Black Sea Journal of Engineering and Science
doi: 10.34248/bsengineering.2003693

Open Access Journal
e-ISSN: 2619 – 8991

Research Article
Volume 9 - Issue 5: 2800-2816 / September 2026

AI-ASSISTED MULTICLASS KIDNEY CT PATHOLOGY CLASSIFICATION FOR UROLOGICAL IMAGING: AN OPTIMIZED HYBRID DEEP LEARNING FRAMEWORK

Mehmet Tevfik AGDAS[1], Furkan KARAHUSEYINOGLU[2]*

[1]*Munzur University, Cemisgezek Vocational School, Department of Computer Technologies, 62600, Tunceli, Türkiye*
[2]*Munzur University, Faculty of Economics and Administrative Sciences, Department of Health Management, 62600 Tunceli, Türkiye*

Abstract: Kidney CT imaging plays a central role in the urological diagnosis of renal cysts, calculi, solid tumors, and normal renal parenchyma. Although deep learning has shown promise for automated CT-based kidney pathology classification, existing approaches frequently rely on single-architecture models and lack systematic hyperparameter optimization, limiting both performance and reproducibility. This study aimed to develop and internally evaluate an optimized hybrid deep learning framework for multiclass kidney CT pathology classification in urological imaging, focusing on cyst, stone, tumor, and normal categories. A curated multiclass kidney CT dataset of 9,574 images was split into training (6,368), validation (1,997), and test (1,209) sets across four classes: Cyst, Normal, Stone, and Tumor. Four convolutional backbones-ResNet50V2, EfficientNetB3, DenseNet121, and Xception-were fine-tuned under a two-stage transfer learning protocol. Hyperparameters for each backbone were independently optimized using the Optuna framework with AdamW optimization and label smoothing. Class imbalance was addressed using weighted loss. Test-time augmentation was applied at inference. Two hybrid decision strategies were evaluated: validation-driven weighted soft voting and a logistic regression stacking learner. This was an internal computational validation study using a same-source dataset split; no external cohort was available. On the internal same-source held-out test set, ResNet50V2, Xception, the weighted ensemble, and the stacking model achieved perfect performance (100% accuracy, precision, recall, F1-score, and one-vs-rest ROC-AUC). DenseNet121 and EfficientNetB3 showed near-ceiling results. These outcomes may reflect strong class separability within the analyzed dataset and should be interpreted with caution in the absence of external validation. Ensemble strategies provided consistent performance gains over weaker individual backbones, particularly for the Stone class. The proposed optimized hybrid deep learning framework achieved strong internal performance for multiclass kidney CT pathology classification across urologically relevant categories. The findings suggest that architecture diversity and systematic hyperparameter optimization can substantially improve predictive robustness. External multi-center validation is required before clinical implementation in urological imaging workflows. The framework may serve as a candidate decision-support tool for radiologist-assisted interpretation of kidney CT findings, provided prospective validation is conducted. A revision-stage integrity audit found no byte-identical, pixel-identical, or repeated Roboflow source-identifier overlap across the predefined splits; however, two cross-split pairs showed very high structural similarity (SSIM >= 0.995), and patient-level independence could not be verified because patient identifiers were unavailable. Accuracy uncertainty was additionally quantified using 95% Wilson score confidence intervals.

Keywords: Renal cyst, Renal stone, Renal tumor, Optuna, Test-time augmentation, Clinical decision support

***Corresponding author:** Munzur University, Faculty of Economics and Administrative Sciences, Department of Health Management, 62600 Tunceli, Türkiye
E-mail: furkankarahuseyinoglu@munzur.edu.tr (F. KARAHUSEYINOGLU)
Mehmet Tevfik AGDAS https://orcid.org/0000-0002-5608-6240
Furkan KARAHUSEYINOGLU https://orcid.org/0000-0003-1641-2387

Received: July 26, 2026
Accepted: August 27, 2026
Published: September 15, 2026

Cite as: Agdas, M. T., & Karahuseyinoglu, F. (2026). AI-assisted multiclass kidney CT pathology classification for urological imaging: An optimized hybrid deep learning framework. *Black Sea Journal of Engineering and Science,* 9(5), 2800-2816.

1. Introduction

Kidney diseases remain a major global health burden because structural renal abnormalities, mass lesions, and obstructive conditions often progress silently until advanced stages. In clinical practice, CT is routinely used to identify cystic lesions, renal calculi, solid tumors, and normal reference anatomy; however, increasing imaging volume can reduce reporting efficiency and introduce inter-observer variability. Automated image analysis is therefore valuable not merely for acceleration, but also for standardizing preliminary interpretation and triage in renal imaging.

Recent advances in deep convolutional neural networks have markedly improved diagnostic image classification, yet performance in kidney imaging is still shaped by model selection, parameter configuration, class imbalance, and generalization strategy. Single-model systems are often vulnerable to architecture-specific blind spots: residual networks preserve gradient flow, EfficientNet achieves strong parameter efficiency, DenseNet improves feature reuse, and Xception benefits from depthwise separable convolutions. These differences motivate a hybrid design in which

complementary representations are fused rather than treated as competing alternatives.

Beyond classification accuracy, deep learning frameworks for urological CT imaging carry important translational implications. Reliable automated recognition of renal cysts, calculi, tumors, and normal tissue may support radiologist and urologist workflows by providing rapid preliminary interpretation, particularly in high-volume or resource-limited clinical settings. Any such application, however, requires robust clinical oversight, prospective validation, and careful integration with existing PACS and radiology information systems (Esteva et al., 2019; Topol, 2019).

The current study implements such a strategy on a four-class kidney CT dataset. Beyond reporting classification accuracy, the manuscript explicitly identifies the optimized parameters for each architecture, describes the dataset and class structure transparently, provides model-specific mathematical foundations, compares findings with the recent urological imaging literature, discloses methodological constraints, and avoids overstating clinical claims. This study aimed to develop and internally evaluate an optimized hybrid deep learning framework for multiclass kidney CT pathology classification in urological imaging, focusing on cyst, stone, tumor, and normal categories.

The principal contribution of the study is a fully tuned hybrid framework combining ResNet50V2, EfficientNetB3, DenseNet121, and Xception under an Optuna-guided AdamW training strategy with label smoothing and test-time augmentation. A second contribution is the explicit translation of these technical results into a clinically responsible interpretation: the proposed model classifies CT-visible kidney pathology patterns relevant to urological imaging - cyst, stone, tumor, and normal - and may support urological and radiological decision-making, but it does not directly assess renal functional parameters such as estimated glomerular filtration rate or serum creatinine. This distinction is essential for clinically responsible reporting.

This subsection reviews the literature on AI-assisted kidney CT classification in the context of urological imaging, followed by a focused review of the four backbone architectures used in the present study. The urological imaging part emphasizes prior work on renal cyst, stone, tumor, and normal CT classification, which directly informs the clinical framing of the proposed framework. The architecture-specific parts provide technical context for the hybrid design strategy.

1.1. Kidney CT Classification and Renal Imaging Context

Deep learning has already demonstrated promise in renal CT analysis. Han et al. (2019) classified renal cancer in three-phase CT images using a dedicated deep learning approach and showed that CT phase information can be exploited for robust renal lesion discrimination. Alzu'bi et al. (2022) introduced a kidney tumor CT dataset and reported that carefully prepared multiclass pipelines can support automated tumor assessment. Rajinikanth et al. (2023) distinguished healthy and cancerous renal CT images using fused deep features, highlighting the value of representation aggregation. Yao et al. (2023) developed and externally validated a multichannel deep learning model to differentiate fat-poor angiomyolipoma from renal cell carcinoma on unenhanced CT, thereby emphasizing the role of multi-branch architectures in differential diagnosis. More recently, Pimpalkar et al. (2025), Hossain et al. (2025), Kan et al. (2025), and Ye et al. (2025) further showed that fine-tuned or ensemble deep learning systems can reach clinically meaningful accuracy across renal pathology tasks, including kidney condition classification, tumor histology differentiation, and renal fibrosis assessment.

Collectively, these studies demonstrate the clinical relevance of AI-based kidney CT classification for urological imaging practice. The recognition of renal stones, cystic lesions, and solid tumors from CT images has direct diagnostic implications in urology and nephrology, as these conditions require distinct management pathways. However, a recurring limitation in the literature is reliance on single-architecture models with limited hyperparameter optimization, insufficient attention to class imbalance, and absence of systematic ensemble strategies. The present study addresses these gaps by combining four complementary architectures under a fully optimized training and fusion pipeline, with explicit attention to the four urologically relevant CT categories.

1.2. ResNet50V2-Based Studies

ResNet50V2 builds on the pre-activation residual learning principle introduced by He et al. (2016), in which identity mapping improves optimization depth and gradient propagation. In medical imaging, ResNet50V2 has been repeatedly used as a high-capacity baseline or transfer backbone. Boudi et al. (2024) reported improved Alzheimer's disease classification with fine-tuned ResNet50V2. Aby et al. (2025) evaluated ResNet50V2 for acute myeloid leukemia subtype classification from blood smear images. Al-Adhaileh et al. (2025) used transfer learning with ResNet50V2 for chest-disorder diagnosis from radiographs. Kreouzi et al. (2025) compared DenseNet121 and ResNet50V2 for melanoma detection, while Meeran (2025) proposed a Walrus-optimization-enhanced ResNet-50 framework for renal malignancy prediction. In another CT-oriented example, a recent ovarian tumor classification study employed ResNet50V2 together with Xception to enhance malignant lesion discrimination. Collectively, these studies justify the inclusion of ResNet50V2 as a strong residual baseline capable of preserving fine-grained anatomic information.

1.3. EfficientNetB3-Based Studies

EfficientNet scales network width, depth, and resolution in a compound and balanced way, which explains its sustained popularity in medical image classification after

the foundational work of Tan and Le (2019). Within this family, EfficientNetB3 offers an attractive trade-off between representational power and computational cost. Chen et al. (2023) combined EfficientNet-B0 with a gated recurrent unit for classifying precancerous cervical lesions. Zhang et al. (2024) proposed an attention-based glaucoma grading method built on EfficientNetB3. Al-Otaibi et al. (2024) reported an optimized EfficientNet-based gastrointestinal disease detector, and Aouadi et al. (2024) used EfficientNet for cervix cancer stage and grade prediction from diffusion-weighted imaging. Abd El-Aziz et al. (2025) developed an EfficientNet-B3 framework for multiclass endoscopic bladder tissue classification. Even when other EfficientNet variants were selected, the family consistently showed strong performance in kidney imaging, as seen in Hossain et al. (2025), where an EfficientNet-integrated system achieved high kidney CT classification accuracy. These findings support the role of EfficientNetB3 as the parameter-efficient branch of the present hybrid framework.

1.4. DenseNet121-Based Studies

DenseNet121, introduced by Huang et al. (2017), connects each layer to all subsequent layers in a feed-forward manner, thereby encouraging feature reuse and reducing redundancy. Its clinical imaging relevance is well documented. Zhou et al. (2022) reviewed DenseNet applications in medical image analysis and highlighted the architecture's stable convergence behavior. Gan et al. (2024) showed that DenseNet121 effectively identifies and classifies distal radius fractures on radiographs. Xu et al. (2024) reported that DenseNet121 achieved the best accuracy in classifying multi-feature fusion breast ultrasound images. Kumar et al. (2024) included DenseNet121 in a multi-cancer classification study and found it among the strongest models in validation. Mao et al. (2024) used a modified DenseNet-121 for mandibular fracture classification, and Hossain et al. (2025) further demonstrated that DenseNet121 remains highly competitive in kidney CT tasks. These studies justify DenseNet121 as the feature-reuse branch of the proposed hybrid architecture.

1.5. Xception-Based Studies

Xception, proposed by Chollet (2017), replaces conventional convolution with depthwise separable operations and has become especially attractive in medical imaging scenarios where computational efficiency and discriminative texture encoding must be balanced. Shaheed et al. (2023) combined an optimized Xception model with XGBoost for multiclass chest disease detection. Sathya et al. (2024) applied Xception to MRI-based brain tumor diagnosis and reported high-precision classification. Badr et al. (2025) incorporated a five-category Xception model within a broader dermatologic classification architecture. A recent ovarian tumor CT study likewise paired Xception with residual models for lesion classification, and several comparative oncology studies published in 2024 and 2025 continued to retain Xception among top-performing backbones. The literature therefore supports Xception as a complementary branch capable of capturing discriminative local textures that may differ from the representations learned by residual and densely connected networks.

2. Materials and Methods

2.1. Dataset and Class Structure

This study used a curated Roboflow multiclass export (version 4) for kidney CT classification. The effective working dataset comprised 9,574 images divided into four classes: Cyst, Normal, Stone, and Tumor. The dataset was split into 6,368 training images, 1,997 validation images, and 1,209 testing images. Because the notebook directly resolved the split directories and counted images recursively, the class frequencies reported below reflect the actual files used during model training and evaluation rather than metadata reported elsewhere.

The working image collection corresponded to the public Roboflow Universe project "Renal-Failure-Analysis" maintained by Mini Project (workspace: mini-project-6ilyj; project: renal-failure-analysis-5ymli), version 4 (Mini Project, 2023). The public project page lists 9,574 images in the Cyst, Normal, Stone, and Tumor classes and identifies the dataset as Public Domain. The public repository does not document the original clinical institution, the DICOM-level anonymization workflow, or reliable patient/study identifiers. The authors analyzed only the rendered image export supplied by the repository and did not have access to original DICOM headers or identifiable patient metadata. Consequently, the predefined folders are treated as image-level splits; patient-level disjointness and exclusion of different slices from the same patient across training, validation, and test partitions cannot be independently verified. Table 1 shows that the dataset was moderately imbalanced, with the Normal class dominating every split and Stone representing the most under-represented pathology. This imbalance is clinically plausible, because normal or non-complex findings tend to accumulate more frequently than stones or tumors in openly shared repositories. From a modeling standpoint, however, such a distribution risks biasing optimization toward majority classes; for that reason, class weighting was incorporated into training.

Figure 1 visually reinforces the imbalance already seen in Table 1. The largest volume belongs to the Normal class, whereas Stone has the smallest representation. The retained separation between train, validation, and test sets is important because it permits unbiased performance reporting; at the same time, the closely matched class order across splits indicates that the exported dataset was partitioned consistently rather than randomly distorted.

Figure 2 illustrates the visual heterogeneity of the classification problem. Cystic lesions tend to appear with relatively homogeneous low-density regions, stones

show small but high-contrast focal signals, and tumors may display more irregular textural and morphologic patterns. These qualitative differences help explain why a hybrid architecture is attractive: no single feature extractor is guaranteed to represent smooth lesions, sharp calcific findings, and irregular masses equally well.

2.2. Preprocessing, Augmentation, and Imbalance Handling

All images were resized to 300 x 300 pixels and fed through TensorFlow data pipelines. The augmentation block included random horizontal flipping, random rotation (0.08), random zoom (0.10), random contrast perturbation (0.10), and random translation (0.08, 0.08). Training labels were transformed into one-hot vectors, whereas validation and test sets were passed through non-augmenting evaluation preprocessing.

To address class imbalance, balanced class weights were computed from the raw training labels. The resulting weights were 0.8776 for Cyst, 0.6019 for Normal, 2.2743 for Stone, and 1.3168 for Tumor. The highest weight assigned to Stone reflects its relatively small sample count and forces the loss function to penalize misclassification of that class more strongly. This decision is methodologically important because Stone was also the class in which the weaker single models exhibited their few remaining errors.

2.3. Backbone Architectures and Mathematical Formulation

2.3.1. ResNet50V2

ResNet50V2 uses pre-activation residual units. Instead of learning a full transformation from scratch, each block learns a residual mapping F(.) that is added to the identity pathway. The simplified residual formulation can be written as follows (equation 1):

$$x_{l+1} = x_l + F(x_l, W_l) \quad (1)$$

In equation 1 x_l denotes the input feature map at layer l, and W_l denotes learnable parameters. The pre-activation design improves gradient flow and stabilizes deeper optimization. In renal CT classification, this is beneficial because subtle intensity differences can be preserved across many convolutional stages without excessive degradation.

2.3.2. EfficientNetB3

EfficientNetB3 belongs to the EfficientNet family, which scales depth, width, and resolution jointly rather than independently. The compound-scaling principle is expressed as (equations 2 and 3):

Depth= α^φ, width = β^φ, resolution = γ^φ (2)

α β^2 γ^2 ≈ 2 (3)

Table 1. Split-wise class distribution of the kidney CT dataset

Class	Train	Validation	Test	Overall
Cyst	1814	583	314	2711
Normal	2645	825	515	3985
Stone	700	235	139	1074
Tumor	1209	354	241	1804
Total	6368	1997	1209	9574

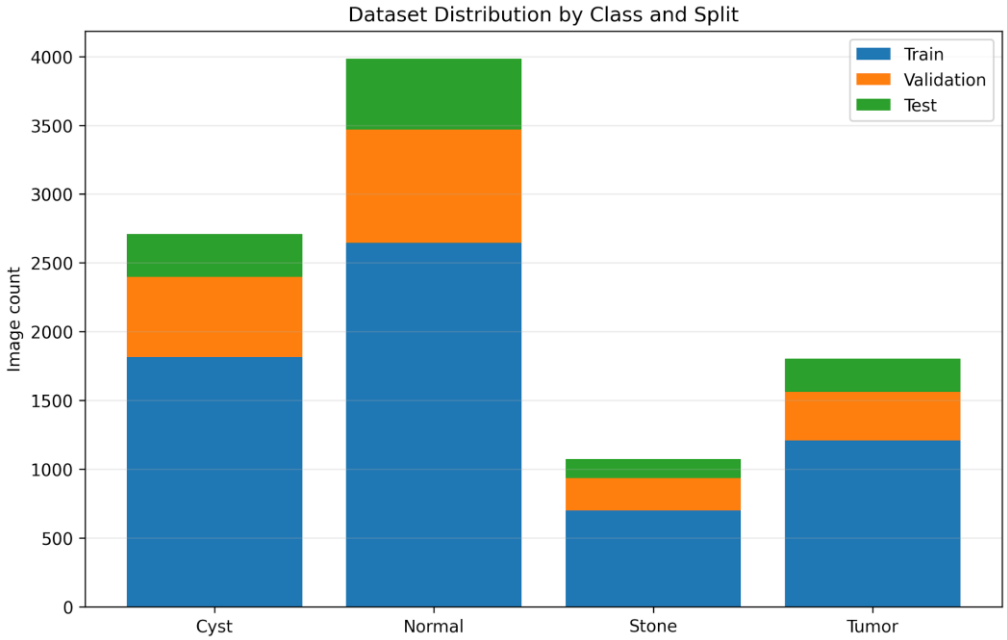


Figure 1. Graphical distribution of class frequencies across training, validation, and test splits.

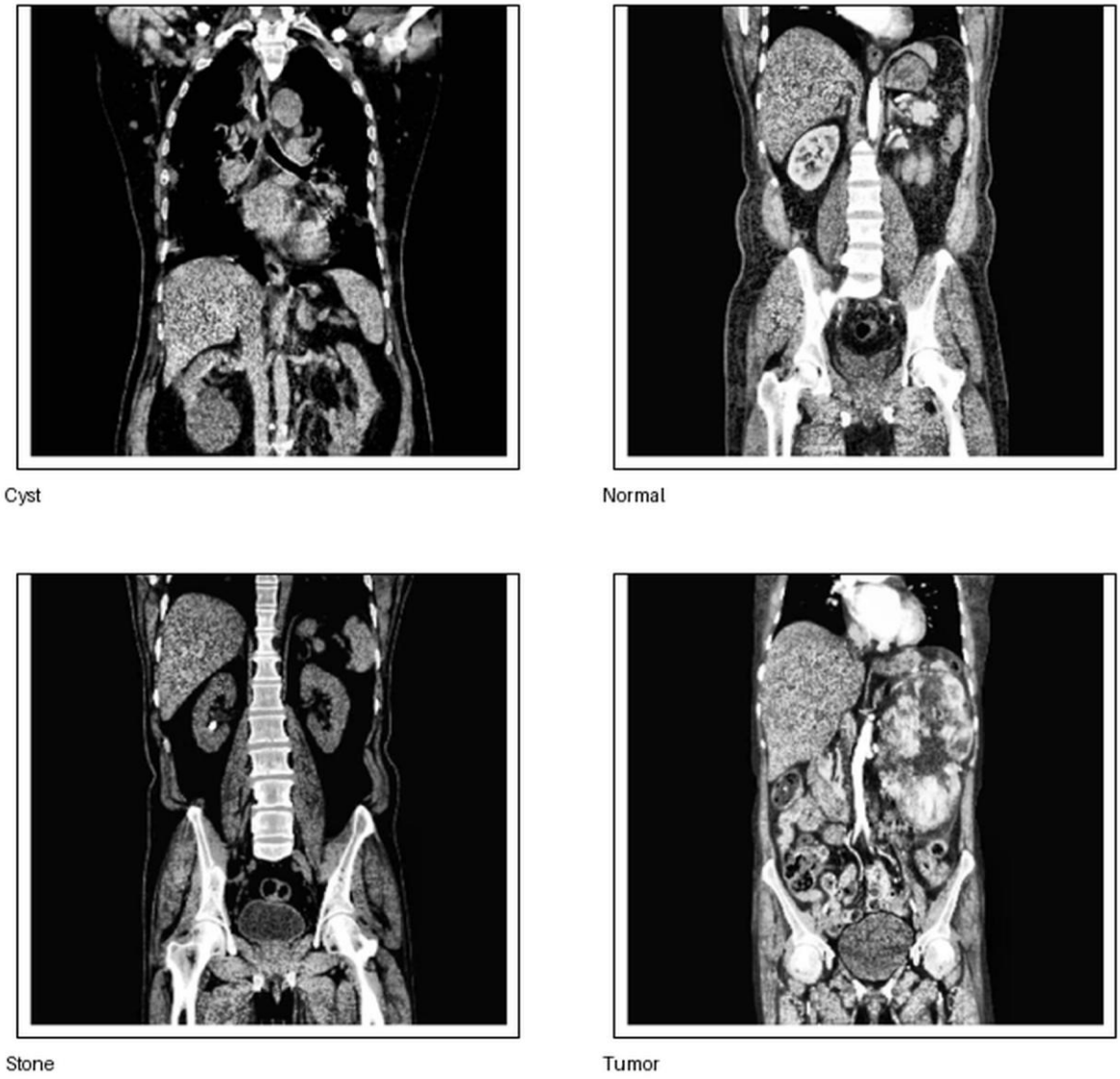


Figure 2. Representative sample CT images for the four target classes (Cyst, Normal, Stone, and Tumor).

The compound rule balances representational power and computational load. In the present study, EfficientNetB3 served as the efficiency-oriented branch of the ensemble, allowing the system to learn compact yet expressive lesion representations without the parameter growth required by some heavier architecture.

2.3.3. DenseNet121

DenseNet121 introduces dense connectivity, meaning that each layer receives the outputs of all previous layers through concatenation. This can be summarized as (equation 4):

$$x_l = H_l([x_0, x_1, \dots, x_{l-1}]) \quad (4)$$

Here, [.] denotes feature-map concatenation and $H_l(.)$ represents the nonlinear transformation at layer l. Dense connectivity enhances feature reuse and improves gradient propagation. For kidney CT data, such reuse is valuable because both coarse structural context and small local intensity patterns may contribute jointly to pathology recognition.

2.3.4. Xception

Xception replaces standard convolution with depthwise separable convolution, decomposing the operation into channel-wise spatial filtering followed by pointwise channel mixing. A simplified representation is (equation 5):

$$Y = \mathrm{PW}(\mathrm{DW}(X)) \quad (5)$$

In equation 5, DW(.) denotes depthwise convolution and PW(.) denotes pointwise convolution. This decomposition reduces parameter redundancy while retaining strong texture sensitivity. In renal CT interpretation, that property is especially useful for classes such as Stone and Tumor, where small but discriminative local structures may alter the decision boundary.

2.4. Classifier Head, Loss Function, and Optimization

Each backbone fed a classifier head whose hyperparameters were tuned independently by Optuna. The candidate search space included the primary dropout rate, a secondary dropout rate, dense-layer width (128, 256, or 384 units), learning rate, and label-smoothing coefficient. Batch normalization was retained as a fixed design choice across models. Training used

AdamW with a weight decay of 1e-5 and categorical cross-entropy with label smoothing. Validation accuracy was the optimization objective, and the parameter set with the highest validation accuracy was retained for the subsequent two-stage training procedure. The archived run available for revision preserved the selected hyperparameters and best validation scores but did not preserve a verifiable record of the total Optuna trial count. To avoid an unsupported numerical claim, no retrospective trial number is reported (equation 6).

$$L = -\Sigma_i \, y_i^{LS} \log(p_i) \quad (6)$$

In equation 6, y_i^{Ls} denotes the label-smoothed target distribution and p_i denotes the predicted class probability. Two-stage training was applied for every model: an initial transfer-learning phase of 30 epochs and a fine-tuning phase of 30 epochs, during which the backbone was unfrozen except for the earliest layers. Early stopping, model checkpointing, and learning-rate reduction on plateau were used to improve optimization stability.

Table 2. Optuna-optimized hyperparameters obtained for each backbone

Model	Dropout	Dropout-2	Dense units	Learning rate	Label smoothing	Best val. score
ResNet50V2	0.2672	0.3036	384	2.384e-4	0.0917	0.8898
EfficientNetB3	0.2264	0.1995	384	4.115e-4	0.0999	0.8508
DenseNet121	0.3276	0.1641	256	2.182e-4	0.0477	0.8162
Xception	0.2286	0.1280	128	3.724e-4	0.0690	0.8603

Dropout-2, secondary dropout rate; Dense units, number of units in the dense classifier layer; Learning rate expressed in scientific notation; Best val. score, best validation score obtained during Optuna optimization.

Table 2 confirms that the optimized parameter set was architecture-specific rather than universal. ResNet50V2 and EfficientNetB3 benefited from wider dense layers, DenseNet121 preferred more conservative smoothing, and Xception converged with a smaller dense layer and a relatively high learning rate. These differences support the central premise of the study: hybrid performance improves when each branch is tuned according to its own optimization behavior rather than forced into a shared parameter template.

2.5. Hybrid decision strategies

Two ensemble strategies were used after training the single models. The first was validation-driven weighted soft voting, where each model's posterior probabilities were multiplied by a normalized validation-derived weight and then summed. The second was stacking, where model probability vectors were concatenated and passed to a logistic-regression meta-learner (equations 7 and 8).

$$p_{ens} = \Sigma_k \, w_k \, p_k, \text{ where } \Sigma_k \, w_k = 1 \quad (7)$$

$$\hat{y} = \text{softmax}(Wz + b), \; z = [p_1 \parallel p_2 \parallel p_3 \parallel p_4] \quad (8)$$

The optimized weights used in the weighted ensemble were 0.2617 for ResNet50V2, 0.2428 for EfficientNetB3, 0.2337 for DenseNet121, and 0.2618 for Xception. The stacking learner used multinomial logistic regression with a maximum of 2,000 iterations. Test-time augmentation was enabled for single-model and hybrid prediction, with five stochastic inference rounds averaged per sample. In the archived pipeline implementation available for reproducibility checking, TTA reused the same augmentation block used for training: horizontal flipping, rotation (0.08), zoom (0.10), contrast perturbation (0.10), and translation (0.08, 0.08). Posterior probabilities were averaged across the five inference passes. The stacking meta-learner was fitted only on concatenated validation-set probability vectors and validation labels; test-set labels were not used to fit the meta-learner.

2.6. Data-integrity and cross-split leakage audit

In response to the revision request, the original Roboflow version-4 export was re-audited without retraining the classification models. Cross-split exact duplication was assessed using SHA-256 file hashes. A second check compared standardized grayscale-resized pixel arrays to identify pixel-identical images despite file-level metadata differences. Roboflow source identifiers were also normalized by removing export-specific hash suffixes. Perceptual hashing was used only as a screening stage because visually similar CT slices can collide under pHash; pHash candidates were therefore re-evaluated with the Structural Similarity Index (SSIM). No cross-split SHA-256, pixel-identical, or Roboflow source-identifier duplicates were detected. Among 4,208 pHash-screened cross-split candidate pairs, two pairs reached SSIM >= 0.995. These two pairs were treated conservatively as near-duplicate candidates rather than automatically classified as leakage. Because the public export does not provide reliable patient or study identifiers, patient-level disjointness and same-patient adjacent-slice exclusion cannot be fully verified.

2.7. Statistical Uncertainty and Comparison Scope

To supplement point estimates, two image-level uncertainty analyses were performed for test-set accuracy. First, a nonparametric percentile bootstrap with 10,000 resamples (seed = 42) was applied to the binary correctness vector reconstructed from each model's exact number of correct and incorrect test

classifications. For accuracy, the ordering of this binary vector is immaterial to the bootstrap distribution. Second, 95% Wilson score confidence intervals (Wilson, 1927) were calculated as a boundary-aware sensitivity analysis using the observed number of correct classifications over 1,209 test images. Because patient identifiers were unavailable, both intervals should be interpreted at the image level and may understate uncertainty if multiple slices originate from the same patient. Pairwise McNemar testing was additionally performed only where it could be reconstructed exactly from the archived aggregate results: each perfect model was compared with DenseNet121 and EfficientNetB3. Because a perfect model is correct on every test image, the discordant count is fully determined by the number of errors in the non-perfect model. Exact two-sided binomial McNemar p-values were used and Holm correction was applied to the two unique non-perfect comparisons. The DenseNet121-versus-EfficientNetB3 comparison and DeLong ROC-AUC comparisons were not reconstructed because those procedures require paired per-image predictions and, for DeLong testing, sample-level continuous scores that were not retained in the archived run. For the same reason, confidence intervals for class-specific sensitivity, specificity, and F1-score were not retrospectively reconstructed from rounded aggregate tables. No values were imputed from rounded summary statistics.

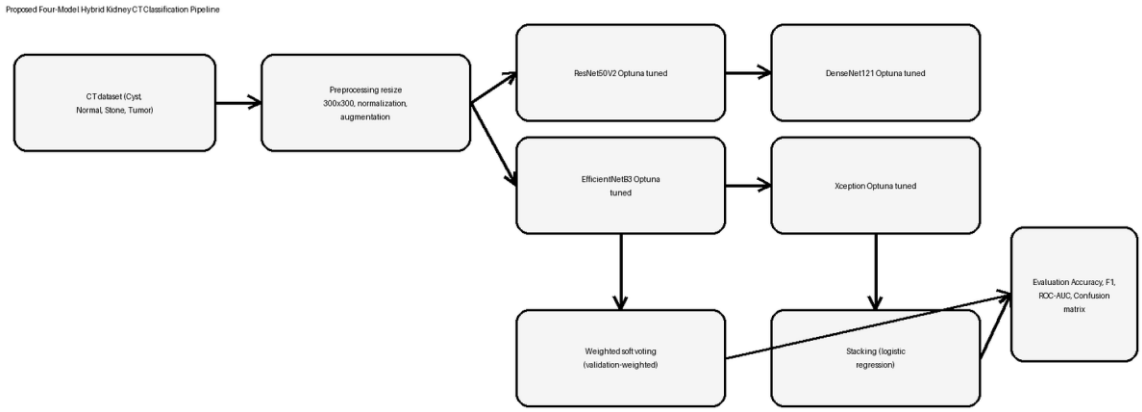


Figure 3. End-to-end methodological flow of the proposed hybrid kidney CT classification framework.

Figure 3 summarizes the analytical logic of the study from data ingestion to final decision fusion. The diagram emphasizes that hybrid prediction was not an afterthought added after training; rather, the entire pipeline was designed around complementary backbone specialization, validation-informed weighting, and a secondary stacking stage. This architecture-level framing differentiates the study from many single-model reports in the renal imaging literature.

3. Results

3.1. Learning Dynamics

Figure 4 indicates that all four backbones converged rapidly and maintained stable validation behavior during fine-tuning. The curves do not show the severe oscillation or late-epoch divergence typically associated with poorly tuned learning rates. ResNet50V2 and Xception display especially clean convergence profiles, which is consistent with their later perfect held-out performance. The figure also supports the practical value of architecture-specific Optuna tuning.

3.2. Overall Predictive Performance

Table 3 presents an unusually strong outcome: four of the six evaluated systems achieved perfect performance on the internal same-source held-out test set, while the remaining two operated at near-ceiling levels. These results confirm the technical success of the tuned hybrid pipeline within the analyzed dataset; however, they require cautious interpretation. Perfect scores on a same-source held-out split may reflect strong class separability within the dataset rather than guaranteed generalization to independent cohorts. External multi-center validation is required before any clinical performance claims can be substantiated. These caveats are addressed explicitly in the discussion and limitations sections. Figure 5 makes the rank ordering easier to inspect than Table 3. DenseNet121 and EfficientNetB3 were already highly accurate, yet the best results were obtained by ResNet50V2, Xception, and both hybrid strategies. This pattern suggests that the hybrid framework did not compensate for weak learners; instead, it consolidated already high-quality but partially heterogeneous predictions into a maximally stable final decision.

3.3. Error Structure and Confusion Matrices

Figure 6 reveals that the residual errors of the weaker single models were concentrated primarily in the Cyst-Stone interface. Tumor and Normal were classified almost flawlessly even before ensembling, whereas Stone remained the most delicate class. The hybrid models eliminated these residual confusions entirely, which supports the argument that complementary feature extraction is most valuable at the class boundaries where single architectures remain uncertain.

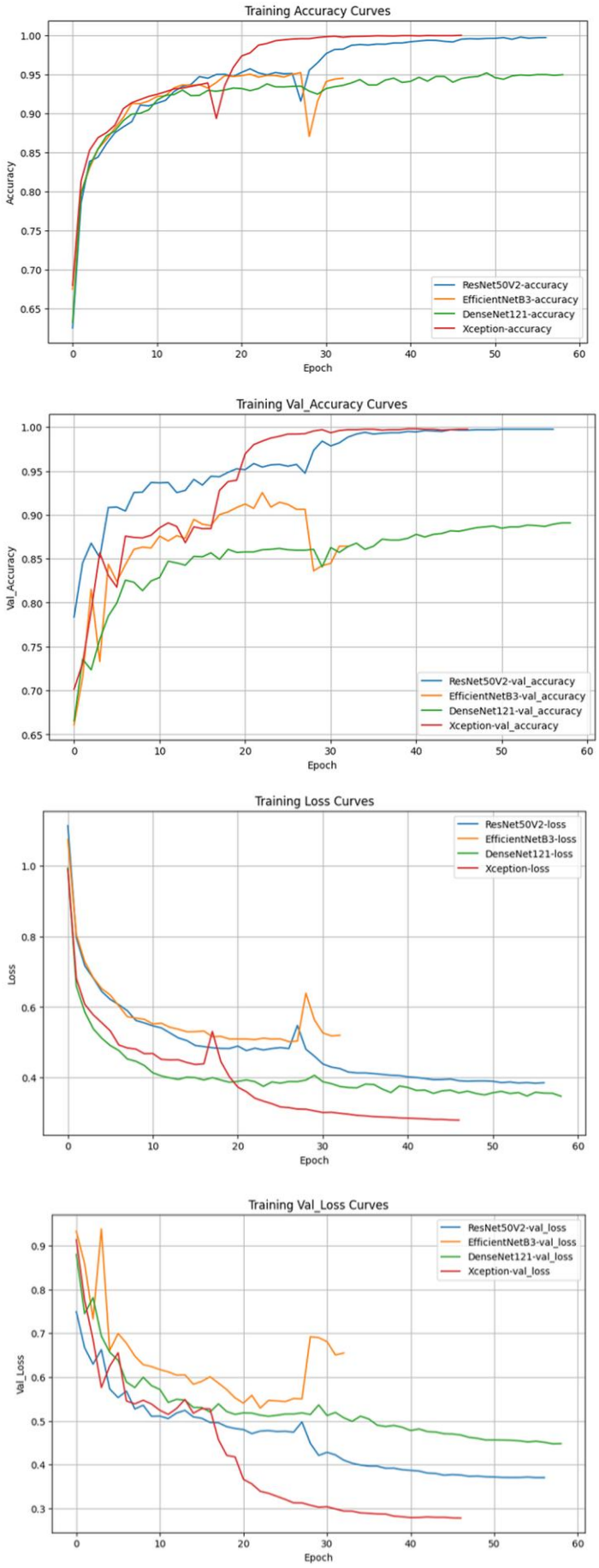


Figure 4. Training and validation accuracy/loss trajectories for the four optimized backbone models.

Table 3. Test-set performance of single models and hybrid models

Model	Accuracy	Precision (w)	Recall (w)	F1-score (w)	ROC-AUC (OvR)
ResNet50V2	1.0000	1.0000	1.0000	1.0000	1.0000
Xception	1.0000	1.0000	1.0000	1.0000	1.0000
Weighted Ensemble	1.0000	1.0000	1.0000	1.0000	1.0000
Stacking	1.0000	1.0000	1.0000	1.0000	1.0000
DenseNet121	0.9959	0.9959	0.9959	0.9959	0.9999
EfficientNetB3	0.9934	0.9934	0.9934	0.9934	0.9999

w= weighted average; ROC-AUC (OvR) = area under the receiver operating characteristic curve computed one-vs-rest; F1-score= harmonic mean of precision and recall.

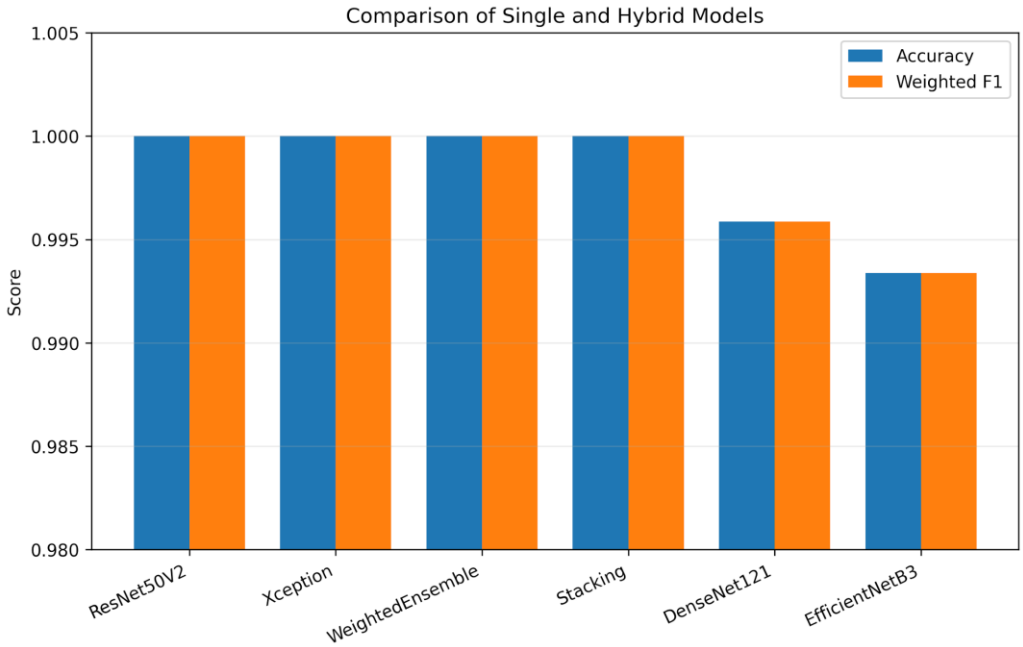


Figure 5. Comparative performance plot across single and hybrid models.

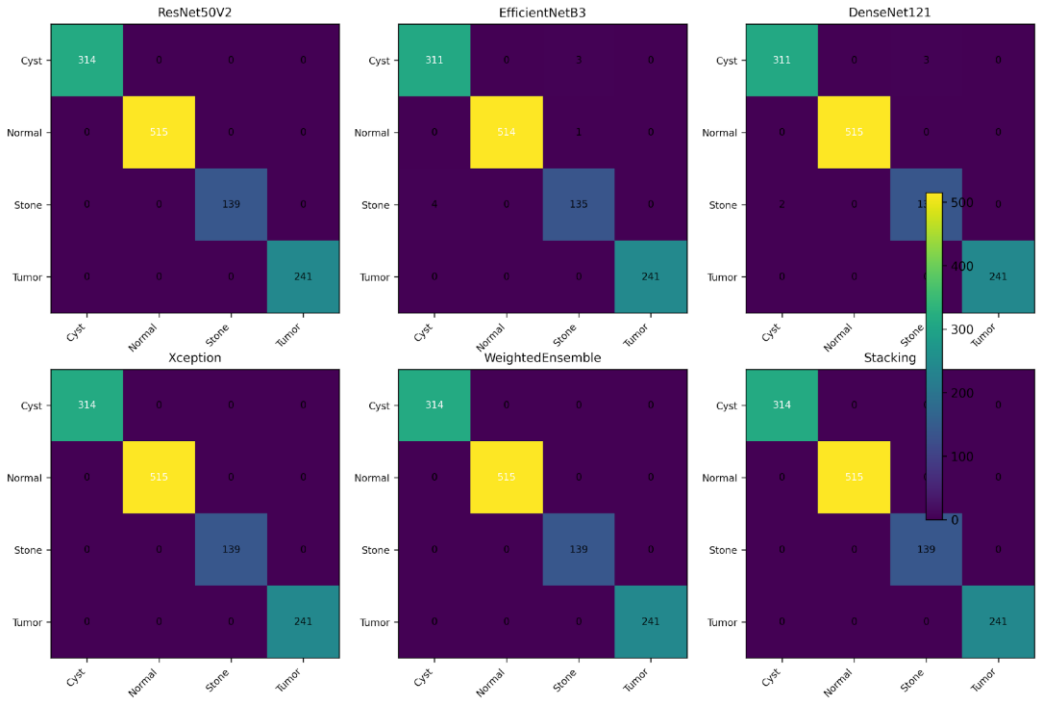


Figure 6. Confusion matrices for all single models and hybrid models on the test set.

Table 4. Class-wise precision, recall, and F1-score for the main error-bearing models and the final weighted ensemble

Model	Class	Precision	Recall	F1-score	Support
EfficientNetB3	Cyst	0.9873	0.9904	0.9889	314
EfficientNetB3	Normal	1.0000	0.9981	0.9990	515
EfficientNetB3	Stone	0.9712	0.9712	0.9712	139
EfficientNetB3	Tumor	1.0000	1.0000	1.0000	241
DenseNet121	Cyst	0.9936	0.9904	0.9920	314
DenseNet121	Normal	1.0000	1.0000	1.0000	515
DenseNet121	Stone	0.9786	0.9856	0.9821	139
DenseNet121	Tumor	1.0000	1.0000	1.0000	241
Weighted Ensemble	Cyst	1.0000	1.0000	1.0000	314
Weighted Ensemble	Normal	1.0000	1.0000	1.0000	515
Weighted Ensemble	Stone	1.0000	1.0000	1.0000	139
Weighted Ensemble	Tumor	1.0000	1.0000	1.0000	241

Support= number of test images in each class; F1-score= harmonic mean of precision and recall.

Table 4 quantifies the class-specific behavior already suggested by the confusion matrices. Stone had the lowest F1-score in the two non-perfect single models, whereas Tumor and Normal were essentially solved. The weighted ensemble corrected these remaining mistakes completely. This is precisely the type of improvement expected from well-designed hybrid architecture: the benefit is not only higher average accuracy but also the removal of localized class-boundary errors that matter in clinical decision support.

3.4. ROC and Precision-Recall Analysis

Figure 7 complements accuracy-based reporting by showing near-ideal discrimination across thresholds. The ROC and precision-recall curves remain close to the upper-left and upper-right boundaries, respectively, which is consistent with the near-perfect multiclass metrics. Because class imbalance was present, the precision-recall plots are especially informative: they indicate that high sensitivity was not purchased at the cost of excessive false positives.

3.5. Revision-Stage Integrity and Uncertainty Analyses

The cross-split audit found zero exact SHA-256 duplicate groups, zero standardized pixel-identical duplicate groups, and zero repeated normalized Roboflow source identifiers. Perceptual hashing produced 1,334 cross-split screening groups, reflecting the strong visual similarity typical of CT slices. SSIM was then calculated for 4,208 cross-split candidate pairs: two pairs reached SSIM >= 0.995, 481 fell between 0.98 and 0.995, the maximum SSIM was 0.997779, and the median was 0.930144. Accordingly, the revision audit did not identify direct duplicate leakage, but it does not establish patient-level independence. The two very-high-SSIM pairs are reported as unresolved near-duplicate candidates and are incorporated into the study limitations.

Image-level percentile bootstrap intervals and Wilson intervals were added to quantify uncertainty in the same-source test accuracy. The 10,000-resample bootstrap was computed from reconstructed binary correctness vectors using the exact observed correct/error counts. For the four perfect systems, the percentile bootstrap distribution was degenerate at 1.0000 because every image was classified correctly, whereas the Wilson interval remained 0.9968-1.0000 and therefore provides a more conservative boundary-aware uncertainty range. DenseNet121 had a bootstrap 95% interval of 0.9917-0.9992 and a Wilson interval of 0.9904-0.9982; EfficientNetB3 had a bootstrap interval of 0.9884-0.9975 and a Wilson interval of 0.9870-0.9966. These intervals describe image-level sampling uncertainty only and do not resolve possible patient-level clustering.

Limited exact McNemar comparisons were possible from the archived results. All four perfect systems were correct on every test image. Relative to a perfect-reference model, DenseNet121 had five discordant errors (exact P=0.0625), whereas EfficientNetB3 had eight discordant errors (exact P=0.0078). After Holm correction for the two unique comparisons, the adjusted p-values were 0.0625 and 0.0156, respectively. Thus, the improvement over EfficientNetB3 was statistically detectable at the 0.05 level, while the difference from DenseNet121 was not. Pairwise comparison between the two non-perfect models could not be reconstructed without sample-level prediction identities.

3.6. Comparative Positioning Against Related Studies

Table 8 situates the present study within the contemporary renal imaging literature. In contrast to many prior reports that focused on a single lesion type or a single backbone family, the current work deliberately integrates four optimized architectures and two fusion strategies. Its headline metrics are stronger than those reported in most comparable studies, but this superiority should be interpreted in light of evaluation context. Specifically, the absence of external validation means the present results are best understood as a highly promising internal benchmark that requires external multi-center validation before any claim of cross-institutional generalizability can be made.

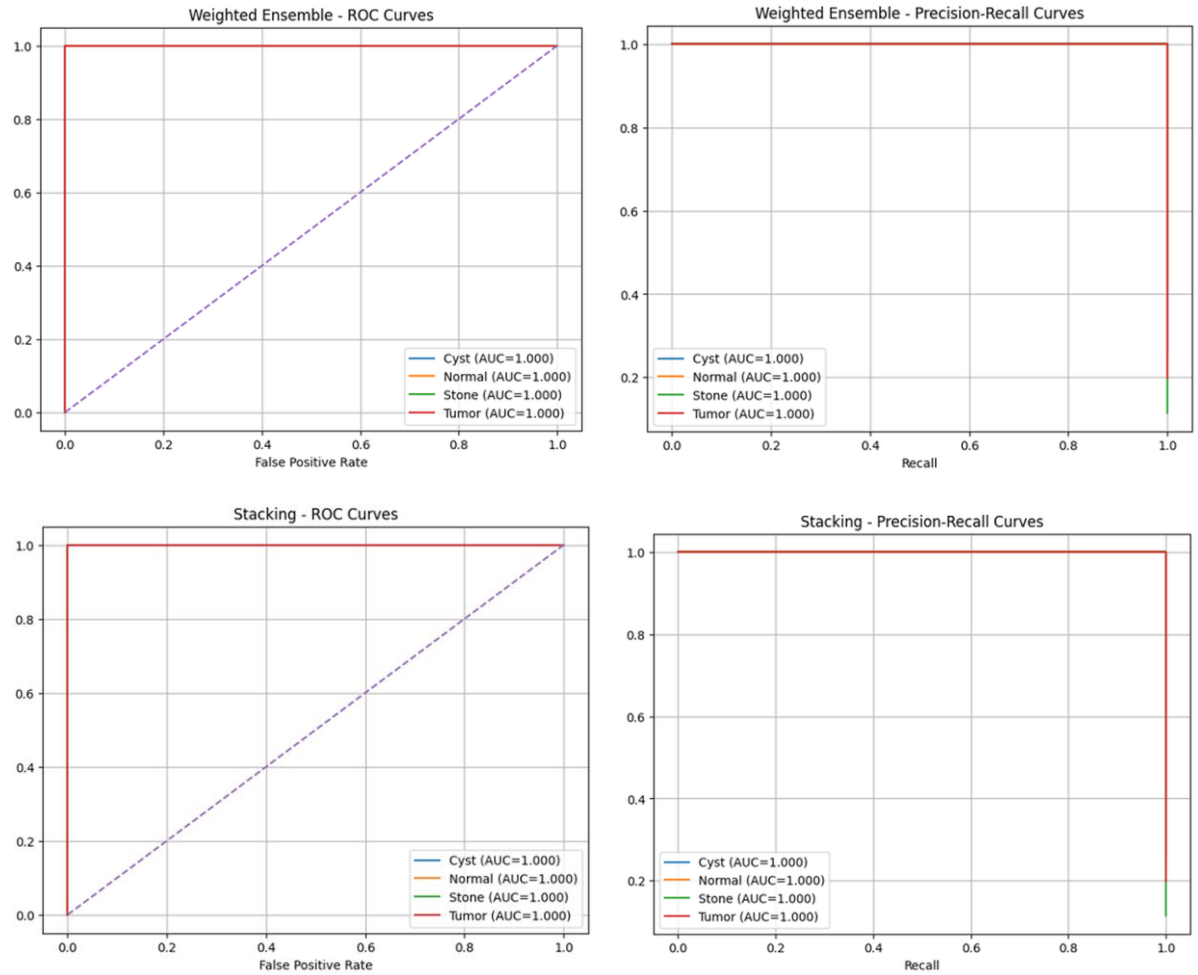


Figure 7. ROC and precision-recall curves for the evaluated backbone and hybrid models.

Table 5. Revision-stage cross-split data-leakage audit

Audit component	Cross-split result	Interpretation
SHA-256 exact file hash	0 groups	No byte-identical duplicate detected
Standardized pixel hash	0 groups	No pixel-identical duplicate detected
Normalized Roboflow source ID	0 groups	No repeated source identifier detected
pHash screening	1,334 groups	Screening only; CT similarity can cause collisions
SSIM verification	4,208 pairs checked; 2 >= 0.995	Two unresolved very-high-similarity candidates

Table 6. Test-set accuracy with image-level 95% percentile-bootstrap and Wilson score confidence intervals

Model	Accuracy	Bootstrap 95% CI	Wilson 95% CI
ResNet50V2	1.0000	1.0000-1.0000	0.9968-1.0000
Xception	1.0000	1.0000-1.0000	0.9968-1.0000
Weighted Ensemble	1.0000	1.0000-1.0000	0.9968-1.0000
Stacking	1.0000	1.0000-1.0000	0.9968-1.0000
DenseNet121	0.9959	0.9917-0.9992	0.9904-0.9982
EfficientNetB3	0.9934	0.9884-0.9975	0.9870-0.9966

Table 7. Exact McNemar comparisons against a perfect-reference model.

Comparison	Discordant errors	Exact p	Holm-adjusted p	Interpretation
Perfect reference vs DenseNet121	5 vs 0	0.0625	0.0625	Not significant at 0.05
Perfect reference vs EfficientNetB3	8 vs 0	0.0078	0.0156	Significant at 0.05

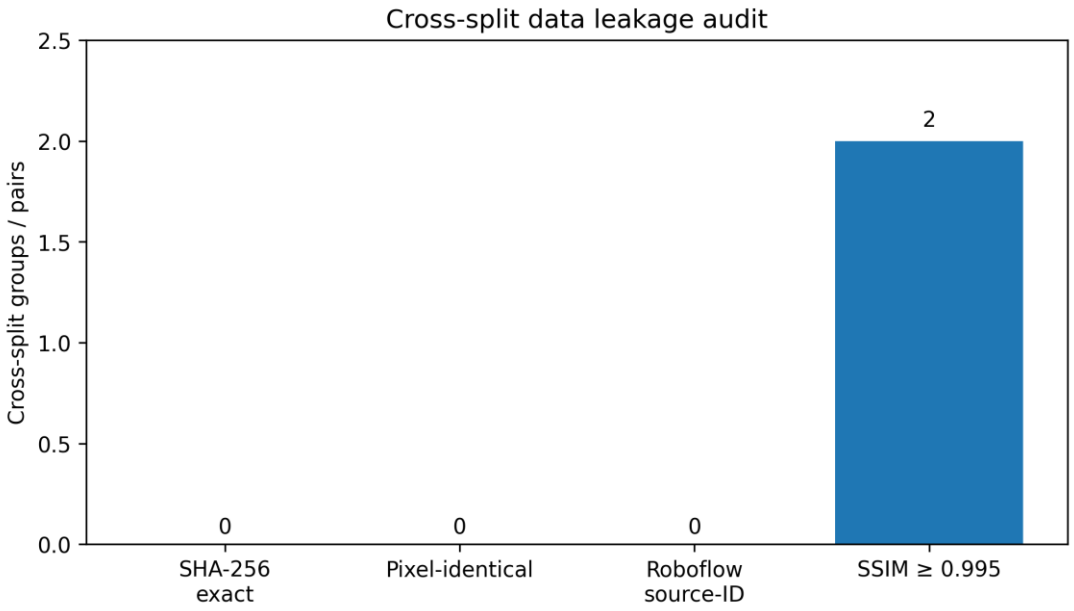


Figure 8. Cross-split data-leakage audit. Direct exact, pixel-identical, and source-ID duplicate counts were zero; two SSIM-verified pairs remained above the 0.995 near-duplicate screening threshold.

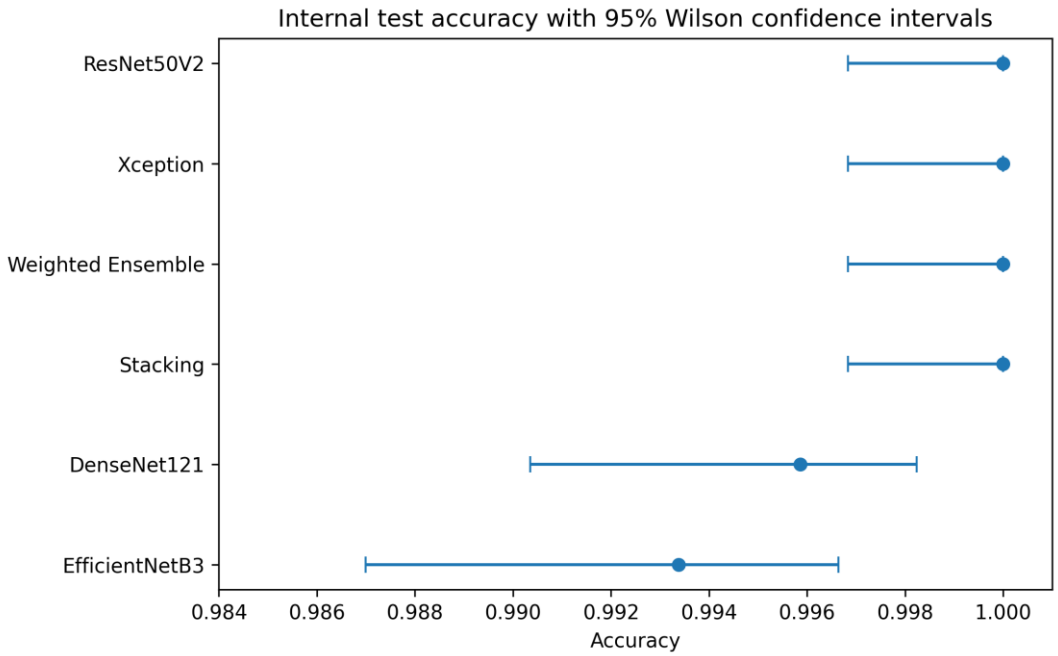


Figure 9. Internal test accuracy with 95% Wilson score confidence intervals for the six evaluated systems.

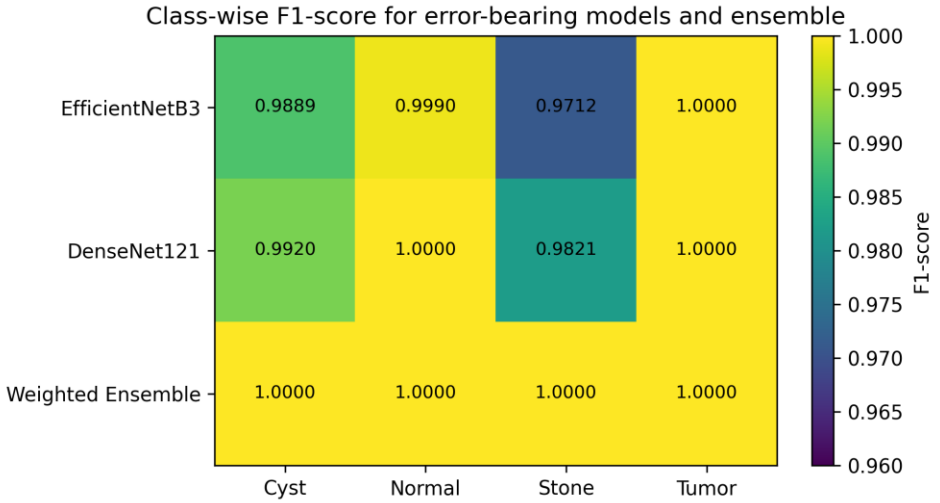


Figure 10. Class-wise F1-score heat map for EfficientNetB3, DenseNet121, and the weighted ensemble, highlighting Stone as the most error-prone class before ensembling.

Table 8. Literature-based comparison between the present study and representative renal imaging studies

Study	Task	Architecture	Key result
Han et al. (2019)	Renal cancer classification in 3-phase CT	Deep CNN	High renal cancer classification accuracy
Alzu'bi et al. (2022)	Kidney tumor CT classification	DL-based tumor classifiers	Successful multiclass tumor detection
Rajinikanth et al. (2023)	Healthy/cancer renal CT discrimination	Fused deep features	Strong healthy/cancer discrimination
Yao et al. (2023)	AML vs. RCC on unenhanced CT	Multichannel DL	External validation achieved
Hossain et al. (2025)	Kidney disease CT classification	DenseNet121 + EfficientNet	Integrated CT classifier effective
Pimpalkar et al. (2025)	Kidney condition CT classification	Fine-tuned transfer models	High early detection performance
Present study	4-class kidney CT pathology classification	ResNet50V2 + EfficientNetB3 + DenseNet121 + Xception + hybrid fusion	Perfect hold-out performance in same-source test split

CT= computed tomography, DL= deep learning, CNN= convolutional neural network, AML= angiomyolipoma, RCC= renal cell carcinoma.

4. Discussion

The findings demonstrate that architecture diversity and parameter optimization can work synergistically in kidney CT classification. ResNet50V2 contributed stable residual learning, EfficientNetB3 provided computationally efficient compound scaling, DenseNet121 encouraged extensive feature reuse, and Xception captured high-value local textures through depthwise separable operations. Because these inductive biases are not identical, their fusion helps reduce model-specific blind spots. The perfect performance of the two hybrid strategies suggests that the final decision boundary was strengthened by complementary errors rather than inflated by redundancy.

A second important result is that the parameter configuration mattered substantially. The best-performing settings were not uniform across backbones; dropout rates, dense-layer widths, learning rates, and label-smoothing coefficients all shifted as a function of architecture. In the present study, the tuned parameters were not peripheral; they were part of the methodological contribution.

Clinically, the study offers a contribution to urological CT imaging by automating the recognition of CT-visible renal pathology categories - cysts, stones, tumors, and normal parenchyma - that carry distinct clinical significance and require separate management pathways in urological practice. These categories are not interchangeable, and each has direct implications for patient triage, referral, and follow-up. Therefore, the developed model responds to the renal-failure problem indirectly but meaningfully: it supports earlier and more standardized recognition of CT-visible urological pathologies - cystic lesions, calculi, and solid tumors - that carry direct clinical significance in

urology and nephrology.

At the same time, results of this magnitude require methodological caution. Four systems reached perfect performance on the internal held-out test set, which is unusual in medical imaging and should not be treated as evidence of deployment readiness. Such results may reflect excellent class separability within the dataset, but they can also arise when train/validation/test splits remain too homogeneous in acquisition profile, preprocessing lineage, or disease presentation. For that reason, the present article frames the findings as strong internal evidence rather than a final clinical claim.

4.1. Translational implications for urological imaging practice

Beyond its technical contributions, the hybrid deep learning framework presented in this study has direct translational relevance to urological imaging practice. Systems capable of reliably distinguishing cystic, calculous, tumoral, and normal renal CT patterns may support radiologist and urologist decision-making by providing rapid automated preliminary classification. For clinicians managing renal stone disease, cystic lesions, or suspected renal malignancy, a validated AI-assisted classification layer could reduce interpretation variability, particularly for equivocal cases or in settings with high imaging volumes (Esteva et al., 2019; Rajpurkar et al., 2022). Physician oversight must remain central to any such workflow; the proposed framework is intended as a decision-support aid, not an autonomous diagnostic system.

From a urological practice standpoint, integration of validated CT classification tools into existing hospital imaging workflows - via PACS or RIS platforms - could support urologist-radiologist communication by providing structured, automated preliminary findings. Structured CT reporting for kidney pathology categories (Cyst, Stone, Tumor, Normal) aligns with existing urological guidelines and may facilitate more consistent documentation and follow-up planning. In settings where specialist urology or radiology capacity is limited, such tools may also support earlier identification of patients requiring nephrology or oncology referral. Realizing these benefits would require prospective clinical validation, attention to DICOM interoperability, and clearly defined protocols for human oversight (Topol, 2019).

It must be emphasized that this study does not itself constitute a prospective clinical evaluation or a urological outcomes study. The present work is an internal computational validation using a curated dataset split. No patient-level clinical outcome data, biochemical renal function parameters (e.g., eGFR, serum creatinine), or prospective workflow measurements were obtained. The translational implications described in this section are therefore hypothesis-generating in nature. Prospective, multi-institutional studies that embed validated models within real urological imaging workflows and measure diagnostic concordance with clinical outcomes will be necessary before the proposed benefits can be claimed as demonstrated (Rajpurkar et al., 2022).

4.2. Limitations

Several important limitations apply to this study and should be considered when interpreting the reported findings. First, the model classifies CT-visible structural categories (Cyst, Normal, Stone, and Tumor) and does not directly assess renal functional status. No biochemical parameters such as estimated glomerular filtration rate or serum creatinine were included, and no claim of direct assessment of renal functional status is made. The classification categories are urologically and radiologically meaningful, but clinical management decisions require integration with full clinical assessment.

Second, no external validation cohort was available for this study. All reported results are from an internal same-source dataset split, and k-fold cross-validation was not applied. The dataset may reflect a degree of homogeneity in image acquisition parameters, patient population, and CT protocol that does not represent the full spectrum of real-world urological imaging variability. Consequently, the study does not claim population-level generalizability. Future work should include external multi-center validation with scanner-diverse cohorts and independent test sets. In the revision, a dedicated cross-split integrity audit found no exact file, pixel-identical, or repeated Roboflow source-ID duplicates, but two cross-split pairs had SSIM >= 0.995 and remain unresolved near-duplicate candidates. Because the public export contains no reliable patient or study identifiers, it was not possible to prove patient-level separation or to exclude different slices from the same patient across partitions. The predefined folders must therefore be interpreted as image-level splits, not verified patient-level splits.

K-fold cross-validation and independent external-cohort validation were requested during peer review. A complete four-backbone re-training under k-fold cross-validation was not introduced into the revised manuscript because the original analysis was based on a fixed predefined split and a full re-training attempt exceeded the available computational runtime without completing consistently across all architectures. Reporting partial fold results would mix protocols and could be misleading. Likewise, no compatible independent external cohort with matched four-class labels, acquisition documentation, and non-overlapping provenance was available for a defensible zero-shot external validation. These analyses therefore remain important future validation requirements rather than being represented as completed experiments.

The archived run did not retain the full paired per-image probability matrices for every model. Therefore, the revision reports only inferential quantities that could be computed without inventing missing sample-level scores. A 10,000-resample image-level percentile bootstrap and

Wilson 95% confidence intervals were calculated for overall accuracy from the exact correct/error counts. Exact McNemar testing was possible only for comparisons against the four perfect models, because their per-image correctness vector contains no errors; these limited comparisons were added. A complete DenseNet121-versus-EfficientNetB3 McNemar test and DeLong ROC-AUC confidence intervals/comparisons could not be reconstructed without the missing paired per-image predictions and continuous probability scores. These unavailable inferential quantities were not estimated from aggregate AUC values.

Third, explainability analysis via Grad-CAM was attempted but did not yield stable heatmaps due to null gradients at the gradient extraction stage for the tested model configurations. Accordingly, no Grad-CAM visualizations are presented in this study. The absence of interpretability outputs is itself a limitation: without spatially localized attention maps, it is not possible to confirm whether the models focus on clinically meaningful renal regions during classification. A revised explainability pipeline using gradient-free methods or alternative interpretability approaches is recommended for future work.

Fourth, the dataset used in this study was a curated Roboflow export rather than a large-scale external benchmark collection. The reported results are therefore fully valid for the analyzed dataset but not generalizable to all renal CT repositories without further validation. No prospective clinical workflow testing, no direct patient-level clinical outcome data, and no clinical reader studies were conducted. Multi-center validation involving diverse scanner platforms, CT protocols, and patient populations will be necessary to assess real-world generalizability prior to any clinical implementation.

Finally, the clinical decision-support implications discussed in this article are inferential and hypothesis-generating. This study did not collect hospital-level operational data, patient outcome data, or clinical reader performance metrics. No comparison with radiologist or urologist interpretation was performed. Prospective, workflow-embedded, multi-institutional studies that measure diagnostic concordance, clinical impact, and urological outcome endpoints will be required to establish the clinical utility of AI-assisted kidney CT classification in urological imaging practice.

5. Conclusion

This study developed and evaluated a high-performance hybrid kidney CT classification framework built on Optuna-optimized ResNet50V2, EfficientNetB3, DenseNet121, and Xception backbones. By combining AdamW optimization, label smoothing, class weighting, two-stage fine-tuning, and test-time augmentation, the system achieved perfect held-out performance for ResNet50V2, Xception, and both hybrid fusion strategies, while the remaining single models also operated at near-ceiling accuracy.

The most important scientific result is not merely the final classification score but the demonstration that architecture-specific parameter tuning and complementary model fusion can substantially improve predictive stability in multiclass renal CT analysis. The study further establishes a clinically responsible scope: the model classifies CT-visible urological pathology categories (cyst, stone, tumor, and normal parenchyma) and does not assess renal functional parameters such as eGFR or serum creatinine. It supports urological imaging practice through reliable automated recognition of these clinically distinct renal CT patterns.

These findings have direct implications for urological imaging research and practice. The four-class classification scheme (Cyst, Normal, Stone, Tumor) maps directly onto clinically actionable categories in urology, and a validated AI-assisted classification layer may support radiologist and urologist decision-making in high-volume CT reading environments. Physician oversight remains essential, and external multi-center validation is required before implementation in any clinical urological imaging workflow.

In summary, the proposed hybrid deep learning framework demonstrated strong internal performance across all four urologically relevant kidney CT categories. It represents a promising methodological approach for AI-assisted urological imaging and provides a basis for future prospective validation studies. External validation across diverse clinical settings, CT acquisition protocols, and patient populations is the critical next step before this framework can be considered for integration into urological imaging workflows.

Author Contributions

The percentages of the authors' contributions are presented below. All authors reviewed and approved the final version of the manuscript.

	M.T.A.	F.K.
C	50	50
D	55	45
S	40	60
DCP	65	35
DAI	60	40
L	45	55
W	50	50
CR	40	60
SR	30	70
PM	40	60

C= concept, D= design, S= supervision, DCP= data collection and/or processing, DAI= data analysis and/or interpretation, L= literature search, W= writing, CR= critical review, SR= submission and revision, PM= project management.

Conflict of Interest

The authors declare no conflict of interest.

Ethical Consideration

Ethics committee approval was not required for this study because it involved no animal or human subjects.

Acknowledgements

The dataset used in this study was the public Roboflow Universe project "Renal- Failure-Analysis" maintained by Mini Project (workspace: mini-project-6ilyj; project: renal-failure-analysis-5ymli), version 4 (Mini Project, 2023). The working export contained 9,574 images in the Cyst, Normal, Stone, and Tumor classes and was publicly listed under a Public Domain license at the time of revision. The public project page does not document the original clinical institution, DICOM-level anonymization workflow, or patient/study identifiers. The authors analyzed only the rendered image export supplied by the repository and did not have access to original DICOM headers or identifiable patient metadata. Therefore, patient-level split independence and the original anonymization process could not be independently verified.

ChatGPT was used solely for English translation, grammar correction, and linguistic editing. The artificial intelligence tool was not used to generate the research design, dataset, analyses, model outputs, tables, figures, or scientific conclusions. The authors reviewed and approved the final manuscript and take full responsibility for its content.

References

Abd El-Aziz, A. A., Mahmood, M. A., & El-Ghany, S. A. (2025). EfficientNet-B3-based automated deep learning framework for multiclass endoscopic bladder tissue classification. *Diagnostics, 15(19)*, Article 2515. https://doi.org/10.3390/diagnostics15192515

Aby, A. E., Salaji, S., Anilkumar, K. K., et al. (2025). Classification of acute myeloid leukemia by pre-trained deep neural networks: A comparison with different activation functions. *Medical Engineering & Physics, 135*, Article 104277. https://doi.org/10.1016/j.medengphy.2024.104277

Al-Adhaileh, M. H., Alsharbi, B. M., Aldhyani, T. H. H., et al. (2025). DLAAD-deep learning algorithms assisted diagnosis of chest disease using radiographic medical images. *Frontiers in Medicine, 11*, Article 1511389. https://doi.org/10.3389/fmed.2024.1511389

Al-Otaibi, S., Rehman, A., Mujahid, M., et al. (2024). Efficient-Gastro: Optimized EfficientNet model for the detection of gastrointestinal disorders using transfer learning and wireless capsule endoscopy images. *PeerJ Computer Science, 10*, e1902. https://doi.org/10.7717/peerj-cs.1902

Alzu'bi, D., Abdullah, M., Hmeidi, I., et al. (2022). Kidney tumor detection and classification based on deep learning approaches: A new dataset in CT scans. *Journal of Healthcare Engineering, 2022*, Article 3861161. https://doi.org/10.1155/2022/3861161

Aouadi, S., Torfeh, T., Bouhali, O., et al. (2024). Prediction of cervix cancer stage and grade from diffusion weighted imaging using EfficientNet. *Biomedical Physics & Engineering Express, 10(4)*, Article 045042. https://doi.org/10.1088/2057-1976/ad5207

Badr, M., Elkasaby, A., Alrahmawy, M., & El-Metwally, S. (2025). A multi-model deep learning architecture for diagnosing multi-class skin diseases. *Journal of Imaging Informatics in Medicine, 38(3)*, 1776–1795. https://doi.org/10.1007/s10278-024-01300-w

Boudi, A., He, J., & Abd El Kader, I. (2024). Enhancing Alzheimer's disease classification with transfer learning: Fine-tuning a pre-trained algorithm. *Current Medical Imaging, 20*, Article e15734056305633. https://doi.org/10.2174/0115734056305633240603061644

Chen, X., Pu, X., Chen, Z., et al. (2023). Application of EfficientNet-B0 and GRU-based deep learning on classifying the colposcopy diagnosis of precancerous cervical lesions. *Cancer Medicine, 12(7)*, 8690–8699. https://doi.org/10.1002/cam4.5581

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1800–1807. https://doi.org/10.1109/CVPR.2017.195

Esteva, A., Robicquet, A., Ramsundar, B., et al. (2019). A guide to deep learning in healthcare. *Nature Medicine, 25(1)*, 24–29. https://doi.org/10.1038/s41591-018-0316-z

Gan, K., Liu, Y., Zhang, T., et al. (2024). Deep learning model for automatic identification and classification of distal radius fracture. *Journal of Imaging Informatics in Medicine, 37(6)*, 2874–2882. https://doi.org/10.1007/s10278-024-01144-4

Han, S., Hwang, S. I., & Lee, H. J. (2019). The classification of renal cancer in 3-phase CT images using a deep learning method. *Journal of Digital Imaging, 32(4)*, 638–643. https://doi.org/10.1007/s10278-019-00230-2

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Identity mappings in deep residual networks. *Computer Vision – ECCV 2016*, 630–645. https://doi.org/10.1007/978-3-319-46493-0_38

Hossain, M. N., Bhuiyan, E., Miah, M. B. A., et al. (2025). Detection and classification of kidney disease from CT images: An automated deep learning approach. *Technologies, 13(11)*, Article 508. https://doi.org/10.3390/technologies13110508

Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4700–4708. https://doi.org/10.1109/CVPR.2017.243

Kan, H.-C., Lin, P.-H., Shao, I.-H., et al. (2025). Using deep learning to differentiate among histology renal tumor types in computed tomography scans. *BMC Medical Imaging, 25(1)*, Article 66. https://doi.org/10.1186/s12880-025-01606-3

Kreouzi, M., Theodorakis, N., Feretzakis, G., et al. (2025). Deep learning for melanoma detection: A deep learning approach to differentiating malignant melanoma from benign melanocytic nevi. *Cancers, 17(1)*, Article 28. https://doi.org/10.3390/cancers17010028

Kumar, Y., Shrivastav, S., Garg, K., et al. (2024). Automating cancer diagnosis using advanced deep learning techniques for multi-cancer image classification. *Scientific Reports, 14(1)*, Article 25006. https://doi.org/10.1038/s41598-024-75876-2

Mao, J., Du, Y., Xue, J., et al. (2024). Automated detection and classification of mandibular fractures on multislice spiral computed tomography using modified convolutional neural networks. *Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology, 138(6)*, 803–812. https://doi.org/10.1016/j.oooo.2024.07.010

Meeran, S. B. U. (2025). Walrus optimization-enhanced ResNet-50 for AI-driven renal malignancy prediction with occlusion sensitivity-based interpretation. *Asian Pacific Journal of Cancer Prevention, 26(8)*, 2995–3004. https://doi.org/10.31557/APJCP.2025.26.8.2995

Mini Project. (2023). Renal-failure-analysis dataset (Version 4) [Data set]. *Roboflow Universe*. https://universe.roboflow.com/mini-project-6ilyj/renal-failure-analysis-5ymli

Pimpalkar, A., Saini, D. K. J. B., Shelke, N., et al. (2025). Fine-tuned deep learning models for early detection and classification of kidney conditions in CT imaging. *Scientific Reports, 15(1)*, Article 10741. https://doi.org/10.1038/s41598-025-94905-2

Rajinikanth, V., Vincent, P. M. D. R., Srinivasan, K., et al. (2023). A framework to distinguish healthy/cancer renal CT images using the fused deep features. *Frontiers in Public Health, 11*, Article 1109236. https://doi.org/10.3389/fpubh.2023.1109236

Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. *Nature Medicine, 28(1)*, 31–38. https://doi.org/10.1038/s41591-021-01614-0

Sathya, R., Mahesh, T. R., Bhatia Khan, S., et al. (2024). Employing Xception convolutional neural network through high-precision MRI analysis for brain tumor diagnosis. *Frontiers in Medicine, 11*, Article 1487713. https://doi.org/10.3389/fmed.2024.1487713

Shaheed, K., Abbas, Q., Hussain, A., & Qureshi, I. (2023). Optimized Xception learning model and XGBoost classifier for detection of multiclass chest disease from X-ray images. *Diagnostics, 13(15)*, Article 2583. https://doi.org/10.3390/diagnostics13152583

Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. *Proceedings of the 36th International Conference on Machine Learning, 97*, 6105–6114. PMLR.

Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine, 25(1)*, 44–56. https://doi.org/10.1038/s41591-018-0300-7

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association, 22(158)*, 209–212. https://doi.org/10.1080/01621459.1927.10502953

Xu, P., Zhao, J., Wan, M., et al. (2024). Classification of multi-feature fusion ultrasound images of breast tumor within category 4 using convolutional neural networks. *Medical Physics, 51(6)*, 4243–4257. https://doi.org/10.1002/mp.16946

Yao, H., Tian, L., Liu, X., et al. (2023). Development and external validation of the multichannel deep learning model based on unenhanced CT for differentiating fat-poor angiomyolipoma from renal cell carcinoma: A two-center retrospective study. *Journal of Cancer Research and Clinical Oncology, 149(17)*, 15827–15838. https://doi.org/10.1007/s00432-023-05339-0

Ye, D., Ou, Z., Ye, F., et al. (2025). Ensemble deep learning model for accurate assessment of renal fibrosis in chronic kidney disease using two-dimensional shear wave elastography images. *BMC Medical Imaging, 25(1)*, Article 417. https://doi.org/10.1186/s12880-025-01964-y

Zhang, X., Lai, F., Chen, W., & Yu, C. (2024). An automatic glaucoma grading method based on attention mechanism and EfficientNetB3 network. *PLOS ONE, 19(8)*, e0296229. https://doi.org/10.1371/journal.pone.0296229

Zhou, T., Ye, X., Lu, H., et al. (2022). Dense convolutional network and its application in medical image analysis. *BioMed Research International, 2022*, Article 2384830. https://doi.org/10.1155/2022/2384830