Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9911

Implementation of MobileNetV2-Based Convolutional Neural Network for Melanoma Detection Using Dermoscopic Images

Herizon[1*], Widia Dartika Putri[2], Fisla Wirda[3], Milda Yuliza[4] , Qonita Salma Alfarizy[5]
[1,2,3,4]Politeknik Negeri Padang, Indonesia
[5] Fakultas Kedokteran Universitas Syiah Kuala, Banda Aceh, Indonesia
[1*]herizonjhon1969@gmail.com, [2]widiadartika@gmail.com, [3]wirdafisla@gmail.com, [4]mildayuliza01@yahoo.com, [5]qonitasalmaalfarizy@gmail.com

open access

***Corresponding Author**

Article History:
Submitted: 26-10-2026
Accepted: 08-11-2026
Published: 11-11-2026

Keywords:
CNN; Dermoscopy Images; MobileNetV2; Skin Cancer; Skin Lesion Classification.

Brilliance: Research of Artificial Intelligence is licensed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0).

ABSTRACT

Melanoma is an aggressive form of skin cancer, making early detection essential for the timely identification of potentially malignant skin lesions and the improvement of diagnostic outcomes. This study aims to evaluate a convolutional neural network (CNN) based on the MobileNetV2 architecture for classifying dermoscopic images into benign and malignant categories using the HAM10000 dataset. Image preprocessing involves resizing and normalization, while additional techniques, including contrast-limited adaptive histogram equalization (CLAHE), hair removal, and soft segmentation, are applied to enhance image quality, reduce visual artifacts, and improve lesion visibility. The model is developed using transfer learning with weights pretrained on ImageNet, followed by training of the classification head and fine-tuning of the final network layers to adapt the extracted features to the target classification task. Model performance is assessed using accuracy, loss, precision, recall, and F1-score to provide a comprehensive evaluation of classification performance. Evaluation on a primary test set comprising 659 images yields an accuracy of 85.13%, a loss of 0.3430, and a sensitivity of 88.33% for the malignant class. A separate evaluation conducted on 128 images using Google Colab achieves an accuracy of 87.50%; these results are reported independently of the primary evaluation. The findings indicate that MobileNetV2 has potential for dermoscopic skin lesion classification, particularly in identifying malignant lesions. However, further evaluation using independent datasets, cross-dataset validation, and computational efficiency analysis is necessary to assess the model's generalizability and reliability before considering its potential application in clinical settings.

INTRODUCTION

Skin cancer, particularly melanoma, is a major health concern because of its potential to metastasize to other tissues and organs. Early detection is therefore essential for improving the prospects of timely diagnosis and treatment. Dermoscopy is a non-invasive imaging technique that enables detailed examination of the morphological structures and visual patterns of skin lesions. However, similarities in the visual characteristics of benign and malignant lesions, together with substantial inter-lesion variability, can complicate the diagnostic process. Naseri and Safaei (2025) reviewed the application of machine learning and deep learning techniques for melanoma diagnosis, while Yang et al. (2025) discussed advances in machine learning methods for image-based skin cancer classification. Meedeniya et al. (2024) also reviewed developments in deep learning-based skin cancer detection, including preprocessing, segmentation, classification, and model interpretability. Collectively, these studies highlight the potential of computational approaches to support the analysis of dermoscopic images.

Convolutional neural network (CNN) are widely used in medical image analysis because they can automatically learn hierarchical visual representations for classification tasks. Wu et al. (2022) reviewed the application of deep learning to skin cancer classification, while Shetty et al. (2022) investigated skin lesion classification using machine learning and CNN-based approaches. Ye et al. (2024) conducted a systematic review and meta-analysis of deep learning algorithms for melanoma detection using dermoscopic images. In addition, Mehmood et al. (2023) developed SBXception, a modified version of the Xception architecture designed with a shallower and wider network structure for efficient skin lesion classification. Although these approaches have demonstrated promising potential, their performance must be assessed in relation to dataset characteristics, classification accuracy, and their ability to correctly identify melanoma cases.

Computational efficiency is crucial for CNN deployment on resource-constrained devices. Öznacar and Kayapunar (2025) demonstrated the applicability of MobileNetV2 for skin cancer classification, highlighting its potential to reduce computational complexity while maintaining classification performance. Indraswari et al. (2022)

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9911

applied MobileNetV2 to melanoma image classification, while Dhuhita et al. (2023) and Jain et al. (2021) investigated MobileNetV2-based approaches and transfer learning for skin cancer classification. Nirupama and Virupakshappa (2025) further extended MobileNetV2 by incorporating attention mechanisms and multiscale features for skin disease classification. Shinde et al. (2023) developed Squeeze-MNet, a MobileNet-based model using transfer learning, and evaluated its prototype on a Raspberry Pi 4. These studies demonstrate the relevance of lightweight CNN architectures for applications involving constrained computing environments. Nevertheless, model selection requires a balanced assessment of classification performance, computational requirements, and inference latency.

Class imbalance presents an additional challenge in melanoma detection because models trained on imbalanced datasets may become biased toward majority classes, potentially reducing their ability to identify minority-class cases. Alam et al. (2022) investigated skin cancer classification using imbalanced datasets, while Hu et al. (2024) examined multiscale feature fusion in conjunction with class-weighted loss functions. Meedeniya et al. (2024) also highlighted data limitations and model evaluation challenges as important considerations in the development of skin cancer detection systems. These findings emphasize the need to account for class distribution and adopt appropriate learning strategies to improve the reliable identification of malignant lesions.

Regarding dataset selection, Tajerian et al. (2023) employed the HAM10000 dataset, which comprises 10,015 dermoscopic images representing seven diagnostic categories of pigmented skin lesions. This dataset provides a relevant benchmark for evaluating deep learning approaches in skin lesion classification. For the present study, the images are grouped into benign and malignant classes according to predefined diagnostic-label mapping criteria. During preprocessing, the images are resized to 224 × 224 pixels, and pixel values are normalized using the preprocessing procedure required by MobileNetV2. The classification model is developed through transfer learning using ImageNet-pretrained weights, followed by fine-tuning to adapt the network to the target classification task. Model performance is evaluated using accuracy, precision, recall (sensitivity), specificity, F1-score, a confusion matrix, and the error rate.

The primary objective of this study is to implement and evaluate a MobileNetV2-based CNN for melanoma detection using dermoscopic images. In addition to evaluating classification performance, the model is deployed on a Raspberry Pi equipped with a webcam to investigate its operational feasibility in a resource-constrained environment. The results are expected to provide insights into the classification capabilities of MobileNetV2 and its potential for deployment in lightweight, dermoscopic image-based melanoma detection systems.

LITERATURE REVIEW

Research on deep learning-based skin lesion classification has advanced through the development and application of diverse convolutional neural network (CNN) architectures, datasets, and image preprocessing strategies. In a systematic review of machine learning and deep learning approaches for skin lesion classification and diagnosis, Kassem et al. (2021) highlighted the widespread adoption of these techniques in computer-aided diagnosis systems. Similarly, Debelee (2023) examined machine learning methods for skin lesion classification and detection, discussing challenges associated with data availability, class imbalance, image segmentation, and classification. Collectively, these studies suggest that the performance of skin lesion classification systems depends not only on the selected CNN architecture but also on dataset characteristics and preprocessing procedures.

In the context of dermoscopic image classification, Shetty et al. (2022) compared machine learning methods and CNN-based approaches for classifying skin lesions using the HAM10000 dataset. Their study reported a maximum accuracy of 95.18% for CNN-based classification across seven skin lesion categories, demonstrating the potential of CNNs to automatically learn discriminative visual features. Furthermore, Mehmood et al. (2023) proposed SBXception, a modified version of the Xception architecture featuring a shallower and wider network structure for efficient skin lesion classification. Meanwhile, Shrivastava et al. (2021) investigated multiclass skin lesion classification using image augmentation and transfer learning. These studies highlight the importance of selecting appropriate network architectures and learning strategies. However, classification performance must also be assessed in relation to class distribution, validation protocols, and the model's ability to identify minority classes. Consequently, accuracy alone is insufficient to establish the effectiveness of a classification model, particularly when the target classes are unevenly represented.

Subsequent research has increasingly explored lightweight CNN architectures to reduce computational requirements while maintaining classification performance. Öznacar and Kayapunar (2025) demonstrated the applicability of MobileNetV2 for skin cancer classification and efficient feature extraction. Indraswari et al. (2022) applied MobileNetV2 to classify melanoma images into benign and malignant categories, while Dhuhita et al. (2023) investigated its application to skin cancer detection. These studies demonstrate the relevance of MobileNetV2 as a lightweight alternative for image-based skin lesion classification. Nevertheless, meaningful comparisons between models require consideration of differences in datasets, class definitions, preprocessing methods, and experimental configurations.

The applicability of lightweight architectures has also been investigated under different image acquisition conditions and hardware configurations. In their MobileSkin study, Yilmaz et al. (2022) evaluated skin lesion classification using images acquired with a handheld dermatoscope connected to a mobile phone, illustrating an

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9911

approach relevant to portable diagnostic support systems. Wang et al. (2023) proposed an enhanced MobileNetV2-based approach incorporating generative adversarial networks for skin lesion classification. In addition, Shinde et al. (2023) developed Squeeze-MNet, a MobileNet-based model that combines hair-removal preprocessing with transfer learning, and evaluated its prototype on a Raspberry Pi 4. These studies indicate that lightweight CNN models can be adapted for deployment on devices with limited computational resources. However, the effectiveness of architectural modifications and their practical suitability for resource-constrained environments must be established through systematic empirical evaluation.

In addition to architectural considerations, class imbalance remains a significant challenge in skin lesion classification. Alam et al. (2022) investigated deep learning-based skin cancer classification using imbalanced datasets, while Hu et al. (2024) examined multiscale feature fusion combined with class-weighted loss functions. These studies highlight the importance of incorporating class distribution into model development and optimization. In melanoma detection, inadequate representation of malignant cases may impair the model's ability to identify the target class, even when overall accuracy appears satisfactory. Therefore, model evaluation should include sensitivity, precision, specificity, and the F1-score alongside accuracy to provide a more comprehensive assessment of classification performance, particularly for the melanoma class.

At a broader level, Wu et al. (2022) discussed challenges in developing deep learning-based skin cancer classification systems, including data limitations, class imbalance, generalization, and computational efficiency. Similarly, Ye et al. (2024) conducted a systematic review and meta-analysis of deep learning algorithms for melanoma detection using dermoscopic images. Collectively, these studies emphasize that model evaluation should extend beyond conventional classification metrics to consider robustness across datasets and testing conditions. In particular, performance on a single dataset may not adequately reflect a model's generalizability to images acquired under different conditions or from different data sources.

Based on this synthesis, the present study employs a MobileNetV2-based CNN with transfer learning for melanoma detection using dermoscopic images. The study evaluates classification performance using multiple metrics rather than accuracy alone and investigates the feasibility of deploying the model on a Raspberry Pi 4 equipped with a webcam. By considering both diagnostic classification performance and implementation on resource-constrained hardware, this study aims to contribute to the development of computationally efficient approaches for dermoscopic image-based melanoma detection.

METHOD

Research Design

This study developed a skin lesion classification system using a fine-tuned MobileNetV2 model and webcam-based image acquisition. The system consists of two parallel pathways: a classification pathway for predicting benign and malignant lesions, and a visualization pathway for displaying the lesion region. The overall system workflow is illustrated in Figure 1.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, Number 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9911

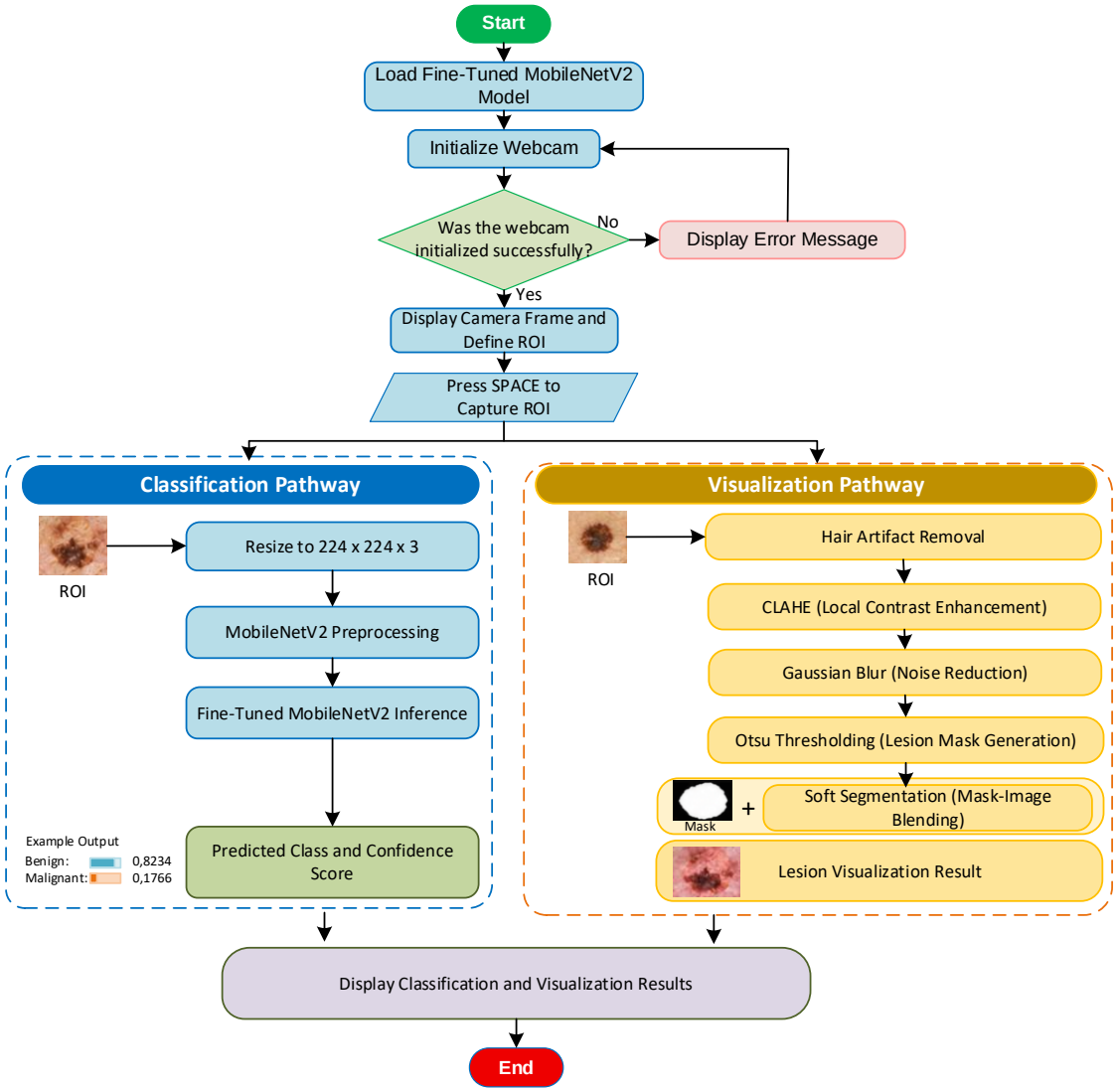


Figure 1. Flowchart of the MobileNetV2-Based Skin Lesion Classification and Visualization System

Figure 1 illustrates the workflow of the proposed system, beginning with the loading of the fine-tuned MobileNetV2 model and webcam initialization. After successful initialization, the system displays the camera frame and region of interest (ROI). When the SPACE key is pressed, the ROI is processed through two parallel pathways. The classification pathway resizes and normalizes the image before performing inference using the fine-tuned MobileNetV2 model to predict the benign or malignant class and its associated confidence score. Meanwhile, the visualization pathway applies hair artifact removal, CLAHE, Gaussian blur, and Otsu thresholding to generate a lesion mask, which is subsequently combined with the original image through soft segmentation. The classification and visualization results are then displayed in the system. The segmentation output serves solely as a visual aid and is not directly used as input to the classification model.

Dataset and Data Partitioning

This study uses the HAM10000 (Human Against Machine with 10,000 training images) dataset, which comprises 10,015 RGB dermoscopic images in JPEG format with a resolution of 600 × 450 pixels. For this study, the images are grouped into two classes: benign and malignant. Of the initial 2,637 images, comprising 1,440 benign and 1,197 malignant images, the development dataset was balanced by reducing the number of benign images to 1,197. This resulted in a balanced dataset of 2,394 images, with 1,197 images in each class.

The balanced dataset was subsequently divided into training and validation sets using an 80:20 split, yielding 1,916 training images and 478 validation images. An independent test set comprising 659 images was kept separate from the training and validation processes and used exclusively for the final model evaluation. This test set contained 359 benign images and 300 malignant images, as presented in Table 1. By keeping the test set separate from model

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9911

development, the evaluation was designed to measure classification performance on images that had not been used to update the model weights during training.

Tabel 1. Dataset Splitting

Step	Benign	Malignant	Total
Before balancing	1.440	1.197	2.637
Afterbalancing	1.197	1.197	2.394
Training	-	-	1.916
Validation	-	-	478
Testing	300	359	659

Table 1 summarizes the dataset distribution and partitioning process. The initial dataset of 2,637 images was balanced by reducing the benign class to 1,197 images, resulting in 2,394 images with equal class representation. The balanced dataset was divided into training (1,916 images) and validation (478 images) sets using an 80:20 ratio. An independent test set of 659 images, comprising 359 benign and 300 malignant images, was reserved exclusively for final evaluation. This separation helps prevent data leakage and provides a more reliable assessment of the model's classification performance.

Image Preprocessing and Augmentation

Before classification, all images undergo preprocessing to meet the input requirements of the MobileNetV2 architecture. The dermoscopic images are resized to 224 × 224 pixels while retaining their three RGB color channels, resulting in an input tensor of dimensions 224 × 224 × 3. Pixel values are then normalized from the range [0, 255] to [0, 1] by dividing each pixel value by 255. This procedure standardizes the input data and facilitates consistent numerical scaling during model training.

To increase data diversity and improve the model's ability to generalize, data augmentation is applied exclusively to the training set. The augmentation operations include image rotation, zooming, and horizontal flipping. These transformations introduce variations in image orientation and scale while preserving the relevant visual characteristics of skin lesions. The validation and test sets are not augmented, allowing model performance to be assessed on images that have not undergone these augmentation transformations.

MobileNetV2 Architecture

MobileNetV2 is employed as the convolutional neural network (CNN) backbone to extract visual features from dermoscopic images. The model is initialized with weights pretrained on ImageNet, and its original classification head is removed by setting include_top=False. In this configuration, MobileNetV2 serves as a feature extractor, while a custom classification head is added to distinguish between the two target classes: benign and malignant.

Each preprocessed image, with dimensions of 224 × 224 × 3, is passed through the convolutional layers and inverted residual blocks of MobileNetV2. The resulting feature maps have dimensions of 7 × 7 × 1280. Global average pooling (GAP) is then applied to aggregate the spatial feature maps into a 1,280-dimensional feature vector, reducing the need for a large number of parameters associated with a conventional flattening operation followed by fully connected layers.

The resulting feature vector is passed through a dropout layer to reduce the risk of overfitting. Subsequently, a dense output layer with a sigmoid activation function generates a probability score between 0 and 1 for binary classification. The predicted class is determined by applying a predefined decision threshold to this score. The threshold and the interpretation of the output probability depend on the class-label encoding used during model training.

Model Training and Fine-Tuning

Model development consists of two stages: classification-head training and fine-tuning. During the first stage, the pretrained MobileNetV2 backbone is frozen, and only the newly added classification head is trained to distinguish between benign and malignant lesions. Training is performed using 1,916 training images and 478 validation images, with a batch size of 32, the Adam optimizer, binary cross-entropy loss, and a maximum of 10 epochs.

During the second stage, selected layers of the MobileNetV2 backbone are unfrozen to enable fine-tuning. Specifically, the trainable section extends from block_14_expand to out_relu, encompassing blocks 14–16 and Conv_1. Batch normalization layers remain frozen to help maintain the pretrained feature statistics, while the unfrozen backbone layers and classification head are optimized jointly. Fine-tuning uses the Adam optimizer with a learning rate of 1×10^{-5}, a batch size of 32, and binary cross-entropy loss. The same training and validation sets are used in both stages.

An independent test set comprising 659 images is excluded from model training, fine-tuning, and model-selection decisions and is reserved for final evaluation. During training, the model's learning behavior is monitored using training accuracy, validation accuracy, training loss, and validation loss. These metrics are used to assess convergence and identify potential overfitting. The independent test set is evaluated only after the model configuration has been finalized.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9911

Visual Segmentation

Visual segmentation is performed to improve the visualization of skin lesions within a region of interest (ROI). The procedure includes hair removal to reduce artifacts caused by hair, contrast-limited adaptive histogram equalization (CLAHE) to enhance local contrast, Gaussian blurring to reduce high-frequency noise, and Otsu thresholding to generate a binary mask of the lesion region. The resulting mask is then combined with the original image using a soft-segmentation procedure to emphasize the selected region while retaining relevant visual information.

Hair removal and ROI-based processing have been explored in previous studies of skin lesion image analysis. In the present study, however, the segmentation procedure is used for supplementary visualization rather than as a direct input to the MobileNetV2 classification model. The classifier receives images processed through resizing and pixel-value normalization, whereas the outputs of hair removal, CLAHE, Gaussian blurring, Otsu thresholding, and soft segmentation are presented separately to illustrate the intermediate image-processing results. Consequently, these visualization procedures are not considered components of the classifier's input pipeline, and their individual effects on classification performance are not evaluated in this study.

Evaluation and Testing

Model evaluation was conducted to assess the ability of the proposed MobileNetV2-based CNN to classify dermoscopic images into benign and malignant categories. The primary evaluation used a designated test set that was kept separate from the training and validation data. Classification performance was assessed using accuracy, precision, recall (sensitivity), the F1-score, a confusion matrix, and the error rate. Accuracy measures the proportion of correctly classified samples across the test set, whereas precision, recall, and the F1-score provide complementary measures of classification performance. The confusion matrix summarizes the numbers of correct and incorrect predictions for each class, while the error rate represents the proportion of incorrectly classified samples.

In addition to the primary evaluation, a supplementary test involving 128 images was conducted to examine the model's classification performance on an additional image set. The results of this supplementary evaluation were analyzed separately from those of the primary test set. Further testing was performed on a Raspberry Pi equipped with a webcam to investigate the feasibility of deploying the model in a resource-constrained hardware environment. During this stage, images were captured in real time and processed by the system to generate classification predictions. This implementation was intended to assess the system's operational feasibility under practical image-acquisition conditions.

The evaluation metrics were calculated from the confusion matrix using four fundamental quantities: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). For binary classification, accuracy, precision, recall, and the F1-score are calculated as follows:

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

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

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

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

The error rate is calculated as the proportion of incorrect predictions relative to the total number of evaluated samples:

$$Error\ rate = \frac{FP+FN}{TP+TN+FP+FN} x100\% \quad (5)$$

For this study, melanoma is treated as the positive class when calculating binary classification metrics. Accordingly, recall corresponds to melanoma sensitivity, while specificity is calculated as (TN/(TN+FP)) and measures the proportion of benign cases correctly identified as negative. These metrics provide a more comprehensive assessment of the model's ability to distinguish malignant from benign lesions.

Following this procedure, the evaluation was conducted under three conditions, including the use of testing data and an additional set of 128 images. The results from each test serve as the basis for analyzing the model's performance in the "Results and Discussion" section.

RESULT

Testing was performed to evaluate the model's ability to classify images that were not used for weight updates during the training process. The testing dataset consists of 659 images, comprising 359 benign images and 300 malignant images, as shown in Table 2. Model performance was evaluated using test accuracy, test loss, a confusion matrix, and a classification report to provide an overview of the classification performance, both overall and for each individual class.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9911

Tabel 2. Evaluation results on the testing data

Parameters	Result
Number of test data	659 images
Accuracy Test	85,13%
Test Losses	0,3430

As shown in Table 2, the fine-tuned MobileNetV2 model achieved a test accuracy of 85.13% and a test loss of 0.3430. The accuracy indicates that the model correctly classified 561 of the 659 test images, while the remaining 98 images were misclassified. The test loss quantifies the discrepancy between the model's predicted probabilities and the ground-truth labels, as measured by the loss function used during evaluation. To further investigate the model's classification performance and identify patterns of misclassification, the test results were subsequently analyzed using a confusion matrix and a classification report, as shown in Figure 2.

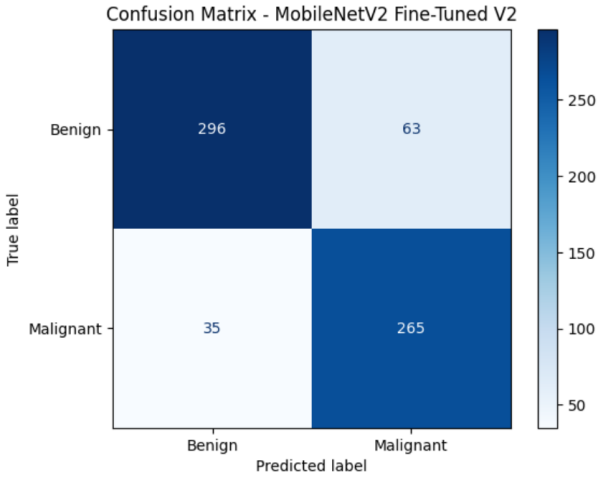


Figure 2. Confusion Matrix MobileNetV2

As shown in Fig. 2, of the 359 images labeled as benign, 296 were correctly classified as benign, whereas 63 were misclassified as malignant. Among the 300 malignant images, 265 were correctly classified as malignant, while the remaining 35 were misclassified as benign. These results indicate that the model produces classification errors in both classes, with a greater number of false-positive predictions (63 cases) than false-negative predictions (35 cases). This pattern suggests that the model is more likely to misclassify benign lesions as malignant than to misclassify malignant lesions as benign. Nevertheless, the clinical significance of these errors should be interpreted in conjunction with class-specific evaluation metrics, particularly sensitivity and specificity.

Tabel 3. Classification Report Fine Tuning

Class	Precision	Recall	F1-score	Support
Benign	0,8943	0,8245	0,8580	359
Malignant	0,8079	0,8833	0,8439	300
Accuracy	-	-	0,8513	659
Macro Average	0,8511	0,8539	0,8510	659
Weighted Average	0,8550	0,8513	0,8516	659

As shown in Table 3, the classification report indicates that the benign class achieved a precision of 89.43%, a recall of 82.45%, and an F1-score of 85.80%. In comparison, the malignant class achieved a precision of 80.79%, a recall of 88.33%, and an F1-score of 84.39%. The recall of 88.33% for the malignant class indicates that the model correctly identified 88.33% of the malignant images in the test set. However, 11.67% of the malignant images were incorrectly classified as benign. Conversely, the higher precision for the benign class indicates that a greater proportion of predictions labeled as benign were correct compared with predictions labeled as malignant.

Overall, the model achieved a macro-average F1-score of 85.10% and a weighted-average F1-score of 85.16%. The small difference between these values suggests that the model's overall F1-score is relatively consistent when the two classes are weighted equally and when their respective sample sizes are taken into account. Nevertheless, the differences in precision and recall between the two classes indicate that classification performance remains uneven. In particular, although the model demonstrates relatively high sensitivity to malignant lesions, some malignant cases are still misclassified as benign. Therefore, future model development should prioritize improving melanoma sensitivity by reducing false-negative predictions, while maintaining adequate specificity for benign lesions.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9911

Additional Testing Using Google Colab

Following the fine-tuning process, an additional evaluation was conducted using Google Colab to assess the model's ability to classify dermoscopic images into two categories: benign and malignant. This evaluation used 128 images, comprising 64 benign and 64 malignant images, as shown in Fig. 3 and 4. The purpose of this additional evaluation was to assess the model's classification performance on this separate set of images.

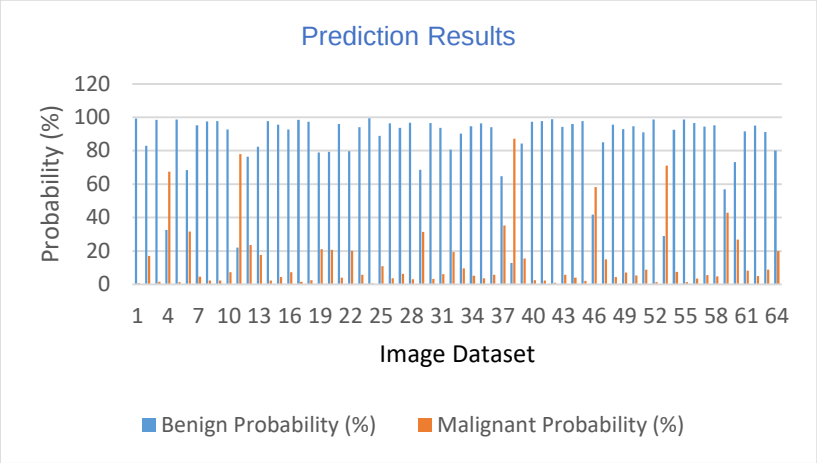


Figure 3. Results of testing benign data using Google Colab

As shown in Fig. 3, the model's classification performance was evaluated using 64 benign images, 59 images were correctly classified as benign, while 5 images were incorrectly classified as malignant. Consequently, the classification success rate for the benign image group reached 92.19%, with an error rate of 7.81%. The calculation for the classification success rate for benign images is as follows:

$$Benign\ Accuracy\ = \frac{59}{64}\ x\ 100\%$$

Benign Accuracy = 92,19%

$$Error = \frac{5}{64}\ x\ 100\%$$

Error = 7,81%

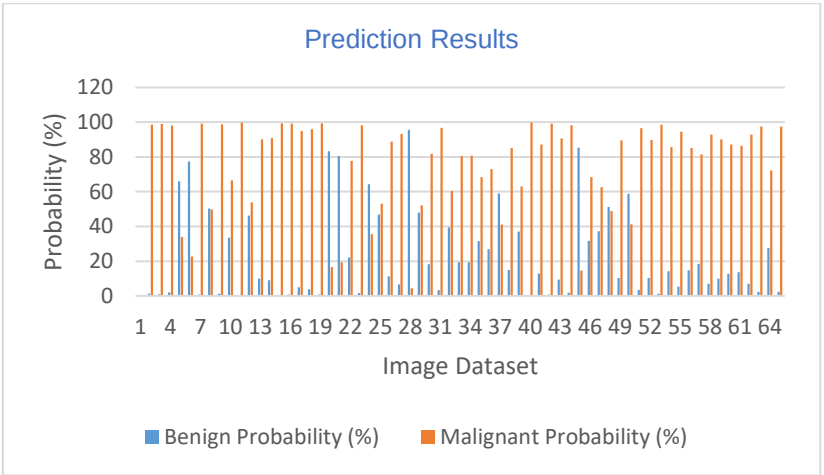


Figure 4. Results of testing on malignant data using Google Colab

As shown in Fig. 4, the model's classification performance was evaluated using 64 malignant images, 53 images were correctly classified as malignant, while the remaining 11 images were incorrectly classified as benign. These results indicate a classification success rate of 82.81% and a misclassification rate of 17.19%. The calculation for the classification success rate for malignant images is as follows:

$$Malignant\ Accuracy\ = \frac{53}{64}\ x\ 100\%$$

Malignant Accuracy = 82,81%

$$Error\ = \frac{11}{64}\ x\ 100\%$$

Error = 17,19%

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9911

Accordingly, the model achieved an overall accuracy of 87.50% and an error rate of 12.50%. The results indicate that the model performed better in classifying benign images than malignant images. Specifically, 11 malignant images were incorrectly classified as benign, compared with five benign images that were incorrectly classified as malignant. These findings suggest that the model's ability to identify malignant lesions requires further improvement, particularly in reducing false-negative predictions. Such errors are especially important in melanoma detection because malignant lesions misclassified as benign may go undetected by the system. Therefore, future improvements should prioritize enhancing sensitivity to malignant lesions while maintaining satisfactory classification performance for benign lesions.

DISCUSSION

The evaluation results show that the fine-tuned MobileNetV2 model achieved a test accuracy of 85.13% and a test loss of 0.3430 on a test set comprising 659 images, including 359 benign and 300 malignant cases. These results demonstrate the model's ability to distinguish between benign and malignant skin lesions. However, overall accuracy alone does not fully characterize classification performance, particularly in a diagnostic context where the ability to identify malignant lesions is of critical importance. Therefore, the results should be interpreted in conjunction with class-specific metrics and the confusion matrix.

The confusion matrix shows that the model correctly classified 296 of the 359 benign images, while the remaining 63 were misclassified as malignant. Among the 300 malignant images, 265 were correctly classified, whereas 35 were misclassified as benign. Thus, the model correctly classified 561 images and misclassified 98, consistent with the reported test accuracy. The greater number of false-positive predictions than false-negative predictions indicates that benign lesions were more frequently classified as malignant than malignant lesions were classified as benign. Although false-positive predictions may lead to unnecessary follow-up examinations, false-negative predictions are particularly concerning in melanoma detection because they may result in malignant lesions being overlooked.

The class-specific metrics provide further insight into the model's performance. The recall for the malignant class was 88.33%, compared with 82.45% for the benign class. This means that the model correctly identified 88.33% of the malignant images in the test set, while 11.67% were incorrectly classified as benign. In a melanoma detection context, these false-negative predictions warrant particular attention because they may contribute to missed identification of potentially malignant lesions. Conversely, the benign class achieved a precision of 89.43%, exceeding the 80.79% precision obtained for the malignant class. This indicates that predictions labeled as benign were more likely to be correct than predictions labeled as malignant. The F1-scores were 85.80% for the benign class and 84.39% for the malignant class, suggesting relatively comparable performance in balancing precision and recall across the two classes. Similarly, the macro-average and weighted-average F1-scores were 85.10% and 85.16%, respectively. Nevertheless, these aggregate metrics should not obscure differences in class-specific performance. In particular, improving sensitivity to malignant lesions while maintaining adequate specificity remains an important direction for further model development.

The preprocessing pipeline includes image resizing and normalization, as well as contrast-limited adaptive histogram equalization (CLAHE), hair removal, and soft segmentation for image enhancement and visualization. CLAHE is intended to improve local contrast, while hair removal aims to reduce visual artifacts caused by hair on the skin surface. Soft segmentation is used to emphasize the lesion region; however, its output is not supplied as a separate input to the classification model. These procedures are intended to facilitate image inspection and address certain image-quality variations. Nevertheless, the overall evaluation results do not establish the individual contribution of each preprocessing operation to classification performance. Ablation studies are therefore needed to determine whether these operations provide measurable improvements in classification accuracy, sensitivity, specificity, or other relevant metrics.

MobileNetV2 was selected as the CNN backbone because it was designed to support computationally efficient image recognition. Its inverted residual blocks and linear bottlenecks enable feature extraction with a relatively lightweight architecture. In this study, fine-tuning was performed by unfreezing selected layers near the end of the pretrained network and applying a learning rate of (1 \times 10^{-5}). This approach allows the trainable layers to adapt their feature representations to dermoscopic images while retaining information learned during pretraining. However, the use of a lightweight architecture does not, by itself, demonstrate that the complete system achieves a particular level of computational efficiency. If efficiency is a key objective, further evaluation should include inference latency, model size, parameter count, memory consumption, and, where available, floating-point operations (FLOPs). Such measurements would provide a more comprehensive assessment of the model's suitability for deployment on resource-constrained devices.

Additional testing was conducted using Google Colab on a set of 128 images, comprising 64 benign and 64 malignant images. The model correctly classified 59 benign images and 53 malignant images, resulting in 112 correct predictions and 16 misclassifications. Accordingly, the overall accuracy was 87.50%. The class-specific classification rates were 92.19% for benign images and 82.81% for malignant images. Specifically, 11 malignant images were incorrectly classified as benign, whereas five benign images were incorrectly classified as malignant. These results indicate that false-negative predictions for the malignant class remain a concern. However, the accuracy obtained from

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9911

this additional evaluation should not be interpreted as an improvement over the 85.13% accuracy achieved in the primary evaluation because the two results were obtained from different test sets. Furthermore, the independence of the additional test images from the training, validation, and primary test sets must be verified before these results can be used to support claims regarding the model's generalization capability.

When comparing the proposed model with previous studies, differences in datasets, network architectures, class configurations, training strategies, preprocessing procedures, and data-partitioning protocols must be considered. Therefore, the test accuracy of 85.13% obtained in this study is not sufficient, on its own, to establish that the model outperforms or performs comparably to previously reported methods. A meaningful comparison requires careful consideration of methodological differences, evaluation metrics, and test-set characteristics. Where possible, comparative experiments should evaluate competing models under the same dataset partitions and experimental conditions. In addition, all references used to support comparisons with previous studies should be verified against the original publications to ensure that the reported findings and interpretations are accurate.

This study has several limitations. First, the use of a public dataset may not adequately represent the diversity of images encountered under real-world acquisition conditions, particularly when images are captured using different devices or under varying illumination conditions. Second, the available evaluation data are insufficient to establish the model's generalization capability across independent datasets and acquisition environments. Third, the misclassification of malignant images as benign highlights the need to improve sensitivity to the malignant class. Fourth, the individual contributions of the preprocessing procedures to classification performance have not been established through ablation studies. Finally, the computational efficiency of the complete system remains to be quantified using measurements such as inference time, model size, and memory consumption.

Future research should address these limitations through external validation using independent datasets, evaluation of images captured with the intended acquisition devices, systematic analysis of misclassified cases, and ablation studies to assess the contribution of individual preprocessing operations. Further experiments should also quantify computational efficiency and evaluate the model under realistic deployment conditions. These investigations would provide stronger evidence regarding the model's robustness to variations in image characteristics and its feasibility for further development as a computer-assisted tool for preliminary skin lesion assessment. The model should not be considered a substitute for professional clinical diagnosis without appropriate clinical validation.

CONCLUSION

This study employed MobileNetV2 with transfer learning and fine-tuning to classify skin lesion images into benign and malignant categories. Evaluation on 659 test images achieved 85.13% accuracy and a test loss of 0.3430, with 561 correct and 98 incorrect predictions. Recall reached 88.33% for malignant lesions and 82.45% for benign lesions. However, 35 malignant images were misclassified as benign, highlighting the need to reduce false-negative predictions.

An additional evaluation on 128 images using Google Colab achieved 87.50% accuracy, with classification rates of 92.19% for benign and 82.81% for malignant images. These results are supplementary and do not demonstrate improved performance or generalization because the test sets differ and their independence must be verified. Further research should include external validation, more diverse datasets, and systematic analysis of misclassified images. Direct deployment tests and measurements of inference time, memory consumption, and computational requirements are also needed to assess suitability for resource-constrained devices. The system is intended to support preliminary skin lesion assessment, not replace professional clinical diagnosis. Further validation under representative real-world conditions is required before potential clinical application.

ACKNOWLEDGMENT

This research was supported by DIPA Politeknik Negeri Padang and BIMA Kemdiktisaintek. The authors gratefully acknowledge the financial support that made this research possible.

REFERENCES

Alam, T. M., et al. (2022). An efficient deep learning-based skin cancer classifier for an imbalanced dataset. Diagnostics, 15, 2115, https://doi.org/10.3390/diagnostics12092115.

Debelee, T. G. (2023). Skin lesion classification and detection using machine learning techniques: A systematic review. Diagnostics, *13*(19), 3147, https://doi.org/10.3390/diagnostics13193147.

Dhuhita, W. M. P., Ubaid, M. Y., & Baita, A. (2023). MobileNet V2 implementation in skin cancer detection. ILKOM Jurnal Ilmiah, 15(3), 498–506. https://doi.org/10.33096/ilkom.v15i3.1702.498-50

Hu, Z., Mei, W., Chen, H., & Hou, W. (2024). Multi-scale feature fusion and class weight loss for skin lesion classification. *Computers in Biology and Medicine, 176*, Article 108594. https://doi.org/10.1016/j.compbiomed.2024.108594.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9911

Indraswari, R., Rokhana, R., & Herulambang, W. (2021). Melanoma image classification based on MobileNetV2 network. Procedia Computer Science, 197, 198–207. https://doi.org/10.1016/j.procs.2021.12.132

Jain, S., Singhania, U., Tripathy, B., Nasr, E. A., & Aboudaif, M. K. (2021). Deep learning-based transfer learning for classification of skin cancer. Sensors, 21, 8142. https://doi.org/10.3390/s21238142.

Kassem, M. A., Hosny, K. M., & Damaševič, R. (2021). Machine learning and deep learning methods for skin lesion classification and diagnosis: A systematic review. Diagnostics 2021, 11, 1390. https://doi.org/10.3390/diagnostics11081390.

Meedeniya, D., De Silva, S., Gamage, L., & Isuranga, U. (2024). Skin cancer identification utilizing deep learning: A survey. The Journal of Engineering Become Vector, https://doi.org/10.1049/ipr2.13219.

Mehmood, A., Gulzar, Y., Ilyas, Q. M., Jabbari, A., & Ahmad, M. (2023). SBXception: A Shallower and Broader Xception Architecture for Efficient Classification of Skin Lesions. Cancers, 15, 3604. https://doi.org/10.3390/cancers15143604.

Naseri, H., & Safaei, A. A. (2025). Diagnosis and prognosis of melanoma from dermoscopy images using machine learning and deep learning: A systematic literature review. *BMC Cancer*. https://doi.org/10.1186/s12885-024-13423-y.

N. S. R., Shrivastava, V. K., & Parvathi, K. (2021). Multiclass skin lesion classification using image augmentation technique and transfer learning models. *International Journal of Intelligent Unmanned Systems, 2*(12), 220–228. https://doi.org/10.1108/IJIUS-02-2021-0010.

Nirupama, & Virupakshappa. (2024). MobileNet-V2: An enhanced skin disease classification by attention and multi-scale features. *Journal of Imaging Informatics in Medicine, 38*(3), 1734–1754. https://doi.org/10.1007/s10278-024-01271-y.

Öznacar, T., & Kayapunar, N. V. (2025). Advanced skin cancer prediction with medical image data using MobileNetV2 deep learning and optimized techniques. Scientific Reports, 15, Article 28962. https://doi.org/10.1038/s41598-025-14963-4

Shetty, B., Fernandes, R., Rodrigues, A. P., Chengoden, R., Bhattacharya, S., & Lakshmanna, K. (2022). Skin lesion classification of dermoscopic images using machine learning and convolutional neural network. *Scientific Reports*. https://doi.org/10.1038/s41598-022-22644-9.

Shinde, R. K., et al. (2023). Squeeze-MNet: Precise skin cancer detection model for low computing IoT devices using transfer learning. Cancers, 15, 12. https://doi.org/10.3390/cancers15010012

Tajerian, A., Kazemian, M., Tajerian, M., & Akhavan Malayeri, A. (2023). Design and validation of a new machine-learning-based diagnostic tool for the differentiation of dermatoscopic skin cancer images. PLOS ONE, 18(4), e0284437. https://doi.org/10.1371/journal.pone.0284437

Wang, H., Qi, Q., Sun, W., Li, X., & Dong, B. (2023). Classification of skin lesions with generative adversarial networks and improved MobileNetV2. International of Journal Imaging System and Technology, https://doi.org/10.1002/ima.22880.

Wu, Y., Chen, B., Zeng, A., Pan, D., Wang, R., & Zhao, S. (2022). Skin cancer classification with deep learning: A systematic review. *Frontiers in Oncology*. https://doi.org/10.3389/fonc.2022.893972.

Yang, G., Luo, S., & Greer, P. (2025). Advancements in skin cancer classification: A review of machine learning techniques in clinical image analysis. *Multimedia Tools and Applications, 84*(11), 9837–9864. https://doi.org/10.1007/s11042-024-19298-2.

Ye, Z., Zhang, D., Zhao, Y., Chen, M., Wang, H., Seery, S., Qu, Y., & Xue, P. (2024). Deep learning algorithms for melanoma detection using dermoscopic images: A systematic review and meta-analysis. *Artificial Intelligence in Medicine*. https://doi.org/10.1016/j.artmed.2024.102934.

Yilmaz, A., Gencoglan, G., Varol, R., Demircali, A. A., Keshavarz, M., & Uvet, H. (2022). MobileSkin: Classification of skin lesion images acquired using mobile phone-attached hand-held dermoscopes. Journal of Clinical Medicine, 11, 5102, https://doi.org/10.3390/ jcm11175102.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.