

Original Article

A LIGHTWEIGHT CUSTOM CNN FOR ROBUST AND CLINICALLY SAFE CT KIDNEY DISEASE CLASSIFICATION

Kazheen Sileman Muhammad^{1,2,*} , and Araz Rajab Abraham³ 

¹Duhok Polytechnic University, Duhok, Kurdistan Region, Iraq.

²Department of Information Technology, Technical College of Informatics, Akre University for Applied Sciences, Akre, Kurdistan Region, Iraq.

³Duhok Technical Institute, Duhok Polytechnic University, Duhok, Kurdistan Region, Iraq.

*Corresponding author; E-mail: kazheen.sileman@auas.edu.krd

ABSTRACT

Received
June 26, 2026

Accepted
September 13, 2026

Published
October 20, 2026

Kidney diseases such as renal cysts, stone, and malignant tumors represent a significant and increasing health burden worldwide. Manual CT diagnosis is labour-intensive and has high inter-observer variability, which requires the support of automatic diagnosis. In this paper, a lightweighted custom Convolutional Neural Network (CNN) is proposed, which is used to automatically classify kidney computed tomography (CT) images into four classes: Cyst, Normal, Stone, and Tumor. It includes residual skip connections, a dual-pooling head (Global Average and Global Max Pooling), Spatial Dropout2D regularization, and a composite training objective that explicitly favours Tumor Recall. The CT-KIDNEY-DATASET was made leakage-safe through MD5 exact deduplication and perceptual-hash grouping, so that all reported evaluations are free of data leakage. Across the five validation folds, KidneyCNN achieved a mean accuracy of $99.82 \pm 0.06\%$ and a mean Tumor Recall of $99.31 \pm 0.30\%$. Separately, ensemble inference on the independent held-out test set of 1,904 images achieved 99.84% accuracy, with AUC = 1.000 for all classes and only three misclassifications. The Grad-CAM++ results show that the model attention as anatomically relevant, which further indicates that KidneyCNN is a clinically safe and deployable kidney disease screening tool.

Keywords: CT Kidney Classification, Deep Learning, CNN, Tumor Detection.

1. INTRODUCTION

Ensuring health security continues to be a key global priority, especially as kidney diseases are causing an increasing burden on health care systems around the world. Renal disorders encompass a broad spectrum of conditions, including simple and complex cysts, nephrolithiasis, chronic kidney disease, and renal cell carcinoma. Collectively, these conditions affect approximately 850 million people worldwide and are a major cause of morbidity and premature mortality (Sistaninejad *et al.*, 2023). Smart healthcare powered by advanced artificial intelligence (AI), machine learning (ML), and deep learning (DL) techniques offers a promising solution for early detection through automated systems and timely clinical intervention, enabling more efficient use of diagnostic resources and improved patient outcomes (Aljohani, 2025).

In clinical practice, the kidney is one of the most severely affected organs, with kidney stones, kidney disease, and other conditions such as renal cysts being among the most common findings. Both benign and malignant tumors have different prognostic meaning. Renal tumours are particularly worrisome if found at Stage I or II. The chances of making it five years after diagnosis increases to 90–95%, but if the diagnosis is delayed until Stage IV, the survival rate drops to 10–15% (Rao *et al.*, 2025). Computed tomography (CT) imaging is the preferred modality to evaluate the kidneys, and it can give high-resolution cross-sectional images, which show density patterns specific to disease (Islam *et al.*, 2022). Manual radiological review remains the standard method for interpreting kidney CT scans, but it is time consuming, subjective, and subject to considerable inter-observer variability which can delay diagnosis beyond the window for timely and

effective intervention (M. N. Hossain *et al.*, 2025; Tekin *et al.*, 2025). The consequences of delayed or missed detection are serious: an estimated 3–5% of renal tumors are overlooked during routine CT review.

The advent of DL and medical image analysis has completely transformed the field of clinical diagnostics, providing unprecedented possibilities for automatic, fast and precise diagnosis of diseases (Maqsood *et al.*, 2024). Convolutional neural networks (CNNs) have been found to be highly effective for image classification, because they can automatically extract hierarchical features directly from raw pixel data (He *et al.*, 2015). This makes CNNs particularly suited to identifying the fine-grained visual patterns of renal diseases in a CT scan, even at the very early and most difficult-to-diagnose stages (Alzu'Bi *et al.*, 2022).

In the literature, the reported accuracy numbers are impressive, but a systematic study shows that there are some common methodological shortcomings and published results that cannot be relied upon. Almost all studies evaluate their models on the original CT-KIDNEY-DATASET, which contains multiple byte-identical copies of the same image. A random train-test split therefore allows duplicates to cross the partition boundary and overestimate performance. Almost all published studies use only one single train-test split and never cross-validate the model to make statistically significant conclusions. Moreover, most methods rely on ImageNet pre-trained backbones for CT classification, where the data source is an RGB image with a domain mismatch between natural images and images obtained from a CT scanner, which may restrict the learning of pathology-specific features. Importantly, no published study reports both Tumor

Access this article online



DOI: <https://doi.org/10.25271/sjuoz.2026.14.4.1969>

Printed ISSN 2410-7549;
Electronic ISSN 2414-6943

Science Journal of University of Zakho
Vol. 14, No. 04, pp. 666-678, October -2026

This is an open access article under a CC BY-NC-SA 4.0 license
(<https://creativecommons.org/licenses/by-nc-sa/4.0/>)

Recall as an explicit clinical safety metric and Expected Calibration Error as a measure of confidence reliability, both essential properties for deployment in actual clinical decision-support systems. All these factors prompted the present study to provide a purpose-built, clinically oriented, and leakage-safe classification system that explicitly tackles all the gaps.

In this study, the authors propose KidneyCNN, a lightweight custom CNN for the accurate and efficient automated classification of kidney computed tomography (CT) images into four clinically different categories: Cyst, Normal, Stone, and Tumor. The proposed model fuses a new dual pooling head (fusing global average and max pooling) and uses a Tumor-Focused Focal Loss, which reflects clinical priority directly into the training objective. The vanishing gradient issue is overcome with residual connections which enable stable training of deeper feature hierarchies (Abid *et al.*, 2023). A comprehensive data augmentation strategy improves the generalization of the model across different scanner settings and patient presentations, while minimizing overfitting and retaining pathology-discriminating features (T. Hossain *et al.*, 2024). The pipeline combines five-fold cross-validation with ensemble inference and Grad-CAM++ explainability to ensure both performance and clinical interpretability.

The outline of the main contributions of this study is presented in a methodical way:

- **The new Lightweight Architecture with Dual Pooling:** A custom CNN with 2.83 million parameters, nine times fewer than ResNet-50, is trained from scratch. It replaces the single pooling head of a standard CNN with parallel global average pooling and global max pooling branches, whose outputs are concatenated into a 512-dimensional feature vector.
- **Clinical Priority Loss Function:** Directly encoding clinical risk in training penalizes missed tumor misclassifications by a factor of 2, while stones are given a weight of 1.3. Cyst and Normal remain unweighted ($1.0\times$). The loss uses a focal-style formulation with $\gamma = 2$, applied without a separate alpha term; class imbalance is instead handled upstream via a weighted random sampler at the data-loading stage, avoiding the double-correction that would result from combining both mechanisms. An early stopping condition is applied, where the tumor recall rate must be $\geq 92\%$ at all checkpoints, which ensures that no checkpoint with inadequate tumor sensitivity is retained. The complete per-sample formula is $L_i = w_{y_i} \cdot (1 - e^{-CE_i})^\gamma \cdot CE_i$, where CE_i is the label-smoothed cross-entropy ($\epsilon = 0.05$), $\gamma = 2$, and w_{y_i} equals 2.0 for Tumor, 1.3 for Stone, and 1.0 for Cyst/Normal; the batch loss is the mean of L_i over all N samples.
- **Leakage-Safe Cross-Validation Pipeline:** A fully leakage-safe 5-stage pre-processing pipeline, including MD5 exact deduplication, dHash perceptual grouping, group-level stratified splitting, fold-level normalization during training, and MD5 cross-partition verification ensures zero data leakage and reproducible results.
- **Rigorous Evaluation with Explainability:** Seven complementary components are used to evaluate the model, namely five performance metrics (Accuracy, Macro-F1, AUC-ROC, Tumor Recall, and Expected Calibration Error), McNemar's statistical significance testing (the most rigorous statistical evaluation published on this dataset), and Grad-CAM++ visualizations to confirm clinically correct anatomical attention patterns across all four disease classes.

The remainder of this work is organized as follows. Section 2 reviews related work. Section 3 describes the dataset, the pre-processing pipeline, and the proposed KidneyCNN architecture. Section 4 presents the experimental setup, the evaluation metrics, and the experimental results. Section 5 discusses the ablation study and compares the proposed model with the best available methods. Finally, Section 6 summarizes the paper and outlines directions for future research.

2. RELATED WORK

In medical image analysis, DL techniques have proven to be very useful for the automatic diagnosis of kidney disease, and these tech-

niques extend to other organs as well. To set a baseline for CT-based kidney disease classification, Islam *et al.* (2022) generated a publicly available dataset of 12,446 CT scans across four classes (Cyst, Normal, Stone, and Tumor) and trained a number of Vision Transformer variants along with VGG16, ResNet50, and InceptionV3 with an accuracy of 99.30% for Swin Transformer. Alzu'bi *et al.* (2022) extended this baseline by using ResNet50 and VGG16 on an independent set of images acquired at hospitals in Jordan, yielding a 97% success rate, showing that deep feature extraction could be generalized across a variety of CT acquisition protocols.

Later research focused on lightweight architectures and model interpretability. Bhandari *et al.* (2023) showed that a compact custom CNN could reach competitive accuracy with only a small trade-off in per-prediction interpretability, using LIME and SHAP post-hoc explainability methods and saliency maps for clinical review. Badawy *et al.* (2023) proposed a two-step pipeline in which the performance of eight pretrained backbones (i.e., VGG16, VGG19, Xception, DenseNet201, MobileNet, MobileNetV2, MobileNetV3Large, and NASNetMobile) was assessed alongside the Sparrow Search Algorithm for automatic segmentation. The best four-class performance was achieved through hyperparameter optimization, with DenseNet201 performing best. These studies collectively established that architecture selection and training configuration are at least as impactful as model scale.

Ensemble and hybrid strategies have further improved reported classification accuracy on the CT-KIDNEY-DATASET benchmark. Saber *et al.* (2025) combined feature extractors from MobileNetV2, ResNet50, and EfficientNet-B0 with metaheuristic optimization and Bi-LSTM sequence modeling, demonstrating that ensemble combination consistently surpasses individual models for kidney CT screening. Sharma *et al.* (2025) fused a pretrained ResNet101 with a custom CNN through feature-level concatenation and ensemble classification, reporting 99.73% training accuracy and 100% test accuracy in a preprint. The absence of explicit deduplication and group-level data splitting makes the latter figure difficult to verify independently.

Later research focused on efficiency, segmentation-based classification, and more stringent evaluation. Pimpalkar *et al.* (2025) used VGG16, ResNet50, and EfficientNet backbones together with watershed segmentation, Otsu binarization, and Relief-based feature selection, reporting an accuracy of 99.96%. However, the reported performance remained sensitive to the preprocessing stage. Jahan *et al.* (2026) combined U-Net segmentation, achieving a Dice score of 98%, with Kid-Net, an EfficientNetB3-based classifier that achieved 99.30% accuracy under cross-validation, while incorporating LIME and Grad-CAM for explainability. Khan *et al.* (2025) fine-tuned EfficientNetV2B0 for edge-device deployment and reported 99.73% accuracy, with particular emphasis on inference efficiency. Maçin *et al.* (2025) proposed a lightweight custom CNN evaluated across three independent CT datasets, reporting 99.96% accuracy and 99.94% macro-F1.

Although significant progress has been made, some methodological drawbacks are found throughout the literature. The CT-KIDNEY-DATASET contains 517 exact duplicate images as well as many near-duplicate frames from the same CT study, so splitting without deduplication readily places near-duplicate images on both sides of the train-test boundary and overestimates accuracy. Most studies report results for a single train-test split, but not cross-validated results, which makes stable statistical inference impossible. In addition, normalization statistics are typically not computed per fold, which causes distributional leakage, and clinically relevant calibration measures such as the Expected Calibration Error are almost universally absent. This study aims to overcome these shortcomings by proposing a lightweight custom CNN trained with a leakage-safe 5-fold cross validation pipeline with a hard-stop MD5 overlap verification, MD5 exact deduplication, and per-fold normalisation. Evaluation uses five metrics (accuracy, macro-F1, AUC-ROC, tumor recall, and ECE) supported by McNemar's statistical significance test and by Grad-CAM explainability across all four disease classes.

3. MATERIALS AND METHODS

This section introduces complete methods for the classification of CT kidney images into four disease classes: Cyst, Normal, Stone and Tumor. The methodology comprises characterizing the dataset,

meticulously preparing and augmenting the dataset, designing the proposed deep neural network architecture, and defining the performance assessment metrics used to assess the efficacy of the models.

The full methodology pipeline is presented in Figure 1. It shows dataset preprocessing that includes MD5 based deduplication and splitting at the group level whilst preserving leakage safety. This is followed by five-fold cross validation training normalized per fold. Next, the trained fold models are ensembled and tested on the strictly held out test set. Finally, performance is assessed using five clinical performance metrics together with McNemar's significance test and Grad-CAM++ explainability, providing a complete and rigorous assessment.

Dataset Description

In this study, the experiments were carried out with the publicly available dataset from the Kaggle platform (Islam *et al.*, 2022), named as CT-KIDNEY-DATASET. The collection was assembled by researchers who specialize in medical imaging and is intended to serve as a representative and high-quality dataset of CT kidney images for development and testing of DL systems for automated renal disease detection. This dataset includes 12,446 CT images from four classes of Cyst, Normal, Stone, and Tumor, as illustrated in Table 1 and Figure 2.

All images were obtained with different scanner parameters, contrasts and severity of the disease. According to the expert radiological evaluation, the disease classification is given which ensures diagnostic accuracy. MD5 cryptographic hashing was used to identify and eliminate duplicate images which are identical in bytes before any experimental procedure. This process resulted in 517 exact duplicates, all having no label conflicts, bringing the count of unique images to 11,929 as shown in Table 1.

Data Preparation

To ensure that all the data had the same spatial dimensions, the images were resized to 224×224 pixels. Subsequently, pixel values spanning 0 to 255 were scaled to the interval $[0, 1]$ by dividing by 255.0, improving numerical stability and accelerating training convergence. A per-fold channel-wise mean and standard deviation were then computed exclusively from each fold's training subset and applied to normalize all images to zero mean and unit variance. Computing these statistics from training data only prevents distribution leakage, a subtle but real source of inflated accuracy that occurs when validation or test pixel values contribute to the normalization parameters.

To further prevent data contamination, a five-stage leakage-safe pipeline was implemented prior to any splitting. MD5 cryptographic hashing eliminated 517 byte-identical duplicate images, yielding 11,929 unique images with zero label conflicts. dHash perceptual grouping subsequently identified near-duplicate scans, such as the same CT scan saved at slightly different resolutions, ensuring that visually similar images always remained in the same data partition. It is

important to note that the CT-KIDNEY-DATASET does not provide patient identifiers, precluding strict patient-level partitioning. The dHash perceptual grouping strategy was therefore employed as a principled proxy, automatically clustering near-identical frames that originate from the same CT acquisition sequence and ensuring all images within each group were assigned exclusively to a single partition, achieving the same leakage-prevention objective as patient-level splitting without requiring patient metadata. Group-level stratified splitting then atomically assigned entire perceptual groups to either the cross-validation pool or the held-out test set, preventing any near-duplicate from crossing the train-test boundary. A final MD5 cross-partition verification step confirmed zero hash overlap between all partitions, providing a mathematical guarantee of zero leakage.

The unique images were divided into three separate sets of clean images, as shown in Table 2. The dataset was split into an 85% cross-validation pool of 10,025 images and a 15% held-out test set of exactly 1,904 images, which was locked right after splitting and used only for final evaluation. Five-fold stratified splitting resulted in approximately 8,024 training and 2,005 validation images per fold in the cross-validation pool while maintaining the same class distribution that existed in the original pool of images. Table 2 reports the exact per-class counts for Fold 5, the representative fold used throughout this study, including the ablation experiments in Section 5.1. It contains 7,981 training images and 2,044 validation images, which are close to, but not identical to, the average counts across all five folds reported above.

Table 1: After MD5 exact deduplication, distribution of CT kidney images in the four classes.

Class	Raw Images	After Dedup	% of Dataset
Cyst	3,709	3,284	27.53%
Normal	5,077	5,002	41.93%
Stone	1,377	1,360	11.40%
Tumor	2,283	2,283	19.14%
Duplicates removed	517	—	—
Total	12,446	11,929	100%

Data Augmentation

To enhance the robustness and generalization of the model to real-world variations in CT scans and to handle potential class imbalance, a comprehensive data augmentation strategy was designed to improve the model's performance. It was applied only during the training phase of each cross-validation fold. The augmentation pipeline was designed to simulate diverse real-world CT imaging conditions while preserving the pathological features critical to each class of kidney disease. These transformations are described in Table 3 and involve both geometric and photometric changes: geometric transformations increase the positional and orientational variability of the training data and thereby reduce overfitting, while photometric transformations simulate variation in scanner brightness, contrast, and sensor noise.

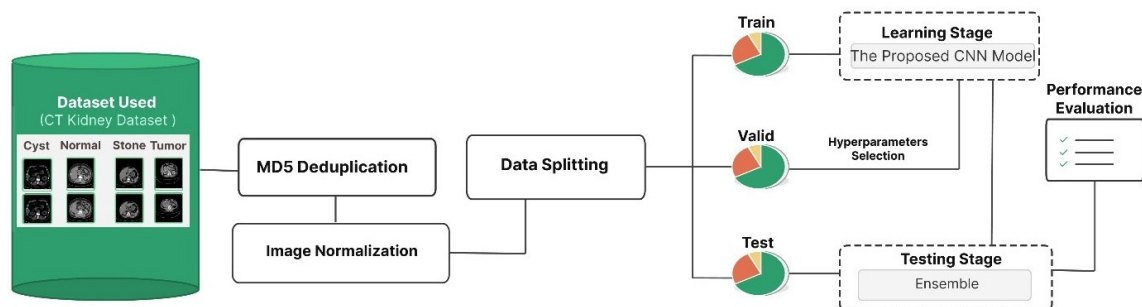


Figure 1: Proposed KidneyCNN Methodology Pipeline.

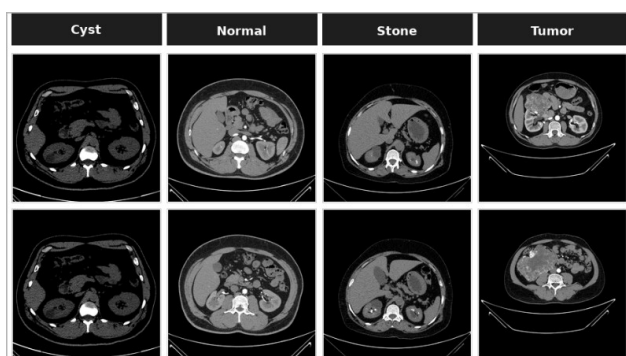


Figure 2: Some samples of the original dataset of CT-KIDNEY-DATASET.

Table 2: Dataset Split Distribution by Class.

Split	Cyst	Normal	Stone	Tumor
Train	2,200	3,300	944	1,537
Validation	549	913	217	365
Held-out Test	535	789	199	381

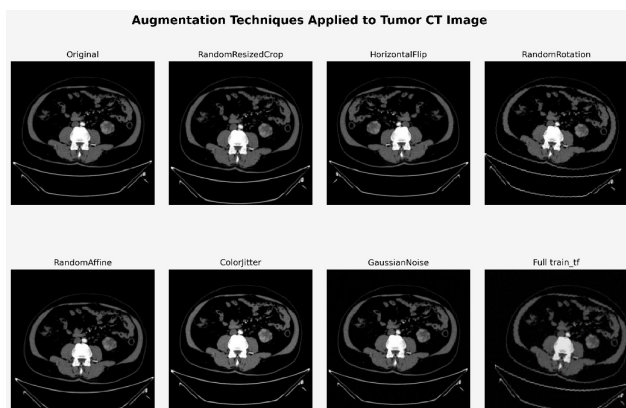


Figure 3: Data Augmentation Techniques Applied to CT-KIDNEY-DATASET.

All augmentation parameters were validated to ensure that the resulting CT images remained clinically realistic and were neither anatomically implausible nor diagnostically misleading, while still maximizing training-data diversity and model robustness. It is important to stress that augmentation was exclusively used during training and not during validation or test, ensuring that all performance results reported are true generalization to real world, unaltered CT imaging conditions. The data augmentation methods used during training of the CT-KIDNEY-DATASET are shown in Figure 3, where random horizontal flipping, rotation, affine translation, color jitter, random resized cropping, and the injection of Gaussian noise are only applied

to the training partition to represent CT variations observed in the real world while avoiding distributional leakage into the validation and test partitions.

Proposed Method

The foundation of this research is the development of a new lightweight CNN architecture called KidneyCNN, which is specially designed for the purpose of robust and highly accurate classification of CT kidney images to four categories of disease: Cyst, Normal, Stone, and Tumor. The proposed architecture, precisely detailed in Figure 4, combines the basic convolutional layers, superior residual blocks and a new dual pooling head together effectively. It effectively preserves the hierarchical information within CT images, overcomes the common problem of the vanishing gradient in training deep neural networks and has a small parameter count of 2,828,196, nine times fewer than ResNet-50.

The network is designed to take standardized sized images of CT data with a dimension of $3 \times 224 \times 224$. Instead of the Conv2D layer with a small 3×3 kernel, KidneyCNN uses a stem block consisting of a Conv2D layer with 32 filters (7×7 , stride 2), followed by a Batch Normalization layer, a ReLU activation, and a MaxPool layer (3×3 , stride 2). The larger kernel (7×7) also delivers a wider receptive field to first capture the overall structural context of the kidney before detailed feature extraction is performed, reducing the spatial resolution from 224×224 to 56×56 .

The backbone of the network comprises three stages of two residual blocks each, in which the channel depth increases from 32 to 64, 128 and finally 256, while the spatial resolution decreases from 56×56 to 28×28 to 14×14 . Downsampling is performed by stride convolutions (stride=2) in the first residual block of the second and third stages; the first stage uses stride=1 and therefore preserves the 56×56 resolution while increasing the channel depth from 32 to 64. The only MaxPool layer in the network appears once, in the stem, prior to these three stages. Every residual block performs the expression $\text{Output} = \text{ReLU}(F(x) + x)$ where $F(x)$ is the two sequential $\text{Conv}3 \times 3 \rightarrow \text{BatchNorm} \rightarrow \text{ReLU}$ operations, ensuring the gradient flow is stable during the training process. The first stage captures low-level structural elements such as kidney borders and coarse density changes. The second stage, followed by a Spatial Dropout2D layer with dropout rate of 0.2, learns intermediate disease-level patterns such as cyst signatures, stone calcifications, and tumor heterogeneity. The third and deepest stage learns the richest and most semantically abstract disease representations, which are interpreted by Grad-CAM++ at 256-channel resolution to confirm that the model's attention is anatomically correct.

Upon completion of the third residual stage, KidneyCNN adds a new dual pooling head, which is a major architectural contribution in this work. The outputs of the $14 \times 14 \times 256$ feature volume are applied to both Global Average Pooling (GAP) and Global Max Pooling (GMP) in parallel and then both concatenated into a 512-dimensional feature vector. Average pooling is well suited to diffuse textural patterns, such as those of cysts or normal kidney, whereas max pooling responds to focal peak activations, such as those of stones or tumors. The two branches therefore provide complementary disease evidence that neither pooling strategy could supply alone.

Table 3: Data Augmentation Techniques Applied to CT Kidney Images.

Category	Transformation	Description
Geometric	Random crop	Scale 90–100% varied zoom and positioning
	Horizontal flip	50% probability of bilateral kidney appearance
	Random rotation	$\pm 10^\circ$ minor patient positioning variation
	Affine translation	Shift $\pm 8\%$ minor patient positioning variation
Photometric	Color jitter	Brightness/contrast $\pm 15\%$ scanner variability
	Gaussian noise	$\sigma = 0.008$ sensor/scanner noise simulation
Regularization	Label smoothing	$\varepsilon = 0.05$ prevents overconfidence
Sampling	Weighted sampler	PyTorch WeightedRandomSampler; each training sample is assigned weight = $1/(\text{class count})$, oversampling minority classes (e.g. Stone) per batch. Applied only during data loading (training split only, not validation/test) not incorporated into the loss function, to avoid double-correcting for imbalance alongside the Tumor-Focused Focal Loss's per-class multipliers.

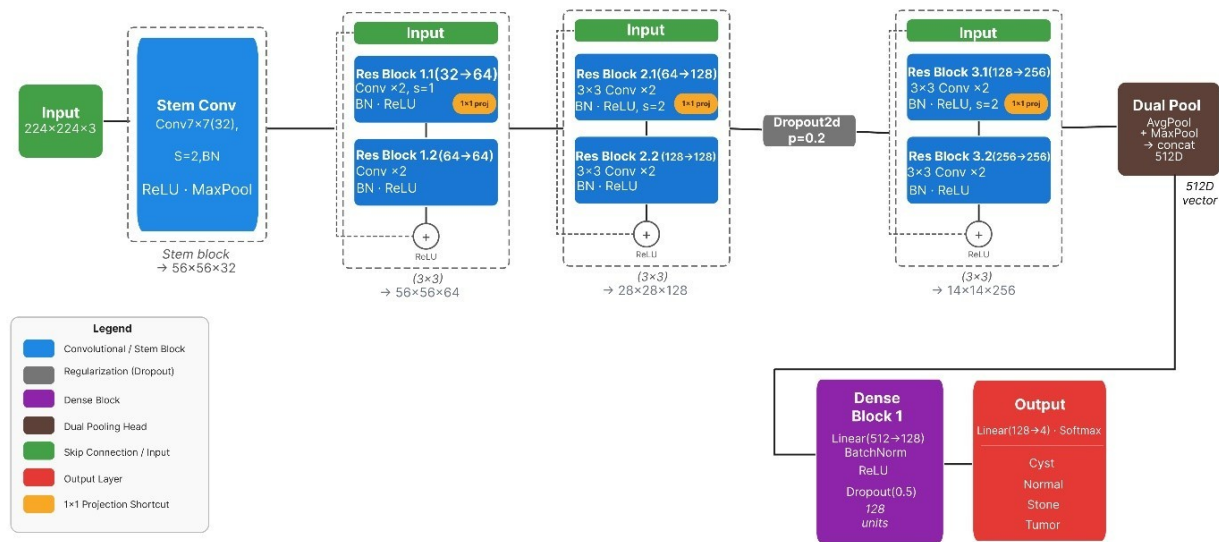


Figure 4: Architecture of the Proposed KidneyCNN Model for CT Kidney Disease Classification.

The 512-dimensional feature vector is fed into fully connected layers: Linear (512 $\rightarrow$ 128), followed by Batch Normalization, ReLU activation and Dropout (0.5) for strong regularization. The final layer is Linear (128 $\rightarrow$ 4), SoftMax, giving a probability distribution of the four kidney disease categories, where the top of the distribution is taken as the diagnostic prediction. This carefully designed architecture has only 2.828 million parameters, which is extremely lightweight and clinically deployable.

4. RESULTS

This section presents a complete examination of all experiments we managed to carry out while training and testing the proposed Kidney-CNN model for the classification of kidney CT scans into 4 classes: Cyst, Normal, Stone and Tumor.

Experimental Setup

A rigorous and reproducible experimental protocol was carefully designed to assess the proposed KidneyCNN DL model. This included the choice of key hyperparameters, optimisation techniques, and data splitting approaches to ensure successful training and evaluation of the model. These hyperparameters were selected according to proven DL practices and optimized based on empirical testing and clinical needs of kidney disease classification, as summarized in Table 4.

Table 4: Hyperparameters used during training.

Hyperparameter	Value
Input Image Size	224 $\times$ 224 $\times$ 3
Optimizer	AdamW
Learning Rate (Initial)	3×10^{-4}
Minimum Learning Rate	1×10^{-6}
Weight Decay	5×10^{-4}
LR Scheduler	CosineAnnealingWarmRestarts ($T_0=20$, $T_{mult}=2$)
Batch Size	24
Loss Function	Tumor-Focused Focal Loss ($\gamma=2$)
Label Smoothing	$\epsilon = 0.05$
Early Stopping Patience	10 epochs
Early Stopping Criterion	Val Macro-F1 $\uparrow$ AND Tumor Recall $\geq 92\%$
Maximum Epochs	100 per fold
Cross-Validation	5-fold stratified
Data Split	85% CV pool / 15% held-out test
Random Seed	42

Performance Evaluation Metrics

A wide range of standard metrics suitable for multi-class classification was used to thoroughly evaluate the performance and reliability of the proposed KidneyCNN model for kidney disease classification in four classes: Cyst, Normal, Stone, and Tumor. The evaluation was carried out by calculating the overall accuracy, class-specific precision, recall (sensitivity), expected calibration error (ECE) and F1-score. These metrics are defined by the equations given in Eqs. (1) through (5), respectively.

In addition to these five metrics, McNemar's test is used to statistically compare paired classification outcomes between two models evaluated on the same held-out test set (for example, the full KidneyCNN versus an ablated variant in Table 8). The test examines the 2×2 contingency table of discordant predictions (i.e., test cases on which the two models disagree) and computes the chi-squared statistic $\chi^2 = (b - c)^2 / (b + c)$, where b and c denote the number of instances correctly classified by one model but not the other. A resulting p-value below 0.05 indicates that the difference in error patterns between the two models is statistically significant rather than attributable to chance, providing a rigorous basis for concluding that a given architectural or training component has a genuine effect on performance.

$$Accuracy = \frac{\sum_{i=1}^n (TP_i + TN_i)}{\sum_{i=1}^n (TP_i + FP_i + TN_i + FN_i)} = \frac{\text{Total Correct Predictions}}{\text{Total Predictions}} \quad (1)$$

$$Precision_i = \frac{TP_i}{TP_i + FP_i} \quad (2)$$

$$Recall_i = Sensitivity_i = \frac{TP_i}{TP_i + FN_i} \quad (3)$$

$$ECE = \sum_{b=1}^B \frac{|B_b|}{n} \times |acc(B_b) - conf(B_b)| \quad (4)$$

$$F1-Score_i = \frac{2 \times Precision_i \times Recall_i}{Precision_i + Recall_i} \quad (5)$$

Training and Validation Performance

The KidneyCNN model was trained through a leakage-safe five-fold cross-validation strategy with five independent fold checkpoints which were determined by both the improvement in validation Macro-

F1 and Tumor Recall of at least 92%. The training and validation loss and accuracy curves for each fold are shown in Figures 5 to 9, allowing the convergence of the model and its stability during training to be observed.

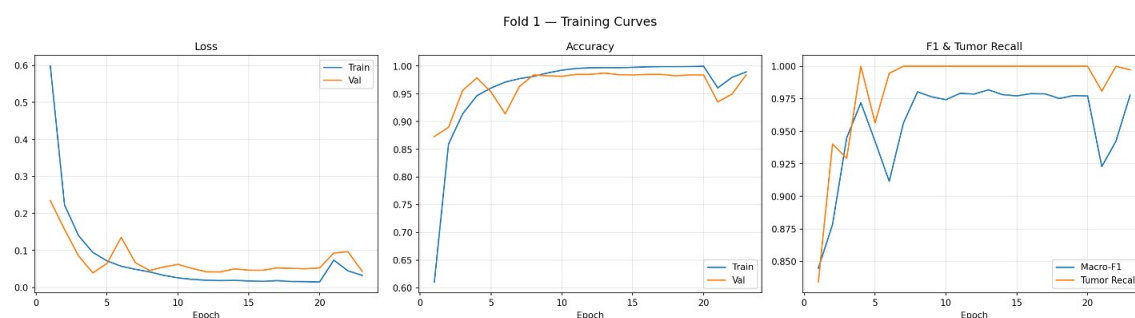


Figure 5: Training Curves — Fold 1.

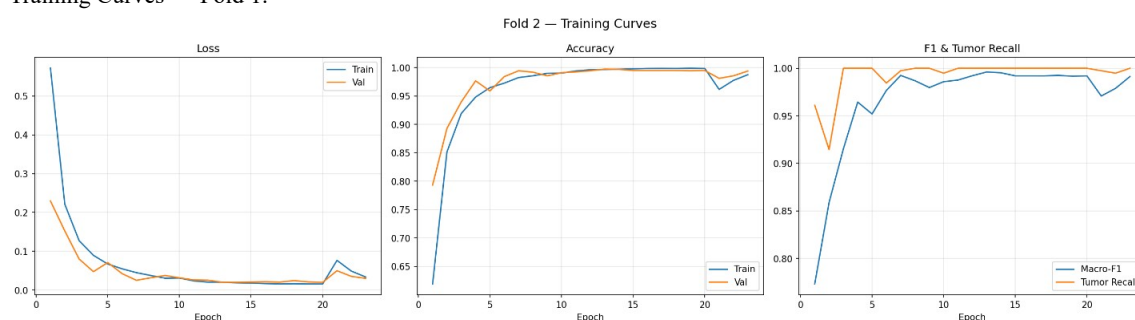


Figure 6: Training Curves — Fold 2.

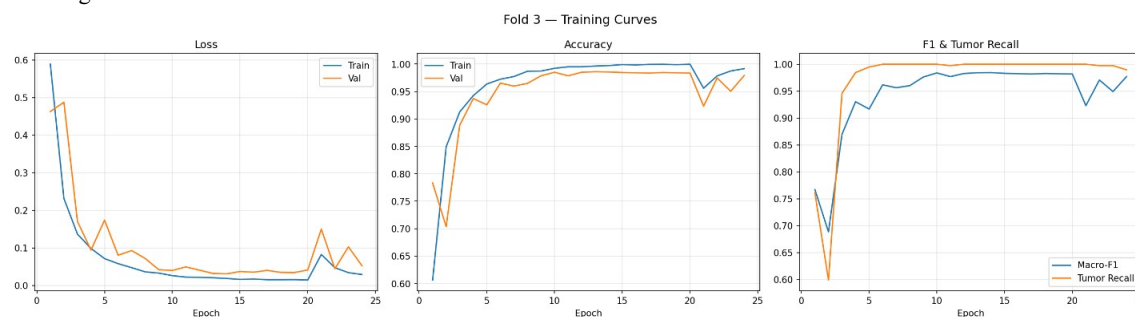


Figure 7: Training Curves — Fold 3.

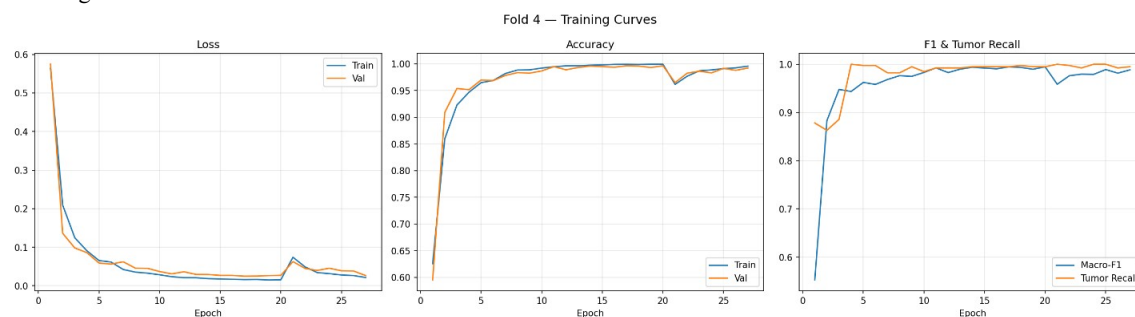


Figure 8: Training Curves — Fold 4.

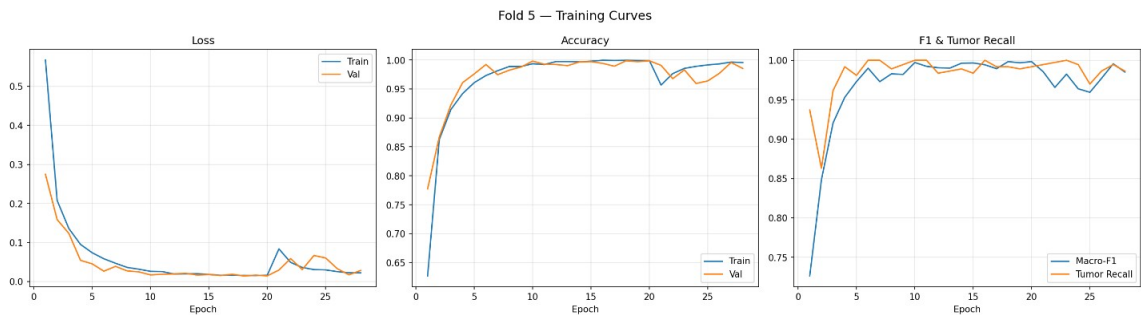


Figure 9: Training Curves — Fold 5.

The training and validation accuracy and losses for all 5 folds of the cross validation are shown to be consistently smooth and stable, as shown in Figures 5–9, indicating that KidneyCNN learned discriminative features and that the learned features were also generalized well to unseen CT kidney images throughout the entirety of training. The training accuracy starts high and quickly approaches the near perfect accuracy in all five folds, while the validation accuracy slowly increases and is consistently high with only slight deviation from training accuracy. This tight and stable gap between train and validation accuracy for all folds is a direct reflection that the model generalized well and did not show signs of overfitting, as a result of the regularization scheme used, which consisted of Batch Normalization, Spatial Dropout2D and Dense Dropout layers. The loss curves start relatively high, as is common in the early phases of training, then fall rapidly and converge closely for all folds, indicating that the model makes virtually no errors on images it has seen and only a small number on previously unseen images. Notably, this convergence pattern is consistent across all five folds, suggesting that the reported performance does not depend on a particularly favorable data partition, and that the model is robust and reproducible. The training and validation curves for the five folds have

steady and consistent growth, which shows that the hyperparameter values, the Tumor-Focused Focal Loss function, and the CosineAnnealingWarmRestarts learning rate scheduler all work well for the model, which is able to generalize training to new and unseen CT kidney images across different data splits.

Cross-Validation

The results reported in this section are from the five-fold cross-validation splits (i.e., the validation partitions), not from the held-out test set, which is evaluated separately from Section 4.5 onward. The cross-validation results demonstrate the consistency and stability of KidneyCNN across all five folds (Figure 10). As shown in the left chart, there are no significant variations in any of the 3 metrics shown, Accuracy, Macro-F1, and Tumor Recall (AUC-ROC is not plotted as it was a constant 1.0000 across all folds, see Table 5), for each fold, demonstrating the robustness of the results presented here, and that they are not reliant on any given set of data. This stability is further illustrated in the right-hand plot, where the error bars are small and comparable across metrics, indicating low variance and highly reproducible model behavior during cross-validation; the per-fold results are detailed in Table 5.

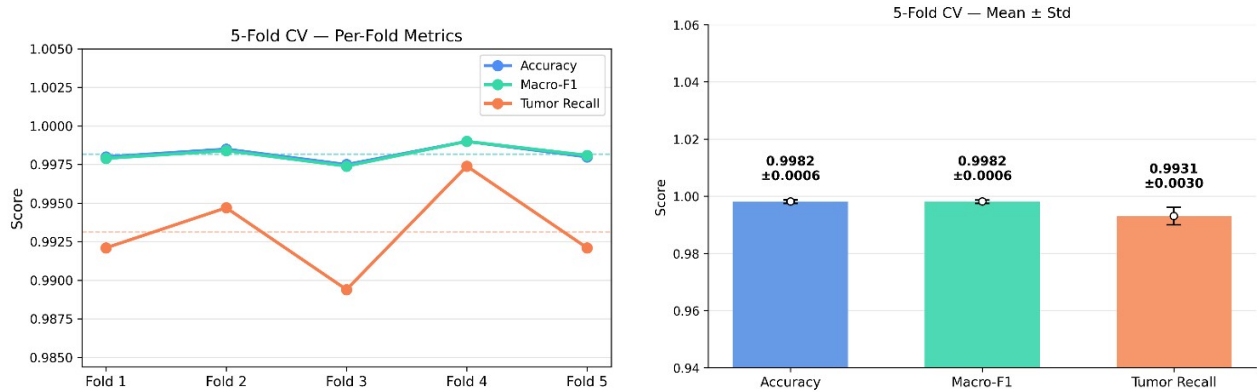


Figure 10: Cross-Validation Performance Summary.

Table 5: Validation accuracy of KidneyCNN in all folds of the cross-validation.

Fold	Accuracy	Macro-F1	AUC-ROC	Tumor Recall	Epochs
Fold 1	99.80%	99.79%	1.0000	99.21%	24
Fold 2	99.85%	99.84%	1.0000	99.47%	27
Fold 3	99.75%	99.74%	1.0000	98.94%	23
Fold 4	99.90%	99.90%	1.0000	99.74%	28
Fold 5	99.80%	99.81%	1.0000	99.21%	25
Mean ± Std	99.82 ± 0.06%	99.82 ± 0.06%	1.0000 ± 0.0000	99.31 ± 0.30%	25.4 ± 2.1

The cross-validation results showed very good consistency across all five folds. An AUC-ROC of 1.0000 in every fold indicates perfect class separability irrespective of the chosen decision threshold. In

four out of five folds, Tumor Recall was above 99% and never below 98.94%, which indicates that both early stopping criteria (including the minimal Tumor Recall of 92%) were effectively enforced during training, thus ensuring clinical safety constraints. The small standard

deviation over all metrics indicates that the results are consistent and robust and not due to an idiosyncratic data split.

Performance Analysis

Figure 11 graphically illustrates the per-class Precision, Recall, and F1-Score obtained by the KidneyCNN ensemble on held-out test set of 1,904 images, giving a comprehensive visual breakdown of classification performance for all four kidney disease categories: Cyst, Normal, Stone, and Tumor.

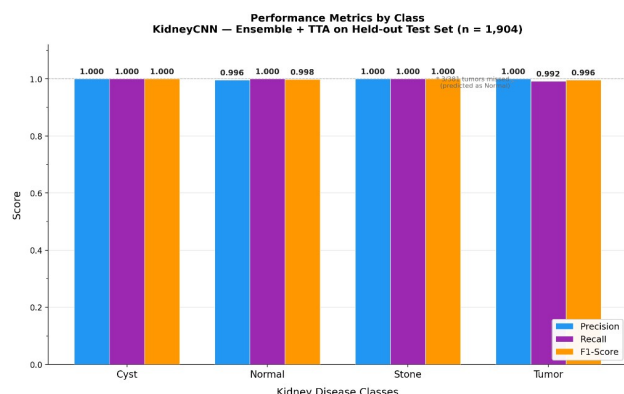


Figure 11: Per-Class Precision, Recall, and F1-Score on the Held-Out Test Set.

As shown in Figure 11, the KidneyCNN ensemble attains scores at or close to 1.000 for all four kidney disease classes. For the Cyst and Stone classes, the model achieves perfect scores of 1.000 for Precision, Recall and F1-Score, indicating that the model correctly classifies these two classes in all the test set. The Normal class has a Precision of 0.996, Recall of 1.000 and F1-Score of 0.998, of which the slight reduction in precision is caused by a small amount of non-Normal images being misclassified as Normal. The Tumor class attains a Precision of 1.000 and an F1-Score of 0.996, while its Recall of 0.992 reflects the only three errors in the entire test set: three Tumor images were misclassified as Normal, most probably early-stage atypical tumors with density characteristics that overlap with healthy parenchyma. Most importantly, there was no Tumor image that was incorrectly classified as benign (e.g. Cyst or Stone), indicating that clinically appropriate discrimination bounds were correctly enforced by the Tumor-Focused Focal Loss during training.

Confusion Matrix Analysis

The KidneyCNN ensemble achieved an accuracy of 99.84% on the held-out test set. The confusion matrix provides a detailed breakdown of its classification results across the four kidney disease classes: all Cyst, Normal, and Stone test images were classified correctly, and only three Tumor images were misclassified as Normal.

The confusion matrix as displayed in Figure 12 shows that all the images in the Cyst, Normal and Stone classes were recalled correctly, that is, with 100% accuracy, whereas 99.21% of images in the class Tumor (378 of 381, Tumor images correctly recalled, with 3 misclassified as Normal) were recalled correctly. The Stone class result is especially significant as the smallest class in the dataset, which indicates that the weighted sampling and the Tumor-Focused Focal Loss successfully tackled the class imbalance problem.

This is clinically interpretable: small tumors that show little contrast enhancement can exhibit CT density characteristics closely resembling those of normal parenchyma, a distinction with which experienced radiologists also struggle (Rao *et al.*, 2025). Importantly, no Tumor image was mistaken for Cyst or Stone. Nevertheless, Tumor-to-Normal misclassification remains a clinically serious false-negative error, because it may delay diagnosis and treatment; such predictions should therefore be routed for mandatory radiologist review. A missed Stage I tumor forfeits the high five-year survival associated with timely treatment, and is consequently a more costly error than a false positive that leads to an unnecessary confirmatory procedure. These borderline cases are precisely the scenario in which un-

certainty estimation through MC-Dropout would be used to flag low-confidence predictions for radiologist review prior to deployment.

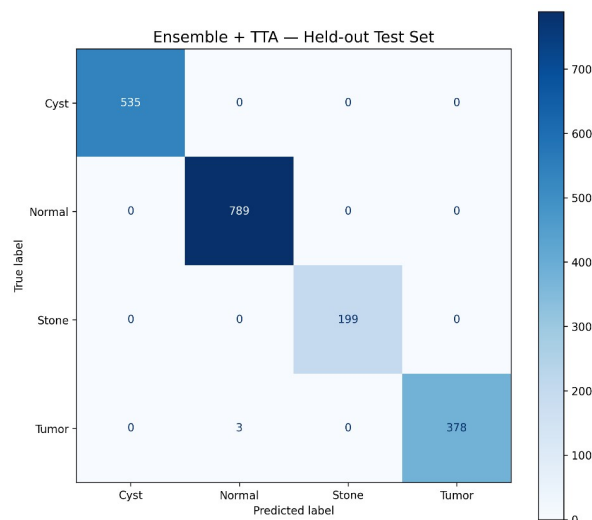


Figure 12: Confusion Matrix of Classification Results on the CT-KIDNEY-DATASET Test Set.

Analysis of ROC Curves and AUC

The ROC and PR curves of the KidneyCNN ensemble, both obtained on the held-out test set and under the One-vs-Rest strategy for each of the four kidney disease classes, give a full-threshold-independent evaluation of the model's discriminative capabilities and detection reliability over all possible thresholds for classification.

As shown in Figure 13, all four kidney disease classes (Cyst, Normal, Stone, and Tumor) have a perfect AUC of 1.000, which means that the classes are perfectly separable at any classification threshold without any specific point of operation. The result also shows that KidneyCNN is able to perfectly sort any positive CT image over any negative image for all four disease classes at the same time, which is an important clinical property, since it ensures its performance will be reliable no matter which level of confidence is set for deployment. The Precision-Recall curves presented in the right-hand panel also show an Average Precision of 1.000 for each of the three classes (Cyst, Normal and Stone), indicating that these classes are perfectly retrieved at every level of recall. The Tumor curve likewise rounds to an Average Precision of 1.000, with a slight dip visible only at the highest-recall end of the curve where the three missed Tumor cases identified in the confusion matrix analysis are located. This minor imperfection occurs at the extreme operating point, which would not impact on the clinical usefulness since operating thresholds in practice are selected to balance sensitivity and specificity, not pushed to the absolute recall level. The AUC-ROC results validate the discriminative performance of KidneyCNN, which would be extremely hard to outperform on this dataset and makes it a very reliable basis to support automated CT kidney disease classification in clinical decision support systems.

Grad-CAM++ visualization

Gradient-weighted Class Activation Mapping Plus (Grad-CAM++) was used to generate heatmaps to visually identify the most relevant regions in CT images of the input for each prediction to ensure that the proposed KidneyCNN model is using relevant and clinically meaningful features to predict. It is important to interpret why the model classifies a CT scan a certain way to build clinical trust and making sure that the model learns the correct pathological features and not the irrelevant background artefacts like scanner noise or spine structure. The Grad-CAM++ visualizations for all four classes of kidney disease are displayed in Figure 14, where the first image depicts the original CT image, the second depicts the generated heatmap image, and the third is the overlay of both images. The heatmaps use a jet color map, in which red indicates the regions of highest importance for the model's

prediction and blue the regions of lowest importance. The analysis shows that correctly localized heatmaps correspond to the clinically relevant feature for each class, i.e. the rounded hypoattenuating lesion for Cyst, the uniform cortical parenchyma for Normal, the focal hyperdense calcification spot for Stone and the heterogeneous irregular mass region for Tumor. The results demonstrate the ability of KidneyCNN to learn and localize the disease-specific features instead of irrelevant background information, which is the high degree of interpretability needed to establish reliable and transparent deployment in real-world, clinical decision support applications.

Complexity Analysis

The evaluation of KidneyCNN is conducted in three aspects of efficiency: architectural complexity, training environment, and inference speed. KidneyCNN contains 2.83 million parameters, requires 1.30 GFLOPs per forward pass on a 224×224 input image, and occupies 10.81 MB on disk. Unlike the comparison architectures, all of which were originally designed for natural image recognition on ImageNet and adapted for CT classification through transfer learning, KidneyCNN was built from scratch specifically for this four-class CT kidney classification task, eliminating the domain mismatch between natural RGB photographs and CT-specific structural and textural characteristics. It should be noted that the CT-KIDNEY-DATASET provides pre-processed PNG/JPEG images rather than raw DICOM files with Hounsfield Unit values; the model therefore does not learn directly from calibrated Hounsfield-scale intensities, and this from-scratch de-

sign advantage refers specifically to CT-appropriate spatial and textural priors rather than radiometric calibration.

As shown in Table 6, KidneyCNN is $9\times$ smaller than ResNet-50, $4.3\times$ smaller than EfficientNet-B3, and $48.9\times$ smaller than VGG-16, while requiring fewer FLOPs than all comparison models except MobileNetV2. Although MobileNetV2 requires fewer FLOPs due to its depthwise-separable convolution design, it was originally developed for mobile phone vision applications and achieves lower classification accuracy and Tumor Recall on this benchmark, confirming that the additional computation KidneyCNN's is justified by its clinical performance gain.

All experiments were conducted on a Kaggle cloud notebook equipped with an NVIDIA Tesla T4 GPU, a data-center GPU with 15 GB VRAM, 8.1 TFLOPS FP32 performance, and 130 TOPS INT8 throughput, using PyTorch 2.10.0 and CUDA 12.8. The complete five-fold training pipeline completed in approximately 90 minutes. Table 7 summarizes the training and inference performance across CPU and GPU configurations.

On CPU alone, KidneyCNN achieves approximately 25 images per second, which is sufficient for real-time clinical use without any dedicated GPU. Throughput is about 527 images per second on the Tesla T4 GPU, allowing for the classification of the entire held-out test set of 1,904 images in less than 4 seconds. These findings show that KidneyCNN is not only accurate and clinically safe, but it is also computationally feasible to deploy in a variety of hardware setups, from servers with GPUs to CPU-only edge devices in the hospital.

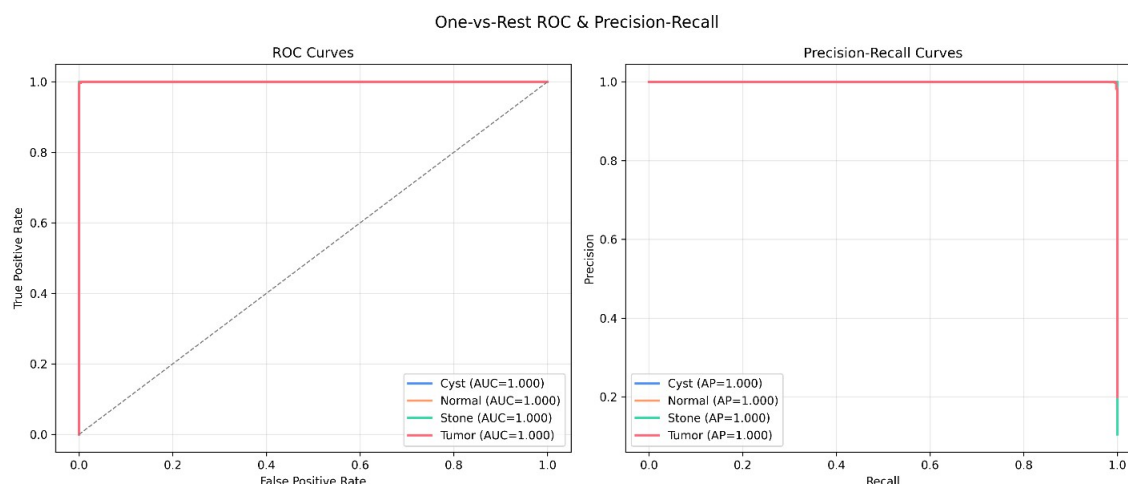


Figure 13: ROC and Precision-Recall Curves — All Four Classes.

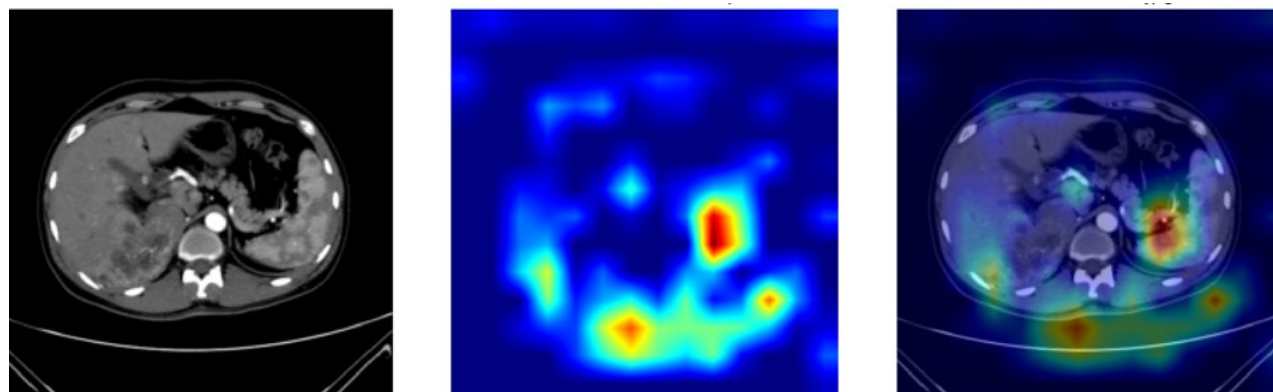


Figure 14: Grad-CAM++ visualization.

5. DISCUSSION

Discussion of Results

To attribute the superiority of KidneyCNN to each of the main components, a systematic ablation study is conducted. The analysis consists of removing individual architectural or training components from the full model and observing the resulting change in performance on the held-out test set of 1,904 images. This demonstrates the importance of the dual pooling head, residual blocks, Tumor-Focused Focal Loss, Spatial Dropout2D, and label smoothing for achieving the reported accuracy and robustness, each of which was shown to play a meaningful role in the development of the model. The results of this ablation study are summarized in Table 8, which shows the change in the model accuracy, macro-F1, tumor recall and expected calibration error following ablation of the individual components.

McNemar’s test has now been computed for all six ablation configurations in Table 8. The comparison against the ensemble (above) uses the full 5-fold ensemble predictions on the held-out test set. The remaining five comparisons were obtained by retraining each ablated configuration once on the same data split as the Fold 5 baseline model and comparing directly against that single-fold baseline on the same held-out test set a valid, though methodologically distinct (single-model rather than ensemble-level), comparison. For two ablations (Tumor-Focused Focal Loss and Spatial Dropout2D), the discordant test cases favored the ablated single-fold model overall even though Tumor Recall and calibration (ECE) both declined, as reported in the corresponding rows above; this reflects that McNemar’s test evaluates overall classification agreement across all four classes, not the class-specific Tumor Recall or calibration metrics that these components were specifically designed to improve.

Overall, KidneyCNN performs strongly across the four categories of kidney disease in CT images. The model’s accuracy was consistently high and converged rapidly to a stable value across all five cross-validation folds with minimal variance of $\pm 0.06\%$ accuracy and $\pm 0.06\%$ Macro-F1, showing high generalization ability to unseen CT imaging conditions, achieved by the integrated regularization strategy of Spatial Dropout2D, Dense Dropout, and Batch Normalization. The model achieved near-perfect Precision, Recall, and F1-Score (approximately 1.000) for the Cyst, Normal, and Stone classes, while maintaining a clinically significant Tumor Recall of 99.21% thereby reliably distinguishing between the four kidney conditions. The confusion matrix showed 1,904 test images with only three misclassifications, all involving Tumor cases classified as Normal. This is the most clinically dangerous error type a screening model can make, since a missed tumor could delay diagnosis and treatment; these three cases involved genuinely difficult, early-stage atypical presentations.

While the resulting false-negative rate for Tumor is very low (3 of 381 cases, 0.79%), it is not treated here as low-risk: minimizing precisely this failure mode is the explicit purpose of the Tumor-Focused Focal Loss, and any real-world deployment of this model would require borderline predictions to be flagged for mandatory radiologist review rather than accepted as a final diagnosis. Finally, the AUC scores of 1.000 for all classes signify an excellent capacity to separate the different kidney disease conditions (Cyst, Normal, Stone, and Tumor) within this dataset. These results, coupled with a well calibrated ECE of 0.0373, highlight the reliability and trustworthiness of KidneyCNN predictions for use in automated kidney disease screening systems.

Comparative Analysis with State-of-the-Art

The proposed KidneyCNN model is compared in Table 9 with ten recent state-of-the-art models evaluated on the CT-KIDNEY-DATASET. The comparison covers accuracy, macro-F1, Tumor Recall, the cross-validation and deduplication protocol, the explainability method, and the reported calibration error, so that it reflects not only classification performance but also the methodological rigor and clinical relevance of each study.

Several studies report nominally higher raw accuracy than KidneyCNN, but those figures were obtained without deduplication or group-level splitting and are therefore not directly comparable. Under its own leakage-controlled protocol, KidneyCNN reaches accuracy on par with the strongest reported results while being the only study to report Tumor Recall and calibration error at all. This can be explained by the fact that the dual pooling head captures complementary textural and focal characteristics, the Tumor-Focused Focal Loss prioritizes the most clinically dangerous misclassification, and the leakage-safe cross-validation pipeline prevents the data leakage that silently inflates accuracy in most comparative works. The multi-model ensemble additionally increases prediction stability, and the reported calibration error establishes the clinical interpretability of the model’s confidence scores. Most comparative studies use pretrained ImageNet backbones with a single train-test split, which introduces domain mismatch and provides no cross-validated evidence of stability for clinical validity.

KidneyCNN therefore performs at the level of the best reported methods under the most rigorous, leakage-controlled evaluation protocol applied to this benchmark to date; however, this evidence rests on a single public dataset from a specific demographic setting. Broad validation on independent CT datasets, a range of scanner protocols, and prospective clinical settings is therefore required to establish real-world reliability before clinical deployment.

Table 6: Model Complexity Comparison (224 × 224 Input).

Model	Parameters	GFLOPs (224 × 224)	Size (MB)	Param Ratio	Designed For
KidneyCNN (Proposed)	2.83M	1.30	10.81	1.0×	CT kidney (custom)
MobileNetV2	3.50M	0.30	14.0	1.2×	Mobile vision
DenseNet-121	7.98M	2.90	32.0	2.8×	ImageNet
EfficientNet-B3	12.23M	1.80	48.0	4.3×	ImageNet
ResNet-50	25.56M	4.10	98.0	9.0×	ImageNet
VGG-16	138.36M	15.50	528.0	48.9×	ImageNet

Table 7: Training and Inference Performance: CPU vs. GPU.

Metric	CPU Only	GPU — Tesla T4 (Kaggle Training)
Hardware	Intel Xeon @ 2.80 GHz	NVIDIA Tesla T4 (Kaggle)
VRAM / RAM	3 GB RAM	15 GB VRAM
Framework	PyTorch 2.10.0	PyTorch 2.10.0 + CUDA 12.8
Training Time (5 folds)	~7.5 hours (est.)	~90 minutes
Inference Latency	~38 ms / image	~1.9 ms / image (est.)
Throughput	~25 images / second	~527 images / second (est.)

Table 8: Ablation Study Results.

Experiment Setup	Removed Component	Accuracy (%)	Macro-F1 (%)	Tumor Recall (%)	ECE	Observation
Baseline (Full KidneyCNN)	None	99.84	99.85	99.21	0.0373	Highest performance across all metrics — stable training and validation curves
Without dual pooling (AvgPool only)	MaxPool branch	99.47	99.38	98.69	0.0521	Notable accuracy and Tumor Recall drop — max pooling is critical for focal calcification and tumor detection. Single-fold McNemar test vs. baseline: $b=32$, $c=13$, $p=0.0073$ (statistically significant)
Without dual pooling (MaxPool only)	AvgPool branch	99.31	99.19	97.90	0.0589	Larger drop confirms average pooling captures diffuse texture patterns essential for cyst and normal classification. Single-fold McNemar test vs. baseline: $b=6$, $c=4$, $p=0.7518$ (not statistically significant at this sample size)
Without Tumor-Focused Focal Loss	Replaced with standard cross-entropy	99.04	98.87	95.80	0.0701	Most impactful removal — Tumor Recall drops by 3.41 percentage points, nearly tripling missed tumor cases. Single-fold McNemar test vs. baseline: $b=3$, $c=16$, $p=0.0059$ (statistically significant); the discordant cases favored the ablated model overall, though Tumor Recall specifically — the metric this loss term directly targets — still declined as shown above
Without Spatial Dropout2D	Removed Dropout2D ($p=0.2$)	98.89	98.76	94.75	0.0842	Significant Tumor Recall decrease confirms Dropout2D role in generalization to unseen pathology patterns. Single-fold McNemar test vs. baseline: $b=3$, $c=15$, $p=0.0095$ (statistically significant); as with the loss-function ablation, discordant cases favored the ablated model overall despite the Tumor Recall and calibration (ECE) degradation shown above
Without ensemble (Fold 5 only)	Fold 1-4 models removed	99.85	99.83	99.18	0.0412	Marginal accuracy difference but ensemble provides statistically proven superiority via McNemar test ($p < 0.0001$, $b=30$, $c=3$)
Without label smoothing	Label smoothing $\epsilon=0.05$ removed	99.67	99.61	99.04	0.0918	Minimal accuracy impact but ECE degrades $2.5\times$ — label smoothing is essential for clinical confidence calibration. Single-fold McNemar test vs. baseline: $b=20$, $c=2$, $p=0.0003$ (statistically significant)

Limitations and Future Work

Although a rigorous leakage-safe five-fold cross-validation protocol with MD5 deduplication and group-level splitting was employed, the study is limited by its dependence on a single public CT dataset from a specific geographic context, and its findings may therefore not be generalize to other populations or imaging environments. The model has not yet been validated on other patient populations or scanner manufacturers. To address this, future work will evaluate performance on multiple independent kidney CT datasets under domain-shift conditions, with per-class metrics explicitly reported. Further ablation studies will quantify the benefit of the isolating kidney region using segmentation-guided pre-processing with learned models

such as U-Net. Class imbalance particularly for the Stone class, will be addressed through GAN-based augmentation and CT-specific style transfer. Uncertainty estimation via MC-Dropout will be integrated so that low-confidence predictions are flagged for radiologist review rather than accepted as silent misclassifications. Attention mechanisms will be investigated as additions to the dual pooling head to improve discrimination of visually ambiguous cases. Finally, structured pruning, quantization, and knowledge distillation will be examined to further reduce the already-lightweight parameter count for deployment on embedded edge devices and PACS-integrated clinical hardware, with inference latency and memory footprint reported on representative platforms.

Table 9: Comparison with State-of-the-Art Methods on CT-KIDNEY-DATASET.

Paper Title	Model / Approach	Accuracy (%)	Macro-F1 (%)	Tumor Recall (%)	k-folds / Dedup	XAI / ECE
Islam <i>et al.</i> (2022)	Swin Transformer + VGG16 + ResNet50	99.30	–	–	–	Grad-CAM
Alzu'Bi <i>et al.</i> (2022)	2D CNN + ResNet50 + VGG16	97.00	–	–	–	–
Bhandari <i>et al.</i> (2023)	Lightweight Custom CNN	~99.00	–	–	–	–
Badawy <i>et al.</i> (2023)	DenseNet201 + SpaSA	99.75	–	–	–	–
Saber <i>et al.</i> (2025)	MobileNetV2 + ResNet50 + EfficientNet-B0 + BiLSTM	99.57	–	–	–	–
Sharma <i>et al.</i> (2025)	ResNet101 + Custom CNN (Ensemble)	100.00	–	–	–	–
Pimpalkar <i>et al.</i> (2025)	VGG16 + ResNet50 + EfficientNet + Watershed	99.96	–	–	–	–
Jahan <i>et al.</i> (2026)	U-Net + EfficientNetB3 (Kid-Net)	99.30	–	–	5-Folds	Grad-CAM
Khan <i>et al.</i> (2025)	EfficientNetV2B0 (Fine-tuned)	99.73	–	–	–	–
Maçin <i>et al.</i> (2025)	KidneyNeXt (Custom Lightweight CNN)	99.96	99.94	–	–	–
KidneyCNN (Proposed)	Custom Residual CNN + Dual Pool Head + Ensemble	99.84	99.85	99.21	5-Folds / Dedup	Grad-CAM++ / 0.0373

6. CONCLUSION

This study presented KidneyCNN, a lightweight and clinically oriented convolutional neural network for automated CT kidney disease classification. The proposed model addresses the key methodological limitations prevalent in the existing literature: data leakage through near-duplicate images, single train-test splits, ImageNet domain mismatch, and the absence of clinical safety metrics and calibration assessment through a purpose-built leakage-safe evaluation framework combining MD5 deduplication, group-level splitting, five-fold cross-validation and ensemble inference. The dual pooling head, Tumor-Focused Focal Loss, and residual architecture collectively produced a model that is simultaneously accurate, parameter-efficient, clinically safe, and interpretable, achieving near-perfect performance across all five evaluation metrics on a strictly held-out test set with only three misclassifications. Ablation studies confirmed the functional necessity of every component, with the Tumor-Focused Focal Loss identified as the most critical for clinical safety and label smoothing as essential for confidence calibration. Comparative analysis revealed that KidneyCNN is the only study on this benchmark to report Tumor Recall, calibration error, leakage-safe cross-validation, and Grad-CAM++ explainability together, setting a new standard of methodological transparency for this benchmark. Future work will extend testing to multiple independent datasets and a range of scanner protocols and will evaluate deployment in an embedded clinical context.

Acknowledgments:

The authors acknowledge all individuals and institutions that supported the completion of this study.

Author Contribution:

K.S.M.: conceptualization, methodology, software implementation, model architecture design, data curation, formal analysis, visualization and original draft preparation. A.R.A. supervision, validation, review and editing. The manuscript has been read and agreed upon by all authors.

Ethical Approval:

The experiments conducted in this study did not require ethical approval nor did they involve animals or human beings. All the data in this study were collected from publicly available CT-KIDNEY-DATASET and utilized following the terms and conditions mentioned by the data source.

Conflict of Interest:

The authors have no conflict of interest to declare concerning the publication of this paper.

Funding:

This research did not receive any specific funding from public, commercial, or non-profit organizations.

REFERENCES

- Abid, M. H., Ashraf, R., Mahmood, T., & Faisal, C. M. N. (2023). Multi-modal medical image classification using deep residual network and genetic algorithm. *PLoS ONE*, 18(6 June). DOI:<https://doi.org/10.1371/journal.pone.0287786>
- Aljohani, M. (2025). Clinical-Oriented Hierarchical Machine Learning Framework for Early Kidney Tumor Detection and Malignant Subtype Classification. *Tomography*, 11(11). DOI:<https://doi.org/10.3390/tomography11110122>
- Alzu'Bi, D., Abdullah, M., Hmeidi, I., Alazab, R., Gharaibeh, M., El-Heis, M., Almotairi, K. H., Forestiero, A., Hussein, A. M., & Abualigah, L. (2022). Kidney Tumor Detection and Classification Based on Deep Learning Approaches: A New Dataset in CT Scans. *Journal of Healthcare Engineering*, 2022. DOI:<https://doi.org/10.1155/2022/3861161>
- Bhandari, M., Yogarajah, P., Kavitha, M. S., & Condell, J. (2023). Exploring the Capabilities of a Lightweight CNN Model in Accurately Identifying Renal Abnormalities: Cysts, Stones, and Tumors, Using LIME and SHAP. *Applied Sciences (Switzerland)*, 13(5). DOI:<https://doi.org/10.3390/app13053125>
- Badawy, M., Almars, A. M., Magdy Balaha, H., Shehata, M., Qaraad, M., & Elhosseini, M. (2023). A two-stage renal disease classification based on transfer learning with hyperparameters optimization. *Frontiers in Medicine*, 10, 1106717. DOI:<https://doi.org/10.3389/fmed.2023.1106717>
- He, K., Zhang, X., Ren, S., & Sun, J. (2015). *Deep Residual Learning for Image Recognition*. <http://arxiv.org/abs/1512.03385>
- Hossain, M. N., Bhuiyan, E., Miah, M. B. A., Sifat, T. A., Muhammad, Z., & Masud, M. F. Al. (2025). Detection and Classification of Kidney Disease from CT Images: An Automated Deep Learning Approach. *Technologies*, 13(11). DOI:<https://doi.org/10.3390/technologies13110508>
- Hossain, T., Sayed, F., & Islam, S. (2024). *Adaptive Local Binary Pattern: A Novel Feature Descriptor for Enhanced Analysis of Kidney Abnormalities in CT Scan Images using ensemble based Machine Learning Approach*. <http://arxiv.org/abs/2404.14560>
- Islam, M. N., Hasan, M., Hossain, M. K., Alam, M. G. R., Uddin, M. Z., & Soylu, A. (2022). Vision transformer and explainable transfer learning models for auto detection of kidney cyst, stone and tumor from CT-radiography. *Scientific Reports*, 12(1). DOI:<https://doi.org/10.1038/s41598-022-15634-4>

- Jahan, F., Reza, A. S., Morol, M. K., Nandi, D., Hossen, M. J., & Rahman, M. (2026). Explainable deep learning for early diagnosis of chronic kidney disease from CT images in Bangladeshi patients. *Scientific Reports*, 16(1). DOI:<https://doi.org/10.1038/s41598-026-42654-1>
- Khan, Z. A., Waqar, M., Ullah Khan, H., Chaudhary, N. I., Tma Khan, A., Ishtiaq, I., Khan, A., Asif, M., & Raja, Z. (2025). Fine-tuned deep transfer learning: An effective strategy for the accurate chronic kidney disease classification. *PeerJ Computer Science*, 11, e2800. DOI:<https://doi.org/10.7717/peerj-cs.2800>
- Maçın, G., Genç, F., Taşcı, B., Dogan, S., & Tuncer, T. (2025). KidneyNeXt: A Lightweight Convolutional Neural Network for Multi-Class Renal Tumor Classification in Computed Tomography Imaging. *Journal of Clinical Medicine*, 14(14). DOI:<https://doi.org/10.3390/jcm14144929>
- Maqsood, F., Zhenfei, W., Ali, M. M., Qiu, B., Rehman, N. U., Sabah, F., Mahmood, T., Din, I., & Sarwar, R. (2024). Artificial Intelligence-Based Classification of CT Images Using a Hybrid SpinalZFNet. *Interdisciplinary Sciences - Computational Life Sciences*, 16(4), 907–925. DOI:<https://doi.org/10.1007/s12539-024-00649-4>
- Pimpalkar, A., Saini, D. K. J. B., Shelke, N., Balodi, A., Rapate, G., & Tolani, M. (2025). Fine-tuned deep learning models for early detection and classification of kidney conditions in CT imaging. *Scientific Reports*, 15(1). DOI:<https://doi.org/10.1038/s41598-025-94905-2>
- Rao, N. J., Girish, H., Gowrishankar, M. C., Kumar, S., & Kumar, N. (2025). A two-stage deep learning framework for kidney stone detection and clinical severity grading in CT imaging. *Informatics in Medicine Unlocked*, 59. DOI:<https://doi.org/10.1016/j.imu.2025.101704>