Journal of Agricultural Sciences (JAS) 32(3):701-717
DOI: 10.15832/ankutbd.1819492

J Agr Sci-Tarim Bili
e-ISSN: 2148-9297
jas.ankara.edu.tr

Explainable Deep Learning for Plant Leaf Diseases: A Comparative Study of Grad-CAM

Adnan Gökten[a*], Erkut Tekeli[b], Hasan Beytullah Dönmez[c]

[a]*Department of Computer Technology, Kozan Vocational School, Çukurova University, Adana, TÜRKIYE*
[b]*Department of Software Engineering, Faculty of Engineering, Muğla Sıtkı Koçman University, Muğla, TÜRKIYE*
[c]*Department of Field Crops, Faculty of Agriculture, Çukurova University, Adana, TÜRKIYE*

ARTICLE INFO
Research Article
Corresponding Author: Adnan Gökten, E-mail: agokten@cu.edu.tr
Received: 07 November 2025 / Revised: 03 February 2026 / Accepted: 28 February 2026 / Online: 28 July 2026

Cite this article
Gökten A, Tekeli E, Dönmez H B (2026). Explainable Deep Learning for Plant Leaf Diseases: A Comparative Study of Grad-CAM. *Journal of Agricultural Sciences (Tarim Bilimleri Dergisi),* 32(3):701-717. DOI: 10.15832/ankutbd.1819492

ABSTRACT

Plant disease detection is critical for sustainable agriculture and food security. While deep learning models achieve high accuracy in leaf disease classification, their black box nature poses limitations for trust and adoption among agricultural practitioners. This study presents a comparative evaluation of three convolutional neural network architectures (ConvNeXt-Tiny, MobileNetV2, and VGG16) for classifying potato, maize, and pepper leaf diseases, with emphasis on explainability through Gradient-weighted Class Activation Mapping (Grad-CAM). The experimental results demonstrate that ConvNeXt-Tiny achieves 99-100% accuracy across all plant species, MobileNetV2 attains 97-100% accuracy with lower computational requirements, and VGG16 yields 97-99.5% accuracy. Grad-CAM visualizations reveal that modern architectures precisely focus on lesion regions, whereas older models occasionally attend to irrelevant features such as leaf veins and edges. Misclassification analysis identifies shadows and natural leaf patterns as primary error sources. This research demonstrates that explainable artificial intelligence is not merely complementary but essential for developing trustworthy agricultural decision support systems.

Keywords: Explainable artificial intelligence, Grad-CAM, Plant disease classification, Agricultural decision support

1. Introduction

Agriculture serves as the foundation of global food security and sustainable development. Plant diseases cause significant yield losses worldwide, and this challenge is further intensified by climate change dynamics (Savary et al. 2019). Environmental stress factors, including water scarcity, temperature fluctuations, and expanding pest distributions, pose serious threats to agricultural production systems. Consequently, early disease detection and accurate classification have become increasingly vital for farmers and agricultural experts (Ferentinos 2018).

Recent advances in artificial intelligence (AI), particularly deep learning-based image processing techniques, have shown promising potential for classifying plant leaf diseases (Liakos et al. 2018; Benos et al. 2021). Deep learning architectures can model nonlinear relationships that surpass the limitations of classical statistical methods, thereby achieving superior accuracy rates (Breiman 2001; Ryo & Rillig 2017). However, these high-performance systems possess a critical limitation: a lack of interpretability. This phenomenon, known as the "black box problem," creates trust issues, particularly in domains such as agriculture, where decisions have direct economic and social consequences (Rudin 2019; Rudin et al. 2022).

Explainable Artificial Intelligence (XAI) addresses this challenge by ensuring that models not only provide high accuracy but also reveal which features and data elements influence their decisions (Doshi-Velez & Kim 2017; Adadi & Berrada 2018). In agricultural applications where end users are farmers and agronomists, model interpretability becomes critically important. Researchers have reported that agricultural practitioners experience difficulty trusting unexplained black box models (Barbedo 2016; Meske & Bunde 2020).

Visualization-based explanation methods are among the most common XAI applications in agriculture. Grad-CAM is widely used for visualizing the CNN decision mechanism. Grad-CAM explains the rationale behind classifications by showing through heatmaps which leaf regions the model focuses on during prediction (Selvaraju et al. 2017). While Grad-CAM has been applied in agricultural image classification studies, it has typically been implemented on single models (Lin et al. 2019; Kundu et al. 2021). Recent research has demonstrated that XAI methods in smart agriculture contexts enhance model decision interpretability beyond mere accuracy scores and provide transparency in leaf disease classification (Wei et al. 2022; Quach et al. 2023). Similarly, Ryo (2022) and Gardezi et al. (2023) emphasized that explainable artificial intelligence methods play an important role in increasing user trust and adoption of agricultural decision support systems.

In agricultural applications, Kundu et al. (2021) showed that Grad-CAM revealed relevant features extracted by their Custom-Net model when classifying pearl millet leaf diseases, and that transfer learning improved feature extraction. Kundu et al. (2022) performed feature visualization using Grad-CAM on their MaizeNet model for maize leaf disease detection and severity estimation, demonstrating that the model correctly focused on disease regions, achieving 98.50% accuracy. In a study by Wei et al. (2022), despite achieving high accuracy on the PlantVillage dataset, Grad-CAM visualizations revealed that in some classes the model focused on leaf edges or background features rather than disease symptoms, highlighting that misfocus can lead to prediction errors.

Additionally, methods such as SHAP (Lundberg & Lee 2017) and LIME (Ribeiro et al. 2016) have been applied in the literature. However, while these methods provide strong explanations, particularly for tabular and numerical data, they are not as intuitive as Grad-CAM in terms of visualization for image-based classifications such as leaf diseases. While SHAP and LIME produce superpixel-based explanations in image classification, Grad-CAM provides more direct and visually comprehensible explanations through heatmaps (Linardatos et al. 2020). For this reason, Grad-CAM has been more frequently preferred in agricultural image classification studies in recent years, particularly for explaining CNN-based models.

XAI is also observed to be on the rise in artificial intelligence research generally. Ryo (2022), examining the increase in published articles between 2011-2021, demonstrated a dramatic increase in XAI and interpretable machine learning topics, especially after 2018. This indicates a growing need for explainability not only in agricultural applications but also in diverse fields such as medicine, climate science, and autonomous systems (Başağaoğlu et al. 2022).

When evaluated in terms of CNN architectures, the literature reveals examination of different models from an explainability perspective. VGG16 (Simonyan & Zisserman 2015), despite being an older architecture, remains a classic model still used for comparison purposes. MobileNetV2 (Sandler et al. 2018) stands out particularly for its applicability in field conditions due to low hardware requirements. ConvNeXt, a modern CNN developed with inspiration from transformer architectures, has attracted attention in recent years with its high accuracy capacity (Liu et al. 2022). However, comparative examination of these three models in the same datasets within an XAI context has remained limited in the literature.

Recent studies have made significant contributions to XAI applications in plant disease classification; however, several gaps remain. Natarajan et al. (2024) achieved 99.95% accuracy using a fine-tuned deep CNN with Grad-CAM on the PlantVillage dataset, but their analysis was limited to a single architecture without comparative evaluation across different model types. Similarly, Karim et al. (2024) deployed a modified MobileNetV3Large model with Grad-CAM for real-time grape leaf disease detection, demonstrating the potential of lightweight architectures for edge deployment; however, their study focused on a single crop and a single model. Alhammad et al. (2025) applied Grad-CAM for potato leaf disease classification but did not provide quantitative metrics for evaluating attention map accuracy. The present study addresses these gaps by: (1) providing a systematic comparison of three architecturally distinct CNN models (modern high-performance, lightweight deployable, and classical baseline) within a unified XAI framework, (2) including both qualitative Grad-CAM visualizations and quantitative IoU-based metrics for objective assessment of attention accuracy, (3) analyzing misclassification patterns to identify model-specific attention biases, and (4) examining multiple crop species to assess cross-dataset generalizability.

This study aims to comparatively examine different CNN architectures and explain their decision mechanisms through the Grad-CAM methodology. For this purpose, three different models were selected: ConvNeXt, MobileNetV2, and VGG16. The original contribution of this study involves going beyond comparing quantitative performance metrics (e.g. accuracy, F1-score) to explain the decision mechanisms of these three models through Grad-CAM visualizations. Secondly, it discusses the results in the context of agricultural decision support systems.

2. Material and Methods

2.1. Dataset

This study utilized the PlantVillage segmented dataset, one of the most widely used open data sources for plant leaf disease classification (Hughes & Salathé 2015). The dataset contains 8479 samples from three plant species (maize, potato, and pepper), encompassing both healthy specimens and various disease classes. The distribution of samples by plant is presented in Table 1. Each class receives balanced representation, enabling comparison of learning tendencies across different classes for the models.

Table 1- Number of samples for maize, potato, and pepper plants in the PlantVillage segmented dataset

Plant	*Sample size*
Maize	3852
Potato	2152
Pepper	2475
Total	8479

The data was segmented into three distinct subsets: training (70%), validation (15%), and test (15%). The partitioning process utilised a random but deterministic selection method, with a seed value of 42. This ensured reproducibility and consistency. Additionally, selecting the segmented version of the dataset clarified the leaf-background distinction, facilitating models to focus solely on biologically meaningful regions. This aspect strengthens the interpretability of explainable artificial intelligence (XAI) techniques.

2.2. Preprocessing

The images were resized to 224×224 pixels and normalized using mean-standard deviation values from the ImageNet dataset. This approach enables more effective use of pre-trained weights during transfer learning (Krizhevsky et al. 2012). Data augmentation was applied during training to enhance model generalization capacity: processes including random rotation, horizontal flipping, brightness variation, and mild blurring simulated image diversity that could be encountered in field conditions. Consequently, models were designed to produce more stable results not only on dataset-specific samples but also in real-world conditions.

2.3. Deep learning architectures

Three CNN architectures representing different generations were used:

• ConvNeXt-Tiny: A modern architecture combining transformer design principles with CNN structure (Liu et al. 2022). Beyond providing high accuracy, it enables the evaluation of contemporary architectures' decision mechanisms in explainability analyses.

• MobileNetV2: An architecture suitable for field use on mobile devices, capable of operating with low computational cost thanks to depthwise separable convolutions (Sandler et al. 2018). This ensures the study has implications not only at the laboratory scale but also for practical agricultural applications.

• VGG16: VGG16, although considered an older architecture, was deliberately selected as a classical baseline. In addition to providing a comparison with earlier literature, it also serves an educational and interpretability purpose. Because many end-users of agricultural decision-support systems (e.g., farmers and agronomists) may not have a technical background in deep learning, using a simpler and older architecture allows clearer visualization of Grad-CAM attention differences. This makes it easier to demonstrate how modern architectures focus more accurately on disease regions compared to traditional networks.

Therefore, the selected models represent three complementary categories: modern high-performance (ConvNeXt), lightweight deployable (MobileNet), and classical baseline (VGG16), enabling both technical and practical evaluation.

All models were initialized with pre-trained weights on ImageNet, and output layers were reconfigured according to appropriate class numbers for each plant species.

2.4. Training strategy

Models were trained using 5-fold cross-validation (K-Fold CV). This approach tests whether models demonstrate stable performance across different subsets without depending on a single data split. Training hyperparameters include: Epochs: 30, Batch size: 32, Optimization: Adam, Learning rate: 0.0001, Loss function: CrossEntropyLoss.

Early stopping (patience=3) was applied to prevent overfitting, terminating training when validation loss ceased improving. Best validation weights were saved for each fold, and the "BEST" model was determined. The final test phase adopted an ensemble approach combining all fold results, with final test scores reported accordingly. Statistical significance between models was determined using paired t-tests on the test accuracy scores from all five folds (α=0.05).

2.5. Evaluation metrics

Performance evaluation used accuracy and weighted F1-score as primary metrics. Additionally, confusion matrices were calculated and visualized to analyze the effects of class imbalances (Sokolova & Lapalme 2009). This enabled detailed examination of model strengths and weaknesses in each class.

2.6. Explainable artificial intelligence analysis

Gradient-weighted Class Activation Mapping (Grad-CAM) was applied to enhance model decision transparency (Selvaraju et al. 2017). Grad-CAM tests whether decision mechanisms are biologically meaningful by visualizing regions that models focus on during classification through heatmaps.

A bidirectional strategy was adopted:

- Correct classifications: Examining which leaf regions models observe to make correct decisions, evaluating focus on lesion areas.

- Misclassifications: Analysis of visual biases (shadows, vein structures, light reflections, etc.) leading to decision errors.
This method evaluated not only accuracy scores but also the reliability of decision processes.

2.7. Quantitative evaluation of Grad-CAM explanations

To complement the qualitative Grad-CAM visualizations, a quantitative explainability analysis was conducted.

Since the PlantVillage segmented dataset does not provide ground-truth lesion masks, a proxy lesion mask was automatically generated.

First, a leaf mask was extracted by separating the leaf region from the dark background using grayscale thresholding and morphological operations. Then, potential lesion regions were estimated within the leaf by detecting color anomalies in the CIELAB color space. Pixels with the highest color deviation (top 10%) were considered proxy lesion areas.

Grad-CAM heatmaps were binarized and compared with these proxy lesion masks. The overlap between the attention regions and lesion areas was measured using Intersection-over-Union (IoU). Additionally, mean activation intensity inside and outside the lesion regions was computed.

This approach enabled an objective and reproducible comparison of the attention behavior of different CNN architectures.

2.8. Implementation environment

All experiments were conducted on Google Colab with GPU (Tesla T4) support. Modeling was performed using PyTorch (Paszke et al. 2019), while Scikit-learn (Pedregosa et al. 2011), Matplotlib (Hunter 2007), and Seaborn (Waskom 2021) libraries were used for data processing and visualizations. The PyTorch-Grad-Cam (Gildenblat 2021) library was utilized for explainability analyses.

3. Results

In this study, disease classification was performed on potato, maize, and pepper leaves using ConvNeXt-Tiny, MobileNetV2, and VGG16 deep learning architectures. The decision mechanisms of the models were analyzed through Grad-CAM visualizations.

3.1. Classification performance

The classification performance of the models across five-fold cross-validation is given in Table 2 and Table 3.

Table 2- F1 values of three different CNN models in classifying potato, maize, and pepper leaf diseases

Dataset	*Model*	*Fold 1*	*Fold 2*	*Fold 3*	*Fold 4*	*Fold 5*	*Mean*	*Std*	*p-value**
Maize	ConvNeXt-Tiny	0.990	0.980	0.979	0.984	0.988	0.984	0.005	-
	MobileNetV2	0.983	0.988	0.976	0.985	0.981	0.983	0.004	0.6010
	VGG16	0.976	0.983	0.963	0.970	0.976	0.974	0.007	0.0375**
Pepper	ConvNeXt-Tiny	1.000	1.000	0.997	0.997	0.997	0.998	0.001	-
	MobileNetV2	0.997	1.000	1.000	0.989	0.997	0.997	0.004	0.4382
	VGG16	0.995	0.995	0.995	0.987	0.997	0.994	0.004	0.0599
Potato	ConvNeXt-Tiny	1.000	1.000	1.000	1.000	1.000	1.000	0.000	-
	MobileNetV2	1.000	1.000	0.997	1.000	1.000	0.999	0.001	0.3739
	VGG16	1.000	1.000	0.991	1.000	1.000	0.998	0.004	0.3739

*: Paired t-test compared to ConvNeXt-Tiny (α=0.05, two-tailed), **: Statistically significant difference (P<0.05)

Table 3- Accuracy (%) of three different CNN models in classifying potato, maize, and pepper leaf diseases

Dataset	*Model*	*Fold 1*	*Fold 2*	*Fold 3*	*Fold 4*	*Fold 5*	*Mean*	*Std*	*p-value**
Maize	ConvNeXt-Tiny	98.969	97.938	97.938	98.454	98.797	98.419	0.477	-
	MobileNetV2	98.282	98.797	97.595	98.454	98.110	98.247	0.445	0.5826
	VGG16	97.595	98.282	96.392	97.079	97.595	97.388	0.702	0.0415**
Pepper	ConvNeXt-Tiny	100.000	100.000	99.733	99.733	99.733	99.840	0.146	-
	MobileNetV2	99.733	100.000	100.000	98.930	99.733	99.679	0.439	0.4261
	VGG16	99.465	99.465	99.465	98.663	99.733	99.358	0.406	0.0533
Potato	ConvNeXt-Tiny	100.000	100.000	100.000	100.000	100.000	100.000	0.000	-
	MobileNetV2	100.000	100.000	99.691	100.000	100.000	99.938	0.138	0.3739
	VGG16	100.000	100.000	99.074	100.000	100.000	99.815	0.414	0.3739

*: Paired t-test compared to ConvNeXt-Tiny (α=0.05, two-tailed), **: Statistically significant difference ($P<0.05$)

As shown in Table 2 and Table 3, to ensure statistical reliability and to evaluate model stability, 5-fold cross-validation results are reported individually for each fold. In addition to the average performance, the standard deviation values are also presented. The low variance across folds indicates that the proposed models exhibit stable and consistent performance rather than achieving high accuracy due to a favorable single Split.

The obtained results reveal that modern deep learning architectures achieve high success in detecting plant leaf diseases. The ConvNeXt-Tiny model exhibited the highest performance across all three plant species. It achieved 98.42% accuracy in the maize dataset, 100% accuracy in potato, and 99.84% accuracy in the pepper datasets. Paired t-tests revealed statistically significant differences between ConvNeXt-Tiny and VGG16 for maize ($P<0.05$) but there is no significant difference between ConvNeXt-Tiny and MobileNetV2.

For potato classification, all three models performed exceptionally well (≥99.81%), with no statistically significant differences between ConvNeXt-Tiny and the others. Similarly, in pepper classification, all three models achieved perfect accuracy (≥99.35%), showing no significant difference ($P>0.05$).

3.2. Class-Based performance analysis

When examining detailed class-based performance metrics, models were found to demonstrate varying success levels across different disease types.

3.2.1. Maize leaf diseases

The maize dataset contains four different classes: Cercospora leaf spot, Common rust, Northern Leaf Blight, and healthy. To better analyse the learning dynamics of the models, the training and validation loss and accuracy values were recorded for each epoch during the training process of maize. Based on these values, learning curves were generated to visualize the convergence behaviour of the models. In addition, class-wise ROC-AUC curves were computed on the test set to evaluate the discriminative performance. According to Figure 1, allow the assessment of model stability, detection of potential overfitting, and examination of the effectiveness of the early stopping strategy.

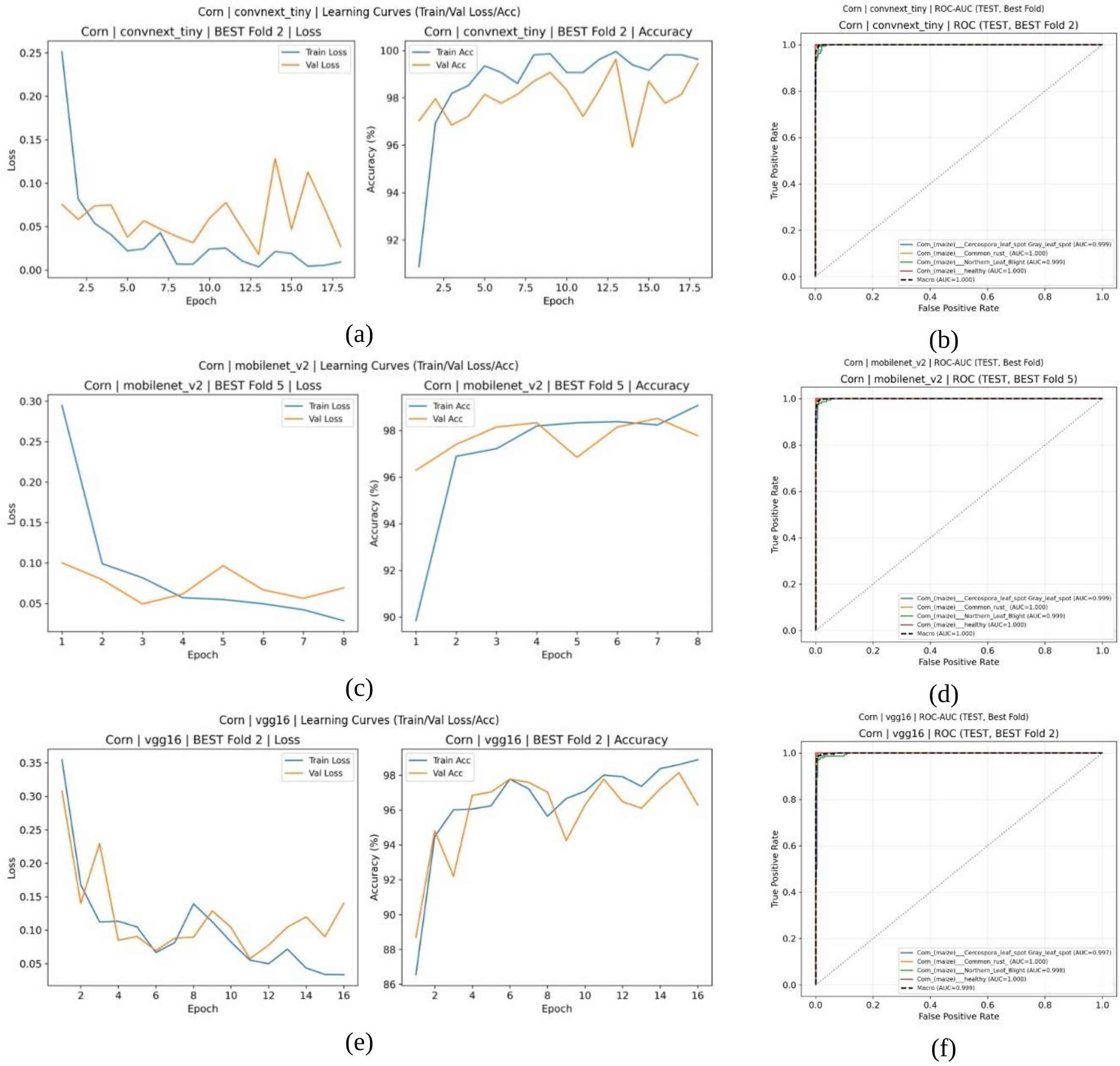


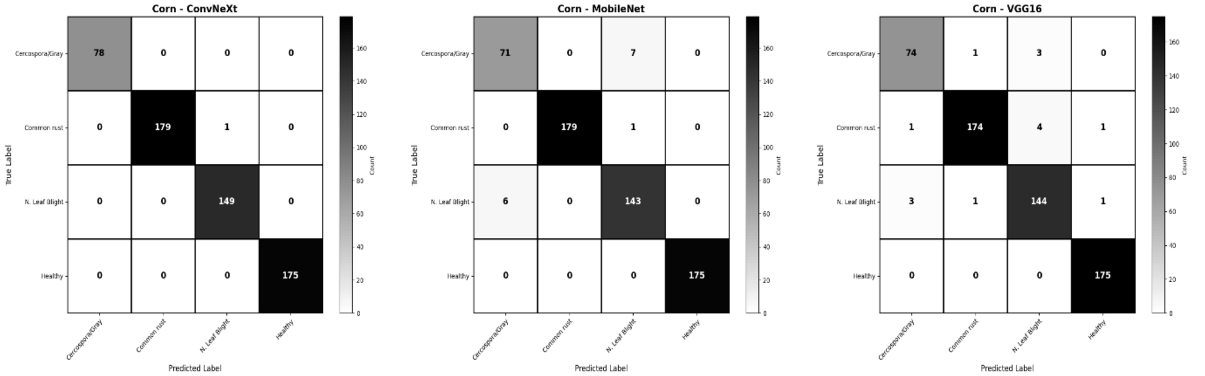
Figure 1- Learning dynamics of maize: (a) learning curves of ConvNeXt-Tiny , (b) ROC curves of ConvNeXt-Tiny, (c) learning curves of MobileNetV2, (d) ROC curves of MobileNetV2, (e) learning curves of VGG16, (f) ROC curves of VGG16

In the ConvNeXt-Tiny model, precision, recall, and F1-score values ranged from 0.99-1.00 across all classes (Table 4).

Table 4- Performance metrics of the ConvNeXt-Tiny model for maize plant disease classification

Classes	*Precision*	*Recall*	*F1-Score*
Maize - Cercospora leaf spot	1.00	1.00	1.00
Maize - Common rust	1.00	0.99	1.00
Maize - Northern Leaf Blight	0.99	1.00	1.00
Maize - Healthy	1.00	1.00	1.00

Confusion matrix analysis revealed that only one Common rust sample was misclassified as Northern Leaf Blight (Fig. 2a).



(a) (b) (c)

Figure 2- Confusion matrix results of three different CNN models for the maize plant: (a) ConvNeXt-Tiny, (b) MobileNetV2, (c) VGG16. Each row represents the actual class, while each column represents the predicted class

In the MobileNetV2 model, the precision of 0.92 and recall of 0.91 in the Cercospora leaf spot class remained lower compared to other classes (Table 5). Examination of the confusion matrix revealed that 7 out of 78 Cercospora samples were misclassified as Northern Leaf Blight, and additionally, 6 Northern Leaf Blight samples were predicted as Cercospora (Fig. 2b). This situation may have resulted from both disease types displaying similar visual symptoms.

Table 5- Performance metrics of the MobileNetV2 model for maize plant disease classification

Classes	*Precision*	*Recall*	*F1-Score*
Maize - Cercospora leaf spot	0.92	0.91	0.92
Maize - Common rust	1.00	0.99	1.00
Maize - Northern Leaf Blight	0.95	0.96	0.95
Maize – Healthy	1.00	1.00	1.00

A similar pattern was observed in the VGG16 model (Table 6). In the Cercospora class, precision was 0.95 and recall was 0.95, with 3 samples incorrectly predicted as Northern Leaf Blight. Additionally, confusion matrix examination revealed that 4 Common rust samples were classified as Northern Leaf Blight, while 3 Northern Leaf Blight samples were classified as Cercospora (Fig. 2c).

Table 6- Performance metrics of the VGG16 model for maize plant disease classification

Classes	*Precision*	*Recall*	*F1-Score*
Maize - Cercospora leaf spot	0.95	0.95	0.95
Maize - Common rust	0.99	0.97	0.98
Maize - Northern Leaf Blight	0.95	0.97	0.96
Maize - Healthy	0.99	1.00	0.99

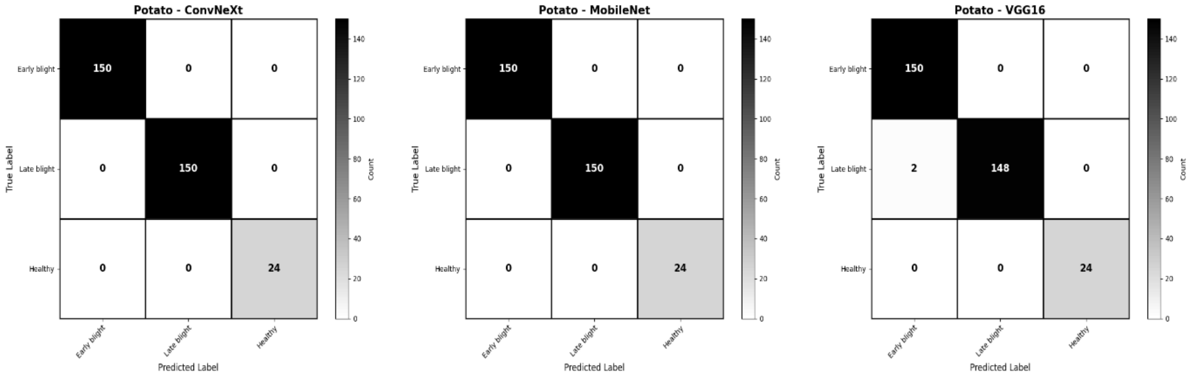
3.2.2. Potato leaf diseases

The potato dataset contains three classes: Early blight, Late blight, and healthy. In ConvNeXt-Tiny and MobileNetV2 models, precision, recall, and F1-score values were 1.00 across all classes (Table 7).

Table 7- Performance metrics of ConvNeXt-Tiny and MobileNetV2 models for potato plant disease classification

Classes	*Precision*	*Recall*	*F1-Score*
Potato - Early blight	1.00	1.00	1.00
Potato - Late blight	1.00	1.00	1.00
Potato - Healthy	1.00	1.00	1.00

Analysis of the confusion matrix revealed that the ConvNeXt-Tiny and MobileNetV2 models achieved perfect classification with no misclassifications for potato diseases (Fig. 3a, 3b). This finding indicates that potato leaf diseases possess more visually discriminative characteristics.



(a) (b) (c)

Figure 3- Confusion matrix results of three different CNN models for potato plant: (a) ConvNeXt-Tiny, (b) MobileNetV2, (c) VGG16. Each row represents the actual class, while each column represents the predicted class

In the VGG16 model, precision, recall, and F1-score values ranged between 0.99 and 1.00 across all classes (Table 8).

Table 8- Performance metrics of the VGG16 model for potato plant disease classification

Classes	*Precision*	*Recall*	*F1-Score*
Potato - Early blight	0.99	1.00	0.99
Potato - Late blight	1.00	0.99	0.99
Potato - Healthy	1.00	1.00	1.00

Analysis of the confusion matrix revealed that the VGG16 model misclassified 2 out of 150 Late blight instances as Early blight (Fig. 3c).

3.2.3. Pepper leaf diseases

The pepper dataset comprises two classes: Bacterial spot and healthy. In the ConvNeXt-Tiny and VGG16 models, precision, recall, and F1-score values were 1.00 across all classes (Table 9).

Table 9- Performance metrics of ConvNeXt-Tiny and VGG16 models for pepper plant disease classification

Classes	*Precision*	*Recall*	*F1-Score*
Pepper - Bacterial spot	1.00	1.00	1.00
Pepper - Healthy	1.00	1.00	1.00

Analysis of the confusion matrix revealed that the ConvNeXt-Tiny and VGG16 models achieved perfect performance in pepper classification, with all samples correctly predicted (Fig. 4a, 4c).

(a) (b) (c)

Figure 4- Confusion matrix results of three different CNN models for pepper plant: (a) ConvNeXt-Tiny, (b) MobileNetV2, (c) VGG16. Each row represents the actual class, while each column represents the predicted class

In the MobileNetV2 model, precision, recall, and F1-score values ranged between 0.99 and 1.00 across all classes (Table 10).

Table 10- MobileNetV2 model performance metrics for pepper plant disease classification

Classes	Precision	Recall	F1-Score
Pepper - Bacterial_spot	1.00	0.99	1.00
Pepper- Healthy	1.00	1.00	1.00

Analysis of the confusion matrix revealed that the MobileNetV2 model misclassified only one Bacterial spot instance as healthy (Fig. 4b).

3.3. Explainability analysis through Grad-CAM visualizations

Grad-CAM visualizations were conducted for each model to provide transparency in model decision-making. These analyses elucidate the specific leaf regions on which the models focus during classification and allow for the evaluation of whether biologically meaningful decision mechanisms are utilized.

The Grad-CAM heat maps for correctly classified samples of maize, potato, and pepper leaves are presented in Figures 5, 6, and 7, respectively.

3.3.1. Maize leaf diseases

In the classification of Northern Leaf Blight disease in maize leaves, the ConvNeXt-Tiny and MobileNetV2 models demonstrated clear focus on lesion regions. Particularly for Northern Leaf Blight disease, regions containing long elliptical lesions were prominently highlighted with high activation values in the heat maps (Fig. 5).

In contrast, the VGG16 model exhibited smaller and more fragmented attention regions. This indicates that VGG16 focuses on more localized and specific leaf features (e.g., small necrotic spots, vein structures). However, this localized focus does not fully capture the overall spread of the disease in some cases (Fig. 5).

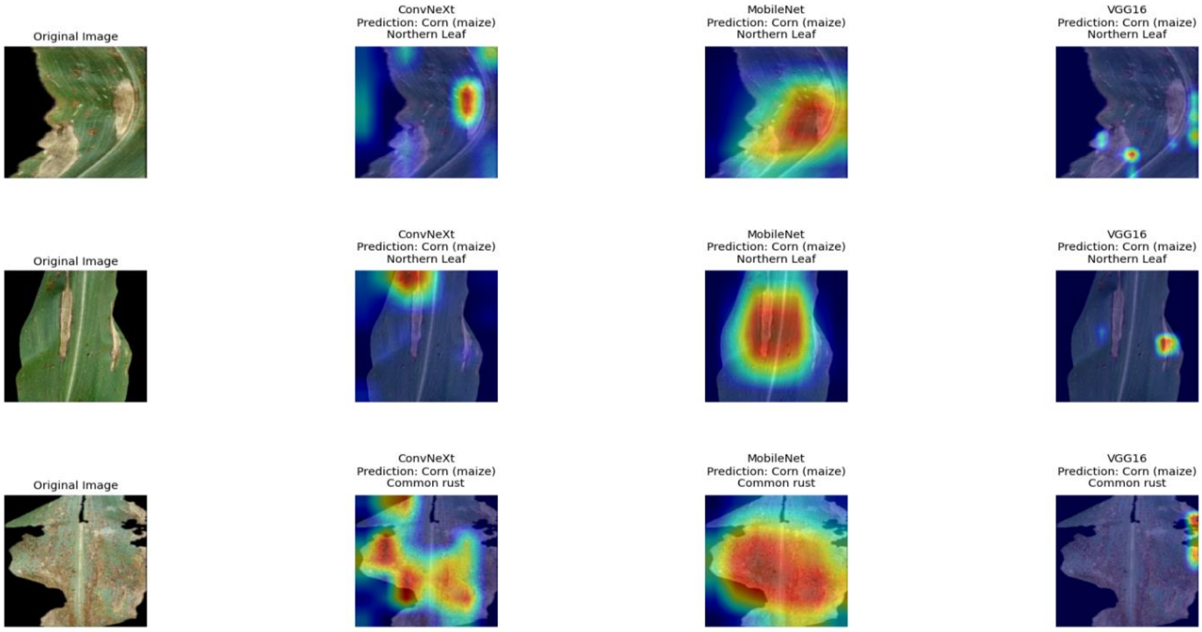


Figure 5- Grad-CAM visualizations for correct classifications of maize leaf disease. From left to right: original image, ConvNeXt-Tiny, MobileNetV2, and VGG16 heat maps. Red regions represent areas where the model exhibits the highest activation

3.3.2. Potato leaf diseases

For potato Early and Late Blight diseases, all models generated highly successful heat maps. The ConvNeXt-Tiny model most precisely delineated the boundaries of disease lesions and attributed high activation values to regions containing necrotic tissues (Fig. 6).

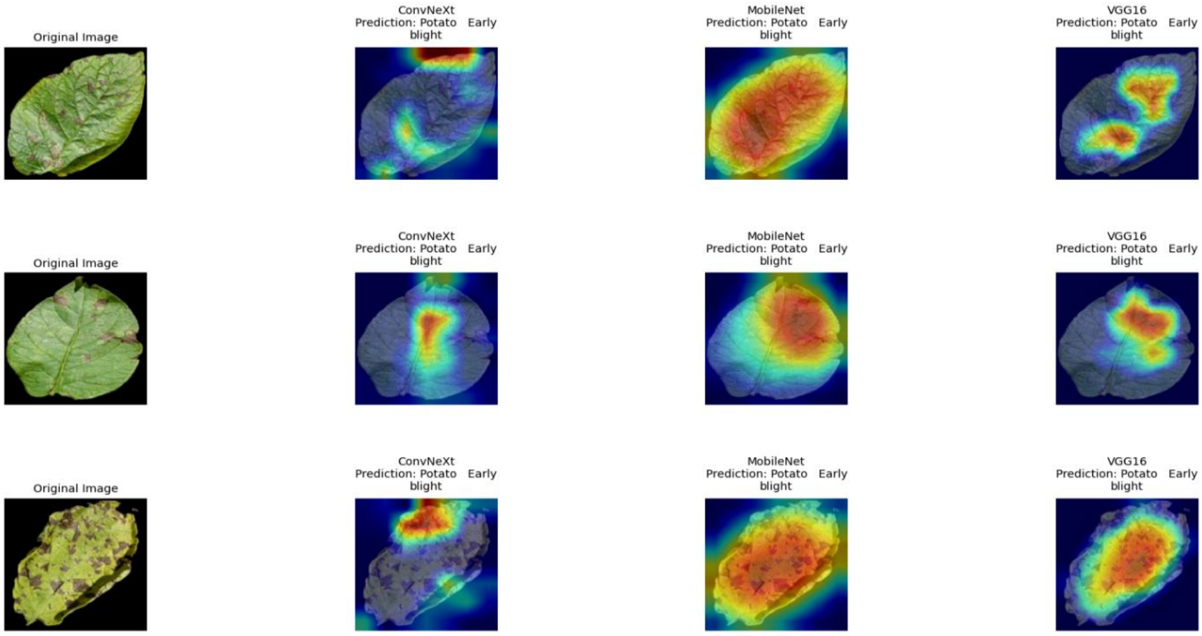


Figure 6- Grad-CAM visualizations for correct classifications of potato Early Blight leaf disease. From left to right: original image, ConvNeXt-Tiny, MobileNetV2, and VGG16 heat maps. Red regions represent areas where the model exhibits the highest activation

MobileNetV2 also focused on diseased regions; however, the attention distribution encompassed a broader area (Fig. 6). This suggests that the model gathers information about tissues surrounding the diseased regions as well.

Although the VGG16 model partially attended to leaf edges and vein structures in some samples, disease symptoms were generally detected successfully (Fig. 6).

3.3.3. Pepper bacterial spot disease

The ConvNeXt-Tiny model demonstrated precise and focused attention on regions containing disease lesions. In the heat maps, it was observed that the red-yellow color tones perfectly corresponded to leaf regions exhibiting bacterial spot symptoms (Fig. 7). This indicates that the model focuses on biologically correct regions when detecting the disease.

The MobileNetV2 model successfully detected lesion regions similarly to ConvNeXt-Tiny; however, the attention distribution spread over a broader area and exhibited a more diffuse pattern. Nevertheless, diseased regions remained at the center of the heat map (Fig. 7).

The VGG16 model exhibited a narrower and more focal attention. In some samples, it was observed that the heat map concentrated only on the lesion center but failed to fully encompass the disease spread area (Fig. 7). Furthermore, the VGG16 model was found to partially attend to leaf edges, suggesting that the model also focuses on certain structural features.

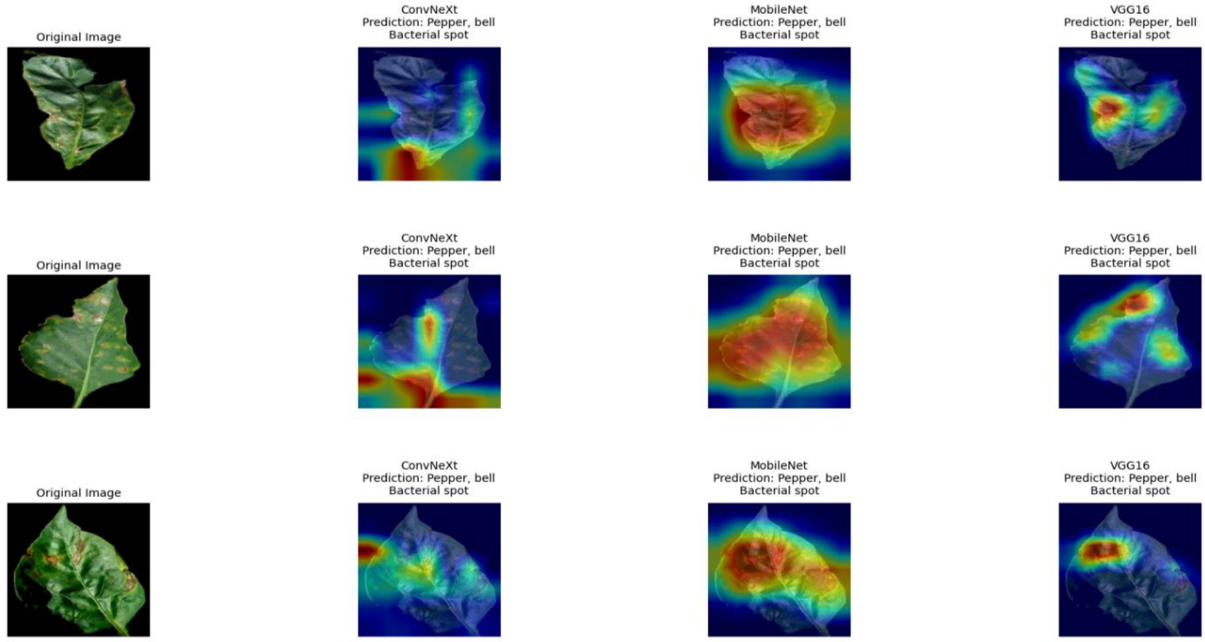


Figure 7- Grad-CAM visualizations for correct classifications of pepper bacterial spot disease. From left to right: original image, ConvNeXt-Tiny, MobileNetV2, and VGG16 heat maps. Red regions represent areas where the model exhibits the highest activation.

3.4. Grad-CAM analysis of misclassifications

Analysis of misclassifications is critically important for understanding which visual features models are overly sensitive to and what types of biases affect them. Grad-CAM heat maps of misclassified samples for maize, potato, and pepper leaves are shown in Figures 8, 9, and 10, respectively.

3.4.1. Maize misclassifications

The most frequently confused disease pairs in the dataset are between Cercospora leaf spot and Northern Leaf Blight. In the first example of Figure 8, Northern Leaf Blight disease was misclassified as Common Rust by VGG16. Upon examination of the Grad-CAM heat map, it is observed that the model focused on small necrotic spots and color changes at the leaf edge. These regions, which show visual similarity to the characteristic small rust-colored pustules of Common Rust disease, led to the model's error.

In the second example, Cercospora leaf spot disease was predicted as Northern Leaf Blight (Fig. 8). The VGG16 heat map focused on longitudinally extending regions of the leaf. For long, elliptical lesion forms can be observed in both disease types, the model is believed to have been influenced by this morphological similarity.

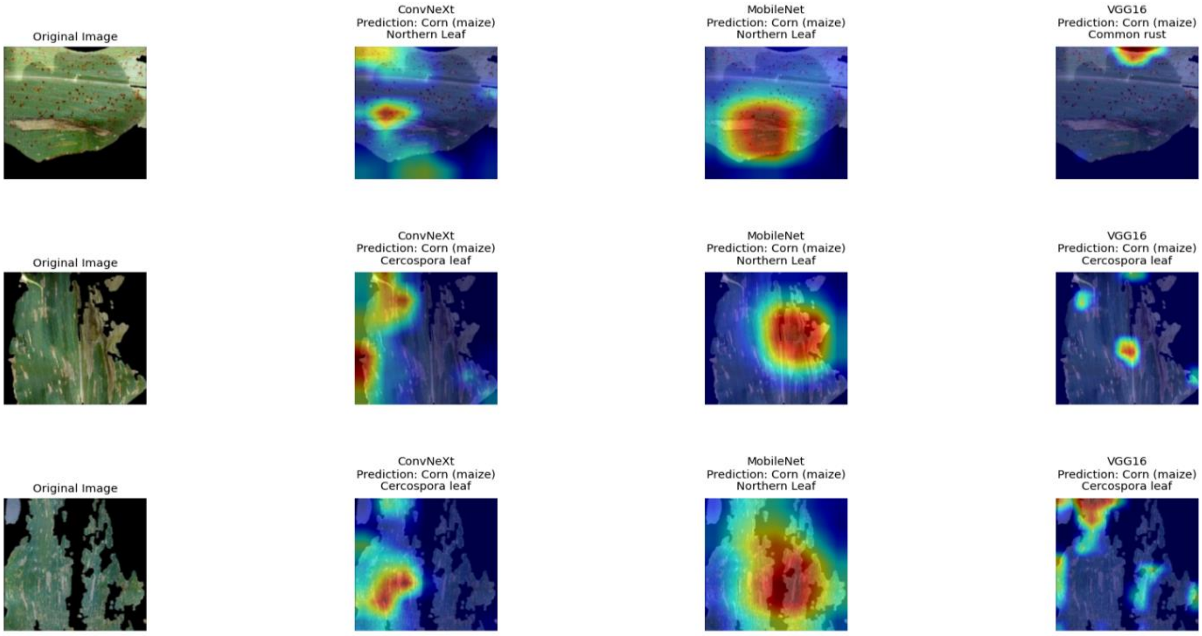


Figure 8- Grad-CAM visualizations for misclassifications of maize leaf diseases. Examples where models confuse disease types exhibiting similar visual symptoms

3.4.2. Potato misclassifications

Misclassifications in the potato dataset are quite limited. Figure 9 shows that the VGG16 model incorrectly predicted 2 Late Blight samples as Early Blight. Grad-CAM analysis revealed that the heat map in these samples focused on leaf edges and chlorotic (yellowing) regions. Since similar chlorosis symptoms can be observed in both disease types, the VGG16 model likely erred due to this visual similarity.

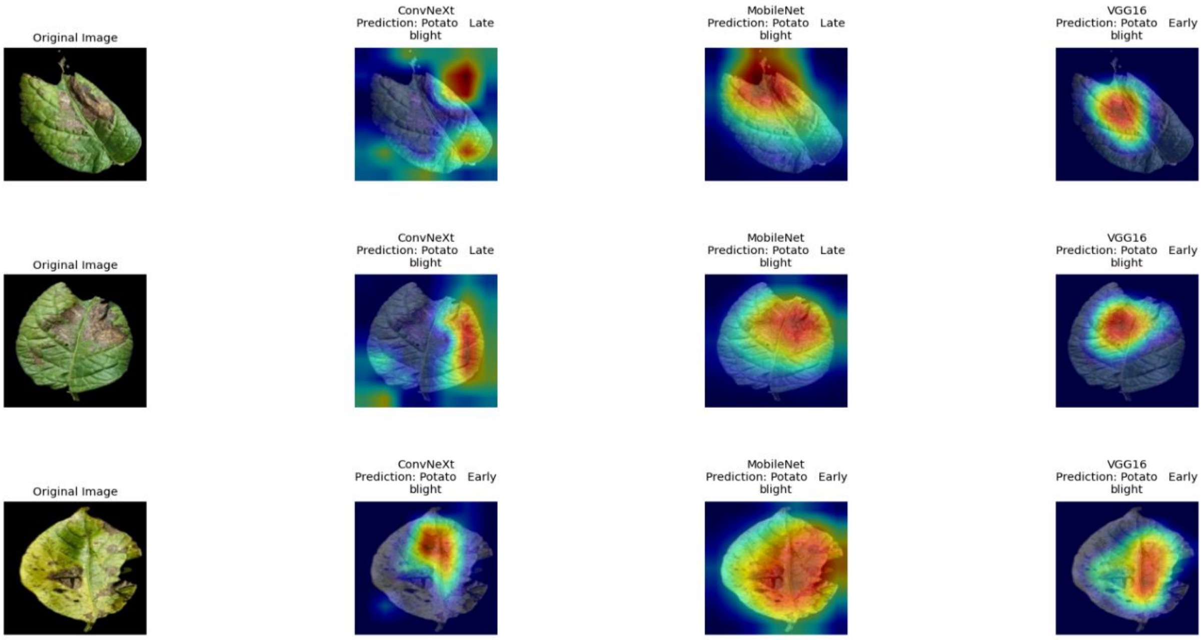


Figure 9- Grad-CAM visualizations for misclassifications of potato leaf diseases. Examples where Late Blight was incorrectly predicted as Early Blight

3.4.3. Pepper misclassifications

In the first row of Figure 10, a healthy pepper leaf is misclassified as bacterial spot by all three models. Upon examination of the Grad-CAM analysis, it is observed that the VGG16 model's heat map concentrated on the leaf center and vein structures. This finding indicates that the model may have interpreted the natural vein pattern and texture of the leaf as disease lesions.

In the second row, another healthy leaf is misclassified (Fig. 10). The heat maps of the MobileNetV2 and VGG16 models are distributed over large areas and focused on the general texture of the leaf surface. This suggests that the factor leading to false positives by the models may be visual similarities arising from specific lighting conditions or pigmentation variations on the leaf.

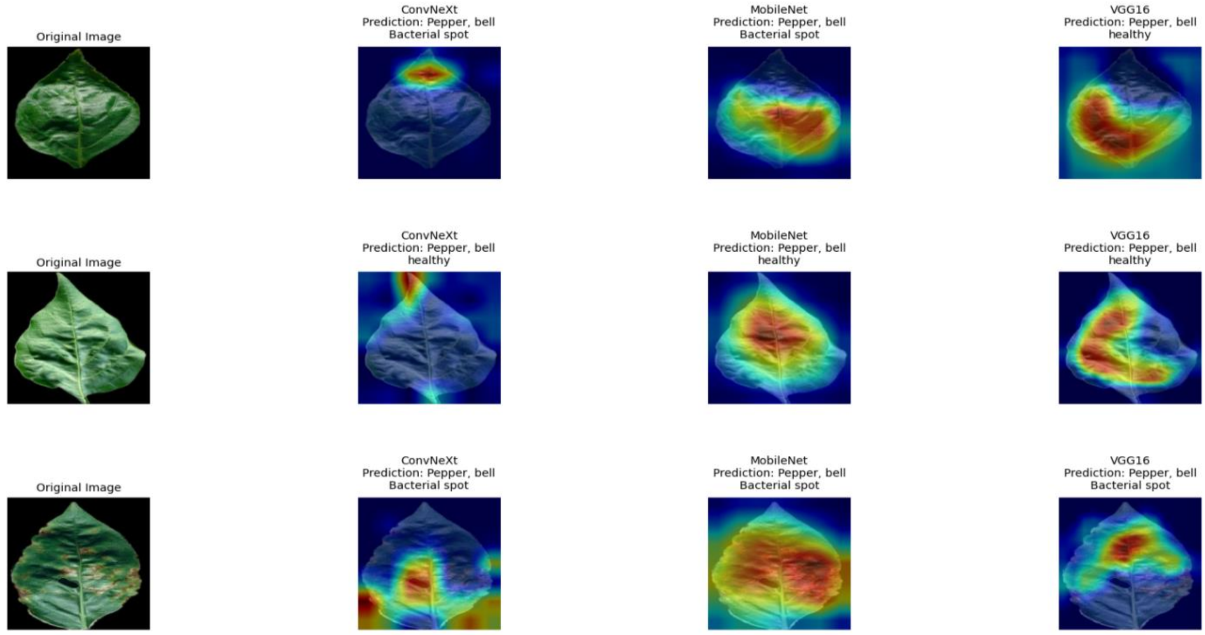


Figure 10- Grad-CAM analysis of misclassifications in pepper leaves. First two rows: misclassification of healthy leaves as bacterial spot

As presented in Figure 11, the Grad-CAM visualizations were further evaluated quantitatively using IoU computed with proxy lesion masks, demonstrating that the proposed models consistently concentrated on disease regions rather than background areas.

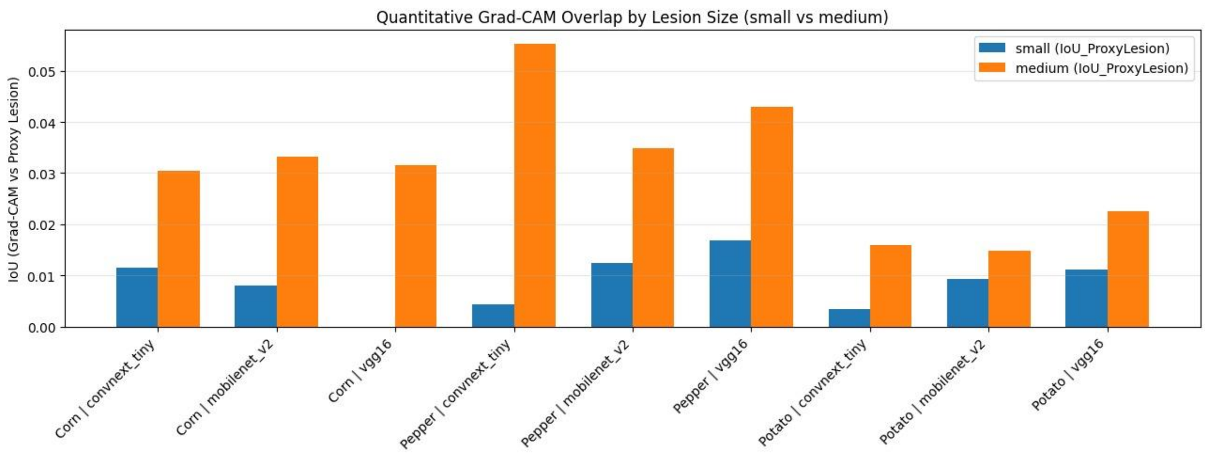


Figure 11- Quantitative evaluation (IoU) of Grad-CAM localization performance using proxy lesion masks across datasets and CNN models

Figure 11 summarizes the quantitative Grad-CAM evaluation results. All models demonstrated measurable overlap between attention maps and lesion regions. ConvNeXt-Tiny achieved the highest Dice and IoU scores, indicating more precise localization of symptomatic regions. MobileNetV2 showed moderate localization performance with lower overlap values, while VGG16 exhibited comparatively weaker spatial alignment. Lesion-size-based analysis further revealed that attention consistency increased for medium and large lesions, whereas very small lesions were more challenging for all models. These findings quantitatively confirm that the networks focus on biologically meaningful regions rather than background artifacts.

4. Discussion

This study compared the performance of three different CNN architectures (ConvNeXt-Tiny, MobileNetV2, VGG16) in classifying plant leaf diseases and evaluated the explainability of model decisions through Grad-CAM visualizations. The findings revealed that not only high classification accuracy but also the interpretability of model decision mechanisms is critically important for agricultural decision support systems.

4.1. Model generalizability and dataset limitations

Although the models demonstrated high accuracy rates, the PlantVillage segmented dataset predominantly consists of images captured under controlled conditions with uniform backgrounds and low environmental complexity. Consequently, the classification task becomes considerably simpler compared to real field images. Therefore, the reported high performance should be interpreted as an upper-bound estimate rather than a reflection of actual field performance. Mohanty et al. (2016) achieved 99.35% accuracy on the PlantVillage dataset collected in a controlled environment; however, when the same model was tested on images from real-world conditions, the accuracy dropped dramatically to 31.40%. Future validation on real field images containing complex backgrounds, varying lighting conditions, and occlusions would enable a more objective assessment of model generalizability (Ahmad et al. 2023).

4.2. A Comparative evaluation of model performance

The accuracy rates obtained in this study are consistent with similar studies in the literature and, in some cases, demonstrate even higher results. ConvNeXt-Tiny achieved 100% accuracy on the potato dataset. This demonstrates the effectiveness of integrating transformer-based design principles into CNN architectures. The ConvNeXt architecture proposed by Liu et al. (2022) combines modern components such as depthwise convolutions, layer normalization, and GELU activation with the classical CNN structure, providing both high accuracy and computational efficiency.

The success rates of MobileNetV2 in the range of 98-100% validate the effectiveness of the inverted residual structure emphasized by Sandler et al. (2018). Particularly, achieving almost 100% accuracy on potato and pepper datasets demonstrates that this lightweight architecture possesses sufficient discriminative feature learning capacity even in complex disease classification tasks. This finding presents significant potential for mobile applications that can operate with low hardware requirements in field conditions.

The performance of the VGG16 model in the range of 97-99% demonstrates that classical CNN architectures still constitute a strong foundation; however, it reveals that modern architectures surpass this model in terms of both accuracy and computational cost. Although the deep and regular structure proposed by Simonyan and Zisserman (2015) was groundbreaking in the field of image classification, it has fallen relatively behind compared to today's more complex architectures.

The statistically significant performance advantage of ConvNeXt-Tiny in maize disease classification ($P<0.001$) can be attributed to its modern architectural components inspired by transformer designs (Liu et al. 2022). However, the lack of significant differences in potato and pepper datasets ($P>0.05$) suggests that these diseases may have more distinctive visual features, making them easier to classify even with lighter architectures.

Importantly, while MobileNetV2 achieved comparable test accuracy to ConvNeXt-Tiny in some cases, Grad-CAM visualizations revealed different attention mechanisms, highlighting that performance metrics alone are insufficient for assessing model reliability.

One of the findings of this study is that the balance between model complexity and explainability must be carefully managed. While ConvNeXt-Tiny provided the highest accuracy, MobileNetV2 demonstrated comparable performance with lower computational cost. Although VGG16 has a simpler structure than both models, it exhibited some weaknesses in Grad-CAM visualizations.

These findings bring Breiman (2001) discussion of the "accuracy-interpretability tradeoff" back to the agenda. However, as observed in this study, modern deep learning architectures can provide satisfactory levels of explainability when used in conjunction with post-hoc explanation methods such as Grad-CAM, while achieving high accuracy. The capacity to model non-linear relationships emphasized by Ryo and Rillig (2017) is the fundamental reason for the superiority of deep learning models in complex agricultural problems.

4.3. Contribution of Grad-CAM visualizations to agricultural applications

Grad-CAM analyses enabled the evaluation of whether the decision mechanisms are biologically meaningful, beyond models achieving high accuracy. One of the most important contributions of this study is demonstrating that three different architectures employ different attention strategies. Alhammad et al. (2025) used Grad-CAM in the classification of potato leaf diseases and emphasized the importance of interpretability. Our study validated this finding across three different models and provided a comparative analysis.

The ConvNeXt-Tiny model demonstrated the most precise and focused attention on disease lesions. This model was able to clearly delineate lesion boundaries and successfully differentiate between diseased and healthy tissues. It was observed that the Grad-CAM method proposed by Selvaraju et al. (2017), when used with modern architectures like ConvNeXt, produces sharper and more discriminative visualizations. This finding aligns with the principles of "transparency and trust" emphasized by Cartolano et al. (2024) in the context of smart agriculture.

The broader and more diffuse attention distribution of the MobileNetV2 model suggests that the model also evaluates contextual information surrounding disease regions. This characteristic may be advantageous in some cases for detecting ambiguous symptoms in early stages of disease. However, an excessively broad attention area implies that the model may sometimes be influenced by regions not directly related to the disease.

The local and focal attention strategy of the VGG16 model demonstrated that the model focuses on smaller and more specific features. However, this approach showed a tendency to fail to capture the overall spread of disease and to attend to regions not directly related to the disease, such as leaf edges and vein structures. This is consistent with the "misattention" problem also reported by Lin et al. (2019) and Wei et al. (2022).

Grad-CAM visualizations provided extremely valuable information for understanding the causes of misclassifications. In cases where healthy leaves were misclassified as diseased, it was observed that models were overly sensitive to the natural vein pattern of the leaf, pigmentation variations, or shadow effects resulting from lighting conditions. This finding points to the problem of "spurious correlation," also emphasized by Kundu et al. (2021).

It is reported that deep learning-based classification models struggle to distinguish diseases with morphologically similar lesion forms (Cercospora leaf spot and Northern Leaf Blight). This situation arises from the fact that these diseases have lesion structures with similar color, shape, and texture characteristics, limiting the discriminative power of the model. Indeed, the frequent confusion between Cercospora leaf spot and Northern Leaf Blight diseases has also been observed in deep transfer learning-based studies (Fraiwan et al. 2022). Similar findings were reported in a study on rice leaf diseases by Rahman et al. (2024), who additionally compared Grad-CAM++, Score-CAM, and SHAP methods and reported that Grad-CAM++ provides more sensitive visualizations.

These results demonstrate that model evaluations based solely on accuracy scores are insufficient and that explainability analyses must be considered in the model improvement process. The "stop explaining black box models" approach emphasized by Rudin (2019) and Rudin et al. (2022) advocates that models to be used especially in critical agricultural decisions must provide both high accuracy and intrinsic interpretability.

4.4. Integration potential into agricultural decision support systems

The findings of this study provide important implications for the design of artificial intelligence systems to be used in plant disease detection. The demonstration by the ConvNeXt-Tiny model of both high accuracy and biologically meaningful attention mechanisms suggests that this model is an architecture that can be preferred in agricultural decision support systems.

However, the low computational cost and high success of MobileNetV2 provide an indispensable advantage, especially for applications that can operate in resource-constrained environments (smartphones, field computers). It has been reported by Xu et al. (2025) that the integration of attention mechanisms with an improved lightweight architecture based on MobileNetV2 achieves high accuracy in agricultural applications and can operate effectively on resource-constrained devices.

The "user trust" factor emphasized by Sabrina et al. (2022) is the most critical point in the integration of explainable artificial intelligence into agricultural systems. For farmers and agricultural experts to see not only a disease label but also the visual findings on which this decision is based is of vital importance for the adoption and reliability of the system. Human-AI interaction studies conducted by Meske & Bunde (2020) have shown that transparency and interpretability are the most important factors increasing user trust.

5. Conclusions

This study achieved high accuracy rates in classifying potato, maize, and pepper leaf diseases using ConvNeXt-Tiny, MobileNetV2, and VGG16 architectures. However, the main contribution of the study is the systematic examination of the explainability of model decisions through Grad-CAM visualizations.

The findings demonstrated that modern deep learning models are not only numerically successful but can also develop biologically meaningful decision mechanisms. While ConvNeXt-Tiny exhibited the strongest performance in terms of both accuracy and explainability, MobileNetV2 emerged as an ideal option for practical applications with low resource requirements. Although VGG16 still constitutes a strong foundation, it showed some limitations compared to modern architectures.

Grad-CAM analysis of misclassifications revealed the visual biases and weaknesses of models, providing valuable insights for future model improvements. These findings emphasize that explainable artificial intelligence in agricultural decision support systems is not an optional feature but a fundamental requirement.

This study makes three important contributions: (1) demonstrating that ConvNeXt-Tiny achieves the highest performance across all datasets with 99.83-100% accuracy and that lesion-focused attention strategy is the most reliable in Grad-CAM analyses, (2) showing that MobileNetV2 with comparable performance (97.59-100%) and low computational cost is an ideal candidate for mobile agricultural applications, and (3) proving through Grad-CAM visualizations that high accuracy alone is insufficient for model reliability. Our findings emphasize that explainability in agricultural decision support systems is not optional but a fundamental requirement, and indicate that future studies should focus on generalization capacity in field conditions and the integration of multiple XAI methods.

Author Declarations

Authors' contributions: Study conception and design: AG, ET; data collection: AG; analysis and interpretation of results: AG, ET, HBD; draft manuscript preparation: AG, ET, HBD. All authors reviewed and approved the final version of the manuscript.

Conflict of interest/Competing interests: The author(s) declare no conflict of interest.

Ethics approval/declarations: No need/ Not applicable.

Funding: This research received no external funding.

Consent to participate: No need/ Not applicable.

Consent for publication: No need/ Not applicable.

Data availability: Data are available on request due to privacy or other restrictions.

References

Adadi A & Berrada M (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access* 6: 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Ahmad A, Saraswat D & El Gamal A (2023). A survey on using deep learning techniques for plant disease diagnosis and recommendations for development of appropriate tools. *Smart Agricultural Technology, 3*, 100083. https://doi.org/10.1016/j.atech.2022.100083

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*, 1449329. https://doi.org/10.3389/frai.2024.1449329

Barbedo J G A (2016). A review on the main challenges in automatic plant disease identification based on visible range images. *Biosystems Engineering* 144: 52–60. https://doi.org/10.1016/j.biosystemseng.2016.01.017

Başağaoğlu H, Chakraborty D, Do Lago C, Gutierrez L, Şahinli M A, Mirchi A, Moriasi D & Şengör S S (2022). A Review on Interpretable and Explainable Artificial Intelligence in Hydroclimatic Applications. *Water*, *14*(8), 1230. https://doi.org/10.3390/w14081230

Benos L, Tagarakis A C, Dolias G, Berruto R, Kateris D & Bochtis D (2021). Machine learning in agriculture: A comprehensive updated review. *Sensors*, *21*(11), 3758. https://doi.org/10.3390/s21113758

Breiman L (2001). Statistical modeling: The two cultures. *Statistical Science*, *16*(3), 199–215.

Cartolano A, Cuzzocrea A & Pilato G (2024). Analyzing and assessing explainable AI models for smart agriculture environments. *Multimedia Tools and Applications*, *83*, 37225–37246. https://doi.org/10.1007/s11042-023-17978-z

Doshi-Velez F & Kim B (2017). Towards a rigorous science of interpretable machine learning. arXiv. https://doi.org/10.48550/arXiv.1702.08608

Ferentinos K P (2018). Deep learning models for plant disease detection and diagnosis. *Computers and Electronics in Agriculture*, *145*, 311–318. https://doi.org/10.1016/j.compag.2018.01.009

Fraiwan M, Dinh A, Wahid K A & Bhowmik P (2022). Classification of corn diseases from leaf images using deep transfer learning. *Plants*, *11*(20), 2668. https://doi.org/10.3390/plants11202668

Gardezi M, Joshi B, Rizzo D M, Ryan, M, Prutzer E, Brugler S & Dadkhah A (2023). Artificial intelligence in farming: Challenges and opportunities for building trust. *Agronomy Journal*, *116*(3), 1217–1228. https://doi.org/10.1002/agj2.21353

Gildenblat J (2021). *PyTorch library for CAM methods*. GitHub. https://github.com/jacobgil/pytorch-grad-cam

Hughes D P & Salathé M (2015). An open access repository of images on plant health to enable the development of mobile disease diagnostics. arXiv. https://doi.org/10.48550/arXiv.1511.08060

Hunter J D (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Karim M J, Goni M O F, Nahiduzzaman M, Ahsan M, Haider J & Kowalski, M (2024). Enhancing agriculture through real-time grape leaf disease classification via an edge device with a lightweight CNN architecture and Grad-CAM. *Scientific Reports, 14,* 16022. https://doi.org/10.1038/s41598-024-66989-9

Krizhevsky A, Sutskever I & Hinton, G E (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems, 25*, 1097-1105. https://doi.org/10.1145/3065386

Kundu N, Rani G, Dhaka V S, Gupta K, Nayak S C, Verma S, Ijaz M F & Woźniak M (2021). IoT and interpretable machine learning based framework for disease prediction in pearl millet. *Sensors*, *21*(16), 5386. https://doi.org/10.3390/s21165386

Kundu N, Rani G, Dhaka V S, Gupta K, Nayaka S C, Vocaturo E & Zumpano E (2022). Disease detection, severity prediction, and crop loss estimation in MaizeCrop using deep learning. *Artificial Intelligence in Agriculture*, *6*, 276–291. https://doi.org/10.1016/j.aiia.2022.11.002

Liakos K G, Busato P, Moshou D, Pearson S & Bochtis D (2018). Machine learning in agriculture: A review. *Sensors*, *18*(8), 2674. https://doi.org/10.3390/s18082674

Lin K, Gong L, Huang Y, Liu C & Pan J (2019). Deep learning-based segmentation and quantification of cucumber powdery mildew using convolutional neural network. *Frontiers in Plant Science*, *10*, 155. https://doi.org/10.3389/fpls.2019.00155

Linardatos P, Papastefanopoulos V & Kotsiantis S (2020). Explainable AI: A review of machine learning interpretability methods. *Entropy*, *23*(1), 18. https://doi.org/10.3390/e23010018

Liu Z, Mao H, Wu C Y, Feichtenhofer C, Darrell T & Xie S (2022). A ConvNet for the 2020s. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11976–11986. https://doi.org/10.1109/CVPR52688.2022.01167

Lundberg S M & Lee S I (2017). A unified approach to interpreting model predictions. arXiv. https://doi.org/10.48550/arXiv.1705.07874

Meske C & Bunde E (2020). Transparency and trust in human-AI interaction: The role of model interpretability, explainability and trust. arXiv. https://doi.org/10.48550/arXiv.2002.01543

Mohanty S P, Hughes D P & Salathe M (2016). Using Deep Learning for Image-Based Plant Disease Detection. *Frontiers in Plant Science, 7*, 1419. https://doi.org/10.3389/fpls.2016.01419

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

Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A, Köpf A, Yang E, DeVito Z, Raison M, Tejani A, Chilamkurthy S, Steiner B, Fang L, … Chintala S (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Advances in Neural Information Processing Systems*, *32*, 8024–8035. https://doi.org/10.48550/arXiv.1912.01703

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M & Duchesnay E (2011). *Scikit-learn: Machine Learning in Python*. https://doi.org/10.48550/arXiv.1201.0490

Quach L D, Quoc K N, Quynh A N, Thai-Nghe N & Nguyen T G (2023). Explainable deep learning models with gradient-weighted class activation mapping for smart agriculture. *IEEE Access, 11*, 83752–83762. https://doi.org/10.1109/ACCESS.2023.3296792

Rahman M M, Yan Z, Aziz M T, Siddick M A B, Truong T, Sharif M M, Datta N, Mahmud T, Pecho R D C & Farid S M (2024). Explainable Deep Transfer Learning Framework for Rice Leaf Disease Diagnosis and Classification. *International Journal of Advanced Computer Science and Applications*, *15*(12), 862–884. https://doi.org/10.14569/IJACSA.2024.0151287

Ribeiro M T, Singh S & Guestrin C (2016). "Why should I trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778

Rudin C (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Rudin C, Chen C, Chen Z, Huang H, Semenova L & Zhong C (2022). Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistics Surveys*, *16*, 1–85. https://doi.org/10.1214/21-SS133

Ryo M (2022). Explainable artificial intelligence and interpretable machine learning for agricultural data analysis. *Artificial Intelligence in Agriculture*, *6*, 257–265. https://doi.org/10.1016/j.aiia.2022.11.003

Ryo M & Rillig M C (2017). Statistically reinforced machine learning for nonlinear patterns and variable interactions. *Ecosphere, 8*(11), e01976. https://doi.org/10.1002/ecs2.1976

Sabrina F, Sohail S, Farid F, Jahan S, Ahamed F & Gordon S (2022). An interpretable artificial intelligence-based smart agriculture system. *Computers, Materials & Continua*, *72*(2), 3777–3797. https://doi.org/10.32604/cmc.2022.026363

Sandler M, Howard A, Zhu M, Zhmoginov A & Chen L C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4510–4520. https://doi.org/10.1109/CVPR.2018.00474

Savary S, Willocquet L, Pethybridge S J, Esker P D, McRoberts N & Nelson A (2019). The global burden of pathogens and pests on major food crops. *Nature Ecology & Evolution*, *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. *Proceedings of the IEEE International Conference on Computer Vision*, 618–626. https://doi.org/10.1109/ICCV.2017.74

Simonyan K & Zisserman A (2015). Very deep convolutional networks for large-scale image recognition. arXiv. https://doi.org/10.48550/arXiv.1409.1556

Sokolova M & Lapalme G (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, *45*(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Waskom M L (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021

Wei K, Chen B, Zhang J, Fan S, Wu K, Liu G & Chen D (2022). Explainable deep learning study for leaf disease classification. *Agronomy*, *12*(5), 1035. https://doi.org/10.3390/agronomy12051035

Xu Y, Li D, Li C, Yuan Z & Dai Z (2025). LiSA-MobileNetV2: an extremely lightweight deep learning model with Swish activation and attention mechanism for accurate rice disease classification. *Frontiers in Plant Science*, *16*, 1619365. https://doi.org/10.3389/fpls.2025.1619365