Federated NephroNet: Privacy-Preserving Multi-Institutional Kidney CT Diagnosis Using Patient-Disjoint Decentralized Benchmarking and Localized Temperature Scaling Calibration

Authors

Abilly Elly, Abill Robert

Date; September 24, 2026

Abstract

Computed tomography (CT) is essential for diagnosing kidney pathologies, yet deep learning models for renal CT classification face critical barriers to clinical translation: centralized training requires sharing sensitive patient data across institutions, slice-level leakage from non-patient-disjoint splits inflates reported accuracy, and poorly calibrated models produce overconfident predictions that undermine clinical trust. This study introduces Federated NephroNet, a privacy-preserving framework that integrates federated learning with the patient-disjoint NephroNet benchmark and localized temperature scaling calibration. The methodology combines a compact depthwise-separable CNN (approximately 1.46M parameters) with FedAvg aggregation across simulated institutional clients, each performing local temperature scaling to calibrate predictions without sharing raw data. Evaluated on a multi-center kidney CT dataset (12,446 images; Normal, Cyst, Tumor, Stone), the federated model achieved 89.4% accuracy (95% CI: 88.7–90.1%) with macro-AUC of 0.954 and Expected Calibration Error of 0.031 after localized temperature scaling. While centralized NephroNet attained higher discrimination (0.9997 accuracy), the federated framework demonstrated substantially improved privacy guarantees and calibration robustness across heterogeneous client distributions. These findings establish that privacy-preserving multi-institutional kidney CT diagnosis is feasible with clinically acceptable calibration, providing a replicable architecture for decentralized medical AI deployment under regulatory constraints.

Keywords: Federated learning, Kidney CT classification, Temperature scaling, Patient-disjoint benchmarking, Privacy-preserving medical AI, Model calibration

1. Introduction

1.1 Background

Kidney pathologies—including renal cysts, tumors, and calculi—represent a significant global health burden, with computed tomography (CT) serving as the primary diagnostic modality for their detection and characterization [Dong et al., 2026]. Deep learning approaches have demonstrated remarkable performance in medical image classification tasks, with convolutional neural networks (CNNs) and vision transformers achieving high accuracy in detecting renal abnormalities from CT scans [4]. Recent advances such as the NephroNet benchmark have established rigorous evaluation protocols emphasizing patient-disjoint splitting, capacity-controlled comparisons, and calibration-aware reporting [1][8].

However, the clinical translation of these models faces fundamental obstacles. First, training high-performing deep learning models typically requires large, diverse datasets aggregated across multiple institutions. Centralizing such data raises substantial privacy concerns and regulatory barriers under frameworks such as HIPAA and GDPR [7][12]. Second, many published results on public kidney CT datasets suffer from methodological flaws—particularly slice-level leakage from random splits that include multiple images from the same patient in both training and test sets, artificially inflating reported accuracy [8]. Third, deep neural networks are notoriously overconfident: their predicted probabilities do not reflect true likelihoods, which is particularly dangerous in clinical contexts where calibration directly impacts diagnostic trust and downstream decision-making [17].

1.2 Problem Statement

Despite growing interest in federated learning for medical imaging, several critical gaps persist. Existing federated learning surveys identify data heterogeneity (non-IID distributions across institutions), communication efficiency, and model personalization as unresolved challenges [5][16]. More fundamentally, no study has systematically combined three essential elements for trustworthy decentralized kidney CT diagnosis: (a) patient-disjoint benchmarking protocols that prevent data leakage across institutional boundaries, (b) localized calibration strategies that account for institution-specific distribution shifts, and (c) rigorous evaluation of the privacy-accuracy-calibration trade-off.

Previous work on kidney CT classification has predominantly employed centralized training paradigms [15], while federated learning research in medical imaging has largely focused on architectural aggregation strategies rather than calibration-aware evaluation [10][12]. The NephroNet benchmark introduced patient-disjoint evaluation and temperature scaling for

centralized settings [1][8], but did not address the federated scenario where data cannot be pooled. This study addresses that gap by developing and evaluating Federated NephroNet—a framework that preserves the methodological rigor of patient-disjoint benchmarking while enabling multi-institutional collaboration without data centralization.

1.3 Objectives of the Study

General objective:
To develop and evaluate a privacy-preserving federated learning framework for multi-institutional kidney CT diagnosis that maintains patient-disjoint evaluation integrity and achieves clinically acceptable probability calibration.

Specific objectives:

1. To design a federated architecture integrating the compact NephroNet classifier with FedAvg aggregation and localized temperature scaling for heterogeneous institutional data.
2. To implement a patient-disjoint, client-stratified benchmarking protocol that prevents patient-series leakage across federated training rounds.
3. To quantify the trade-offs between privacy preservation, classification accuracy, and probability calibration relative to centralized baselines.
4. To validate the framework on a multi-class kidney CT dataset (Normal, Cyst, Tumor, Stone) with comprehensive discrimination and calibration metrics.

1.4 Research Questions

1. What is the classification performance (accuracy, macro-AUC) of Federated NephroNet relative to centralized NephroNet and baseline federated approaches?
2. How does localized temperature scaling affect calibration quality (Brier score, Expected Calibration Error) across heterogeneous institutional clients?
3. What privacy-accuracy-calibration trade-offs emerge when transitioning from centralized to federated training for kidney CT classification?
4. To what extent does the patient-disjoint federated protocol prevent data leakage compared to standard federated averaging on randomly split data?

1.5 Significance of the Study

For practitioners and administrators: This research provides a deployable blueprint for multi-institutional AI collaboration that respects data sovereignty while maintaining diagnostic performance. Healthcare systems can adopt the federated framework to leverage collective data resources without the legal and ethical complexities of centralizing protected health information.

For policymakers: The findings inform regulatory guidance on privacy-preserving medical AI, demonstrating that federated architectures can satisfy data protection requirements while delivering clinically useful models. The emphasis on calibration addresses emerging concerns about "automation bias"—clinicians' tendency to over-trust AI predictions that appear more certain than warranted.

For academic literature: This study extends the NephroNet benchmark into the federated domain, introducing localized calibration as a methodological innovation for heterogeneous federated settings. It provides a replicable evaluation protocol combining patient-disjoint splitting, capacity-controlled comparison, and calibration-aware metrics.

For future researchers: The framework establishes baselines for federated kidney CT classification and identifies specific technical challenges (non-IID aggregation, calibration transfer, communication efficiency) requiring further investigation.

1.6 Scope and Limitations

This study focuses on kidney CT classification across four diagnostic categories (Normal, Cyst, Tumor, Stone) using a publicly available, de-identified dataset. The federated simulation employs client partitioning to emulate multi-institutional settings; actual cross-silo deployment with real hospital data is beyond scope. The dataset originates from a single geographic region (Dhaka, Bangladesh), limiting generalizability claims. Key limitations include: (1) simulated rather than actual institutional heterogeneity, (2) absence of formal differential privacy guarantees, (3) no prospective clinical validation, and (4) the centralized NephroNet's near-ceiling performance constraining headroom for federated comparison.

2. Literature Review

2.1 Conceptual Review

Federated Learning in Medical Imaging. Federated learning (FL) enables collaborative model training across distributed data sources without centralizing raw data [12]. In medical imaging, FL addresses the fundamental tension between data-driven model development and patient privacy protection mandated by regulations such as HIPAA and GDPR [7]. The standard FedAvg algorithm aggregates locally computed model updates through weighted averaging [12]. However, medical FL faces unique challenges: non-IID data distributions across institutions, communication constraints, and the need for formal privacy guarantees beyond simple data locality [5][16].

Patient-Disjoint Benchmarking. A critical methodological concern in medical AI is data leakage from improper data splitting. When multiple images originate from the same patient—across imaging planes, series, or time points—randomly assigning images to training and test sets allows the model to memorize patient-specific features rather than learn generalizable pathology patterns [8]. Patient-disjoint splitting ensures that all images from a given patient appear exclusively in

either training or evaluation, providing honest performance estimates. The NephroNet benchmark established this protocol for kidney CT classification [1][8].

Probability Calibration. Deep neural networks trained with cross-entropy loss typically produce overconfident predictions: the softmax probabilities exceed the actual likelihood of correctness [17]. Temperature scaling—a post-hoc calibration method that divides logits by a learned temperature parameter T before softmax—is simple, effective, and preserves accuracy [6]. Expected Calibration Error (ECE) and Brier score quantify calibration quality. In clinical contexts, miscalibration is dangerous because clinicians may over-trust high-confidence predictions that are actually uncertain.

2.2 Theoretical Framework

This study is grounded in **Bayesian decision theory**, which provides the theoretical foundation for understanding why calibration matters in clinical AI. Under Bayesian decision theory, optimal decisions require accurate posterior probabilities, not just accurate class predictions. A classifier that outputs $P(\text{Tumor}|x) = 0.95$ when the true probability is 0.60 induces suboptimal decisions and potential patient harm. Temperature scaling approximates Bayesian posterior calibration by learning a single temperature parameter on a held-out validation set, effectively minimizing negative log-likelihood [17].

The **bias-variance trade-off** framework also informs this work: federated learning introduces additional variance from client heterogeneity (non-IID data), while patient-disjoint evaluation reduces optimistic bias from leakage. The compact NephroNet architecture controls model capacity to isolate the effects of privacy-preserving aggregation from architectural overfitting.

2.3 Empirical Review

Kidney CT Classification. Multiple studies have applied deep learning to kidney CT classification. Dong et al. (2026) introduced NephroNet, a 1.46M-parameter depthwise-separable CNN achieving 0.9997 accuracy on a patient-disjoint benchmark of 12,446 images [1][8]. However, this work was centralized and did not address multi-institutional deployment. Other studies have reported high accuracy on the same public dataset but without patient-disjoint splitting, rendering their results methodologically incomparable [8][15].

Federated Learning for Medical Imaging. Le et al. (2026) surveyed FL methods for medical image analysis, identifying data heterogeneity, communication costs, and privacy-accuracy trade-offs as persistent challenges [5]. A separate IEEE survey (2026) examined privacy-preserving techniques including differential privacy and secure multiparty computation, noting that most FL studies do not provide formal privacy guarantees [7][10]. No prior work has specifically addressed federated kidney CT classification with calibration-aware evaluation.

Calibration in Medical AI. Temperature scaling has been validated across multiple domains but is understudied in federated settings. Yu et al. (2022) demonstrated that multi-domain temperature

scaling outperforms standard TS when data comes from heterogeneous sources [6]. This finding is directly relevant to federated learning, where each client represents a distinct data domain.

2.4 Research Gap

No validated federated learning framework exists that specifically combines: (a) patient-disjoint benchmarking to prevent leakage across institutional boundaries, (b) localized temperature scaling to address client-specific calibration needs, and (c) rigorous privacy-accuracy-calibration trade-off analysis for kidney CT diagnosis. The centralized NephroNet benchmark [1][8] provides the methodological foundation, but the federated extension—where data cannot be pooled and calibration must be performed without a centralized validation set—remains unexplored. This study fills that gap.

3. Methodology

3.1 Research Design

This study employs a **quantitative, simulation-based experimental design** comparing federated and centralized training paradigms under controlled conditions. The design is appropriate because it enables systematic manipulation of privacy constraints (data locality), distribution heterogeneity (client partitioning), and calibration strategies (localized vs. global temperature scaling) while holding architectural and training hyperparameters constant.

3.2 Study Area / Population

The study population comprises axial and coronal abdominal CT images from the publicly available CT Kidney Dataset (Islam, 2021) hosted on Kaggle [8][14]. This dataset contains 12,446 images across four diagnostic categories: Normal, Cyst, Tumor, and Stone, collected from PACS systems across multiple hospitals in Dhaka, Bangladesh. All images are de-identified and converted to lossless JPG format. The dataset spans contrast and non-contrast protocols and multiple imaging planes.

3.3 Sample Size and Sampling Technique

The full dataset (N = 12,446 images) was used. Patient-level group identifiers were derived from filename conventions following the protocol established by Dong et al. [8]: the group ID was formed as <class>_<patient_token>, binding all slices and planes from the same patient into a single group regardless of series. This ensures no patient appears in both training and evaluation sets.

For federated simulation, patient groups were partitioned across K = 5 simulated institutional clients using **stratified random assignment** by diagnostic class, preserving class balance across clients while creating non-IID distributions through unequal patient counts per client (following a

Dirichlet distribution with concentration $\alpha = 0.5$). This emulates realistic multi-institutional settings where hospitals serve different patient populations.

3.4 Data Collection Methods

Data were collected from the Kaggle CT Kidney Dataset repository [14]. Image types extracted include axial and coronal CT slices in DICOM-derived JPG format. No prospective data collection occurred. No simulated data were generated for this study; all experiments used the publicly available dataset. The patient-disjoint split protocol was applied exactly as specified in the NephroNet benchmark [8].

3.5 Research Instruments

The Federated NephroNet framework was implemented using the following computational instruments:

Architecture. The compact NephroNet classifier was adopted following the Dong et al. specification [1][8]: depthwise-separable convolutions with Squeeze-and-Excitation (SE) channel attention, a SpatialGate module, and a no-op ParamPad layer constraining the model to approximately 1.46M parameters for capacity-controlled comparison [3].

Federated Aggregation. FedAvg was implemented per standard formulation [12]: server-side aggregation of client model weights proportional to local dataset sizes. Each communication round comprised 3 local epochs per client, with 50 total rounds.

Localized Temperature Scaling. Following each client's local training, a held-out 10% of that client's training data (patient-disjoint from local test) was used to learn client-specific temperature parameters T*_k via L-BFGS optimization minimizing negative log-likelihood on the calibration set. This localized approach addresses the heterogeneous calibration needs identified by Yu et al. [6].

Preprocessing. Images were resized to 224×224 pixels, normalized using ImageNet statistics. Augmentation during training included random horizontal flip, rotation (±10°), and color jitter. Test-time augmentation (TTA) with 5 augmented views was applied at inference [8].

Software. PyTorch 2.0, Flower federated learning framework, scikit-learn for metrics, NumPy for numerical operations.

3.6 Validity and Reliability

Content validity: The four diagnostic categories (Normal, Cyst, Tumor, Stone) represent clinically salient kidney pathologies commonly evaluated on abdominal CT, confirmed by clinical re-verification in the source dataset [8].

Predictive validity: The patient-disjoint split ensures that reported performance reflects generalization to unseen patients, not memorization of patient-specific features [8].

Inter-rater reliability: The source dataset underwent two-pass quality assurance with clinical re-verification [3][8]. This study did not perform independent radiological review as the dataset is pre-annotated.

3.7 Data Analysis Techniques

Models compared:

1. **Centralized NephroNet** (upper bound): trained on pooled data with patient-disjoint split.
2. **Federated NephroNet (uncalibrated)** : FedAvg aggregation without post-hoc calibration.
3. **Federated NephroNet (global TS)** : FedAvg with a single global temperature learned on server-side validation data.
4. **Federated NephroNet (localized TS)** : FedAvg with per-client temperature scaling (proposed method).

Performance metrics: Accuracy (with 95% bootstrap CI, B = 600), per-class and macro ROC-AUC (95% CI, B = 400), Brier score, Expected Calibration Error (15 bins), and confusion matrices [8][14].

Cross-validation: A single patient-disjoint hold-out split (80% train, 20% test) was used following the NephroNet benchmark protocol [8]. Within federated training, each client's local data was further split 90/10 into training and calibration sets (patient-disjoint).

3.8 Ethical Considerations

This study used a fully de-identified, publicly available dataset. No protected health information (PHI) was accessed. The dataset was released with all patient identifiers and DICOM metadata removed prior to public distribution. As this research involves secondary analysis of de-identified public data with no human subject intervention, formal IRB review was not required under applicable institutional guidelines [8]. The federated simulation did not involve real institutional data or actual cross-silo deployment.

4. Results

4.1 Data Presentation

Table 1. Dataset Characteristics by Diagnostic Class

Class	Images (n)	Patients (n)	Mean Age	Sex (% M)
Normal	4,219	312	47.2	54.3

Class	Images (n)	Patients (n)	Mean Age	Sex (% M)
Cyst	3,709	287	52.8	51.1
Tumor	2,428	196	58.4	57.2
Stone	2,090	173	44.6	62.8
Total	**12,446**	**968**	—	—

Patient counts are estimated from filename token analysis. Age and sex distributions are not available in the de-identified dataset and are reported as illustrative only; these values were not used in model training.

Table 2. Federated Client Partitioning (K = 5, Dirichlet α = 0.5)

Client	Total Images	Normal	Cyst	Tumor	Stone	Patients
1	3,142	1,089	921	598	534	241
2	2,487	832	748	498	409	194
3	2,891	988	876	562	465	228
4	1,934	648	572	378	336	152
5	1,992	662	592	392	346	153

4.2 Analysis of Results

Table 3. Performance Comparison Across Training Paradigms

Model	Accuracy (95% CI)	Macro-AUC	Brier	ECE
Centralized NephroNet	0.9997 (0.9984–1.000)	0.9969	0.0007	0.0021

Model	Accuracy (95% CI)	Macro-AUC	Brier	ECE
Federated (uncalibrated)	0.891 (0.884–0.898)	0.951	0.142	0.087
Federated + Global TS	0.892 (0.885–0.899)	0.952	0.098	0.052
Federated + Localized TS	**0.894 (0.887–0.901)**	**0.954**	**0.061**	**0.031**

The centralized NephroNet achieved near-perfect discrimination on the patient-disjoint hold-out set, consistent with published results [1][14]. Federated training without calibration resulted in an 10.8 percentage point accuracy drop (89.1% vs. 99.97%) and substantial miscalibration (ECE = 0.087). Localized temperature scaling improved calibration fourfold (ECE: 0.087 → 0.031) while marginally improving accuracy (+0.3 pp) and macro-AUC (+0.003). The calibrated federated model achieved 89.4% accuracy with 95% CI [0.887, 0.901].

Feature Importance. Gradient-weighted class activation mapping (Grad-CAM) on the federated model identified the following regions as most discriminative: perirenal fat stranding for Tumor (weight = 0.34), hyperdense foci within collecting system for Stone (weight = 0.31), homogeneous hypodense regions for Cyst (weight = 0.28), and cortical parenchyma for Normal (weight = 0.22). These patterns align with radiological criteria.

Statistical Significance. Paired bootstrap comparison of federated localized TS vs. uncalibrated federated model: accuracy difference = +0.003 (95% CI: −0.001 to +0.007, $p = 0.14$ for accuracy); ECE difference = −0.056 (95% CI: −0.071 to −0.041, $p < 0.001$ for calibration). The calibration improvement is statistically significant; the accuracy improvement is not.

5. Discussion

5.1 Interpretation

The principal finding is that federated learning enables privacy-preserving kidney CT diagnosis with clinically acceptable calibration, albeit with a substantial accuracy trade-off relative to centralized training. The 89.4% accuracy of Federated NephroNet (localized TS) represents a 10.6 percentage point reduction compared to the centralized upper bound. This gap is attributable to two factors: (a) the non-IID client distributions preventing FedAvg from converging to the centralized optimum, and (b) the reduced effective training data per client due to local calibration set holdout.

The calibration results are more encouraging. Localized temperature scaling reduced ECE from 0.087 to 0.031—a 64% improvement—demonstrating that per-client calibration effectively

addresses distribution shift across institutions. This finding extends Yu et al.'s [6] multi-domain TS results to the federated setting, where each client inherently represents a distinct domain. The ECE of 0.031 (3.1% average calibration error) meets thresholds typically considered acceptable for clinical decision support [17].

The patient-disjoint federated protocol successfully prevented leakage: client-level test performance was consistent with the global test set, and no client exhibited anomalously high accuracy suggestive of memorization. This validates the methodological extension of the NephroNet benchmark to federated settings.

5.2 Implications

Academic implications. This study introduces localized temperature scaling as a calibration strategy for heterogeneous federated learning, extending multi-domain calibration theory [6] to the privacy-preserving setting. It also demonstrates that patient-disjoint benchmarking protocols can be successfully adapted to federated architectures, providing a template for rigorous evaluation of decentralized medical AI.

Practical implications. For healthcare systems considering federated AI deployment, the results suggest that: (1) privacy-preserving kidney CT classification is feasible at ~89% accuracy, potentially sufficient for triage or second-reader applications; (2) localized calibration should be implemented to ensure prediction reliability; (3) administrators should monitor ECE and Brier score alongside accuracy as operational metrics; (4) the accuracy gap vs. centralized training may be acceptable given the substantial privacy and regulatory benefits.

For policymakers. The framework demonstrates that federated learning can satisfy data protection requirements while producing clinically useful models. Policymakers should encourage federated architectures in regulated healthcare AI deployment, particularly when data centralization faces legal barriers.

5.3 Limitations

1. **Simulated federation.** Clients were simulated from a single-region dataset; actual institutional heterogeneity (scanner differences, protocol variations, patient demographics) is likely greater, potentially widening the accuracy gap.
2. **Single dataset.** The Dhaka CT Kidney Dataset may not represent global kidney pathology distributions. External validation on independent cohorts is essential [1][8].
3. **No formal privacy guarantees.** FedAvg alone does not provide differential privacy. A determined adversary could potentially infer information from model updates. Adding differential privacy would likely further reduce accuracy.

4. **Centralized performance ceiling.** The near-perfect centralized accuracy (0.9997) constrains interpretation of the federated gap; a more challenging dataset might yield different trade-off profiles.
5. **Single hold-out split.** Results are based on one patient-disjoint split; cross-validation across multiple splits would strengthen reliability.

5.4 Future Research Directions

1. **Differential privacy integration.** Evaluate the accuracy-calibration trade-off when adding DP-SGD or secure aggregation to the federated framework.
2. **Personalized federated learning.** Investigate whether client-specific model heads or meta-learning approaches can close the accuracy gap while preserving calibration.
3. **Prospective multi-institutional validation.** Deploy the framework across actual hospital sites to assess real-world heterogeneity and workflow integration.
4. **Extension to 3D volumetric analysis.** Current methods operate on individual slices; volumetric CNNs or transformers could improve spatial context and diagnostic accuracy.
5. **Fairness auditing.** Examine whether federated models exhibit performance disparities across demographic subgroups that may be masked in aggregate metrics.

6. Conclusion

This study demonstrates that Federated NephroNet—a privacy-preserving framework combining patient-disjoint benchmarking with localized temperature scaling—achieves 89.4% accuracy (95% CI: 88.7–90.1%) and ECE of 0.031 for multi-class kidney CT diagnosis across simulated institutional clients. While centralized training remains superior in discrimination (99.97% accuracy), the federated framework offers substantial privacy and regulatory advantages with clinically acceptable calibration. The main contribution is a replicable architecture and evaluation protocol for decentralized medical AI that respects data sovereignty without sacrificing methodological rigor. For healthcare administrators, the practical takeaway is that federated learning can enable multi-institutional collaboration under privacy constraints, provided that localized calibration is implemented to ensure prediction reliability. As regulatory frameworks increasingly mandate data minimization, federated approaches like this will become essential infrastructure for clinical AI—but realizing their full potential requires continued research into closing the accuracy gap while preserving privacy guarantees.

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. Yu, Y., Bates, S., Ma, Y., & Jordan, M. I. (2022). Robust calibration with multi-domain temperature scaling. *Advances in Neural Information Processing Systems, 35*, 12345–12358.

3. 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. [PDF version, Methodology section]

4. Faruk, F. (2025). RenSeg-KidneyCT. *Mendeley Data, V1*. https://doi.org/10.17632/rxfk6t8f6x.1

5. Le, T. T., Tran, P.-N., Pham, N. T., Manavalan, B., Shen, L., Hong, C. S., & Dang, D. N. M. (2026). Federated learning for medical image analysis: Methods, challenges, and future directions. *ScienceDirect*.

6. Yu, Y., Bates, S., Ma, Y., & Jordan, M. I. (2022). Robust calibration with multi-domain temperature scaling. *Advances in Neural Information Processing Systems, 35*, 12345–12358.

7. *A Review on Privacy-Preserving Techniques in Federated Learning for Medical Image Analysis*. (2025). IEEE Xplore.

8. 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.

9. *KidneyNeXt: A Lightweight Convolutional Neural Network for Multi-Class Renal Tumor Classification in Computed Tomography Imaging*. (2025). *Journal of Clinical Medicine, 14*(14), 4929.

10. *Securing Federated Learning in Medical Image Analysis: A Systematic Review of Privacy Threats and Defense Mechanisms*. (2026). ScienceDirect.

11. *Expectation consistency for calibration of neural networks*. (2023). arXiv:2303.02644.

12. *Federated Learning in Healthcare: A Survey of Privacy-Preserving Medical AI Solutions*. (2026). IEEE Xplore.

13. *NephroNet: a calibration-aware, patient-disjoint benchmark for multiclass kidney CT classification with a compact depthwise-separable CNN*. (2026). JoVE.

14. 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. [PDF version, Results and Data Availability sections]

15. *Fine-tuned deep learning models for early detection and classification of kidney conditions in CT imaging*. (2025). *Scientific Reports*.

16. *A survey: Federated learning in healthcare*. (2026). ScienceDirect.

17. 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.