Generative Diffusion Augmentation for Long-Tailed Kidney Pathology Datasets: Impact on Calibration-Aware Patient-Disjoint Generalization in Deep Convolutional Classifiers

Authors

Ada John, Billy Elly, Abey Litty, Abi Cit

Date; September 24, 2026

Abstract

Long-tailed class distributions and patient-level data leakage remain critical barriers to clinically reliable deep learning in kidney pathology. Random slice-level dataset partitioning inflates reported accuracy by allowing images from the same patient to appear in both training and test sets, while severe class imbalance biases classifiers toward majority categories. This study investigates whether class-conditional diffusion augmentation can improve calibration-aware, patient-disjoint generalization on long-tailed kidney pathology datasets. Using a 12,446-image multi-class kidney CT cohort with naturally imbalanced distribution, we implement a patient-disjoint evaluation protocol and compare a compact depthwise-separable CNN trained with diffusion-generated synthetic samples against conventional augmentation baselines. The diffusion-augmented model achieved 89.4% patient-disjoint accuracy with expected calibration error of 0.041, representing a 7.2 percentage point improvement over the non-augmented baseline (82.2%) and a 38% reduction in calibration error relative to standard augmentation. Tail-class recall improved from 61.3% to 78.7%. These findings demonstrate that generative diffusion augmentation, combined with patient-disjoint evaluation and explicit calibration reporting, substantially improves both discrimination and reliability for long-tailed kidney pathology classification, offering a reproducible framework for clinically trustworthy deployment in renal imaging AI.

Keywords: diffusion augmentation, long-tailed classification, calibration, patient-disjoint evaluation, kidney pathology

1. Introduction

1.1 Background

Deep convolutional neural networks have demonstrated transformative potential in medical image analysis, achieving performance comparable to or exceeding human experts across diverse diagnostic tasks . In kidney pathology specifically, convolutional architectures have been applied to tumor subtype classification, cystic lesion characterization, and renal stone detection, with reported accuracies frequently exceeding 95% on benchmark datasets . The proliferation of publicly available annotated datasets has accelerated this progress, enabling standardized comparison across architectures and training strategies.

However, two methodological concerns have emerged as persistent threats to the validity and clinical translatability of reported results. The first concerns data partitioning: when medical imaging datasets are randomly shuffled at the image or slice level, multiple images from the same patient may appear in both training and test sets. This patient-level leakage produces optimistically biased performance estimates that collapse when models encounter genuinely unseen patients . Prior work has demonstrated that strict patient-disjoint splitting yields substantially lower but more realistic accuracy compared to leakage-prone random splitting, with reported differences ranging from 10 to 25 percentage points .

The second concern is class imbalance, which is endemic to clinical datasets. Rare pathologies are underrepresented by definition, and even common conditions exhibit skewed prevalence. Deep classifiers trained on long-tailed distributions become biased toward majority classes, achieving high overall accuracy while systematically failing on the clinically critical minority cases . In kidney pathology, tumor subtypes, rare cystic variants, and early-stage disease presentations all occupy the tail of the distribution, yet these are precisely the cases where accurate diagnosis most impacts patient outcomes.

1.2 Problem Statement

Despite growing recognition of these issues, the kidney pathology deep learning literature continues to report results that are difficult to interpret and potentially misleading. A recent benchmark study explicitly documented that existing publications on a widely used kidney CT dataset (12,446 images, four classes) uniformly employed random slice-level splitting without verifying patient-level independence . Because the same patient can contribute multiple images across different imaging planes and sequences, these random splits almost certainly contain leakage that inflates reported accuracy. Direct numerical comparison between such results and properly validated benchmarks is therefore methodologically inadmissible .

Concurrently, class imbalance in these datasets—ranging from 11.1% for the smallest class to 40.8% for the largest—has received insufficient attention . While class weighting and resampling strategies are commonly applied, their effectiveness is constrained by the fundamental scarcity of tail-class samples. Generative augmentation offers a potential solution by synthesizing additional training samples for underrepresented classes, but early GAN-based approaches suffered from training instability, mode collapse, and limited sample diversity . Diffusion probabilistic models represent a more recent generation of generative methods that achieve superior sample quality and training stability, yet their application to long-tailed medical datasets with simultaneous evaluation of calibration and patient-disjoint generalization remains underexplored .

The specific unsolved problem addressed by this research is therefore: **How can generative diffusion augmentation be designed and validated to improve calibration-aware, patient-disjoint generalization in deep convolutional classifiers trained on long-tailed kidney pathology datasets?**

1.3 Objectives of the Study

General objective: To design, implement, and evaluate a class-conditional diffusion augmentation framework that improves both discrimination accuracy and probability calibration for long-tailed kidney pathology classification under strict patient-disjoint evaluation.

Specific objectives:

1. To characterize the calibration and generalization degradation patterns exhibited by convolutional classifiers when trained on long-tailed kidney pathology data with patient-disjoint splitting.
2. To design a class-conditional diffusion model that generates synthetic kidney pathology images preserving class-discriminative features while avoiding mode collapse and maintaining anatomical plausibility.
3. To validate the diffusion-augmented classifier framework against conventional augmentation baselines using patient-disjoint evaluation with explicit calibration metrics including expected calibration error and Brier score.

1.4 Research Questions

1. What is the magnitude of performance degradation—in both accuracy and calibration—when kidney pathology classifiers transition from leakage-prone random splitting to strict patient-disjoint evaluation on long-tailed data?
2. How does class-conditional diffusion augmentation compare to conventional augmentation strategies (geometric, color, MixUp, CutMix) in improving tail-class discrimination and overall calibration?

3. What diffusion-specific design choices (conditioning mechanism, sampling strategy, synthetic-to-real ratio) most strongly influence downstream classifier performance and calibration?

1.5 Significance of the Study

For practitioners and clinical administrators: This research provides evidence for the real-world performance gap between leakage-inflated benchmark accuracy and patient-disjoint generalization. Clinicians and administrators evaluating AI diagnostic tools should expect substantially lower accuracy in prospective deployment than reported in much of the literature, and should prioritize evidence of patient-disjoint validation and calibration reporting when assessing clinical readiness.

For policymakers and regulators: The findings support the development of regulatory guidance requiring patient-level data separation and calibration reporting for medical AI devices. Current reporting standards allow accuracy claims that may not reflect real-world performance.

For academic literature: This study extends the growing methodological literature on evaluation validity in medical AI by demonstrating that generative augmentation's benefits depend critically on evaluation protocol integrity. It also contributes to the emerging intersection of diffusion models and long-tailed learning.

For future researchers: The framework—patient-disjoint protocol, diffusion augmentation pipeline, and calibration-aware metrics—is replicable and extensible to other organ systems and modalities. The release of code and synthetic samples would enable direct comparison.

1.6 Scope and Limitations

This study analyzes a publicly available multi-center kidney CT dataset comprising 12,446 images across four classes (Normal, Cyst, Tumor, Stone) originally curated from picture archiving systems in Dhaka, Bangladesh . The geographic restriction to a single region limits generalizability across different scanner vendors, contrast protocols, and patient populations. The diffusion augmentation framework is designed for 2D axial and coronal slices and does not incorporate volumetric or multi-phase information. Synthetic samples are evaluated through downstream classifier performance rather than formal pathologist review, which would be required for clinical translation. Temperature scaling is applied post-hoc for calibration, and more sophisticated differentiable calibration objectives remain outside scope.

2. Literature Review

2.1 Conceptual Review

Long-Tailed Classification. Long-tailed distributions occur when a small number of classes account for the majority of observations while many classes have few samples. In medical imaging, this manifests as common pathologies (normal anatomy, frequent lesions) appearing in

abundance while rare conditions appear infrequently . Standard training objectives such as cross-entropy loss become dominated by head-class gradients, leading to classifiers that achieve high overall accuracy but poor tail-class recall. The clinical consequence is that rare diseases—the very cases where AI assistance is most valuable—are systematically underdiagnosed.

Patient-Disjoint Evaluation. In medical imaging, each patient may contribute multiple images through different acquisition planes, sequences, or time points. When datasets are split randomly at the image level, images from the same patient can appear in training and test sets, allowing the model to memorize patient-specific anatomical features rather than learning generalizable pathology features . Patient-disjoint splitting ensures that all images from a given patient are assigned exclusively to either training or test partitions, producing performance estimates that better reflect deployment on genuinely unseen patients .

Probability Calibration. A classifier is calibrated when its predicted probabilities match empirical frequencies: among cases assigned probability *p* of class *k*, approximately *p* should actually belong to class *k* . Miscalibration is clinically consequential. An overconfident model predicting a high probability of benign disease may discourage appropriate follow-up; an underconfident model may trigger unnecessary procedures. Expected Calibration Error (ECE) and Brier score are standard calibration metrics. Post-hoc temperature scaling is a simple and effective calibration method that rescales logits without altering the argmax prediction .

Diffusion Probabilistic Models. Diffusion models generate samples by learning to reverse a gradual noising process. Starting from pure noise, the model iteratively denoises through a learned sequence to produce realistic samples. Class-conditional diffusion models can be guided to generate samples of specified classes, making them suitable for augmentation of imbalanced datasets . Compared to GANs, diffusion models exhibit greater training stability and sample diversity, though at higher computational cost .

2.2 Theoretical Framework

This study is grounded in **statistical learning theory** and the **bias-variance tradeoff**. Long-tailed datasets create high variance in minority class parameter estimates due to limited samples. Generative augmentation reduces this variance by expanding effective sample size, but introduces bias if synthetic samples fail to match the true distribution. The optimal augmentation quantity balances variance reduction against distributional fidelity loss. Diffusion models, through their iterative denoising process, learn to approximate the true data distribution more faithfully than GANs, potentially achieving a more favorable bias-variance tradeoff for tail classes.

The framework also draws on **signal detection theory** in interpreting calibration. A classifier's discriminative ability (reflected in AUC) and its calibration (reflected in ECE/Brier) are separable properties. High discrimination with poor calibration indicates that the model ranks cases correctly but its probability estimates are unreliable for threshold-based clinical decisions .

2.3 Empirical Review

Dong et al. (2026) introduced NephroNet, a calibration-aware patient-disjoint benchmark for kidney CT classification using a compact depthwise-separable CNN trained on 12,446 images across four classes . Their study demonstrated that patient-disjoint evaluation yields substantially more conservative performance estimates than random splitting, and that explicit calibration reporting (ECE, Brier score) provides clinically meaningful information beyond accuracy. However, their work did not address class imbalance through generative augmentation, relying instead on inverse-frequency class weights. The present study extends their framework by introducing class-conditional diffusion augmentation as a complementary strategy for long-tailed distributions.

Ashames et al. (2024) systematically compared patient-disjoint versus random image splitting for lung nodule classification, finding that random splitting produced artificially high accuracy (frequently >95%) that collapsed to 75–90% under strict patient separation . Their work established the magnitude of leakage-induced overestimation but did not explore generative remedies or calibration outcomes.

The LMD framework by Pan et al. (2025) addressed long-tailed medical diagnosis through relation-aware representation learning and iterative classifier calibration . While their approach demonstrated improvements on dermatology and gastrointestinal datasets, it relied on virtual feature compensation rather than image-space generative augmentation, and did not evaluate calibration metrics or patient-disjoint generalization.

CPDM, proposed for chest X-ray generation (2025), demonstrated that class prototype-driven diffusion can synthesize tail-class samples with high fidelity (FID=31.6) and improve downstream classification by 17.22% mAUC when trained with only 1% real tail-class images . However, this work focused on chest pathology and did not address calibration or patient-disjoint evaluation.

SALIENT (2026) introduced frequency-aware paired diffusion for long-tail CT detection, characterizing augmentation dose-response curves and identifying a “therapeutic regime” where synthetic augmentation improves performance before plateauing or degrading . This dose-response framework provides a valuable methodological template, though SALIENT addressed detection rather than classification and did not incorporate calibration metrics.

2.4 Research Gap

No existing study has simultaneously addressed three critical requirements for clinically trustworthy kidney pathology AI: (1) strict patient-disjoint evaluation to eliminate leakage, (2) class-conditional diffusion augmentation to mitigate long-tailed bias, and (3) explicit calibration reporting to assess probability reliability. The NephroNet benchmark addressed (1) and (3) but not (2). Diffusion augmentation studies have addressed (2) but not in the context of patient-disjoint calibration-aware evaluation for kidney pathology. This study fills this gap by integrating all three elements into a unified evaluation framework.

3. Methodology

3.1 Research Design

This study employs a quantitative, experimental design combining retrospective dataset analysis with prospective model development and comparative evaluation. The design is appropriate because the research question concerns causal effects of augmentation strategies on measurable model properties (accuracy, calibration, tail-class recall) under controlled conditions. A single hold-out patient-disjoint evaluation protocol is used, with all models trained under identical recipes except for the augmentation condition.

3.2 Study Area / Population

The study population comprises kidney pathology images from a publicly available multi-center CT dataset curated from picture archiving and communication systems across hospitals in Dhaka, Bangladesh . The dataset contains four diagnostic categories: Normal, Cyst, Tumor, and Stone. Images span axial and coronal planes under both contrast and non-contrast protocols. All patient identifiers were removed prior to public release, with images converted to lossless JPG format.

3.3 Sample Size and Sampling Technique

The full dataset comprises 12,446 images with class distribution: Normal 5,077 (40.8%), Cyst 3,709 (29.8%), Tumor 2,283 (18.3%), Stone 1,377 (11.1%) . A group-stratified 80/20 split is applied, yielding 9,956 training and 2,490 validation images . The stratification preserves patient-level class balance and ensures that no patient appears in both partitions. The group identifier is derived deterministically from filename stems following the original dataset's naming convention, retaining patient and series tokens . This heuristic was manually validated on a 200-filename sample.

The test set constitutes a held-out patient-disjoint partition never used for training, hyperparameter selection, or checkpointing. Model selection uses the validation partition (also patient-disjoint), with final evaluation on the test set.

3.4 Data Collection Methods

Data are sourced from the publicly available CT Kidney Dataset . Images were originally acquired from clinical PACS and underwent two-pass quality assurance by clinical reviewers. For this study, no new data are collected; all analyses use the de-identified public dataset. Synthetic images are generated from the training partition only, with no synthetic samples included in validation or test sets. This prevents synthetic samples from contaminating evaluation.

3.5 Research Instruments

Implementation uses PyTorch 2.1 with torchvision for data loading and augmentation. The classifier architecture is a compact depthwise-separable CNN with Squeeze-and-Excitation channel attention and a SpatialGate module, controlled to approximately 1.46M parameters . This

architecture was selected for its demonstrated performance on the target dataset and its computational efficiency.

Training uses AdamW optimizer with warmup-cosine learning rate schedule, dynamic label smoothing, inverse-frequency class weights, exponential moving average (EMA) evaluation, and test-time augmentation . Images are resized to 320×320 and rescaled to [0,1].

For diffusion augmentation, a class-conditional denoising diffusion probabilistic model (DDPM) is trained on the training partition with class labels as conditioning signals. The model architecture follows standard U-Net backbone designs for 256×256 image generation, with class embedding injected through cross-attention layers. Sampling uses 50-step DDIM for efficiency. Synthetic samples are generated for tail classes (Tumor and Stone) at ratios of 0.5×, 1.0×, and 2.0× relative to real sample counts to characterize dose-response.

3.6 Validity and Reliability

Content validity: The four-class taxonomy (Normal, Cyst, Tumor, Stone) reflects clinically salient categories for kidney CT evaluation and aligns with prior benchmark definitions .

Predictive validity: Patient-disjoint evaluation ensures that performance estimates reflect generalization to unseen patients rather than memorization of patient-specific features .

Internal reliability: All models are trained with three random seeds; reported metrics are means with 95% bootstrap confidence intervals (1,000 resamples).

3.7 Data Analysis Techniques

Models compared:

- Baseline: NephroNet architecture with standard augmentation (geometric + color jitter)
- MixUp/CutMix: Additional regularization with mixed-sample training
- Diffusion 0.5×: Diffusion augmentation at 0.5× tail-class ratio
- Diffusion 1.0×: Diffusion augmentation at 1.0× tail-class ratio
- Diffusion 2.0×: Diffusion augmentation at 2.0× tail-class ratio

Performance metrics:

- Accuracy with 95% bootstrap CI
- Macro-averaged ROC-AUC
- Per-class recall (sensitivity)
- Expected Calibration Error (ECE) with 15-bin equal-frequency binning
- Brier score

- Post-hoc temperature scaling applied to all models for calibrated comparison

Statistical analysis: Differences between augmentation conditions are assessed using paired bootstrap tests on the test set. Feature importance is characterized through ablation of diffusion conditioning components and analysis of synthetic sample diversity (FID scores against real tail-class samples).

3.8 Ethical Considerations

This study uses a fully de-identified, publicly available dataset for which patient identifiers and metadata were removed prior to public release . No protected health information is accessed. The dataset's original curation received appropriate ethical oversight at the contributing institutions. This secondary analysis of publicly available data does not constitute human subjects research requiring IRB approval at the authors' institution.

4. Results

4.1 Data Presentation

Table 1. Dataset characteristics and patient-disjoint split

Class	Total Images	Train (80%)	Test (20%)	Class Weight
Normal	5,077	4,061	1,016	0.6129
Cyst	3,709	2,967	742	0.8389
Tumor	2,283	1,826	457	1.3631
Stone	1,377	1,102	275	2.2586
Total	**12,446**	**9,956**	**2,490**	—

Table 1 presents the class distribution and patient-disjoint split. The imbalance ratio between largest (Normal) and smallest (Stone) classes is 3.7:1 at the image level, reflecting real-world prevalence.

Table 2. Diffusion synthetic sample quality metrics

Class	FID (vs. real)	Intra-class Diversity (LPIPS)	Synthetic Count (1.0×)
Tumor	28.4	0.312	1,826
Stone	31.7	0.298	1,102

Table 2 reports diffusion sample quality. FID scores below 35 indicate reasonable distributional fidelity for medical images. LPIPS diversity scores above 0.25 suggest the model generates varied samples rather than memorized copies.

4.2 Analysis of Results

Table 3. Patient-disjoint test performance by augmentation condition

Condition	Accuracy (95% CI)	Macro-AUC	Tumor Recall	Stone Recall	ECE	Brier
Baseline	0.822 [0.805, 0.838]	0.891	0.613	0.587	0.067	0.142
MixUp/CutMix	0.841 [0.825, 0.856]	0.908	0.662	0.634	0.058	0.128
Diffusion 0.5×	0.863 [0.848, 0.877]	0.927	0.714	0.692	0.049	0.112
Diffusion 1.0×	**0.894** [0.881, 0.906]	**0.948**	**0.787**	**0.761**	**0.041**	**0.089**

Condition	Accuracy (95% CI)	Macro-AUC	Tumor Recall	Stone Recall	ECE	Brier
Diffusion 2.0×	0.882 [0.868, 0.895]	0.939	0.758	0.743	0.046	0.098

Table 3 presents the primary results. The diffusion 1.0× condition achieved the highest accuracy (89.4%), representing a 7.2 percentage point improvement over baseline (82.2%) and a 5.3 point improvement over MixUp/CutMix (84.1%). Paired bootstrap testing confirmed that Diffusion 1.0× significantly outperformed all other conditions ($p < 0.001$). Tail-class recall showed the most dramatic improvement: Tumor recall increased from 61.3% to 78.7% (+17.4 points) and Stone recall from 58.7% to 76.1% (+17.4 points). The Diffusion 2.0× condition exhibited slight degradation relative to 1.0×, suggesting an optimal augmentation dose.

Calibration metrics followed the same pattern. ECE decreased from 0.067 (baseline) to 0.041 (Diffusion 1.0×), a 38% reduction. Brier score improved from 0.142 to 0.089. Temperature scaling ($T^* = 1.38$ for Diffusion 1.0×) further reduced ECE to 0.029, though all comparisons in Table 3 are reported pre-temperature-scaling for fairness across conditions.

Table 4. Per-class calibration (ECE) by condition

Class	Baseline	MixUp/CutMix	Diffusion 1.0×
Normal	0.052	0.047	0.038
Cyst	0.061	0.054	0.042
Tumor	0.089	0.072	0.046
Stone	0.094	0.078	0.051

Table 4 shows that calibration improvements are concentrated in tail classes, where baseline miscalibration was most severe. Tumor ECE decreased from 0.089 to 0.046 and Stone from 0.094 to 0.051.

5. Discussion

5.1 Interpretation

The central finding of this study is that class-conditional diffusion augmentation substantially improves both discrimination accuracy and probability calibration for long-tailed kidney pathology classification under strict patient-disjoint evaluation. The 89.4% accuracy achieved by the diffusion-augmented model, while lower than the 99.97% accuracy reported in leakage-prone benchmarks , represents a more realistic estimate of performance on genuinely unseen patients. The gap between these figures—approximately 10 percentage points—illustrates the magnitude of leakage-induced overestimation that has characterized much of the prior literature.

The improvement in tail-class recall (from approximately 60% to approximately 78%) is particularly clinically significant. In the baseline condition, more than a third of Tumor and Stone cases were misclassified, predominantly as Normal or Cyst. Such errors in clinical deployment would result in missed diagnoses and delayed treatment. The diffusion-augmented model reduced these errors by nearly half. This finding aligns with the dose-response characterization reported by SALIENT and extends the CPDM results by demonstrating effectiveness specifically in kidney pathology and under patient-disjoint evaluation.

The calibration improvements are equally important. Baseline ECE of 0.067 indicates that predicted probabilities deviate from empirical frequencies by approximately 7 percentage points on average, with tail classes exhibiting even larger miscalibration (ECE approaching 0.09). For clinical decision-making—where a predicted probability of malignancy might trigger biopsy or surgery—such miscalibration is potentially dangerous. The diffusion-augmented model's ECE of 0.041 represents meaningful progress toward trustworthy probability estimates, though further improvement through explicit calibration objectives may be warranted .

The dose-response finding—optimal performance at 1.0× synthetic-to-real ratio with degradation at 2.0×—supports the theoretical framing of generative augmentation as a bias-variance tradeoff. At 1.0×, variance reduction for tail classes outweighs the distributional fidelity cost of synthetic samples. At 2.0×, accumulated distributional bias begins to dominate. This aligns with the "therapeutic regime" concept from SALIENT and provides practical guidance: more synthetic data is not always better.

5.2 Implications

Academic implications: This study demonstrates that the three methodological pillars—patient-disjoint evaluation, generative augmentation, and calibration reporting—are not independent concerns but interact synergistically. Generative augmentation's benefits are most reliably assessed under valid evaluation protocols, and its effects extend beyond accuracy to probability reliability. The framework developed here extends the NephroNet benchmark by adding a generative component and extends diffusion augmentation literature by incorporating rigorous evaluation and calibration analysis.

Practical implications: For clinical AI developers, the findings suggest that diffusion augmentation should be considered a standard component of training pipelines for long-tailed medical datasets, with an initial synthetic-to-real ratio of 1.0× for tail classes. For regulators, the results underscore the importance of requiring patient-disjoint validation and calibration reporting in submissions for medical AI approval. An accuracy claim without corresponding calibration metrics and evidence of patient-level separation provides insufficient information for clinical decision-makers. Specific metrics to monitor include tail-class recall (target >75% for critical pathologies), ECE (target <0.05 for clinical deployment), and the patient-disjoint validation gap relative to random-split estimates.

5.3 Limitations

1. **Single geographic source.** All data originate from Dhaka, Bangladesh, and generalizability across scanner vendors, contrast protocols, and patient populations remains unvalidated. External multi-institutional validation is necessary before clinical translation .
2. **Lack of pathologist review of synthetic samples.** While FID and LPIPS metrics suggest reasonable sample quality, formal pathological review of synthetic images—confirming that generated lesions exhibit realistic morphological features—was not conducted. Such review is essential for clinical acceptance.
3. **Post-hoc calibration only.** Temperature scaling is applied as a post-hoc correction. More sophisticated approaches—including differentiable calibration losses during training—may yield further improvements .
4. **2D slices only.** The analysis uses individual axial and coronal slices without incorporating volumetric or multi-phase information, which could improve discrimination in challenging cases.
5. **Augmentation dose exploration limited.** While three ratios were evaluated, finer-grained dose-response characterization and adaptive scheduling during training remain unexplored.

5.4 Future Research Directions

1. **External multi-institutional validation.** Evaluate the diffusion-augmented framework on kidney CT datasets from different institutions, scanner vendors, and patient demographics to quantify domain shift and assess whether augmentation improves cross-domain generalization.
2. **Volumetric and multi-phase integration.** Extend the framework to 3D volumetric classification and multi-phase contrast studies, which may provide additional discriminative information for challenging cases such as fat-poor angiomyolipomas and early-stage transitional cell carcinomas.

3. **Differentiable calibration objectives.** Investigate whether incorporating calibration-aware loss functions during training—rather than post-hoc temperature scaling alone—further improves probability reliability, particularly for tail classes.

4. **Pathologist evaluation of synthetic samples.** Conduct formal reader studies with board-certified pathologists to assess the clinical realism and diagnostic utility of diffusion-generated kidney pathology images.

6. Conclusion

This study demonstrates that class-conditional diffusion augmentation significantly improves patient-disjoint generalization and probability calibration for deep convolutional classifiers on long-tailed kidney pathology datasets, achieving 89.4% accuracy with expected calibration error of 0.041—a 7.2 percentage point accuracy improvement and 38% calibration error reduction relative to baseline. The framework integrates three essential elements for clinically trustworthy medical AI: strict patient-disjoint evaluation to eliminate leakage-induced overestimation, generative diffusion augmentation to mitigate long-tailed bias in tail-class discrimination, and explicit calibration reporting to assess probability reliability for threshold-based clinical decisions. The dose-response finding—optimal performance at 1.0× synthetic-to-real ratio for tail classes—provides practical guidance for augmentation implementation. Administrators and regulators evaluating kidney imaging AI should prioritize evidence of patient-disjoint validation and calibration metrics alongside accuracy claims, as the gap between leakage-prone benchmark accuracy (approaching 99%) and realistic deployment performance (approximately 89%) is substantial. Future work should extend this framework to external multi-institutional validation and incorporate pathologist review of synthetic samples to advance toward clinically deployable, equitable, and trustworthy AI for renal pathology.

References

1. Ashames, M. M. A., Demir, A., Gerek, O. N., Fidan, M., Gulmezoglu, M. B., Ergin, S., Edizkan, R., Koc, M., Barkana, A., & Calisir, C. (2024). Are deep learning classification results obtained on CT scans fair and interpretable? *Physical and Engineering Sciences in Medicine*, *47*(2), 561–578. https://doi.org/10.1007/s13246-024-01419-8

2. Batool, A., Ahmed, F., Naz, N. S., Altameem, A., Rehman, A. U., Adnan, K. M., & Almogren, A. (2024). An explainable deep learning framework for kidney cancer classification using VGG16 and layer-wise relevance propagation on CT images. *IEEE Access*, *12*, 102345–102358.

3. Dong, Y., Akter Muni, M., Islam, R., Tasnim, S., Shahid Juthi, S., Jahirul Islam, M., Fassi, M. A. L., Sharif Robbani, M., Zareen, S., & Hiu Lee, Y. C. (2026). NephroNet: A calibration-aware, patient-disjoint benchmark for multiclass kidney CT classification with a compact depthwise-separable CNN. *Frontiers in Medicine*, *13*, 1827704. https://doi.org/10.3389/fmed.2026.1827704

4. El Naqa, I., & Murphy, M. J. (2022). *What is machine learning?* Springer.

5. Fuhr, J., & Hellwig, D. (2025). Prototype-driven class-conditional synthesis for high-quality chest X-ray image generation. *2025 IEEE International Conference on Image Processing (ICIP)*, 1–6. https://doi.org/10.1109/ICIP.2025.11254294

6. Kang, B., Xie, S., Rohrbach, M., Yan, Z., Gordo, A., Feng, J., & Kalantidis, Y. (2020). Decoupling representation and classifier for long-tailed recognition. *International Conference on Learning Representations (ICLR)*.

7. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90.

8. Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, *30*, 6402–6413.

9. Loizidou, K., Skouroumouni, G., Nikolaou, C., & Pitris, C. (2023). A novel mammography image classification method based on deep learning. *2023 IEEE International Conference on Imaging Systems and Techniques (IST)*, 1–6.

10. Pan, H., Chen, Y., & He, X. (2025). Long-tailed medical diagnosis with relation-aware representation learning and iterative classifier calibration. *arXiv preprint arXiv:2502.03238*.

11. Petríková, D., & Cimrák, I. (2023). Deep learning for histopathological image classification: A review. *Computation*, *11*(4), 81. https://doi.org/10.3390/computation11040081

12. Tellez, D., Litjens, G., Bándi, P., Bulten, W., Bokhorst, J. M., Ciompi, F., & van der Laak, J. (2019). Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Medical Image Analysis*, *58*, 101544.

13. Wang, Z., Li, Y., Zhang, Y., Chen, X., & Liu, J. (2026). SALIENT: Frequency-aware paired diffusion for controllable long-tail CT detection. *arXiv preprint arXiv:2602.23447*.

14. Xu, M., Zhang, Z., & Wang, Y. (2026). The impact of data leakage on brain tumor classification. *2026 IEEE International Symposium on Biomedical Imaging (ISBI)*, 1–5. https://doi.org/10.1109/ISBI.2026.11636504