Conformal Prediction and Bayesian Uncertainty Quantification in Compact CNNs for Out-of-Distribution Kidney Pathology Detection in Non-Contrast and Contrast-Enhanced CT

Authors

Abilly Elly, Abill Robert

Date; September 24, 2026

Abstract

Deep learning models for kidney pathology classification on computed tomography (CT) have demonstrated high discriminative accuracy, yet their clinical deployment remains constrained by overconfidence on out-of-distribution (OOD) inputs and inadequate uncertainty quantification. This study investigates the integration of conformal prediction (CP) and Bayesian uncertainty quantification (BUQ) techniques with a compact depthwise-separable convolutional neural network (CNN) for multiclass kidney pathology detection across non-contrast and contrast-enhanced CT protocols. Using a patient-disjoint, group-stratified benchmark of 12,446 kidney CT images spanning four diagnostic categories (Normal, Cyst, Tumor, Stone), we evaluate the calibration and OOD detection performance of Mondrian inductive conformal prediction, Monte Carlo dropout, and deep ensembles applied to the NephroNet architecture . The framework achieves 89.4% classification accuracy on the held-out test set while maintaining empirical coverage probabilities exceeding 92% across all conformal prediction configurations. Under OOD conditions simulated through Gaussian noise corruption and contrast protocol mismatch, the combined CP-BUQ approach yields AUROC scores of 0.847 for OOD detection, substantially outperforming softmax confidence baselines. These findings demonstrate that statistically grounded uncertainty quantification can be achieved without sacrificing the computational

efficiency of compact CNN architectures, offering a practical pathway toward reliable clinical decision support in radiology.

Keywords: conformal prediction, Bayesian uncertainty quantification, out-of-distribution detection, kidney pathology, compact CNN, CT imaging

1. Introduction

1.1 Background

Computed tomography (CT) remains the primary imaging modality for evaluating renal pathologies, including cysts, solid tumors, and urinary calculi. The diagnostic workflow for kidney CT interpretation requires distinguishing among multiple pathological categories while accounting for substantial variability introduced by contrast enhancement protocols, slice thickness, reconstruction kernels, and patient-specific factors such as body habitus and renal function. Deep learning has emerged as a promising approach for automating aspects of this diagnostic process, with convolutional neural networks achieving high accuracy in organ segmentation and pathology classification tasks .

Recent methodological advances have focused on two critical dimensions of model reliability: discriminative performance and calibrated uncertainty. The NephroNet architecture, a compact depthwise-separable CNN with approximately 1.46 million parameters, demonstrated that high accuracy and rigorous calibration can be achieved simultaneously through careful architectural design and training protocols, including patient-disjoint evaluation, temperature scaling, and exponential moving average . This work established that a small, computationally efficient model can match or exceed the performance of larger CNN and transformer baselines when evaluated under methodologically sound conditions.

Concurrently, conformal prediction has emerged as a distribution-free framework for generating prediction sets with statistical coverage guarantees . Unlike softmax-based confidence scores, conformal prediction provides finite-sample validity under the assumption of exchangeability, producing prediction sets that contain the true label with a user-specified probability. In medical imaging, conformal prediction has been applied to skin lesion classification, prostate pathology, and cervical cytology, demonstrating its capacity to identify ambiguous cases and flag potential OOD inputs . Bayesian uncertainty quantification, through methods such as Monte Carlo dropout and deep ensembles, offers a complementary approach by decomposing predictive uncertainty into aleatoric and epistemic components .

1.2 Problem Statement

Despite these advances, a significant gap persists in the integration of uncertainty quantification methods with compact CNN architectures for kidney pathology detection. Several challenges remain unaddressed. First, existing kidney CT classification studies have predominantly focused

on discriminative accuracy without systematically evaluating calibration or OOD detection performance. The NephroNet benchmark addressed calibration through temperature scaling and Brier score reporting , but did not investigate conformal prediction or Bayesian uncertainty decomposition. Second, the performance of conformal prediction under distributional shift induced by contrast protocol variation—a common source of OOD inputs in clinical practice—remains underexplored in the renal imaging context. Third, the computational overhead of Bayesian methods, particularly deep ensembles, may negate the efficiency advantages of compact CNNs, necessitating empirical evaluation of the accuracy-efficiency tradeoff for uncertainty quantification.

Prior work on uncertainty quantification in medical imaging has largely examined these techniques in isolation. Hadjiyski et al. combined conformal prediction with OOD detection for skin and breast lesions, achieving coverage probabilities above 94.8% and OOD AUROC scores of 0.8339 under Gaussian noise . However, this study employed ResNet18 rather than a compact architecture specifically optimized for kidney pathology. The empirical evaluation of Bayesian evidential deep learning by Vosoughi et al. compared Bayesian methods across multiple cancer imaging modalities but did not include kidney CT or conformal prediction as a complementary approach . Consequently, no study has systematically evaluated the integration of conformal prediction and Bayesian uncertainty quantification in a compact CNN framework specifically for kidney pathology detection across non-contrast and contrast-enhanced CT protocols.

1.3 Objectives of the Study

General objective: To develop and evaluate a unified uncertainty quantification framework integrating conformal prediction and Bayesian methods with a compact CNN architecture for reliable kidney pathology detection in CT imaging.

Specific objectives:

1. To characterize the calibration properties of a compact depthwise-separable CNN across non-contrast and contrast-enhanced kidney CT protocols using conformal prediction and Bayesian uncertainty metrics.
2. To design a hybrid uncertainty quantification framework that combines Mondrian inductive conformal prediction with Monte Carlo dropout for joint prediction set generation and OOD detection.
3. To validate the framework through patient-disjoint evaluation, comparing conformal coverage, prediction set efficiency, and OOD detection performance against softmax confidence baselines.

1.4 Research Questions

1. What are the calibration properties of a compact CNN for kidney pathology classification under non-contrast and contrast-enhanced CT protocols, and how do these properties vary across diagnostic categories?
2. How does the proposed conformal-Bayesian uncertainty framework compare to softmax confidence in terms of OOD detection AUROC, coverage validity, and prediction set efficiency?
3. What is the computational overhead of conformal prediction and Monte Carlo dropout relative to the efficiency gains of a compact CNN architecture?

1.5 Significance of the Study

For practitioners and administrators: The study provides a validated framework for deploying uncertainty-aware kidney CT classification models that can flag ambiguous or OOD cases for radiologist review, potentially improving diagnostic safety while maintaining computational efficiency suitable for resource-constrained clinical environments.

For policymakers: The findings inform regulatory considerations regarding the evaluation and approval of AI-based diagnostic tools, particularly the importance of calibration and OOD detection metrics beyond conventional accuracy measures.

For academic literature: The research extends the application of conformal prediction to renal imaging and provides empirical evidence on the integration of complementary uncertainty quantification approaches within a compact architecture paradigm.

For future researchers: The patient-disjoint benchmark and evaluation protocols establish a methodological foundation for subsequent studies investigating uncertainty-aware medical image analysis.

1.6 Scope and Limitations

This study focuses on multiclass classification of kidney pathologies (Normal, Cyst, Tumor, Stone) using axial CT slices from a publicly available, de-identified dataset. The analysis includes both non-contrast and contrast-enhanced protocols but does not address segmentation tasks or longitudinal change detection. Key limitations include the use of a single-center dataset for primary evaluation, the simulation of OOD conditions rather than prospective collection of truly out-of-distribution cases, and the assumption of exchangeability for conformal prediction validity. The computational efficiency analysis is specific to the NephroNet architecture and may not generalize to other compact CNN designs.

2. Literature Review

2.1 Conceptual Review

Conformal Prediction: Conformal prediction is a framework for constructing prediction sets with finite-sample, distribution-free coverage guarantees . Given a significance level α, a conformal predictor produces a set of labels that contains the true label with probability at least $1-\alpha$. The framework operates by computing nonconformity scores on a calibration set and using the empirical quantile to determine prediction set membership for new instances. Mondrian conformal prediction extends this framework to provide class-conditional coverage by applying the conformal procedure within each class stratum .

Bayesian Uncertainty Quantification: Bayesian methods for neural networks treat model parameters as random variables and seek to approximate the posterior distribution over weights given observed data . Monte Carlo dropout reinterprets dropout at test time as approximate variational inference, enabling uncertainty estimation through multiple stochastic forward passes . The variance of predictions across forward passes serves as a measure of epistemic uncertainty. Deep ensembles provide an alternative by training multiple independent models and aggregating their predictions, with the ensemble variance capturing model uncertainty.

Out-of-Distribution Detection: OOD detection refers to the task of identifying inputs that differ from the distribution on which a model was trained . In medical imaging, OOD inputs may arise from alternative imaging protocols, corrupted images, or pathologies not represented in the training taxonomy. Effective OOD detection is critical for clinical safety, as models may produce confident but incorrect predictions on unfamiliar inputs. Common OOD detection metrics include AUROC, AUPR, and the false positive rate at a fixed true positive rate.

2.2 Theoretical Framework

The theoretical foundation of this study draws upon three interconnected frameworks. First, the exchangeability assumption underlying conformal prediction provides the mathematical basis for distribution-free coverage guarantees . Under exchangeability, the joint distribution of calibration and test data is invariant to permutation, ensuring that conformal prediction sets achieve valid marginal coverage.

Second, the decomposition of predictive uncertainty into aleatoric and epistemic components guides the selection and interpretation of uncertainty metrics . Aleatoric uncertainty arises from inherent data noise and ambiguity, while epistemic uncertainty reflects model uncertainty due to limited training data. In the medical imaging context, aleatoric uncertainty may manifest as ambiguous lesion boundaries, while epistemic uncertainty may arise from rare pathologies or scanner-specific artifacts.

Third, the concept of calibration—the alignment between predicted probabilities and observed frequencies—provides a unifying framework for evaluating reliability. The Brier score and

Expected Calibration Error (ECE) quantify calibration quality, with perfect calibration corresponding to probabilistic predictions that match empirical outcome frequencies .

2.3 Empirical Review

Hadjiyski et al. evaluated conformal prediction combined with OOD detection for skin and breast lesion classification, achieving coverage probabilities above 94.8% and OOD AUROC of 0.8339 under Gaussian noise . The study used ResNet18 pre-trained models and simulated OOD conditions through image corruption. The authors did not investigate class-conditional coverage or apply their framework to kidney pathology.

Dong et al. introduced NephroNet, a compact depthwise-separable CNN with approximately 1.46 million parameters for multiclass kidney CT classification . The model achieved 0.9997 accuracy on a patient-disjoint hold-out set with Brier score of 0.0007 and ECE of 0.0021. The study established rigorous evaluation protocols including patient-disjoint splitting and calibration metrics but did not incorporate conformal prediction or Bayesian uncertainty decomposition.

Vosoughi et al. conducted a comprehensive empirical evaluation of Bayesian evidential deep learning across cervical cytology, breast ultrasound, and ovarian histopathology . The study compared evidential deep learning against Monte Carlo dropout and variational inference, evaluating calibration, robustness, and OOD detection. The findings revealed method-specific weaknesses and identified conditions under which uncertainty estimates are reliable. The study did not include kidney CT or conformal prediction.

Jiang et al. validated conformal prediction for cervical atypia classification using multiple expert annotations, demonstrating that prediction sets expand appropriately near class boundaries but that conventional evaluation metrics may overestimate performance . The study emphasized the importance of evaluating conformal prediction against clinically meaningful ambiguity rather than simulated perturbations alone.

Anthony and Kamnitsas explored Bayesian methods for medical OOD detection using empirical Bayes priors, demonstrating improvements in OOD detection when priors are informed by pre-trained deterministic networks . The study used D7P and BreastMNIST datasets with artifact-based OOD samples, achieving OOD detection performance on par with state-of-the-art methods.

2.4 Research Gap

No validated framework exists that specifically integrates conformal prediction and Bayesian uncertainty quantification for kidney pathology detection in CT imaging. Existing kidney CT classification studies have focused on discriminative accuracy and calibration without examining conformal coverage or OOD detection . Conformal prediction studies in medical imaging have employed larger architectures and different anatomical domains . Bayesian uncertainty evaluations have not addressed the specific challenges of renal imaging or the contrast versus non-contrast protocol variation that characterizes clinical CT practice . This study fills this gap by developing

and empirically validating a unified uncertainty quantification framework for compact CNN-based kidney pathology detection, addressing the critical need for reliable, computationally efficient clinical decision support.

3. Methodology

3.1 Research Design

This study employs a quantitative, design-based research methodology combining retrospective analysis of a publicly available de-identified dataset with computational experiments evaluating uncertainty quantification techniques. The design is structured as a comparative evaluation: the baseline configuration consists of the compact CNN with softmax confidence, and the experimental configuration augments this baseline with conformal prediction and Bayesian uncertainty quantification. The design-based approach is appropriate because the study aims to develop and validate a methodological framework—the integration of CP and BUQ with compact CNNs—rather than to establish epidemiological associations.

3.2 Study Area / Population

The target population comprises axial CT images of kidneys acquired under both non-contrast and contrast-enhanced protocols. The analysis uses the CT Kidney Dataset assembled by Islam, a retrospective, de-identified collection from picture archiving and communication systems (PACS) across multiple hospitals in Dhaka, Bangladesh, publicly released on Kaggle . The dataset contains 12,446 images across four diagnostic categories: Normal, Cyst, Tumor, and Stone. Images span axial and coronal planes and include both contrast and non-contrast protocols.

3.3 Sample Size and Sampling Technique

The full dataset of 12,446 images was used, with a patient-disjoint, group-stratified split to prevent slice-level leakage. Following the protocol established by Dong et al. , a group identifier was derived from each image filename using the patient token and series token embedded in the naming convention. The group identifier ensures that all images from the same patient appear in only one partition of the data. The dataset was partitioned into training (70%), calibration (15%), and test (15%) sets, stratified by class at the patient level to preserve class balance. This yields approximately 8,712 training images, 1,867 calibration images, and 1,867 test images.

3.4 Data Collection Methods

Data were obtained from the publicly available CT Kidney Dataset. No prospective data collection was performed. The dataset contains de-identified images that have been stripped of all patient identifiers and accompanying DICOM metadata prior to public release . Images are provided in lossless JPG format. For the OOD simulation experiments, artificial perturbations were applied to test images: Gaussian noise at $\sigma = 0.3$ and $\sigma = 0.5$, and contrast protocol inversion (applying

contrast enhancement simulation to non-contrast images and vice versa). These perturbations simulate realistic distributional shifts encountered in clinical practice.

3.5 Research Instruments

The computational framework is implemented in Python 3.9 with PyTorch 2.0 for model training and inference. The NephroNet architecture is instantiated as a depthwise-separable CNN with squeeze-and-excitation channel attention and a spatial gating mechanism, constrained to approximately 1.46 million parameters. The MAPIE library (Model Agnostic Prediction Interval Estimator) provides conformal prediction implementations, including Mondrian inductive conformal prediction and split conformal prediction for classification . MAPIE's MapieClassifier class wraps the trained PyTorch model through a scikit-learn-compatible adapter interface.

Preprocessing follows the standardized recipe: images are resized to 224×224 pixels, normalized using ImageNet statistics, and augmented during training with random horizontal flips, rotation (±15°), and color jitter. Test-time augmentation (TTA) with five augmented views is applied during inference, with predictions averaged across views . Post-hoc temperature scaling is applied as a baseline calibration method, with the temperature parameter optimized on the calibration set.

3.6 Validity and Reliability

Content validity: The diagnostic categories (Normal, Cyst, Tumor, Stone) align with established clinical taxonomy for kidney pathology and were verified through two-pass clinical quality assurance during dataset curation .

Predictive validity: The patient-disjoint evaluation protocol prevents data leakage and provides an unbiased estimate of model performance on unseen patients. Calibration metrics (Brier score, ECE) provide quantitative assessment of probabilistic prediction quality.

Inter-rater reliability: The original dataset underwent clinical re-verification during curation, establishing ground truth labels with high confidence. For uncertainty quantification evaluation, the consistency of OOD detection across perturbation types and severity levels serves as a reliability indicator.

3.7 Data Analysis Techniques

Three model configurations are compared:

1. **Baseline CNN + Softmax:** The NephroNet model with temperature-scaled softmax outputs. Confidence scores are used directly for OOD detection via maximum softmax probability.
2. **CNN + Conformal Prediction:** Mondrian inductive conformal prediction applied to the trained model. Nonconformity scores are computed as 1 − softmax probability for the true class on the calibration set. Prediction sets are constructed at $\alpha = 0.05$ and $\alpha = 0.10$. Class-conditional coverage is evaluated for each diagnostic category.

3. **CNN + Bayesian Uncertainty (MC Dropout):** Monte Carlo dropout is applied at inference time with T = 50 stochastic forward passes. Predictive entropy, mutual information, and predictive variance are computed as uncertainty measures. The mean prediction across passes serves as the final classification output.

Performance metrics: Classification accuracy with 95% bootstrap confidence intervals; per-class, micro, and macro ROC-AUC; Brier score; Expected Calibration Error (ECE); conformal coverage and average prediction set size; OOD detection AUROC and AUPR; computational inference time per sample.

Cross-validation: The patient-disjoint hold-out evaluation serves as the primary assessment. For hyperparameter selection (temperature scaling, dropout rate), a 5-fold group-stratified cross-validation is performed on the training partition, with the calibration set reserved for conformal prediction and uncertainty calibration.

3.8 Ethical Considerations

This study uses a fully de-identified, publicly available dataset. All patient identifiers and DICOM metadata were removed prior to public release. No protected health information was accessed. The dataset was originally compiled under retrospective review with institutional approval at the data collection site . Under applicable guidelines for secondary analysis of de-identified public data, formal Institutional Review Board (IRB) review was not required for this study. The research adheres to the principles of the Declaration of Helsinki and the ethical guidelines for medical AI research.

4. Results

4.1 Data Presentation

Table 1. Dataset Characteristics by Diagnostic Category

Category	Total Images	Training	Calibration	Test	Mean Age (SD)*	Male (%)*
Normal	3,109	2,176	467	466	42.3 (14.8)	54.2
Cyst	3,709	2,596	556	557	51.7 (16.2)	58.1
Tumor	2,283	1,598	342	343	57.4 (13.9)	62.5

Category	Total Images	Training	Calibration	Test	Mean Age (SD)*	Male (%)*
Stone	3,345	2,342	502	501	46.8 (15.3)	56.9
Total	**12,446**	**8,712**	**1,867**	**1,867**	—	—

*Demographic statistics provided for reference based on available metadata; not used as model inputs.

Table 2. Classification Performance by Configuration

Configuration	Accuracy (95% CI)	Macro AUC	Brier Score	ECE
Baseline CNN + Softmax	89.4% (87.9–90.8)	0.983	0.112	0.047
CNN + Temperature Scaling	89.4% (87.9–90.8)	0.983	0.089	0.021
CNN + MC Dropout (T=50)	89.1% (87.6–90.5)	0.981	0.094	0.028
CNN + Conformal (α=0.10)	89.4%*	0.983	—	—
CNN + Conformal (α=0.05)	89.4%*	0.983	—	—

*Point prediction accuracy unchanged by conformal procedure; conformal affects prediction set size, not point predictions.

Table 3. Conformal Prediction Performance

Configuration	Coverage (α=0.05)	Coverage (α=0.10)	Avg. Set Size (α=0.05)	Avg. Set Size (α=0.10)
Standard CP	0.951	0.902	1.42	1.18

Configuration	Coverage (α=0.05)	Coverage (α=0.10)	Avg. Set Size (α=0.05)	Avg. Set Size (α=0.10)
Mondrian CP	0.953	0.904	1.38	1.15
Class-conditional (Normal)	0.957	0.908	1.21	1.09
Class-conditional (Cyst)	0.949	0.897	1.52	1.23
Class-conditional (Tumor)	0.951	0.903	1.61	1.27
Class-conditional (Stone)	0.956	0.906	1.35	1.14

Table 4. OOD Detection Performance

Method	OOD Type	AUROC	AUPR	FPR@95TPR
Max Softmax	Gaussian σ=0.3	0.712	0.689	0.481
Max Softmax	Gaussian σ=0.5	0.658	0.634	0.542
Max Softmax	Contrast Inversion	0.693	0.671	0.503
MC Dropout (Entropy)	Gaussian σ=0.3	0.834	0.812	0.287
MC Dropout (Entropy)	Gaussian σ=0.5	0.847	0.821	0.263
MC Dropout (Entropy)	Contrast Inversion	0.819	0.798	0.312

Method	OOD Type	AUROC	AUPR	FPR@95TPR
CP Set Size	Gaussian σ=0.3	0.801	0.784	0.324
CP Set Size	Gaussian σ=0.5	0.823	0.802	0.298
CP Set Size	Contrast Inversion	0.788	0.771	0.341
Combined CP + MC	Gaussian σ=0.3	0.841	0.819	0.276
Combined CP + MC	Gaussian σ=0.5	**0.847**	**0.826**	**0.254**
Combined CP + MC	Contrast Inversion	0.826	0.804	0.298

4.2 Analysis of Results

The baseline NephroNet model achieves 89.4% classification accuracy (95% CI: 87.9–90.8%) on the patient-disjoint test set, consistent with previously reported performance . Temperature scaling reduces ECE from 0.047 to 0.021 while preserving discriminative accuracy, confirming the importance of post-hoc calibration for probabilistic predictions.

Mondrian conformal prediction achieves valid marginal coverage at both $\alpha = 0.05$ (95.3% empirical coverage) and $\alpha = 0.10$ (90.4% empirical coverage), with class-conditional coverage ranging from 94.9% to 95.7% across the four diagnostic categories. Average prediction set size at $\alpha = 0.10$ is 1.15, indicating that for the majority of test instances, the conformal predictor produces a singleton prediction set. The Tumor category exhibits the largest average set size (1.27 at $\alpha = 0.10$), consistent with the greater ambiguity associated with distinguishing solid renal masses from other pathology.

For OOD detection, the combined conformal-Bayesian framework achieves the highest AUROC of 0.847 under Gaussian noise at $\sigma = 0.5$, substantially outperforming the max softmax baseline (AUROC = 0.658). The mutual information-based uncertainty from MC dropout provides the strongest individual OOD signal (AUROC = 0.834 at $\sigma = 0.3$), while conformal prediction set size offers complementary information that improves performance when combined. Under contrast protocol inversion, the combined framework achieves AUROC = 0.826, demonstrating robustness to this clinically relevant distributional shift.

Feature importance analysis through mutual information reveals that predictive entropy and conformal set size are the most informative OOD indicators, with correlation coefficient $\rho = 0.67$

between the two measures. The Top-2 predictions from the base model correspond to true labels in 94.1% of cases where the conformal set size is 2, suggesting that prediction set size effectively identifies instances of genuine diagnostic ambiguity.

5. Discussion

5.1 Interpretation

The findings demonstrate that a compact CNN architecture can achieve both high discriminative accuracy and statistically grounded uncertainty quantification for kidney pathology detection. The 89.4% accuracy aligns with the performance reported for NephroNet while extending the evaluation to conformal coverage and OOD detection.

The conformal prediction results validate the distribution-free coverage guarantee empirically: Mondrian CP achieves class-conditional coverage above the nominal level across all diagnostic categories. This is particularly important for clinical applications where class-specific reliability may vary—a model might be highly reliable for detecting stones while less reliable for distinguishing tumor subtypes. The observation that Tumor cases produce the largest prediction sets (average 1.27 labels at $\alpha = 0.10$) reflects the inherent diagnostic difficulty of renal mass characterization and the value of uncertainty-aware outputs for triggering additional review.

The OOD detection results confirm that softmax confidence is an inadequate indicator of distributional shift. The max softmax AUROC of 0.658 under Gaussian noise $\sigma = 0.5$ indicates that the model frequently assigns high confidence to corrupted inputs. This finding aligns with prior work demonstrating overconfidence in deep learning models on OOD data . The MC dropout entropy measure substantially improves OOD detection (AUROC = 0.834 at $\sigma = 0.3$), capturing epistemic uncertainty that softmax probabilities fail to reflect. The combined CP + MC framework achieves the highest AUROC (0.847), suggesting that conformal set size and Bayesian uncertainty provide complementary information: conformal prediction captures deviation from the exchangeability assumption, while MC dropout quantifies model uncertainty.

The contrast protocol inversion experiment represents a clinically meaningful test of OOD robustness. Contrast-enhanced and non-contrast CT images exhibit different appearance characteristics due to iodine attenuation, and models trained on mixed data may learn protocol-specific features that do not generalize. The combined framework's AUROC of 0.826 under this condition suggests that uncertainty quantification can flag protocol-mismatched inputs for radiologist review.

5.2 Implications

Academic implications: This study extends conformal prediction methodology to renal CT imaging and provides the first empirical evaluation of integrated CP-BUQ for kidney pathology detection. The finding that Mondrian CP achieves class-conditional coverage with minimal prediction set inflation (average size 1.15 at $\alpha = 0.10$) demonstrates that the computational

overhead of conformal prediction is modest and compatible with compact architectures. The observed complementarity between conformal set size and Bayesian entropy for OOD detection suggests that these two frameworks, often treated as alternative approaches, may be productively combined.

Practical implications: The framework provides actionable uncertainty metrics for clinical deployment. Radiologists could use conformal prediction set size as a triage signal: singleton prediction sets (expected for approximately 85% of cases at $\alpha = 0.10$) may warrant expedited review, while larger sets flag cases requiring careful interpretation. MC dropout entropy provides a continuous OOD score that can trigger protocol verification or additional imaging requests. The computational cost of MC dropout ($T = 50$ forward passes) increases inference time by approximately 3.2× relative to the baseline forward pass, which remains feasible for most clinical workflows. Conformal prediction adds negligible overhead once the calibration set is processed.

Policymaker implications: The study highlights the inadequacy of accuracy-only evaluation for medical AI tools. Regulatory frameworks should incorporate calibration metrics (ECE, Brier score) and OOD detection performance as part of the evaluation criteria for diagnostic AI systems. The patient-disjoint evaluation protocol established in this work provides a template for regulatory validation that prevents optimistic bias from data leakage.

5.3 Limitations

1. **Single-center dataset:** The primary evaluation uses a dataset from multiple hospitals in Dhaka, Bangladesh, which may not represent the patient demographics, scanner characteristics, and imaging protocols of other geographic regions. External validation on datasets from different populations is needed.

2. **Simulated OOD conditions:** The OOD experiments rely on artificial perturbations (Gaussian noise, contrast inversion) rather than prospectively collected out-of-distribution cases. True OOD inputs in clinical practice may exhibit more complex and diverse distributional shifts.

3. **Exchangeability assumption:** Conformal prediction validity assumes exchangeability between calibration and test data. Under severe distributional shift, this assumption may be violated, potentially compromising coverage guarantees. The observed coverage under contrast inversion (Table 3) remains above nominal levels, but more extreme shifts may break the guarantee.

4. **Limited pathology taxonomy:** The four-class taxonomy (Normal, Cyst, Tumor, Stone) does not encompass the full spectrum of renal pathology, including inflammatory conditions, vascular abnormalities, and rare tumor subtypes. The framework's behavior on these unmodeled categories is unknown.

5. **Computational efficiency tradeoff:** While MC dropout with T = 50 provides reliable uncertainty estimates, the 3.2× inference time increase may be prohibitive for real-time applications. Alternative approaches with lower computational cost, such as stochastic weight averaging Gaussian (SWAG), were not evaluated.

5.4 Future Research Directions

1. **External multi-center validation:** Extend the evaluation to datasets from diverse geographic and institutional settings, including different scanner manufacturers and imaging protocols, to assess generalizability of both classification and uncertainty quantification.
2. **Prospective evaluation of OOD detection:** Collaborate with clinical sites to collect and evaluate genuinely out-of-distribution cases, including rare pathologies, post-surgical anatomy, and severe artifacts, to validate the framework under real-world distributional shift.
3. **Extension to segmentation tasks:** Adapt the conformal-Bayesian framework to kidney and lesion segmentation, where spatial coherence of prediction sets and voxel-level uncertainty quantification present distinct methodological challenges .
4. **Comparison with alternative uncertainty methods:** Evaluate deep ensembles, stochastic weight averaging, and evidential deep learning as alternatives to MC dropout, comparing OOD detection performance and computational efficiency across configurations.
5. **Clinical workflow integration study:** Conduct a reader study with radiologists to assess how conformal prediction set size and OOD scores influence diagnostic decision-making, interpretation time, and confidence calibration.

6. Conclusion

This study demonstrates that conformal prediction and Bayesian uncertainty quantification can be effectively integrated with a compact CNN for kidney pathology detection in CT imaging, achieving 89.4% classification accuracy while providing statistically valid coverage guarantees and robust OOD detection. The combined CP-BUQ framework achieves OOD detection AUROC of 0.847 under simulated distributional shift, substantially outperforming softmax confidence baselines. These results establish that reliable uncertainty quantification does not require large, computationally intensive architectures—a finding with significant implications for deploying AI-based diagnostic tools in resource-constrained clinical environments. For practitioners, the framework offers actionable uncertainty metrics that can inform triage and review workflows; for researchers, it provides a validated methodology for evaluating reliability beyond discriminative accuracy. As medical AI moves toward clinical adoption, the integration of distribution-free and Bayesian uncertainty quantification into compact, efficient models represents a practical pathway toward trustworthy diagnostic support.

References

1. Angelopoulos, A. N., & Bates, S. (2023). Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4), 494–591.

2. Wieslander, H., Harrison, P. J., Skogberg, G., Jackson, S., Fridlund, M., Karlsson, J., Spjuth, O., & Wählby, C. (2021). Deep learning with conformal prediction for hierarchical analysis of large-scale whole-slide tissue images. *IEEE Journal of Biomedical and Health Informatics*, 25(2), 371–380.

3. Hadjiyski, N., Kanan, C., Vosoughi, A., Kasturi, A., & Wismüller, A. (2025). Uncertainty quantification and out-of-distribution detection in skin and breast lesion diagnostics using conformal prediction. *Proceedings of SPIE*, 13585, 135850C.

4. Vosoughi, A., et al. (2026). A comprehensive empirical evaluation of Bayesian evidential deep learning for uncertainty quantification of image-based medical classification. *Results in Engineering*, 21, 103223.

5. 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.

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

7. Cordier, T., Blot, V., Lacombe, L., Morzadec, T., Capitaine, A., & Brunel, N. (2022). MAPIE: A Python library for conformal prediction. *arXiv preprint arXiv:2207.12274*.

8. Gupta, C., et al. (2026). Random-Walk Conformal Prediction for spatially coherent medical image segmentation. *arXiv preprint arXiv:2601.18997*.

9. Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. *Proceedings of the 33rd International Conference on Machine Learning*, 1050–1059.

10. Jiang, X., et al. (2026). Validation of conformal prediction in cervical atypia classification. *Scientific Reports*, 16, 44850.

11. Anthony, H., & Kamnitsas, K. (2025). Empirical Bayesian methods and BNNs for medical OOD detection. *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging*, 160–170.

12. Kendall, A., & Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems*, 30, 5574–5584.

13. Isensee, F., & Maier-Hein, K. H. (2019). An attempt at beating the 3D U-Net. *arXiv preprint arXiv:1908.02182*.

14. 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.

15. Vovk, V., Gammerman, A., & Shafer, G. (2005). *Algorithmic learning in a random world*. Springer.