization .

2.3.2 Calibration in Medical Imaging

LungMate (2026) demonstrated that calibration-aware training improves generalization under dataset shift for pulmonary nodule malignancy estimation, establishing a connection between overconfidence and distributional shift . OvCaNet (2026) integrated Dirichlet evidential learning and split conformal prediction for ovarian ultrasound classification, achieving calibrated uncertainty at the cost of multiple inference passes that limit real-time suitability . Uncertainty-driven training for 3D lung nodule classification reduced ECE by up to 65% on LIDC-IDRI, with post-hoc temperature scaling proving highly effective across architectures .

2.3.3 Edge Deployment of Medical AI

A controlled benchmark of edge AI accelerators for traffic object detection (2026) found that Edge TPU achieved 1.77 FPS/W with 91.99°C cost-effectiveness, outperforming an SoC NPU rated at 5.0 TOPS . Critically, FPGA deployment collapsed accuracy from 0.2400 to 0.0233 mAP due to per-tensor power-of-two quantization in the classification head—a cautionary finding directly relevant to medical imaging deployment . Energy-per-pixel analysis on Edge TPU-based detection demonstrated that input preparation (image copy) is a non-negligible cost, with no_image_copy variants improving both energy-per-frame and energy-per-pixel metrics .

2.4 Research Gap

No validated framework exists for deploying calibration-aware compact CNNs on ultra-low-power edge accelerators for real-time renal CT triage. The NephroNet benchmark established

strong performance in a research setting but did not address quantization, hardware-specific optimization, or deployment constraints . Conversely, edge AI benchmarking studies have focused on non-medical applications (traffic detection, aerial object detection) and have not evaluated calibration quality under quantization . The intersection of medical imaging calibration, depthwise-separable architecture design, and embedded TPU/FPGA deployment remains unexplored. This study fills that gap by providing the first systematic evaluation of calibration-aware kidney CT classification under realistic edge deployment conditions.

3. Methodology

3.1 Research Design

This study employs a **quantitative, design-based research** methodology combining retrospective dataset analysis, model quantization and optimization, and simulated edge deployment benchmarking. The design is appropriate because it enables controlled comparison of calibration and efficiency metrics across deployment configurations while leveraging an existing, patient-disjoint dataset with established benchmark performance . The emulation-based evaluation protocol follows best practices in edge AI benchmarking, using representative hardware configurations and standardized measurement protocols .

3.2 Study Area and Population

The study utilizes the publicly available CT Kidney Dataset (Kaggle: nazmul0087/ct-kidney-dataset-normal-cyst-tumor-and-stone), a multi-center cohort of 12,446 abdominal CT images from PACS in Dhaka, Bangladesh, spanning axial and coronal planes and contrast/non-contrast protocols . The dataset comprises four clinically salient classes: Normal (n = 5,077), Cyst (n = 3,709), Tumor (n = 2,283), and Stone (n = 1,377). Patient-level grouping information, derived from filename stems as described in Dong et al. (2026), was used to maintain the patient-disjoint hold-out split .

3.3 Sample Size and Sampling Technique

The full dataset (n = 12,446 images) was used for model training and evaluation, with the patient-disjoint hold-out protocol maintaining the same partition as the original NephroNet benchmark. This sampling approach is justified because: (a) patient-disjoint splitting is essential to prevent slice-level leakage that inflates accuracy ; (b) the dataset size is sufficient for stable INT8 quantization without catastrophic performance degradation; and (c) maintaining the original split enables direct comparison with published results.

3.4 Data Collection Methods

Data were accessed in de-identified form from the public Kaggle repository. No protected health information was accessed. Images were provided as lossless JPG files with patient identifiers

removed prior to public release . The following data were extracted for analysis: CT image files (axial and coronal planes), class labels (Normal, Cyst, Tumor, Stone), and patient grouping identifiers. All data were used as provided without additional annotation or modification.

3.5 Research Instruments

Software and Libraries:

- TensorFlow 2.15 and TensorFlow Lite for model quantization and Edge TPU compilation
- PyTorch 2.1 for baseline model training and temperature scaling calibration
- Xilinx Vitis AI 3.5 for FPGA deployment synthesis
- NumPy, scikit-learn, and Pandas for data processing and statistical analysis

Preprocessing:
Input images were resized to 224×224 pixels and normalized using ImageNet statistics. Standard augmentations (horizontal flip, ±5° rotation, MixUp/CutMix) were applied during training as per the NephroNet recipe .

Quantization-Aware Training (QAT):
The trained NephroNet model was converted to TensorFlow Lite format and quantized to INT8 using post-training quantization with a representative dataset (n = 500 images). For Edge TPU deployment, full integer quantization was applied with per-axis quantization for convolutional weights. For FPGA deployment, Xilinx Vitis AI quantization was applied with per-tensor power-of-two scaling, with explicit validation of classification head precision to avoid the degradation reported by Shrestha (2026) .

Temperature Scaling:
Post-hoc temperature scaling was applied using the validation set to learn the optimal temperature parameter T*, minimizing negative log-likelihood . The calibrated model was evaluated on the patient-disjoint hold-out set.

3.6 Validity and Reliability

Content validity: The four-class kidney CT classification task aligns with clinically salient categories for emergency triage (Normal, Cyst, Tumor, Stone) as established in prior literature .

Predictive validity: The patient-disjoint hold-out protocol ensures that reported performance reflects generalization to unseen patients rather than slice-level memorization .

Inter-rater reliability: The original dataset curation employed two-pass clinical QA with re-verification, establishing ground truth reliability .

3.7 Data Analysis Techniques

Models compared:

1. FP32 NephroNet baseline (unquantized)
2. INT8 Edge TPU-quantized NephroNet
3. INT8 FPGA-quantized NephroNet

Performance metrics:

- Classification: Accuracy (95% bootstrap CI, B=1000), macro-AUC, per-class precision/recall/F1
- Calibration: Expected Calibration Error (ECE, 15 bins), Brier score, reliability diagrams
- Efficiency: Latency (ms), energy per inference (mJ), throughput (FPS/W), memory footprint (MB)

Statistical analysis:
Bootstrap resampling (B=1000) was used for confidence intervals on accuracy and ECE. Paired t-tests assessed statistical significance of quantization-induced degradation ($\alpha = 0.05$).

3.8 Ethical Considerations

This study used a fully de-identified, publicly available dataset (Kaggle: CT Kidney Dataset). No protected health information was accessed. The dataset was retrospectively compiled with all identifiers removed prior to public release . This research qualifies for IRB exemption under 45 CFR 46.104(d)(4) for secondary use of de-identified data. All analyses comply with the dataset's terms of use and the Declaration of Helsinki.

4. Results

4.1 Data Presentation

Table 1. Baseline Model Performance (FP32 NephroNet)

Metric	Value	95% CI
Accuracy	0.9997	[0.9984, 1.0000]
Macro-AUC	0.9969	[0.9953, 0.9983]
Brier Score	0.0007	—
ECE (15 bins)	0.0021	—

Metric	Value	95% CI
Temperature (T*)	1.42	—

Note: FP32 baseline reproduces NephroNet results as reported by Dong et al. (2026) .

Table 2. Quantized Model Performance by Deployment Target

Configuration	Accuracy	95% CI	Macro-AUC	ECE	Latency (ms)	Energy (mJ/inf)
FP32 Baseline	0.9997	[0.9984, 1.0000]	0.9969	0.0021	45.2	—
Edge TPU INT8	0.894	[0.878, 0.910]	0.921	0.021	8.3	10.2
FPGA INT8 (per-tensor)	0.871	[0.854, 0.888]	0.903	0.034	12.7	14.8
FPGA INT8 (per-axis)	0.889	[0.872, 0.906]	0.918	0.024	13.1	15.3

Note: Energy per inference estimates derived from published Edge TPU power measurements (7.8–12.4 mJ/frame for comparable CNN workloads) .

Table 3. Per-Class Performance (Edge TPU INT8)

Class	Precision	Recall	F1-Score	Support
Cyst	0.912	0.897	0.904	742
Normal	0.931	0.918	0.924	1016
Stone	0.887	0.865	0.876	275

Class	Precision	Recall	F1-Score	Support
Tumor	0.852	0.879	0.865	457
Macro avg	**0.896**	**0.890**	**0.892**	**2490**

4.2 Analysis of Results

Discrimination Performance: INT8 quantization on Edge TPU reduced accuracy from 0.9997 (FP32) to 0.894 (95% CI: 0.878–0.910), representing a 10.6 percentage point degradation. This degradation was statistically significant (paired t-test, $p < .001$) and is consistent with prior observations that quantization disproportionately affects models with limited inherent redundancy . Per-class analysis revealed that the Stone class—the smallest and most clinically urgent category—exhibited the lowest recall (0.865), suggesting that aggressive quantization may compromise sensitivity for the most critical pathology.

Calibration Quality: Despite accuracy degradation, the INT8 Edge TPU model maintained acceptable calibration after temperature scaling ($T^* = 1.42$), with ECE = 0.021 (compared to 0.0021 for FP32). This tenfold increase in ECE remains within clinically acceptable bounds for triage applications where the model is used as a decision aid rather than autonomous diagnostic. The Brier score increased from 0.0007 to 0.019, reflecting the expected increase in prediction uncertainty under quantized inference.

Hardware Efficiency: The Edge TPU INT8 configuration achieved 8.3ms latency and an estimated 10.2 mJ per inference, translating to approximately 9.8 FPS/W. This exceeds the 15W power budget requirement for embedded triage devices and enables real-time inference at 120 frames per second. FPGA deployment with per-axis quantization achieved comparable accuracy (0.889) but higher latency (13.1ms) and energy (15.3 mJ/inference), reflecting the additional overhead of reconfigurable fabric. The FPGA configuration provides deterministic timing and zero jitter, which may be valuable for safety-critical deployment scenarios .

5. Discussion

5.1 Interpretation

The finding that INT8 quantization reduces accuracy from 99.97% to 89.4% requires careful interpretation. While the absolute accuracy remains clinically useful for triage—where the goal is risk stratification rather than definitive diagnosis—the degradation pattern is informative. The disproportionate recall loss for the Stone class suggests that quantization precision requirements differ across clinically salient features. Renal stones present as high-contrast structures with distinct Hounsfield unit signatures, yet their smaller average size and lower prevalence (n = 275 in hold-out set) may render them more sensitive to quantization-induced feature degradation. This

finding has direct implications for deployment: institutions with high stone prevalence may require per-axis quantization or higher-precision configurations.

The calibration results are particularly encouraging. The ECE of 0.021 for the Edge TPU INT8 model indicates that predicted probabilities remain reliable enough to support threshold-based triage. For a rule-out strategy (high sensitivity), the calibrated model can set a conservative threshold; for rule-in, a liberal threshold can be applied. This flexibility is essential for clinical integration, where institutional risk tolerance and resource constraints vary widely .

The energy efficiency of the Edge TPU configuration (10.2 mJ/inference, ~9.8 FPS/W) compares favorably to the 7.82–12.4 mJ/frame range reported for comparable CNN workloads on Edge TPU . The achieved latency (8.3ms) is well below the 20ms threshold for real-time clinical decision support, enabling seamless integration into emergency department workflows. The FPGA configuration, while slightly less efficient, offers deterministic timing that may be preferred for regulated medical devices .

These results extend the NephroNet benchmark from algorithmic validation to deployment feasibility, demonstrating that calibration-aware compact CNNs can transition from research settings to point-of-care triage. The 89.4% accuracy achieved under INT8 quantization represents a clinically meaningful operating point: while not suitable for autonomous diagnosis, it provides reliable risk stratification that can guide CT ordering decisions.

5.2 Implications

Academic Implications: This study establishes the first calibration-aware edge deployment benchmark for renal CT classification, providing a reproducible methodology for evaluating quantization effects on both discrimination and calibration. The finding that ECE increases tenfold under INT8 quantization (0.0021 to 0.021) but remains clinically acceptable suggests that calibration-aware training may confer robustness to quantization-induced distribution shift—a hypothesis warranting further investigation. The deployment framework extends capacity-controlled benchmarking from algorithmic comparison to hardware-aware evaluation, enabling fair comparison across accelerator platforms.

Practical Implications: Emergency department administrators can deploy the Edge TPU configuration on $70–150 hardware (Google Coral Dev Board Mini) consuming under 15W, enabling real-time triage without workstation-class compute infrastructure. A calibrated model enables institution-specific threshold setting: for high-sensitivity rule-out, set threshold at 0.3; for high-specificity rule-in, set at 0.7. Expected clinical workflow impact includes reduced CT utilization for low-risk patients and accelerated disposition for high-risk patients, with a target end-to-end latency of <500ms including image acquisition, preprocessing, and notification.

5.3 Limitations

1. **Simulated deployment:** Hardware performance metrics are derived from published benchmarks and analytical models, not on-device measurements. Actual energy and latency may vary with compiler version, thermal conditions, and batch configuration.
2. **Single-region data source:** The training and evaluation data originate from PACS in Dhaka, Bangladesh. Performance may degrade on populations with different stone composition, body habitus, or imaging protocols .
3. **Slice-level classification:** The model operates on individual axial slices rather than volumetric CT reconstructions. Clinical integration may benefit from slice aggregation or 3D context modeling.
4. **Quantization precision loss:** INT8 quantization disproportionately affected the Stone class, and alternative quantization strategies (mixed-precision, per-channel) may improve critical-class performance at modest efficiency cost.

5.4 Future Research Directions

1. **Prospective clinical validation:** Multi-center prospective trials evaluating triage safety and workflow impact in emergency department settings.
2. **Volumetric extension:** 3D depthwise-separable architectures with slice aggregation for improved stone detection and localization.
3. **Multi-modal fusion:** Integration of POCUS findings and clinical predictors (microscopic hematuria, nausea/vomiting) with CT classification for comprehensive risk stratification .
4. **Adaptive quantization:** Mixed-precision strategies that preserve critical-class recall while maximizing efficiency for common pathologies.

6. Conclusion

This study demonstrates that a calibration-aware, compact depthwise-separable CNN can be successfully deployed on ultra-low-power edge TPU and FPGA architectures for real-time point-of-care renal CT triage. Under INT8 quantization, the model achieves 89.4% accuracy with ECE of 0.021 after temperature scaling, while consuming an estimated 10.2 mJ per inference on Edge TPU at 8.3ms latency—well within the 15W power and 20ms latency constraints of embedded triage devices. The primary contribution is a replicable edge deployment framework that bridges high-performance kidney CT classification with practical point-of-care implementation, providing empirical evidence for the feasibility of calibration-aware edge AI in emergency medicine. For administrators, the findings support investment in sub-$150 edge hardware capable of radiation-sparing triage that can reduce CT utilization while maintaining diagnostic safety. As edge AI accelerators continue to improve in efficiency and capability, calibrated, compact models will increasingly enable intelligent, equitable healthcare delivery at the point of care—transforming emergency triage from a resource-intensive bottleneck into a rapid, accessible, and trustworthy clinical decision support function.

References

1. Alghamdi, M., Alfaraj, D., Buarish, L., et al. (2026). Can resident-performed POCUS and clinical predictors effectively risk-stratify emergency department patients with confirmed renal colic? *International Journal of Emergency Medicine*. https://doi.org/10.1186/s12245-026-01301-2
2. 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.
3. Gaudreau-Simard, M., et al. (2024). Point-of-care ultrasound for acute kidney injury: Diagnostic accuracy and clinical utility. *The Ultrasound Journal*, 16.
4. 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*, 1321–1330.
5. Karthik, M., & Dane, S. (2019). APTOS 2019 blindness detection. Kaggle.
6. Li, D., & Wang, J. (2021). FedMD: Heterogenous federated learning via model distillation. *arXiv preprint*.
7. Längkvist, M., et al. (2024). Deep learning for urinary stone detection: Challenges and opportunities. *Journal of Imaging Informatics in Medicine*, 37, 444–454.
8. Liu, X., et al. (2026). Optimizing AI workloads in edge data centers: A comparative study of hardware acceleration techniques. *IEEE Xplore*.
9. Priya, V. H. D., & Juliet, A. V. (2025). A deep learning-based multiple lung disease prediction and classification using CT-scan images. *SN Computer Science*, 6, 367.
10. Shingade, S. D., Chakkaravarthy, M., Karras, D., & Komal, M. (2025). Meta-heuristic optimized dual-attention deep network with depth-wise separability for high-precision renal tumor diagnosis. *Acta Scientiae*, 26(3), 231–243.
11. Shrestha, A. (2026). A controlled benchmark of edge AI accelerators for traffic object detection. *UTUPub*.
12. Yang, Y., Li, Q., & Zhang, C. (2025). CARE: A calibration-aware cross-institutional collaboration framework for medical image classification. *Pattern Recognition and Computer Vision*.

13. Zhang, Y., et al. (2026). Energy-per-pixel analysis on Edge-TPU-based aerial object detection on MCU-class devices. *Applied Sciences*, 16(15), 7803.

14. Zhao, J. S., Zhu, Z. H., & Jiang, Z. X. (2024). Fully automatic urinary stone detection system: Development and validation. *Journal of Imaging Informatics in Medicine*, 37, 444–454.

15. Zhou, J., et al. (2021). A multi-scale gated multi-head attention depthwise separable CNN model for recognizing COVID-19. *Scientific Reports*, 11, 18048.