Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi

Pamukkale University Journal of Engineering Sciences

Localization evaluation of CAM based explainability techniques for plant disease detection

Bitki hastalığı tespiti için CAM tabanlı açıklanabilirlik yöntemlerinin yerelleştirme değerlendirmesi

Duygu SİNANÇ TERZİ[1*]

[1]Department of Computer Engineering, Amasya University, Amasya, Türkiye.
duygusinanc@gazi.edu.t

Received/Geliş Tarihi: 01.10.2024
Accepted/Kabul Tarihi: 20.05.2025
Revision/Düzeltme Tarihi: 09.04.2025
doi: 10.5505/pajes.2025.50955
Research Article/Araştırma Makalesi

Abstract

In recent years, computer vision technologies have played a critical role in precision agriculture, leveraging robotics and artificial intelligence to automate tasks in crop production. While image-based applications hold promise, model interpretability remains a significant challenge. Explainable artificial intelligence aims to address this by providing plant scientists with interpretable, reliable information, improving the understanding of plant diseases. This study focuses on integrating explainability metrics into model evaluation, with a detailed analysis of explainability methods applied to plant disease classification models. Using Class Activation Mapping based visualization methods with architectures such as EfficientNet, MobileNet, ResNet, and ShuffleNet, trained on a public plant disease dataset, the study assessed both classification success and model explainability. Localization results were derived from an energy-based perspective, assessing how well saliency maps aligned with bounding boxes of diseased areas. The findings reveal that feature dimensions and positions in the images significantly influence classification outcomes, highlighting the importance of precise annotations during data labeling. This study uncovers potential biases in disease detection and emphasizes the need for explainability metrics in evaluating deep learning models, paving the way for more accurate and efficient plant disease detection techniques.

Keywords: precision agriculture, computer vision, deep learning, plant disease classification, explainability

Öz

Son yıllarda bilgisayarla görme teknolojileri, hassas tarımda kritik bir rol oynamış, robotik ve yapay zekayı kullanarak mahsul üretiminde görevleri otomatikleştirmiştir. Görüntü tabanlı uygulamalar umut vadetse de, modelin yorumlanabilirliği önemli bir zorluk olmaya devam etmektedir. Açıklanabilir yapay zeka, bitki bilimcilerine yorumlanabilir ve güvenilir bilgiler sunarak bitki hastalıklarının anlaşılmasını geliştirmeyi hedeflemektedir. Bu çalışma, açıklanabilirlik metriklerinin model değerlendirmesine entegrasyonuna odaklanmakta ve bitki hastalığı sınıflandırma modellerine uygulanan açıklanabilirlik yöntemlerinin detaylı bir analizini sunmaktadır. EfficientNet, MobileNet, ResNet ve ShuffleNet gibi mimarilerle, açık bir bitki hastalığı veri seti üzerinde eğitilmiş Sınıf Aktivasyon Haritalama tabanlı görselleştirme yöntemleri kullanılarak hem sınıflandırma başarısı hem de modelin açıklanabilirliği değerlendirilmiştir. Lokalizasyon sonuçları, dikkat haritalarının hastalıklı bölgeleri etiketleyen sınırlayıcı kutularla ne kadar uyumlu olduğunu enerji tabanlı bir perspektiften değerlendirerek elde edilmiştir. Bulgular, görüntülerden çıkarılan özelliklerin boyutları ve konumlarının sınıflandırma sonuçlarını önemli ölçüde etkilediğini göstermektedir ve veri etiketleme aşamasında doğru anotasyonların önemini vurgulamaktadır. Bu çalışma, hastalık tespitindeki olası yanlılıkları ortaya çıkarmakta ve derin öğrenme modellerinin değerlendirilmesinde açıklanabilirlik metriklerinin gerekliliğini vurgulayarak, bitki hastalıklarının daha doğru ve verimli bir şekilde tespit edilmesi için derin öğrenme tekniklerinin optimize edilmesine zemin hazırlamaktadır.

Anahtar kelimeler: hassas tarım, bilgisayarlı görme, derin öğrenme, bitki hastalığı sınıflandırması, açıklanabilirlik

1 Introduction

In the era of Agriculture 4.0, precision agriculture stands out as a pivotal paradigm, holding the potential to bring about a profound transformation in agricultural methodologies. This revolutionary approach is centered around the adoption of advanced monitoring and intervention technologies. The primary objective is not only to support production efficiency but also to address environmental concerns by minimizing negative impacts associated with traditional farming practices [1]. By leveraging cutting-edge tools, precision agriculture aims to usher in a sustainable and technologically advanced era for the agricultural sector.

Utilizing digital images to analyze and understand the environment, computer vision technologies offer precise, location-specific insights into crops and their surrounding ecosystems. These technologies utilize a variety of sensing modalities to capture detailed information about plant health, growth patterns, and environmental conditions. RGB imaging enables visual inspection of crops, identifying visible symptoms of stress, disease, or nutrient deficiencies. Near-infrared, multispectral, and hyperspectral imaging go beyond the visible spectrum to detect subtle physiological and biochemical changes, offering early warnings of plant health issues.

Deep learning (DL) empowers computer vision with advanced pattern recognition capabilities. The notable advantage of DL lies in its ability to automatically extract features from raw data and form higher-level features through the composition of lower-level ones. The highly hierarchical structure and expansive learning capacity of DL models excel in classification and prediction tasks, demonstrating flexibility in addressing complex data analysis challenges. Applied to precision

[*]Corresponding author/Yazışılan Yazar

agriculture, DL enables data-driven, automated, and informed decision-making processes. DL finds diverse applications in agriculture, significantly extending its impact across various domains. In plant disease detection, DL facilitates automated and precise identification of diseases, enabling timely intervention and effective disease management [2]. For pest detection, DL is employed to accurately identify pests affecting crops, thus improving pest control strategies. DL aids in weed identification, allowing farmers to implement targeted weed control measures and enhance crop yields. Furthermore, DL models are utilized for leaf classification based on specific characteristics, assisting in the identification and classification of plant species [3]. Plant phenology recognition through DL enables precise timing of agricultural activities aligned with plant growth stages, optimizing resource management. DL also supports the segmentation of roots and soil, offering comprehensive analysis of below-ground structures and their impact on plant health. In crop yield estimation, DL analyses multiple factors to provide accurate yield forecasts, assisting farmers in planning and resource allocation. Additionally, DL facilitates the automated counting of fruits, providing valuable data for harvest planning and yield estimation [32]. In the realm of autonomous farming machinery, DL enhances obstacle detection, ensuring efficient and safe operations. Land cover classification through DL contributes to the monitoring and assessment of agricultural ecosystems and land use patterns [33]. DL is also applied to monitor animal welfare in agricultural settings, ensuring the well-being of livestock. For water management, DL enables precise allocation based on crop requirements, thus conserving water and improving irrigation practices. Finally, DL assists in soil management by analyzing soil composition and optimizing fertilization strategies, thereby enhancing soil health and crop productivity.

Among the above, plant diseases, which lead to reduced harvest yields, constitute a serious threat that not only affects the livelihoods of individual farmers but also has broader implications for the economic stability of nations. Plant diseases are physiological disorders that negatively impact growth, development, and overall health, resulting from the interaction between the host, causal agent, and environment [4]. The broad categorization normally seen among plant diseases falls into two classes: namely biotic and abiotic diseases. Biotic diseases come from living organisms, such as fungi, bacteria, and viruses, while abiotic diseases are caused by non-living agents: environmental conditions, chemicals, and mechanical injury. Biotic diseases are far more aggressive and contagious than abiotic diseases, which are not usually that threatening but preventable.

In the accurate and timely detection of diseases and estimation of severity, DL architectures offer promising solutions that minimize economic losses, enhance food safety, and promote environmentally friendly farming [5]. Despite the enhanced speed and accuracy offered by DL-based methods, they often operate as black-boxes. This means their internal decision-making processes are not easily understood by humans, creating uncertainties about how decisions are made and the principles guiding these models. Explainable Artificial Intelligence (XAI) addresses this challenge by focusing on developing methods and techniques to make the decision-making process of DL models more transparent and interpretable [6]. Integrating XAI with DL models for plant disease detection offers following four benefits.

- ✓ Explain to Discover: Questioning for explanations serves as a valuable method to acquire inherent information and understand the underlying task in plant disease detection. XAI is a potent tool for verifying and gaining new insights into the complexities of plant diseases, leading to a more reliable solution.
- ✓ Explain to Justify: The increasing debates over biased or unfair outcomes in plant disease detection highlight the need for explanations to ensure trustworthy decisions. XAI is essential for providing motives or rationalizations for specific decisions, especially in cases of unexpected outcomes.
- ✓ Explain to Control: Explanation is not just for justifying outcomes but also for preventing errors in the plant disease detection process. Understanding more about the behavior of the system provides superior visibility over vulnerabilities and faults.
- ✓ Explain to Improve: Continuous improvement of DL models used in plant disease detection is a key motivation for explaining algorithms. Models that can be explained and understood are more easily improved.

The integration of XAI techniques in plant disease detection has gained significant attention, with researchers exploring various visualization methods to enhance model interpretability. Among these, Class Activation Mapping (CAM)-based approaches have emerged as powerful tools for uncovering the decision-making processes of CNNs. For instance, Toda and Okura conducted an initial comprehensive analysis to understand the learning process of CNNs during the diagnosis of plant diseases, employing various neuron-wise and layer-wise visualization methods [19]. Their findings showed that Grad-CAM is one of the most effective and cost-efficient methods for generating attention maps.

Building on this finding, in this study, a range of CAM-based visualization methods were implemented using EfficientNet, MobileNet, ResNet, and ShuffleNet architectures, all trained with a publicly available plant disease image dataset. The evaluation of the results focused on two key aspects: classification success and explainability. Classification success was measured using standard metrics such as accuracy, precision, recall, and f1-score to ensure the robustness of the models in correctly identifying plant diseases. Explainability was assessed by analyzing the clarity and interpretability of the generated heatmaps, determining how well the highlighted regions corresponded to the actual diseased areas of the plants as labeled in the dataset. To derive localization evaluation results, explainability maps were analyzed and the bounding boxes in the validation set were considered from an energy-based perspective. Instead of focusing only on the peak value, the amount of energy from the saliency map that fell within the bounding box of the target object was assessed. This approach provided deeper insights into how the models make decisions and helped uncover any potential biases present in the detection of plant diseases. By thoroughly examining and discussing these biases, this study contributes to a more comprehensive understanding of the challenges and limitations in plant disease detection using deep learning models. This, in turn, can facilitate the optimization of deep learning techniques for enhanced accuracy and efficiency in plant disease detection. Moreover, it underscores the necessity of incorporating explainability metrics when evaluating the success of deep learning models in future studies.

Table 1. Summary of the literature on explainability for the detection of plant diseases.

Reference	**Year**	**Dataset**	**Deep Learning Architecture**	**Best Result (%)**		**Explainability Approach**	**Explainability Purpose**
				accuracy	**f1-Score**		
[7]	2018	Self-collected soybean images	DCNN	94.13	*	Top-K high-resolution feature maps	identify and quantify foliar stresses
[8]	2019	Self-collected charcoal rot stem images	3D DCNN	95.73	87	Saliency map	provide physiological insight into model predictions
[9]	2019	PlantVillage	VGG16	*	*	Teacher/Student	provide an interpretable architecture
[10]	2021	PlantVillage	VGG16, VGG19, ResNet50, Inception, MobileNet, MobileNetV2, EfficientNet	92.49	98.42	Grad-CAM, LIME	eliminate false positives
[11]	2022	Self-collected thermal plant images, Paddy crop	CNN	98.55	80.25	CAM	improve classification performance
[12]	2022	Plant Village, AI Challenger 2018	VGG, GoogLeNet, ResNet	99.89	*	SmoothGrad, LIME, GradCAM	clarify the focus of the model in feature extraction
[13]	2022	PlantVillage	Xception+Unet	*	99.1	ResTS	create top-quality visualizations to identify specific spots
[14]	2023	Self-collected omics data and hyperspectral images	CNN	95.5	95.4	saliency maps, activation maximization	visualize the internal representations of the model
[15]	2023	Apple, Embrapa, Maize, PlantVillage, Rice	VGG+ViT	98.86	98.85	Grad-CAM, LIME	interpret prediction results
[16]	2023	Cassava	VOLO, EfficientNetV2S, RESNEXT50	90.5	*	SHAP	generate user-level explainability
[17]	2023	Sunflower	CNN, VGG19, InceptionV3, Xception, ResNet v2, MobileNet, DenseNet201, MobileNetV2, VGG16	93	93	LIME	understand misclassifications
[18]	2023	Self-collected maize images	DenseNet, EfficientNet, MobileNetV3, ShuffleNetV2	*	96.049	LayerCAM, ScoreCAM, AblationCAM, XGradCAM	improve model interpretability

*: Not applied

2 Literature review

Recent research in the literature emphasizes the integration of XAI with DL models for plant disease detection, providing various benefits.

Ghosal et al. constructed an explainable deep learning framework for identifying, classifying, and quantifying plant stress [7]. Nagasubramanian et al. deployed a novel 3D deep convolutional neural network to leverage the spatial and spectral dimensions of 3D hyperspectral images for classification, and then utilized saliency maps to visualize the most sensitive pixel locations [8]. Brahimi et al. introduced a new trainable visualization technique for classifying plant diseases, utilizing a Convolutional Neural Network (CNN) structure with two deep classifiers, namely the Teacher and the Student, where the combined representation of both serves as a proxy for visualizing the most important image regions [9]. This technique demonstrated superior performance compared to existing visualization methods, as evaluated by the area over perturbation curve. Arvind et al. developed a pipeline to interpret and validate predicted outputs using explainability techniques after classifying plant diseases with transfer learning on both original and augmented data [10]. Batchuluun et al. utilized a convolutional neural network coupled with a residual network and incorporated a class activation map to

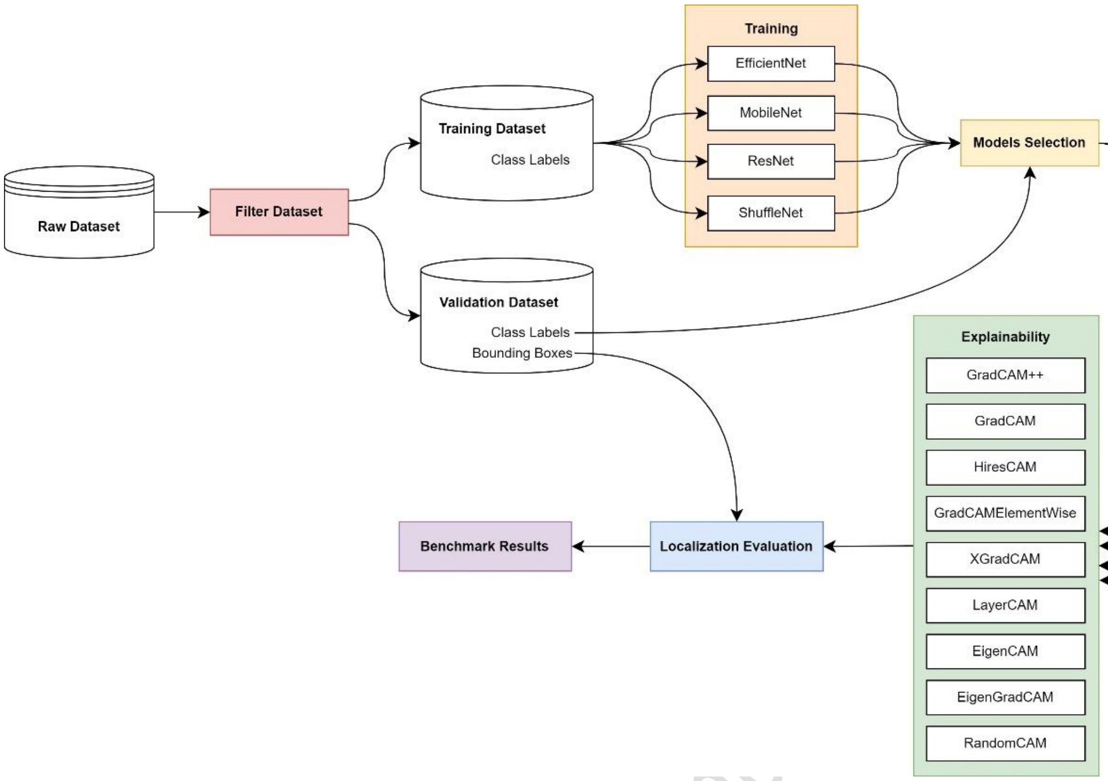

Figure 1. The flowchart of this study.

enhance the performance of plant and crop disease classification using thermal images [11]. Wei et al. examined the interpretability of different classification models using the fruit disease leaves dataset through three experiments: the first focused on classifying fruit and pest species, the second addressed fruit disease classification, and the third examined fruit type classification [12]. Shah et al. enhanced [9] and created an architecture called ResTS (Residual Teacher/Student), featuring two classifiers and a decoder, which serves as both a visualization and classification tool [13]. Shoaib et al. introduced a novel explainable gradient-based approach, EG-CNN, which integrates gene expression data with image data to attain a comprehensive comprehension of plant diseases [14]. Thakur et al. developed a hybrid model that combines the strengths of a vision transformer with the feature extraction capability of convolutional neural networks for lightweight disease classification and then evaluated the explainability of its predictions [15]. Chhetri et al. proposed a generic approach that combines semantic technology and deep learning to enhance prediction accuracy and generate user-friendly explanations [16]. Ghosh et al. developed a hybrid model combining transfer learning and a simple CNN for detecting sunflower diseases, using explainability techniques to analyze misclassifications by generating perturbations in model behavior [17]. Yang et al. introduced a framework for classifying and localizing maize leaf spot diseases using weakly supervised learning, combining lightweight convolutional neural networks with interpretable algorithms to achieve high classification accuracy and fast detection speeds, while also localizing disease spots [18].

These studies collectively enhance the understanding of how models classify diseases, uncover potential biases, and validate predictions, ultimately leading to more accurate and reliable plant disease detection systems. Among these approaches, CAM-based methods have been widely adopted due to their demonstrated superiority in many studies. Building on this, the present study implements a range of CAM-based visualization techniques to evaluate localization performance—specifically, assessing how well the highlighted regions align with the actual diseased areas of the plants as annotated in the dataset.

3 Materials and methods

This study aims to evaluate the decision-making processes of the models by comprehensively applying various explainability methods to deep learning models used in the classification of plant diseases. Fig. 1 encapsulates this entire process, illustrating the flow from the raw dataset through training, model selection, application of explainability methods, and finally, the evaluation of the results. This comprehensive approach aims to enhance the understanding of model decision-making in the context of plant disease classification, ultimately contributing to the development of more transparent and interpretable deep learning models.

3.1 Dataset

High-quality datasets with accurately labelled images are critical for training models to recognize and classify plant diseases effectively. The diversity of the dataset, encompassing various disease types, environmental conditions, and different plant species, enhances the model's ability to generalize across different scenarios, thereby improving its robustness and applicability in real-world settings. Public datasets play a crucial role in meeting the needs of deep learning models for detecting plant diseases. They provide a valuable resource for

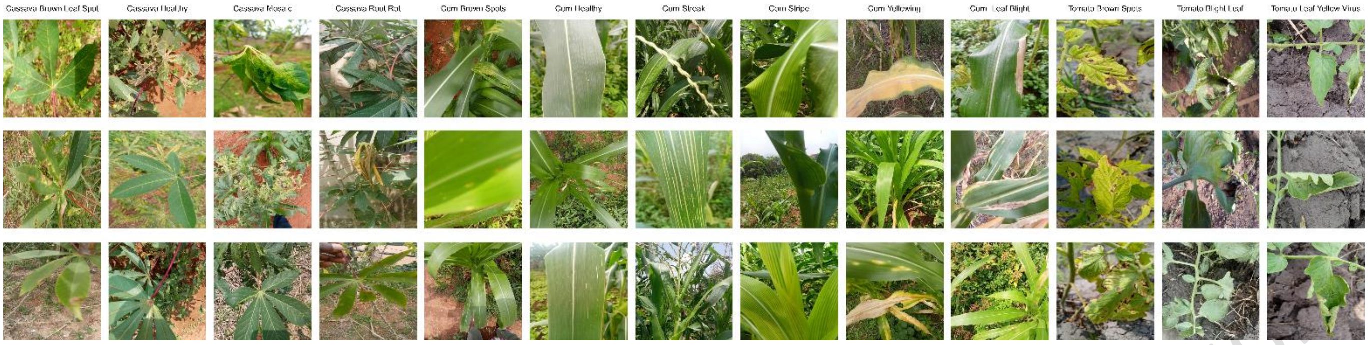

Figure 2. Representative images from the FieldPlant

researchers and practitioners, enabling the development and benchmarking of models without the necessity for costly and time-consuming data collection processes.

To generate an explainability score, a ground truth is needed, so a dataset labelled with bounding boxes was used. The FieldPlant [20] dataset was used in this study due to its inclusion of plant disease images sourced directly from fields, accompanied by manual annotations added to individual leaves within each image, overseen by plant pathologists. Field Plant originally includes 8629 individually annotated leaves spanning 27 disease classes. The impact of class imbalance on model training was reduced by filtering out classes with very few representations in the existing dataset. Accordingly, classes with fewer than 50 examples were excluded from the study. Thirteen different classes with 50 or more samples were utilized: cassava brown leaf spot, cassava healthy, cassava mosaic, cassava root rot, corn brown spots, corn healthy, corn streak, corn stripe, corn yellowing, corn leaf blight, tomato brown spots, tomato blight leaf, and tomato leaf yellow virus. Fig. 2 presents randomly selected example images for these classes. For each class, three images are shown that depict various health issues and disease symptoms.

The data for the classification process was divided into training and validation sets using a random stratified split to ensure class balance. Specifically, 80% of the data was allocated for training, while the remaining 20% was reserved for validation.

The training dataset includes class labels, which are essential for teaching the model to recognize different classes. In contrast, the validation dataset contains both class labels and bounding boxes. This distinction is crucial as it allows for a comprehensive evaluation of the model's performance. While the class labels enable assessment of classification accuracy, the inclusion of bounding boxes facilitates evaluation of the model's localization capabilities. This dual evaluation approach ensures that the model is not only accurate in identifying classes but also proficient in accurately locating the objects within those classes.

3.2 Deep learning models

CNN is a deep learning architecture designed specifically for processing structured grid data such as images. Their hierarchical structure of convolutional and pooling layers allows them to learn spatial hierarchies in data, making them powerful for a variety of tasks in computer vision. The development of various CNN architectures has enabled more efficient, deeper, and accurate models suitable for diverse applications, from mobile devices to large-scale image classification tasks [21].

There are several types of CNN architecture, each designed to address specific challenges or improve performance. To classify plant diseases, four deep learning models utilizing CNN architecture were employed: EfficientNet, MobileNet, ResNet, and ShuffleNet. EfficientNet [25] is designed to achieve superior performance by utilizing a compound scaling method, which carefully balances network depth, width, and resolution, thus improving accuracy while maintaining efficiency in terms of model size and computational cost. MobileNet [26] uses depthwise separable convolutions, a technique that splits the standard convolution operation into two lighter operations, drastically reducing computational complexity while still allowing the model to capture intricate features. ResNet [27] introduced the concept of residual learning, where shortcut connections are added between layers, allowing gradients to flow more easily during backpropagation. This helps address the vanishing gradient problem and facilitates the training of very deep neural networks. ShuffleNet [28] employs pointwise group convolution and channel shuffle operations to reduce computation cost while maintaining accuracy.

All models were trained on the training set using the MMPreTrain [22] framework. The hyperparameter configurations used for the models are as provided in Table 2. This paper does not aim to demonstrate superior classification success but rather to provide a localization evaluation of CAM-based explainability techniques. Consequently, extensive hyperparameter optimization was not conducted for the deep learning architectures. They were executed with default model training settings, except those specified in the table.

3.3 Explainability

Explainability refers to the ability to interpret and understand the decision-making processes of deep learning models. Various types of explainability methods exist, including those that assign importance scores to input features indicating their contributions to the model's output, approximate the model locally with a simpler surrogate model, or generate rules that describe the model's behavior. CAM is a technique used in deep learning, particularly in CNNs, to identify and visualize the regions of an input image that are most relevant for predicting a specific class. CAM helps to understand which parts of the image contribute most to the model's decision, thereby providing interpretability to otherwise black-box models.

Table 2. The hyperparameter configurations.

Model	Backbone	Optimizer	Learning Rate	Augmentation	Batch Size	Epoch Number
EfficientNet	EfficientNet B0	AdamW	0.001	ShiftScaleRotate RandomBrightnessContrast RandomFlip	16	50
MobileNet	MobileNetV2	AdamW				
ResNet	ResNet 101	SGD				
ShuffleNet	ShuffleNetV1	AdamW				

In this study, nine distinct CAM-based methods were employed: GradCAM, GradCAM++, HiResCAM, GradCAMElementWise, XGradCAM, LayerCAM, EigenCAM, EigenGradCAM, and RandomCAM. Grad-CAM operates by computing the gradients of the predicted class with respect to the feature maps of the final convolutional layer [23]. The gradients are averaged to obtain weights for each feature map channel, which are used to create a weighted sum of the feature maps. After applying a ReLU to retain positive contributions, the resulting heatmap highlights the image regions most influential in the model's decision.

To formalize this process mathematically, let us consider how Grad-CAM computes its visual explanations. Given an input image, the CNN generates feature maps A^k from a convolutional layer. These feature maps are used to compute the score y^c for the target class c. Firstly, the gradients of the score y^c with respect to the feature maps A^k are computed. These gradients indicate the importance of each feature map for the target class. Eq. (1) represents the gradient of the score for class c with respect to the (i, j)-th element of the k -th feature map. The gradients are then globally averaged over the spatial dimensions to obtain the importance weights α_k^c for each feature map k. Z is the number of pixels in the feature map. The importance weights α_k^c are then used to perform a weighted sum of the forward activation maps A^k. The ReLU function is applied to the result to ensure that only the positive influences on the class score are considered. This produces the final GradCAM heatmap $L_{GradCAM}^c$, which highlights the regions in the input image that are most relevant for predicting class c.

$$\frac{\partial\, y^c}{\partial\, A_{ij}^k} \qquad (1)$$

$$\alpha_k^c = \frac{1}{Z}\sum_i \sum_j \frac{\partial\, y^c}{\partial\, A_{ij}^k} \qquad (2)$$

$$L_{GradCAM}^c = ReLU\ (\sum_k \alpha_k^c\, A^k) \qquad (3)$$

GradCAM++ is an enhanced version of GradCAM that seeks to overcome key limitations of the original method, particularly in accurately localizing features and dealing with multiple instances of the same class within an image [29]. HiResCAM takes a different approach by performing an element-wise multiplication of activations and gradients. GradCAMElementWise also conducts element-wise multiplication of activations with gradients, followed by a ReLU operation before summation. XGradCAM scales the gradients using normalized activations, effectively weighting the gradients by the relative importance of activation patterns. ScoreCAM departs from gradient-based methods by adopting a perturbation-based strategy. It generates importance maps by masking the input image with scaled activation maps and measuring the drop in the output score for the target class [30]. EigenCAM extracts the first principal component of the 2D activations, revealing overall feature importance regardless of class discrimination [31]. EigenGradCAM builds upon the principles of EigenCAM but reintroduces class-specific gradients to infuse the resulting activation maps with discriminative information. RandomCAM generates CAMs with random uniform values within a specified range for spatial activations, providing a baseline for comparison.

When generating explainability maps, the latest convolution block of each model was utilized. This choice is crucial, as the final convolutional layers often capture the most abstract and representative features relevant to classification tasks. According to MMPreTrain model summary, for EfficientNet Layer 6, for MobileNet the final convolution block (conv2), for ResNet Layer 4, and for ShuffleNet the average of convolution and batch normalization layers belonging to Layer 2.3 were employed.

$$Proportion = \frac{\sum L_{(i.j)\in bbox}^c}{\sum L_{(i.j)\in bbox}^c + \sum L_{(i.j)\notin bbox}^c} \qquad (4)$$

Bounding boxes serve as crucial components in training object detection algorithms, particularly those designed to ascertain the precise locations of objects within images. In this study, bounding boxes were used as a key element in the evaluation of explainability. Explainability maps were used to derive localization evaluation results based on bounding boxes in the validation set from an energy-based perspective. Instead of solely focusing on the maximum point, the amount of energy from the saliency map that overlaps with the bounding box of the target object [24] was considered shown in Eq. (4).

4 Results

The performance of the deep learning models and the effectiveness of the explainability methods are assessed using both benchmark classification results and localization evaluations. Firstly, the classification performance of each model is examined to ensure a fair basis for comparison. To compare the classification results, precision, recall and f1-score of each model were used. Given that the plant disease detection problem under consideration is a multi-class classification, the macro f1-score, computed by averaging the f1-scores of all classes, was employed for evaluation.

Table 3. Classification results.

Model	precision	recall	macro f1-score
EfficientNet	**90.90**	88.36	89.17
Mobilenet	82.48	77.21	78.55
ResNet	89.29	**91.32**	**90.15**
ShuffleNet	88.68	88.67	88.20

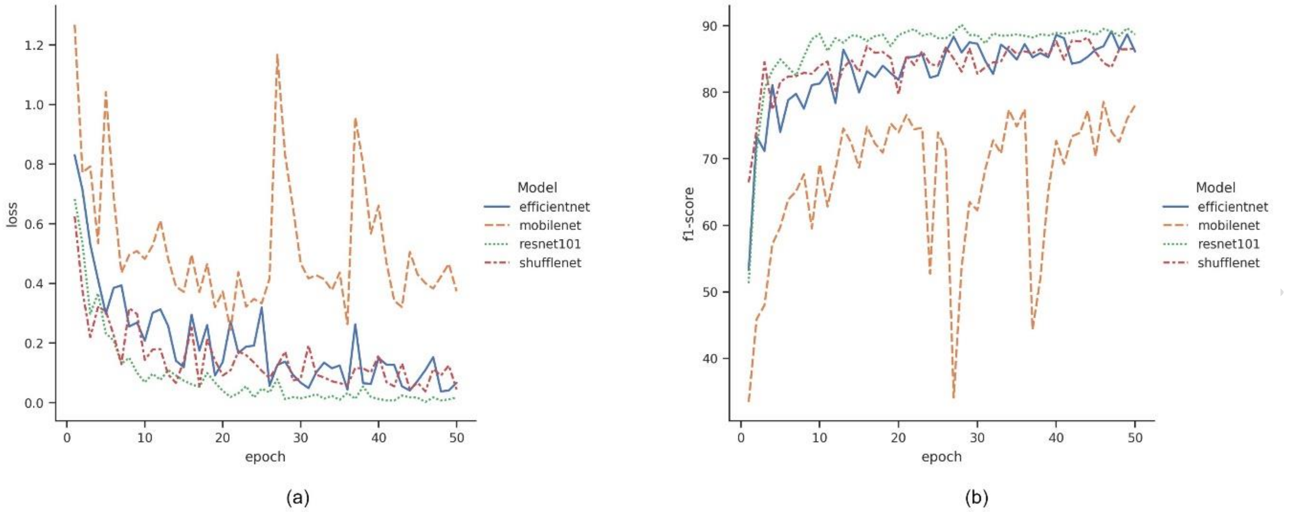

Figure 3. (a) training loss and (b) validation f1-scores of models.

As shown in Table 3, ResNet exhibited the highest recall and f1-score among the models, indicating its superior ability to correctly classify positive samples and maintain a balance between precision and recall. EfficientNet also performed well, particularly in terms of precision, suggesting its effectiveness in minimizing false positives. While ShuffleNet produced results close to those of previous models, Mobilenet lagged slightly behind in all metrics.

The training loss and validation f1-scores for the deep learning models over fifty epochs are provided to observe the learning progression of the models and their ability to generalize to new data. In Fig. 3(a), the training loss for all models demonstrates a general downward trend, indicating effective learning. However, variations in stability and convergence rates are observed. ResNet exhibits the smoothest decline in loss, with consistent learning over time and a low final loss value. EfficientNet shows a slightly more fluctuating loss curve compared to ResNet, indicating some instability in the training process. ShuffleNet struggles with maintaining consistent learning and optimization. In Fig. 3(b) presents the validation f1-scores, highlighting the generalization capabilities of the models. ResNet consistently outperforms the other models. EfficientNet also performs well, with a steady increase in f1-score, closely trailing ResNet. ShuffleNet maintains competitive performance, slightly below EfficientNet. MobileNet has not produced satisfactory results, as evidenced by its high training loss and low validation F1-scores, both of which indicate suboptimal model performance and poor convergence. Despite its weaker performance, MobileNet was deliberately included in the analysis to provide a comparative baseline and to examine how explainability methods behave when applied to a relatively underperforming model.

The localization-based explainability proportions of the models for each class are given in Tables 4-7. The mean of the nine different explainability localization evaluations produced by each deep learning architecture for each disease class is given in italics in the tables for the highest value. The highest mean localization proportion produced by any deep learning architecture for each disease class is given in bold in the tables.

EfficientNet demonstrates the highest mean localization proportion for cassava brown leaf spot at 0.78, while tomato brown spots exhibit the lowest proportion at 0.48. For MobileNet, the class with the highest average localization proportion is corn stripe with 0.86, whereas the lowest proportion is observed for tomato brown spots with 0.50. ResNet shows the highest average localization proportion of 0.85 for corn stripe, while tomato brown spots have the lowest proportion at 0.52. ShuffleNet achieves the highest average localization proportion of 0.75 for cassava brown leaf spot, while corn yellowing exhibits the lowest proportion of 0.42. ResNet has the highest average localization proportions across all explainability techniques for all classes except corn stripe, with MobileNet leading in this specific class. This indicates that ResNet consistently outperforms other models in terms of the quality of explanations generated by various CAM-based techniques. EfficientNet also tends to produce consistent and high localization accuracy across most disease classes. Finally, ShuffleNet exhibits the lowest localization performance overall, facing particular challenges in certain disease classes. This suggests that it may require further optimization to improve.

Additionally, when explainability and classification performance are evaluated together, it becomes evident that high classification scores do not always correspond to high explainability. For instance, although EfficientNet demonstrates strong classification capabilities, its localization proportions are relatively low. Conversely, MobileNet, which exhibits the lowest classification performance among the models, does not produce the lowest localization proportions.

The localization proportions produced by each explainability technique vary depending on the model and class. EigenCAM and EigenGradCAM consistently yield high values, with the highest localization proportion achieved by MobileNet on the corn stripe class, reaching 94% using EigenCAM. On the other hand, XGradCAM and RandomCAM produce lower values. The lowest localization proportion was observed with EfficientNet on the tomato brown spots class, achieving only 26% with XGradCAM. This variation suggests that some techniques are more effective at accurately highlighting relevant regions

Table 4. EfficientNet explainaiblity localization evaluation.

	GradCAM++	GradCAM	HiResCAM	GradCAMElementWise	XGradCAM	LayerCAM	EigenCAM	EigenGradCAM	RandomCAM	Mean
cassava brown leaf spot	0.81	0.79	0.82	0.79	0.64	0.81	0.79	0.86	0.71	*0.78*
cassava healthy	0.69	0.59	0.69	0.65	0.61	0.66	0.64	0.68	0.58	0.64
cassava mosaic	0.73	0.66	0.75	0.69	0.54	0.71	0.68	0.75	0.59	0.68
cassava root rot	0.65	0.49	0.64	0.63	0.45	0.66	0.62	0.74	0.51	0.60
corn brown spots	0.56	0.57	0.60	0.56	0.45	0.58	0.54	0.63	0.48	0.55
corn healthy	0.69	0.69	0.61	0.68	0.70	0.68	0.72	0.65	0.71	0.68
corn streak	0.74	0.72	0.76	0.68	0.58	0.69	0.65	0.74	0.57	0.68
corn stripe	0.74	0.76	0.76	0.72	0.69	0.73	0.71	0.79	0.70	0.73
corn yellowing	0.57	0.60	0.66	0.53	0.33	0.55	0.46	0.63	0.42	0.53
corn leaf blight	0.68	0.63	0.71	0.63	0.53	0.65	0.60	0.72	0.56	0.64
tomato brown spots	0.51	0.47	0.56	0.50	0.26	0.53	0.44	0.65	0.36	0.48
tomato blight leaf	0.58	0.48	0.59	0.55	0.42	0.57	0.52	0.59	0.45	0.53
tomato leaf yellow virus	0.66	0.61	0.79	0.62	0.44	0.66	0.56	0.90	0.43	0.63

Table 5. MobileNet explainaiblity localization evaluation.

	GradCAM++	GradCAM	HiResCAM	GradCAMElementWise	XGradCAM	LayerCAM	EigenCAM	EigenGradCAM	RandomCAM	Mean
cassava brown leaf spot	0.84	0.78	0.82	0.82	0.77	0.84	0.91	0.80	0.74	0.81
cassava healthy	0.74	0.72	0.73	0.72	0.67	0.73	0.74	0.71	0.61	0.71
cassava mosaic	0.71	0.69	0.69	0.71	0.68	0.72	0.79	0.68	0.63	0.70
cassava root rot	0.56	0.51	0.51	0.53	0.60	0.54	0.63	0.45	0.42	0.53
corn brown spots	0.70	0.64	0.64	0.62	0.59	0.64	0.72	0.64	0.56	0.64
corn healthy	0.76	0.72	0.81	0.75	0.73	0.79	0.71	0.73	0.64	0.74
corn streak	0.71	0.77	0.77	0.73	0.71	0.75	0.80	0.80	0.63	0.74
corn stripe	0.92	0.87	0.87	0.84	0.80	0.87	0.94	0.91	0.74	***0.86***
corn yellowing	0.53	0.53	0.49	0.49	0.49	0.51	0.63	0.51	0.46	0.52
corn leaf blight	0.72	0.67	0.68	0.65	0.64	0.67	0.75	0.66	0.58	0.67
tomato brown spots	0.46	0.42	0.49	0.51	0.43	0.50	0.56	0.69	0.40	0.50
tomato blight leaf	0.56	0.51	0.51	0.56	0.51	0.55	0.67	0.66	0.51	0.56
tomato leaf yellow virus	0.52	0.47	0.45	0.54	0.50	0.51	0.79	0.70	0.43	0.54

Table 6. ResNet explainaiblity localization evaluation.

	GradCAM++	GradCAM	HiResCAM	GradCAMElementWise	XGradCAM	LayerCAM	EigenCAM	EigenGradCAM	RandomCAM	Mean
cassava brown leaf spot	0.86	0.85	0.85	0.84	0.80	0.85	0.89	0.89	0.74	**0.84**
cassava healthy	0.74	0.74	0.74	0.73	0.69	0.73	0.72	0.78	0.65	**0.72**
cassava mosaic	0.78	0.77	0.76	0.75	0.68	0.75	0.81	0.84	0.62	**0.75**
cassava root rot	0.64	0.63	0.64	0.63	0.59	0.63	0.68	0.75	0.56	**0.64**
corn brown spots	0.65	0.64	0.64	0.63	0.61	0.64	0.69	0.71	0.54	**0.64**
corn healthy	0.85	0.85	0.85	0.81	0.76	0.83	0.83	0.89	0.68	**0.82**
corn streak	0.77	0.77	0.77	0.75	0.73	0.75	0.82	0.82	0.66	**0.76**
corn stripe	0.87	0.86	0.86	0.84	0.81	0.85	0.91	0.92	0.74	*0.85*
corn yellowing	0.55	0.55	0.55	0.53	0.50	0.54	0.60	0.65	0.44	**0.55**
corn leaf blight	0.72	0.72	0.71	0.69	0.64	0.70	0.78	0.79	0.59	**0.70**
tomato brown spots	0.51	0.51	0.51	0.49	0.43	0.50	0.65	0.65	0.39	**0.52**
tomato blight leaf	0.64	0.64	0.64	0.62	0.56	0.63	0.78	0.77	0.48	**0.64**
tomato leaf yellow virus	0.68	0.67	0.69	0.66	0.53	0.67	0.86	0.86	0.53	**0.68**

Table 7. ShuffleNet explainaiblity localization evaluation.

	GradCAM++	GradCAM	HiResCAM	GradCAMElementWise	XGradCAM	LayerCAM	EigenCAM	EigenGradCAM	RandomCAM	Mean
cassava brown leaf spot	0.66	0.76	0.72	0.76	0.76	0.74	0.81	0.78	0.73	*0.75*
cassava healthy	0.56	0.67	0.63	0.63	0.61	0.59	0.75	0.68	0.64	0.64
cassava mosaic	0.57	0.62	0.64	0.65	0.65	0.65	0.71	0.72	0.63	0.65
cassava root rot	0.46	0.55	0.51	0.52	0.49	0.51	0.49	0.49	0.42	0.49
corn brown spots	0.49	0.50	0.52	0.54	0.52	0.57	0.44	0.55	0.49	0.51
corn healthy	0.61	0.57	0.57	0.61	0.60	0.56	0.67	0.74	0.66	0.62
corn streak	0.62	0.70	0.65	0.65	0.58	0.73	0.53	0.70	0.61	0.64
corn stripe	0.67	0.71	0.65	0.69	0.70	0.70	0.72	0.72	0.70	0.70
corn yellowing	0.34	0.32	0.37	0.45	0.41	0.41	0.50	0.57	0.44	0.42
corn leaf blight	0.54	0.54	0.56	0.59	0.56	0.60	0.54	0.57	0.57	0.56
tomato brown spots	0.35	0.43	0.44	0.56	0.41	0.63	0.48	0.76	0.38	0.49
tomato blight leaf	0.42	0.48	0.52	0.53	0.42	0.58	0.39	0.55	0.40	0.48
tomato leaf yellow virus	0.42	0.57	0.70	0.66	0.50	0.76	0.40	0.75	0.42	0.57

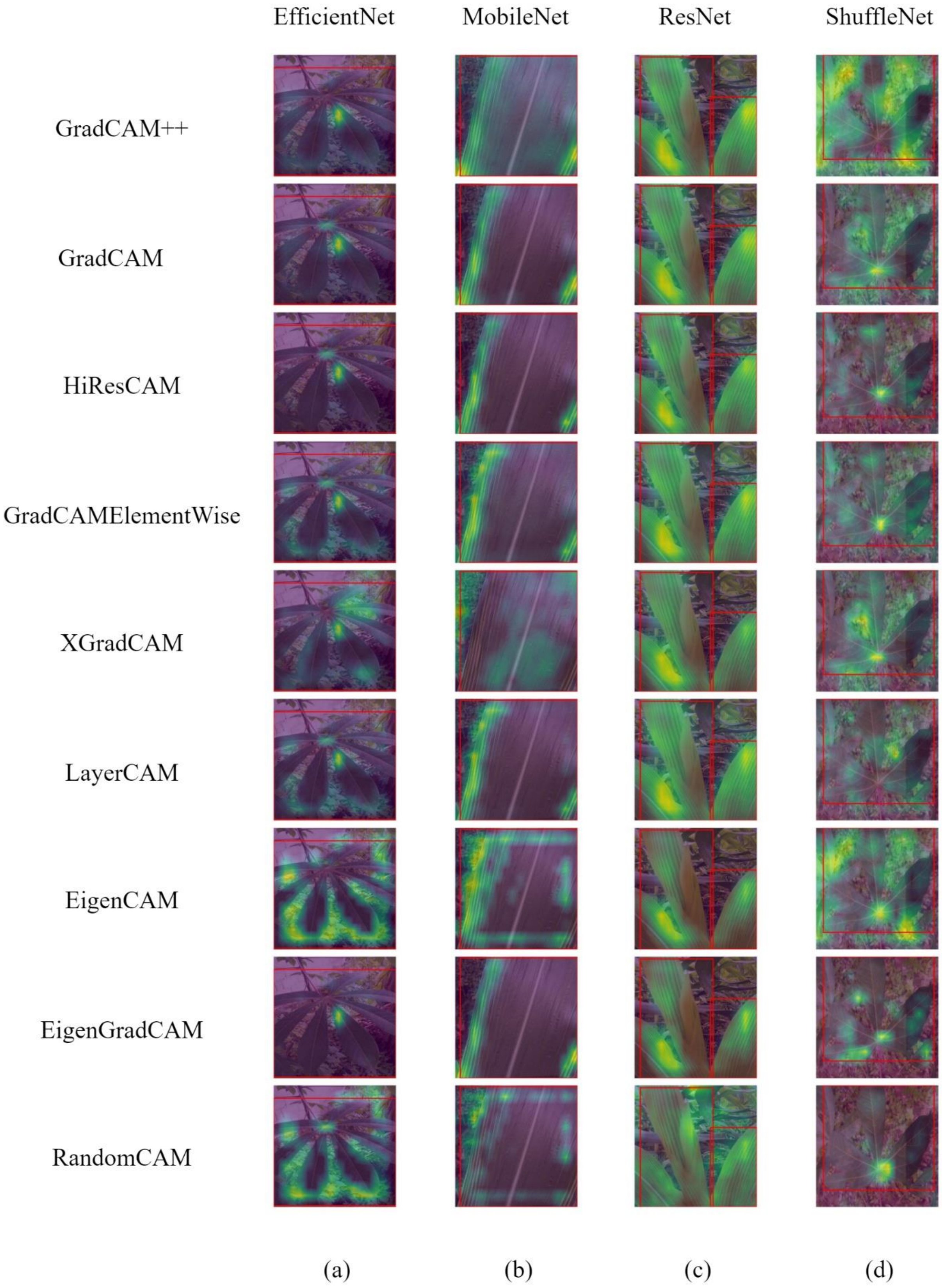

Figure 4. Explainability maps for the classes where the models show the highest localization evaluation proportion.

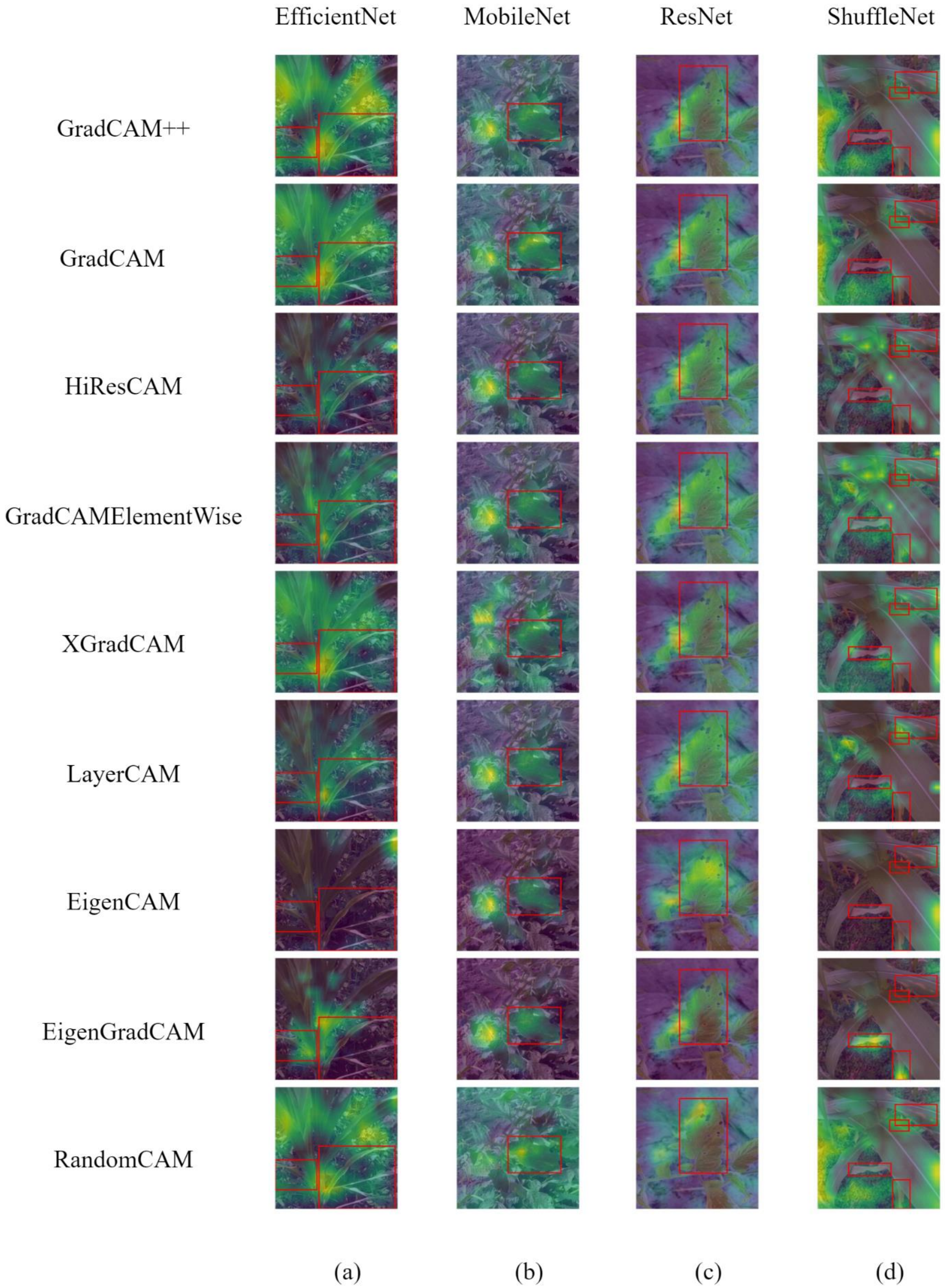

Figure 5. Explainability maps for the classes where the models showed the lowest success localization evaluation proportion.

within images, regardless of the model or class. These results demonstrate the potential of these techniques as valuable tools for enhancing the interpretability of deep learning models in various classification tasks.

Explainability maps of randomly selected images from classes with the highest and lowest localization evaluation proportions for each model are shown in Fig. 4 and Fig. 5. These figures highlight the regions of interest identified by each model during predictions. The following outcomes emerged from evaluating these images based on their localization proportions.

✓ Deep learning models capture the colors and textures of lesions associated with particular diseases during classification, mirroring human decision-making processes.

✓ The larger area of the bounding box results in features extracted for classification originating from a broader visual area (Fig. 4(a), Fig. 4(b), Fig. 4(d)). Consequently, this leads to a high proportion for the respective class. For example, when most of the image comprises plants belonging to the specific anomaly, it contributes to the higher proportion.

✓ As seen in Fig. 4(c), the high proportion is also related to the robustness of the model. Extracting features from the correct area in classification ensures a high proportion. When the model accurately focuses on the diseased parts of the plant, it demonstrates its reliability and robustness in disease detection, which is crucial for practical applications in agriculture.

✓ When features are extracted from inappropriate regions in classification, possibly due to bias, it results in a lower proportion (Fig. 5(b)). For instance, the model may focus on the background or other plants rather than the anomaly present in the target plant. This misdirection results in inaccurate classifications and diminished model performance.

✓ Models may generate poor explainability proportion due to the presence of characteristic features of the relevant class scattered across multiple areas in the image (indicated by multiple bounding boxes, Fig. 5(a)). In such instances, while features extracted from a smaller area might be adequate for classification, the abundance of ground truth areas could lead to a decrease in the proportion. For instance, if anomalies are present in three different leaves of a plant, the model might classify them using only the features extracted from a single leaf. This highlights the challenge of ensuring that all relevant features are considered for accurate classification.

✓ The explainability proportion is adversely impacted when essential features required for classification are localized within a small area (indicated by multiple small bounding boxes, Fig. 5(d)). In such scenarios, the model encounters difficulties in decision-making, as it grapples with identifying the relevant regions, thus compromising its robustness. For instance, this occurs when an anomaly in a plant is confined to a tiny portion of a leaf. The struggle of model to focus on such small, yet critical areas, results in lower proportions.

✓ Finally, as observed in the example shown in Fig. 5(c), even if the necessary features for the classifier are extracted from the correct area, the proportion may still be low if the ground truth area is a subset of this larger area. For example, extracting features from the entire leaf for an anomaly located in a smaller area of the leaf. This issue underscores the importance of precise localization in feature extraction to ensure high explainability proportions and accurate disease detection.

Overall, the size of the bounding box significantly influences the explainability proportion—larger bounding boxes typically yield higher proportions because they encompass more relevant visual features, especially when the majority of the image depicts the anomaly. High localization proportions also indicate model robustness, as seen when the model focuses accurately on the diseased areas, reinforcing its reliability in real-world agricultural scenarios. Conversely, when the model extracts features from irrelevant regions—due to bias or misdirection—the proportion drops, leading to misclassifications. In some cases, scattered anomaly features across the image lower explainability even if the classification is correct, because the model may only use a subset of the relevant information. Similarly, when anomalies are confined to very small regions, the model struggles to correctly identify and extract features, negatively impacting both the proportion and robustness. Lastly, even when correct areas are targeted, if the model uses a larger-than-necessary region, explainability suffers, emphasizing the need for precise localization.

Although the explanations offer valuable insights, it is important to note that their reliability is inherently tied to the performance of the underlying model. If the model's accuracy is compromised, the explanations may become misleading or inconsistent. Consequently, the effectiveness of model explanations depends on the model's overall performance, and caution should be exercised when interpreting explanations generated by suboptimal models.

5 Conclusions

By leveraging vast amounts of data, DL models can accurately identify and classify a wide range of plant diseases, often surpassing traditional methods in speed and accuracy. However, evaluating the effectiveness of these models reveals that relying solely on classification performance is insufficient. This limitation arises from the varied data used for model training, which is collected from different devices like smartphones and cameras under inconsistent environmental conditions. The lack of standardization in real-world scenarios leads to disparities in lighting, zoom levels, resolutions, and other factors, introducing biases into the dataset. Consequently, models trained on such biased datasets inherently reflect these biases, ultimately limiting their ability to generalize well to unseen data.

In this paper, the evaluation was expanded beyond conventional classification metrics. In addition to assessing the models based on their classification performance, CAM-based explainability scores were also employed. This process involves comparing the explainability maps against the ground truth provided by the bounding boxes in the validation dataset to assess how accurately the models can localize the areas of interest related to plant diseases. By analyzing the localization proportions, it becomes possible to not only assess the accuracy and reliability of the models' predictions but also to uncover flaws or limitations within the dataset itself. This dual evaluation process provides a comprehensive understanding of the model performance and highlights areas where the dataset may need improvement. Potential improvements may encompass enriching data collection methodologies through the inclusion of more diverse samples, ensuring uniformity in imaging and labeling conditions for consistency, and implementing cutting-edge preprocessing techniques to further elevate dataset quality.

This paper highlights the critical importance of incorporating explainability metrics into the evaluation framework for deep learning models. Doing so offers transparent insight into the decision-making mechanisms of the model, thereby enhancing trust and understanding among stakeholders regarding the results. This transparency enables farmers and agronomists to make informed adjustments based on their knowledge and local conditions. Moreover, decisions based on model predictions can have far-reaching consequences. From optimizing crop yields to minimizing pesticide usage and conserving water resources, the impact of these decisions extends beyond individual farms to broader environmental and socio-economic landscapes. Therefore, ensuring the transparency and interpretability of deep learning models becomes imperative for fostering responsible and sustainable agricultural practices.

The proposed approach not only streamlines the identification of potential biases and errors in the detection process but also fosters a deeper investigation into how these issues affect model performance.

6 Ethics committee approval and conflict of interest statement

"There is no need for permission from the ethics committee for the article prepared."

"There is no conflict of interest with any person or institution in the article prepared."

7 Author contribution statements

Author 1 contributed to the idea formation, design, literature review, result evaluation, and content writing of the article.

8 References

[1] da Silveira F, Lerme, FH, Amaral FG. "An overview of agriculture 4.0 development: Systematic review of descriptions, technologies, barriers, advantages, and disadvantages". *Computers and Electronics in Agriculture*, 189, 106405, 2021.

[2] Albahar M. "A Survey on Deep Learning and Its Impact on Agriculture: Challenges and Opportunities". *Agriculture*, 13(3), 540, 2023.

[3] Saranya T, Deisy C, Sridevi S, Anbananthen KSM. "A comparative study of deep learning and Internet of Things for precision agriculture". *Engineering Applications of Artificial Intelligence*, 122, 106034, 2023.

[4] Ahmad A, Saraswat D, El Gamal A. "A survey on using deep learning techniques for plant disease diagnosis and recommendations for development of appropriate tools". Smart Agricultural Technology, 3, 100083, 2023.

[5] Abade A, Ferreira PA, de Barros Vidal F. "Plant diseases recognition on images using convolutional neural networks: A systematic review". *Computers and Electronics in Agriculture*, 185, 106125, 2021.

[6] Ding W, Abdel-Basset M, Hawash H, Ali AM. "Explainability of artificial intelligence methods, applications and challenges: A comprehensive survey". *Information Sciences*. 615, 238-292, 2022.

[7] Ghosal S, Blystone D, Singh AK, Ganapathysubramanian B, Singh A, Sarkar S. "An explainable deep machine vision framework for plant stress phenotyping". *Proceedings of the National Academy of Sciences*, 115(18), 4613-4618, 2018.

[8] Nagasubramanian K, Jones S, Singh AK, Sarkar S, Singh A, Ganapathysubramanian B. "Plant disease identification using explainable 3D deep learning on hyperspectral images". *Plant Methods*, 15, 1-10, 2019.

[9] Brahimi M, Mahmoudi S, Boukhalfa K, Moussaoui A. "Deep interpretable architecture for plant diseases classification". *2019 IEEE Signal Processing: Algorithms, Architectures, Arrangements, and Applications*, 111-116, Poznan, Poland, 18-20 September 2019.

[10] Arvind C, Totla A, Jain T, Sinha N, Jyothi R, Aditya K, Farhan M, Sumukh G, Ak G. "Deep Learning Based Plant Disease Classification with Explainable AI and Mitigation Recommendation". *2021 IEEE Symposium Series on Computational Intelligence*, 1-8, Orlando, USA, 05-07 December 2021.

[11] Batchuluun G, Nam SH, Park KR. "Deep learning-based plant classification and crop disease classification by thermal camera". *Journal of King Saud University-Computer and Information Sciences*, 34(10), 10474-10486, 2022.

[12] Wei K, Chen B, Zhang J, Fan S, Wu K, Liu G, Chen D. "Explainable deep learning study for leaf disease classification". *Agronomy*, 12(5), 1035, 2022.

[13] Shah D, Trivedi V, Sheth V, Shah A, Chauhan U. "ResTS: Residual deep interpretable architecture for plant disease detection". *Information Processing in Agriculture*, 9(2), 212-223, 2022.

[14] Shoaib M, Shah B, Sayed N, Ali F, Ullah R, Hussain I. "Deep learning for plant bioinformatics: an explainable gradient-based approach for disease detection". *Frontiers in Plant Science*, 14, 2023.

[15] Thakur P S, Chaturvedi S, Khanna P, Sheorey T, Ojha A. "Vision transformer meets convolutional neural network for plant disease classification". *Ecological Informatics*, 77, 102245, 2023.

[16] Chhetri TR, Hohenegger A, Fensel A, Kasali MA, Adekunle AA. "Towards improving prediction accuracy and user-level explainability using deep learning and knowledge graphs: A study on cassava disease". *Expert Systems with Applications*, 233, 120955, 2023.

[17] Ghosh P, Mondal AK, Chatterjee S, Masud M, Meshref H, Bairagi AK. "Recognition of sunflower diseases using hybrid deep learning and its explainability with AI". *Mathematics*, 11(10), 2241, 2023.

[18] Yang S, Xing Z, Wang H, Gao X, Dong X, Yao Y, Zhang R, Zhang X, Li S, Zhao Y, Liu Z. "Classification and localization of maize leaf spot disease based on weakly supervised learning". *Frontiers in Plant Science*, 14, 1128399, 2023.

[19] Toda Y, Okura F. "How convolutional neural networks diagnose plant disease". *Plant Phenomics*, 9237136, 2019.

[20] Moupojou E, Tagne A, Retraint F, Tadonkemwa A, Wilfried D, Tapamo H, Nkenlifack M. "FieldPlant: A dataset of field plant images for plant disease detection and classification with deep learning". *IEEE Access*, 11, 35398-35410, 2023.

[21] Karahan T, Nabiyev V. "Plant identification with convolutional neural networks and transfer learning". *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi*, 27(5), 638-645, 2021

[22] OpenMMLab. "MMPreTrain". https://github.com/open-mmlab/mmpretrain (08.10.2024).

[23] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. "Grad-CAM: visual explanations from deep

networks via gradient-based localization". *International Journal of Computer Vision*, 128, 336-359, 2020.

[24] Wang H, Wang Z, Du M, Yang F, Zhang Z, Ding S, Mardziel, P, Hu X. "Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks". *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 24-25, Seattle, USA, 14-19 June 2020.

[25] Tan M., Quoc L. "Efficientnet: Rethinking model scaling for convolutional neural networks". *2019 International conference on machine learning,* 1-11, Long Beach, California, 9-15 June 2019.

[26] Sandler M, Howard A, Zhu M, Zhmoginov A, Chen C. "MobileNetV2: Inverted Residuals and Linear Bottlenecks," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4510-4520, Salt Lake City, USA, 18-23 June 2018.

[27] He K, Zhang X, Ren S, Sun J. "Deep Residual Learning for Image Recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770-778, Las Vegas, USA, 27-30 June 2016.

[28] Zhang X, Zhou X, Lin M, J. Sun, "ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6848-6856, Salt Lake City, USA, 18-23 June 2018.

[29] Chattopadhay A, Sarkar A, Howlader P, Balasubramanian VN. "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks". *2018 IEEE winter conference on applications of computer vision (WACV)*, 839-847, Lake Tahoe, USA 12-15 March 2018.

[30] Wang H, Wang Z, Du M, Yang F, Zhang Z, Ding S, Hu X. "Score-CAM: Score-weighted visual explanations for convolutional neural networks." *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW),* 111-119, 14-19 June 2020.

[31] Muhammad MB, Yeasin M. "Eigen-cam: Class activation map using principal components." *2020 IEEE International Joint Conference on Neural Networks (IJCNN)*, 1-7, Glasgow, UK, 19-24 July 2020.

[32] Farjon G, Liu H, Yael E. "Deep-learning-based counting methods, datasets, and applications in agriculture: A review." *Precision Agriculture* ,24, 1683-1711, 2023.

[33] Chakraborty SK, Chandel NS, Jat D, Tiwari MK, Rajwade YA, Subeesh, A. "Deep learning approaches and interventions for futuristic engineering in agriculture." *Neural Computing and Applications,* 34, 20539-20573, 2022.