International Journal of Informatics and Data Science
ISSN 3026-7315 (media online)
Volume 3, No 2, June 2026, Page: 127−136
DOI: doi.org/10.64366/ijids.v3i2.576
https://journals.adaresearch.or.id/ijids/index

Transfer Learning Performance and Efficiency of VGG16 and MobileNetV2 on Three-Class Animals-10 Classification

Teti Desyani[*], Hendri Ardiansyah, Oktaviyanus

Faculty of Computer Science, Informatics Engineering Study Program, Universitas Pamulang, South Tangerang, Indonesia
Email: [1,*]dosen00839@unpam.ac.id, [2]dosen00832@unpam.ac.id, [3]oktaviyanus02@gmail.com
Correspondence Author Email: dosen00839@unpam.ac.id
Received: 12/05/2026, Revised: 03/06/2026, Accepted: 18/06/2026, Published: 30/06/2026

Abstract–Animal image classification remains challenging because variations in pose, illumination, object scale, viewing angle, and background can affect model predictions. In addition, high-capacity convolutional neural networks may impose substantial computational and storage requirements. This study compares the classification performance and computational efficiency of VGG16 and MobileNetV2 using transfer learning. Dog, Cat, and Sheep images were selected from the Animals-10 dataset to represent visually related companion animals and a visually distinct livestock category while allowing balanced sampling. Following image validation, duplicate checking, and class balancing, the final dataset contained 5,004 images, with 1,668 images per class. The dataset was divided using an identical stratified 80:20 split for both models, with 20% of the training portion used for validation. Both models used ImageNet weights, 224 × 224-pixel inputs, identical classification heads, data augmentation, and equivalent training configurations. MobileNetV2 achieved 97.80% accuracy, a 97.81% macro F1-score, and a 99.88% macro ROC-AUC, whereas VGG16 achieved 96.80%, 96.81%, and 99.67%, respectively. MobileNetV2 reduced training time by 66.03%, inference time by 59.56%, and weight size by 77.83%. However, the exact McNemar test indicated no statistically significant difference in classification performance ($p = 0.064$). Therefore, MobileNetV2 provided substantially greater computational efficiency while maintaining classification performance comparable to VGG16, making it a more practical option for resource-constrained animal image classification systems.

Keywords: Animal Image Classification; MobileNetV2; Transfer Learning; VGG16; Computational Efficiency

1. INTRODUCTION

The development of computer vision and deep learning has expanded the application of automated image classification across various fields, including wildlife monitoring, biodiversity conservation, livestock management, human-animal conflict mitigation, and camera-trap system development. Conventional animal monitoring generally requires images to be manually observed, sorted, and labeled. This process becomes increasingly difficult as the volume of collected images continues to grow. Studies on visual animal monitoring have shown that deep learning has been employed to support animal detection, tracking, pose estimation, species recognition, and behavior classification [1]. These developments demonstrate that automated image processing can reduce the burden of manual analysis and accelerate the extraction of information from visual data [2].

Animal image classification continues to face several challenges because object appearance can vary with pose, viewing angle, size, illumination, image quality, and background conditions. Images collected from natural environments may also contain noise, partially visible objects, or other visual characteristics that are shared across classes. Battu and Reddy Lakshmi [3] demonstrated that the classification of animal images obtained from camera traps must address diverse environmental conditions and variations in label quality. Tøn et al. [4] similarly explained that unfavourable viewing angles, illumination, and image quality can limit classification performance when a system relies exclusively on visual information. Their study improved wildlife classification accuracy from 98.4% to 98.9% by combining images with metadata. These findings suggest that the quality of visual representations and model design are essential factors in producing reliable classifications.

Convolutional Neural Networks (CNNs) are widely used for image classification because they can learn hierarchical visual features directly from data. Early convolutional layers generally identify simple features, such as edges, colours, and textures, whereas deeper layers form more complex object representations.[5] However, training a CNN from scratch requires a large dataset, considerable training time, and substantial computational resources. Transfer learning offers an approach to reducing these requirements by leveraging weights from models previously trained on large-scale datasets. Nazir and Kaleem [6] applied transfer learning to camera-trap image classification and demonstrated that pretrained models could support image processing on edge devices. A pretrained deep learning approach has also been employed for animal classification in the context of automated chick sex determination [7]. These findings suggest that transfer learning can be applied to various image classification tasks when the availability of data and computational resources is limited.

VGG16 is one of the CNN architectures frequently used in transfer-learning applications. Its deep and relatively uniform arrangement of convolutional layers enables it to produce strong visual feature representations. VGG16 has been applied to various image-classification problems and has been extended through architectural modification and integration with other models. Kumar and Kumar [8] developed Efficient-VGG16 for X-ray image classification using transfer learning and evaluated it across several data-splitting scenarios. Another study combined a VGG16-based feature extractor with an integration of LeNet and MobileNetV2 to improve medical image classification [9], [10]. Despite its feature-extraction capability, VGG16 requires a relatively large number of

ISSN 3026-7315 (media online)
Volume 3, No 2, June 2026, Page: 127−136
DOI: doi.org/10.64366/ijids.v3i2.576
https://journals.adaresearch.or.id/ijids/index

parameters, computational operations, and storage capacity. These characteristics may restrict its implementation on devices with limited memory and processing resources. MobileNetV2 was developed as a lightweight architecture using depthwise separable convolutions, inverted residual blocks, and linear bottlenecks. These components reduce computational operations while retaining the model's ability to extract meaningful visual features. Indraswari et al. [11] used MobileNetV2 for melanoma image classification and demonstrated its potential for further implementation on mobile devices. MobileNetV2 has also been enhanced through the integration of attention mechanisms and optimization methods for medical image classification [12]. In animal image classification, Nazir and Kaleem [6] reported that MobileNetV2 achieved a precision of 0.95, recall of 0.96, and an F1-score of 0.95 on one of the camera-trap datasets used in their study. These results demonstrate that a lightweight architecture can deliver competitive performance while meeting deployment requirements for edge devices.

Computational efficiency is important because high classification accuracy alone does not necessarily indicate that a model is suitable for practical implementation. A model with a large number of parameters may require longer inference time, greater storage capacity, and more powerful computing resources. Studies on lightweight CNNs in various domains have therefore emphasised the need to maintain predictive performance while reducing model complexity. [13] [14]. In animal image classification, computational efficiency affects the ability of a system to produce timely predictions on camera traps, mobile applications, and livestock-monitoring devices. Consequently, model comparisons should consider not only accuracy, precision, recall, and F1-score but also training time, inference time, parameter count, and model weight size.[15]

The quality and distribution of a dataset can also influence classification results. Class imbalance may cause a model to favour the majority class, meaning that overall accuracy may not adequately represent performance across individual classes. Image augmentation can increase the variation of training data [16], however, augmentation should be applied only to the training subset so that the validation and testing subsets continue to represent unmodified images. Previous research has indicated that increasing training-data diversity can help models learn more general visual representations [17]. Furthermore, training, validation, and testing subsets must be separated consistently and checked for duplicate images to prevent data leakage and overly optimistic evaluation results.

Previous studies have demonstrated the effectiveness of VGG16, MobileNetV2, and transfer learning across multiple image-classification domains [6], [8], [9], [11], [12]. However, many studies have primarily focused on improving accuracy within a particular domain, constructing hybrid architectures, or modifying only one model. Direct comparisons between VGG16 and MobileNetV2 have not always employed an identical data partition, classification head, training configuration, evaluation device, and efficiency-measurement procedure. In addition, differences in accuracy are frequently reported only descriptively, without paired statistical testing of the predictions produced by the compared models.[18] A small numerical difference therefore cannot independently establish that one model has significantly different classification performance.

This study selects Dog, Cat, and Sheep images from the Animals-10 dataset. Dog and Cat were selected as visually related companion-animal categories that may share characteristics such as fur texture, body contours, poses, and indoor or outdoor backgrounds. Sheep was included as a livestock category with more distinctive characteristics, while still presenting potentially overlapping visual cues, including fur-like texture, four-legged body structure, and outdoor backgrounds. The availability of images in these three classes also enables balanced sampling of 1,668 images per class without synthetic oversampling. This selection provides a focused experimental setting for examining model behaviour across both visually similar and relatively distinct animal categories. However, because only three of the ten Animals-10 classes are included, the findings are interpreted specifically within this controlled three-class classification setting.

Following class balancing, the final dataset consisted of 5,004 images. A stratified 80:20 split was applied, and the same partition was used for both models. VGG16 and MobileNetV2 were implemented using ImageNet-pretrained weights, an input resolution of 224×224pixels, identical classification heads, equivalent augmentation operations, the same optimiser and learning rate, an equal batch size, and the same early-stopping mechanism. This controlled configuration was designed to minimise the influence of unequal experimental treatments on the comparison.

The models were evaluated using accuracy, macro precision, macro recall, macro F1-score, macro Receiver Operating Characteristic–Area Under the Curve (ROC–AUC), and confusion matrices. Computational efficiency was assessed based on total training time, average time per epoch, inference time, inference time per image, parameter count, and model weight size. Because both models generated predictions for the same testing images, the exact McNemar test was employed to determine whether their classification outcomes differed significantly. Therefore, the contribution of this study is a controlled comparison of a relatively high-capacity model and a lightweight model by simultaneously considering classification performance, computational efficiency, and statistical significance. The findings are expected to provide an empirical basis for selecting an appropriate architecture for animal image-classification systems, particularly in environments with limited computational and storage resources.

International Journal of Informatics and Data Science
ISSN 3026-7315 (media online)
Volume 3, No 2, June 2026, Page: 127−136
DOI: doi.org/10.64366/ijids.v3i2.576
https://journals.adaresearch.or.id/ijids/index

2. RESEARCH METHODOLOGY

This study employed a comparative experimental method to evaluate the classification performance and computational efficiency of VGG16 and MobileNetV2 in classifying Dog, Cat, and Sheep images. The research procedure included dataset selection and validation, class balancing, stratified data partitioning, image preprocessing and augmentation, [3], [4]. transfer-learning implementation, model training, performance evaluation, efficiency measurement, and statistical testing. Transfer learning was selected because it enables pretrained models to reuse visual representations learned from a large-scale dataset without retraining the entire convolutional network from scratch [6].

2.1 Research Stages and Research Objects

The research stages are presented in Figure 1. The experiment began by selecting three classes from the Animals-10 dataset: *cane*, *gatto*, and *pecora*, which were mapped to Dog, Cat, and Sheep, respectively.

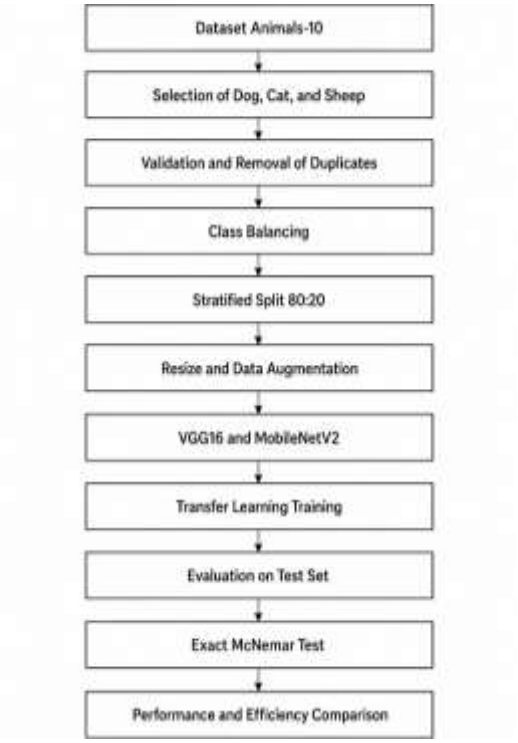

Figure 1. Research stages for comparing VGG16 and MobileNetV2

As illustrated in Figure 1, the validated dataset was balanced through random undersampling using a random seed of 42. The number of images was adjusted according to the smallest class, resulting in 1,668 images for each class and a total of 5,004 images. The original dataset contained 4,863 Dog images, 1,668 Cat images, and 1,820 Sheep images. Table 1 presents the distribution before and after class balancing.

Table 1. Dataset distribution before and after class balancing

Class	Original Folder	Initial Number	Final Number
Dog	cane	4.863	1.668
Cat	gatto	1.668	1.668
Sheep	pecora	1.820	1.668
Total		8.351	5.004

As shown in Table 1, class balancing removed the dominance of the Dog class and provided equal class contributions during training and evaluation. The balanced dataset was then divided using a stratified 80:20 hold-out procedure with random_state=42. This process produced 4,003 training images and 1,001 testing images. Subsequently, 20% of the training portion was allocated to validation, resulting in 3,202 training images and 801 validation images. Examination of the data partitions showed no overlap between the training, validation, and testing subsets.

A single stratified split was used to ensure that both architectures were trained and evaluated on identical samples while maintaining manageable computational requirements. The fixed random seed supported experimental reproducibility. Nevertheless, the absence of repeated splits or cross-validation means that the reported results represent the specified data partition. This limitation is considered when interpreting the findings.

ISSN 3026-7315 (media online)
Volume 3, No 2, June 2026, Page: 127−136
DOI: doi.org/10.64366/ijids.v3i2.576
https://journals.adaresearch.or.id/ijids/index

2.2 Comparative Method for VGG16 and MobileNetV2

All images were resized to 224×224pixels with three RGB channels using bilinear interpolation. Augmentation was applied only to the training images and consisted of horizontal flipping, random rotation with a factor of 0.10, random zoom with a factor of 0.10, and random contrast adjustment with a factor of 0.10. Validation and testing images were not augmented. VGG16 preprocessing converted RGB values to BGR and subtracted the ImageNet channel means, whereas MobileNetV2 preprocessing scaled pixel values to the range $[-, 11]$.

Both models used ImageNet-pretrained weights with include_top=False. Their convolutional bases were frozen and employed as feature extractors. An identical classification head was attached to each model, consisting of Global Average Pooling, a dense layer with 256 units and ReLU activation, dropout with a rate of 0.5, and a three-unit softmax output layer. This configuration ensured that the pretrained architecture was the primary experimental difference.

The models were trained in a Kaggle Notebook using two GPUs through MirroredStrategy. Adam was used with an initial learning rate of 1×10^{-4}, a batch size of 32, a maximum of 20 epochs, and sparse categorical cross-entropy loss. Early stopping monitored validation loss with a patience of five epochs and restored the best model weights. ReduceLROnPlateau used a factor of 0.2, a patience of two epochs, and a minimum learning rate of 1×10^{-7}. The complete configuration is summarised in Table 2.

Table 2. Experimental configurations for VGG16 and MobileNetV2

Parameter	VGG16	MobileNetV2
Pretrained weights	ImageNet	ImageNet
Input size	224 × 224 × 3	224 × 224 × 3
Convolutional base	Frozen	Frozen
Global Average Pooling	Yes	Yes
Dense layer	256, ReLU	256, ReLU
Dropout	0.5	0.5
Output layer	3, Softmax	3, Softmax
Optimizer	Adam	Adam
Initial learning rate	1×10^{-4}	1×10^{-4}
Batch size	32	32
Maximum epochs	20	20
Early stopping	Patience of 5	Patience of 5
Random seed	42	42

The final evaluation used the same 1,001 testing images in an identical order. Classification performance was evaluated using accuracy, macro precision, macro recall, macro F1-score, macro ROC–AUC with the one-versus-rest approach, and confusion matrices. Computational efficiency was assessed using total training time, average time per epoch, inference time, inference time per image, parameter count, and model weight size. Although the fixed hold-out partition supported an identical comparison between the two models, repeated data splits or cross-validation could provide a broader estimate of model stability and generalisation.[19] [20]

The exact McNemar test was used because both models produced paired predictions for the same testing images. The significance level was set at $\alpha = 0.05$. A result of $p < 0.05$indicated a statistically significant difference in classification outcomes, whereas $p \geq 0.05$indicated that the observed difference was not statistically significant.

3. RESULT AND DISCUSSION

3.1 Research Results

The experiments compared VGG16 and MobileNetV2 using the same training, validation, and testing subsets. Both models were trained with ImageNet-pretrained convolutional bases, identical classification heads, and equivalent optimisation configurations. Therefore, differences in the results primarily reflected the characteristics of the two convolutional architectures.

VGG16 completed 20 training epochs before the early-stopping criterion was satisfied. The best model weights were obtained at epoch 15, when the lowest validation loss of 0.0627 was recorded. At the end of the training process, VGG16 achieved a training accuracy of 97.02% and a validation accuracy of 97.42%. Its total training time was 283.43 seconds, corresponding to an average of 14.17 seconds per epoch.

Figure 2 presents the accuracy and loss curves of VGG16. The training accuracy increased considerably during the initial epochs and subsequently became more stable. A similar pattern appeared in the validation accuracy, which remained close to the training accuracy throughout most of the training process. Although several fluctuations occurred, the difference between the two curves was relatively small. This pattern indicates that VGG16 learned relevant representations without showing severe overfitting [1], [3].

International Journal of Informatics and Data Science
ISSN 3026-7315 (media online)
Volume 3, No 2, June 2026, Page: 127−136
DOI: doi.org/10.64366/ijids.v3i2.576
https://journals.adaresearch.or.id/ijids/index

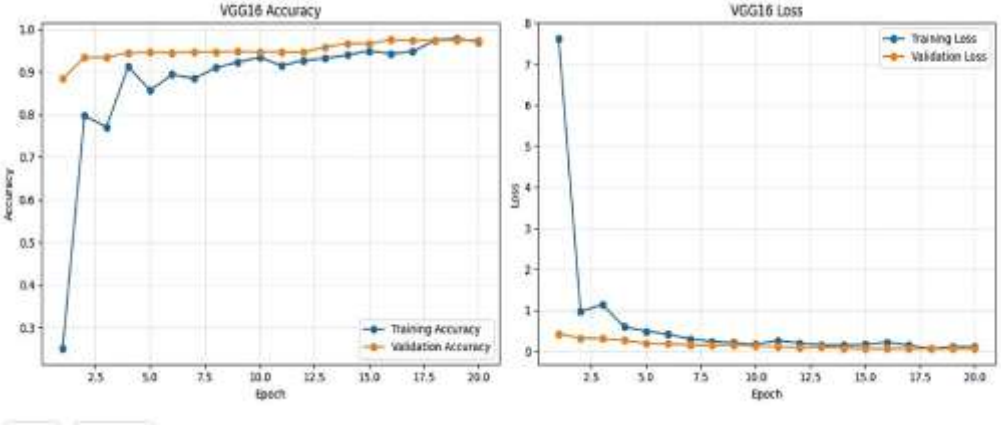

Figure 2. Training and validation curves of VGG16

The validation accuracy was occasionally higher than the training accuracy. This result is reasonable because data augmentation and dropout were activated only during training. Augmentation increased the difficulty of the training images by introducing variations in orientation, scale, and contrast. Dropout also temporarily deactivated some units in the classification head. During validation, both mechanisms were disabled, allowing the model to process unmodified images using the complete network.

MobileNetV2 reached the early-stopping criterion after 13 epochs. Its best model weights were obtained at epoch 8, with a validation loss of 0.0841 and a validation accuracy of 97.94%. At the final epoch, MobileNetV2 achieved a training accuracy of 96.83% and a validation accuracy of 98.04%. The total training time was 96.29 seconds, with an average of 7.41 seconds per epoch.

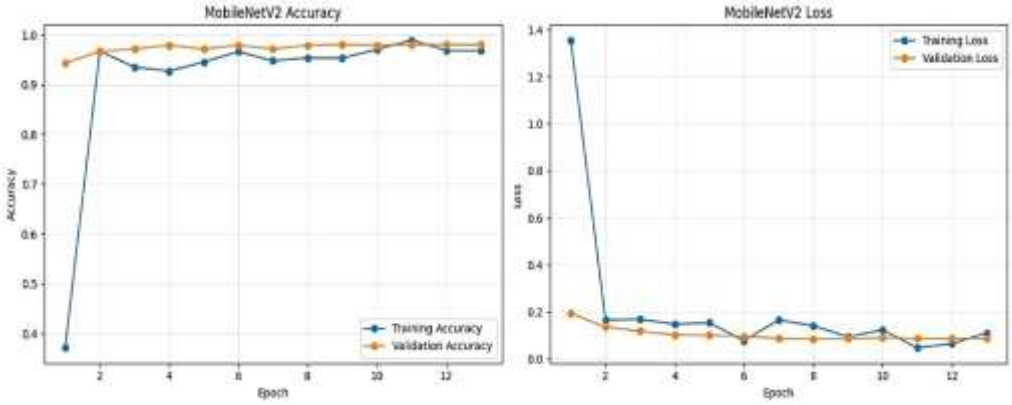

Figure 3. Training and validation curves of MobileNetV2

As illustrated in Figure 3, MobileNetV2 reached a high accuracy level during the early stages of training. Its training and validation curves subsequently remained relatively stable. Validation accuracy was again occasionally higher than training accuracy because augmentation and dropout were applied only during training. The loss curves showed several minor fluctuations but did not indicate continuously increasing validation loss. Thus, MobileNetV2 also demonstrated satisfactory generalisation within the specified data partition.

The best validation loss of VGG16 was lower than that of MobileNetV2, whereas the best validation accuracy of MobileNetV2 was slightly higher. This difference is possible because validation loss considers the confidence assigned to all classes, whereas accuracy only considers whether the class with the highest probability is correct. A model can produce more correct classifications while assigning less confident probabilities to some samples, resulting in higher accuracy but not necessarily lower loss.

The final evaluation was performed on the same 1,001 testing images. Table 3 presents the classification metrics obtained by VGG16 and MobileNetV2.

Table 3. Classification performance of VGG16 and MobileNetV2

Metric	VGG16	MobileNetV2	Difference
Accuracy	96.80%	97.80%	1.00 pp
Macro precision	96.85%	97.81%	0.96 pp
Macro recall	96.80%	97.80%	1.00 pp
Macro F1-score	96.81%	97.81%	1.00 pp

International Journal of Informatics and Data Science
ISSN 3026-7315 (media online)
Volume 3, No 2, June 2026, Page: 127−136
DOI: doi.org/10.64366/ijids.v3i2.576
https://journals.adaresearch.or.id/ijids/index

Metric	VGG16	MobileNetV2	Difference
Macro ROC–AUC	99.67%	99.88%	0.21 pp

As shown in Table 3, both architectures produced high classification performance. MobileNetV2 achieved an accuracy of 97.80%, whereas VGG16 achieved 96.80%. MobileNetV2 therefore classified ten additional testing images correctly. The differences in macro precision, macro recall, and macro F1-score were approximately one percentage point. The small differences among these macro metrics also indicate that the models performed consistently across the three balanced classes.

Both models produced macro ROC–AUC values exceeding 99%. VGG16 achieved 99.67%, while MobileNetV2 achieved 99.88%. These results show that both models had a strong ability to distinguish the three animal categories across different classification thresholds. However, a high ROC–AUC does not mean that every image was correctly classified. ROC–AUC measures class-separation capability based on predicted probabilities, whereas accuracy evaluates final class decisions.

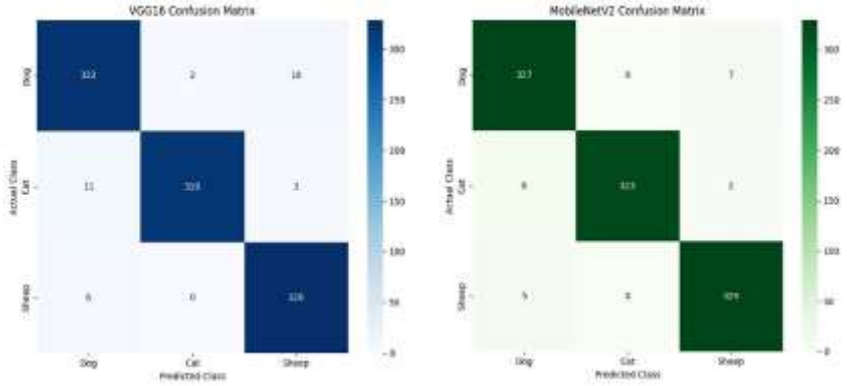

Figure 4. Confusion matrices of VGG16 (a) and MobileNetV2 (b)

The confusion matrices of both models are presented in Figure 4. As shown in Figure 4(a), VGG16 correctly classified 322 of 334 Dog images, 319 of 333 Cat images, and 328 of 334 Sheep images.. The confusion matrices of both models are presented in Figure 4. As shown in Figure 4(a), VGG16 correctly classified 322 of 334 Dog images, 319 of 333 Cat images, and 328 of 334 Sheep images. The model produced 32 incorrect predictions in total. Most errors involved Dog and Cat images, indicating that similarities in fur texture, body shape, pose, and background could influence the classification results.

Figure 4(b) shows that MobileNetV2 correctly classified 327 Dog images, 323 Cat images, and 329 Sheep images, producing 22 incorrect predictions. Compared with VGG16, MobileNetV2 generated five additional correct predictions for Dog, four for Cat, and one for Sheep. Nevertheless, the error patterns of both models were relatively similar. Sheep was the easiest class to classify, whereas several Dog and Cat images remained difficult because of their partially overlapping visual characteristics.

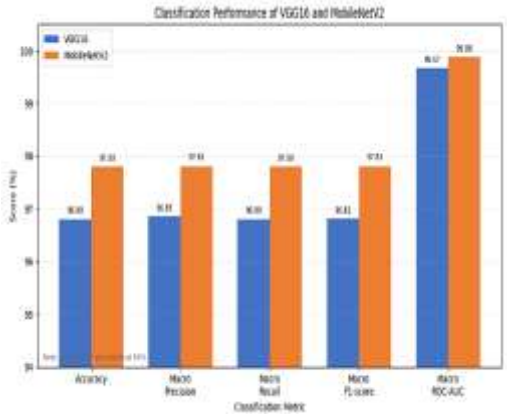

Figure 5. Comparison of classification metrics between VGG16 and MobileNetV2

Figure 5 compares the classification metrics achieved by VGG16 and MobileNetV2. MobileNetV2 obtained numerically higher accuracy, macro precision, macro recall, macro F1-score, and macro ROC–AUC. However, the differences ranged from only 0.21 to 1.00 percentage points. The vertical axis begins at 94% to improve the visibility of these small differences; therefore, the graph should be interpreted together with the exact values in Table 3. Moreover, this descriptive comparison does not demonstrate statistical superiority because the exact McNemar test produced $p = 0.0639$, indicating no statistically significant difference between the classification outcomes of the two models.

ISSN 3026-7315 (media online)
Volume 3, No 2, June 2026, Page: 127−136
DOI: doi.org/10.64366/ijids.v3i2.576
https://journals.adaresearch.or.id/ijids/index

Table 4. Comparison of class-level F1-scores

Class	VGG16	MobileNetV2
Dog	95.69%	97.03%
Cat	97.55%	98.48%
Sheep	97.19%	97.92%
Macro average	96.81%	97.81%

The errors produced by both models frequently involved predictions toward the Dog class and confusion between Dog and Sheep images. VGG16 also showed notable confusion between Cat and Dog. These errors may have been influenced by similarities in fur texture, body position, object scale, and background characteristics. These errors may have been caused by similarities in fur texture, dominant colors, body positions, or particular backgrounds. Some images may also have shown the animal at a relatively small scale, causing the models to receive more visual information from the background than from the primary object. In contrast Cat achieved the highest class-level F1-score for both models, whereas Sheep produced the highest number of correct predictions. These metrics describe different aspects of classification performance and should therefore be interpreted separately. One possible explanation is that the facial characteristics, eyes, ears, and body structure of cats were easier to distinguish from those of the other two classes.

In addition to classification performance, this study measured time and storage efficiency. The efficiency comparison is presented in Table 5. MobileNetV2 completed training in 96.29 seconds, whereas VGG16 required 283.43 seconds. Based on these total training times, MobileNetV2 reduced the training time by 66.03%. However, the two models completed different numbers of epochs because of early stopping. Therefore, the average time per epoch was used as an additional basis for comparison. MobileNetV2 required 7.41 seconds per epoch, whereas VGG16 required 14.17 seconds, representing a reduction of 47.71%.

Table 5. Efficiency comparison of VGG16 and MobileNetV2

Indicator	VGG16	MobileNetV2	Reduction
Training epochs	20	13	35.00%
Training time (seconds)	283.43	96.29	66.03%
Time per epoch (seconds)	14.17	7.41	47.71%
Inference time for 1,001 images (seconds)	9.2604	3.7446	59.56%
Inference time per image (ms)	9.2511	3.7408	59.56%
Weight file size (MB)	57.73	12.80	77.83%

Table 5 shows a clear efficiency advantage for MobileNetV2. Its training required 96.29 seconds, compared with 283.43 seconds for VGG16. Thus, MobileNetV2 reduced total training time by 66.03%. This reduction was influenced by both the faster average processing time per epoch and the smaller number of epochs completed before early stopping. For this reason, average time per epoch provides an additional architecture-oriented comparison. MobileNetV2 required 7.41 seconds per epoch, whereas VGG16 required 14.17 seconds.

MobileNetV2 also required 3.745 seconds to generate predictions for the complete testing subset, compared with 9.260 seconds for VGG16. The corresponding inference times were 3.741 and 9.251 milliseconds per image. Therefore, MobileNetV2 reduced inference time by 59.56% under the same single-GPU measurement procedure. Its weight file was 12.80 MB, representing a 77.83% reduction compared with the 57.73 MB VGG16 weight file.

3.2 Discussion

The results demonstrate that MobileNetV2 produced numerically higher classification performance than VGG16 across all test metrics. The improvement of approximately one percentage point in accuracy was relatively small but was accompanied by consistent improvements in Macro Precision, Macro Recall, Macro F1-score, and Macro ROC-AUC. This consistency indicates that the improvement was not attributable to only one class. The confusion matrices also showed increases in the number of correct predictions for the dog, cat, and sheep classes. Nevertheless, the exact McNemar test indicated that the improvement was not statistically significant at the 0.05 level. Therefore, the findings should not be interpreted as statistical proof that MobileNetV2 has superior classification capability. A more appropriate interpretation is that MobileNetV2 provides classification performance that is at least comparable to that of VGG16 while offering substantial efficiency advantages.

This finding is relevant to the study conducted by Nazir and Kaleem [6], who applied transfer learning to animal image classification on an edge-based camera-trap platform. Their study reported that MobileNetV2 achieved a precision of 0.95, a recall of 0.96, and an F1-score of 0.95 on one of the datasets used. The present study obtained a Macro Precision of 0.9784, a Macro Recall of 0.9780, and a Macro F1-score of 0.9781. These higher values cannot be directly interpreted as superior performance because of differences in the datasets, number of classes, data distributions, and image-capture environments. The present dataset contained only three balanced classes, whereas camera-trap images generally exhibit greater environmental variation and more complex object quality. Nevertheless, both studies demonstrate that MobileNetV2 can produce high performance while maintaining relatively low computational requirements.

ISSN 3026-7315 (media online)
Volume 3, No 2, June 2026, Page: 127−136
DOI: doi.org/10.64366/ijids.v3i2.576
https://journals.adaresearch.or.id/ijids/index

Tøn et al. [4] reported that the performance of deep neural networks for wildlife classification is affected by image quality, illumination, viewing angle, and environmental conditions. The addition of metadata in their study increased accuracy from 98.4% to 98.9%. This finding indicates that visual information alone may be limited when objects are not clearly visible. The present study used only image information without metadata but still achieved an accuracy of 97.8022% using MobileNetV2. The difference between the studies lies in the types of datasets and information sources employed. The findings of Tøn et al. suggest that the performance achieved in the present study could potentially be improved through metadata integration, particularly if the method is applied to camera traps or animal-monitoring systems in real-world environments.

Battu and Reddy Lakshmi [3] emphasized that animal image classification faces challenges associated with natural environmental conditions and label quality. Their study employed a deep neural network approach to classify noisy camera-trap images. In contrast, the present experiment performed file validation, duplicate detection, and class balancing before training. These procedures produced more controlled experimental conditions but simultaneously limited the generalizability of the findings to more diverse datasets. The high accuracy achieved by both models in this study was therefore likely influenced not only by their architectural capabilities but also by the use of three balanced and relatively distinguishable classes.

The MobileNetV2 results also support the study by Indraswari et al. [11], who used MobileNetV2 for image classification and considered the architecture suitable for implementation on mobile devices. Although their research objects differed from those examined in the present study, both studies demonstrate that MobileNetV2 can use transfer learning to produce effective classification without requiring a large architecture. Differences in performance across datasets demonstrate that model capability remains dependent on object characteristics, the number of classes, preprocessing procedures, and data quality. Therefore, the present findings should be regarded as evidence specific to the three selected Animals-10 classes rather than a guarantee that identical performance will be obtained on every animal image dataset.

VGG16 nevertheless demonstrated strong performance, achieving an accuracy of 96.8032% and a Macro ROC-AUC of 99.6687%. This result is consistent with the use of VGG16 as a feature extractor in various transfer learning studies [6]. VGG16 even achieved a lower best validation loss than MobileNetV2, at 0.062700 compared with 0.084099. However, this lower validation loss did not result in higher test accuracy. This condition may occur because validation loss and test accuracy measure different aspects of model performance. Validation loss considers predicted probabilities on the validation set, whereas test accuracy measures only the proportion of correct classifications on the test set. Differences between the validation and test samples may also cause the relative performance ranking of the two models to change.

The primary advantage of MobileNetV2 was its efficiency. Its inference time was 59.56% lower, and its weight file was 77.83% smaller than that of VGG16. This finding is consistent with the development of lightweight convolutional neural networks, which aims to reduce computational costs without causing a substantial loss of performance [7]. MobileNetV2 uses depthwise separable convolutions, inverted residuals, and linear bottlenecks, enabling convolutional operations to be performed more efficiently. In contrast, VGG16 employs a sequence of standard convolutions with large feature maps, resulting in greater computational and parameter-storage requirements.

The practical implication of these results is that MobileNetV2 is more suitable for animal image classification applications that require rapid responses or operate on devices with limited capacity. Possible applications include camera-trap devices, mobile applications, livestock-monitoring systems, and edge devices that process images without continuously transmitting them to a server. This efficiency may reduce processing time, storage requirements, bandwidth consumption, and delays in obtaining classification results. The present findings are also consistent with studies of deep learning-based animal monitoring that identify model efficiency as an important requirement for real-world deployment [8], [9]

The McNemar (p)-value of 0.063915 provides important context for interpreting the difference in accuracy between the two models. Although MobileNetV2 corrected 17 predictions that were incorrect for VGG16 and lost only seven predictions that VGG16 classified correctly, the number of discordant pairs was insufficient to cross the statistical significance threshold of 0.05. This result does not mean that the two models are identical; instead, it indicates that the evidence obtained from the test set was not sufficiently strong to establish a statistically significant difference in classification performance. Therefore, the recommendation to use MobileNetV2 is more strongly supported by its combination of numerical performance and computational efficiency than by its accuracy difference alone.

The contribution of this study lies in the implementation of a controlled comparison protocol. Both models used identical training, validation, and test images; the same classification head; equivalent training configurations; and inference evaluations performed on the same device. In addition to classification metrics, this study measured training time, inference time, and weight file size and subsequently verified differences in the paired predictions using the exact McNemar test. This approach complements previous studies that generally emphasized accuracy without necessarily testing the statistical significance of performance differences or simultaneously evaluating computational costs.

This study was limited to three classes from the Animals-10 dataset and a single data split generated using a random seed of 42. Class balancing improved the fairness of the evaluation but reduced the number of dog images

ISSN 3026-7315 (media online)
Volume 3, No 2, June 2026, Page: 127−136
DOI: doi.org/10.64366/ijids.v3i2.576
https://journals.adaresearch.or.id/ijids/index

used and did not represent the imbalanced class distributions that may occur under real-world conditions. SHA-256 examination could identify identical files but could not detect every possible near-duplicate image resulting from resizing, compression, or cropping. Furthermore, the computational times were obtained in the Kaggle environment and may have been affected by the hardware configuration and system workload at the time of the experiment.

Future studies could expand the number of classes, use multiple random seeds or stratified cross-validation, and group near-duplicate images before dividing the dataset. Testing on edge devices is also required to confirm the advantages of MobileNetV2 under real-world implementation conditions. Grad-CAM analysis could be employed to verify that the models make decisions based on relevant animal body regions rather than background patterns. Considering classification performance, computational efficiency, and statistical testing simultaneously, MobileNetV2 represents a more proportionate choice for developing lightweight and responsive animal image classification systems.

4. CONCLUSION

This study compared the classification performance and computational efficiency of VGG16 and MobileNetV2 using transfer learning for the classification of Dog, Cat, and Sheep images selected from the Animals-10 dataset. Following class balancing, 5,004 images were used, consisting of 1,668 images per class. Both models were trained and evaluated using identical data partitions, classification heads, and training configurations. MobileNetV2 achieved an accuracy of 97.80%, a macro F1-score of 97.81%, and a macro ROC–AUC of 99.88%, whereas VGG16 achieved 96.80%, 96.81%, and 99.67%, respectively. Although MobileNetV2 produced numerically higher classification metrics, the exact McNemar test yielded a p-value of 0.0639. Therefore, the difference in their classification outcomes was not statistically significant at the 5% significance level. The results indicate that MobileNetV2 provided classification performance comparable to that of VGG16 rather than statistically superior performance. The principal advantage of MobileNetV2 was its computational efficiency. MobileNetV2 reduced total training time by 66.03%, inference time by 59.56%, and model weight size by 77.83% compared with VGG16. These reductions demonstrate that a lightweight architecture can maintain high classification performance without requiring the computational and storage resources associated with a larger architecture. Consequently, MobileNetV2 represents a more practical choice for animal image-classification applications with limited processing capacity, storage, or response-time requirements. The findings are limited to three balanced Animals-10 classes, a single stratified data partition, frozen convolutional bases, and the specified experimental environment. Future studies should evaluate all Animals-10 classes, employ repeated stratified splits or cross-validation, and investigate partial fine-tuning to examine the stability and generalisability of the findings. Additional experiments on edge or mobile devices are also recommended to evaluate memory consumption, energy use, inference latency, and real-world deployment performance.

5. DECLARATIONS

5.1 Author Contributions

Conceptualization: T.D., H.A., and O;
Methodology: T.D., H.A., and O;
Software: T.D., H.A., and O;
Validation: T.D., H.A., and O;
Formal Analysis: T.D., H.A., and O;
Investigation: T.D., H.A., and O;
Resources: T.D., H.A., and O;
Data Curation: T.D., and O;
Writing Original Draft Preparation: T.D.;
Writing Review and Editing: T.D., H.A., and O;
Visualization: T.D., and H.A.;
All authors have read and agreed to the published version of the manuscript.

5.2 Data Availability

The dataset used in this study are available from the corresponding author upon reasonable request.

5.3 Funding

This research received no external funding. The publication and manuscript processing fees associated with this article, if applicable, will be covered by the authors.

5.4 Institutional Review Board Statement

Not applicable. This study used a publicly available secondary dataset and did not involve human participants, clinical interventions, or the collection of personally identifiable information.

ISSN 3026-7315 (media online)
Volume 3, No 2, June 2026, Page: 127−136
DOI: doi.org/10.64366/ijids.v3i2.576
https://journals.adaresearch.or.id/ijids/index

5.5 Informed Consent Statement

Not applicable. This study did not involve direct participation of human subjects or the collection of informed consent data.

5.6 Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

5.7 Generative AI and AI-assisted Technologies in the Writing Process

During the preparation of this manuscript, the authors used generative AI tools to assist in improving language clarity, grammatical accuracy, and the organization of the manuscript. The authors critically reviewed, verified, and edited all AI-assisted content and take full responsibility for the accuracy, integrity, and final content of the published work.

REFERENCES

[1] R. A. Rajagukguk *et al.*, "Deep learning for visual animal monitoring (detection, tracking, pose estimation, and behavior classification): A comprehensive review," *Elsevier B.V.*, 2025, Dec, doi: 10.1016/j.atech.2025.101539.
[2] E. Fazzari, D. Romano, F. Falchi, and C. Stefanini, "Animal behavior analysis methods using deep learning: A survey," *Expert Syst. Appl.*, vol. 289, Sep. 2025, doi: 10.1016/j.eswa.2025.128330.
[3] T. Battu and D. S. Reddy Lakshmi, "Animal image identification and classification using deep neural networks techniques," *Measurement: Sensors*, vol. 25, Feb. 2023, doi: 10.1016/j.measen.2022.100611.
[4] A. Tøn, A. Ahmed, A. S. Imran, M. Ullah, and R. M. A. Azad, "Metadata augmented deep neural networks for wild animal classification," *Ecol. Inform.*, vol. 83, Nov. 2024, doi: 10.1016/j.ecoinf.2024.102805.
[5] E. González Fernández, A. L. Sandoval Orozco, and L. J. García Villalba, "A multi-channel approach for detecting tampering in colour filter images," *Expert Syst. Appl.*, vol. 230, Nov. 2023, doi: 10.1016/j.eswa.2023.120498.
[6] S. Nazir and M. Kaleem, "Object classification and visualization with edge artificial intelligence for a customized camera trap platform," *Ecol. Inform.*, vol. 79, Mar. 2024, doi: 10.1016/j.ecoinf.2023.102453.
[7] J. Saetiew *et al.*, "Automated chick gender determination using optical coherence tomography and deep learning," *Poult. Sci.*, vol. 104, no. 5, May 2025, doi: 10.1016/j.psj.2025.105033.
[8] S. Kumar and H. Kumar, "Efficient-VGG16: A Novel Ensemble Method for the Classification of COVID-19 X-ray Images in Contrast to Machine and Transfer Learning," in *Procedia Computer Science*, Elsevier B.V., 2024, pp. 1289–1299. doi: 10.1016/j.procs.2024.04.122.
[9] N. T J, "An enhanced deep learning framework for prostate cancer detection using modified VGG16 and LeNet-MobileNetV2 integration," *Results in Engineering*, vol. 27, Sep. 2025, doi: 10.1016/j.rineng.2025.106918.
[10] A. Ali Linkon *et al.*, "Evaluation of Feature Transformation and Machine Learning Models on Early Detection of Diabetes Mellitus," *IEEE Access*, vol. 12, no. October, pp. 165425–165440, 2024, doi: 10.1109/ACCESS.2024.3488743.
[11] R. Indraswari, R. Rokhana, and W. Herulambang, "Melanoma image classification based on MobileNetV2 network," in *Procedia Computer Science*, Elsevier B.V., 2021, pp. 198–207. doi: 10.1016/j.procs.2021.12.132.
[12] O. F. Altal *et al.*, "Hybrid attention-enhanced MobileNetV2 with particle swarm optimization for endometrial cancer classification in CT images," *Inform. Med. Unlocked*, vol. 57, Jan. 2025, doi: 10.1016/j.imu.2025.101662.
[13] A. Rácz and A. Gere, "Comparison of missing value imputation tools for machine learning models based on product development cases studies," *LWT*, vol. 221, Apr. 2025, doi: 10.1016/j.lwt.2025.117585.
[14] J. Wan and B. Yong, "Automatic extraction of surface water based on lightweight convolutional neural network," *Ecotoxicol. Environ. Saf.*, vol. 256, May 2023, doi: 10.1016/j.ecoenv.2023.114843.
[15] M. A. I. Aquil and W. H. W. Ishak, "Evaluation of scratch and pre-trained convolutional neural networks for the classification of tomato plant diseases," *IAES International Journal of Artificial Intelligence*, vol. 10, no. 2, pp. 467–475, Jun. 2021, doi: 10.11591/IJAI.V10.I2.PP467-475.
[16] X. Liu, B. Wang, S. Jin, and Z. Song, "Missing value interpolation algorithm for long-term temperature observation data based on data augmentation multiple interpolation method," *Results in Engineering*, vol. 27, Sep. 2025, doi: 10.1016/j.rineng.2025.106211.
[17] B. Z. Wubineh, L. Jeleń, and A. Rusiecki, "DCGAN-based Cytology Image Augmentation for Cervical Cancer Cell Classification Using Transfer Learning," in *Procedia Computer Science*, Elsevier B.V., 2025, pp. 1003–1011. doi: 10.1016/j.procs.2025.02.206.
[18] Q. Aini, N. Lutfiani, H. Kusumah, and M. S. Zahran, "Deteksi dan Pengenalan Objek Dengan Model Machine Learning: Model Yolo," *CESS (Journal of Computer Engineering, System and Science)*, vol. 6, no. 2, p. 192, 2021, doi: 10.24114/cess.v6i2.25840.
[19] O. A. Montesinos López, A. Montesinos López, and J. Crossa, "Overfitting, Model Tuning, and Evaluation of Prediction Performance," in *Multivariate Statistical Machine Learning Methods for Genomic Prediction*, Springer International Publishing, 2022, pp. 109–139. doi: 10.1007/978-3-030-89010-0_4.
[20] A. Peryanto, A. Yudhana, and R. Umar, "Klasifikasi Citra Menggunakan Convolutional Neural Network dan K Fold Cross Validation", JAIC, vol. 4, no. 1, pp. 45–51, May 2020. doi: doi.org/10.30871/jaic.v4i1.2017