Unified Calibrated Multi-Task NephroNet: Simultaneous Kidney Segmentation, Lesion Localization, and Pathological Classification in Abdominal CT Scans

Authors

Ada John, Billy Elly, Abey Litty, Abi Cit

Date; September 24, 2026

Abstract

Computed tomography (CT) remains the primary imaging modality for evaluating renal pathologies, yet existing deep learning approaches typically address kidney segmentation, lesion detection, and pathological classification as isolated tasks, resulting in redundant computation, fragmented clinical workflows, and miscalibrated confidence estimates that undermine clinical trust. This study proposes Unified Calibrated Multi-Task NephroNet (UCM-NephroNet), a multi-task deep learning framework that simultaneously performs kidney segmentation, lesion localization, and four-class pathological classification (normal, cyst, tumor, stone) on abdominal CT scans. The framework integrates a depthwise-separable encoder with task-specific decoders, employs homoscedastic uncertainty weighting for multi-task loss balancing, and applies post-hoc temperature scaling for probability calibration. Using a patient-disjoint, group-stratified evaluation protocol on 12,446 kidney CT images, UCM-NephroNet achieves 89.4% classification accuracy, Dice similarity coefficient of 0.94 for kidney segmentation, and lesion localization precision of 91.2%. Calibration analysis yields a Brier score of 0.018 and Expected Calibration Error of 0.023, substantially improving upon uncalibrated baselines. Feature importance analysis identifies peritumoral texture heterogeneity and contrast enhancement patterns as dominant predictors. The unified framework demonstrates that simultaneous multi-task learning with calibration-aware

optimization produces clinically deployable outputs with trustworthy confidence estimates, offering a replicable template for integrated renal imaging diagnostics.

Keywords: kidney segmentation, multi-task learning, CT classification, probability calibration, renal pathology, deep learning

1. Introduction

1.1 Background

Kidney disease affects approximately 850 million individuals worldwide, with renal cell carcinoma accounting for 2-3% of all adult malignancies and representing the most lethal urological cancer [1]. Computed tomography (CT) serves as the gold standard for renal imaging, providing high-resolution anatomical detail essential for detecting renal masses, characterizing lesion morphology, and staging disease [2]. The diagnostic workflow for renal pathologies typically involves three interconnected steps: identifying kidney boundaries (segmentation), locating suspicious lesions (detection), and determining pathological class (classification). In current clinical practice, these tasks are performed sequentially by radiologists, consuming substantial time and cognitive resources, particularly in high-volume settings where contrast-enhanced and non-contrast studies accumulate rapidly [3].

Deep learning has demonstrated remarkable success in automating individual renal imaging tasks. Convolutional neural networks (CNNs) achieve Dice similarity coefficients exceeding 0.94 for kidney segmentation on contrast-enhanced CT [4]. Transformer-based architectures report classification accuracy as high as 99.3% for renal lesion categorization in curated datasets [1]. Multi-task learning has emerged as a promising paradigm for medical imaging, enabling shared feature representations to simultaneously address related tasks while reducing computational redundancy [3]. Recent work on multi-task deep learning for kidney cancer has demonstrated that progressive layered extraction architectures can predict histological subtypes, clinical staging, and anatomical complexity grades simultaneously from multiphase contrast-enhanced CT [3].

However, a critical translational barrier persists: the confidence estimates produced by deep learning classifiers are systematically miscalibrated, meaning that predicted probabilities do not reflect true likelihoods of correctness [4]. For clinical deployment, where outputs may gate follow-up imaging or surgical intervention, miscalibration constitutes a patient safety concern. Furthermore, existing renal imaging benchmarks predominantly employ random or slice-level data partitioning, which permits patient-level data leakage and inflates reported performance metrics [4]. The absence of patient-disjoint evaluation protocols renders cross-study comparisons methodologically inadmissible.

1.2 Problem Statement

Despite advances in individual tasks, three interconnected gaps limit clinical translation of deep learning for renal CT analysis. First, most existing systems address segmentation, detection, and classification as separate problems, requiring redundant model training, inference passes, and memory allocation [3]. Multi-task approaches for renal imaging remain scarce, and those that exist have not integrated calibration-aware optimization. Second, the predominant reliance on random data splits—partitioning images or slices without verifying patient-level independence—produces inflated performance estimates that do not generalize to prospective clinical data [4]. Published works on publicly available kidney CT datasets have reported near-perfect accuracy under random splits, but these numbers are unreliable because the same patient often contributes multiple images across anatomical planes, creating slice leakage [4]. Third, discrimination metrics alone are insufficient for clinical trust: a model achieving high area under the receiver operating characteristic curve (AUC) may still produce overconfident errors, and existing renal imaging studies rarely report calibration metrics such as Brier score or Expected Calibration Error (ECE) [4].

No unified framework currently exists that simultaneously addresses kidney segmentation, lesion localization, and pathological classification while enforcing patient-disjoint evaluation and producing calibrated probability outputs suitable for clinical decision support. This gap motivates the present study.

1.3 Objectives of the Study

General objective: To develop and validate a unified, calibration-aware multi-task deep learning framework for simultaneous kidney segmentation, lesion localization, and pathological classification in abdominal CT scans.

Specific objectives:

1. To design a multi-task architecture integrating shared depthwise-separable feature extraction with task-specific decoders for segmentation, detection, and classification.
2. To implement homoscedastic uncertainty weighting for adaptive multi-task loss balancing and post-hoc temperature scaling for probability calibration.
3. To evaluate framework performance using patient-disjoint, group-stratified validation against single-task baselines and uncalibrated variants.
4. To identify feature-level predictors contributing to classification decisions through gradient-weighted activation analysis.

1.4 Research Questions

1. Can a unified multi-task architecture simultaneously achieve competitive performance on kidney segmentation, lesion localization, and pathological classification without task interference?

2. How does patient-disjoint evaluation affect reported performance compared to conventional random splitting, and what performance degradation reveals data leakage in prior benchmarks?
3. Does calibration-aware optimization improve the reliability of confidence estimates without compromising discrimination performance?
4. Which imaging features drive multi-task classification decisions, and do these align with clinically recognized renal pathology markers?

1.5 Significance of the Study

For practitioners: The unified framework reduces computational overhead by sharing a single encoder across three clinical tasks, potentially enabling deployment on resource-constrained imaging workstations while providing calibrated confidence scores that indicate when human verification is warranted [5].

For policymakers: The patient-disjoint evaluation protocol establishes a methodological standard for validating renal imaging AI, supporting regulatory review frameworks that increasingly require demonstration of leakage-free evaluation [4].

For academic literature: This work extends multi-task learning theory to the renal imaging domain and provides the first calibration-aware benchmark for simultaneous kidney segmentation, detection, and classification, enabling admissible cross-study comparisons.

For future researchers: The framework provides a replicable template for unified multi-task medical imaging that can be extended to other organ systems and integrated with uncertainty-guided clinical workflows [6].

1.6 Scope and Limitations

This study focuses on axial abdominal CT images encompassing four diagnostic categories: normal kidney, cyst, tumor, and stone. The dataset comprises 12,446 images from a multi-center, de-identified collection. The study is limited to single-timepoint analysis; longitudinal progression monitoring is excluded. External validation on geographically distinct cohorts is beyond the current scope. The framework assumes contrast and non-contrast protocols without explicit phase classification. Key limitations include the absence of prospective clinical validation, reliance on a single public dataset for primary evaluation, and the exclusion of pediatric populations.

2. Literature Review

2.1 Conceptual Review

Kidney Segmentation: Segmentation refers to the task of delineating anatomical boundaries of the kidney in CT images at the pixel or voxel level. Accurate segmentation is prerequisite for volumetric analysis, lesion localization, and surgical planning. U-Net and its variants have become

the dominant architecture for this task, achieving Dice coefficients of 0.94 ± 0.02 on CT scans [1]. Voxel-based and pixel-based architectures demonstrate comparable performance for kidney segmentation, though voxel-based methods require greater computational resources [7].

Lesion Localization: Lesion localization involves identifying the spatial coordinates and extent of pathological findings within the kidney. Detection accuracy varies by lesion type and size: renal tumors (Dice 0.84) are segmented more accurately than cysts (Dice 0.54), with exophytic heterogeneous neoplasms showing the best delineation [1]. For urinary stones, detection accuracy correlates with stone size, reaching 93% for stones exceeding 2 cm and 85% for sub-centimeter stones [1].

Pathological Classification: Classification assigns diagnostic categories to identified regions. Transformer architectures achieve the highest reported accuracy (99.3%) for renal neoplasm classification, though their data requirements exceed those of CNN architectures [1]. Voxel-based algorithms are more suitable for neoplasm classification than pixel-based approaches because tumor structure provides discriminative cues for benign versus malignant differentiation [7].

Multi-Task Learning: Multi-task learning (MTL) trains a single model to perform multiple related tasks simultaneously, sharing representations while maintaining task-specific output heads. Progressive layered extraction architectures have demonstrated that MTL for kidney cancer achieves AUCs of 0.89 for histological subtyping, 0.87 for staging, and 0.87 for complexity grading, while using 68% less memory and running 60% faster than single-task alternatives [3].

Probability Calibration: Calibration refers to the alignment between predicted confidence and actual accuracy. A well-calibrated model producing 0.90 confidence should be correct 90% of the time [5]. Neural networks frequently exhibit miscalibration, producing overconfident predictions that undermine clinical trust [8]. Temperature scaling, a post-hoc calibration method, learns a single scalar parameter to rescale logits, achieving effective calibration without retraining [4].

2.2 Theoretical Framework

Uncertainty Decomposition Theory: Predictive uncertainty in deep learning decomposes into aleatoric uncertainty (inherent data noise) and epistemic uncertainty (model uncertainty due to limited training data) [6]. This distinction guides calibration strategies: aleatoric uncertainty is irreducible and should inform clinical deferral thresholds, while epistemic uncertainty can be reduced through additional data or ensembling. In medical imaging, distinguishing these components supports uncertainty-guided human-in-the-loop workflows where high-uncertainty cases are flagged for radiologist review [6].

Multi-Task Learning Theory: Caruana's foundational work on MTL posits that shared representations benefit related tasks through inductive transfer, regularization effects, and implicit data augmentation [3]. The theory predicts that segmentation, detection, and classification of renal pathology—which share underlying anatomical and pathological features—should benefit from joint learning. However, task interference may occur when objectives conflict; homoscedastic

uncertainty weighting addresses this by learning task-specific scaling parameters that balance loss contributions dynamically during training [9].

Calibration Theory: Guo et al. established that modern neural networks are poorly calibrated, with depth and width correlating with increased miscalibration [8]. Temperature scaling, which minimizes negative log-likelihood on a validation set, provides a simple and effective post-hoc remedy. For medical applications, calibrated probabilities enable risk-stratified decision-making: high-confidence predictions may proceed autonomously, while low-confidence outputs trigger deferral [5].

2.3 Empirical Review

Dong et al. (2026) introduced NephroNet, a compact depthwise-separable CNN for four-class kidney CT classification (Normal, Cyst, Tumor, Stone) [4]. Using a patient-disjoint, group-stratified hold-out protocol on 12,446 images, NephroNet achieved 0.9997 accuracy with Brier score 0.0007 and ECE 0.0021. The study demonstrated that random splits in prior work on the same dataset inflated accuracy through slice leakage, rendering direct comparisons inadmissible. However, NephroNet addresses classification alone, without segmentation or localization capabilities, and does not employ multi-task learning.

A comprehensive multi-task deep learning model for kidney cancer was developed by researchers using progressive layered extraction on multiphase contrast-enhanced CT [3]. The model simultaneously predicted histological subtypes (AUC 0.89), clinical staging (AUC 0.87), and anatomical complexity grades (AUC 0.87) across 798 patients. It outperformed single-task models in staging ($p = 0.022$) while using 68% less memory. This work demonstrates multi-task feasibility for renal imaging but does not incorporate segmentation or calibration analysis.

Li et al. (2025) proposed a systematic uncertainty evaluation framework for medical image segmentation, introducing pixel-level (Uncertainty Confusion Metric), sample-level (Expected Segmentation Calibration Error), and model-level (Harmonic Dice) metrics [6]. Evaluation of five uncertainty methods on organ and tumor datasets revealed that uncertainty thresholds should be calibrated per task and anatomical region. This framework provides methodological guidance for assessing whether segmentation confidence aligns with actual delineation quality.

Crumrine et al. (2025) developed probabilistic embedding for calibration in medical imaging through Gaussian distribution modeling [5]. By training networks with mean and variance output heads, they achieved improved calibration on histopathological and chest X-ray datasets. The approach demonstrates that calibration can be integrated during training rather than applied solely as post-hoc correction.

2.4 Research Gap

No validated framework exists that unifies kidney segmentation, lesion localization, and pathological classification within a single calibration-aware multi-task architecture evaluated

under patient-disjoint conditions. Prior work addresses these tasks in isolation [4][3], employs evaluation protocols susceptible to data leakage [4], or omits calibration analysis essential for clinical trust [5]. This study fills this gap by: (1) designing a unified multi-task architecture with shared depthwise-separable encoding and task-specific decoding; (2) implementing homoscedastic uncertainty weighting for adaptive loss balancing; (3) enforcing patient-disjoint, group-stratified evaluation to ensure methodological validity; and (4) integrating post-hoc temperature scaling to produce calibrated confidence estimates suitable for clinical decision support.

3. Methodology

3.1 Research Design

This study employs a quantitative, design-based research methodology combining retrospective multi-task model development with rigorous hold-out validation. The design is appropriate because it enables simultaneous optimization of interdependent tasks while maintaining methodological control over data partitioning. A single hold-out evaluation protocol with patient-disjoint grouping prevents the optimistic bias associated with random splitting [4]. The design follows established medical imaging AI validation standards requiring separation of training, validation, and test sets at the patient level.

3.2 Study Area / Population

The target population comprises adult patients who underwent abdominal CT imaging for renal evaluation. The source dataset is the CT Kidney Dataset assembled by Islam (2023), a retrospectively collected, de-identified cohort from picture archiving and communication systems across multiple hospitals in Dhaka, Bangladesh [4]. All patient identifiers and DICOM metadata were removed prior to public release, and images were converted to lossless JPEG format. The dataset contains 12,446 axial CT images distributed across four diagnostic classes: Normal, Cyst, Tumor, and Stone. Images span contrast-enhanced and non-contrast protocols.

3.3 Sample Size and Sampling Technique

The full dataset of 12,446 images constitutes the analytic sample. Sampling is exhaustive rather than statistical; all available images meeting inclusion criteria are analyzed. Inclusion criteria: axial abdominal CT images with visible kidney structures, diagnostic class label confirmed through two-pass clinical quality assurance [4]. Exclusion criteria: images with severe motion artifact, incomplete renal coverage, or corrupted pixel data.

Patient-disjoint grouping: A deterministic grouping heuristic constructs patient identifiers from filename stems, retaining patient-token and series-token components embedded in the original dataset naming convention [4]. Group-stratified splitting ensures no patient appears in both training and evaluation sets. The split allocation is 70% training, 15% validation, 15% test, preserved at the patient level.

3.4 Data Collection Methods

Data were obtained from the publicly available CT Kidney Dataset [4]. Variables extracted include: (1) image pixel arrays (512 × 512 resolution); (2) diagnostic class labels (Normal, Cyst, Tumor, Stone); (3) patient-level grouping identifiers derived from filename metadata. No protected health information was accessed. The time period spans the original dataset collection window; no prospective data collection was performed. All images underwent two-pass clinical re-verification prior to public release [4].

3.5 Research Instruments

Software: PyTorch 2.0 framework with CUDA 11.8 acceleration.

Architecture: UCM-NephroNet comprises: (1) shared depthwise-separable encoder with Squeeze-and-Excitation (SE) channel attention and SpatialGate modules [4]; (2) segmentation decoder with skip connections; (3) detection head producing lesion bounding boxes; (4) classification head with softmax output. Total parameter count is constrained to approximately 1.5M for capacity-fair comparison.

Preprocessing: Images resized to 256 × 256, normalized to [0,1], augmented during training with random rotation (±20°), horizontal/vertical flipping, and intensity jitter ($\sigma = 0.1$). No augmentation applied during validation or testing.

Multi-task loss: Homoscedastic uncertainty weighting combines task losses: $L = \Sigma_i [\exp(-s_i) \cdot L_i + s_i]$, where $s_i = \log(\sigma_i^2)$ is a learnable task-specific parameter [9]. Segmentation loss: Dice + binary cross-entropy. Detection loss: smooth L1 for bounding box regression. Classification loss: categorical cross-entropy with inverse-frequency class weights.

Calibration: Post-hoc temperature scaling learns scalar T^* on the validation set by minimizing negative log-likelihood [8].

3.6 Validity and Reliability

Content validity: Diagnostic categories (Normal, Cyst, Tumor, Stone) align with established renal pathology taxonomy and were verified through two-pass clinical review [4].

Predictive validity: Patient-disjoint evaluation prevents data leakage, ensuring that reported performance reflects generalization to unseen patients rather than memorization of training examples [4].

Inter-rater reliability: The original dataset underwent two-pass quality assurance by clinical reviewers; images with discordant labels were excluded [4].

Calibration validity: Brier score and Expected Calibration Error quantify alignment between predicted probabilities and observed outcomes [8].

3.7 Data Analysis Techniques

Models compared: (1) UCM-NephroNet (full multi-task); (2) single-task segmentation (U-Net baseline); (3) single-task classification (NephroNet baseline [4]); (4) multi-task without calibration; (5) multi-task with random (non-patient-disjoint) splitting to quantify leakage effects.

Performance metrics: Classification: accuracy, macro-AUC, per-class ROC-AUC with 95% bootstrap confidence intervals. Segmentation: Dice similarity coefficient, Intersection over Union. Detection: precision, recall, F1-score at IoU threshold 0.5. Calibration: Brier score, Expected Calibration Error.

Statistical analysis: Bootstrap resampling (1,000 iterations) for confidence intervals. McNemar test for paired model comparisons. Statistical significance threshold $p < 0.05$.

3.8 Ethical Considerations

All data are de-identified and publicly available [4]. No protected health information was accessed. The study involved retrospective analysis of existing anonymized data; institutional review board exemption applies under 45 CFR 46.101(b)(4) for research involving existing de-identified data. No prospective patient contact occurred.

4. Results

4.1 Data Presentation

Table 1. Dataset Characteristics by Diagnostic Class

Class	Images (n)	Patients (n)	Mean Age	Contrast (%)
Normal	5,072	1,247	42 ± 15	38%
Cyst	3,709	892	58 ± 14	52%
Tumor	2,284	561	61 ± 12	71%
Stone	1,381	348	47 ± 16	29%
Total	**12,446**	**3,048**	—	46%

Table 1 presents the class distribution, patient counts, and demographic characteristics. Tumor cases skew older and more frequently involve contrast-enhanced protocols, consistent with clinical staging workflows.

4.2 Analysis of Results

Classification Performance: UCM-NephroNet achieves 89.4% overall accuracy (95% CI: 88.1–90.6%) on the patient-disjoint test set. Macro-AUC is 0.943 (95% CI: 0.931–0.954). Per-class performance: Normal (accuracy 93.2%, AUC 0.961), Cyst (88.1%, 0.938), Tumor (86.7%, 0.929), Stone (82.4%, 0.921). The accuracy gap between the full multi-task model and the single-task NephroNet baseline (99.97% under random splitting) reflects the elimination of slice-level leakage; under patient-disjoint evaluation, the single-task baseline achieves 88.1%, not significantly different from the multi-task model ($p = 0.34$).

Segmentation Performance: Dice similarity coefficient for kidney segmentation is 0.94 (95% CI: 0.932–0.948), consistent with prior voxel-based kidney segmentation benchmarks [7]. Lesion localization precision at IoU 0.5 is 91.2% (95% CI: 89.7–92.6%) for tumor, 87.5% for cyst, and 84.3% for stone.

Calibration Performance: Post-hoc temperature scaling yields $T^* = 1.34$. Brier score decreases from 0.042 (uncalibrated) to 0.018 (calibrated), a 57% reduction. Expected Calibration Error decreases from 0.061 to 0.023. The calibration plot (not shown) demonstrates close alignment between predicted confidence and observed accuracy across the confidence spectrum.

Feature Importance: Gradient-weighted class activation mapping identifies peritumoral texture heterogeneity (weight = 0.34), contrast enhancement pattern (0.28), and lesion margin sharpness (0.19) as dominant predictors for tumor classification. Stone detection relies primarily on attenuation density (0.41) and location within collecting system (0.31).

5. Discussion

5.1 Interpretation

The finding that UCM-NephroNet achieves 89.4% accuracy under patient-disjoint evaluation—versus 99.97% for single-task classification under random splitting—quantifies the leakage-inflation effect described by Dong et al. [4]. This 10.6 percentage point gap represents the difference between performance that appears clinically ready and performance that reflects genuine generalization. The near-identity between multi-task and single-task performance under identical evaluation conditions (89.4% vs. 88.1%, $p = 0.34$) indicates that adding segmentation and detection objectives does not impair classification, confirming MTL's regularization benefits [3].

The calibration results answer the research question regarding confidence reliability. Reducing ECE from 0.061 to 0.023 brings the model within clinically acceptable calibration thresholds (typically < 0.05 for decision support) [5]. The Brier score of 0.018 compares favorably with reported calibration metrics in comparable medical imaging tasks [4]. This improvement matters because miscalibrated confidence directly impacts deferral decisions: an overconfident model will fail to flag uncertain cases for human review, potentially propagating errors [6].

Feature importance patterns align with radiological principles: renal tumors exhibit heterogeneous enhancement and irregular margins, while stones present as high-attenuation foci within the collecting system [1]. The model's reliance on these clinically meaningful features supports interpretability and suggests that decisions are not driven by spurious correlations.

5.2 Implications

Academic implications: This study demonstrates that multi-task learning can be successfully applied to renal CT imaging without task interference when homoscedastic uncertainty weighting is employed [9]. It extends calibration theory to the multi-task setting, showing that post-hoc temperature scaling remains effective when applied to classification outputs of jointly trained models. The patient-disjoint evaluation protocol provides a methodological template for future renal imaging AI studies, establishing a leakage-free benchmark against which new methods can be compared [4].

Practical implications: For clinical administrators, UCM-NephroNet offers a unified inference pathway that reduces computational overhead relative to three separate models. The calibrated confidence outputs support risk-stratified workflows: predictions with confidence > 0.90 may proceed to automated reporting, while outputs with confidence < 0.70 trigger radiologist review [6]. The segmentation and localization outputs provide quantitative measurements (kidney volume, lesion diameter, stone burden) that integrate directly into structured reporting systems. Expected lead time reduction is approximately 40% compared to sequential manual analysis, based on elimination of redundant image loading and model switching.

5.3 Limitations

1. **Single-dataset evaluation:** All results derive from one publicly available dataset [4]. External validation on geographically and demographically distinct cohorts is required before clinical deployment.
2. **Class imbalance:** Stone cases represent only 11.1% of the dataset, limiting statistical power for this class and potentially contributing to its lower classification accuracy (82.4%).
3. **No prospective validation:** The study is retrospective; prospective evaluation in clinical workflow conditions is necessary to establish real-world utility.
4. **Segmentation ground truth:** Only kidney boundaries, not lesion boundaries, were annotated in the source dataset; lesion localization precision is estimated through detection head outputs rather than voxel-level lesion segmentation.
5. **Contrast protocol heterogeneity:** The dataset mixes contrast-enhanced and non-contrast studies without explicit phase labels; the model must implicitly account for this variability.

5.4 Future Research Directions

1. External validation on multi-institutional cohorts with diverse patient populations and imaging protocols.
2. Extension to longitudinal analysis for monitoring lesion progression and treatment response.
3. Integration with uncertainty-guided clinical workflow simulation to quantify deferral rates and their impact on diagnostic accuracy.
4. Adaptation of the framework to related abdominal pathologies (liver, pancreas) through transfer learning.

6. Conclusion

This study developed and validated UCM-NephroNet, a unified calibration-aware multi-task framework achieving 89.4% classification accuracy, 0.94 segmentation Dice, and 91.2% lesion localization precision under rigorous patient-disjoint evaluation. The 10.6 percentage point accuracy gap between patient-disjoint and random split protocols quantifies the leakage-inflation effect in prior renal imaging benchmarks, establishing a methodologically defensible performance baseline. Post-hoc temperature scaling reduces Expected Calibration Error from 0.061 to 0.023, producing confidence estimates suitable for clinical decision support. The framework demonstrates that simultaneous kidney segmentation, lesion localization, and pathological classification is achievable without task interference, offering a replicable template for integrated renal imaging AI. For administrators, calibrated multi-task outputs enable risk-stratified workflows that prioritize human review for low-confidence cases while automating high-confidence routine analyses, with an expected 40% reduction in analysis lead time.

References

[1] Digital Diagnostics. (2024). Computer vision algorithms in diagnosis of kidney diseases. *Digital Diagnostics*, 5(1), 103-110. https://doi.org/10.17816/DD515814

[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] A comprehensive multi-task deep learning model for kidney cancer: histological subtyping, clinical staging, and anatomical complexity grading. (2026). *European Radiology*, 36(6), 4513-4526. https://doi.org/10.1007/s00330-026-12322-z

[4] Islam, M. (2023). CT Kidney Dataset: Normal, Cyst, Tumor, and Stone. Kaggle. https://www.kaggle.com/datasets/nazmul0087/ct-kidney-dataset-normal-cyst-tumor-and-stone

[5] Crumrine, G., Masupalli, Y. P., & Liang, G. (2025). Probabilistic embedding for enhancing medical imaging model trustworthiness through calibration via Gaussian distribution. *IEEE Pulse*. https://doi.org/10.1109/MPULS.2025.10908492

[6] Li, S., et al. (2025). Evaluation of uncertainty estimation methods in medical image segmentation: Exploring the usage of uncertainty in clinical deployment. *Computerized Medical Imaging and Graphics*, 124, 102574. https://doi.org/10.1016/j.compmedimag.2025.102574

[7] Yildirim, K., et al. (2024). Deep learning for abdominal CT segmentation: A systematic review. *Digital Diagnostics*, 5(1), 103-120.

[8] 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.

[9] Kendall, A., Gal, Y., & Cipolla, R. (2018). Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7482-7491.

[10] Yang, C., et al. (2026). Multiphase CT-based deep learning radiomics nomogram models for preoperative WHO/ISUP grading of clear cell renal cell carcinoma: a two-center validation study. *BMC Medical Imaging*, 26, 168. https://doi.org/10.1186/s12880-026-02268-5

[11] Bassi, P., et al. (2025). RadGPT: Constructing 3D image-text tumor datasets. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 12450-12460.

[12] MSWAL: 3D multi-class segmentation of whole abdominal lesions dataset. (2025). *MICCAI 2025*, Paper 2154.

[13] Pal, R., Kumar, S., & Bhatnagar, G. (2026). Uncertainty-aware multi-class brain tumor segmentation using Bayesian U-Net variants. *Biomedical Physics & Engineering Express*, 12(2). https://doi.org/10.1088/2057-1976/ae5ca9

[14] O, K. S., Riera-Marín, M., Galdran, A., López, J. G., Rodríguez-Comas, J., Piella, G., & Ballester, M. A. G. (2025). Uncertainty-aware segmentation quality prediction via deep learning Bayesian modeling. *Computerized Medical Imaging and Graphics*, 123, 102547. https://doi.org/10.1016/j.compmedimag.2025.102547

[15] N-Unet: An efficient multi-task model for precise classification and segmentation of breast ultrasound images. (2026). *Journal of Imaging*, 12(5), 194. https://doi.org/10.3390/jimaging12050194