IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 15, No. 6, December 2026, pp. 5764~5773
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i6.pp5764-5773

Robust tomato diseases detection against background variations with attentive region proposal networks

Laila Mukhrina Azizi[1], Ade Ramdan[2], Vicky Zilvan[2], Evandri[2], Hilman Ferdinandus Pardede[2], Ana Hadiana[3], Dicky Rianto Prajitno[2]
[1]Department of Statistics, Faculty of Mathematics and Natural Sciences, Syiah Kuala University, Banda Aceh, Indonesia
[2]Research Center for Artificial Intelligence and Cyber Security, National Research and Innovation Agency, Bandung, Indonesia
[3]Research Center for Data and Information Science, National Research and Innovation Agency, Bandung, Indonesia

Article Info

Article history:

Received Jul 3, 2025
Revised Aug 24, 2026
Accepted Sep 20, 2026

Keywords:

Attention module
Densely connected convolutional network 201
Extreme inception
Region proposal network
Tomato disease detection

ABSTRACT

Studies of using artificial intelligence to detect plant diseases have achieved good accuracies when evaluated under controlled conditions. However, their robustness is still a challenge when there exist background variations on the data. This paper proposes the use of attentive region proposal networks (ARPN) to improve the robustness of plant disease detection. ARPN combines region proposal networks (RPN) with an attention mechanism added on top of the network. Seven convolutional neural network (CNN) models, namely mobile network version 2 (MobileNetV2), visual geometry group 16 layers (VGG16), VGG19, extreme inception (Xception), densely connected convolutional network 201 layers (DenseNet201), residual network 50 layers (ResNet50), and inception version 3 (InceptionV3), were employed as feature extractors and the proposed ARPN architecture was placed on top of them. Evaluations on various types of attention mechanisms revealed that the convolutional bottleneck attention model (BAM) was the most effective. The model was evaluated on a tomato-specific subset derived from the PlantDoc dataset. The experiments show that the proposed framework could improve the model's robustness against background variations. The ablation analysis shows that applying RPN alone on top of CNN models improves the model's robustness. DenseNet201 achieves the best performance, showing a 2.31% improvement with RPN. Adding an attention mechanism could further improve the results, with the highest accuracy of 60.48% achieved by Xception.

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

Corresponding Author:

Evandri
Research Center for Artificial Intelligence and Cyber Security, National Research and Innovation Agency
Samaun Samadikun Science and Technology Area, Sangkuriang Street, Bandung 40135, Indonesia
Email: evan001@brin.go.id

1. INTRODUCTION

Plant diseases are a major cause of crop losses [1] and quality degradation of crop products [2]. The study found that most of the diseases could be identified from visual inspections of their leaves [3]. However, such identifications usually require trained experts for accurate results [4]. The use of artificial intelligence and computer vision with convolutional neural networks (CNN) has increased rapidly and shows tremendous performance under controlled conditions [5]. For instance, visual geometry group 16 layers (VGG16), residual network (ResNet), and inception version 3 (InceptionV3) are used in [6] and the study reported that an accuracy of 98.49% can be achieved on the PlantVillage dataset. PlantVillage datasets are popular datasets for benchmarking plant disease detection. It comprises of leave images of infected plants. In this dataset, the data are collected under controlled conditions. Many studies have reported the evaluation of their method on

the PlantVillage dataset or its subset. Some studies explore some pre-trained models like VGG16, InceptionV3, and mobile network (MobileNet), and a custom CNN on a subset of the same dataset, and achieve an accuracy of 91.2% [7]. A custom CNN is also proposed in [8] with an accuracy of 88.17%. Modifying CNN such as dense networks is reported in [9]. In this study, dense residual networks are modified for tomato plants, and a performance of more than 95% is achieved. Using wide networks such as Google network (GoogleNet), Inception, and ResNet is reported in [10] and the results of near perfect accuracy can be achieved with an area under the curve (AUC) of 99.72% and an accuracy of 99.39%. Fine-grained visual classification tasks with tomato leaf images with three networks: a location network, a feedback network, and a classification network are employed in [11] and find it even more effective than pretrained CNN models.

However, the challenges of applying the model in real environments remain. Due to background variations, CNN often fail to achieve good results. There are several sources of background variations that may exist on images under real environmental conditions. Some of them are variations of changes in illumination conditions, angle of image acquisition, and the existence of other background objects. These conditions cause mismatches with training conditions where the lighting, angle, and background are usually controlled, causing the models to fail to generalize in testing conditions. For instance, in [12], a comparative study on using CNN for plant disease classification in real environments showed a significant drop with results under 40% accuracy.

Some studies have tried to improve the robustness of CNN models with various computer vision approaches by applying object detection and segmentation before image classification [13]. Deep learning-based image detectors such as you only look once version 5 (YOLOv5) is implemented in [14] with some success to improve the robustness of the models. The study applies ResNet50 and MobileNetV2 on the PlantDoc dataset. But compared to the results on PlantVillage, it is significantly worse. The PlantDoc dataset is a dataset of images of plant leaves that are infected by diseases similar to PlantVillage but are collected from the internet. Hence, the images have various backgrounds and watermarks in them. Other object detection methods, i.e. faster region-based convolutional neural network (R-CNN), and k-means clustering for segmentation are employed with VGG16 to improve the robustness of deep learning models. An improvement of 2.71% is reported in the study [15]. Other studies [16], employ R-CNN but with a ResNet34 backbone as CNN.

While studies have suggested that the use of object detection like YOLO and Faster-RCNN has improved the robustness of the models against background variations, these studies use large architectures of them, increasing the computational load for training. In RCNN, the backbone for identification is called regional proposal networks (RPNs) [17]. It works by identifying nearby information and proposing areas that are suggested to be the same objects. Applying object detection alone is also not that effective without any image segmentation because some of the background images still remaining on the detected areas. This may cause additional computations for image segmentation and background removal.

Recent studies have continuously addressed the challenge of background variations in in-the-wild plant disease imagery. For instance, approaches using vegetative feature isolation before classification have proven effective in minimizing background noise and enhancing model performance [18]. This improvement in accuracy is further driven by the use of advanced CNN architectures, such as extreme inception (Xception), combined with multi-scale feature extraction and fusion methods [19]. The reliability of the Xception model in extracting complex patterns has been further validated in related crop studies through the application of wrapper feature selection and feature concatenation techniques [20]. While these segmentation and feature engineering techniques show promising results, they often require complex preprocessing pipelines. Therefore, to provide a more dynamic and spatially focused alternative, this study proposes the integration of RPN with an attention mechanism.

Attention mechanisms have been incorporated into CNN models to improve plant disease classification. For example, Bhujel *et al.* [21] employed the convolutional block attention module (CBAM), while squeeze-and-excitation (SE) blocks were adopted in [22]. The latter improved accuracy on the tomato plant dataset from 96.81% to 99.24%. In these studies, the robustness problems are not evaluated.

This study develops frameworks for robust plant disease detection by incorporating RPN into CNN architectures and adding attention mechanisms to them. Applying attention mechanisms may reduce the need for image segmentation due to their ability to focus on specific image regions, thereby reducing the impact of background images on classification. Various attention mechanisms are explored, namely CBAM, SE, and bottleneck attention model (BAM). The proposed method is named attentive region proposal networks (ARPN). Seven CNN models are explored as backbone architectures for feature extraction in the RPN.

2. METHOD

The block diagram of the proposed method is shown in Figure 1. It comprises data preprocessing, ARPN modules, and classifiers. The ARPN modules consist of an RPN CNN backbone, RPN, and attention

mechanisms. The outputs of ARPN are then fed into a fully connected layer for classification. Seven types of CNN backbones and three types of attention mechanisms are explored. The CNN backbones include MobileNetV2, VGG16, VGG19, Xception, densely connected convolutional network 201 layers (DenseNet201), ResNet50, and InceptionV3, while the attention mechanisms include CBAM, SE, and BAM. During data preprocessing, all images are first resized to 224 × 224 × 3. After RGB color features are extracted, the features are normalized to the range of 0–1. The processed data are then fed into the CNN backbones, which act as feature learners to extract useful features from the data. For all CNN backbones, pre-trained models are used as the initial models, with their classifier layers removed.

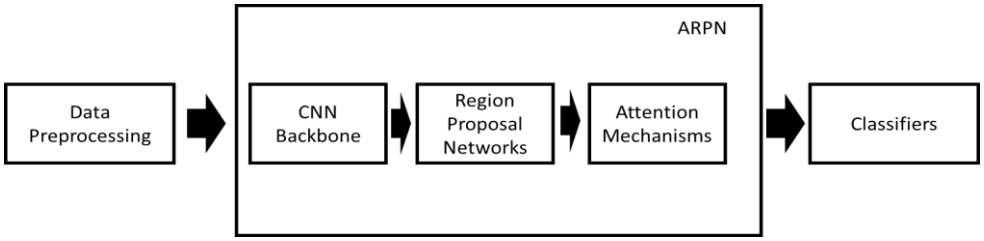


Figure 1. Block diagram of the proposed method

In this study, seven CNN architectures, namely MobileNetV2, VGG16, VGG19, Xception, DenseNet201, ResNet50, and InceptionV3, were employed as backbone networks for the RPN to detect tomato leaf diseases. The selection of multiple backbones is motivated by differences in architectural design and feature extraction capability across CNN models. Previous studies have shown that pre-trained CNNs such as DenseNet201, MobileNetV2, ResNet50, VGG16, and Xception are effective in extracting discriminative features from plant leaf images, contributing to improved disease detection performance depending on the dataset and training configuration [23].

VGG16 and VGG19 extract stable low-level and mid-level visual features through deep convolutional layers [24], whereas ResNet50 employs residual connections to facilitate learning in deep networks and improve feature representation [25]. InceptionV3 and Xception are designed to capture multi-scale features efficiently, making them suitable for images with varying lesion sizes and textures [26]. DenseNet201 enhances feature reuse and gradient flow through dense connectivity [27], whereas MobileNetV2 offers a lightweight architecture with competitive accuracy, enabling an analysis of the trade-off between detection performance and computational complexity [28].

The RPN is a convolutional network that produces a proposal area of the target object [17]. It operates by estimating the boundaries of the object and calculating the probabilities of the proposed region having the target object. Given input data, the ratio between the width and height of them are calculated, and the scale, which is the size of the image, is determined. Then, with every feature map from the convolutional layers and each point as an anchor, anchor boxes are generated and then slid across the entire feature map. Then each anchor box is classified, and using the intersection over union metric, the probability of the anchor as foreground or background is determined. The output of the RPN is bounding box predictions and anchor box classification that can be used as the boundary to crop the images into the target of the interesting part of the images. In this case, most of the bounding boxes are expected to correspond to the infected areas of the leaves.

Figure 2 shows an RPN built on top of a CNN backbone: an input RGB image (e.g., 224 × 224 × 3) is first converted into a compact feature map that encodes high-level visual patterns. A sliding-window 3 × 3 conv produces a 256-d intermediate representation at each feature-map location, and two 1 × 1 heads output "2k objectness scores" (object vs. background) and "4k bounding-box offsets" for "k anchor boxes" of different scales/aspect ratios. These scores and offsets are then used to generate refined region proposals that indicate candidate lesion/object locations.

Attention mechanisms are applied after the RPN block. By having an attention mechanism, the networks learn to compute attention weights representing the importance of various parts of the given data [29]. While originally used for time-dependent data, many studies have shown that attention mechanisms benefit spatial data as well. In spatial data like images, the attention mechanism could learn and focus on certain areas of the image. Adding an attention mechanism on top of RPN is expected to enable the networks to focus on the regions of interest identified by RPN while reducing attention to the remaining background areas. Three types of attention mechanisms: the SE module, the CBAM, and the BAM are explored.

CBAM combines channel and spatial attention, where channel attention aims to find the attended channel while spatial attention focuses on finding areas in the spatial domain that reflect greater importance [30]. BAM is a novel attention mechanism that aims to find key features from the feature map. In a sense, it works like an encoder but can focus on the key features instead of finding a compressed representation of the data [31]. The SE apply global pooling on features to find squeezed features (like bottleneck features) [32]. After the proposed ARPN module, two fully connected layers are used as classifiers, with the final fully connected layer serving as the output layer. The ARPN output is first flattened before being passed to the fully connected layers.

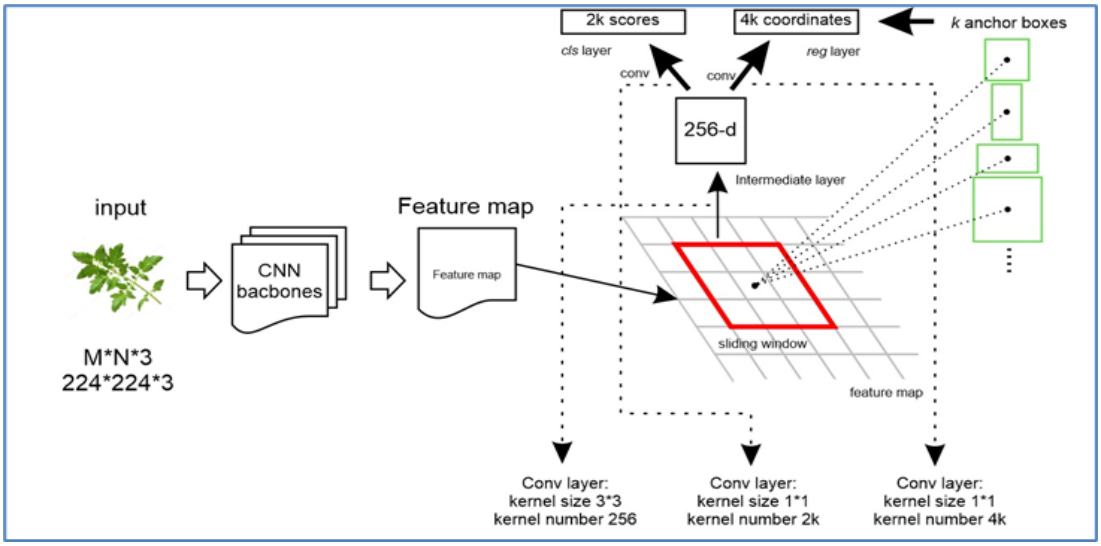


Figure 2. RPN architecture

3. RESULTS AND DISCUSSION

3.1. Setup

In this study, the PlantDoc tomato subset consists of 637 images. The dataset is split into training and testing sets using a fixed ratio (e.g., 80% for training and 20% for testing) to ensure a fair and consistent evaluation of model performance. The relatively small size of the dataset reflects real-world conditions, where labeled data for plant disease detection are often limited. To enhance model generalization, real-time data augmentation—specifically 20% shear, 20% zoom, and random horizontal flipping—was applied exclusively to the training set. Meanwhile, all images across both the training and testing sets were uniformly resized to a spatial resolution of 224 × 224 pixels and normalized to a [0, 1] range using a 1/255 scaling factor, ensuring that the final evaluation was conducted on undistorted data.

The proposed method was evaluated using a subset of the PlantDoc dataset [13]. PlantDoc is a dataset that comprises leaf images of infected plants collected from Google Images and represents images captured under real-world conditions. Unlike PlantVillage, the PlantDoc dataset contains images with various backgrounds and watermarks. Therefore, it is widely used to evaluate the robustness of models against background variations. The dataset comprises seven classes of tomato plant diseases commonly found in tomato plants, with a total of 637 images. Figure 3 shows sample images from the PlantDoc dataset before and after preprocessing, where Figure 3(a) presents the original images and Figure 3(b) shows the resized and standardized images used for training.

The dataset itself has been divided into test and training set. The hyperparameters include two learning rates, 10^{-4} and 10^{-5}, the adaptive moment estimation (Adam) optimizer, a batch size of 32, and 50 training epochs. Transfer learning is used for each backbone model by using pre-trained models that are trained using image network (ImageNet) and excluding the top layers (classifier layers). The entire model was fine-tuned during training.

3.2. Evaluation

First, the performance of each CNN backbone was evaluated with and without RPN. Table 1 summarizes their accuracies. The results demonstrate that the addition of RPN generally enhances the robustness and accuracy of the models, albeit with an expected increase in computational cost, as indicated by the higher giga floating-point operations per second (GFLOPS). The performance improves for most CNN

backbones. However, lower performance is observed only for InceptionV3, ResNet50, and Xception when using a learning rate of 10^{-4}. Without RPN, DenseNet201 achieves the highest accuracy. With RPN, DenseNet201 also achieves the highest accuracy of 58.13%, representing a 2.32% improvement. However, the largest improvement is achieved by VGG16, with an absolute increase of 30.23% when using RPN. VGG19 also shows a significant improvement with RPN integration. The performance of RPN-integrated models is highly sensitive to the selected learning rate. These findings indicate that although RPN effectively guides the network to focus on relevant disease regions, optimal hyperparameter tuning is crucial to fully harness its capabilities.



(a)



(b)

Figure 3. Samples of PlantDoc datasets before and after preprocessing (a) before preprocessing and (b) after preprocessing

The accuracy differences presented in Table 1 can be further explained by empirical observations of the models' convergence behavior during the training phase. For the majority of the CNN backbones (such as DenseNet201 and InceptionV3), models trained with a 10^{-4} learning rate exhibit faster convergence and lower validation loss than those trained with 10^{-5}, indicating more effective optimization and better generalization. In these cases, the smaller learning rate tends to produce slower and stagnant loss reduction, increasing the risk of getting trapped in local minima. However, it is crucial to note that this behavior is not universal; architectures with distinct structural characteristics, particularly Xception and the VGG networks, demonstrated the opposite trend. For these models, a smaller learning rate was necessary to prevent unstable gradient updates, thereby achieving optimal stability and higher peak accuracy.

Table 1. Comparison of CNN backbone performance (accuracy) with and without RPN

CNN backbone	Learning rate	Without RPN	GFLOPS	With RPN (%)	GFLOPS
MobileNetV2	10^{-4}	40.31	613.5	46.51	974.0
	10^{-5}	39.53		43.41	
VGG16	10^{-4}	21.70	3.074×10^{4}	22.48	3.09×10^{4}
	10^{-5}	21.70		51.93	
VGG19	10^{-4}	21.70	3.906×10^{4}	26.35	3.923×10^{4}
	10^{-5}	41.08		48.83	
Xception	10^{-4}	53.48	9.134×10^{3}	47.28	9.669×10^{3}
	10^{-5}	51.93		56.59	
DenseNet201	10^{-4}	55.81	8.74×10^{3}	58.13	9.137×10^{3}
	10^{-5}	52.71		55.03	
ResNet50	10^{-4}	40.31	7.751×10^{3}	38.75	8.286×10^{3}
	10^{-5}	37.98		36.00	
InceptionV3	10^{-4}	53.48	5.694×10^{3}	50.38	5.966×10^{3}
	10^{-5}	49.61		48.83	

The performance of the proposed methods using various attention mechanisms is presented in Table 2. Adding an attention mechanism improves the robustness of some models. However, a comparison of Tables 1 and 2 indicates that incorporating attention mechanisms does not necessarily improve performance compared with using RPN alone. Improvements are observed when VGG19, MobileNetV2, Xception, and

ResNet50 are used as CNN backbones with SE attention. Interestingly, although ResNet50 and InceptionV3 exhibit lower performance when using RPN alone, their performance can be improved by incorporating attention mechanisms.

Both BAM and CBAM generally perform better than SE as attention mechanisms. BAM generally outperforms RPN alone, providing greater improvements in model robustness on average than RPN alone and RPN combined with SE. Almost all CNN backbones show improved robustness with BAM at certain learning rates. Among all the evaluated attention mechanisms, CBAM achieves the best overall performance. Similar to BAM, CBAM improves the robustness of almost all models. The highest accuracy across all evaluated models is achieved by Xception with CBAM, reaching 60.48%. Similar to the results in Table 1, a learning rate of 10^{-4} is also found to be optimal.

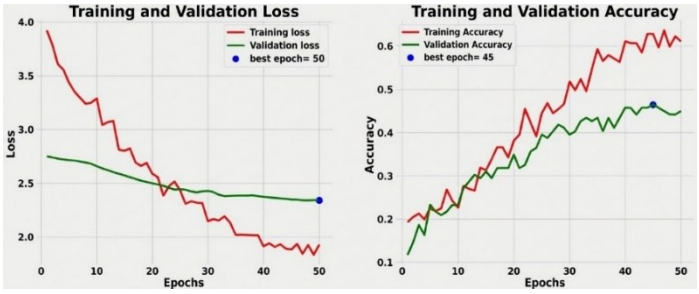
Table 2. The performance (accuracy) of the proposed method APRN on various attention mechanisms

CNN backbone	Learning rate	RPN + SE (%)	GFLOPS	RPN + BAM (%)	GFLOPS	RPN + CBAM (%)	GFLOPS
MobileNetV2	10^{-4}	36.43	974.5	45.73	975.3	39.53	974.1
	10^{-5}	44.96		49.61		48.06	
VGG16	10^{-4}	21.00	3.09×10^{4}	44.18	3.09×10^{4}	41.86	3.09 × 104
	10^{-5}	46.00		55.03		50.38	
VGG19	10^{-4}	21.70	3.923×10^{4}	41.86	3.923×0^{4}	48.48	3.923×10^{4}
	10^{-5}	51.00		50.38		44.18	
Xception	10^{-4}	54.26	9.67×10^{3}	55.03	9.67×10^{3}	60.48	9.669×10^{3}
	10^{-5}	55.03		52.71		55.03	
DenseNet201	10^{-4}	57.36	9.138×10^{3}	49.61	9.139×10^{3}	57.36	9.138×10^{3}
	10^{-5}	44.86		55.03		58.91	
ResNet50	10^{-4}	27.13	8.287×10^{3}	34.88	8.287×10^{3}	32.55	8.286×10^{3}
	10^{-5}	41.86		41.08		42.63	
InceptionV3	10^{-4}	44.96	5.967×10^{3}	53.48	5.967×10^{3}	49.61	5.966×10^{3}
	10^{-5}	55.03		51.93		53.48	

Figure 4 shows the convergence curve for Xception + RPN + CBAM. The learning curves compare the same Xception + RPN + CBAM architecture under two learning rates: 10^{-4} as in Figure 4(a) and 10^{-5} as in Figure 4(b). With lr 10^{-4}, the model converges rapidly and achieves higher validation accuracy (best around epoch ~38, ≈0.60), indicating better discriminative performance on the validation set. However, the validation loss reaches its minimum earlier (around epoch ~14) and then increases while training loss continues to decrease, suggesting overfitting and a growing generalization gap (high training accuracy vs. much lower validation accuracy).



(a)



(b)

Figure 4. Convergence analysis of Xception + RPN + CBAM (a) learning rate 10^{-4} and (b) learning rate 10^{-5}

In contrast, lr = 10-5 (bottom) yields a smoother and more stable optimization, with validation loss decreasing consistently up to the final epochs (best ≈50), but the validation accuracy remains lower (best around epoch ~45, ≈0.45), indicating underfitting or insufficient learning at this step size. Overall, $lr = 1e-4$ provides better validation accuracy but requires careful regularization/early stopping, whereas $lr = 1e-5$ improves stability but may require longer training or a learning-rate schedule to reach comparable validation performance. Figure 5 compares the confusion matrices of the Xception backbone under three configurations: the baseline model without RPN (Figure 5(a)), the model with RPN only (Figure 5(b)), and the proposed ARPN with CBAM (Figure 5(c)).

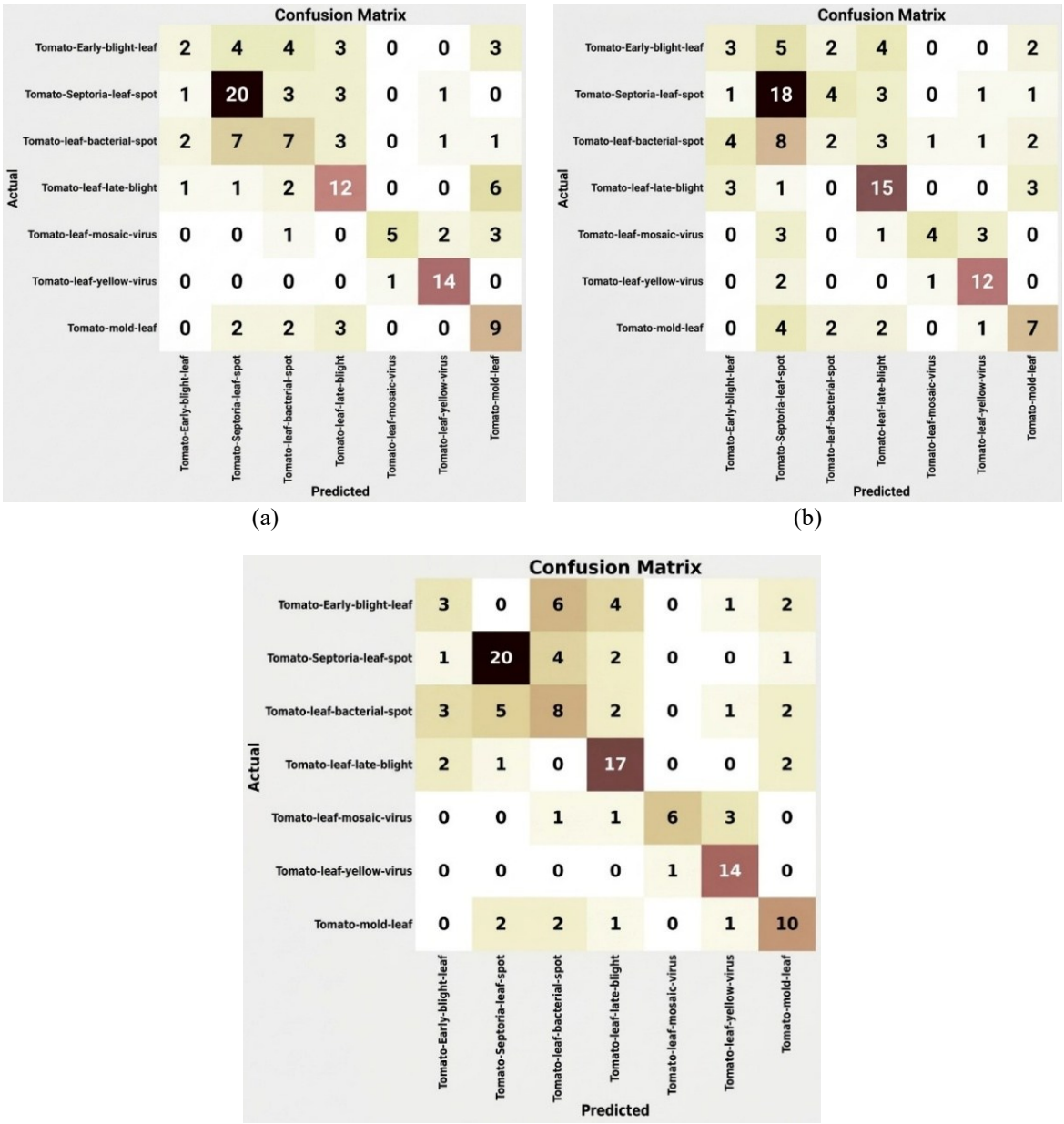


Figure 5. Confusion matrices of Xception as CNN backbones: (a) without RPN, (b) with RPN only, and (c) with ARPN (CBAM as attention mechanism)

The proposed ARPN improves the classification performance across most disease classes, although misclassifications remain among visually similar diseases. CBAM achieved the best performance, suggesting that its combined channel and spatial attention is more effective at suppressing background interference while

emphasizing disease-relevant features. It is important to emphasize that further research is still needed in this area. Although the proposed method improves performance by more than 10%, the current best accuracy remains relatively low. This limitation may also be attributed to the limited availability of training data. Incorporating and combining PlantVillage data may benefit future studies. However, due to different data acquisition conditions, various explorations on signal processing and modeling adaptation must be performed to find the best strategies for improving the model robustness.

4. CONCLUSION

This study proposes a framework to improve the robustness of plant disease detection using ARPNs. Although attention mechanisms are not originally designed to enhance model robustness, their ability to focus on specific image regions that provide the most informative features for disease identification can be utilized to reduce the influence of background images. The experimental results confirm this capability. Among the evaluated models, BAM and CBAM are found to be the most effective attention mechanisms for this task. Overall, the robustness of the models is improved.

FUNDING INFORMATION

The paper is fully funded by Rumah Program Artificial Intelligence, Big Data, dan Teknologi Komputasi Untuk Biodiversitas Dan Citra Satelit from Organisasi Riset Elektronika dan Informatika, National Research and Innovation Agency (BRIN), 2023, Number: 1/III.6/HK/2023.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Laila Mukhrina Azizi	✓	✓	✓		✓	✓		✓	✓	✓	✓			
Ade Ramdan	✓	✓				✓				✓				
Vicky Zilvan			✓	✓	✓	✓				✓				
Evandri			✓	✓	✓	✓		✓		✓				
Hilman Ferdinandus Pardede	✓									✓		✓		
Ana Hadiana			✓			✓				✓	✓			
Dicky Rianto Prajitno			✓			✓				✓	✓			

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The datasets used in this study are publicly accessible and were obtained from open-source repositories in Kaggel, "PlantDoc" at https://www.kaggle.com/datasets/nirmalsankalana/PlantDoc-dataset.

REFERENCES

[1] S. Savary, L. Willocquet, S. J. Pethybridge, P. Esker, N. McRoberts, and A. Nelson, "The global burden of pathogens and pests on major food crops," *Nature Ecology and Evolution*, vol. 3, no. 3, pp. 430–439, 2019, doi: 10.1038/s41559-018-0793-y.
[2] S. Coulibaly, B. K.-Foguem, D. Kamissoko, and D. Traore, "Deep neural networks with transfer learning in millet crop images," *Computers in Industry*, vol. 108, pp. 115–120, 2019, doi: 10.1016/j.compind.2019.02.003.
[3] S. P. Mohanty, D. P. Hughes, and M. Salathé, "Using deep learning for image-based plant disease detection," *Frontiers in Plant Science*, vol. 7, 2016, doi: 10.3389/fpls.2016.01419.
[4] T. A. Salih, A. J. Ali, and M. N. Ahmed, "Deep learning convolution neural network to detect and classify tomato plant leaf diseases," *OALib*, vol. 07, no. 05, pp. 1–12, 2020, doi: 10.4236/oalib.1106296.
[5] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015, doi: 10.1038/nature14539.

[6] N. K. Trivedi *et al.*, "Early detection and classification of tomato leaf disease using high-performance deep neural network," *Sensors*, vol. 21, no. 23, 2021, doi: 10.3390/s21237987.
[7] M. Agarwal, A. Singh, S. Arjaria, A. Sinha, and S. Gupta, "ToLeD: tomato leaf disease detection using convolution neural network," in *Procedia Computer Science*, 2020, vol. 167, pp. 293–301, doi: 10.1016/j.procs.2020.03.225.
[8] G. Sakkarvarthi, G. W. Sathianesan, V. S. Murugan, A. J. Reddy, P. Jayagopal, and M. Elsisi, "Detection and classification of tomato crop disease using convolutional neural network," *Electronics*, vol. 11, no. 21, 2022, doi: 10.3390/electronics11213618.
[9] C. Zhou, S. Zhou, J. Xing, and J. Song, "Tomato leaf disease identification by restructured deep residual dense network," *IEEE Access*, vol. 9, pp. 28822–28831, 2021, doi: 10.1109/ACCESS.2021.3058947.
[10] V. M.-Gutiérrez *et al.*, "Comparison of convolutional neural network architectures for classification of tomato plant diseases," *Applied Sciences*, vol. 10, no. 4, 2020, doi: 10.3390/app10041245.
[11] G. Yang, G. Chen, Y. He, Z. Yan, Y. Guo, and J. Ding, "Self-supervised collaborative multi-network for fine-grained visual categorization of tomato diseases," *IEEE Access*, vol. 8, pp. 211912–211923, 2020, doi: 10.1109/ACCESS.2020.3039345.
[12] D. Singh, N. Jain, P. Jain, P. Kayal, S. Kumawat, and N. Batra, "PlantDoc: a dataset for visual plant disease detection," in *ACM International Conference Proceeding Series*, 2020, pp. 249–253, doi: 10.1145/3371158.3371196.
[13] Q. Wang, F. Qi, M. Sun, J. Qu, and J. Xue, "Identification of tomato disease types and detection of infected areas based on deep convolutional neural networks and object detection techniques," *Computational Intelligence and Neuroscience*, 2019, doi: 10.1155/2019/9142753.
[14] V. Balafas, E. Karantoumanis, M. Louta, and N. Ploskas, "Machine learning and deep learning for plant disease classification and detection," *IEEE Access*, vol. 11, pp. 114352–114377, 2023, doi: 10.1109/ACCESS.2023.3324722.
[15] Y. Zhang, C. Song, and D. Zhang, "Deep learning-based object detection improvement for tomato disease," *IEEE Access*, vol. 8, pp. 56607–56614, 2020, doi: 10.1109/ACCESS.2020.2982456.
[16] M. Nawaz *et al.*, "A robust deep learning approach for tomato plant leaf disease localization and classification," *Scientific Reports*, vol. 12, no. 1, pp. 1–18, 2022, doi: 10.1038/s41598-022-21498-5.
[17] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, vol. 39, no. 6, pp. 1137–1149, doi: 10.1109/TPAMI.2016.2577031.
[18] S. Saleem, M. I. Sharif, M. I. Sharif, M. Z. Sajid, and F. Marinello, "Comparison of deep learning models for multi-crop leaf disease detection with enhanced vegetative feature isolation and definition of a new hybrid architecture," *Agronomy*, vol. 14, no. 10, 2024, doi: 10.3390/agronomy14102230.
[19] S. Saleem, A. Hussain, N. Majeed, Z. Akhtar, and K. Siddique, "A multi-scale feature extraction and fusion deep learning method for classification of wheat diseases," *Journal of Computer Science*, vol. 21, no. 1, pp. 34–42, 2025, doi: 10.3844/jcssp.2025.34.42.
[20] M. A. Naeem *et al.*, "Deep learning-based approach for identification of potato leaf diseases using wrapper feature selection and feature concatenation," *Journal of Computer Science*, vol. 21, no. 5, pp. 1202–1209, 2025, doi: 10.3844/jcssp.2025.1202.1209.
[21] A. Bhujel, N. E. Kim, E. Arulmozhi, J. K. Basak, and H. T. Kim, "A lightweight attention-based convolutional neural networks for tomato leaf disease classification," *Agriculture*, vol. 12, no. 2, 2022, doi: 10.3390/agriculture12020228.
[22] S. Zhao, Y. Peng, J. Liu, and S. Wu, "Tomato leaf disease diagnosis based on improved convolution neural network by attention module," *Agriculture*, vol. 11, no. 7, 2021, doi: 10.3390/agriculture11070651.
[23] T. S. Alam, C. B. Jowthi, and A. Pathak, "Comparing pre-trained models for efficient leaf disease detection: a study on custom CNN," *Journal of Electrical Systems and Information Technology*, vol. 11, no. 1, pp. 1–26, 2024, doi: 10.1186/s43067-024-00137-1.
[24] F. Chollet, "Xception: deep learning with depthwise separable convolutions," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1800–1807, doi: 10.1109/CVPR.2017.195.
[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.
[26] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9, doi: 10.1109/CVPR.2015.7298594.
[27] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2261–2269, 2017, doi: 10.1109/CVPR.2017.243.
[28] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, "MobileNetV2: inverted residuals and linear bottlenecks," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520, doi: 10.1109/CVPR.2018.00474.
[29] M. H. Guo *et al.*, "Attention mechanisms in computer vision: a survey," *Computational Visual Media*, vol. 8, no. 3, pp. 331–368, 2022, doi: 10.1007/s41095-022-0271-y.
[30] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: convolutional block attention module," in *Computer Vision – (ECCV 2018)*, 2018, pp. 3–19, doi: 10.1007/978-3-030-01234-2_1.
[31] J. Park, S. Woo, J. Y. Lee, and I. S. Kweon, "A simple and light-weight attention module for convolutional neural networks," *International Journal of Computer Vision*, vol. 128, no. 4, pp. 783–798, 2020, doi: 10.1007/s11263-019-01283-0.
[32] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141, doi: 10.1109/CVPR.2018.00745.

BIOGRAPHIES OF AUTHORS



Laila Mukhrina Azizi received her bachelor's degree in Statistics from Faculty of Mathematics and Natural Sciences, Syiah Kuala University. Her research interests include forecasting and deep learning, particularly in applications for plant disease detection. She can be contacted at email: laila_mz@mhs.usk.ac.id.



Ade Ramdan is a researcher at the Research Center for Artificial Intelligence and Cyber Security, National Research and Innovation Agency (BRIN), Republic of Indonesia. He obtained his bachelor's degree from Bandung Institute of Technology in 2008 and his master's degree from STMIK-LIKMI Indonesia in 2023. His research interests include artificial intelligence, machine learning, pattern recognition, image processing, data mining, and embedded systems. He can be contacted at email: ader004@brin.go.id.



Vicky Zilvan is a researcher at the Research Center for Artificial Intelligence and Cyber Security, National Research and Innovation Agency (BRIN), Republic of Indonesia. He received his bachelor's degree in Computer Science from the Bogor Agricultural Institute and his master's degree in Informatics from the Bandung Institute of Technology. His areas of interest include artificial intelligence, machine learning, pattern recognition, speech processing, and data mining. He can be contacted at email: vick001@brin.go.id.



Evandri is a researcher at the Research Center for Artificial Intelligence and Cyber Security, National Research and Innovation Agency (BRIN), Republic of Indonesia. He received his bachelor's and master's degrees in Information Science from the University of Fukui, Japan. His areas of interest include artificial intelligence, machine learning, and pattern recognition. He can be contacted at email: evan001@brin.go.id.



Hilman Ferdinandus Pardede received his bachelor's degree in Electrical Engineering from the University of Indonesia. His master's degree is from the University of Western Australia and his doctoral degree is obtained from Tokyo Institute of Technology in Computer Science. He is currently a research professor at the Research Center for Artificial Intelligence and Cyber Security, the National Research and Innovation Agency (BRIN), Republic of Indonesia. His research interests include machine learning, multimedia signal processing and pattern recognition, and artificial intelligence. He can be contacted at email: hilm003@brin.go.id.



Ana Hadiana is a researcher at the Research Center for Information and Data Science, National Research and Innovation Agency, Republic of Indonesia. He obtained his bachelor degree (B.Eng.) from Fukui University, Japan in 1993, his master's degree (M.Eng.) from Fukui University, Japan, in 1995, and his doctorate (Dr.Eng.) from Shinsyu University in 2004. His research interests include enterprise architecture, data mining, user interface, and kansei/affective engineering. He can be contacted at email: anah001@brin.go.id.



Dicky Rianto Prajitno is a researcher at the Research Center for Artificial Intelligence and Cyber Security, National Research and Innovation Agency (BRIN), Republic of Indonesia. He received his bachelor's degree in Industrial Control Systems and his master's degree in Multimedia and Game Technology from the Bandung Institute of Technology. His areas of interest include computer vision, pattern recognition, artificial intelligence, and machine learning. He can be contacted at email: dick003@brin.go.id.