COFFEE LEAF DISEASE CLASSIFICATION USING THREE LIGHTWEIGHT DEEP LEARNING MODELS: ROBUSTNESS EVALUATION AGAINST IMAGE DEGRADATION AND GRAD-CAM VISUALIZATION

TRAN HUU QUYET[1]
[1]Faculty of Engineering Technology, Hung Vuong University, Phu Tho Province, Vietnam
Email address: tranhuuquyet@hvu.edu.vn
**Corresponding author: Tran Huu Quyet*
http://doi.org/10.51453/3093-3706/2026/1478

ARTICLE INFO

Received: 15/3/2026
Revised: 22/4/2026
Published: 28/4/2026

KEYWORDS

Coffee leaf disease;
Lightweight deep learning;
Transfer learning;
MobileNetV3;
EfficientNet;
Robustness evaluation;
Grad-CAM.

ABSTRACT

This study evaluates three lightweight deep learning models, namely MobileNetV3-Small, MobileNetV3-Large, and EfficientNet-B0, for coffee leaf disease classification using the Ethiopian Coffee Leaf Disease dataset. The dataset includes four classes: Cerscospora, Healthy, Leaf rust, and Phoma, with 9,180 training images, 1,620 validation images, and 1,200 test images. All models were trained using transfer learning with ImageNet-pretrained weights, and the input images were resized to 224 × 224 pixels. Model performance was first assessed on the original test set and then further evaluated under two image degradation conditions, including blur and low-light. Grad-CAM was applied to the model with the highest overall performance to visualize image regions contributing to its predictions. The experimental results show that MobileNetV3-Large achieved the highest weighted-average F1-score on the original test images, reaching approximately 0,9992. Under degraded image conditions, the same model also obtained the highest weighted-average F1-scores, with 0,9114 under blur and 0,9464 under low-light. EfficientNet-B0 showed comparable performance on the original and low-light images but was more affected by blur. MobileNetV3-Small achieved high performance on the original images but showed a larger performance decrease under low-light conditions. Grad-CAM visualizations indicate that MobileNetV3-Large mainly focused on diseased leaf regions, while several blurred samples showed less distinct activation patterns, corresponding to some misclassification cases. The results indicate that, for lightweight deep learning models in coffee leaf disease classification, evaluation should consider not only classification performance on original images but also robustness under degraded image conditions and model interpretability.

1. INTRODUCTION

Coffee is an economically important crop in many tropical countries, and its growth and productivity can be substantially affected by foliar diseases. Diseases such as leaf rust, leaf spot, and other biologically induced leaf symptoms may reduce photosynthetic capacity and adversely affect crop yield and quality if they are not detected in a timely manner. In practical production settings, coffee leaf disease recognition is commonly performed through visual inspection by farmers or agronomic experts. Although this approach remains useful in field practice, it is strongly influenced by individual experience, observation conditions, and the visibility of disease symptoms on leaf surfaces. These limitations have encouraged the development of automated image-based methods to support faster and more consistent disease detection in smart agriculture systems.

In recent years, deep learning, particularly convolutional neural networks (CNNs), has been widely applied to image-based plant disease recognition. Unlike conventional approaches that rely on manually designed features such as shape, color, or texture descriptors, deep learning models can learn discriminative representations directly from image data. The study by Mohanty et al. is among the early works demonstrating the applicability of deep learning to image-based plant disease detection, while recent review studies have further highlighted the increasing role of computer vision and deep learning in precision agriculture [1], [2].

For coffee leaf disease recognition, several studies have employed CNNs, transfer learning, and related deep learning approaches for disease classification and severity assessment. Esgario et al. investigated the classification and severity estimation of biotic stress in coffee leaves using deep learning. Montalbo and Hernandez applied deep convolutional models to classify Barako coffee leaf diseases. Yebasse et al. combined coffee disease classification with visualization methods, whereas Pham et al. proposed artificial intelligence-based solutions for coffee leaf disease classification [3], [4], [5], [6]. These studies provide empirical evidence for the feasibility of deep learning in coffee leaf disease recognition and form a basis for further experimental evaluation of model efficiency and applicability under practical imaging conditions.

In addition to classification accuracy, model size and computational complexity are important factors when disease recognition systems are intended for deployment on resource-constrained devices, such as mobile phones, edge devices, or field-based diagnostic systems. Lightweight architectures, including MobileNetV3 and EfficientNet, have been designed to balance computational efficiency and representational capacity. MobileNetV3 was developed with mobile and embedded applications in mind, while EfficientNet introduced a compound scaling strategy to systematically balance network depth, width, and input resolution [7], [8]. In the context of coffee leaf disease recognition, edge-based deployment has also received research attention, indicating the need for compact models that can operate reliably outside laboratory environments [9].

However, many existing studies on plant disease classification, including those on coffee leaf diseases, mainly evaluate model performance on original images or images captured under relatively controlled conditions. In real field environments, images may suffer from quality degradation caused by hand motion, defocus, low illumination, unstable viewing angles, or varying environmental conditions. Such factors can alter the visual characteristics of diseased regions and reduce model recognition performance. Therefore, in addition to reporting accuracy or F1-score on original test images, it is necessary to evaluate model robustness under common image degradation conditions, such as blur and low-light. This evaluation direction is consistent with robustness studies that examine neural network performance under common corruptions and perturbations in image data [10].

Another important issue in applying deep learning to plant disease diagnosis is model interpretability. Although CNNs can achieve high classification performance, their decision-making process is often difficult to inspect directly. Grad-CAM is a gradient-based visualization technique that identifies image regions with substantial contributions to a model prediction [11]. In coffee leaf disease classification, Grad-CAM visualization can be used to examine whether a model focuses on diseased leaf regions and to support the analysis of incorrect or unstable predictions under degraded image conditions [5].

Based on these considerations, this study conducts an experimental evaluation using the Ethiopian Coffee Leaf Disease dataset, which contains four classes: Cerscospora, Healthy, Leaf rust, and Phoma [12]. Three lightweight deep learning models are compared, including MobileNetV3-Small, MobileNetV3-Large, and EfficientNet-B0. The models are trained using transfer learning with ImageNet-pretrained weights and are evaluated on the original test set as well as two degraded versions of the test images, namely blur and low-light. In addition, Grad-CAM is applied to the model with the highest overall performance to visualize the image regions used by the model during prediction.

The main contributions of this study are threefold. First, it compares three lightweight deep learning models for four-class coffee leaf disease classification under a unified experimental setting. Second, it evaluates model robustness under two common image degradation conditions, blur and low-light, rather than relying only on performance measured on original images. Third, it uses Grad-CAM to visually analyze the decision regions of the best-performing model, including correctly classified samples and degraded or misclassified samples. This study does not propose a new network architecture; instead, it focuses on an application-oriented experimental evaluation to provide additional evidence on the use of lightweight deep learning models for coffee leaf disease classification.

2. RELATED WORKS

Image-based plant disease recognition has been extensively studied with the development of deep learning and computer vision. Unlike conventional methods that often depend on hand-crafted features, convolutional neural networks (CNNs) can learn relevant visual representations directly from leaf images, including color patterns, lesion shapes, and surface textures. Mohanty et al. reported the applicability of deep learning to image-based plant disease detection, providing an early basis for the use of CNNs in precision agriculture [1]. Recent review studies have also indicated that deep learning and computer vision are increasingly used for plant disease detection, classification, and monitoring across different crop types [2].

In the context of coffee leaf diseases, several studies have applied CNNs and transfer learning for disease classification and severity assessment. Esgario et al. investigated the classification and severity estimation of biotic stress in coffee leaves using deep learning [3]. Montalbo and Hernandez employed deep convolutional models for classifying Barako coffee leaf diseases, while Yebasse et al. combined coffee disease classification with visualization methods to support model interpretation [4] [5]. More recent studies have further extended this research direction through artificial intelligence-based approaches, transfer learning, and specialized deep learning models for coffee leaf disease detection and classification [6], [13], [14]. Araaf et al. focused on coffee leaf rust detection and implemented an edge device for pruning infected leaves, showing the relevance of deep learning models for field-oriented disease management [15]. In addition to model development, specialized datasets such as BRACOL have contributed to research on the identification and quantification of coffee leaf diseases and pests [16].

Beyond classification accuracy, practical deployment requires models with compact size, fast inference, and suitability for resource-constrained devices. MobileNetV3 is a lightweight CNN architecture designed for mobile and embedded environments [7]. EfficientNet is another widely used architecture that applies a systematic model scaling strategy to balance network depth, width, and input resolution [8]. In coffee disease classification, edge-based deployment has also been explored, indicating the need to evaluate lightweight models for practical coffee leaf disease recognition tasks [9].

However, a common limitation of previous studies is that model evaluation is often conducted mainly on original images or images captured under relatively stable conditions. In real-world environments, leaf images may be blurred, poorly illuminated, or otherwise degraded due to non-uniform image acquisition conditions. These changes may affect disease-related visual features and reduce recognition performance. Therefore, robustness evaluation under common image degradation conditions is necessary, particularly when models are expected to operate in field conditions [10].

Model interpretability is also relevant for computer-aided plant disease diagnosis. Grad-CAM provides a gradient-based visualization mechanism for identifying image regions that contribute to CNN predictions [11]. In coffee leaf disease classification, such visualization can help examine whether the model attends to diseased leaf regions and can support the analysis of both correct predictions and misclassified cases under degraded image conditions [5].

Overall, previous studies have demonstrated the applicability of deep learning to plant disease recognition in general and coffee leaf disease classification in particular. Existing research has addressed disease classification, severity estimation, model visualization, coffee leaf rust detection, edge deployment, and dataset construction. Nevertheless, limited experimental work has jointly examined three aspects within the same setting: comparison of lightweight deep learning models, robustness evaluation under image degradation conditions such as blur and low-light, and Grad-CAM-based interpretation. Therefore, this study compares MobileNetV3-Small, MobileNetV3-Large, and EfficientNet-B0 on the Ethiopian Coffee Leaf Disease dataset, while also evaluating robustness and analyzing Grad-CAM visualizations to provide an application-oriented assessment of lightweight models for coffee leaf disease classification.

3. PROPOSED METHODOLOGY

This section describes the experimental procedure used to compare three lightweight deep learning models for coffee leaf disease classification. Since this study does not propose a new network architecture, the methodology is designed as a unified evaluation framework, including dataset preparation, image preprocessing, transfer learning-based model training, evaluation on original images, robustness assessment under degraded image conditions, and Grad-CAM-based visualization. The overall workflow of the study is illustrated in Figure 1.

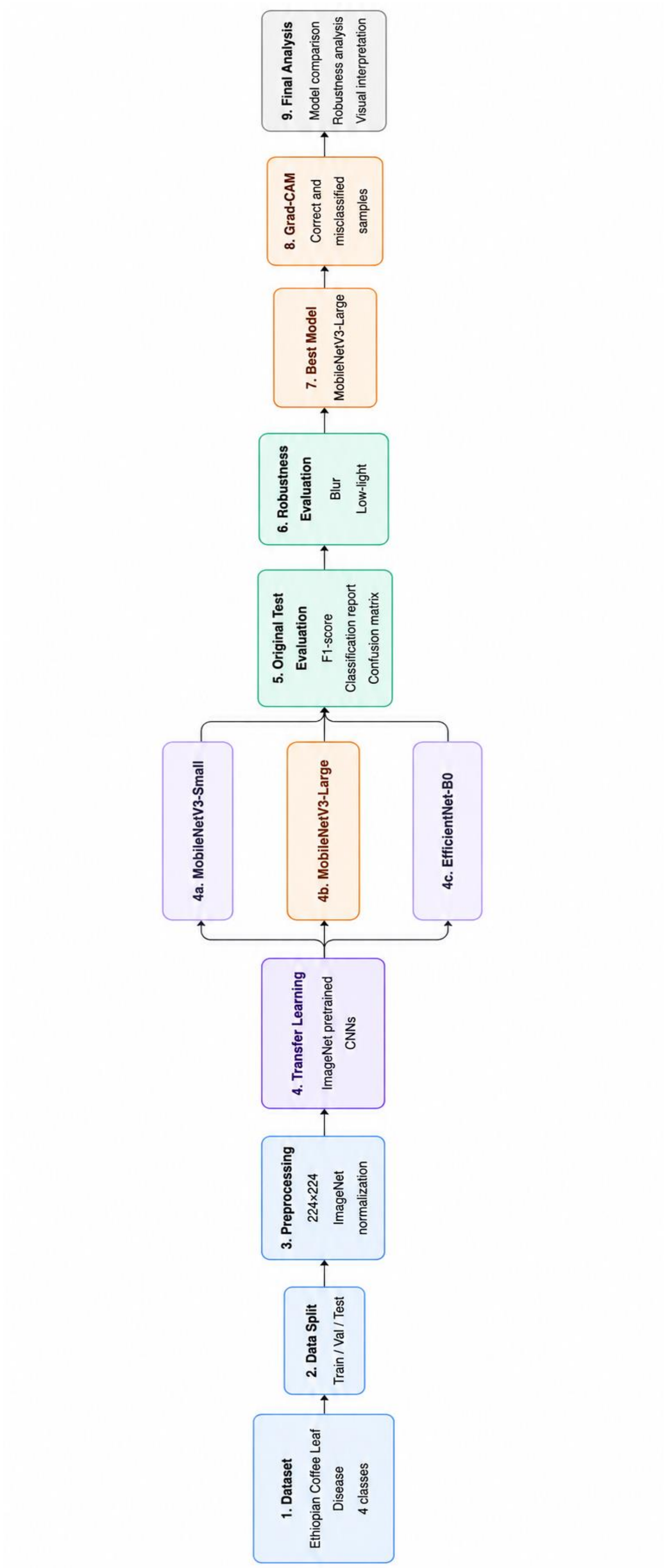

Figure 1. Overall experimental workflow, including data preparation, training of three lightweight deep learning models, evaluation on original and degraded images, selection of the best-performing model, and Grad-CAM visualization.

3.1. Dataset

This study used the Ethiopian Coffee Leaf Disease dataset published on Kaggle [12]. The dataset contains coffee leaf images from four classes: Cerscospora, Healthy, Leaf rust, and Phoma. In the experimental environment, the data were organized into two main directories, namely train aug and test. The train aug directory contained 2,700 images for each class. These images were divided into 2,295 images per class for training and 405 images per class for validation. The test directory was kept unchanged and used as the official test set, with 300 images for each class. In total, the experiment used 9,180 training images, 1,620 validation images, and 1,200 test images. The data split is summarized in Table 1.

Table 1. Experimental data split

Dataset	Images per class	Number of classes	Total images
Train	2295	4	9180
Validation	405	4	1620
Test	300	4	1200

The test set was kept separate from the training and validation processes to ensure that the final evaluation was performed on unseen data. This design helps provide a more objective assessment of model performance after training and model selection.

3.2. Image preprocessing

All input images were resized to 224 × 224 pixels to match the input requirements of the CNN architectures used in this study. The images were then converted into tensors and normalized using ImageNet statistics. The same preprocessing procedure was applied consistently to the training, validation, and test sets to ensure comparability across all experimental stages.

The training set was used to update model parameters, while the validation set was used to monitor training performance and select the best model checkpoint. The test set was used only for the final evaluation. This separation reduces the risk of biased evaluation caused by repeated use of test data during model training or model selection.

3.3. Lightweight deep learning models

Three lightweight deep learning models were selected for comparison: MobileNetV3-Small, MobileNetV3-Large, and EfficientNet-B0. MobileNetV3 is a family of CNN architectures designed for mobile and resource-constrained environments. MobileNetV3-Small represents a more compact configuration, whereas MobileNetV3-Large provides higher representational capacity [7].

EfficientNet-B0 was included as another lightweight baseline for comparison with MobileNetV3. The EfficientNet architecture is based on a compound scaling strategy that jointly scales network depth, width, and input resolution to balance accuracy and computational cost [8].

In this study, all three models were initialized with ImageNet-pretrained weights. The final classification layer of each model was modified to match the four-class coffee leaf disease classification task. Transfer learning was used to exploit visual representations learned from a large-scale dataset while reducing the training cost compared with training the models from scratch.

3.4. Training setup

The models were trained using PyTorch and torchvision in a Google Colab GPU environment. CrossEntropyLoss was used as the loss function because the task is a multi-class classification problem. The Adam optimizer was applied with a learning rate of 1e-4. Each model was trained for 8 epochs with a batch size of 32. After each epoch, the model was evaluated on the validation set, and the checkpoint with the best validation performance was saved for the testing stage.

Table 2. Training configuration

Component	Value
Framework	PyTorch, torchvision
Experimental environment	Google Colab GPU
Input image size	224 × 224
Training strategy	Transfer learning
Initial weights	ImageNet pretrained
Optimizer	Adam
Learning rate	1e-4
Loss function	CrossEntropyLoss
Batch size	32
Number of epochs	8

The same training configuration was applied to all three models to ensure a fair comparison. Therefore, differences in performance mainly reflect differences among the model architectures rather than variations in training settings.

3.5. Evaluation protocol and robustness assessment

After training, the models were first evaluated on the original test set. The evaluation outputs included the F1-score, classification report, and confusion matrix. The evaluation outputs included the weighted-average F1-score, classification report, and confusion matrix. The weighted-average F1-score was used as the main comparison metric because it jointly considers precision and recall while taking the number of samples in each class into account, making it suitable for multi-class classification tasks.

In addition to evaluation on original images, this study assessed model robustness under two common image degradation conditions: blur and low-light. Specifically, two additional test sets were generated from the original test set: one containing blurred images and the other containing low-light images. The three trained models were then evaluated on these degraded test sets, and their F1-scores were compared with the results obtained on the original images.

The use of blur and low-light conditions was intended to approximate common image acquisition issues in field environments, such as defocus caused by hand motion or insufficient illumination under

Email: sjttu@tqu.edu.vn

unstable lighting conditions. This evaluation approach is consistent with robustness studies that examine neural network performance under common image corruptions and perturbations [10].

3.6. Grad-CAM visualization

To complement the quantitative evaluation, Grad-CAM was used to visualize image regions that contributed to the classification decisions of the model. Grad-CAM generates gradient-based heatmaps that highlight the areas on which a CNN focuses when making predictions [11].

In this study, Grad-CAM was applied only to the model with the highest overall performance. Based on the experimental results, MobileNetV3-Large was selected for visualization. The visualization set included correctly classified samples from different classes and selected degraded or misclassified samples. The purpose of this step was to examine whether the model focused on disease-related regions of the leaves and to support the analysis of misclassification cases under degraded image conditions.

Overall, the methodological framework was designed to evaluate the models from three perspectives: classification performance on original images, robustness under degraded image conditions, and visual interpretability through Grad-CAM.

4. RESEARCH RESULTS

This section reports the experimental findings obtained from the three lightweight deep learning models on the Ethiopian Coffee Leaf Disease dataset. The evaluation was conducted from three complementary perspectives: classification performance on the original test set, robustness under degraded imaging conditions, and visual interpretability using Grad-CAM. The F1-score was used as the primary metric for comparing model performance.

4.1. Overall performance on original test images

The three models were first evaluated on the original test set, which contained 1,200 images in total, with 300 images for each class. The classification results are summarized in Table 3.

Table 3. Classification performance on the original test set

Model	Original F1-score
MobileNetV3-Small	0,9916
MobileNetV3-Large	0,9992
EfficientNet-B0	0,9958

As shown in Table 3, all three lightweight models achieved high F1-scores on the original test images. MobileNetV3-Large obtained the highest F1-score, reaching 0,9992, which indicates that the model was able to distinguish the four coffee leaf classes with only a small number of errors under normal image conditions. EfficientNet-B0 also achieved a high classification result, with an F1-score of 09958, while MobileNetV3-Small obtained an F1-score of 0,9916. Although MobileNetV3-Small performed slightly below the other two models, its result remains highly competitive given its more compact architecture.

These results suggest that transfer learning with lightweight convolutional neural networks yielded high classification performance for coffee leaf disease classification when the input images are clear and well captured. However, high performance on clean test images alone does not necessarily guarantee reliable performance in practical field conditions. In real agricultural environments, images may be affected by blur, insufficient lighting, or other acquisition-related issues. Therefore, additional robustness evaluation is needed to better assess the practical applicability of these models.

4.2. Confusion matrix analysis

To further examine the prediction behavior of each model, confusion matrices were generated for the original test set. The results are presented in Figure 2.

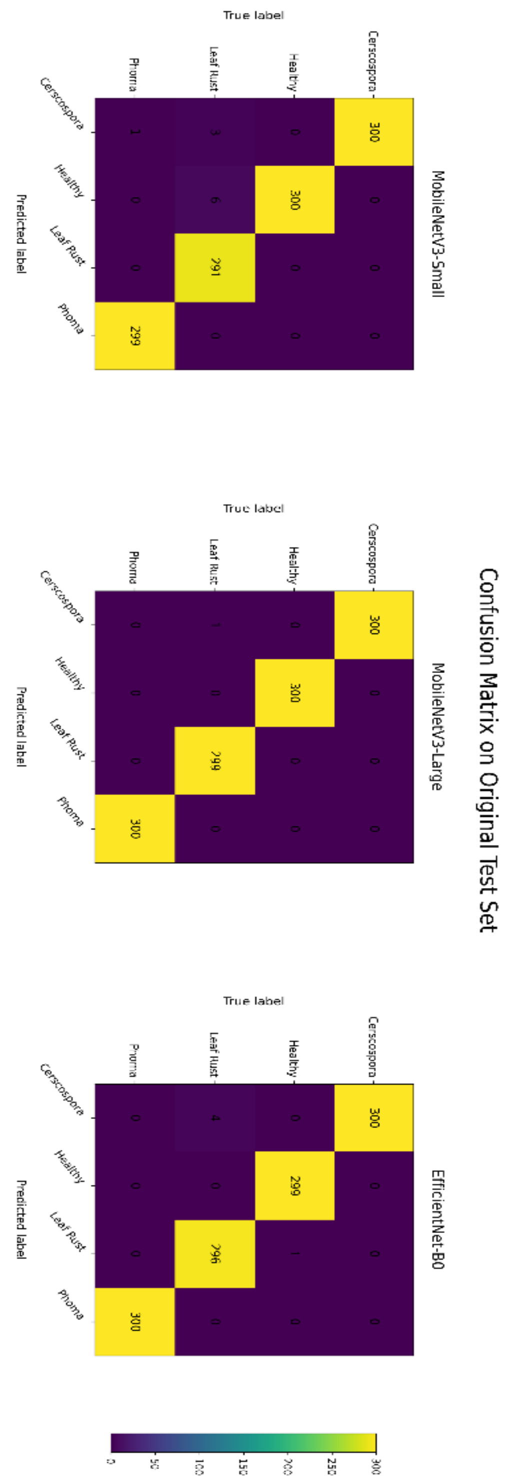

Figure 2. Confusion matrices of the three lightweight deep learning models on the original test set

The confusion matrices show that the predictions of all three models were highly concentrated along the main diagonal, confirming their strong classification performance on the original test images. MobileNetV3-Small correctly classified all samples from the Cerscospora and Healthy classes. For the Leaf rust class, 291 out of 300 samples were correctly predicted, while three samples were misclassified as Cerscospora and six samples were misclassified as Healthy. For the Phoma class, 299 out of 300 samples were correctly classified, with only one sample being misclassified as Cerscospora.

MobileNetV3-Large produced the most accurate confusion matrix among the three models. It correctly classified all 300 samples from the Cerscospora, Healthy, and Phoma classes. Only one Leaf rust sample was misclassified as Cerscospora, while the remaining 299 Leaf rust samples were correctly identified. This near-perfect prediction pattern is consistent with the highest F1-score reported for MobileNetV3-Large in Table 3.

EfficientNet-B0 also showed high performance across the four classes. The model correctly classified all samples from the Cerscospora and Phoma classes. For the Healthy class, 299 samples were correctly classified, while one sample was predicted as Leaf rust. For the Leaf rust class, 296 samples were correctly identified, and four samples were misclassified as Cerscospora.

Overall, the remaining errors were mainly related to the Leaf rust class, particularly its confusion with Cerscospora. This suggests that some Leaf rust samples may share similar visual characteristics with Cerscospora, such as lesion-like spots, color variation, or disease patterns on the leaf surface. Among the three models, MobileNetV3-Large showed the most reliable prediction behavior on the original test set.

4.3. Robustness evaluation under image degradation

To evaluate model robustness, two degraded versions of the original test set were created: a blurred image set and a low-light image set. The same trained models were then tested on these degraded images without further fine-tuning. The F1-score results across the three testing conditions are presented in Table 4.

Table 4. F1-score results on original and degraded test images

Model	Original F1	Blur F1	Low-light F1
MobileNetV3-Small	0,9916	0,8854	0,7816
MobileNetV3-Large	0,9992	0,9114	0,9464
EfficientNet-B0	0,9958	0,8177	0,9438

The robustness results show that image degradation had a clear negative impact on model performance. Under the blur condition, the F1-score decreased for all three models compared with their performance on the original test set. MobileNetV3-Large remained the best-performing model, achieving an F1-score of 0,9114. MobileNetV3-Small achieved 0,8854, while EfficientNet-B0 showed the largest performance drop, decreasing to 0,8177. This finding indicates that blur can substantially reduce the discriminative visual information required for coffee leaf disease classification, especially for EfficientNet-B0 in this experimental setting.

The results under the low-light condition reveal a different pattern. MobileNetV3-Large again achieved the highest F1-score, reaching 0,9464. EfficientNet-B0 followed closely with an F1-

score of 0,9438, suggesting that it maintained classification performance when image brightness was reduced. In contrast, MobileNetV3-Small experienced a considerable decline, with its F1-score dropping to 0,7816. This suggests that the smaller MobileNetV3 variant was less stable under insufficient illumination.

Taken together, MobileNetV3-Large demonstrated the most consistent performance across all evaluation conditions. It achieved the highest F1-score on the original test set and maintained the best performance under both blur and low-light degradation. These results indicate that MobileNetV3-Large provides a stronger balance between representational capacity and robustness to image quality degradation than the other two lightweight models evaluated in this study.

4.4. Grad-CAM visualization

Following the quantitative evaluation, Grad-CAM was applied to MobileNetV3-Large, which was selected as the best-performing model. The purpose of this analysis was to visually inspect the image regions that contributed most strongly to the model's predictions. The Grad-CAM results are shown in Figure 3.

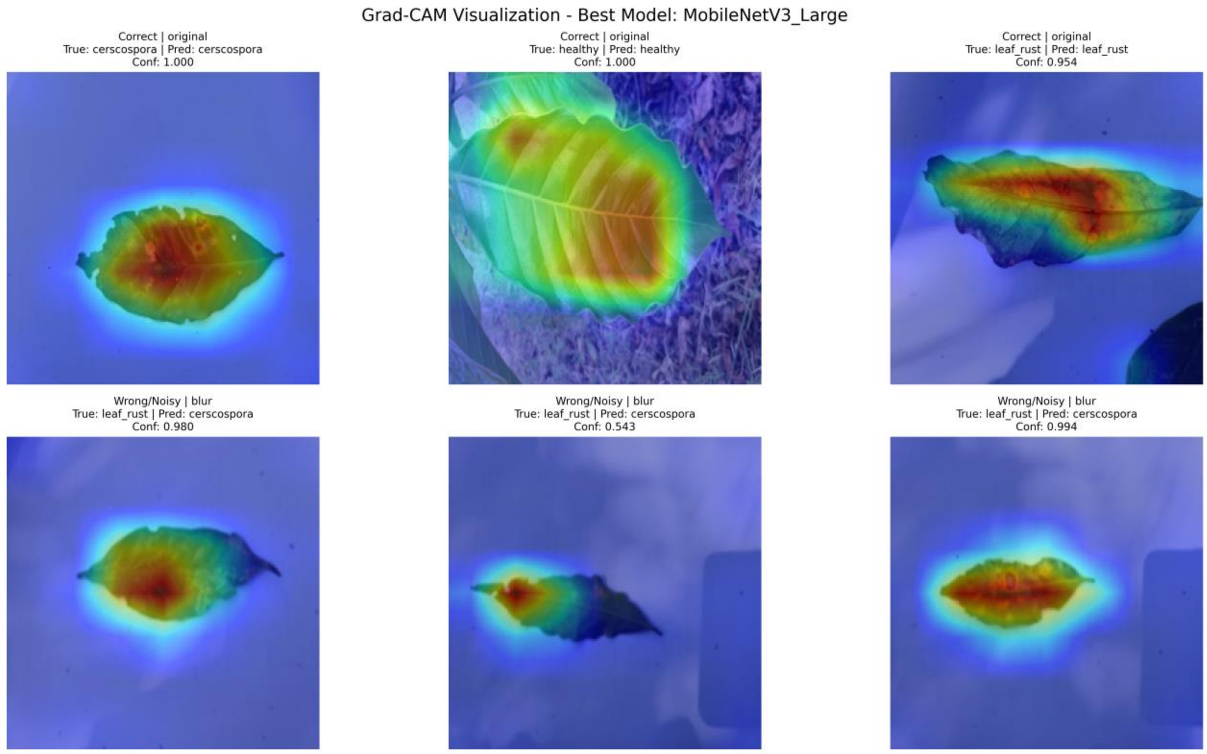

Figure 3. Grad-CAM visualization of MobileNetV3-Large on correctly classified samples and degraded or misclassified samples

The Grad-CAM visualizations show that, in correctly classified samples, MobileNetV3-Large generally focused on meaningful regions of the leaf rather than on the background. The highlighted areas often corresponded to visible disease-related patterns, including lesion regions, color changes, and abnormal surface structures. This suggests that the model learned relevant visual cues for distinguishing among the coffee leaf disease classes.

For degraded images, particularly blurred samples, the heatmaps indicate that the model still tended to attend to symptomatic regions on the leaf. However, blur reduced the clarity of lesion boundaries and disease-related textures, making the visual differences between classes less distinct.

As a result, the model became more prone to confusing visually similar classes. This behavior was especially evident in several blurred Leaf rust samples that were misclassified as Cerscospora.

These visualization results provide additional support for the robustness evaluation. The classification errors were not solely a consequence of model capacity; they were also related to the visibility and clarity of disease features in the input images. When disease-related patterns became less distinguishable due to image degradation, the model's decision-making became less reliable. This finding highlights the importance of evaluating deep learning models under non-ideal imaging conditions, particularly when such models are intended for real-world agricultural applications.

4.5. Discussion

The experimental results demonstrate that lightweight deep learning models can achieve highly accurate classification of coffee leaf diseases under normal image conditions. The high performance of all three models confirms the effectiveness of transfer learning in this task. By using ImageNet-pretrained weights, the models were able to leverage general visual representations and adapt them to the coffee leaf disease classification problem through fine-tuning.

Nevertheless, the robustness evaluation shows that performance on clean test images provides only a partial view of model reliability. When the input images were degraded by blur or low-light conditions, the performance of all models declined to varying degrees. MobileNetV3-Large showed the most stable behavior across both degradation types. EfficientNet-B0 remained robust under low-light conditions but was more sensitive to blur. MobileNetV3-Small, despite achieving high performance on the original test set, showed a substantial decline under low-light conditions.

From a practical perspective, MobileNetV3-Large was the most favorable model in this experimental setting among the three candidates evaluated in this study. It achieved the best overall classification performance while also maintaining lower performance degradation under degraded imaging conditions. This is particularly important for field-based plant disease recognition, where image quality can vary due to camera motion, focus errors, lighting changes, and uncontrolled acquisition environments.

The Grad-CAM results further strengthen this interpretation by showing that MobileNetV3-Large generally relied on disease-related regions of the leaves when making predictions. At the same time, the visualizations also indicate that degraded image quality can weaken the visibility of discriminative symptoms, thereby increasing the possibility of misclassification. Therefore, combining quantitative metrics with visual interpretability provides a more comprehensive understanding of model behavior than using classification performance alone.

In summary, MobileNetV3-Large achieved the best overall performance among the three lightweight models investigated in this study. It obtained the highest F1-score on the original test set and demonstrated lower performance degradation under both blur and low-light conditions. The Grad-CAM analysis also showed that the model tended to focus on lesion-related regions, supporting the interpretability of its predictions and providing useful insight into the causes of misclassification under degraded image conditions.

5. CONCLUSION AND FUTURE DEVELOPMENT

This study compared three lightweight deep learning models, namely MobileNetV3-Small, MobileNetV3-Large, and EfficientNet-B0, for coffee leaf disease classification using the Ethiopian Coffee Leaf Disease dataset. The models were trained using a transfer learning strategy and evaluated on four image classes: Cerscospora, Healthy, Leaf rust, and Phoma. In addition to performance evaluation on the original test set, this study assessed model robustness under two degraded image conditions, blur and low-light, and applied Grad-CAM to visualize the decision regions of the model with the highest overall performance.

The experimental results show that all three models achieved high F1-scores on the original test images. MobileNetV3-Large obtained the highest F1-score, reaching approximately 0,9992, followed by EfficientNet-B0 with an F1-score of approximately 0,9958 and MobileNetV3-Small with an F1-score of approximately 0,9916. The confusion matrix results also showed that MobileNetV3-Large produced only a small number of misclassifications on the original test set, indicating its ability to distinguish the four coffee leaf classes under stable image conditions.

In the robustness evaluation, the classification performance of all models decreased when the input images were degraded. However, the degree of performance reduction varied across architectures. MobileNetV3-Large maintained the highest F1-scores under both blur and low-light conditions, with values of approximately 0,9114 and 0,9464, respectively. EfficientNet-B0 maintained high performance under low-light conditions but showed a larger decrease under blur, whereas MobileNetV3-Small was less stable, particularly under low-light conditions. These results indicate that evaluation on original images alone is insufficient when the models are intended for use in practical field environments.

The Grad-CAM results for MobileNetV3-Large showed that the model generally focused on leaf regions and areas associated with visible disease symptoms when making predictions. In several blurred samples, disease-related features became less distinct, which contributed to misclassifications between Leaf rust and Cerscospora. This indicates that Grad-CAM can support model interpretation and provide useful insight into the causes of selected misclassification cases.

Future work may extend this study by evaluating the models on more diverse field-collected datasets and by considering additional image degradation factors, such as changes in viewing angle, complex backgrounds, and occlusion. Further evaluation should also include deployment-related metrics, such as model size, inference speed, and computational cost. In addition, data augmentation strategies and robustness-oriented training methods may be investigated to improve model performance under unstable image acquisition conditions.

REFERENCES

1. S. P. Mohanty, D. P. Hughes, and M. Salathé, “Using Deep Learning for Image-Based Plant Disease Detection,” Front. Plant Sci., vol. 7, Sep. 2016, doi: 10.3389/fpls.2016.01419.

2. A. Upadhyay et al., “Deep learning and computer vision in plant disease detection: a comprehensive review of techniques, models, and trends in precision agriculture,” Artif. Intell. Rev., vol. 58, no. 3, p. 92, Jan. 2025, doi: 10.1007/s10462-024-11100-x.

3. J. G. M. Esgario, R. A. Krohling, and J. A. Ventura, "Deep learning for classification and severity estimation of coffee leaf biotic stress," Comput. Electron. Agric., vol. 169, p. 105162, Feb. 2020, doi: 10.1016/j.compag.2019.105162.

4. F. J. P. Montalbo and A. A. Hernandez, "Classifying Barako coffee leaf diseases using deep convolutional models," Int. J. Adv. Intell. Inform., vol. 6, no. 2, pp. 197–209, Jul. 2020, doi: 10.26555/ijain.v6i2.495.

5. M. Yebasse, B. Shimelis, H. Warku, J. Ko, and A. Et., "Coffee Disease Visualization and Classification," Plants, 2021, doi: 10.3390/plants10061257.

6. T. C. Pham, V. D. Nguyen, C. H. Le, M. Packianather, and V.-D. Hoang, "Artificial intelligence-based solutions for coffee leaf disease classification," IOP Conf. Ser. Earth Environ. Sci., vol. 1278, no. 1, p. 012004, Oct. 2023, doi: 10.1088/1755-1315/1278/1/012004.

7. A. Howard et al., "Searching for MobileNetV3," arXiv.org, 2019. [Online]. Available: https://arxiv.org/abs/1905.02244v5. Accessed: May 21, 2026.

8. M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," arXiv.org, 2019. [Online]. Available: https://arxiv.org/abs/1905.11946v5. Accessed: May 21, 2026.

9. J. V. Y. Bordin Yamashita and J. P. R. R. Leite, "Coffee disease classification at the edge using deep learning," Smart Agric. Technol., vol. 4, p. 100183, Aug. 2023, doi: 10.1016/j.atech.2023.100183.

10. D. Hendrycks and T. Dietterich, "Benchmarking Neural Network Robustness to Common Corruptions and Perturbations," arXiv.org, 2019, doi: 10.48550/ARXIV.1903.12261.

11. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," arXiv.org, 2016, doi: 10.48550/ARXIV.1610.02391.

12. B. Yoseph, "ETHIOPIAN COFFEE LEAF DISEASE," Kaggle, 2026. [Online]. Available: https://www.kaggle.com/datasets/biniyamyoseph/ethiopian-coffee-leaf-disease. Accessed: May 21, 2026.

13. N. Mansouri, H. Guessmi, and A. Alkhalil, "A deep learning model for detection and classification of coffee-leaf diseases using the transfer-learning technique," Int. J. Adv. Intell. Inform., vol. 10, no. 3, pp. 379–396, Aug. 2024, doi: 10.26555/ijain.v10i3.1573.

14. A. Topal, B. Tunga, and E. B. Tirkolaee, "DeepEMPR: coffee leaf disease detection with deep learning and enhanced multivariance product representation Mercedes," PeerJ Comput. Sci., vol. 10, p. e2406, 2024, doi: 10.7717/peerj-cs.2406.

15. R. T. Araaf, A. Minn, and T. Ahamed, "Coffee Leaf Rust Disease Detection and Implementation of an Edge Device for Pruning Infected Leaves via Deep Learning Algorithms," Sensors, vol. 24, no. 24, p. 8018, Jan. 2024, doi: 10.3390/s24248018.

16. R. A. Krohling, G. J. M. Esgario, and J. A. Ventura, "BRACOL - A Brazilian Arabica Coffee Leaf images dataset to identification and quantification of coffee diseases and pests," vol. 1, Nov. 2019, doi: 10.17632/yy2k5y8mxg.1.