Indonesian Journal of Data and Science

Volume 7 Issue 2
ISSN 2715-9936
https://doi.org/10.56705/ijodas.v7i2.443

Research Article

Effect of Spatial, Intensity, and Hybrid Augmentation on Kidney CT Image Classification

Ardha Ardhana Putra Agustavada [1]; Aji Prasetya Wibawa [2, *]; Dafa Fadhilah Hilmi [3]; Abdullah Sholum [4]; Felix Andika Dwiyanto [5]

[1] Universitas Negeri Malang, Malang, Indonesia, ardha.ardhana.2205356@students.um.ac.id
[2] Universitas Negeri Malang, Malang, Indonesia, aji.prasetya.ft@um.ac.id
[3] Universitas Negeri Malang, Malang, Indonesia, abdullah.sholum.2205356@students.um.ac.id
[4] Universitas Negeri Malang, Malang, Indonesia, dafa.fadhilah.2205356@students.um.ac.id
[5] AGH University of Kraków, al. Adama Mickiewicza 30, Kraków 30-059, Poland, dwiyanto@agh.edu.pl

Correspondence should be addressed to Aji Prasetya Wibawa; aji.prasetya.ft@um.ac.id

Received 10 March 2026; Accepted 29 July 2026; Published 31 July 2026

Abstract:
Introduction: Kidney diseases remain a major global health challenge, and computed tomography (CT) provides detailed renal imaging that supports accurate diagnosis. Although data augmentation is commonly used to improve deep learning performance on limited medical datasets, the effects of different augmentation strategies on image characteristics and model learning behavior remain insufficiently understood. **Method**: This study evaluated spatial, intensity, and hybrid augmentation for kidney CT image classification using a custom CNN, MobileNetV2, and EfficientNet-B0. A public dataset containing 12,446 CT images across Normal, Cyst, Tumor, and Stone classes was partitioned using stratified sampling into training, validation, and test sets. Four scenarios—baseline, spatial, intensity, and hybrid augmentation—were evaluated across three independent random seeds using accuracy, precision, recall, F1-score, and AUC. **Results and Discussion**: The baseline achieved mean accuracies of 99.96%, 98.32%, and 94.06% for CNN, MobileNetV2, and EfficientNet-B0, respectively. Intensity augmentation slightly improved CNN accuracy to 99.99% and consistently produced smaller performance degradation and more stable convergence than spatial and hybrid augmentation. Spatial and hybrid transformations generally reduced classification performance, indicating that excessive geometric changes may disrupt diagnostically relevant anatomical features. **Conclusion**: Baseline training provided the best overall performance, while intensity augmentation was the most effective augmentation strategy, demonstrating that preserving anatomically meaningful image characteristics is more important than indiscriminately increasing data diversity.

Keywords: Kidney CT Images; Data Augmentation; Intensity Augmentation; Deep Learning; Medical Image Classification; Learning Behavior Analysis.

Dataset link: https://www.kaggle.com/datasets/nazmul0087/ct-kidney-dataset-normal-cyst-tumor-and-stone

1. Introduction

Kidney diseases remain a major global health concern, affecting millions of individuals worldwide and contributing significantly to morbidity, mortality, and healthcare costs [1]. Early diagnosis is essential for preventing disease progression and improving patient outcomes [2]. Among various diagnostic imaging modalities, Computed Tomography (CT) has become an important clinical tool due to its ability to provide detailed anatomical information and high-resolution visualization of renal structures. Consequently, CT imaging plays a crucial role in supporting the diagnosis and assessment of kidney-related abnormalities [3].

Recent advances in deep learning have significantly transformed medical image analysis by enabling the automatic extraction of complex features from imaging data [4]. Architectures such as Convolutional Neural Networks (CNN), MobileNetV2, and EfficientNet-B0 have been widely adopted for medical image classification due to their strong

representation learning capabilities [5], [6]. While CNN serves as a fundamental architecture for hierarchical feature extraction, MobileNetV2 emphasizes computational efficiency through depth wise separable convolutions, and EfficientNet-B0 improves performance through compound scaling of network dimensions. These architectural differences may influence how each model responds to variations introduced during training.

Despite their success, deep learning models require large and diverse datasets to achieve robust performance [7]. However, obtaining extensive annotated medical imaging datasets remains challenging because of privacy concerns, annotation costs, and the need for expert interpretation [8]. As a result, limited data availability often increases the risk of overfitting and reduces model generalizability, particularly in kidney CT image classification where anatomical and pathological variations can be substantial [9].

To address this issue, data augmentation has become a widely adopted strategy for increasing dataset diversity without acquiring additional data [10]. Among the most common approaches are spatial augmentation, which applies geometric transformations such as rotation and flipping, and intensity augmentation, which modifies image properties such as brightness, contrast, and noise levels [11]. Hybrid augmentation combines both approaches to enhance geometric and photometric diversity simultaneously [12]. These techniques can improve model robustness by exposing networks to a wider range of image variations during training.

Data augmentation is widely used to enhance the performance of deep learning models in medical image classification [13]. However, most existing studies focus primarily on predictive performance, offering limited analysis of how different augmentation strategies alter image characteristics and influence the learning behavior of various deep learning architectures [14]. Furthermore, comparative studies examining spatial, intensity-based, and hybrid augmentation across multiple architectures are limited. As a result, there is an incomplete understanding of how these augmentation strategies affect kidney CT image characteristics and classification outcomes. Unlike previous medical image augmentation studies that primarily evaluate classification accuracy using a single augmentation strategy or architecture, this study systematically compares baseline, spatial, intensity, and hybrid augmentation on kidney CT images while analyzing their influence on learning behavior across CNN, MobileNetV2, and EfficientNet-B0 under identical experimental settings.

This research addresses this gap by systematically investigating the effects of online augmentation strategies on kidney CT image classification. The study evaluates spatial, intensity, and hybrid augmentation and examines their impact on the classification performance of CNN, MobileNetV2, and EfficientNet-B0. Beyond comparing predictive performance, this research analyzes architecture-specific learning behavior, including convergence patterns and generalization characteristics, to provide a deeper understanding of how different augmentation strategies influence feature learning in kidney CT image classification. The scope of this study is limited to kidney CT image classification using three deep learning architectures and three categories of online augmentation strategies. Therefore, the findings may not be directly generalized to other imaging modalities, disease domains, augmentation techniques, or neural network architectures beyond those considered in this research.

2. Method

This study investigates the impact of various data augmentation strategies on the classification performance of deep learning models for kidney CT scan images. The experimental workflow comprises five main stages: dataset acquisition, image preprocessing, data augmentation, model development, and performance evaluation. Three classification architectures, a custom CNN, MobileNetV2, and EfficientNetB0, were utilized to assess the effectiveness of spatial, intensity, and hybrid augmentation approaches. All experiments were performed in the Kaggle computational environment, which offers cloud-based resources including two NVIDIA Tesla T4 GPUs (32 GB total VRAM), 4-core CPU, and 29 GB RAM. This hardware configuration facilitated efficient training and evaluation of multiple deep learning scenarios under consistent computational conditions [15]. The overall research workflow is depicted in **Figure 1**.

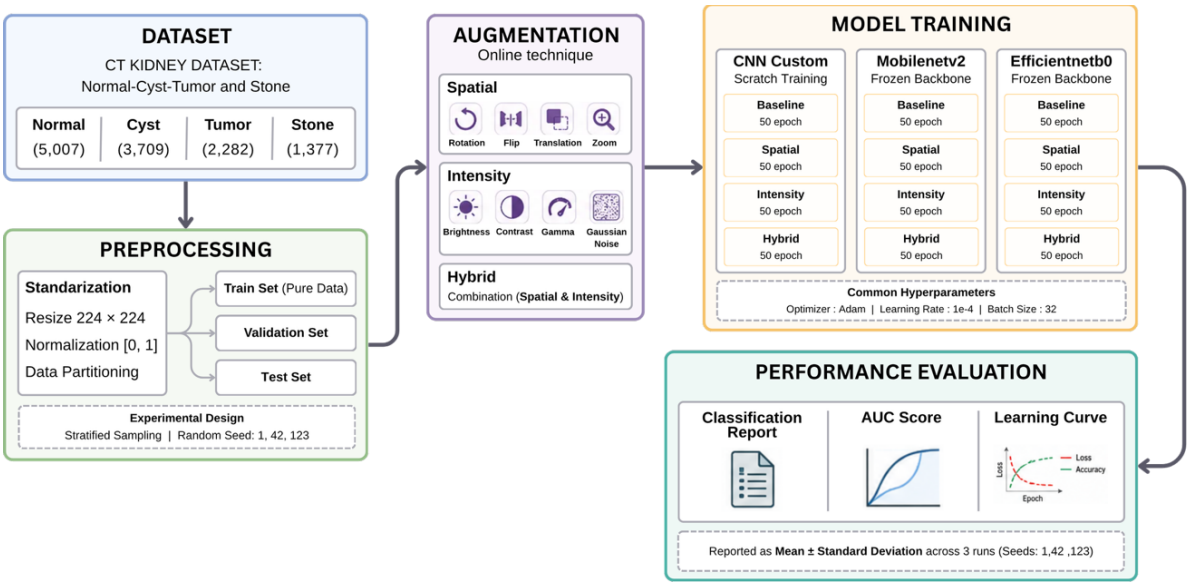
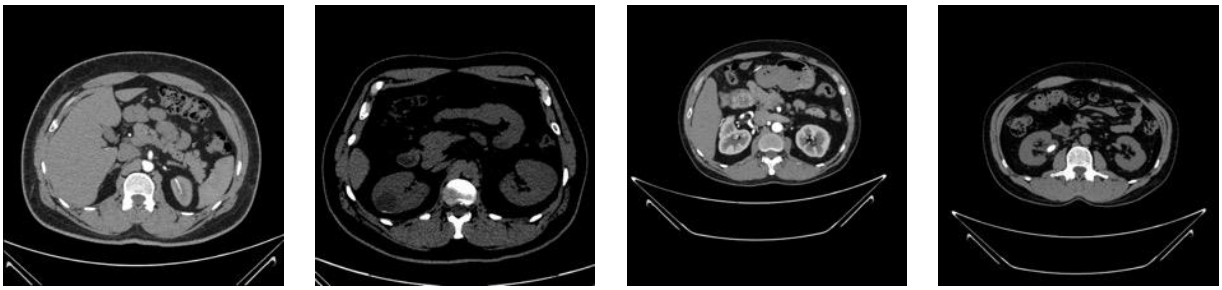


Figure 1. Research Workflow Diagram

Dataset

This study utilizes the CT Kidney Dataset: Normal, Cyst, Tumor, and Stone, which is publicly available on the Kaggle platform. The dataset contains kidney CT images collected from the PACS (Picture Archiving and Communication System) of several hospitals in Dhaka, Bangladesh. The images were annotated by medical experts into four diagnostic categories: Normal, Cyst, Tumor, and Stone. Originally stored in DICOM format, all scans underwent anonymization to remove patient-identifiable information and were subsequently converted to JPG format through a lossless conversion process. The image labels were further verified by radiologists and medical technologists to ensure annotation reliability. The dataset is distributed under the Data files © Original Authors license, allowing its use for academic and research purposes with appropriate attribution to the original creators. However, the dataset does not provide patient identifiers or examination metadata. Consequently, patient-wise data partitioning cannot be performed, making it impossible to verify whether images from the same patient are distributed across different data subsets.



(a) (b) (c) (d)

Figure 2. Representative CT scan images of the four kidney condition classes: (a) Normal, (b) Cyst, (c) Tumor, and (d) Stone.

Figure 2 presents representative CT scan images from the four diagnostic categories, namely (a) Normal, (b) Cyst, (c) Tumor, and (d) Stone. The dataset comprises 12,446 images, including 5,077 Normal images (40.79%), 3,709 Cyst images (29.80%), 2,283 Tumor images (18.34%), and 1,377 Stone images (11.07%). The class distribution is moderately imbalanced, with the Stone category containing the fewest samples. To minimize the risk of data leakage, the dataset was examined for duplicate and visually apparent near-duplicate images before model development. No obvious duplicate images were identified during this inspection. Nevertheless, because patient identifiers are

unavailable, potential patient overlap cannot be completely ruled out. Despite this limitation, the dataset provides sufficient image diversity and sample size to support the development and evaluation of automated kidney disease classification models.

Experimental Design

A standardized experimental protocol was adopted to ensure a consistent and reliable comparison of augmentation strategies across all deep learning models [16]. The protocol defines the dataset partitioning, repeated experimental runs, and evaluation procedure used throughout this study.

The dataset was partitioned into training, validation, and testing subsets using an overall ratio of 60:20:20 through a two-stage stratified sampling procedure. First, the dataset was divided into 80% training-validation and 20% testing subsets using stratified sampling to preserve the original class distribution. Subsequently, the training-validation subset was further divided into 75% training and 25% validation, again using stratified sampling. This procedure resulted in approximately 60% training, 20% validation, and 20% testing while maintaining class balance across all subsets. The class distribution before and after the two-stage stratified data partitioning is summarized in **Table 1**, confirming that the original class proportions were maintained across all data subsets.

Table 1. Class Distribution Across the Training, Validation, And Testing Sets

Class	Original	Train	Validation	Test
Normal	5077	3046	1015	1016
Cyst	3709	2225	742	742
Tumor	2283	1369	457	457
Stone	1377	827	275	275
Total	**12446**	**7467**	**2489**	**2490**

To improve the reliability of the experimental results, each experimental scenario was repeated three independent times using different random seeds (1, 42, and 123). For each run, the complete pipeline-including dataset partitioning, model training, and performance evaluation-was executed independently under identical experimental settings.

For a fair comparison, all augmentation strategies (Baseline, Spatial, Intensity, and Hybrid) were evaluated using the same dataset partitions, optimizer, learning rate, batch size, number of training epochs, and evaluation metrics across the CNN, MobileNetV2, and EfficientNetB0 architectures. Model performance was evaluated exclusively on the held-out test set using Accuracy, Precision, Recall, F1-score, and Area Under the ROC Curve (AUC). All quantitative results are reported as mean ± standard deviation (Mean ± SD) across the three independent runs.

Preprocessing

The preprocessing stage was performed to standardize the kidney CT images before model training. This stage consisted of two main steps: image resizing and pixel normalization. First, all CT images were resized to 224×224 pixels in order to produce a uniform input dimension for the deep learning model. Standardizing the image size is essential because the original dataset may contain images with different spatial resolutions, while the model requires a fixed input shape. By resizing all images to the same dimensions, the learning process can be performed more consistently and efficiently.

After resizing, the pixel values of all images were normalized to the range [0, 1] by dividing each pixel intensity value by 255. This normalization step was applied to the entire dataset to reduce the scale of the input values and improve numerical stability during model training. In addition, normalization helps the optimization process converge more effectively by preventing large variations in pixel intensity values across images [17]. As a result, the model can learn image features more efficiently and achieve more stable performance during training and evaluation.

Augmentation

Data augmentation was employed to investigate the influence of different image transformation strategies on kidney CT image classification performance. In medical image analysis, augmentation is widely adopted to increase data diversity, reduce overfitting, and improve the generalization capability of deep learning models. Nevertheless, because CT images contain clinically important anatomical structures, inappropriate image transformations may alter discriminative visual characteristics and adversely affect classification performance. Therefore, this study evaluated three augmentation strategies, namely spatial augmentation, intensity augmentation, and hybrid augmentation [18].

Spatial augmentation performs geometric transformations that modify the spatial arrangement of image content while preserving the corresponding class label. In contrast, intensity augmentation applies pixel-level manipulations without changing the anatomical structure of the image. Hybrid augmentation combines both approaches within a single augmentation pipeline to generate more diverse training samples by simultaneously introducing geometric and intensity variations [19]. The implementation details and parameter settings of each augmentation strategy are summarized in **Table 2**.

Table 2. Augmentation Strategies Implemented in This Study

Strategy	Transformation	Parameter
Spatial	Random Rotation	±15°
	Random Flip	Horizontal
	Random Translation	Horizontal and vertical shift ±8%
	Random Zoom	Zoom range ±10%
Intensity	Random Brightness	max_delta = 0.15
	Random Contrast	0.80 to 1.20
	Gamma Adjustment	γ = 0.80 to 1.20
	Gaussian Noise	μ = 0.0, σ = 0.02
Hybrid	Spatial + Intensity	Combination of all transformations

As summarized in **Table 2**, spatial augmentation was implemented using a Keras Sequential preprocessing pipeline consisting of Random Rotation, Random Flip, Random Translation, and Random Zoom layers. Random Rotation rotated each image within ±15°, whereas Random Flip performed horizontal image flipping. In addition, Random Translation shifted the image horizontally and vertically by up to ±8%, while Random Zoom applied scaling variations within ±10%. These geometric transformations were intended to simulate variations in object orientation, position, and scale that may occur during CT image acquisition without altering the semantic class of the image.

Intensity augmentation focused on pixel-level transformations while preserving the anatomical structure of kidney CT images. This strategy employed TensorFlow image processing functions, including random brightness adjustment with a maximum delta of 0.15, random contrast adjustment within the range of 0.80 to 1.20, random gamma correction using gamma values between 0.80 and 1.20, and Gaussian noise injection with a mean of 0.0 and a standard deviation of 0.02. Following these operations, pixel values were clipped to the normalized interval of [0.0, 1.0] to maintain numerical stability and ensure consistent input distributions during model training.

Hybrid augmentation integrated both spatial and intensity transformations into a unified augmentation pipeline. In this strategy, geometric transformations were first applied to modify the spatial characteristics of the image, followed by pixel-level intensity manipulations to generate additional appearance variations. By combining these two transformation categories, hybrid augmentation produced training samples with simultaneous geometric and intensity changes, thereby providing greater data diversity than either augmentation strategy individually.

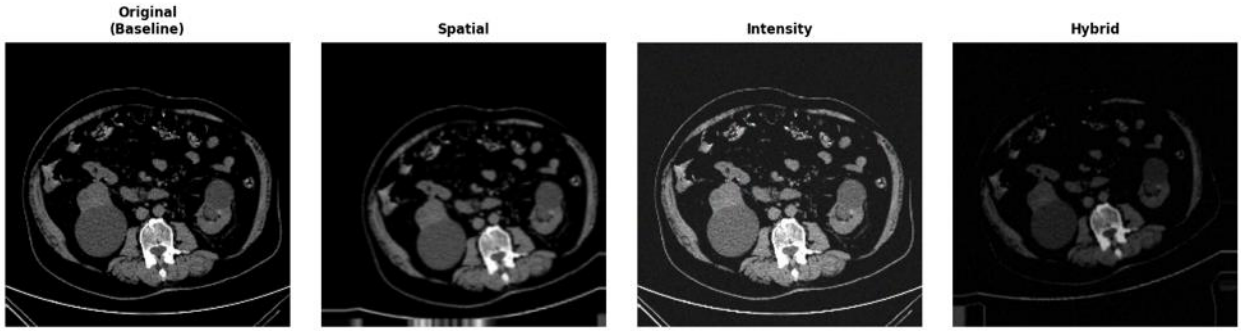


Figure 3. Examples of augmented images.

To improve computational efficiency and avoid permanently increasing dataset size, all augmentation strategies were implemented using an online augmentation scheme within the TensorFlow data pipeline. Rather than generating augmented images prior to training, image transformations were applied dynamically during each training iteration, allowing different augmented versions of the same image to be created across multiple epochs [20]. The online augmentation pipeline was integrated into the TensorFlow tf.data workflow through a custom dataset construction function. Meanwhile, the validation and testing datasets were directly batched and prefetched without passing through the augmentation functions. Consequently, all model evaluations were performed using the original, non-augmented images, ensuring an unbiased comparison of classification performance while preserving the consistency of the experimental protocol across CNN, MobileNetV2, and EfficientNetB0.

Model Architecture

This study employed three deep learning architectures for kidney CT image classification, namely a custom CNN, MobileNetV2, and EfficientNetB0. All models received input images of size $224 \times 224 \times 3$, where the original CT images were converted into three-channel format to match the input requirements of the architectures. The use of these three models was intended to compare the performance of a simple custom CNN with two transfer learning architectures that have been widely adopted in image classification tasks. A summary of the architectural configuration of each model is presented in **Table 2**.

The first model was a custom CNN designed specifically for this study. The architecture consisted of two convolutional blocks followed by a classification layer. The first convolutional layer used 28 filters with a 3×3 kernel and was followed by a ReLU activation and max-pooling operation. The second convolutional layer applied 64 filters with the same 3×3 kernel, followed by ReLU and max-pooling. The resulting feature maps were then flattened and directly connected to a Dense layer with four output neurons and a softmax activation function. No Batch Normalization or Dropout layers were incorporated because the model was intentionally designed as a simple baseline architecture to evaluate the effectiveness of the proposed augmentation strategies without additional regularization mechanisms. This architecture represents a relatively simple baseline model that learns visual features directly from the kidney CT images without relying on pretrained feature representations. The high baseline performance obtained by this architecture is further discussed in the Discussion section in relation to the characteristics and complexity of the dataset as well as the study limitations.

The second model was MobileNetV2, which utilized pretrained ImageNet weights. In this model, the MobileNetV2 backbone was used as a fixed feature extractor by freezing the base model during training. The original top classification layers were removed by setting include_top=False, and the extracted feature maps were processed using a GlobalAveragePooling2D layer. The resulting feature vector was then connected to a Dense layer with four output neurons and softmax activation. The backbone was intentionally kept frozen to ensure that all experimental scenarios were evaluated under the same feature extraction setting, allowing the comparison to focus on the impact of the proposed augmentation methods rather than differences introduced by backbone fine-tuning. In addition, freezing the pretrained backbone reduces the number of trainable parameters and lowers the risk of overfitting when training on a relatively limited medical imaging dataset. Compared with the custom CNN, MobileNetV2 is a more efficient

architecture because it employs depthwise separable convolutions and inverted residual blocks, which reduce computational complexity while maintaining effective feature extraction capability [21].

The third model was EfficientNetB0, which also used pretrained ImageNet weights and a frozen backbone during training. In this study, EfficientNetB0 was implemented through a TF Hub KerasLayer backbone, which internally performs Global Average Pooling and produces a 1280-dimensional feature vector. Similar to MobileNetV2, the backbone was not fine-tuned so that all augmentation scenarios could be evaluated using an identical pretrained feature extractor, thereby providing a fair comparison across architectures while reducing the risk of overfitting caused by the relatively small number of trainable medical images. To improve regularization before classification, a Dropout layer with a rate of 0.5 was added, followed by a Dense layer with four output neurons and softmax activation. In contrast to MobileNetV2, EfficientNetB0 includes dropout in its classification head to reduce overfitting. In terms of architectural concept, EfficientNetB0 differs from the other two models by using a compound scaling strategy, which balances network depth, width, and input resolution to achieve efficient performance [21].

To ensure a fair comparison, all scenarios across the three architectures were trained using the same hyperparameter settings. The training process employed the Adam optimizer with a learning rate of 1×10^{-4}, a batch size of 32, and 50 epochs for all experimental scenarios. These settings were consistently maintained across the custom CNN, MobileNetV2, and EfficientNetB0 so that differences in augmentation performance could be attributed primarily to the characteristics of the model architectures, rather than variations in training configurations.

Table 3. Summary of Model Architectures and Hyperparameter Settings Used in This Study

Model	Input Size	Main Architecture	Classification Head	Hyperparameters
CNN	224×224×3	Conv2D (28, 3×3) + ReLU + MaxPooling; Conv2D (64, 3×3) + ReLU + MaxPooling; Flatten	Dense (4, Softmax)	optimizer = Adam, learning rate = 1 × 10^{-4}, batch size = 32, epochs = 50 all scenario
MobileNetV2	224×224×3	Backbone with pretrained ImageNet weights, include_top=False, base model frozen	GlobalAveragePooling2D + Dense(4, Softmax)	optimizer = Adam, learning rate = 1 × 10^{-4}, batch size = 32, epochs = 50 all scenario
EfficientNetB0	224×224×3	Backbone with pretrained ImageNet weights, internal Global Average Pooling	Dropout(0.5) + Dense(4, Softmax)	optimizer = Adam, learning rate = 1 × 10^{-4}, batch size = 32, epochs = 50 all scenario

Performance Evaluation

Model performance was evaluated using five classification metrics: accuracy, precision, recall, F1-score, and Area Under the Curve (AUC). The final performance of each model is reported as the mean ± standard deviation (Mean ± SD) across the three independent experimental runs. All metrics are calculated based on the values in the confusion matrix, which consists of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). The mathematical formulations of accuracy, precision, recall, and F1-score are presented in Equations (1)-(4).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

Accuracy measures the proportion of correctly classified samples among all observations and provides an overall assessment of classification performance, as shown in Equation (1). Precision, defined in Equation (2), quantifies the

proportion of positive predictions that are actually correct, reflecting the model's ability to minimize false positive errors. Recall, presented in Equation (3), measures the proportion of actual positive samples that are successfully identified by the model and indicates its capability to reduce false negative errors. Meanwhile, the F1-score in Equation (4) represents the harmonic mean of precision and recall, providing a balanced evaluation when both metrics are equally important, particularly in classification tasks where class distributions may be uneven [22].

In addition, this study employs the Area Under the Receiver Operating Characteristic Curve (AUC) to evaluate the model's ability to distinguish between classes across different classification thresholds. An AUC value close to 1 indicates excellent discriminative capability, whereas an AUC value near 0.5 indicates performance comparable to random guessing [23]. Beyond classification metrics, learning curve analysis was performed to assess the training dynamics and generalization behavior of each model throughout the training process. A model is considered to exhibit good generalization when the training and validation curves converge with a small performance gap while maintaining stable loss reduction [24]. Therefore, learning curve analysis complements the quantitative evaluation metrics by providing a deeper understanding of how different augmentation strategies influence the learning behavior of the model.

3. Result and Discussion

This section examines the effects of various augmentation strategies on kidney CT-scan classification performance using CNN, MobileNetV2, and EfficientNetB0 architectures. The analysis encompasses overall performance comparisons, evaluations of learning behavior through accuracy-loss curves, and assessments of confusion matrices to identify class-level prediction patterns. The findings offer insights into the appropriateness of each augmentation strategy for medical image classification tasks.

Overall Performance Comparison

Table 4 and **Figure 4** present a comparison of the performance of CNN, MobileNetV2, and EfficientNetB0 under baseline, spatial, intensity, and hybrid augmentation settings. The findings indicate that intensity augmentation maintained model performance more effectively than spatial or hybrid augmentation. However, the baseline scenario consistently yielded the highest performance across all evaluated architectures.

Table 4. Overall Performance of All Experimental Scenarios Mean ± SD

Model	Method	Accuracy (%)	Precision	Recall	F1-Score	AUC Score
CNN	Baseline	99.96 ± 0.04	0.9996 ± 0.0004	0.9996 ± 0.0004	0.9996 ± 0.0004	1.0 ± 0.0
CNN	Spatial	86.37 ± 3.29	0.8871 ± 0.0155	0.8637 ± 0.0329	0.8637 ± 0.0283	0.9777 ± 0.0035
CNN	Intensity	99.99 ± 0.02	0.9999 ± 0.0002	0.9999 ± 0.0002	0.9999 ± 0.0002	0.9999 ± 0.0
CNN	Hybrid	70.62 ± 1.78	0.8053 ± 0.0117	0.7062 ± 0.0178	0.6849 ± 0.0073	0.9432 ± 0.0019
MobileNetV2	Baseline	98.32 ± 0.06	0.9833 ± 0.0006	0.9832 ± 0.0006	0.9832 ± 0.0006	0.9993 ± 0.0002
MobileNetV2	Spatial	71.94 ± 2.03	0.7979 ± 0.0109	0.7194 ± 0.0203	0.7199 ± 0.0207	0.9477 ± 0.0037
MobileNetV2	Intensity	91.86 ± 0.44	0.9198 ± 0.0042	0.9186 ± 0.0044	0.9180 ± 0.0044	0.9912 ± 0.0002
MobileNetV2	Hybrid	82.28 ± 0.3	0.8246 ± 0.002	0.8228 ± 0.003	0.8137 ± 0.0037	0.9581 ± 0.0049
EfficientNetB0	Baseline	94.06 ± 0.58	0.9423 ± 0.0063	0.9406 ± 0.0058	0.9379 ± 0.0061	0.9935 ± 0.0005
EfficientNetB0	Spatial	74.83 ± 2.22	0.8035 ± 0.0128	0.7483 ± 0.0222	0.7500 ± 0.0201	0.9565 ± 0.0009
EfficientNetB0	Intensity	86.64 ± 0.19	0.8663 ± 0.0018	0.8664 ± 0.0019	0.8650 ± 0.0021	0.9772 ± 0.0013
EfficientNetB0	Hybrid	79.69 ± 0.8	.0.8201 ± 0.0052	0.7969 ± 0.008	0.8007 ± 0.0067	0.9533 ± 0.001

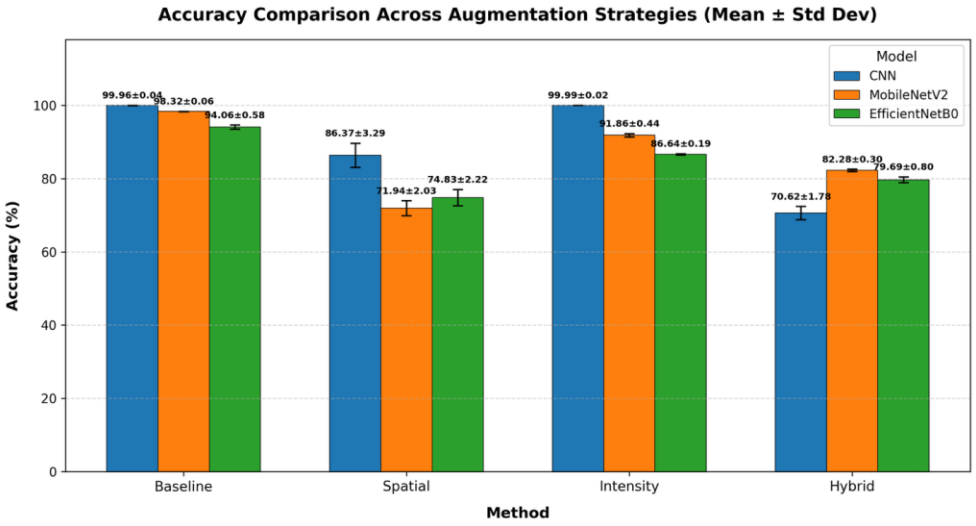


Figure 4. Accuracy Comparison Across Augmentation Strategies (Mean ± Std Dev).

Based on **Table 4** and **Figure 4**, it is observed that the effect of augmentation strategies on model accuracy is not uniform but depends on the architecture used. Across all models, the baseline scenario yielded the highest accuracy compared to other augmentation methods. This outcome will be consistently referred to as the "baseline accuracy" throughout the remainder of this section. The CNN achieved an accuracy of 99.96%, followed by MobileNetV2 at 98.32% and Efficient NetB0 at 94.06%. These values will serve as the primary reference point in subsequent discussions. This result indicates that the original data already optimally represents class characteristics, meaning that applying transformations to the training data does not always improve the model's classification performance.

The application of spatial augmentation is correlated with a consistent decline in performance across all tested architectures, contrary to the general expectation that data diversification enhances a model's generalization capabilities. The most substantial drop in accuracy occurred in the CNN, decreasing from 99.96% to 86.37%, followed by MobileNetV2 from 98.32% to 71.94%, and Efficient NetB0 from 94.06% to 74.83%. Equivalent degradation was also observed in precision, recall, and F1-score values. Nevertheless, the AUC values of the three models remained within the range of 0.9477-0.9777, indicating that the ability to distinguish between classes was maintained across various decision thresholds, even though the rate of label prediction errors increased significantly.

In contrast to the degradation profile exhibited by spatial augmentation, intensity augmentation demonstrated far superior performance stability and maintained classification accuracy at a competitive level across all architectures. For the CNN, applying this method yielded an accuracy of 99.99%, with precision, recall, and F1-score values each reaching 0.9999, while the AUC was 0.9999-just 0.01% below the baseline scenario. MobileNetV2 and Efficient NetB0 also demonstrated comparative advantages over other augmentation methods, achieving accuracies of 91.86% and 86.64%, respectively.

Among all augmentation strategies evaluated, hybrid augmentation-which simultaneously applies spatial and intensity transformations-consistently exhibited the poorest performance across most experimental conditions. The largest relative decline was observed in the CNN architecture, with accuracy falling to 70.62% and the corresponding F1-score dropping to 0.6849, marking the lowest value recorded throughout the entire experimental series. MobileNetV2 and Efficient NetB0 recorded accuracies of 82.28% and 79.69%, respectively. Both models still fell short of the results achieved in the intensity augmentation scenario, despite maintaining AUC values within the range of 0.9432-0.9581.

From a cross-architecture comparison perspective, CNN demonstrated the highest resilience to intensity augmentation, as evidenced by an accuracy retention of 99.99% and an AUC of 0.9999. Conversely, the two transfer-learning-based models-MobileNetV2 and EfficientNetB0-exhibited greater sensitivity to all tested augmentation scenarios. Nevertheless, the AUC values for both models remained above the 0.94 threshold across all methods, confirming that their basic discriminative capabilities were preserved despite a decline in classification accuracy.

Overall, these findings indicate that intensity augmentation was the most effective augmentation strategy for preserving classification performance across the evaluated models, although it did not consistently outperform the baseline scenario. Meanwhile, spatial and hybrid augmentation proved counterproductive, as they have the potential to obscure the visual features that form the foundation of decision-making in the classification process.

Impact of Augmentation on CNN

Figure 4 illustrates the learning dynamics of a CNN architecture under various augmentation strategies, while **Figure 5** displays the confusion matrix for the augmentation scenario that yields the smallest performance degradation; these two figures collectively facilitate the evaluation of the augmentations applied to the custom CNN, and they must be preserved in any subsequent analysis to maintain argument coherence and scope integrity.

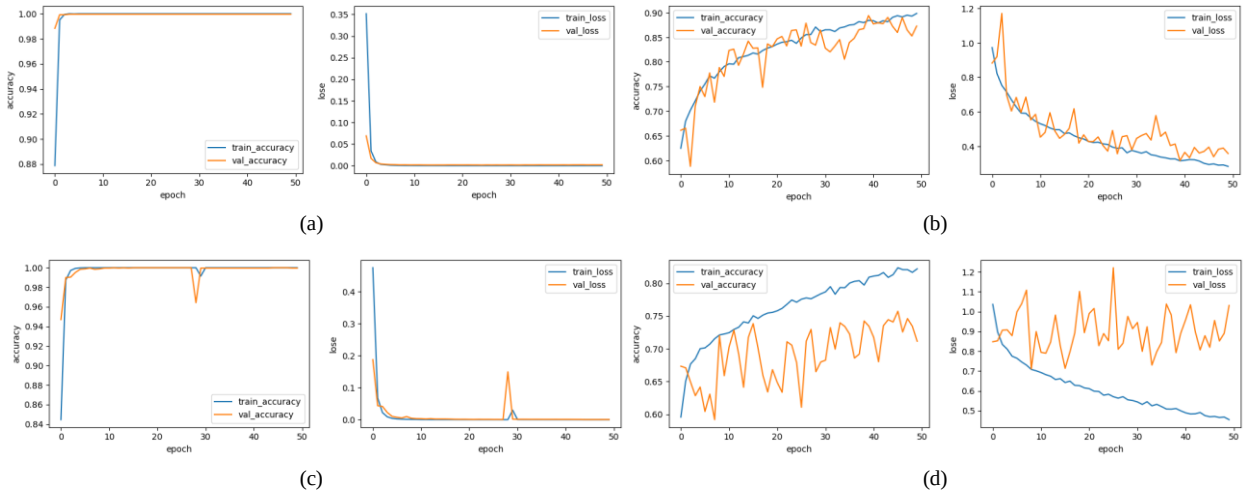


Figure 5. Training and validation performance (accuracy and loss) of CNN architecture across different augmentation strategies (Run 1, random seed = 1): (a) Baseline, (b) Spatial, (c) Intensity, and (d) Hybrid. Left panels show accuracy; right panels show loss.

As shown in **Figure 5a**, the baseline scenario exhibits a highly efficient learning curve: the training accuracy reaches a value close to 1.00 even before the fifth epoch, starting from an initial value of 0.84, and the validation accuracy also stabilizes at this peak value, remaining consistent throughout the remaining iterations. Analogous dynamics are observed in the loss functions, as the training loss and validation loss both transition from approximately 0.35 and 0.10, respectively, toward values close to zero. This indicates the achievement of optimal convergence with nearly perfect discriminative capacity.

The application of spatial augmentation (**Figure 5b**) resulted in a training profile characterized by relative instability compared to the baseline condition. Although the training loss showed a fairly substantial decrease (from approximately 1.00 to 0.30), the model's generalization capacity appeared to be hindered, as reflected by the validation loss fluctuating widely within the range of 0.30 to 1.20. Concurrently, training accuracy increased gradually from approximately 0.60 to 0.90, whereas validation accuracy fluctuated within the range of 0.60-0.88 until the end of training.

In contrast, the intensity augmentation depicted in **Figure 5c** produced a training pattern that closely resembled the baseline conditions. Training accuracy transitioned rapidly from 0.84 to a value approaching perfection within the first few epochs, while validation accuracy remained mostly within the range of 0.98-1.00. Although a notable transient spike occurred around epoch 28 where validation accuracy briefly dipped and validation loss spiked up to 0.25 both components of the loss function generally continued to show a stable downward trajectory toward zero.

Among all conditions evaluated, the hybrid augmentation scenario in Fig. 4d exhibits the most degraded learning profile, and the training accuracy, which rose to only 0.82 from an initial value of 0.60, is the lowest across all scenarios; it is accompanied by a validation accuracy that remains confined to the range 0.60-0.75 and exhibits

significant fluctuations. More diagnostically concerning is the behavior of the validation loss: while the training loss shows a steady decrease from 1.10 to 0.20, the validation loss oscillates in the range of 0.40-1.30 without any identifiable convergence trend. Overall, the findings in **Figure 2** indicate that not all augmentation strategies have an equivalent impact on CNN training dynamics. Intensity augmentation was found to produce learning stability closest to the baseline conditions, while hybrid augmentation resulted in the most significant degradation in generalization among all tested scenarios.

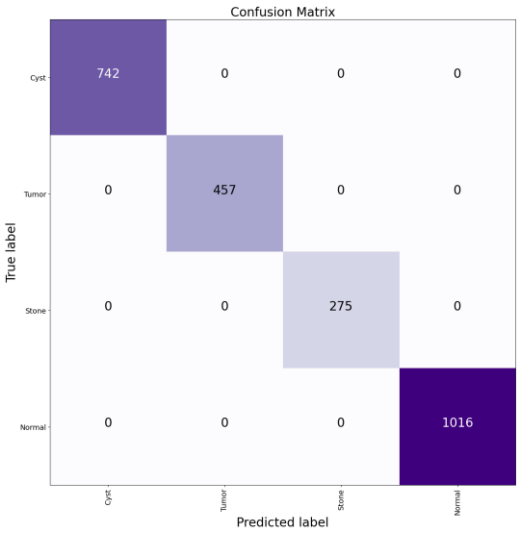
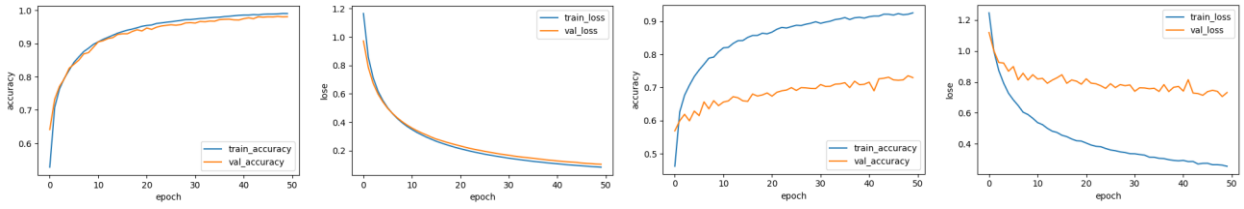


Figure 6. Confusion matrix of CNN using the Intensity augmentation strategy (representative Run 1, random seed = 1)

The confusion matrix in **Figure 6** illustrates the performance of the CNN with intensity augmentation, which achieved perfect predictions for this specific representative run across a total of 2,490 test images. All four classes Cyst (742/742), Tumor (457/457), Stone (275/275), and Normal (1016/1016) were classified without a single error. Given the perfectly concentrated distribution of values along the main diagonal of the confusion matrix and the absence of any non-diagonal elements, these results confirm the model's near-perfect discriminative ability across all evaluated categories.

Impact of Augmentation on MobileNetV2

Figure 7 presents the learning behavior of MobileNetV2 under baseline, spatial, intensity, and hybrid augmentation conditions using Run 1, and **Figure 7** presents the confusion matrix for intensity augmentation, which is reported smallest performance degradation for the MobileNetV2 architecture; the following training and validation curves therefore provide further illustrative evidence of how different augmentation strategies influence the model's capacity to learn discriminative features from kidney CT scans.



(a) (b)

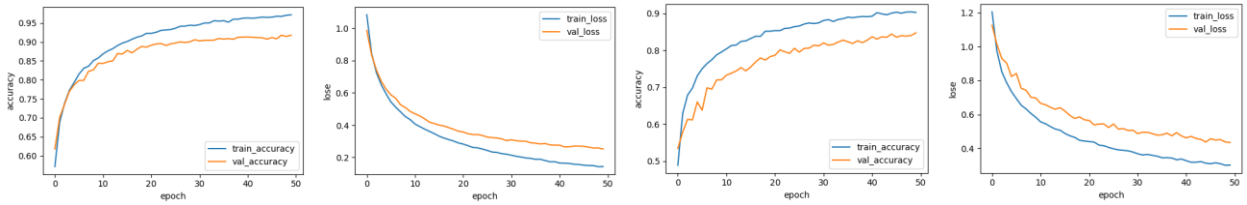


Figure 7. Training and validation performance (accuracy and loss) of MobileNetV2 across different augmentation strategies (Run 1, random seed = 1): (a) Baseline, (b) Spatial, (c) Intensity, and (d) Hybrid. Left panels show accuracy; right panels show loss.

Figure 6 illustrates the significant differences in the learning dynamics of MobileNetV2 among the four augmentation strategies tested. In the baseline scenario (**Figure 6a**), both the training and validation loss curves show a steady downward trend toward a value of approximately 0.10, with the gap between the curves remaining small throughout the training session. Accordingly, training and validation accuracy increase consistently until both reach the range of 0.98-0.99 by the 50th epoch.

The spatial augmentation scenario depicted in **Figure 6b** is characterized by a substantial gap between the training and validation domains, and the training loss decreases to around 0.2, whereas the validation loss remains in the 0.70 range-a much larger difference compared to the baseline scenario. A similar divergence is observed in the accuracy metrics: training accuracy reached around 0.92, while validation accuracy plateaus at only 0.73, suggesting that spatial transformations substantially reduce the model's generalization capability.

Intensity augmentation (**Figure 6c**) produced a more stable learning profile and was closer to the baseline scenario than spatial augmentation. The training and validation loss curves decrease consistently throughout training, reaching approximately 0.15 and 0.26, respectively, by the final epoch. Likewise, training and validation accuracies increase steadily to approximately 0.97 and 0.92, respectively. Compared with spatial augmentation, the smaller gap between the training and validation curves indicates improved generalization, suggesting that intensity-based transformations preserve diagnostically relevant image characteristics more effectively while still introducing useful variability during training.

The hybrid augmentation scenario in **Figure 6d** yields performance that falls between spatial augmentation and intensity augmentation. The training loss decreases gradually to around 0.30, while the validation loss remains at around 0.45; training and validation accuracies reach approximately 0.90 and 0.85, respectively, in the final epoch. The gap between the training and validation domains in this method is smaller than in the spatial augmentation scenario, although the convergence rate achieved is still below that of the baseline and intensity augmentation. Overall, Fig. 6 shows that intensity augmentation is the strategy most compatible with the MobileNetV2 architecture, producing learning behavior closest to the baseline conditions, while spatial augmentation results in the greatest performance degradation among all tested methods.

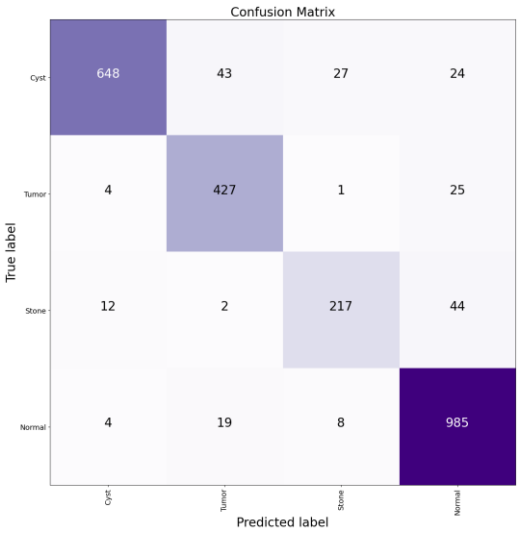


Figure 8. Confusion matrix of MobileNetV2 using the Intensity augmentation strategy (representative Run 1, random seed = 1)

Figure 8 presents the confusion matrix for MobileNetV2 with intensity augmentation using Run 1, demonstrating variations in classification performance across the four classes while maintaining a dominant main diagonal. The Normal class achieved the highest performance, with 985 of 1,016 samples correctly classified, while only 31 samples were misclassified into the Cyst (4), Tumor (19), and Stone (8) classes. The Cyst class recorded 648 of 742 correct predictions, with most errors distributed across the Tumor (43), Stone (27), and Normal (24) classes. The Tumor class achieved 427 of 457 correct predictions, with the majority of misclassifications directed to the Normal class (25 samples), followed by the Cyst class (4). The Stone class remained the most challenging category, with 217 of 275 samples correctly classified, while 44 samples were incorrectly predicted as the Normal class. Overall, the confusion matrix indicates that the Normal class achieved the highest classification performance, whereas the Stone class remained the most challenging to classify. The frequent misclassification of Cyst, Tumor, and Stone samples as Normal suggests that these classes share overlapping visual characteristics with normal kidney CT images.

Impact of Augmentation on EfficientNetB0

Figure 9 illustrates the learning behavior of EfficientNetB0 under baseline, spatial, intensity, and hybrid augmentation settings, and Fig. 9 presents the confusion matrices corresponding to the augmentation scenarios that yield the smallest performance degradation; these figures are used to evaluate the impact of augmentation on the EfficientNetB0 architecture.

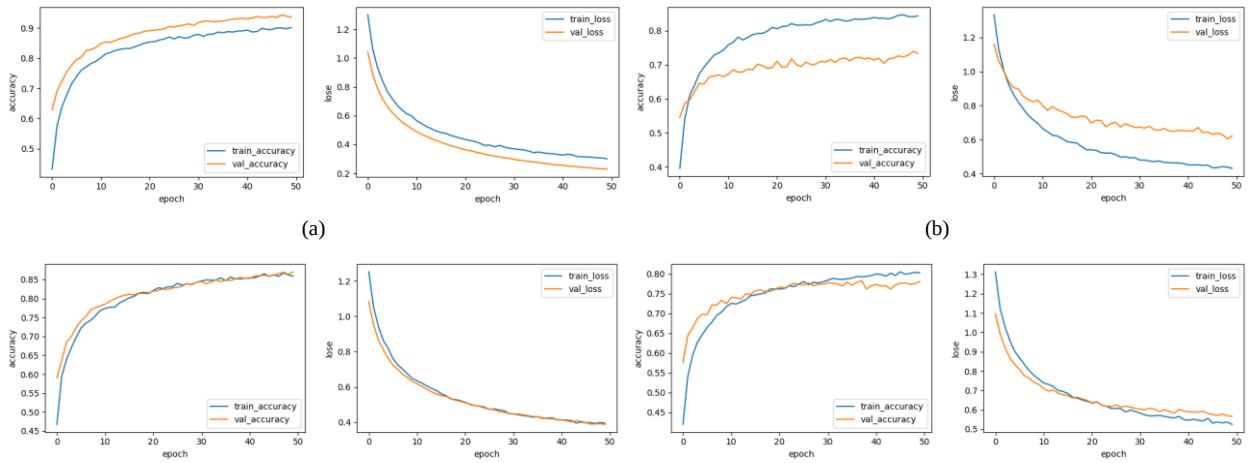


Figure 9. Training and validation performance (accuracy and loss) of EfficientNetB0 across different augmentation strategies (Run 1, random seed = 1): (a) Baseline, (b) Spatial, (c) Intensity, and (d) Hybrid. Left panels show accuracy; right panels show loss.

Figure 9 illustrates the significant differences in the learning patterns of EfficientNetB0 across the four augmentation strategies tested. In the baseline scenario depicted in Figure 9a, the training accuracy increased from approximately 0.40 in the early epochs to approximately 0.90 by the end of training, whereas the validation accuracy exhibited a steeper improvement and attained a higher value of approximately 0.95. Both the training and validation loss curves simultaneously decreased from values above 1.0 to approximately 0.3 and 0.24, respectively, and the validation loss fell below the training loss by the end of the session.

The spatial augmentation configuration (Figure 9b) is characterized by increasing divergence between the training and validation curves as training progresses, and the decrease in training loss from approximately 1.25 to 0.43 was not proportionally matched by the validation loss, which ended in the range of 0.61, thereby indicating a larger gap compared with the baseline configuration. Concurrently, the training accuracy increased from approximately 0.44 to 0.84, while the validation accuracy only reached about 0.72 at epoch 50.

Intensity augmentation (Figure 9c) produced the most distinctive learning profile among all scenarios, characterized by training and validation accuracy curves that were nearly coincident throughout the training process. The training and validation loss curves declined in parallel from values around 1.1-1 toward the 0.40 range, with nearly identical trajectories in both domains, and, consequently, the training and validation accuracy increased steadily from the 0.40-0.55 range in the early epochs to around 0.86 in the final epoch.

The hybrid augmentation scenario in Figure 9d exhibits relatively stable learning dynamics, as both the training and validation loss curves decrease together from values above 1.1 to the 0.56 range, indicating that the gap between the curves remains small throughout training. Training accuracy increased from approximately 0.38 to 0.80, while validation accuracy followed a similar trend and slightly surpassed it, reaching approximately 0.81 at the end of training. Although the convergence level achieved was still below that of the baseline scenario, this profile of a small gap between the curves resembled the pattern observed in the baseline scenario. Overall, Fig. 8 demonstrates that intensity augmentation yields the most consistent generalization behavior on EfficientNetB0, while spatial augmentation results in the greatest training-validation divergence, and hybrid augmentation maintains learning stability despite underperforming compared to the baseline and intensity augmentation.

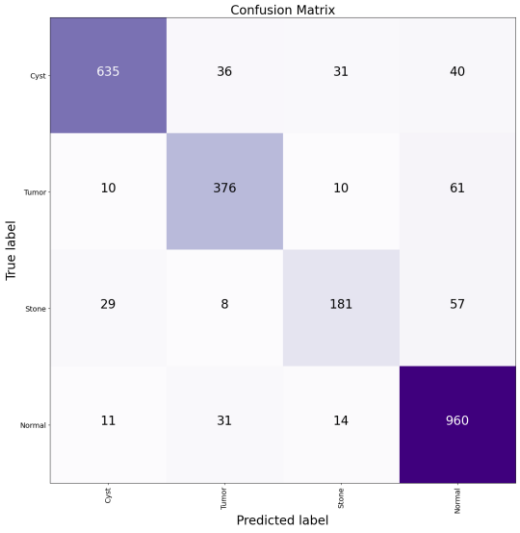


Figure 10. Confusion matrix of EfficientNetB0 using the Intensity augmentation strategy (representative Run 1, random seed = 1)

Figure 10 presents the confusion matrix for EfficientNetB0 with intensity augmentation, demonstrating substantial performance variations across classes while the main diagonal remains dominant. The Normal class recorded the highest performance (960/1016 correct, 56 misclassified into other classes), followed by Cyst (635/742 correct; misclassifications: Tumor 36, Stone 31, Normal 40) and Tumor (376/457 correct, with the highest number of misclassifications 61 into Normal). The lowest performance was recorded in the Stone class: out of 275 samples, only 181 were classified correctly, while 57 were incorrectly predicted as Normal, 29 as Cyst, and 8 as Tumor. The high misclassification rate in the Stone class (94 out of 275 samples, or 34.2%) and the consistent pattern of errors favoring the Normal class across all classes indicate that the most substantial visual discrimination challenge occurs between the Stone and Normal classes.

Discussion

The results of this study show that the baseline scenario generally achieved the highest classification performance across the evaluated architectures. Although intensity augmentation produced a marginally higher mean accuracy than the baseline for CNN (99.99 ± 0.02 vs. 99.96 ± 0.04), baseline performance remained superior for MobileNetV2 and EfficientNetB0, indicating that augmentation did not provide consistent performance improvements across different model architectures. Among the evaluated augmentation strategies, intensity augmentation consistently preserved classification performance better than spatial and hybrid augmentation while producing a stable learning process. These findings indicate that intensity-based variations, such as adjustments of brightness and contrast, can expand data diversity without altering the anatomical characteristics of the kidneys, which constitute the primary source of diagnostic information in CT-scan images. Consequently, intensity augmentation introduces appearance variability while maintaining the structural information required for feature extraction. In contrast, spatial transformations such as rotation, flipping, translation, and zoom may modify the anatomical orientation and spatial relationships of renal structures, potentially generating image variations that are less representative of actual CT acquisition conditions. As a result, the learned feature representations become less consistent, leading to more fluctuating learning behavior and reduced classification performance [25]. This observation is consistent with the experimental results, where spatial and hybrid augmentation consistently produced larger performance degradation than intensity augmentation across all evaluated models. Thus, the effectiveness of augmentation in medical image classification is determined not only by its ability to generate more diverse data but also by its ability to preserve the integrity of anatomical structures relevant to feature extraction and model decision-making.

An interesting finding of this study is that the custom CNN consistently outperformed the transfer learning models (MobileNetV2 and EfficientNetB0) under all evaluated scenarios, including the baseline configuration. One possible explanation is that the pretrained ImageNet features used by MobileNetV2 and EfficientNetB0 were learned from natural RGB images, which differ substantially from grayscale kidney CT images. Moreover, the pretrained backbones were intentionally kept frozen throughout all experiments to ensure a fair comparison of augmentation strategies by preventing performance differences caused by model-specific fine-tuning. Under this controlled setting, the lightweight CNN trained directly on the target dataset may have learned task-specific representations that were more suitable for distinguishing kidney CT characteristics than the fixed pretrained features. Nevertheless, the consistently high baseline performance also suggests that the original dataset already provides highly discriminative visual patterns, reducing the potential benefit of additional data augmentation.

The findings of this study have practical implications for the development of Computer-Aided Diagnosis (CAD) systems based on renal CT scans. Rather than improving performance beyond the original dataset, the results demonstrate that intensity augmentation is the most suitable augmentation strategy when additional data variability is required, as it consistently preserved classification performance better than the other evaluated augmentation methods. Furthermore, the results of this study indicate that the selection of an augmentation strategy is a critical factor that can influence the stability of learning and the model's generalization ability in medical imaging applications. The repeated experimental protocol using three independent random seeds also produced relatively small standard deviations across all performance metrics, indicating that the observed performance trends were consistent and not substantially affected by random initialization.

This study still has several limitations. Although the experimental reliability was improved by employing repeated runs with three different random seeds and reporting the results as Mean ± Standard Deviation, the publicly available CT Kidney Dataset does not provide patient identifiers. Consequently, patient-level data partitioning could not be performed, and the possibility of patient overlap between dataset partitions cannot be completely excluded. Furthermore, the evaluation focused only on three groups of augmentation strategies: spatial, intensity, and hybrid augmentation. In addition, the transfer learning models were evaluated using frozen pretrained backbones to isolate the effect of augmentation, whereas fine-tuning was not investigated. Future research may explore augmentation techniques specifically designed for CT images-such as Contrast Limited Adaptive Histogram Equalization (CLAHE), windowing augmentation, or CT acquisition noise simulation-to gain a deeper understanding of the impact of augmentation on classification performance [26]. Future studies should also evaluate patient-level data partitioning, external validation datasets, and fine-tuning strategies to further improve the generalizability of the findings.

Overall, this study demonstrates that the effectiveness of augmentation strategies in kidney CT-scan image classification depends not only on increasing data diversity but also on preserving anatomically relevant information for robust feature learning. Although intensity augmentation produced a marginally higher mean accuracy than the baseline for CNN, the baseline remained superior for MobileNetV2 and EfficientNetB0, indicating that augmentation did not consistently across architectures. Nevertheless, among the evaluated augmentation strategies, intensity augmentation most effectively preserved classification performance while maintaining stable learning behavior. These findings highlight the importance of selecting augmentation strategies that are compatible with the intrinsic characteristics of medical images and provide practical guidance for developing more reliable deep learning-based computer-aided diagnosis systems for kidney CT imaging.

4. Conclusion

This study evaluated the effects of three augmentation strategies, namely spatial, intensity, and hybrid augmentation, on kidney CT image classification using CNN, MobileNetV2, and EfficientNetB0. The results show that, although intensity augmentation achieved a marginally higher mean accuracy than the baseline for CNN, the baseline configuration remained superior for MobileNetV2 and EfficientNetB0 and therefore served as the best overall training strategy. Among the evaluated augmentation methods, intensity augmentation consistently produced smaller performance degradation relative to the baseline and exhibited more stable convergence than the spatial and hybrid augmentation strategies. These findings indicate that the effectiveness of augmentation depends not only on increasing data diversity but also on preserving anatomically relevant image characteristics that support robust feature learning. Therefore, the findings should be interpreted as evidence of performance preservation rather than overall performance improvement through augmentation.

This study still has several limitations. Although experimental reliability was improved through repeated runs using three random seeds and reporting the results as Mean ± Standard Deviation, the publicly available CT Kidney Dataset does not provide patient identifiers, making patient-level data partitioning impossible and preventing complete exclusion of potential patient overlap between dataset partitions. Furthermore, the evaluation was limited to three augmentation strategy groups (spatial, intensity, and hybrid), and the transfer learning models were assessed using frozen pretrained backbones without investigating fine-tuning. Future research should explore augmentation techniques specifically designed for CT images, such as Contrast Limited Adaptive Histogram Equalization (CLAHE), windowing augmentation, and CT acquisition noise simulation, while also incorporating patient-level data partitioning, external validation datasets, and fine-tuning strategies to further improve model generalizability. Overall, these findings provide empirical evidence that selecting augmentation strategies consistent with the intrinsic characteristics of medical images is essential for developing robust and reliable deep learning-based computer-aided diagnostic systems.

References:

[1] A. Francis *et al.*, “Chronic kidney disease and the global public health agenda: an international consensus,” *Nat. Rev. Nephrol.*, vol. 20, no. 7, pp. 473–485, Jul. 2024, doi: https://doi.org/10.1038/s41581-024-00820-6.

[2] R. Setyati *et al.*, "The Importance of Early Detection in Disease Management," *Journal of World Future Medicine*, vol. 2, no. 1, pp. 51–63, 2024, doi: https://doi.org/10.55849/health.v2i1.692.

[3] S. S. R. Vulasala *et al.*, "Computed tomography of renal emergencies: A comprehensive diagnostic guide for radiology residents," *World J. Nephrol.*, vol. 15, no. 1, Mar. 2026, doi: https://doi.org/10.5527/wjn.v15.i1.114185.

[4] I. D. Mienye, T. G. Swart, G. Obaido, M. Jordan, and P. Ilono, "Deep Convolutional Neural Networks in Medical Image Analysis: A Review," *Information*, vol. 16, no. 3, p. 195, Mar. 2025, doi: https://doi.org/10.3390/info16030195.

[5] H. E. Kim, A. Cosa-Linan, N. Santhanam, M. Jannesari, M. E. Maros, and T. Ganslandt, "Transfer learning for medical image classification: a literature review," *BMC Med. Imaging*, vol. 22, no. 1, p. 69, Dec. 2022, doi: https://doi.org/10.1186/s12880-022-00793-7.

[6] H. Khachnaoui, B. Chikhaoui, N. Khlifa, and R. Mabrouk, "Enhanced Parkinson's Disease Diagnosis Through Convolutional Neural Network Models Applied to SPECT DaTSCAN Images," *IEEE Access*, vol. 11, pp. 91157–91172, 2023, doi: https://doi.org/10.1109/ACCESS.2023.3308075.

[7] G. Tummalapalli, O. Gurrapu, K. N. Kumar, J. Venkata Suman, A. V. Rao, and M. Prabhu, "Deep Learning Approaches for Enhancing Image Classification Accuracy in Medical Imaging," in *2025 Devices for Integrated Circuit (DevIC)*, IEEE, Apr. 2025, pp. 16–21. doi: https://doi.org/10.1109/DevIC63749.2025.11012289.

[8] J. C. L. Chow, "Machine learning in cancer imaging for enhanced precision in diagnosis and therapy," *Discover Computing*, vol. 29, no. 1, p. 186, Mar. 2026, doi: https://doi.org/10.1007/s10791-026-10078-0.

[9] F. Jahan, A. S. Reza, M. K. Morol, D. Nandi, Md. J. Hossen, and M. Rahman, "Explainable deep learning for early diagnosis of chronic kidney disease from CT images in Bangladeshi patients," *Sci. Rep.*, vol. 16, no. 1, p. 14819, Mar. 2026, doi: https://doi.org/10.1038/s41598-026-42654-1.

[10] Y. Dang *et al.*, "Data Augmentation for Sequential Recommendation: A Survey," *IEEE Trans. Knowl. Data Eng.*, vol. 38, no. 8, pp. 4938–4957, Aug. 2026, doi: https://doi.org/10.1109/TKDE.2026.3692969.

[11] N. Kozah, F. Dornaika, J. Charafeddine, and J. El Jaam, "Data Augmentation Techniques for Medical Image Segmentation – A Review," in *2024 International Conference on Computer and Applications (ICCA)*, IEEE, Dec. 2024, pp. 1–8. doi: https://doi.org/10.1109/ICCA62237.2024.10927851.

[12] F. Garcea, A. Serra, F. Lamberti, and L. Morra, "Data augmentation for medical imaging: A systematic literature review," *Comput. Biol. Med.*, vol. 152, p. 106391, Jan. 2023, doi: https://doi.org/10.1016/j.compbiomed.2022.106391.

[13] K. Rais, M. Amroune, M. Y. Haouam, A. Benmachiche, and S. Abid, "Dynamic feature context activation and data augmentation for enhanced medical image segmentation," *Multimed. Tools Appl.*, vol. 85, no. 2, p. 131, Feb. 2026, doi: https://doi.org/10.1007/s11042-026-21296-5.

[14] J. Majidpour and H. Beitollahi, "A Comprehensive Examination of Machine Learning and Deep Learning Approaches for Breast Cancer Detection, Classification, Segmentation, Augmentation, and Feature Selection," *Archives of Computational Methods in Engineering*, vol. 33, no. 2, pp. 1913–1944, Mar. 2026, doi: https://doi.org/10.1007/s11831-025-10359-9.

[15] P. Dabove, M. Daud, and L. Olivotto, "Revolutionizing urban mapping: deep learning and data fusion strategies for accurate building footprint segmentation," *Sci. Rep.*, vol. 14, no. 1, p. 13510, Jun. 2024, doi: https://doi.org/10.1038/s41598-024-64231-0.

[16] D. T. Pham, T. V. Tran, X. Zhu, and H. N. Pham, "Optimising deep learning for building extraction: Dataset efficiency and model backbones under data constraints," *Remote Sens. Appl.*, vol. 41, p. 101876, Jan. 2026, doi: https://doi.org/10.1016/j.rsase.2026.101876.

[17] M. T. R *et al.*, "Transformative Breast Cancer Diagnosis using CNNs with Optimized ReduceLROnPlateau and Early Stopping Enhancements," *International Journal of Computational Intelligence Systems*, vol. 17, no. 1, p. 14, Jan. 2024, doi: https://doi.org/10.1007/s44196-023-00397-1.

[18] S. M. Rayavarapu, T. S. Prasanthi, S. R. Gottapu, and A. Singam, "A Comprehensive Overview on Data Augmentation Techniques for Medical Images," in *2024 5th International Conference on Electronics and Sustainable Communication Systems (ICESC)*, IEEE, Aug. 2024, pp. 1324–1329. doi: https://doi.org/10.1109/ICESC60852.2024.10690109.

[19] T. Islam, Md. S. Hafiz, J. R. Jim, Md. M. Kabir, and M. F. Mridha, "A systematic review of deep learning data augmentation in medical imaging: Recent advances and future research directions," *Healthcare Analytics*, vol. 5, p. 100340, Jun. 2024, doi: https://doi.org/10.1016/j.health.2024.100340.

[20] S. T. Marc, R. Belavkin, D. Windridge, and X. Gao, "An Evolutionary Approach to Automated Class-Specific Data Augmentation for Image Classification," 2024, pp. 170–185. doi: https://doi.org/10.1007/978-3-031-50320-7_12.

[21] R. Jain, R. Sharma, D. Tiwari, K. Joshi, and V. Jain, "Enhanced Classification of Intel Images Using Refined EfficientNet and MobileNetV2 Frameworks," in *2023 4th International Conference on Intelligent Technologies (CONIT)*, IEEE, Jun. 2024, pp. 1–4. doi: https://doi.org/10.1109/CONIT61985.2024.10627673.

[22] A. Tilevik, "Classification and Performance Metrics," in *Multivariate Statistics and Machine Learning in R For Beginners*, Cham: Springer Nature Switzerland, 2025, pp. 147–170. doi: https://doi.org/10.1007/978-3-032-01851-9_10.

[23] Y. Xia and J. Sun, "Area under the Receiver Operating Characteristic Curve (AUC-ROC)," in *Machine Learning for Microbiome Statistics*, Boca Raton: Chapman and Hall/CRC, 2026, pp. 504–519. doi: https://doi.org/10.1201/9781003610281-17.

[24] O. T. Turan, M. Loog, and D. M. J. Tax, "Generalization performance distributions along learning curves," *Pattern Recognit. Lett.*, vol. 201, pp. 29–36, Mar. 2026, doi: https://doi.org/10.1016/j.patrec.2026.01.003.

[25] J. Zhu, T. Ma, J. Li, C. Zeng, Q. Fu, and D. Jing, "ETD-Det: Oriented Object Detection With Spectral Diffusion Encoding and Extended-Gaussian Decoding," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 64, pp. 1–16, 2026, doi: https://doi.org/10.1109/TGRS.2026.3663195.

[26] S. Krishnendu and M. Biradar, "Enhancing diagnostic information in abdominal computed tomography (CT) images through optimized image enhancement techniques," *Phys. Eng. Sci. Med.*, vol. 49, no. 1, pp. 475–487, Mar. 2026, doi: https://doi.org/10.1007/s13246-025-01679-y.