E-ISSN: 2963-4946
Vol. 1 No. 1 2022, pp. 40-56
DOI: 10.59888/ajosh.v4i10.736

Asian Journal of Social and Humanities

Comparison of EfficientNetV2-S and ConvNeXt V2-Tiny for Potato Leaf Condition Classification with Grad-CAM

Yoga Pamungkas
Sekolah Tinggi Manajemen Informatika dan Komputer (STMIK) IKMI Cirebon, Indonesia
Corresponding Author: yogapamungkas2304@gmail.com

Keywords:
image classification; potato leaf; EfficientNetV2-S; ConvNeXt V2-Tiny; Grad-CAM

Abstract

Background: Variations in background, lighting, and viewing angle make potato leaf condition classification difficult in uncontrolled environments.
Objective: This study compared EfficientNetV2-S and ConvNeXt V2-Tiny on seven classes in terms of classification performance and inference efficiency and examined the image regions contributing to predictions using Gradient-weighted Class Activation Mapping (Grad-CAM).
Method: Of 3,076 initial images, 12 with label conflicts were quarantined and 49 identical copies removed, leaving 3,015 images, split into 2,412 training, 301 validation, and 302 test images with similar images kept in the same partition. Both models were trained using transfer learning, weighted cross-entropy, and three seeds.
Results: EfficientNetV2-S achieved a mean test accuracy of 83.22% and macro-F1 of 79.86%, whereas ConvNeXt V2-Tiny achieved 87.97% and 85.21%. The macro-F1 difference of 5.34 percentage points had a 95% confidence interval of 2.43–8.50 percentage points based on a paired group bootstrap, although the per-run advantage ranged from 1.06 to 11.06 points and the seed-42 McNemar test was not significant ($p \approx 0.39$). On an NVIDIA T4 GPU, ConvNeXt V2-Tiny required 11.70 versus 25.39 ms per image. Grad-CAM visualizations showed different activation spreads for the two models on example test images.
Conclusion: ConvNeXt V2-Tiny delivered higher performance and lower latency in this internal test and is the more practical candidate for server-based diagnosis, although cross-location evaluation and spatial validation against lesion annotations are still required.

This is an open access article under the CC BY-SA license.

Copyright © 2026 by Author.
Published by
Pelopor Publikaasi Akademika

INTRODUCTION

Potato is a staple crop whose yield and quality are threatened by a wide range of diseases and pests, and early recognition of their symptoms is an important part of crop management (Gülmez, 2025). The condition of potato leaves can be recognized from visual symptoms such as spots and changes in colour. Images taken in uncontrolled environments also contain variations in lighting, viewing angle, and background. These variations can make classes harder to separate and give a model the opportunity to exploit image elements outside the leaf symptoms.

Globally, about 375 million tonnes of potatoes were harvested from approximately 17.8 million hectares in 2022 (Park et al., 2025), and pathogens and pests are estimated to reduce global potato yields by about 17.2%, with a range of 8.1–21.0% across regions

(Savary et al., 2019). In Indonesia, Central Java is the second-largest potato-producing province after East Java, with about 277.7 thousand tonnes or 20.4% of national production (Sahara et al., 2024).

In highland production centres such as those in Central Java, expert diagnosis is not always available in the field, so delays in recognizing symptoms can translate into yield and income losses for growers. Shabrina et al., (2024) provided a field dataset covering seven categories, including healthy leaves, so that model evaluation can include several types of disorders under diverse image-acquisition conditions. The review by Shoaib et al., (2023) summarized the development of deep learning methods for plant disease recognition. In the context of potato, Gülmez, (2025) highlighted the variability of symptoms, the limitations of datasets, and the challenges of generalization under different image-acquisition conditions.

Many earlier studies relied on images captured in the laboratory with uniform backgrounds, such as those in PlantVillage, which are easier to classify than field images and may lead to optimistic estimates of performance (Xu et al., 2024). Nazir et al., (2023) developed EfficientPNet for potato leaf disease classification. Alhammad et al., (2025) combined image classification with an explainable AI approach. Sinamenye et al., (2025) evaluated a hybrid EfficientNetV2B3 and Vision Transformer model for potato disease classification. Other studies have used the same field dataset.

Shabrina et al., (2024) reported an accuracy of 73.63% for EfficientNetV2B3 on this dataset, compared with 98.15% on PlantVillage. Sangar & Rajasekar, (2025) reported 87.82% by combining EfficientNet-LITE features with an optimized support vector machine, and Rofiqi et al., (2026) reported 89.5% for MobileNetV3-Large after augmentation, class balancing, and Optuna hyperparameter tuning. These studies differ in partitioning, augmentation, and the treatment of similar images, which limits the direct comparison of their scores. In model comparisons, differences in partitions, the degree of image similarity, and the class distribution can affect the interpretation of scores (Xu et al., 2023, 2024). For example, nearly identical images in the training and test partitions can make performance estimates overly optimistic. Table 1 summarizes the studies that are closest to the present work.

Table 1. Summary of related studies

Study	Data and classes	Method	Reported result
Nazir et al., (2023)	PlantVillage potato images (10,800); early blight, late blight, healthy	EfficientPNet: EfficientNetV2 with spatial-channel attention	Accuracy 98.12%
Alhammad et al., (2025)	Public potato leaf disease dataset	Transfer learning with Grad-CAM explanations	Validation accuracy 97%; test accuracy 98%
Sinamenye et al., (2025)	Field potato leaf dataset (3,076 images; seven classes)	Hybrid EfficientNetV2B3 + Vision Transformer	Accuracy 85.06%

Study	Data and classes	Method	Reported result
Shabrina et al., (2024)	Field potato leaf dataset (3,076 images; seven classes)	EfficientNetV2B3 baseline	Accuracy 73.63%
Sangar & Rajasekar, (2025)	Field potato leaf dataset (3,076 images; seven classes)	EfficientNet-LITE features + optimized SVM (KE-SVM)	Accuracy 87.82%
Rofiqi et al., (2026)	Field potato leaf dataset (3,076 images; seven classes)	MobileNetV3-Large with augmentation, SMOTE, and Optuna tuning	Accuracy 89.5%; F1-score 90.4%
Indra et al., (2026)	Sugarcane leaves (2,948 images; five classes); field test set of 427 images	ConvNeXt V2-Tiny, six-fold ensemble; EfficientNetV2-S and ResNet-50 baselines	Test accuracy 98.39%; EfficientNetV2-S baseline 83.37%
This study	Field potato leaf dataset after curation (3,015 images; seven classes)	EfficientNetV2-S vs. ConvNeXt V2-Tiny; three seeds; grouped split	Mean test accuracy 83.22% vs. 87.97%

Vo et al., (2025) compared EfficientNetV2 and ConvNeXt for plant disease classification. Indra et al., (2026) used ConvNeXt V2-Tiny for sugarcane leaf disease classification with an ensemble approach. Gondal et al., (2026) studied plant disease classification under low-light conditions using knowledge distillation. Differences in crops, image conditions, and evaluation protocols limit the comparison of scores across studies. Table 1 shows, for example, that accuracies above 98% were obtained on PlantVillage or with fold ensembles, whereas a hybrid model on the same field potato dataset used here reached 85.06%.

Classification scores alone do not show whether a model bases its decisions on disease symptoms or on incidental elements such as soil, hands, or neighbouring plants that may be correlated with a class in a particular dataset. Visual explanation methods are therefore increasingly reported alongside accuracy in plant disease studies. Gradient-weighted Class Activation Mapping (Grad-CAM) localizes the image regions that support a class score by weighting the last convolutional feature maps with the gradients of that score (Selvaraju et al., 2017).

Alhammad et al., (2025) used Grad-CAM to make the decisions of a potato leaf classifier more interpretable, Natarajan et al., (2024) used explainable visualizations to examine plant disease predictions, and Kondaveeti & Simhadri, (2025) argued for combining qualitative inspection with quantitative evaluation of the explanations. In field images with cluttered backgrounds, such checks are particularly important because the risk of shortcut learning is higher than in laboratory images. Shortcut learning refers to decision rules that perform well on a benchmark but fail to transfer to new conditions (Geirhos et al., 2020), and saliency maps can remain largely unchanged even when the upper-layer parameters of the model are randomized (Binder et al., 2023).

Visual explanations should therefore be accompanied by checks of their reliability. This study compares EfficientNetV2-S and ConvNeXt V2-Tiny on the same partition after an audit of label conflicts and image similarity. Its contribution is a paired evaluation of classification performance, inference cost, and the activation patterns of the two model configurations. Macro-F1 is used to give equal weight to every class in the performance summary, whereas Grad-CAM is used to examine the spatial contributions to predictions. Specifically, the study addresses three questions: First, which of the two models achieves higher classification performance on a curated, similarity-aware partition of the field potato leaf dataset, and how large and uncertain is the difference? Second, how do the two models compare in parameter count and GPU forward time? Third, how do the Grad-CAM activation patterns of the two models differ, and how far can these maps be trusted?.

The novelty of this study lies in combining a label-audited, similarity-aware partition of a public field dataset with a multi-seed paired comparison that reports the uncertainty of the difference, inference cost, and diagnostics of the visual explanations together, aspects that are rarely reported jointly in the potato studies summarized in Table 1. The results are expected to benefit researchers by providing a more conservative reference point for model comparison on this dataset, and to benefit developers of digital crop-protection services by indicating which of two widely available backbones offers a better balance between accuracy and latency. The broader implication is that model selection for field diagnosis should rest on leakage-controlled evaluation, uncertainty-aware comparison, and checks of explanation reliability rather than on a single accuracy figure.

METHOD

The research workflow is shown in Figure 1. The study used a quantitative comparative experiment in which two pretrained backbones were trained under the same data partition, preprocessing, loss function, and training schedule, and were then compared using paired statistics and visual explanations.

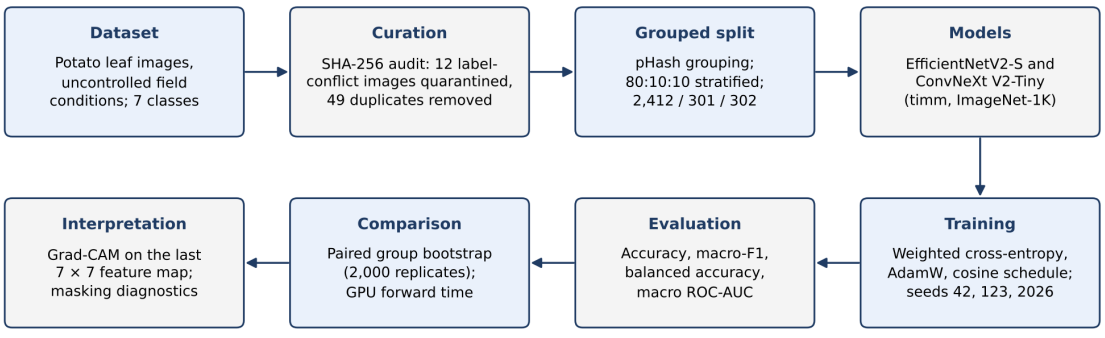


Figure 1. Research Workflow

Dataset and Data Curation

The experiment used the Potato Leaf Disease Dataset in Uncontrolled Environment (Shabrina et al., 2024). The dataset contains seven classes: leaves affected by bacteria, fungi, nematodes, pests, phytophthora, and viruses, and healthy leaves. As described by Sinamenye et al., (2025), the images were captured with smartphone cameras under real

farming conditions in Central Java, Indonesia, and were labelled by plant disease experts. All 3,076 images found could be read.

Pixel-level SHA-256 hashes were used to check for identical images. Six groups containing 12 identical images with different labels were quarantined. After 49 pixel-identical copies were removed, 3,015 unique images were retained (Table 2). Labels followed the class directories. The spelling Phytophthora in this article refers to the folder label Phytopthora in the notebook.

Table 2. Data curation steps

Step	Procedure	Images affected	Images remaining
Initial inventory	All files read successfully	–	3,076
Label-conflict check	Identical pixel hashes with different labels (six groups) quarantined	12	3,064
Duplicate removal	Pixel-identical copies with the same label removed	49	3,015
Similarity grouping	64-bit perceptual hash, Hamming distance ≤ 4; 36 candidate pairs grouped	Grouped, not removed	3,015

Source: Data Processed

A 64-bit perceptual hash with a maximum Hamming distance of four identified 36 candidate pairs of similar images. These pairs were grouped before the data were split at a ratio close to 80:10:10 with class stratification and split seed 2026. Each group was placed intact in a single partition. Checks in the notebook showed no overlap of pixel hashes or groups between partitions. Plant identity and photography session were not used in the grouping. The resulting class distribution is shown in Table 3, together with the class weights used in the loss function.

Table 3. Image distribution after curation and class weights

Class	Training	Validation	Test	Total	Class weight
Bacteria	455	57	57	569	0.757
Fungi	589	75	76	740	0.585
Healthy	160	19	20	199	2.154
Nematode	53	7	7	67	6.501
Pest	483	59	59	601	0.713
Phytophthora	249	29	31	309	1.384
Virus	423	55	52	530	0.815
Total	2,412	301	302	3,015	–

Note. Class weights were computed from the training counts using Equation (4).

The distribution is clearly imbalanced. Fungi is the largest class with 740 images, whereas Nematode has only 67 images, an eleven-fold difference. With only seven Nematode images in each of the validation and test partitions, a single prediction changes the Nematode recall on the test set by more than 14 percentage points, which must be kept in mind when per-class results are interpreted.

Models

EfficientNetV2-S uses an architecture that combines MBConv and Fused-MBConv blocks (Tan & Le, 2021). In Fused-MBConv, which is used in the early stages, the expansion 1 × 1 convolution and the depthwise 3 × 3 convolution of MBConv are replaced by a single regular 3 × 3 convolution, which makes better use of modern accelerators at high spatial resolutions. The later stages use MBConv blocks with depthwise convolutions and squeeze-and-excitation. The family was obtained through training-aware neural architecture search and scaling, with the goal of improving both accuracy and training speed (Tan & Le, 2021).

ConvNeXt V2 builds on ConvNeXt, a fully convolutional network modernized with large depthwise kernels, inverted bottlenecks, and other design choices adopted from vision Transformers (Liu et al., 2022). ConvNeXt V2-Tiny uses Global Response Normalization (GRN) and weights from a fully convolutional masked autoencoder (FCMAE) pretraining scheme that were subsequently fine-tuned on ImageNet-1K (Woo et al., 2023). In FCMAE pretraining, a large proportion of image patches is masked and the network learns to reconstruct them, which encourages representations that do not depend on any single local region. GRN was introduced to increase the contrast and selectivity of channels by normalizing each channel by its global response. For a feature map X with channels X_i, GRN aggregates the spatial L2 norm of each channel, normalizes it by the mean over channels, and rescales the input, as in Equations (1)–(3), where γ and β are learnable parameters (Woo et al., 2023).

$$G(X)_i = \|X_i\|_2 \tag{1}$$

$$N(G(X))_i = G(X)_i / \left((1/C) \Sigma_j G(X)_j \right) \tag{2}$$

$$X_i' = \gamma \cdot X_i \cdot N(G(X))_i + \beta + X_i \tag{3}$$

The implementation used the timm initial weights tf_efficientnetv2_s.in1k and convnextv2_tiny.fcmae_ft_in1k. Both backbones were retrained together with a classification head consisting of dropout 0.2 and a linear layer with seven outputs. Because the two models differ in both architecture and pretraining scheme, the comparison concerns two complete configurations rather than architecture alone.

Preprocessing and Training

RGB images were resized to 224 × 224 pixels and normalized according to the configuration of the initial weights of each backbone. For the training data, augmentation included horizontal and vertical flips, rotation of up to 15°, and random changes in

brightness, contrast, and saturation. Validation and testing used deterministic transformations.

Training was carried out on an NVIDIA Tesla T4 with a batch size of 8 and mixed precision. The AdamW optimizer used an initial learning rate of 0.0001 for the backbone and 0.001 for the classification head, a weight decay of 0.0001, and cosine annealing scheduling. The class weights in the weighted cross-entropy function were computed as in Equation (4), where N is the number of training images, K is the number of classes, and n_k is the number of training images in class k. The weighted loss for a training image with true class y and predicted probability p_y is given in Equation (5).

$$w_k = N / (K \times n_k) \quad (4)$$

$$L = - w_y \log p_y \quad (5)$$

Each model was trained with seeds 42, 123, and 2026 on the fixed partition. The training limit was 15 epochs per run, with a patience of three and a min_delta of 0.0001. The checkpoint was selected based on the lowest unweighted validation cross-entropy. The training settings are summarized in Table 4. The same optimizer, learning rates, schedule, and input resolution were deliberately applied to both backbones so that the comparison reflects the two pretrained configurations under one controlled fine-tuning protocol; no per-model hyperparameter search was performed. The input size of 224 × 224 pixels matches the fine-tuning resolution of convnextv2_tiny.fcmae_ft_in1k but is below the 300–384 pixel range used to train and evaluate EfficientNetV2-S (Tan & Le, 2021), which may disadvantage EfficientNetV2-S. This common resolution was retained to keep memory use and forward-time measurements comparable on the same GPU, and its possible effect is treated as a limitation of the study.

Table 4. Training configuration

Component	**Setting**
Initial weights	tf_efficientnetv2_s.in1k; convnextv2_tiny.fcmae_ft_in1k (timm)
Classification head	Dropout 0.2 + linear layer with seven outputs
Input and augmentation	224 × 224 RGB; flips, rotation ≤ 15°, colour jitter (training only)
Hardware and precision	NVIDIA Tesla T4; mixed precision; batch size 8
Optimizer	AdamW; learning rate 0.0001 (backbone) and 0.001 (head); weight decay 0.0001
Scheduler	Cosine annealing
Loss	Weighted cross-entropy, $w_k = N/(K \times n_k)$
Epochs and early stopping	Maximum 15; patience 3; min_delta 0.0001
Checkpoint selection	Lowest unweighted validation cross-entropy
Seeds	42, 123, and 2026 on a fixed partition

Source: Data Processed

Evaluation and Visualization

Performance was summarized using accuracy, macro-F1, balanced accuracy, and one-vs-rest macr ROC-AUC. With $F1_k$ and $recall_k$ denoting the F1-score and recall of

class k, macro-F1 and balanced accuracy are defined in Equations (6) and (7). Both metrics give every class the same weight regardless of its size, which makes them more informative than accuracy for the imbalanced distribution in Table 3.

$$\text{Macro-F1} = (1/K)\ \Sigma_k\ F1_k \tag{6}$$

$$\text{Balanced accuracy} = (1/K)\ \Sigma_k\ \text{Recall}_k \tag{7}$$

As a reference, a baseline always predicted the most frequent class in the training data. Differences between the models were estimated using a paired group bootstrap on the test data with 2,000 valid replicates. In each replicate, the test groups were resampled with replacement, both models were evaluated on the same resampled groups, and the scores were averaged over the three seeds; the distribution of the replicate differences was used to obtain the 95% confidence interval. Bootstrapping the test set to obtain the confidence interval followed Varoquaux & Colliot, (2023).

A replicate was considered valid only when every class was represented in the resampled test set, so that per-class F1 and macro-F1 were defined; invalid replicates were discarded and resampling continued until 2,000 valid replicates were obtained. For the seed-42 pair, paired correctness was additionally compared with an exact McNemar test on the discordant predictions (Rainio et al., 2024). Resampling groups rather than individual images respects the grouping used during splitting.

Efficiency was measured as GPU forward time using synthetic inputs, FP16 autocast, 10 warm-up iterations, and 50 measurement iterations.Grad-CAM was computed Selvaraju et al., (2017) from the gradients of the target-class logit with respect to the last spatial feature map (7 × 7) of each backbone, and the maps were then upsampled and overlaid on the image. For a feature map A^k and the logit y^c of class c, the channel weights and the activation map are given in Equations (8) and (9), where Z is the number of spatial positions.

$$\alpha_k^c = (1/Z)\ \Sigma_i\ \Sigma_j\ \partial y^c / \partial A_{ij}^k \tag{8}$$

$$L_{\text{Grad-CAM}}^c = \text{ReLU}(\Sigma_k\ \alpha_k^c\ A^k) \tag{9}$$

Visualization used seed 42 on fixed test samples and examples of misclassification. Additional diagnostics included randomization of the classification head, following the model-randomization sanity check examined by Binder et al., (2023), and a comparison of the drop in probability between occluding the 20% of pixels with the highest activation and occluding a shifted mask of the same size.

RESULTS AND DISCUSSION

Classification Performance

ConvNeXt V2-Tiny recorded higher test accuracy and macro-F1 in all three seeds. The macro-F1 difference between the models ranged from 1.06 to 11.06 percentage points, so the size of the advantage depended on the run. The highest test accuracy in a single run was 90.40% for ConvNeXt V2-Tiny with seed 2026. The main summary uses the mean of the three runs, with checkpoints selected on the validation data (Table 5).

Table 5. Validation and test performance in percent

Metric	EfficientNetV2-S	ConvNeXt V2-Tiny
Validation accuracy	83.83 ± 1.38	88.37 ± 2.18
Validation macro-F1	83.53 ± 1.18	88.22 ± 2.41
Test accuracy	83.22 ± 2.68	87.97 ± 2.49
Test macro-F1	79.86 ± 2.60	85.21 ± 3.79
Test balanced accuracy	82.52 ± 1.79	86.02 ± 3.60
Test macro ROC-AUC	97.74 ± 0.78	98.56 ± 0.20

Note. Values are presented as mean ± standard deviation over three seeds.

ConvNeXt V2-Tiny was ahead by 4.75 percentage points in accuracy (95% CI 2.55–7.06) and 5.34 percentage points in macro-F1 (95% CI 2.43–8.50). Both intervals lie above zero in this internal test. The intervals describe the uncertainty due to resampling the test groups with models that had been trained on a single fixed partition; they do not capture the variation that would arise from re-splitting the data or from collecting images at other locations. The majority baseline, which always chose Fungi, achieved an accuracy of only 25.17% and a macro-F1 of 5.74%.

For both models, test macro-F1 was lower than test accuracy, by 3.36 percentage points for EfficientNetV2-S and 2.76 percentage points for ConvNeXt V2-Tiny. Because macro-F1 weights each class equally, this gap indicates that performance on some of the smaller classes was weaker than on the larger ones, which is consistent with the very small number of Nematode and Healthy test images. The standard deviation of test macro-F1 for ConvNeXt V2-Tiny (3.79) was also larger than its standard deviation of test accuracy (2.49), suggesting that the small classes contributed much of the run-to-run variation.

The macro ROC-AUC values of both models were high, at 97.74% and 98.56%, and much closer to each other than their macro-F1 values. ROC-AUC measures how well the predicted scores rank images of a class above images of other classes across all possible thresholds, whereas macro-F1 depends on the single class finally chosen by the arg-max decision. The combination of high ROC-AUC and lower macro-F1 therefore suggests that both models often assigned relatively high scores to the correct class even when another class received the highest score, which is typical when visually similar disorders, such as bacterial and fungal lesions, compete for the same image.

The variation across seeds was substantial. The standard deviations of test accuracy were 2.68 and 2.49 percentage points, and the macro-F1 advantage of ConvNeXt V2-Tiny varied from 1.06 to 11.06 percentage points depending on the seed. A comparison based on any single run could thus have reported either a negligible or a very large advantage. This observation supports the practice of reporting several seeds, together with an uncertainty estimate for the difference, when two architectures are compared on a dataset of this size.

Paired Analysis of Predictions

With seed 42, both models correctly predicted 246 of the 302 images and were both wrong on 22 images. ConvNeXt V2-Tiny was correct on 20 images that EfficientNetV2-S misclassified, whereas EfficientNetV2-S was correct on 14 images that ConvNeXt V2-Tiny misclassified (Table 6). These differing errors show that the model with the higher aggregate score can still fail on images that the comparison model recognizes.

Table 6. Paired correctness of the two models on the test set (seed 42)

	ConvNeXt V2-Tiny correct	**ConvNeXt V2-Tiny wrong**	**Total**
EfficientNetV2-S correct	246	14	260
EfficientNetV2-S wrong	20	22	42
Total	266	36	302

Source: Data processed

From these counts, the seed-42 test accuracy is 86.09% for EfficientNetV2-S and 88.08% for ConvNeXt V2-Tiny. An exact McNemar test on the 34 discordant images (20 versus 14) gives a two-sided p-value of approximately 0.39. For this single run, the difference in correctness between the models is therefore not statistically distinguishable from chance, even though the group bootstrap of the three-seed means placed the difference above zero. The two analyses answer different questions: the McNemar test concerns one pair of trained models, whereas the bootstrap summarizes the mean of three runs per model.

These results complement the study of Sinamenye et al., (2025), which reported 85.06% accuracy for a hybrid model on the same field dataset, by comparing two models on a similarity-based partition. A direct comparison of scores with that study requires the same partition and evaluation procedure. The direction of the difference is also consistent with Indra et al., (2026), in which ConvNeXt V2-Tiny outperformed an EfficientNetV2-S baseline on sugarcane leaf images, although the crop, data, and protocol differ. On the same field dataset, Shabrina et al., (2024) reported 73.63% for EfficientNetV2B3.

Sangar & Rajasekar, (2025) reported 87.82% for EfficientNet-LITE with an optimized support vector machine, and Rofiqi et al., (2026) reported 89.5% for a tuned MobileNetV3-Large. The mean accuracy of ConvNeXt V2-Tiny in this study (87.97%) falls within this range even though similar images were kept within single partitions, whereas the EfficientNetV2-S result (83.22%) is clearly higher than the EfficientNetV2B3 baseline reported with the dataset.

Because these studies used different partitions and augmentation strategies, and in some scenarios applied augmentation before splitting (Rofiqi et al., 2026), the differences provide context rather than a ranking. Loss weighting gave greater weight to classes with fewer training images, with weights ranging from 0.585 for Fungi to 6.501 for Nematode (Table 3). The contribution of weighting to the scores has not been measured separately because both models were trained with the same weighting scheme.

Inference Efficiency

ConvNeXt V2-Tiny has 27.87 million parameters, compared with 20.19 million for EfficientNetV2-S. On the Tesla T4 with a batch size of 1, the mean forward time of ConvNeXt V2-Tiny was 11.70 ms per image, whereas EfficientNetV2-S required 25.39 ms. With a batch size of 8, the throughputs were 349.71 and 302.64 images per second, respectively (Table 7). ConvNeXt V2-Tiny was faster in this GPU configuration despite having more parameters. The measurements cover forward computation with tensors already on the GPU; image loading, preprocessing, and data transfer were excluded.

Table 7. Parameters and GPU forward-time measurements on an NVIDIA Tesla T4

Measure	EfficientNetV2-S	ConvNeXt V2-Tiny
Parameters (millions)	20.19	27.87
Forward time, batch 1 (ms per image)	25.39	11.70
Throughput, batch 8 (images per second)	302.64	349.71

Source: Data processed

At a batch size of 1, ConvNeXt V2-Tiny was about 2.2 times faster than EfficientNetV2-S, whereas at a batch size of 8 its throughput advantage shrank to about 1.16 times. A plausible explanation is that EfficientNetV2-S contains many more sequential layers, including depthwise convolutions and squeeze-and-excitation operations, whose cost at small batch sizes is dominated by kernel-launch and memory-access overheads rather than arithmetic; larger batches amortize these overheads. This explanation was not tested directly, and the results show that parameter count alone is a poor proxy for latency on a given device.

Grad-CAM Interpretation

In the Bacteria example in Figure 2, both models predicted the correct label. The activation of EfficientNetV2-S spread more widely around the leaf in the left-centre part of the image, whereas that of ConvNeXt V2-Tiny was more concentrated in that area. This comparison shows a difference in spatial focus for the same prediction. Agreement with the lesion requires spatial annotations that can serve as a reference.

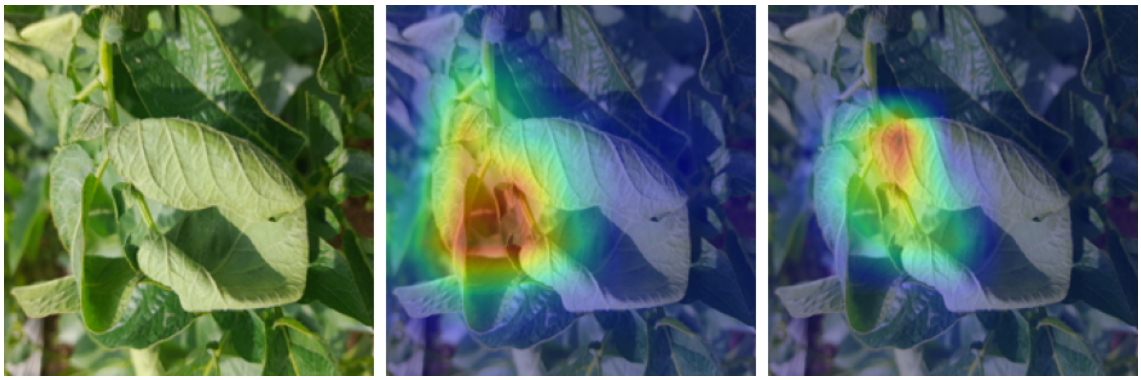


Figure 2. Grad-CAM example for one Bacteria test image with seed 42. From left to right: the original image, the EfficientNetV2-S overlay, and the ConvNeXt V2-Tiny overlay. Warm colours indicate relatively higher activation in each map

In three samples per model, occluding the most highly activated area reduced the probability of the target class more than the average reduction with the shifted mask. The Spearman correlation between the original maps and the maps after randomization of the classification head varied from −0.03 to 0.77. These exploratory results show different responses across samples. Masking can also produce images that differ from the original distribution, so changes in probability need to be read in the context of that procedure. Because the diagnostics were computed on only three samples per model, they are reported as exploratory and are not used to support conclusions about either model; extending them to the full test set is required before the reliability of the maps can be quantified.

The randomization diagnostic deserves particular attention. If a map remains highly correlated with the original after the classification head has been randomized, the map reflects properties of the backbone features more than the decision of the trained classifier. The upper value of 0.77 therefore suggests that, for some samples, the visual pattern may be only weakly tied to the learned class decision, whereas values near zero indicate maps that changed substantially. With only three samples per model, these diagnostics cannot be generalized, but they illustrate why visual plausibility alone is not sufficient evidence of explanation quality.

This observation agrees with Binder et al., (2023), who showed that explanation maps can change only to a limited extent when the upper layers of a network are randomized, and with the argument of Geirhos et al., (2020) that apparently correct predictions may rely on shortcuts that only targeted tests can reveal. Alhammad et al., (2025) used Grad-CAM to illustrate the decisions of a potato leaf classifier, Natarajan et al., (2024) applied XAI visualization to examine plant disease predictions, whereas Kondaveeti & Simhadri, (2025) combined visual and quantitative assessment on rice leaf images. In this study, activation maps were used to examine the focus of the models, and the diagnostic results were read according to the number of samples examined.

Discussion of the Model Difference

Several properties of ConvNeXt V2-Tiny may contribute to its higher performance on the field images, although the present design cannot separate them. First, FCMAE pretraining reconstructs masked regions of the input, which may produce features that are less dependent on any single local cue and more robust to clutter and partial occlusion, which are common in uncontrolled backgrounds (Woo et al., 2023). Second, ConvNeXt V2-Tiny has about 38% more parameters than EfficientNetV2-S, which provides more capacity. Third, its larger 7 × 7 depthwise kernels and GRN may favour the aggregation of context over wider regions. These explanations are hypotheses that would require ablation, for example comparing ConvNeXt V2 weights with and without FCMAE pretraining, to test.

From a practical perspective, the combination of higher accuracy and lower single-image latency on the tested GPU makes ConvNeXt V2-Tiny the more attractive of the two configurations for a server-based diagnostic service. The ranking may differ on

mobile CPUs or edge accelerators, where depthwise convolutions and memory bandwidth behave differently, so latency should be measured on the intended deployment device before a choice is made.

The grouped split used in this study also has consequences for comparison with the literature. When nearly identical images are allowed to fall into both the training and test partitions, test accuracy partly measures memorization, and the estimated advantage of one model over another can be distorted (Xu et al., 2023). By quarantining label conflicts, removing duplicates, and keeping perceptually similar images in the same partition, this study accepts a potentially lower but more defensible estimate. The scores reported here should therefore not be compared directly with results obtained on randomly split versions of the same dataset.

Because the dataset was collected in Central Java with smartphone cameras, the results are directly relevant to the development of smartphone-based tools for Indonesian potato growers. Such tools would face additional variation not represented in the internal test set, including other regions, cultivars, growth stages, and lighting conditions; low-light acquisition, for instance, has been shown to require dedicated modelling strategies (Gondal et al., 2026). Field deployment would also need a way to handle images that do not contain a potato leaf or that show several disorders at once, which the seven-class formulation does not cover.

A methodological detail also affects the interpretation of the training curves. The models were optimized with weighted cross-entropy, but checkpoints were selected using the unweighted validation cross-entropy. The selection criterion therefore favours checkpoints that are well calibrated on the overall validation distribution, which is dominated by the larger classes, rather than checkpoints that maximize performance on the small classes. Selecting checkpoints by validation macro-F1 instead could change the chosen epochs and the per-class balance of errors, and is a reasonable alternative to examine in future experiments.

Limitations

This comparison applies to two complete configurations with their respective initial weighting schemes, and the influence of architecture and pretraining has not been separated. The Nematode class has only seven test images, so per-class scores are sensitive to a few predictions. The pHash grouping controls detected similarity, while the possibility of photos from the same plant or session remains. External testing has not been carried out, and the scope of the results is limited to the internal test partition. The two models shared one set of hyperparameters and a 224 × 224 input, which may not be optimal for EfficientNetV2-S.

The analysis used three seeds on a single fixed partition, so the bootstrap intervals do not capture variability due to re-splitting; repeated grouped splits or grouped k-fold cross-validation would provide a more robust estimate. Per-class results, confusion matrices, and training and validation curves were not included, so convergence within 15 epochs and the per-class balance of errors could not be examined directly.

CONCLUSION

ConvNeXt V2-Tiny achieved a test accuracy of 87.97% and a macro-F1 of 85.21%, exceeding EfficientNetV2-S, which achieved 83.22% and 79.86%. The macro-F1 difference of 5.34 percentage points had a 95% group-bootstrap interval of 2.43–8.50 points, although the advantage varied by seed and the seed-42 McNemar test was not significant. ConvNeXt V2-Tiny was also faster on the single Tesla T4 GPU tested (11.70 vs. 25.39 ms per image), and Grad-CAM showed differences in activation spread between the models. These findings indicate that ConvNeXt V2-Tiny is the more suitable candidate for server-based potato leaf diagnosis, while external location-based testing, measurement on deployment devices, and lesion-level validation of Grad-CAM are needed before field use. Methodologically, the results show that model comparisons on small field datasets should rely on leakage-controlled partitions, several training seeds, and paired uncertainty estimates.

FUNDING

The author states that no financial support was received for the implementation of this research or the publication of this article.

CONFLICT OF INTEREST

The author states that this research was conducted without any commercial or financial relationship that could be interpreted as a potential conflict of interest.

AUTHOR CONTRIBUTION

Yoga Pamungkas: conceptualization, methodology, software, formal analysis, visualization, and writing original draft, review, and editing.

GENERATIVE ARTIFICIAL INTELLIGENCE USAGE STATEMENT

The authors state that generative artificial intelligence (AI) was used in the preparation of this manuscript. The authors acknowledge the use of generative AI to help improve language clarity and improve the grammatical structure of manuscripts. The use is carried out in compliance with applicable ethical guidelines regarding the use and disclosure of generative AI in scientific research. Each author has made substantial intellectual contributions to this work. The entire content of the manuscript has been carefully reviewed to ensure its accuracy, and the author is solely responsible for the integrity and final content of the manuscript

REFERENCES

Alhammad, S. M., Khafaga, D. S., El-hady, W. M., Samy, F. M., & Hosny, K. M. (2025). Deep learning and explainable AI for classification of potato leaf diseases. *Frontiers in Artificial Intelligence*, *7*. https://doi.org/10.3389/frai.2024.1449329

Binder, A., Weber, L., Lapuschkin, S., Montavon, G., Müller, K.-R., & Samek, W. (2023). Shortcomings of top-down randomization-based sanity checks for evaluations of deep neural network explanations. *2023 IEEE/CVF Conference on*

Computer Vision and Pattern Recognition (CVPR), 16143–16152. https://doi.org/10.1109/CVPR52729.2023.01549

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, *2*(11), 665–673. https://doi.org/10.1038/s42256-020-00257-z

Gondal, H. A. H., Im, S. J., Tariq, M. H., Irfan, M., Usman, M., Zeshan, M. A., & Park, K. R. (2026). Seeing through the dark: adaptive switch-guided knowledge distillation via heterogeneous teachers for robust plant disease diagnosis. *Journal of King Saud University Computer and Information Sciences*, *38*(7), 826. https://doi.org/10.1007/s44443-026-01238-9

Gülmez, B. (2025). A Comprehensive Review of Convolutional Neural Networks based Disease Detection Strategies in Potato Agriculture. *Potato Research*, *68*(2), 1295–1329. https://doi.org/10.1007/s11540-024-09786-1

Indra, A., Yunita, F., & Usman. (2026). Sugarcane plant disease classification based on leaf image using ConvNeXt V2 deep learning model. *Intechno Journal: Information Technology Journal*, *8*(1), 22–30. https://doi.org/10.24076/intechnojournal.2026v8i1.2876

Kondaveeti, H. K., & Simhadri, C. G. (2025). Evaluation of deep learning models using explainable AI with qualitative and quantitative analysis for rice leaf disease detection. *Scientific Reports*, *15*, 31850. https://doi.org/10.1038/s41598-025-14306-3

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A ConvNet for the 2020s. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11966–11976. https://doi.org/10.1109/CVPR52688.2022.01167

Natarajan, S., Chakrabarti, P., & Margala, M. (2024). Robust diagnosis and meta visualizations of plant diseases through deep neural architecture with explainable AI. *Scientific Reports*, *14*, 13695. https://doi.org/10.1038/s41598-024-64601-8

Nazir, T., Iqbal, M. M., Jabbar, S., Hussain, A., & Albathan, M. (2023). EfficientPNet—An optimized and efficient deep learning approach for classifying disease of potato plant leaves. *Agriculture*, *13*(4), 841. https://doi.org/10.3390/agriculture13040841

Park, S., Lee, S., Sohn, S.-I., Park, T., & Shin, K. (2025). Specific detection of genetically modified potatoes containing asparagine synthetase-1 and polyphenol oxidase 5 genes derived from potato. *GM Crops & Food*, *16*(1). https://doi.org/10.1080/21645698.2025.2488085

Rainio, O., Teuho, J., & Klén, R. (2024). Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, *14*(1), 6086. https://doi.org/10.1038/s41598-024-56706-x

Rofiqi, H. K., Noersasongko, E., Winarno, S., & Soeleman, M. A. (2026). Augmentation strategy and hyperparameter optimization using Optuna for potato leaf disease classification in uncontrolled environment. *Jurnal Teknik Informatika (JUTIF)*, *7*(2), 743–759. https://doi.org/10.52436/1.jutif.2026.7.2.4898

Sahara, D., Setiani, C., Swastika, D. K. S., Asnawi, R., Sugiman, S. B., Fadwiwati, A. Y., Suhendrata, T., Syam, A., Supriyo, A., Atman, & Dewi, T. (2024). Strategy formulation of agricultural machinery development for sustainable potato farming in upland of Banjarnegara Regency, Central Java, Indonesia. *Heliyon*, *10*(21), e39637. https://doi.org/10.1016/j.heliyon.2024.e39637

Sangar, G., & Rajasekar, V. (2025). Optimized classification of potato leaf disease using EfficientNet-LITE and KE-SVM in diverse environments. *Frontiers in Plant Science*, *16*, 1499909. https://doi.org/10.3389/fpls.2025.1499909

Savary, S., Willocquet, L., Pethybridge, S. J., Esker, P., McRoberts, N., & Nelson, A. (2019). The global burden of pathogens and pests on major food crops. *Nature Ecology & Evolution*, *3*(3), 430–439. https://doi.org/10.1038/s41559-018-0793-y

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *2017 IEEE International Conference on Computer Vision (ICCV)*, 618–626. https://doi.org/10.1109/ICCV.2017.74

Shabrina, N. H., Indarti, S., Maharani, R., Kristiyanti, D. A., Irmawati, Prastomo, N., & Adilah M, T. (2024). A novel dataset of potato leaf disease in uncontrolled environment. *Data in Brief*, *52*, 109955. https://doi.org/10.1016/j.dib.2023.109955

Shoaib, M., Shah, B., El-Sappagh, S., Ali, A., Ullah, A., Alenezi, F., Gechev, T., Hussain, T., & Ali, F. (2023). An advanced deep learning models-based plant disease detection: A review of recent research. *Frontiers in Plant Science*, *14*, 1158933. https://doi.org/10.3389/fpls.2023.1158933

Sinamenye, J. H., Chatterjee, A., & Shrestha, R. (2025). Potato plant disease detection: Leveraging hybrid deep learning models. *BMC Plant Biology*, *25*, 647. https://doi.org/10.1186/s12870-025-06679-4

Tan, M., & Le, Q. V. (2021). EfficientNetV2: Smaller models and faster training. *Proceedings of the 38th International Conference on Machine Learning, Proceedings of Machine Learning Research*, *139*, 10096–10106.

Varoquaux, G., & Colliot, O. (2023). Evaluating machine learning models and their diagnostic value. In O. Colliot (Ed.), *Machine learning for brain disorders* (Vol. 197, pp. 601–630). Humana. https://doi.org/10.1007/978-1-0716-3195-9_20

Vo, T. K. A., Chau, N. V., & Vo, T. N. (2025). A comparative analysis of modern CNN architectures: EfficientNetV2 vs. ConvNeXt for plant disease classification. *2025 RIVF International Conference on Computing and Communication Technologies (RIVF)*, 844–847. https://doi.org/10.1109/RIVF68649.2025.11365177

Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I. S., & Xie, S. (2023). ConvNeXt V2: Co-designing and scaling ConvNets with masked autoencoders. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16133–16142. https://doi.org/10.1109/CVPR52729.2023.01548

Xu, M., Kim, H., Yang, J., Fuentes, A., Meng, Y., Yoon, S., Kim, T., & Park, D. S. (2023). Embracing limited and imperfect training datasets: Opportunities and challenges in plant disease recognition using deep learning. *Frontiers in Plant Science*, *14*,

1225409. https://doi.org/10.3389/fpls.2023.1225409

Xu, M., Park, J.-E., Lee, J., Yang, J., & Yoon, S. (2024). Plant disease recognition datasets in the age of deep learning: Challenges and opportunities. *Frontiers in Plant Science*, *15*, 1452551. https://doi.org/10.3389/fpls.2024.1452551