Indonesian Journal of Electronics, Electromedical Engineering, and Medical Informatics e-ISSN: 2656-8624
Homepage: https://ijeeemi.org/; Vol. 8, No. 2, pp. 172-188, May 2026

RESEARCH PAPER

OPEN ACCES

A Comparative Analysis of Lightweight Deep Learning Models for CT-Based Kidney Disease Classification to Support Early Detection in Geriatric Care

Ardha Ardhana Putra Agustavada[1], Aji Prasetya Wibawa[1], Abdullah Sholum[1], Dafa Fadhilah Hilmi[1], and Felix Andika Dwiyanto[2]

[1] Department of Electrical Engineering and Informatics, Universitas Negeri Malang, Malang, Indonesia
[2] Faculty of Computer Science, AGH University of Kraków, al. Adama Mickiewicza 30, Kraków 30-059, Poland

Abstract

Kidney diseases, including cysts, stones, and tumors, are common among older adults and often progress asymptomatically, leading to delayed diagnoses. Manual interpretation of CT images by clinicians is labor-intensive and can vary significantly between observers, especially in high-volume settings. This study aims to develop and evaluate an artificial intelligence–based decision support system for multiclass kidney disease classification with an emphasis on robustness, computational efficiency, and clinical feasibility in elderly healthcare environments. The study proposes a medical informatics evaluation framework that integrates standard performance metrics with learning dynamics, overfitting analysis, and error distribution assessments to ensure reliable model selection. Three architectures were evaluated: a conventional CNN, MobileNet-V2, and EfficientNet-B0. Experiments were conducted on a publicly available dataset containing 12,446 CT images across four classes (Normal, Cyst, Stone, and Tumor). Models were trained under varying epoch settings and evaluated using weighted accuracy, precision, recall, F1-score, AUC, learning curve analysis, and confusion matrix assessment. The results indicate that the conventional CNN achieved perfect numerical performance but exhibited rapid convergence and early metric saturation, limiting the interpretability of generalization under the current dataset configuration. EfficientNet-B0 showed stable yet conservative performance, whereas MobileNet-V2 achieved near-optimal accuracy with gradual convergence, minimal overfitting, and superior computational efficiency. At the optimal configuration (epoch 50), MobileNet-V2 achieved an accuracy of 1.00, precision of 1.00, recall of 1.00, F1-score of 1.00, and an AUC of 0.9997. These findings suggest that lightweight architectures, particularly MobileNet-V2, offer a practical solution for CT-based kidney disease decision support, while acknowledging the need for patient-level and multi-institutional validation.

Paper History

Received Jan. 29, 2026
Revised Feb 20, 2026
Accepted March 10, 2026
Published April 5, 2026

Keywords

Kidney Disease Classification;
Computed Tomography (CT);
Lightweight Neural Networks;
Medical Image Analysis;
Deep Learning

Author Email

ardha.ardhana.2205356@students.um.ac.id
aji.prasetya.ft@um.ac.id
abdullah.sholum.2205356@students.um.ac.id
dafa.fadhilah.2205356@students.um.ac.id
dwiyanto@agh.edu.pl

I. Introduction

Kidney disease represents one of the most significant and growing health burdens in the elderly population, with prevalence increasing in parallel with aging-related physiological deterioration of renal function and prolonged exposure to comorbid risk factors such as hypertension and diabetes mellitus [1]. In this population, renal pathologies including cysts, calculi, and neoplasms frequently develop without overt clinical symptoms during their early stages, resulting in delayed diagnosis and intervention [2]. Such diagnostic delays are particularly detrimental in older adults, as age-related physiological decline may accelerate disease progression and complicate therapeutic interventions [3]. Consequently, strategies that enable the timely and reliable detection of kidney disease are essential for improving clinical outcomes and strengthening geriatric healthcare systems.

Computed tomography (CT) is widely utilized in clinical practice for the evaluation of renal abnormalities due to its high spatial resolution and ability to provide detailed anatomical information [4]. However, interpreting kidney CT images in the routine clinical setting remains challenging [5]. Radiologists are often confronted with increasing imaging volumes, limited availability of subspecialty expertise, particularly in primary and

Corresponding author: Aji Prasetya Wibawa, aji.prasetya.ft@um.ac.id, Department of Electrical Engineering and Informatics, Universitas Negeri Malang, Malang, Indonesia.
DOI: https://doi.org/10.35882/ijeeemi.v8i2.325

secondary healthcare facilities, and inherent inter-observer variability in image assessment [6]. These constraints hinder early detection and limit the scalability of renal disease screening programs, despite the clinical benefits of early diagnosis and continuous monitoring.

In response to these challenges, deep learning-based approaches, particularly convolutional neural networks (CNNs), have been increasingly applied to automate the classification of kidney diseases from CT images [7]. Numerous studies have reported exceptionally high accuracy values, often approaching perfect performance. An emerging concern in medical imaging research is the "high-performance illusion," in which reported results appear strong but do not translate to clinical settings [8]. This issue frequently arises from dataset-related factors, including data leakage caused by image-level splitting without patient-wise grouping, repeated anatomical slices from the same examination, and limited demographic diversity. In such cases, models may exploit data redundancy or homogeneous image characteristics, leading to overly optimistic performance estimates that do not reflect real-world deployment conditions [9]. Therefore, high numerical accuracy alone is insufficient to indicate clinical reliability, as previously reported in studies discussing data leakage effects in medical imaging model evaluation [10]. To reduce potential evaluation bias, this study explicitly separates the training, validation, and test subsets during model development and performance assessment.

Many previous studies also rely on increasingly complex deep learning architectures without adequate consideration of computational requirements or deployment feasibility [11]. This limitation is particularly important for healthcare facilities with limited computational infrastructure, where diagnostic systems must operate with limited memory and processing power [12]. In this study, lightweight deep learning models refer to neural network architectures designed with reduced parameter count, lower memory consumption, and shorter inference time while maintaining competitive classification performance. Models with these characteristics are more suitable for integration into clinical decision support systems, especially in resource-limited environments such as regional hospitals and outpatient diagnostic units [13]. To address these practical constraints, several studies have begun exploring lightweight deep learning architectures that balance predictive performance with computational efficiency.

Lightweight architectures such as MobileNet and EfficientNet have been proposed as alternatives due to their parameter efficiency and computational practicality [14]. These architectures aim to maintain classification performance while minimizing resource demand, which supports their potential use in routine clinical workflows [15]. Nevertheless, existing research rarely examines how architectural selection affects learning stability, convergence behavior, and resistance to overfitting. The relationship between model selection and deployment readiness for early detection and long-term monitoring of kidney disease in older populations also remains insufficiently studied [16]. Only a limited number of studies have systematically examined the implications of high-performance illusion in public kidney CT datasets and its effect on clinical decision support reliability.

Despite the growing body of research on deep learning–based kidney disease classification from CT images, several critical gaps remain from a medical informatics and clinical decision support perspective [17]. Existing studies predominantly emphasize peak classification accuracy while overlooking system-level evaluation aspects essential for real-world deployment, such as training stability, convergence behavior, susceptibility to overfitting, and error distribution patterns [18]. Moreover, limited attention has been given to how different architectural choices affect the robustness, computational efficiency, and reliability of AI-assisted diagnostic support systems under realistic clinical constraints [19] As a result, there is a lack of comprehensive evaluation frameworks that assess not only model performance but also methodological validity and deployment readiness within medical imaging informatics systems.

Unlike most previous studies that primarily report peak classification accuracy, the proposed framework evaluates deep learning models using a multi-aspect assessment strategy. The framework simultaneously analyzes classification performance, learning stability, parameter efficiency, and inference efficiency within a single evaluation protocol. By integrating computational efficiency metrics with learning behavior analysis, this study positions model suitability not only in terms of predictive accuracy but also practical deploy ability in clinical decision support systems.

To address this gap, this study proposes a structured evaluation framework for an AI-based medical imaging decision-support system for multiclass kidney disease classification from CT images. The framework systematically compares three representative deep learning architectures, namely a conventional CNN, MobileNet-V2, and EfficientNet-B0, which reflect varying levels of model complexity and computational demand [20]. Rather than focusing solely on final classification accuracy, the proposed framework integrates quantitative performance metrics, learning curve analysis, convergence stability assessment, overfitting evaluation, and confusion matrix–based error analysis to provide a holistic understanding of model behavior within a clinical decision support context [21]. Through this comprehensive evaluation strategy, the study aims to better understand how model characteristics influence practical deployment in clinical environments.

Therefore, this study aims to develop and evaluate an AI-based medical imaging decision support framework that assesses the robustness, computational efficiency, and clinical feasibility of deep learning models for CT-based kidney disease classification. The findings are intended to support early detection and monitoring strategies applicable to aging populations, rather than being limited to a specific age-defined cohort [22]. By emphasizing both diagnostic performance and

Corresponding author: Aji Prasetya Wibawa, aji.prasetya.ft@um.ac.id, Department of Electrical Engineering and Informatics, Universitas Negeri Malang, Malang, Indonesia.
DOI: https://doi.org/10.35882/ijeeemi.v8i2.325

implementation feasibility, the proposed framework seeks to bridge the gap between algorithmic development and real-world clinical application.

This study contributes to the field of medical informatics by developing a system-oriented evaluation framework for AI-assisted kidney disease classification using CT images, designed to support clinical decision-making [23]. A comparative analysis of deep learning architectures with distinct computational characteristics is conducted to highlight trade-offs among classification accuracy, generalization stability, and resource efficiency in medical imaging decision support systems [24]. Furthermore, this work provides an in-depth investigation of learning dynamics, overfitting behavior, and error distribution patterns, revealing the limitations of accuracy-centric evaluation when applied to public kidney CT datasets [25]. Finally, practical insights are provided on selecting lightweight deep learning models, particularly MobileNet-V2, to facilitate the development of deployable, reliable, and sustainable AI-based diagnostic support systems in resource-constrained clinical environments [26]. Taken together, these contributions aim to advance the design of clinically deployable AI systems that balance predictive performance with operational feasibility in healthcare settings.

II. Materials and Method

This study employs a structured research methodology to develop a deep neural network (DNN) model for multiclass classification of kidney disease images. The workflow commences with the acquisition of kidney disease images from the Kaggle platform, followed by a preprocessing stage that encompasses image preprocessing and data partitioning. Subsequently, three deep learning architectures, CNN, MobileNet-V2, and EfficientNet-B0, are utilized. Each deep learning architecture was subsequently trained according to a predetermined epoch schedule, with consistent optimizer and loss function configurations to ensure a fair comparison. Model performance was evaluated using standard classification metrics, namely accuracy, precision, recall, F1-score, and the area under the curve (AUC), to quantify the models' capacity to discriminate between classes. All experiments were conducted on a workstation equipped with an AMD Ryzen 5 4600H six-core processor, NVIDIA GeForce GTX 1650 GPU with 4GB VRAM, and 16GB DDR4 RAM. This hardware configuration is maintained consistently across all experiments to ensure reproducibility and equitable computational benchmarking. The methodological framework is depicted in Fig. 1.

A. Dataset

This study utilizes a dataset of kidney computed tomography (CT) images derived from the public *CT Kidney Dataset: Normal-Cyst-Tumor and Stone*, available on the Kaggle platform. The dataset was originally collected from the Picture Archiving and Communication System (PACS) of multiple hospitals in Dhaka, Bangladesh. The CT studies were carefully selected and subsequently verified by a radiologist and a medical technologist to ensure diagnostic correctness before inclusion in the dataset. The dataset comprises abdominal CT images centered on the renal region, and it has been annotated into four pathological categories: Cyst, Normal, Stone, and Tumor.

In total, the dataset comprises 12,446 unique CT images, and the class distribution is not entirely balanced. The Normal class contains 5,077 images, the Cyst class 3,709 images, and the Tumor class 2,283 images. In contrast, the Stone class contains the fewest images, with 1,377 images. This imbalance, which mirrors real-world medical data, can bias classification models, particularly when identifying minority classes [27]. Therefore, appropriate evaluation strategies and model design considerations are necessary to ensure that classification performance remains reliable across both majority and minority classes

As the dataset is publicly available and provided at the image level, the experimental evaluation conducted in this study is intended to assess model behavior and comparative robustness under controlled conditions, rather than to establish definitive clinical diagnostic performance.

Although the dataset was collected from multiple hospitals and reflects real-world clinical imaging practice, it does not include demographic metadata such as patient age. Therefore, its representativeness for geriatric-specific CT imaging cannot be explicitly confirmed, and age-specific generalization remains a limitation of this study. To illustrate the structure and characteristics of the employed data, a sample dataset is presented in Fig. 2. The figure shows sample CT images along with their corresponding class labels indicating the kidney conditions. This example illustrates the relationship between the input images and the clinical interpretation of these labels in the context of kidney pathology.

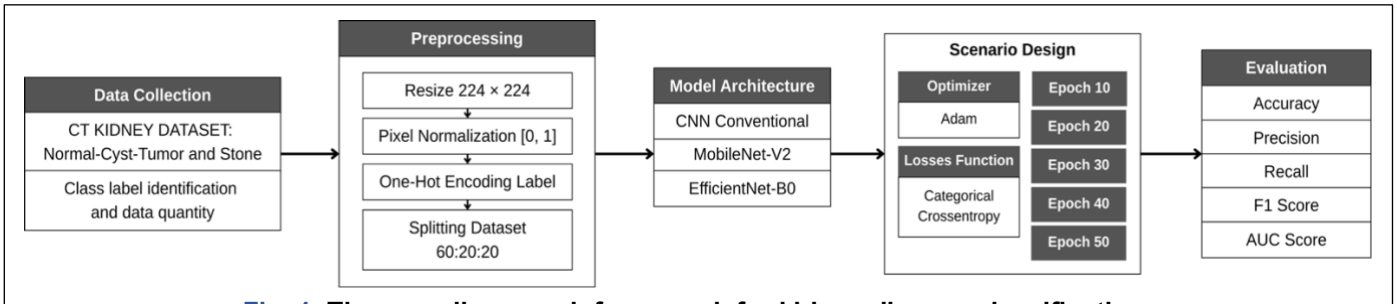


Fig. 1. The overall research framework for kidney disease classification.

Corresponding author: Aji Prasetya Wibawa, aji.prasetya.ft@um.ac.id, Department of Electrical Engineering and Informatics, Universitas Negeri Malang, Malang, Indonesia.
DOI: https://doi.org/10.35882/ijeeemi.v8i2.325

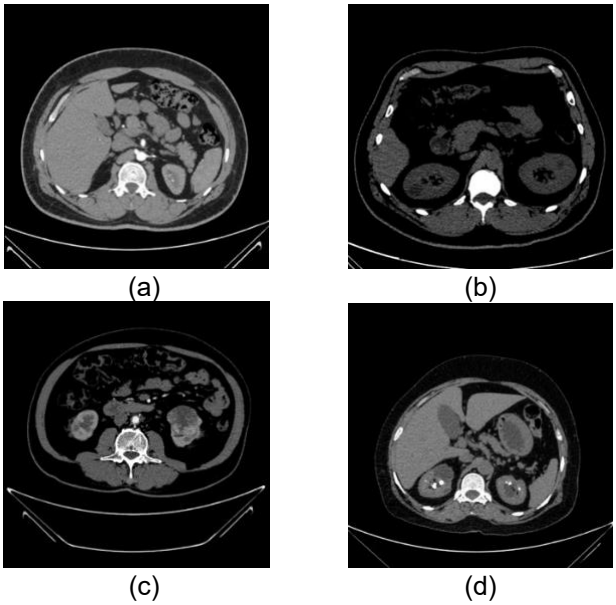


Fig. 2 Example of CT kidney image: (a) Normal, (b) Cyst, (c) Tumor, (d) Stone.

B. Data Preprocessing

The image preprocessing stage was conducted to ensure uniformity in the input format across all evaluated architectures and to maintain experimental consistency for baseline comparison. All CT scan images were resized to a fixed spatial resolution of 224×224 pixels using the $\mathrm{cv2.resize()}$ function from OpenCV. This dimension was selected because it serves as the standard input for most deep learning architectures [28], including MobileNet-V2 and EfficientNet-B0, while balancing visual fidelity and computational efficiency.

In addition to resizing, the pixel values were normalized by dividing each intensity value by 255.0, thereby constraining them to the range [0, 1]. This normalization aims to reduce scale differences between features, accelerate convergence during training, and enhance the numerical stability of the model [29]. The kidney disease category labels were converted to one-hot-encoded vectors to enable multiclass classification with a Softmax output layer. No data augmentation techniques were applied during preprocessing. This decision was intentionally made to preserve consistency across all evaluated models and to ensure a fair baseline comparison of architectural efficiency without introducing additional variability from augmentation strategies.

The preprocessed dataset was partitioned at the image level using the $\mathrm{train\_test\_split}$ function from the Scikit-learn library with the parameter $\mathrm{shuffle} = \mathrm{True}$. The dataset was first split into training and test sets with a 20% test split. The remaining 80% was subsequently split again to obtain a validation subset with a proportion equivalent to 20% of the total dataset. This procedure yielded a final distribution of 60:20:20, corresponding to 7,467 training images, 2,489 validation images, and 2,490 testing images. The dataset provides only image-level metadata, without patient identifiers. Therefore, patient-level grouping during partitioning was not feasible.

C. Model Architecture

This study implements three deep neural network architectures for kidney CT image classification: a conventional Sequential CNN, MobileNet-V2, and EfficientNet-B0. The overall architectural configurations and the structural differences among the three models are illustrated in Fig. 3. The figure provides a visual representation of the feature extraction pipelines, pooling mechanisms, and classification heads employed in each architecture, thereby clarifying the comparative design strategy adopted in this study.

The first architecture is a conventional convolutional neural network (CNN) arranged sequentially to progressively extract hierarchical spatial features from CT images of size 224×224. The convolution operation on layer l is mathematically defined in Eq. (1) [30].

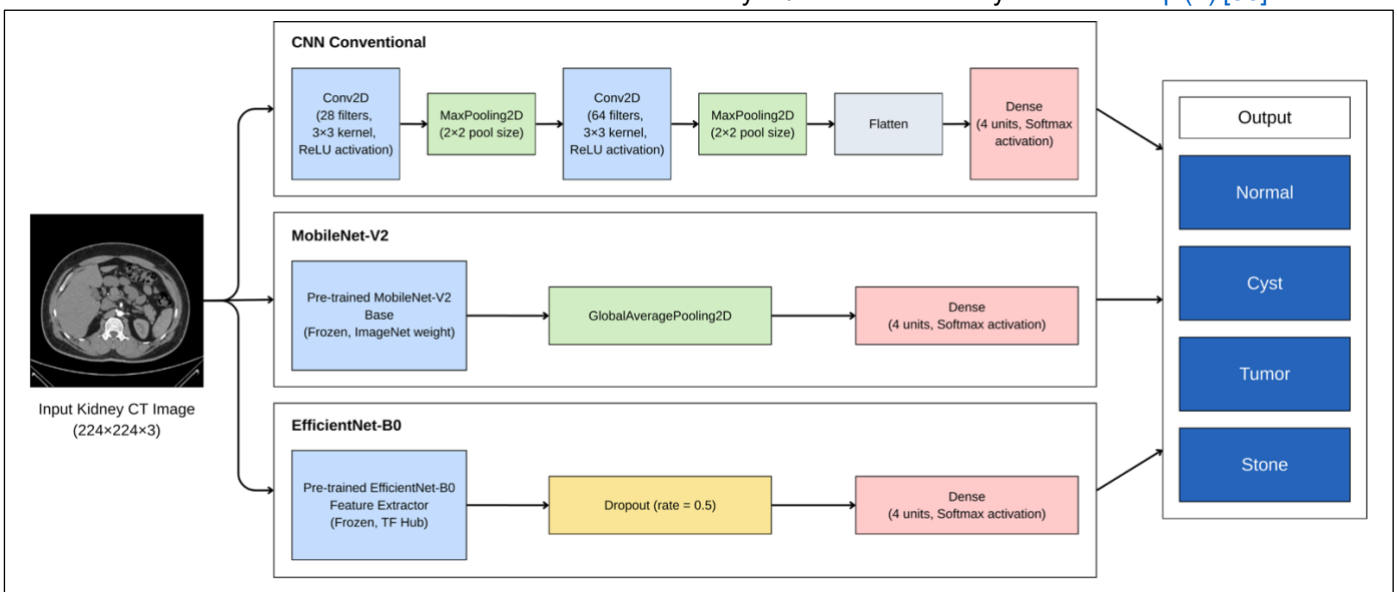


Fig. 3 Architecture Diagram that visually represents each model architecture.

Corresponding author: Aji Prasetya Wibawa, aji.prasetya.ft@um.ac.id, Department of Electrical Engineering and Informatics, Universitas Negeri Malang, Malang, Indonesia.
DOI: https://doi.org/10.35882/ijeeemi.v8i2.325

$$X_{i,j,k}^{(l)} = \sigma\left(\sum_{m=1}^{C_{l-1}}\sum_{u=1}^{K}\sum_{v=1}^{K} W_{u,v,m,k}^{(l)} \cdot X_{i+u,j+v,m}^{(l-1)} + b_k^{(l)}\right) \quad (1)$$

where $X_{i,j,k}^{(l)}$ denotes the activation value at spatial position (i,j) of the k-th feature map in the l-th convolutional layer. The term $X_{i+u,j+v,m}^{(l-1)}$ represents the input feature value from the $(l-1)$-th layer at channel m within the receptive field. $W_{u,v,m,k}^{(l)}$ corresponds to the convolutional kernel weight connecting the m-th input channel to the k-th output feature map at layer l, while $b_k^{(l)}$ denotes the bias term associated with the k-th filter. The indices u and v define the spatial dimensions of the convolution kernel with size $K \times K$, and C_{l-1} indicates the number of input channels from the previous layer. Finally, $\sigma(\cdot)$ represents the nonlinear activation function applied after the convolution operation.

In the first convolutional layer, $K = 3$ and the number of filters $k = 28$, producing 28 feature maps as shown in Fig. 3. The selection of 28 filters in the first layer is intentionally adopted as a manual parameter-reduction strategy to control early-stage redundancy and reduce computational complexity before feature expansion in deeper layers. The second convolutional layer applies $K = 3$ with 64 filters ($k = 64$), increasing representational capacity for deeper feature abstraction. Spatial downsampling is performed using max pooling defined in Eq. (2) [31].

$$P_{i,j,k} = \max_{(u,v)\in\Omega} X_{i+u,j+v,k} \quad (2)$$

where $P_{i,j,k}$ represents the pooled output value at spatial location (i,j) for the k-th feature map. The variable $X_{i+u,j+v,k}$ denotes the input activation within the local pooling region centered around the spatial position. The indices u and v define the coordinates within the pooling window. At the same time, Ω denotes the set of spatial locations within the pooling region. After the second pooling layer, the feature maps are flattened into a 1D vector of 186,624 units and passed to a Dense layer with 4 output neurons, using Softmax activation for multiclass classification. Notably, no dropout regularization is applied in the conventional CNN, allowing the model to retain full representational capacity during training [32].

The second architecture is MobileNet-V2, which uses transfer learning. As illustrated in Fig. 3, the model consists of a pre-trained MobileNet-V2 base network initialized with ImageNet weights [33], followed by a GlobalAveragePooling2D layer and a Dense classification layer with four Softmax units. All convolutional layers in the MobileNet-V2 base are frozen ($\mathrm{trainable} = \mathrm{False}$) to preserve the learned depthwise separable convolution representations. The GlobalAveragePooling2D layer converts the final feature maps into a 1D feature vector while reducing the number of parameters and mitigating overfitting compared to a flattening operation [34]. The final Dense layer with four neurons corresponds to the four kidney CT classes. It is the only trainable component in the model. This design reduces training complexity while effectively leveraging robust feature representations learned from large-scale pre-training.

The third architecture employs EfficientNet-B0 as a pre-trained feature extractor. As shown in Fig. 3, the model uses the EfficientNet-B0 feature vector module from TensorFlow Hub, which directly produces a 1280-dimensional feature representation from a $224 \times 224 \times 3$ input image. EfficientNet-B0 is designed using compound scaling principles that balance network depth, width, and input resolution. Since spatial aggregation is performed internally by the feature-vector module, no explicit GlobalAveragePooling2D layer is included in the Sequential configuration. A Dropout layer with a rate of 0.5 is applied after feature extraction to reduce feature co-adaptation and improve generalization. Finally, a Dense layer with four Softmax neurons maps the extracted features to the kidney CT categories.

D. Scenario Design

All training experiments were executed using the Adam (*Adaptive Moment Estimation*) optimizer. This optimizer was selected for its ability to adjust the learning rate for each parameter adaptively. This property is highly effective for handling data with RSE or noisy gradients [35]. In this study, the Adam optimizer was implemented with the default TensorFlow configuration, using $\eta = 0.001$, $\beta 1 = 0.9$, $\beta 2 = 0.999$, and $\varepsilon = 10^{-7}$. This optimization procedure is specified by Eq. (3) [36].

$$w_{t+1} = w_t - \frac{\eta}{\sqrt{\hat{v}_t} + \epsilon}\hat{m}_t \quad (3)$$

where w_t denotes the model parameter vector at iteration t, and w_{t+1} represents the updated parameter after the optimization step. The term η refers to the learning rate that controls the step size during parameter updates. The variables $\hat{m}_t$ and $\hat{v}_t$ correspond to the bias-corrected first moment (mean) and second moment (uncentered variance) estimates of the gradients, respectively. The parameter ϵ is a small constant added for numerical stability to prevent division by zero during the update process.

The optimization procedure also employs the Categorical Crossentropy loss function, which is well-suited to multiclass classification with one-hot-ended labels [37]. Eq. (4) defines this loss, which quantifies the divergence between the model's predicted probability distribution $(\hat{y})$ and the true label distribution (y) [38].

$$L_{CE} = -\sum_{i=1}^{C} yi \cdot log\,(\hat{y}i) \quad (4)$$

where L_{CE} denotes the categorical cross-entropy loss value. The variable C represents the total number of classes in the classification problem. The term yi corresponds to the true label indicator for class i, which is encoded using a one-hot representation. The predicted probability for class i produced by the Softmax output layer is denoted by $\hat{y}i$. The logarithmic function $log(\cdot)$ measures the negative log-likelihood of the predicted probability for the correct class, thereby penalizing incorrect predictions and encouraging accurate probabilistic classification.

Corresponding author: Aji Prasetya Wibawa, aji.prasetya.ft@um.ac.id, Department of Electrical Engineering and Informatics, Universitas Negeri Malang, Malang, Indonesia.
DOI: https://doi.org/10.35882/ijeeemi.v8i2.325

Categorical cross-entropy was selected because the classification task involves four mutually exclusive classes with one-hot-encoded labels, which is mathematically consistent with Softmax output activation. Specifically, the Softmax function produces a normalized probability distribution over the four classes (\sum$_i$ ŷ$_i$ = 1), and Categorical Crossentropy directly measures the negative log-likelihood of the correct class, thereby encouraging probabilistic calibration of predictions. Although the dataset exhibits class imbalance, no class weighting was applied to maintain consistent comparative evaluation across architectures. This design choice ensures that performance differences reflect architectural behavior rather than loss reweighting strategies. Model robustness under imbalance conditions was instead assessed through weighted Precision, Recall, and F1-score metrics. To analyze model stability and learning dynamics, the experiment was designed with five distinct training duration conditions: 10, 20, 30, 40, and 50 epochs. All training conditions were applied consistently across the three model architectures, without employing early stopping or a learning rate scheduler. This range of epoch counts enables the identification of the model's convergence point, beyond which accuracy improvement slows or ceases (saturation). It also facilitates the detection of tendencies toward overfitting or underfitting [39]. An overview of all experimental configurations is provided in Table 1.

Table 1. Experimental design scenarios for analyzing model stability across varying training durations.

Architecture	Configuration	Epoch
CNN Conventional	Adam, Categorical Crossentropy	10, 20, 30, 40, 50
MobileNet-V2	Adam, Categorical Crossentropy	10, 20, 30, 40, 50
EfficientNet-B0	Adam, Categorical Crossentropy	10, 20, 30, 40, 50

E. Evaluation

The evaluation framework in this study is designed to analyze not only classification performance but also model learning characteristics, convergence stability, and error behavior across different architectural designs. The objective is to compare the robustness and feasibility of deep learning models for kidney disease classification in a computational and methodological context, rather than to claim final clinical diagnostic accuracy. The model's performance was evaluated using several quantitative metrics: accuracy, precision, recall, F1-score, and Area Under the Curve (AUC). These metrics are derived from the confusion matrix, which summarizes correct and incorrect predictions for each class in a multiclass setting. True Positive (TP) denotes positive instances correctly classified, True Negative (TN) denotes negative instances correctly identified, False Positive (FP) occurs when a negative instance is misclassified as positive, and False Negative (FN) occurs when a positive instance is misclassified as negative, as used in the evaluation metrics defined in Eq. (5)–Eq. (8) [40]. Together, these evaluation metrics provide a comprehensive basis for assessing both the predictive capability and the learning behavior of the models within the proposed evaluation framework.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

$$Precision = \frac{TP}{TP + FP} \quad (6)$$

$$Recall = \frac{TP}{TP + FN} \quad (7)$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (8)$$

Eq. (5) to Eq. (8) define the following calculations. Accuracy quantifies the proportion of correct predictions relative to the entire test sample. Precision quantifies the proportion of correctly predicted positive instances among all predicted positives. Recall (Sensitivity) quantifies the proportion of actual positive instances that are correctly identified. F1-Score provides a balanced summary when a trade-off between Precision and Recall is present. These four metrics are particularly important given the class imbalance in the employed dataset. Additionally, the AUC score is used to quantify the model's ability to discriminate between classes, as shown in the Receiver Operating Characteristic (ROC) curve [41]. An AUC value near 1 indicates excellent discriminatory ability, whereas a value near 0.5 indicates performance equivalent to random guessing.

In addition to numerical metrics, learning curve analysis and confusion matrix visualization are employed to identify potential overfitting, premature convergence, and class-specific misclassification patterns, providing a more comprehensive assessment of model reliability. Furthermore, to assess whether the observed performance differences between model architectures are statistically significant, a paired t-test was conducted using accuracy values obtained across epoch configurations (10-50 epochs). Each epoch configuration was treated as a paired observation between models. The statistical significance level was set at $\alpha = 0.05$.

III. Results

This section presents the results of evaluating the performance of three deep learning architectures employed in the study, namely a conventional CNN, MobileNet-V2, and EfficientNet-B0, for the task of classifying kidney CT images into four classes (Normal, Cyst, Stone, and Tumor). The evaluation was conducted across multiple training epoch configurations, ranging from 10 to 50 epochs. Model performance was quantified using standard metrics, including accuracy, precision, recall, F1-score, and AUC score.

A. Model Performances

Corresponding author: Aji Prasetya Wibawa, aji.prasetya.ft@um.ac.id, Department of Electrical Engineering and Informatics, Universitas Negeri Malang, Malang, Indonesia.
DOI: https://doi.org/10.35882/ijeeemi.v8i2.325

The quantitative performance of each architecture is summarized in Table 2 for CNN, Table 3 for MobileNet-V2, and Table 4 for EfficientNet-B0. The tables report metric values obtained at different epoch configurations, providing a comparative overview of classification performance across the three models.

Table 2. Evaluation results for the CNN architecture scenario.

	Acc	Pre	Rec	F1-Sc	AUC
Epoch 10	1.00	1.00	1.00	1.00	1.00
Epoch 20	1.00	1.00	1.00	1.00	1.00
Epoch 30	1.00	1.00	1.00	1.00	1.00
Epoch 40	1.00	1.00	1.00	1.00	1.00
Epoch 50	1.00	1.00	1.00	1.00	1.00

As shown in Table 2, CNN exhibits identical performance metrics across all evaluated epoch configurations. From epoch 10 to epoch 50, the model attains accuracy, precision, recall, and F1-score values of 1.00. The AUC remains constant at 1.00 across all training scenarios. The observed consistency across these metrics indicates that the CNN's classification performance remains unchanged as the number of epochs increases. No numerical variation is observed in any of the evaluation metrics, whether at the beginning or at the conclusion of training. Thus, based on the quantitative metrics analyzed, the CNN achieves the highest classification performance across all epoch configurations evaluated in this study.

Table 3. Evaluation results for the MobileNet-V2 architecture scenario.

	Acc	Pre	Rec	F1-Sc	AUC
Epoch 10	0.98	0.99	0.99	0.99	0.9991
Epoch 20	0.99	0.99	0.99	0.99	0.9997
Epoch 30	0.99	0.99	0.99	0.99	0.9998
Epoch 40	0.99	1.00	1.00	1.00	0.9998
Epoch 50	1.00	1.00	1.00	1.00	0.9997

The performance evaluation of the MobileNet-V2 architecture is summarized in Table 3. At epoch 10, MobileNet-V2 attained an accuracy of 0.98, with precision, recall, and F1-score values of 0.99, and an AUC value of 0.9991. As the number of epochs increased, the accuracy reached 0.99 at 20 and 30 epochs. In contrast, the precision, recall, and F1-score remained in the range of 0.99. At 40 epochs, the precision, recall, and F1-score values reached 1.00, while the accuracy remained at 0.99, with an AUC of 0.9998. At epoch 50, the model attained an accuracy of 1.00, with precision, recall, and F1-score also reaching 1.00. The AUC at this epoch was recorded as 0.9997. Overall, the results in Table 3 demonstrate a gradual improvement in MobileNet-V2 performance as the number of epochs increases, with evaluation metrics approaching or attaining their maximum values at the highest training epoch.

Table 4. Evaluation results for the EfficientNet-B0 architecture scenario.

	Acc	Pre	Rec	F1-Sc	AUC
Epoch 10	0.97	0.97	0.97	0.97	0.9948
Epoch 20	0.97	0.97	0.97	0.97	0.9979
Epoch 30	0.98	0.98	0.98	0.98	0.9988
Epoch 40	0.98	0.98	0.98	0.98	0.9988
Epoch 50	0.98	0.99	0.99	0.98	0.9992

As depicted in Table 4, at epoch 10, the EfficientNet-B0 architecture attained an accuracy of 0.97, with precision, recall, and F1-score each equal to 0.97, and an AUC value of 0.9948. At epoch 20, the accuracy and F1-score remained at 0.97, whereas the AUC increased to 0.9979. Further improvements were observed at epoch 30, when the accuracy, precision, recall, and F1-score increased to 0.98, accompanied by an AUC of 0.9988. These metrics remained stable until epoch 40. At epoch 50, precision and recall increased to 0.99, while accuracy and F1-score remained at 0.98; the AUC value in this configuration reached 0.9992. Collectively, these results demonstrate that EfficientNet-B0 exhibits relatively stable performance across all epoch configurations, with gradual improvements in metrics and consistently high AUC values throughout training.

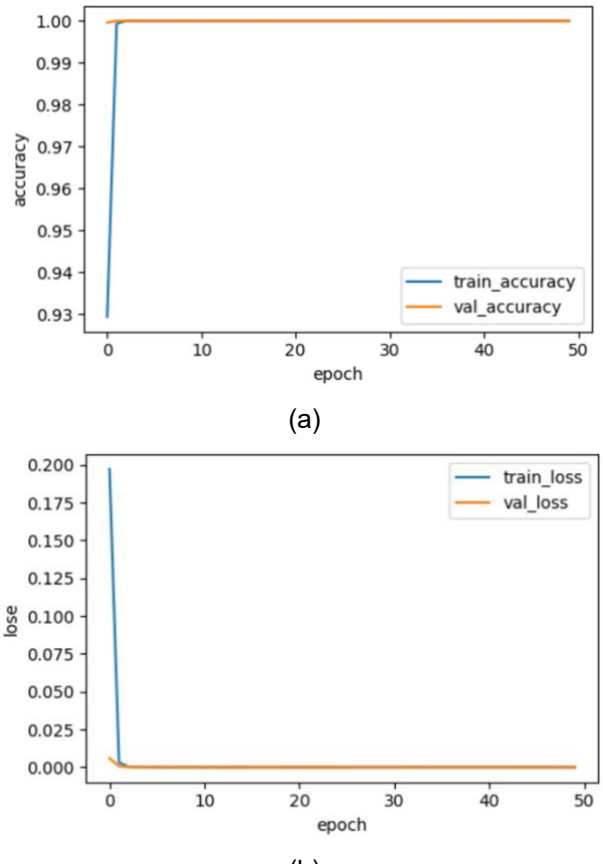
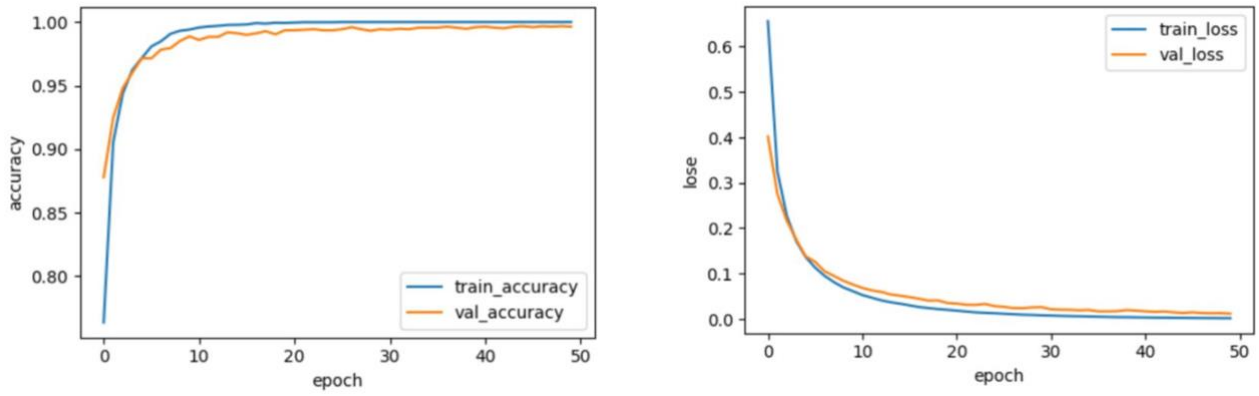


Fig. 4 Learning curve of the CNN architecture: (a) accuracy curves, (b) loss curves.

Corresponding author: Aji Prasetya Wibawa, aji.prasetya.ft@um.ac.id, Department of Electrical Engineering and Informatics, Universitas Negeri Malang, Malang, Indonesia.
DOI: https://doi.org/10.35882/ijeeemi.v8i2.325



(a) (b)

Fig. 5 Learning curve of the MobileNet-V2 architecture: (a) accuracy curves, (b) loss curves.

B. Learning Curves

Learning curves are presented to analyze training dynamics over 50 epochs by examining the evolution of accuracy and loss on both the training and validation sets under identical experimental conditions. Fig. 4, Fig. 5, and Fig. 6 illustrate the comparative learning behavior of the CNN, MobileNet-V2, and EfficientNet-B0 architectures, highlighting their convergence patterns and generalization characteristics throughout training. Fig. 4 illustrates the learning curve of the CNN architecture, showing accuracy and loss values for the training and validation datasets over 50 epochs. At the initial epoch, the training accuracy begins at approximately 0.93 and rapidly increases to values close to 1.00 within the first few epochs. The validation accuracy follows a similar pattern, reaching values near 1.00 at an early stage of training. After this point, both training and validation accuracy curves remain constant and overlap closely until the final epoch.

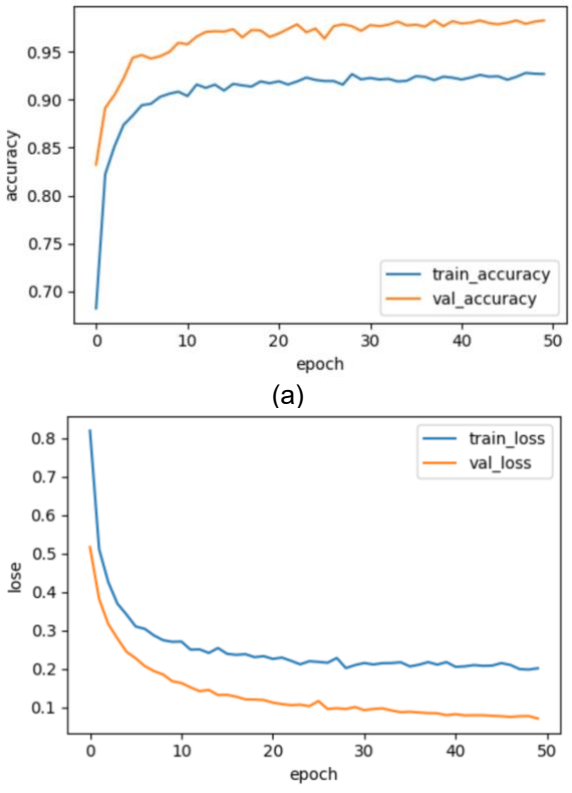
The loss curves presented in Fig. 4 show a sharp decline during the early epochs. The training loss decreases from an initial value of approximately 0.20 to a value near zero within the first few epochs. Similarly, the validation loss rapidly converges toward zero and remains stable throughout the remainder of the training process. After the initial convergence phase, no significant fluctuations or divergences are observed between the training and validation loss curves.

The near-perfect overlap between training and validation curves from the early epochs indicates immediate saturation of the optimization process. The absence of a measurable generalization gap across 50 epochs suggests that performance stabilizes at its maximum value shortly after training begins.

Fig. 5 presents the learning curves of the MobileNet-V2 architecture, showing the progression of training and validation accuracy and loss values over 50 epochs. At the initial epoch, the training accuracy is approximately 0.7, while the validation accuracy is around 0.88. Both accuracy curves increase progressively during the early epochs, reaching values above 0.95 within the first 5 to 10 epochs.

As training continues, the training and validation accuracy curves gradually approach 1.00. Minor fluctuations are observed in the validation accuracy curve during the mid-epochs of training. However, the overall trend remains stable, with both curves converging toward similar values by the final epochs.

The corresponding loss curves show a steady decrease throughout training. The training loss decreases from an initial value of approximately 0.65 to values close to zero by the later epochs. The validation loss follows a similar downward trend, decreasing consistently and remaining slightly higher than the training loss during most epochs. Both loss curves approach minimal values toward the end of training without abrupt changes.



(a)

(b)

Fig. 6 Learning curve of the EfficientNet-B0 architecture: (a) accuracy curves, (b) loss curves.

Corresponding author: Aji Prasetya Wibawa, aji.prasetya.ft@um.ac.id, Department of Electrical Engineering and Informatics, Universitas Negeri Malang, Malang, Indonesia.
DOI: https://doi.org/10.35882/ijeeemi.v8i2.325

Fig. 6 shows the learning curves for the EfficientNet-B0 architecture over 50 training epochs. At the initial epoch, the training accuracy is approximately 0.68. It increases progressively through the early training stages, exceeding 0.90 within the first 10 epochs. The validation accuracy follows a similar upward trend and stabilizes around 0.97-0.98 in the later epochs. Throughout the training process, the validation accuracy consistently exceeds the training accuracy, resulting in a persistent performance gap between the two curves. The corresponding loss curves demonstrate a steady decrease across epochs. The training loss declines from an initial value above 0.80 to approximately 0.20 by the final epochs. In contrast, the validation loss decreases from around 0.50 to below 0.10. Notably, the validation loss remains consistently lower than the training loss across all epochs. Both curves exhibit smooth convergence without abrupt oscillations.

Overall, the CNN learning curve demonstrates rapid convergence with early saturation, as both accuracy and loss reach asymptotic values within the initial epochs and remain stable thereafter. Meanwhile, the MobileNet-V2 learning curve shows a gradual increase in accuracy and continuous reduction in loss, with training and validation curves remaining closely aligned and parallel throughout the training process. EfficientNet-B0 exhibits progressive convergence without early saturation, maintaining smoothly decreasing loss curves while consistently presenting a persistent gap in which validation accuracy remains higher and validation loss lower than the corresponding training curves across epochs. Collectively, these learning dynamics serve as empirical indicators of model generalization behavior, as reflected by the degree of convergence stability, curve alignment, and the presence or absence of performance divergence under identical experimental settings.

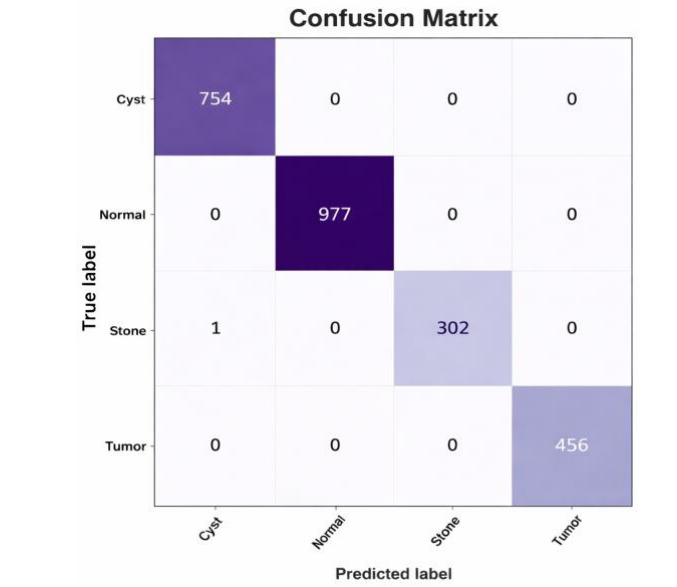


Fig. 7 Confusion matrix of the CNN architecture at 50 epochs.

C. Confusion Matrix

The confusion matrices provide a class-level evaluation of model performance by detailing the distribution of correct and incorrect predictions across the four kidney disease categories. Unlike aggregated metrics, this representation reveals inter-class misclassification patterns and highlights class-specific predictive behavior. Fig. 7, Fig. 8, and Fig. 9 show the confusion matrices for the CNN, MobileNet-V2, and EfficientNet-B0 architectures, respectively, each evaluated at 50 epochs, corresponding to their optimal performance configurations. Fig. 7 presents the confusion matrix of the CNN architecture evaluated at 50 epochs. The model correctly classifies all 754 Cyst images, 977 Normal images, and 456 Tumor images, with no observed misclassifications. In the Stone class, 302 of 303 samples are correctly predicted, with 1 misclassified as Cyst. Overall, only one misclassification is observed, occurring between the Stone and Cyst categories.

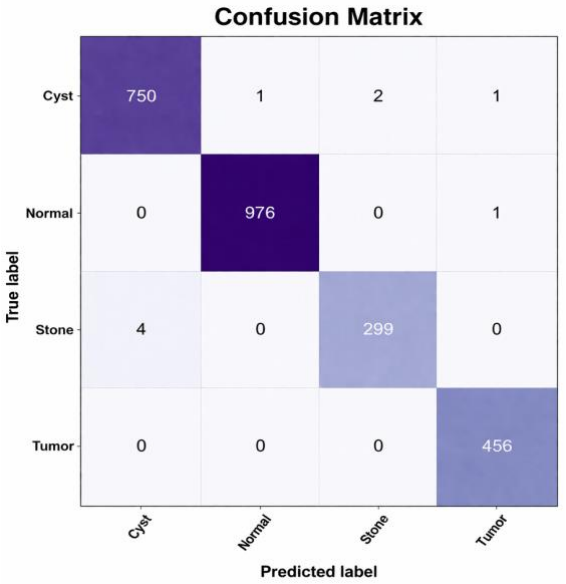


Fig. 8 Confusion matrix of the MobileNet-V2 architecture at 50 epochs.

Fig. 8 presents the confusion matrix of the MobileNet-V2 architecture evaluated at 50 epochs. In the Cyst class, 750 of 754 images are correctly classified, with 1 misclassified as Normal, 2 as Stone, and 1 as Tumor. For the Normal class, 976 images are correctly predicted, with a single misclassification into the Tumor category. In the Stone class, 299 of 303 samples are correctly classified, while 4 are misclassified as Cyst. All 456 Tumor images are correctly identified, with no observed misclassifications. Overall, misclassification events are limited and primarily occur between the Cyst and Stone categories. Fig. 9 presents the confusion matrix of the EfficientNet-B0 architecture evaluated at 50 epochs. In the Cyst class, 752 out of 754 images are correctly classified, with two misclassified as Normal. In the Normal class, 976 images are correctly classified, with 1 misclassification into the Tumor class. In the Tumor class, 446 out of 456 images are correctly identified, with four misclassified as Cyst and six as Normal. The Stone class correctly classifies 279 of 303 samples, while 10 are

Corresponding author: Aji Prasetya Wibawa, aji.prasetya.ft@um.ac.id, Department of Electrical Engineering and Informatics, Universitas Negeri Malang, Malang, Indonesia.
DOI: https://doi.org/10.35882/ijeeemi.v8i2.325

misclassified as Cyst, 10 as Normal, and 4 as Tumor. Misclassification is distributed across multiple class pairs, with the highest frequency observed in the Stone class.

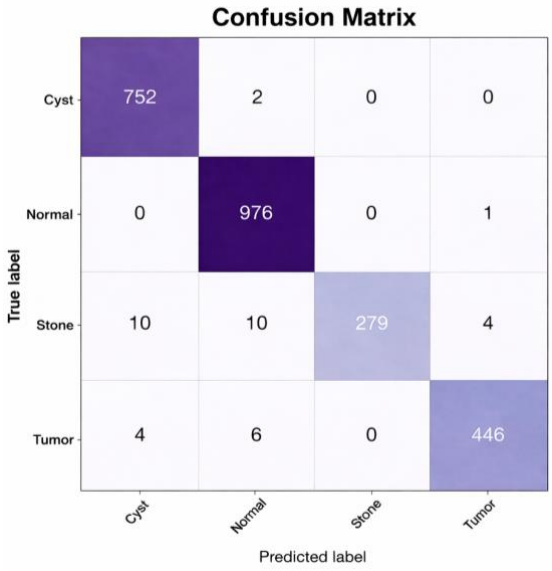


Fig. 9 Confusion matrix of the EfficientNet-B0 architecture at 50 epochs.

Overall, the confusion matrices reveal distinct inter-class misclassification patterns across the evaluated architectures. In all models, misclassification is predominantly observed between the Cyst and Stone categories. The EfficientNet-B0 architecture exhibits the highest error dispersion, particularly in the Stone class. Notably, the Stone category, which represents the smallest test subset, shows the highest number of misclassifications in this architecture. In contrast, the CNN architecture demonstrates only a single misclassification, whereas MobileNet-V2 shows limited errors, primarily between Cyst and Stone.

D. Comparative Performance Summary

To provide an overall comparison of model performance, the best-performing configuration for each evaluated architecture is summarized in Table 5. The table presents the optimal epoch setting and corresponding evaluation metrics for CNN, MobileNet-V2, and EfficientNet-B0, allowing direct comparison of their quantitative classification performance under their respective best conditions. In addition to classification metrics, computational characteristics, including total parameters, trainable parameters, model size, training time, and testing time, are incorporated to enable a comprehensive assessment of efficiency and performance trade-offs. Based on the results shown in Table 5, the CNN model achieves perfect values across all reported metrics, including accuracy, precision, recall, F1-score, and AUC, with these results obtained consistently across epochs 10 to 50. MobileNet-V2 also reaches perfect values for accuracy, precision, recall, and F1-score at epoch 50. In contrast, its AUC value remains very close to unity at 0.9998. EfficientNet-B0 attains its best performance at epoch 50, achieving an accuracy of 0.98, precision and recall values of 0.99, an F1-score of 0.98, and an AUC value of 0.9992.

Table 5. Comparison of the best-performing models across different architectures.

	CNN	MobileNet-V2	EfficientNet-B0
Epoch	10-50	50	50
Acc	1.00	1.00	0.98
Pre	1.00	1.00	0.99
Rec	1.00	1.00	0.99
F1-Sc	1.00	1.00	0.98
AUC	1.0000	0.9998	0.9992
Total Parameters	763,476	2,263,108	4,054,688
Trainable Parameters	763,476 (100%)	5,124 (0.23%)	5,124 (0.13%)
Model Size (MB)	2.91 MB	8.63 MB	15.47 MB
Training Time (50 epochs)	4,270 s	4,684 s	6,538 s
Testing Time	≈10-11 s	≈30 s	≈37 s

From a computational perspective, the CNN exhibits the lowest total parameter count (763,476) and the shortest testing time (approximately 10-11 seconds). However, all parameters in the CNN model are trainable, meaning that 100% of the network weights are updated during training. In contrast, MobileNet-V2 contains 2,263,108 total parameters, yet only 5,124 parameters (0.23%) are trainable under the transfer learning configuration. Similarly, EfficientNet-B0 contains 4,054,688 total parameters, with only 5,124 trainable (0.13%). This indicates that both transfer learning-based architectures rely primarily on pre-trained feature representations, while updating only a very small fraction of parameters during fine-tuning.

Although CNNs demonstrate shorter training and inference times in absolute terms, MobileNet-V2 achieves equivalent peak classification performance while requiring substantially fewer trainable parameters. Compared to EfficientNet-B0, MobileNet-V2 also achieves higher accuracy (1.00 vs. 0.98), lower model size (8.63 MB vs. 15.47 MB), shorter training time (4,684 s vs. 6,538 s at 50 epochs), and reduced inference time (≈30 s vs. ≈37 s).

Therefore, when considering both classification metrics and computational characteristics, CNN and MobileNet-V2 achieve equivalent peak performance across most evaluation metrics, whereas EfficientNet-B0 exhibits slightly lower accuracy and F1 scores. Although CNN demonstrates the shortest training and testing times, MobileNet-V2 achieves comparable predictive performance while requiring only 0.23% of its total parameters to be optimized under the transfer learning configuration. The AUC values for all architectures remain

Corresponding author: Aji Prasetya Wibawa, aji.prasetya.ft@um.ac.id, Department of Electrical Engineering and Informatics, Universitas Negeri Malang, Malang, Indonesia.
DOI: https://doi.org/10.35882/ijeeemi.v8i2.325

consistently high, indicating strong discrimination across the evaluated models.

E. Statistical Significance Analysis

To further examine whether the observed performance differences among the evaluated architectures are statistically significant, a paired t-test was conducted on the accuracy values obtained across training epochs (10–50). The results of the statistical comparison are summarized in Table 6.

Table 6. Paired t-test Results for Model Accuracy Comparison.

Model Comparison	t-value	p-value	Significance
CNN vs MobileNet-V2	3.1623	0.0341	Significant
CNN vs EfficientNet-B0	9.7980	0.0006	Extremely Significant
MobileNet-V2 vs EfficientNet-B0	5.7155	0.0046	Very Significant

The statistical analysis indicates that all pairwise model comparisons yield p-values below the 0.05 significance level. The difference between CNN and MobileNet-V2 is statistically significant ($p = 0.0341$), suggesting that CNN achieves a higher mean accuracy across epochs. A stronger level of significance is observed between CNN and EfficientNet-B0 ($p = 0.0006$), indicating a highly significant difference in performance. Similarly, the comparison between MobileNet-V2 and EfficientNet-B0 shows a statistically significant difference ($p = 0.0046$), confirming that MobileNet-V2 outperforms EfficientNet-B0 in terms of mean accuracy across the evaluated epochs. Overall, these findings provide inferential support for the earlier comparative results and strengthen the conclusion regarding the relative suitability of the evaluated architectures for multiclass kidney CT image classification.

IV. Discussion

The findings of this study indicate that high classification performance in CT-based kidney disease detection does not inherently reflect robust generalization capability. Although all evaluated models achieve near-perfect metrics, their learning behaviors differ substantially, as reflected in the contrast between immediate saturation in CNN, progressive convergence in MobileNet-V2, and conservative stabilization in EfficientNet-B0. This distinction is critical in medical image analysis, where models must generalize across heterogeneous patient data rather than optimize dataset-specific patterns. The evaluation framework employed in this study enables this distinction by analyzing multi-epoch consistency, training–validation alignment, and class-level error distribution, thereby providing a more reliable basis for assessing model suitability in clinical decision support systems.

As shown in Table 2, the conventional CNN model reaches identical performance results across all evaluated epoch configurations, with accuracy, precision, recall, F1-score, and AUC consistently reaching 1.00 from epoch 10 to 50. This lack of variance in classification performance indicates immediate metric saturation, suggesting that the model's optimization process plateaus shortly after training commences. The learning curve in Fig. 4 further supports this observation, as accuracy rapidly increases from 0.93 to 1.00 and training loss sharply declines from 0.20 to near zero within the initial epochs, indicating that feature representation is established early. Although the confusion matrix in Fig. 7 reveals a near-perfect error rate of approximately 0.04%, resulting from a single misclassification in the Stone class, this high level of fitting should be interpreted with caution. The absence of patient-level data separation in the current dataset introduces a potential source of data leakage, whereby the model may learn patient-specific visual patterns rather than generalized disease features [42]. Therefore, the perfect AUC value of 1.00 is likely influenced by evaluation constraints and dataset characteristics, and should not be regarded as a definitive indicator of clinical readiness [43].

As presented in Table 3, MobileNet-V2 achieves an incremental increase in accuracy from 0.98 at epoch 10 to 1.00 at epoch 50, representing an absolute gain of 0.02. This non-zero variance indicates progressive refinement of feature representations, suggesting that the model avoids premature metric saturation by continuously optimizing its weights [44]. The learning curve in Fig. 5 further supports this trend; the smooth convergence of validation accuracy from 0.88 to nearly 1.00, accompanied by a steadily decreasing loss, demonstrates controlled optimization without divergence. The confusion matrix in Fig. 8 confirms this learning pattern [45], reporting a low error rate of 0.36% (9 misclassifications), primarily between the Cyst and Stone classes. This misclassification pattern corresponds to the visual similarities in texture and intensity observed in renal CT imaging, suggesting that the model's errors are attributable to clinical imaging characteristics rather than random algorithmic failure. This structured misclassification indicates that MobileNet-V2 achieves controlled convergence, with feature representations refined progressively across epochs [46]. This behavior is attributed to the architectural design of MobileNet-V2, which employs depthwise separable convolutions and frozen pretrained layers to reduce parameter redundancy, thereby enabling gradual optimization while maintaining stable generalization performance [47].

Based on Table 4, EfficientNet-B0 demonstrates an incremental accuracy gain from 0.97 at epoch 10 to 0.98 at epoch 50, representing an absolute gain of 0.01. This gradual metric progression indicates that the model undergoes a controlled optimization process without reaching an early performance plateau. The learning curve in Fig. 6 supports this, showing a persistent validation–training gap where the validation loss remains consistently lower than the training loss. This non-divergent trajectory indicates that the model performs under strong regularization, where feature representations are constrained to reduce overfitting [48].

Corresponding author: Aji Prasetya Wibawa, aji.prasetya.ft@um.ac.id, Department of Electrical Engineering and Informatics, Universitas Negeri Malang, Malang, Indonesia.
DOI: https://doi.org/10.35882/ijeeemi.v8i2.325

Table 7. Comparison of the proposed method with previous kidney CT image classification studies.

Study	Method	Acc (%)	Pre	Rec	F1-Sc
Hossain et al., 2023 [56]	CNN-Custom	98.66	0.98	0.98	0.98
Pande et al., 2024 [52]	YOLO-v8	82.52	0.86	0.75	0.76
Ekpar et al., 2024 [57]	CNN-based	97.00	0.96	0.96	0.96
Canbay et al., 2024 [50]	MobileNet	99.83	0.99	0.99	0.99
Refaee et al. 2025 [51]	ResNet50	83.04	0.81	0.78	0.78
Hossain et al., 2025 [58]	DenseNet121 + EfficientNet-B0	99.24	0.99	0.99	0.99
Proposed Model	MobileNet-V2	100.0	1.00	1.00	1.00

The confusion matrix in Fig. 9 reveals 24 misclassifications in the Stone class, representing an accuracy difference of approximately 7.65% compared to the Cyst class. This distributed error pattern suggests that EfficientNet-B0 prioritizes global representation stability instead of class-specific optimization. The combination of incremental gains and persistent gap alignment indicates that compound scaling and dropout mechanisms yield conservative convergence dynamics, resulting in a stable yet less aggressive optimization profile than MobileNet-V2.

The comparative analysis underscores that selecting an optimal architecture for medical informatics requires consideration beyond peak classification accuracy [49]. As detailed in Table 5, although the conventional CNN achieves the highest numerical performance, its reliance on full-parameter training and rapid convergence reduces methodological robustness under the current dataset conditions. In contrast, MobileNet-V2 offers greater stability for Clinical Decision Support Systems (CDSS) due to its controlled convergence dynamics along with minimal trainable parameter overhead of 0.23% (5,124 parameters), resulting in a more effective balance between predictive accuracy and resource efficiency. Statistical significance testing in Table 6 further supports this choice, as all pairwise comparisons yield p-values below 0.05. These findings confirm that the observed performance differences are attributable to the models' inherent architectural learning behaviors rather than random variation [49], thereby establishing MobileNet-V2 as a computationally efficient and methodologically robust approach for kidney disease detection.

Table 7 presents a comparative evaluation of the proposed MobileNet-V2 framework against recent deep learning methods applied to kidney CT datasets. Achieving a peak accuracy of 1.00 places the model alongside top-performing models such as Canbay et al. (99.83%) [50]. However, this numerical saturation indicates that accuracy alone is insufficient for differentiation due to dataset homogeneity. The principal methodological advantage of this study is the reduction of trainable parameters to 0.23% (5,124), demonstrating that near-perfect classification performance can be achieved with minimal computational overhead. This efficiency is essential for real-time deployment in resource-limited geriatric care settings. In contrast, the larger performance gap observed in detection-oriented architectures, such as YOLO-v8 (82.52%) and ResNet50 (83.04%), underscores fundamental differences in task formulation. Localization-driven optimization in these models may not completely capture the fine-grained textural and intensity patterns required for multiclass classification [51][52], which are critical for clinical practitioners to distinguish subtle renal pathologies from normal age-related physiological changes. These findings show that MobileNet-V2 is not only a high-accuracy model but also a methodologically stable alternative that balances diagnostic precision with the functional feasibility required for long-term monitoring in elderly populations.

From a practical and clinical perspective, the lightweight architectural design and parameter-optimization strategy of MobileNet-V2 facilitate its integration into Clinical Decision Support Systems (CDSS), particularly within healthcare facilities that possess limited computational resources. As indicated in Table 5, although MobileNet-V2 has a higher total parameter count than a conventional CNN, only 0.23% (5,124) of its parameters are trainable. In contrast, conventional CNN models require updates to all parameters, which increases the computational complexity and hardware requirements during fine-tuning. This efficiency enables MobileNet-V2 to function as a sustainable assistive tool in high-volume radiology workflows [53]. Its controlled convergence and reduced adaptation burden provide scalability benefits that outweigh minor differences in inference latency. In geriatric renal health, where age-related physiological changes may obscure pathological patterns and delay diagnosis, the model's consistent and objective outputs assist clinicians in distinguishing subtle renal indicators from normal aging variations [54]. Integration into CDSS platforms can enhance screening capacity and improve consistency in clinical assessment [55], thereby supporting early detection strategies and long-term clinical outcomes among elderly populations.

Despite the promising findings reported in this study, several limitations should be acknowledged. First, the experiments rely on a publicly available kidney CT dataset without strict patient-level separation. Under such conditions, images from the same patient or highly similar CT slices may appear in both the training and test subsets, potentially introducing implicit data leakage and optimistic bias in performance estimates. Consequently,

Corresponding author: Aji Prasetya Wibawa, aji.prasetya.ft@um.ac.id, Department of Electrical Engineering and Informatics, Universitas Negeri Malang, Malang, Indonesia.
DOI: https://doi.org/10.35882/ijeeemi.v8i2.325

the near-perfect metrics reported earlier may partly reflect residual similarity between samples rather than absolute disease-level generalization. Second, the evaluation was conducted on a single dataset, which may limit the generalizability of the proposed model when applied to heterogeneous clinical environments with different imaging protocols, scanner configurations, or diverse geriatric demographics. Third, the experimental framework did not incorporate extensive data augmentation or cross-validation strategies, which may reduce robustness to distributional shifts during deployment. Although the multi-epoch evaluation framework provides insight into learning stability, it cannot fully substitute for external multi-center validation. Future studies should therefore prioritize evaluation on multi-institutional datasets with strict patient-wise splitting, incorporate more diverse augmentation strategies, and explore explainable artificial intelligence (XAI) techniques to enhance model transparency and clinician trust in geriatric screening. Addressing these limitations will be essential to ensure that future AI-based medical imaging systems achieve reliable generalization and practical applicability in real-world clinical settings.

Overall, this study demonstrates that robust evaluation and critical interpretation are essential when assessing deep learning models for medical image analysis. By integrating convergence analysis, validation alignment, error distribution assessment, and statistical consistency across epochs, the findings substantiate MobileNet-V2, with an accuracy of 1.00 and an AUC of 0.9997 under the evaluated setting, as a methodologically reliable and practically feasible solution for CT-based kidney disease detection.

V. Conclusion

This study evaluates and compares three deep learning architectures CNN, MobileNet-V2, and EfficientNet-B0for CT image–based kidney disease classification. The results demonstrate that, although the conventional CNN achieves perfect numerical performance, its overly rapid convergence behavior and early metric saturation limit the interpretability of its generalization validity under the current dataset configuration. In contrast, MobileNet-V2 exhibits the most balanced performance across accuracy, learning stability, computational efficiency, and methodologically sound generalization. At the same time, EfficientNet-B0 demonstrates more conservative yet stable performance, attributable to its regularization mechanisms. Based on a comprehensive analysis of evaluation metrics, learning curves, and confusion matrices, MobileNet-V2 emerges as the most methodologically reliable architecture within the current evaluation framework for CT-based kidney disease detection in research and decision-support contexts, particularly for applications related to elderly renal health (accuracy 100%). These findings emphasize that model selection for medical imaging should not rely solely on peak accuracy but should consider learning behavior, robustness, and practical feasibility. While further validation using strictly patient-wise and multi-institutional datasets remains necessary, the current findings provide a methodologically grounded basis for lightweight model deployment in CT-based renal screening research. Accordingly, this study contributes to the development of lightweight, reliable AI-based diagnostic support frameworks that can inform future efforts to detect kidney disease early and improve its management in aging populations.

References

[1] H. Bouarich, A. Chávez Guillén, and D. Rodríguez Puyol, “Kidney and hypertension in older adults,” *Medicina Clínica (English Edition)*, vol. 157, no. 4, pp. 178–184, Aug. 2021, doi: 10.1016/j.medcle.2021.02.005.

[2] D. Mekahli *et al.*, “Clinical practice recommendations for kidney involvement in tuberous sclerosis complex: a consensus statement by the ERKNet Working Group for Autosomal Dominant Structural Kidney Disorders and the ERA Genes & Kidney Working Group,” *Nat. Rev. Nephrol.*, vol. 20, no. 6, pp. 402–420, Jun. 2024, doi: 10.1038/s41581-024-00818-0.

[3] G. Alfano *et al.*, “Rethinking Chronic Kidney Disease in the Aging Population,” ., vol. 12, no. 11, p. 1724, Oct. 2022, doi: 10.3390/life12111724.

[4] Z. Chen *et al.*, “Comprehensive 3D Analysis of the Renal System and Stones: Segmenting and Registering Non-Contrast and Contrast Computed Tomography Images,” *Information Systems Frontiers*, vol. 27, no. 1, pp. 97–111, Feb. 2025, doi: 10.1007/s10796-024-10485-y.

[5] G. Sharma, V. Anand, R. Chauhan, H. S. Pokhariya, S. Gupta, and G. Sunil, “Revolutionizing Kidney Disease Diagnosis: A Comprehensive CNN-Based Framework for Multi-Class CT Classification,” in *2024 IEEE International Conference on Information Technology, Electronics and Intelligent Communication Systems (ICITEICS)*, IEEE, Jun. 2024, pp. 1–6. doi: 10.1109/ICITEICS61368.2024.10624992.

[6] R. Kumar *et al.*, “Intelligence Architectures and Machine Learning Applications in Contemporary Spine Care,” *Bioengineering*, vol. 12, no. 9, p. 967, Sep. 2025, doi: 10.3390/bioengineering12090967.

[7] M. N. Hossain, E. Bhuiyan, M. B. A. Miah, T. A. Sifat, Z. Muhammad, and M. F. Al Masud, “Detection and Classification of Kidney Disease from CT Images: An Automated Deep Learning Approach,” *Technologies (Basel).*, vol. 13, no. 11, p. 508, Nov. 2025, doi: 10.3390/technologies13110508.

[8] T. Biloborodova, B. Brosnan, I. Skarga-Bandurova, and D. J. Strauss, “Generalization Ability in Medical Image Analysis with Small-Scale Imbalanced Datasets: Insights from Neural Network Learning,” 2023, pp. 234–246. doi: 10.1007/978-3-031-49011-8_19.

Corresponding author: Aji Prasetya Wibawa, aji.prasetya.ft@um.ac.id, Department of Electrical Engineering and Informatics, Universitas Negeri Malang, Malang, Indonesia.
DOI: https://doi.org/10.35882/ijeeemi.v8i2.325

[9] S. C. Haynes, P. Johnston, and E. Elyan, "Generalisation challenges in deep learning models for medical imagery: insights from external validation of COVID-19 classifiers," *Multimed. Tools Appl.*, vol. 83, no. 31, pp. 76753–76772, Feb. 2024, doi: 10.1007/s11042-024-18543-y.

[10] I. E. Tampu, A. Eklund, and N. Haj-Hosseini, "Inflation of test accuracy due to data leakage in deep learning-based classification of OCT images," *Sci. Data*, vol. 9, no. 1, p. 580, Sep. 2022, doi: 10.1038/s41597-022-01618-6.

[11] I. D. Mienye and T. G. Swart, "A Comprehensive Review of Deep Learning: Architectures, Recent Advances, and Applications," *Information*, vol. 15, no. 12, p. 755, Nov. 2024, doi: 10.3390/info15120755.

[12] B. Almadani, H. Kaisar, I. R. Thoker, and F. Aliyu, "A Systematic Survey of Distributed Decision Support Systems in Healthcare," *Systems*, vol. 13, no. 3, p. 157, Feb. 2025, doi: 10.3390/systems13030157.

[13] S. Sanchez-Martinez *et al.*, "Machine Learning for Clinical Decision-Making: Challenges and Opportunities in Cardiovascular Imaging," *Front. Cardiovasc. Med.*, vol. 8, Jan. 2022, doi: 10.3389/fcvm.2021.765693.

[14] G. Menghani, "Efficient Deep Learning: A Survey on Making Deep Learning Models Smaller, Faster, and Better," *ACM Comput. Surv.*, vol. 55, no. 12, pp. 1–37, Dec. 2023, doi: 10.1145/3578938.

[15] L. J. Basile, N. Carbonara, R. Pellegrino, and U. Panniello, "Business intelligence in the healthcare industry: The utilization of a data-driven approach to support clinical decision making," *Technovation*, vol. 120, p. 102482, Feb. 2023, doi: 10.1016/j.technovation.2022.102482.

[16] A. S. Mahmoud, O. Lamouchi, and S. Belghith, "Advancements in Machine Learning and Deep Learning for Early Diagnosis of Chronic Kidney Diseases: A Comprehensive Review," *Babylonian Journal of Machine Learning*, vol. 2024, pp. 149–156, Sep. 2024, doi: 10.58496/BJML/2024/015.

[17] M. Gharaibeh *et al.*, "Radiology Imaging Scans for Early Diagnosis of Kidney Tumors: A Review of Data Analytics-Based Machine Learning and Deep Learning Approaches," *Big Data and Cognitive Computing*, vol. 6, no. 1, p. 29, Mar. 2022, doi: 10.3390/bdcc6010029.

[18] G. Jeyalakshmi, F. V. Lloyd, K. Subbulakshmi, and G. Vinudevi, "A Deep Learning Approach for Predicting Chronic Kidney Disease in Electronic Health Records in Neural Nephrology," 2024, pp. 197–216. doi: 10.4018/979-8-3693-8659-0.ch011.

[19] M. Rashid and M. Sharma, "AI-Assisted Diagnosis and Treatment Planning A Discussion of How AI Can Assist Healthcare Professionals in Making More Accurate Diagnoses and Treatment Plans for Diseases," in *AI in Disease Detection*, Wiley, 2025, pp. 313–336. doi: 10.1002/9781394278695.ch14.

[20] M. T. Islam, M. A. Rahman, Md. T. R. Mazumder, and S. H. Shourov, "COMPARATIVE ANALYSIS OF NEURAL NETWORK ARCHITECTURES FOR MEDICAL IMAGE CLASSIFICATION: EVALUATING PERFORMANCE ACROSS DIVERSE MODELS," *American Journal of Advanced Technology and Engineering Solutions*, vol. 04, no. 01, pp. 01–42, Mar. 2024, doi: 10.63125/feed1x52.

[21] F. Mohr and J. N. van Rijn, "Learning curves for decision making in supervised machine learning: a survey," *Mach. Learn.*, vol. 113, no. 11–12, pp. 8371–8425, Dec. 2024, doi: 10.1007/s10994-024-06619-7.

[22] M. Aljaafari, S. E. El-Deep, A. A. Abohany, and S. E. Sorour, "Integrating Innovation in Healthcare: The Evolution of 'CURA's' AI-Driven Virtual Wards for Enhanced Diabetes and Kidney Disease Monitoring," *IEEE Access*, vol. 12, pp. 126389–126414, 2024, doi: 10.1109/ACCESS.2024.3451369.

[23] A. Padhi, A. Agarwal, S. K. Saxena, and C. D. S. Katoch, "Transforming clinical virology with AI, machine learning and deep learning: a comprehensive review and outlook," *Virusdisease*, vol. 34, no. 3, pp. 345–355, Sep. 2023, doi: 10.1007/s13337-023-00841-y.

[24] K. Blagec, J. Kraiger, W. Frühwirt, and M. Samwald, "Benchmark datasets driving artificial intelligence development fail to capture the needs of medical professionals," *J. Biomed. Inform.*, vol. 137, p. 104274, Jan. 2023, doi: 10.1016/j.jbi.2022.104274.

[25] D. Vrbaški, B. Vesin, and K. Mangaroska, "Machine Learning for Chronic Kidney Disease Detection from Planar and SPECT Scintigraphy: A Scoping Review," *Applied Sciences*, vol. 15, no. 12, p. 6841, Jun. 2025, doi: 10.3390/app15126841.

[26] H. Bechinia, D. Benmerzoug, and N. Khlifa, "Approach Based Lightweight Custom Convolutional Neural Network and Fine-Tuned MobileNet-V2 for ECG Arrhythmia Signals Classification," *IEEE Access*, vol. 12, pp. 40827–40841, 2024, doi: 10.1109/ACCESS.2024.3378730.

[27] E. Tasci, Y. Zhuge, K. Camphausen, and A. V. Krauze, "Bias and Class Imbalance in Oncologic DataTowards Inclusive and Transferrable AI in Large Scale Oncology Data Sets," *Cancers (Basel).*, vol. 14, no. 12, p. 2897, Jun. 2022, doi: 10.3390/cancers14122897.

[28] A. Alsajri and A. V. Hacimahmud, "Review of deep learning: Convolutional Neural Network Algorithm," *Babylonian Journal of Machine Learning*, vol. 2023, pp. 19–25, Apr. 2023, doi: 10.58496/BJML/2023/004.

[29] L. Shen, Y. Sun, Z. Yu, L. Ding, X. Tian, and D. Tao, "On Efficient Training of Large-Scale Deep Learning Models," *ACM Comput. Surv.*, vol. 57, no. 3, pp. 1–36, Mar. 2025, doi: 10.1145/3700439.

[30] H. Polat and H. Danaei Mehr, "Classification of Pulmonary CT Images by Using Hybrid 3D-Deep

Corresponding author: Aji Prasetya Wibawa, aji.prasetya.ft@um.ac.id, Department of Electrical Engineering and Informatics, Universitas Negeri Malang, Malang, Indonesia.
DOI: https://doi.org/10.35882/ijeeemi.v8i2.325

Convolutional Neural Network Architecture," *Applied Sciences*, vol. 9, no. 5, p. 940, Mar. 2019, doi: 10.3390/app9050940.
[31] L. Zhao and Z. Zhang, "A improved pooling method for convolutional neural networks," *Sci. Rep.*, vol. 14, no. 1, p. 1589, Jan. 2024, doi: 10.1038/s41598-024-51258-6.
[32] I. Salehin and D.-K. Kang, "A Review on Dropout Regularization Approaches for Deep Neural Networks within the Scholarly Domain," *Electronics (Basel).*, vol. 12, no. 14, p. 3106, Jul. 2023, doi: 10.3390/electronics12143106.
[33] J. Glory Precious, S. P. Angeline Kirubha, and I. Keren Evangeline, "Deployment of a Mobile Application Using a Novel Deep Neural Network and Advanced Pre-Trained Models for the Identification of Brain Tumours," *IETE J. Res.*, vol. 69, no. 10, pp. 6902–6914, Oct. 2023, doi: 10.1080/03772063.2022.2083027.
[34] M. Kia, S. Sadeghi, H. Safarpour, M. Kamsari, S. Jafarzadeh Ghoushchi, and R. Ranjbarzadeh, "Innovative fusion of VGG16, MobileNet, EfficientNet, AlexNet, and ResNet50 for MRI-based brain tumor identification," *Iran Journal of Computer Science*, vol. 8, no. 1, pp. 185–215, Mar. 2025, doi: 10.1007/s42044-024-00216-6.
[35] M. Reyad, A. M. Sarhan, and M. Arafa, "A modified Adam algorithm for deep neural network optimization," *Neural Comput. Appl.*, vol. 35, no. 23, pp. 17095–17112, Aug. 2023, doi: 10.1007/s00521-023-08568-z.
[36] M. Reyad, A. M. Sarhan, and M. Arafa, "A modified Adam algorithm for deep neural network optimization," *Neural Comput. Appl.*, vol. 35, no. 23, pp. 17095–17112, Aug. 2023, doi: 10.1007/s00521-023-08568-z.
[37] R. Chandrasekaran, F. Asgareinjad, J. Morris, and T. Rosing, "Multi-Label Classification With Hyperdimensional Representations," *IEEE Access*, vol. 11, pp. 108458–108474, 2023, doi: 10.1109/ACCESS.2023.3299881.
[38] N. Sampathila *et al.*, "Customized Deep Learning Classifier for Detection of Acute Lymphoblastic Leukemia Using Blood Smear Images," *Healthcare*, vol. 10, no. 10, p. 1812, Sep. 2022, doi: 10.3390/healthcare10101812.
[39] H. Li, G. K. Rajbahadur, D. Lin, C.-P. Bezemer, and Z. M. Jiang, "Keeping Deep Learning Models in Check: A History-Based Approach to Mitigate Overfitting," *IEEE Access*, vol. 12, pp. 70676–70689, 2024, doi: 10.1109/ACCESS.2024.3402543.
[40] K. M. Sujon, R. Hassan, K. Choi, and M. A. Samad, "Accuracy, precision, recall, f1-score, or MCC? empirical evidence from advanced statistics, ML, and XAI for evaluating business predictive models," *J. Big Data*, vol. 12, no. 1, p. 268, Dec. 2025, doi: 10.1186/s40537-025-01313-4.
[41] C. Mosquera, L. Ferrer, D. H. Milone, D. Luna, and E. Ferrante, "Class imbalance on medical image classification: towards better evaluation practices for discrimination and calibration performance," *Eur. Radiol.*, vol. 34, no. 12, pp. 7895–7903, Jun. 2024, doi: 10.1007/s00330-024-10834-0.
[42] J. Bernett *et al.*, "Guiding questions to avoid data leakage in biological machine learning applications," *Nat. Methods*, vol. 21, no. 8, pp. 1444–1453, Aug. 2024, doi: 10.1038/s41592-024-02362-y.
[43] M. Bausch *et al.*, "Distinct neuronal populations in the human brain combine content and context," *Nature*, vol. 650, no. 8102, pp. 690–700, Feb. 2026, doi: 10.1038/s41586-025-09910-2.
[44] A. M. Duhaim, N. S. Baqer, and M. A. Fadhel, "Developing a trustworthy and explainable framework for classifying skin lesions through transfer learning and attention mechanisms," *Comput. Biol. Chem.*, vol. 122, p. 108914, Jun. 2026, doi: 10.1016/j.compbiolchem.2026.108914.
[45] W. Obaid, A. Hussain, T. Rabie, D. H. Abd, and W. Mansoor, "Multi-model deep learning approach for the classification of kidney diseases using medical images," *Inform. Med. Unlocked*, vol. 57, p. 101663, 2025, doi: 10.1016/j.imu.2025.101663.
[46] R. Hou, J. Y. Lo, J. R. Marks, E. S. Hwang, and L. J. Grimm, "Classification performance bias between training and test sets in a limited mammography dataset," *PLoS One*, vol. 19, no. 2, p. e0282402, Feb. 2024, doi: 10.1371/journal.pone.0282402.
[47] N. Rachburee and W. Punlumjeak, "Lotus species classification using transfer learning based on VGG16, ResNet152V2, and MobileNetV2," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 11, no. 4, p. 1344, Dec. 2022, doi: 10.11591/ijai.v11.i4.pp1344-1352.
[48] A. Mohi, L. Kareem, and A. A. Hassan, "EfficientNet-B0 Deep Learning Model for Accurate Classification of Intrabony Lesions in Dental Panoramic Radiographs," May 05, 2025. doi: 10.21203/rs.3.rs-6207574/v1.
[49] L. K. Singh, M. Khanna, H. Monga, R. singh, and G. Pandey, "Nature-Inspired Algorithms-Based Optimal Features Selection Strategy for COVID-19 Detection Using Medical Images," *New Gener. Comput.*, vol. 42, no. 4, pp. 761–824, Nov. 2024, doi: 10.1007/s00354-024-00255-4.
[50] Y. Canbay, S. Adsiz, and P. Canbay, "Privacy-Preserving Transfer Learning Framework for Kidney Disease Detection," *Applied Sciences*, vol. 14, no. 19, p. 8629, Sep. 2024, doi: 10.3390/app14198629.
[51] E. A. Refaee, "Digital transformation in healthcare: leveraging machine learning for predictive analytics in chronic kidney diseases prevention," *Journal of Innovative Digital Transformation*, vol. 2, no. 3, pp. 217–232, Nov. 2025, doi: 10.1108/JIDT-03-2025-0012.
[52] S. D. Pande and R. Agarwal, "Multi-Class Kidney Abnormalities Detecting Novel System Through

Corresponding author: Aji Prasetya Wibawa, aji.prasetya.ft@um.ac.id, Department of Electrical Engineering and Informatics, Universitas Negeri Malang, Malang, Indonesia.
DOI: https://doi.org/10.35882/ijeeemi.v8i2.325

Computed Tomography," *IEEE Access*, vol. 12, pp. 21147–21155, 2024, doi: 10.1109/ACCESS.2024.3351181.
[53] K. Eskandar, "Artificial Intelligence in CT Imaging: A Systematic Review of Diagnostic Accuracy, Clinical Decision–Support Impact, and Integration Pathways," *iRADIOLOGY*, vol. 3, no. 6, pp. 434–445, Dec. 2025, doi: 10.1002/ird3.70046.
[54] R. Tenchov, J. M. Sasso, X. Wang, and Q. A. Zhou, "Aging Hallmarks and Progression and Age-Related Diseases: A Landscape View of Research Advancement," *ACS Chem. Neurosci.*, vol. 15, no. 1, pp. 1–30, Jan. 2024, doi: 10.1021/acschemneuro.3c00531.
[55] D. M. Connaughton and A. J. Mallett, "Models of Care for the Implementation of Genetic Testing in Nephrology," *Semin. Nephrol.*, vol. 45, no. 5, p. 151649, Sep. 2025, doi: 10.1016/j.semnephrol.2025.151649.
[56] M. S. Hossain, S. M. Nazmul Hassan, M. Al-Amin, Md. N. Rahaman, R. Hossain, and M. I. Hossain, "Kidney Disease Detection from CT Images using a customized CNN model and Deep Learning," in *2023 International Conference on Advances in Intelligent Computing and Applications (AICAPS)*, IEEE, Feb. 2023, pp. 1–6. doi: 10.1109/AICAPS57044.2023.10074314.
[57] F. E. Ekpar, "Image-based Chronic Kidney Disease Diagnosis Using 2D Convolutional Neural Networks in the Context of a Comprehensive Artificial Intelligence-Driven Healthcare System," *MOLECULAR SCIENCES AND APPLICATIONS*, vol. 4, pp. 135–143, Nov. 2024, doi: 10.37394/232023.2024.4.13.
[58] M. N. Hossain, E. Bhuiyan, M. B. A. Miah, T. A. Sifat, Z. Muhammad, and M. F. Al Masud, "Detection and Classification of Kidney Disease from CT Images: An Automated Deep Learning Approach," *Technologies (Basel).*, vol. 13, no. 11, p. 508, Nov. 2025, doi: 10.3390/technologies13110508.

Author Biography

ARDHA ARDHANA PUTRA AGUSTAVADA is an undergraduate student in the Department of Electrical Engineering and Informatics at Universitas Negeri Malang. He has a strong academic interest in artificial intelligence. Throughout his studies, he has been actively involved in developing web-based and AI-driven systems. His research interests include threat detection and intelligent educational platforms. In addition to his academic pursuits, he has experience in modern web development technologies, which support his interdisciplinary approach to problem-solving. He is committed to expanding his expertise in intelligent systems. He aims to contribute to research and innovation in applied artificial intelligence.

AJI PRASETYA WIBAWA is a Professor in Knowledge Engineering and Data Science at Universitas Negeri Malang (UM). He received his Ph.D. degree in Electrical and Information Engineering from the University of South Australia (UniSA). He is currently the Head of the Scientific Publication Unit at UM, Indonesia. He serves as the Research Group Leader for Knowledge Engineering and Data Science (KEDS). He is also an active member of the Association for Scientific Computing and Engineering (ASCEE). His research interests span artificial intelligence (AI), data mining, machine translation, natural language processing (NLP), Javanese language computing, and social informatics. Prof. Wibawa's scholarly contributions are reflected in his editorial and publication involvement with journals such as Knowledge Engineering and Data Science (KEDS) and the Bulletin of Social Informatics Theory and Application (BUSINTA), where he promotes interdisciplinary and ethical research in intelligent systems and computational knowledge discovery.

ABDULLAH SHOLUM is an undergraduate student in the Department of Informatics Engineering and Electrical Engineering at Universitas Negeri Malang. His academic interests include information technology, computational systems, and the integration of software and hardware for engineering applications. He has experience in programming, system development, applied electronics, and data processing, with a focus on algorithmic design and technology-driven problem solving. His research interests center on applying modern computing techniques to enhance efficiency and innovation in engineering and information systems.

DAFA FADHILAH HILMI is an undergraduate student in the Department of Electrical Engineering and Informatics at Universitas Negeri Malang. His academic focus integrates electrical engineering and informatics, particularly in computational systems, data processing, and system design. He has engaged in academic activities related to data analysis and software-based solutions, emphasizing analytical reasoning and structured problem-solving. His interests include the development of efficient information systems and the production of clear, systematic, and research-oriented academic work.

Corresponding author: Aji Prasetya Wibawa, aji.prasetya.ft@um.ac.id, Department of Electrical Engineering and Informatics, Universitas Negeri Malang, Malang, Indonesia.
DOI: https://doi.org/10.35882/ijeeemi.v8i2.325

FELIX ANDIKA DWIYANTO was born in Malang in 1994. He is now pursuing a doctoral degree at AGH University of Science and Technology in Krakow, Poland. He received a bachelor's degree in informatics engineering education and a master's degree in vocational education from Universitas Negeri Malang (UM). He is active in scientific writing, publishing in international conferences and journals across multiple countries, including Indonesia, Germany, and Thailand. His field of study is the development of information technology-based learning media, information systems, and data structures. He is active as the Editor of the Belantika Pendidikan journal.

Corresponding author: Aji Prasetya Wibawa, aji.prasetya.ft@um.ac.id, Department of Electrical Engineering and Informatics, Universitas Negeri Malang, Malang, Indonesia.
DOI: https://doi.org/10.35882/ijeeemi.v8i2.325