## Research Paper

# Explainable Multi-Scale Attention-Enhanced Vision Transformer for Robust Detection of Pest-Related Food Safety Hazards

**Prerana Nilesh Khairnar[1*], Bharathi Alagiri[2], Karthikeyan Kasi[3], Tupili Sangeetha[4], Shalini Ramaraju[5], Vivek Marotrao Korde[6]**

1. *Computer Engineering, Sir Visvevaraya Institute of Technology, Chincholi, Nashik, Maharashtra, India.*
2. *Business Architecture Team Lead, Department of Software Engineering, Accenture, No. 40, Chakrapani Street, Mayor Sa Ganesan Kudil, Chennai, Tamil Nadu, India.*
3. *Solution Consultant, Software Engineering, Tata Consultancy Services Pvt Ltd, 12/6 Model Hutment Road Extn, CIT Nagar, Chennai, Tamil Nadu, India.*
4. *Computer Science and Engineering, R.M.D Engineering College, R.S.M Nagar, Kavaraipettai, Tamil Nadu, India.*
5. *Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vadeswaram, AP, India.*
6. *Mechanical Engineering, Yeshwantrao Chavan College of Engineering (YCCE), Nagpur, Hingna Road, Wanadongri, Nagpur, Maharashtra, India.*

**Citation** Khaimar PN, Alagiri B, Kasi K, Sangeetha T, Ramaraju Sh, Korde VM. Explainable Multi-Scale Attention-Enhanced Vision Transformer for Robust Detection of Pest-Related Food Safety Hazards. International Journal of Medical Toxicology and Forensic Medicine. 2026; 16:E53178.

https://doi.org/10.22037/ijmtfm.v16.53178

**Article info:**

**Received:** 14 August, 2026

**Accepted:** 04 September, 2026

**Published:** 23 September, 2026

## A B S T R A C T

**Background:** Agricultural diseases and insect pests pose a major threat to crop production. Initially, identifying and detecting pests could effectively reduce the economic impacts they cause. Some agricultural experts are concerned with deep learning techniques, as they have exhibited critical accuracy in image classification.

**Methods:** This study introduces a novel Multi-Scale Attention Enhanced Transformer Network for Pest Detection in Precision Agriculture (MSAETN-PDPA). The primary objective of the proposed framework is to effectively identify and classify categories of agricultural pests in agricultural field images. The proposed model employs a multi-scale attention-enhanced multi-axis vision transformer network to extract robust features from the input image. GradCAM++ is utilized to interpret the decision-making process of the deep learning technique employed for agricultural pest classification. Extensive simulation analysis is performed to ensure the improved proficiency of the MSAETN-PDPA approach.

**Results:** The comparative evaluation showed that the MSAETN-PDPA technique achieved 98.73% accuracy, outperforming existing methods.

**Conclusion:** The proposed model is useful for identifying food-related safety hazards.

**Keywords:**

Agricultural pests, Image analysis, Deep learning, Multi-Axis vision transformer, Hyperparameter optimization

*** Corresponding Author:**

***Prerana Nilesh Khairnar, PhD***

***Computer Engineering, Sir Visvevaraya Institute of Technology, Chincholi, Nashik, Maharashtra, India.***

***E-mail:*** *prerananileshkhairnar@gmail.com*

## Introduction

Agriculture acts a significant part in assuring food safety and the sustainable growth of society. Moreover, crop diseases and pests have a substantial effect on crop quality and yields, compromising the productivity and effectiveness of agricultural fields [1]. Pest and disease management is the most significant work of futuristic agriculture. Also, Agriculture is one of the world's critical economic sectors, employing 1.23 billion people in off-farm and on-farm activities [2]. Moreover, the growing demand for more and higher-quality crops has led to increased resource consumption, which could harm the environment. Applying Smart Agricultural practices is vital to overcoming these issues in agricultural production, including those related to food security, productivity, sustainability, and environmental impacts. Initially, detecting pest insects is a primary objective to confirm healthy crop development; this reduces the risk of yield loss and enables the application of precision treatments, which offer economic and environmental benefits. Now, this operation is primarily carried out by specialists, who must monitor traps in the field to assess the presence of species that pose considerable financial losses. This process is prone to detection delays and requires greater human intervention.

To overcome these challenges, this study presents a Multi-Scale Attention Enhanced Transformer Network for Pest Detection in Precision Agriculture (MSAETN-PDPA). The significant contribution of this research is as follows:

- Introduce an MSA-MaxNet approach that combines a multi-axis vision transformer (MaxViT) with multi-scale attention (convolutional block attention module (CBAM) and multi-scale convolutional block attention module (MCBAM)), which extracts fine-grained local and global features from agricultural field images.

- Utilize a graph convolutional network (GCN) for pest classification.

- Integrate the AdaBelief optimizer for parameter optimization.

- Incorporate explainable artificial intelligence (XAI) using GradCAM++ that offers visual interpretations of model decisions.

- Conduct simulations and experimentation on the benchmark Agricultural Pests Image Dataset.

### Existing Approaches for Agricultural Pest Detection

This section provides an overview of current progress in agricultural pest detection, highlighting associated research gaps. Venkateswara and Padmanabhan [3] presented a novel automated method that integrates DL into smart farming for nuisance observation to address this problem. Amrani et al. [4] proposed an AI-based framework for identifying beetles in crop images acquired using consumer-grade RGB cameras. Li et al. [5] introduced a comprehensive integrated pest management solution that uses DL for semi-automatic pest identification and an efficient device for pest-control decision-making. Zhang et al. [6] proposed a framework for standard ML models for disease and pest detection and their limitations, and focused on developing DL techniques, covering three mainstream frameworks: Vision Transformer (ViT) algorithms, CNNs, and CNN–Transformer hybrid systems.

Abinaya and Daniel [7] introduced a new technique for pest identification and regulation, enabling detailed soil examination and the avoidance of soil-borne afflictions in sugarcane regions through a combined model. Yin et al. [8] presented an innovative crop pest and disease identification system that depends on a fuzzy neural network (FNN) and multi-stage feature fusion in remote sensing images. Butt [9] determined the structure and execution capability of DL methods for an automatic model for detecting plant diseases employing leaf images. Baeza-Moreno et al. [10] proposed an innovative computer vision (CV) approach for identifying insect pests on tree and plant leaves in real-world settings, integrating DL with standard image processing methods.

Even though recent advances in pest detection have achieved effective performance, various challenges remain unresolved for dynamic deployment in real-world agricultural systems. To address these challenges, this study presents MSAETN-PDPA, a multi-scale attention-improved transformer method incorporated with GCN classification for effective pest detection. The model uses the AdaBelief optimizer and GradCAM++ for interpretation to enhance classifier performance and improve model explainability.

## Materials and Methods

### Proposed MSAETN-PDPA Framework

The proposed MSAETN-PDPA model comprises a systematic DL pipeline developed for precise

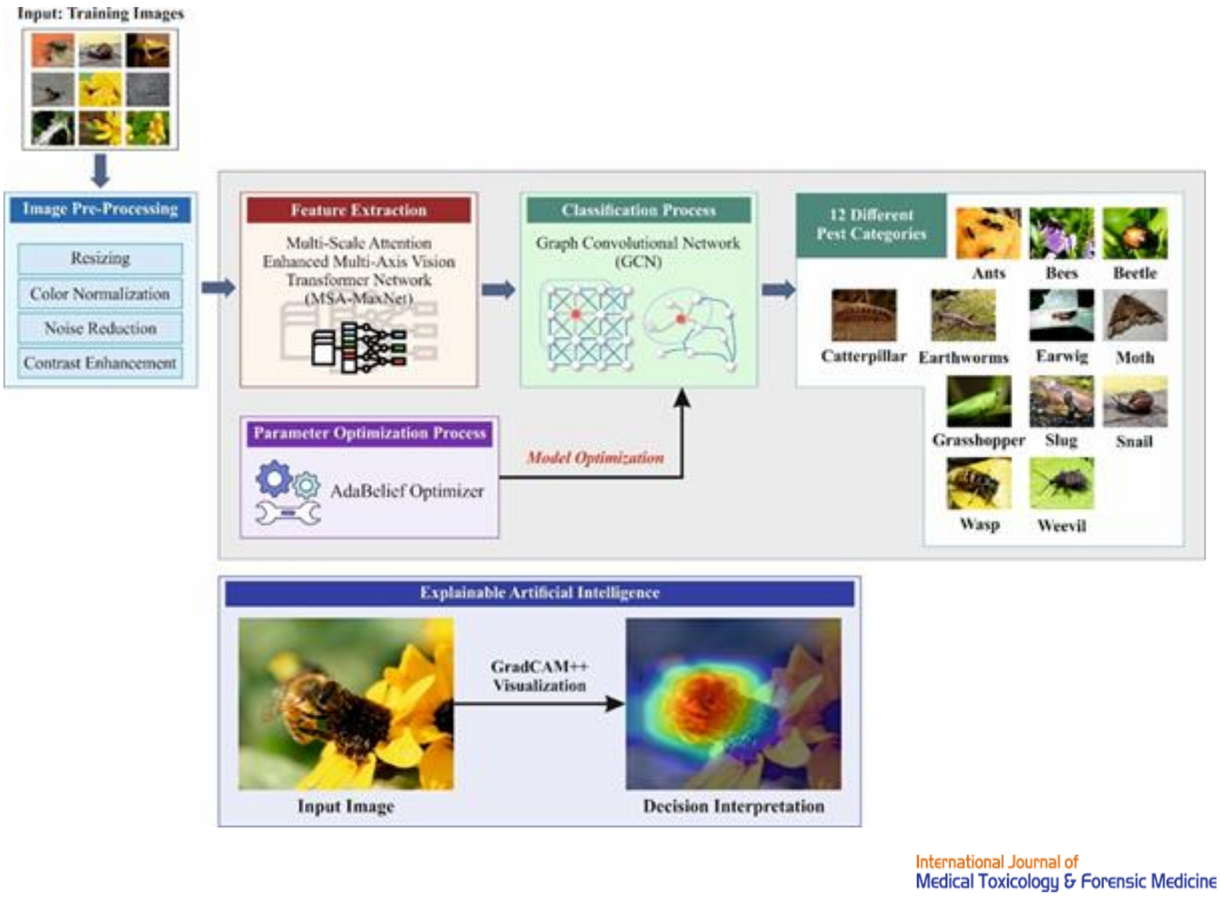


**Figure 1.** Workflow of the presented MSAETN-PDPA framework.

agricultural pest identification and classification. Figure 1 demonstrates the workflow of the presented MSAETN-PDPA model. As shown in Figure 1, the MSAETN-PDPA model initially undergoes preprocessing, which includes image resizing, color normalization, noise reduction using non-local means (NLM) filtering, and contrast enhancement with CLAHE. Feature extraction is performed using an MSA-MaxNet. The captured features are passed into the GCN. To further enhance classifier performance, the AdaBelief optimizer is integrated. Finally, GradCAM++ is used for model interpretation, highlighting the region that contributes most to the model's predictions.

### *Image Preprocessing Pipeline*

Image preprocessing is crucial, as it enhances image quality and prepares images for analysis. In this study, resizing, color normalization, noise reduction, and contrast enhancement are used.

### *Image resizing*

Primarily, all images are resized to 224x224 pixels because they have varying dimensions. This normalization is required due to differences in pixel sizes across the images involved and incorporates a wide range of pest image samples collected from various real-world sources, such as Flickr.

### *Color Normalization*

Consider an input RGB image represented as given below,

$$I(u,v) = [R(u,v), G(u,v), B(u,v)] \quad (1)$$

In this, $R(u,v)$, $G(u,v)$, and $B(u,v)$ stand for the red, green, and blue color levels for every pixel $(u,v)$. First, the intensity of all the channels is normalized. This process is crucial for reducing differences and making agricultural pest images reliable, enabling better processing.

### *Noise Reduction*

The NLM filtering method assigns a weight to each pixel in the test image based on the similarity of its neighboring pixels. Conversely, the algorithm prioritizes similar pixel neighbors by allocating them greater weight. This similarity-based weighting module operates exclusively within a defined search window surrounding the target pixel.

$$I_{den}(u) = \sum_{v\in\Omega} w(u,v)I(v), \sum_{v\in\Omega} w(u,v) = 1 \quad (2)$$

### *Contrast Enhancement*

The image contrast is improved through CLAHE, successfully mitigating the challenges of noise amplification; a predominant limitation is observed in standard histogram equalization (HE) techniques. The intensification in CLAHE is controlled by applying a clipping function to HE at user-defined values called clip limits. CLAHE was configured with a clip limit of 2.0 and an 8 × 8 tile grid, while Non-Local Means filtering was performed using a 7 × 7 search window and 3 × 3 comparison window with a filtering strength of h = 10.

### Feature Learning via MSA-MaxNet

Following image preprocessing, the improved pest images are passed through the proposed MSA-MaxNet-based feature-learning model, yielding discriminative spatial and semantic representations for accurate pest classification. The MSA-MaxNet model is used to extract salient features and enable accurate classification of agricultural pests. The architecture of MSA-MaxNet consists of MCBAM, CBAM, and an encoder [11]. This system starts with a stem layer that separates the input image into isolated patches. The encoding includes a sequence of MBConv layers monitored by MaxViT modules, which are recurrent in every phase to gradually acquire and encode features from the input imagery. Lastly, the MaxViT module is employed to generate discriminative deep feature representations for downstream classification tasks.

MaxViT Block: The MaxViT block contains two primary consecutive components: grid attention and block attention. Each mechanism separates the features into sections. This method enhances local features and incorporates global context, thereby improving feature extraction. Afterward, the sections are combined and returned to their original form.

### Graph Convolutional Pest Classification Network

The feature embeddings extracted from the MSA-MaxNet module are provided as input to the Graph Convolutional Pest Classification Network. During this phase, the extracted image features are treated as graph nodes, and their relationships are modeled using graph convolution operations to classify agricultural pests into distinct classes. GCNs are DL methods designed to process non-Euclidean data, such as graph-structured inputs. They are widely used in supervised learning methods to extract useful representations by leveraging node features and graph structure across multiple layers of graph convolutions [12].

Node feature construction: To ensure efficient learning across visual feature representations, the node feature structure uses deep feature embeddings that can be removed from the network architecture.

The performance of the GCN classification network relies on efficient optimization of its learnable parameters. Hence, the AdaBelief optimizer is implemented to optimize the network parameters and improve classification performance and convergence stability.

## Results

In this section, the performance evaluations of the MSAETN-PDPA approach are provided in detail on the agricultural pest image dataset. The tests demonstrate the feasibility and applicability of the MSAETN-PDPA methodology for real-time classification. The Agricultural Pests Image Dataset [13, 14] has been utilized to evaluate the performance of the MSAETN-PDPA technique. This dataset is a collection of images of 12 diverse types of agricultural pests, including Ants, Bees, Beetles, Caterpillars, Earthworms, Earwigs, Grasshoppers, Moths, Slugs, Snails, Wasps, and Weevils. The images can be obtained on Flickr using the API and resized to have a maximum width or height of 300px. It contains 12 pest classes with diverse variations in shape, color, and size, enabling robust model training and evaluation under different conditions. The dataset contains 5494 samples under 12 classes, as shown in Table 1.

**Table 1.** Dataset details.

| **“Pests”** | **“Images”** |
|---|---|
| “Ants” | “499” |
| “Bees” | “500” |
| “Beetle” | “416” |
| “Catterpillar” | “434” |
| “Earthworms” | “323” |
| “Earwig” | “466” |
| “Grasshopper” | “485” |
| “Moth” | “497” |
| “Slug” | “391” |
| “Snail” | “500” |
| “Wasp” | “498” |
| “Weevil” | “485” |
| **“Total”** | **“5494”** |

International Journal of Medical Toxicology & Forensic Medicine

The comparison results and computational analysis (CT) of the MSAETN-PDPA methodology against

**Table 2.** Comparison of outcomes of the MSAETN-PDPA method with other models.

| Methodologies | Accuracy | Loss Rate (%) | Precision | Recall | F-measure | Computational Time (sec) |
|---|---|---|---|---|---|---|
| Greenhouse pest | 91.00 | 9.00 | 89.04 | 86.92 | 85.99 | 3.12 |
| DenseNet201 | 92.43 | 7.57 | 90.51 | 86.51 | 86.51 | 1.95 |
| GAEnsemble | 95.16 | 4.84 | 86.62 | 88.27 | 88.01 | 4.74 |
| AlexNet | 81.68 | 18.32 | 86.22 | 90.25 | 89.81 | 3.70 |
| EfficientNet | 94.99 | 5.01 | 85.99 | 86.03 | 89.01 | 1.93 |
| ViT b 16 | 96.13 | 3.87 | 87.34 | 89.75 | 90.14 | 4.04 |
| ShuffleNet-V2 | 83.73 | 16.27 | 86.87 | 86.30 | 85.10 | 1.75 |
| MSAETN-PDPA | 98.73 | 1.27 | 92.12 | 92.12 | 92.10 | 0.34 |



existing approaches are presented in Table 2 [14-16]. The outcome implies that the MSAETN-PDPA method obtains better performance. Based on CT, the proposed MSAETN-PDPA model records minimal CT of 0.34sec, whereas the Greenhouse pest, DenseNet201, GAEnsemble, AlexNet, EfficientNet, ViT-b16, and ShuffleNet-V2 methodologies attain maximal CT. The MSAETN-PDPA approach has achieved superior results: an accuracy of 98.73%, a precision of 92.12%, a recall of 92.12%, and an F-measure of 92.10%.

## Discussion

Figure 2 shows the visual interpretation of GradCAM++ with other XAI techniques under various classes. The first column displays the real insect images from various classes, whereas the remaining columns show the corresponding heatmaps. Warmer colors (red/yellow) imply areas that contribute most to the prediction of the model. Across classes, each approach usually concentrated on the major body and distinctive characteristics of the insects, confirming meaningful feature learning. Grad-CAM provides coarse localization, whereas Score-CAM produces smoother, more object-aligned attention maps. Grad-CAM++ further refines localization, emphasizing finer discrimination details.

Table 3 represents the ablation study of the MSAETN-PDPA model under various measures. The performance indicated that the Multi-Scale Attention (MSA) + GCN, Multi-Axis Vision Transformer (MAViT) + GCN, MSA + MAViT + GCN, Multi-Scale Attention-Enhanced MAViT + GCN (Without Preprocessing), and Preprocessing + MAViT + GCN (Without Multi-Scale Attention) methodologies achieve least outcome with different metrics, where the MSAETN-PDPA (Preprocessing + Multi-Scale Attention-Enhanced MAViT + GCN) model accomplish higher performance with accuracy of 98.73%, precision of 92.12%, recall of 92.12%, and F-measure of 92.10%.

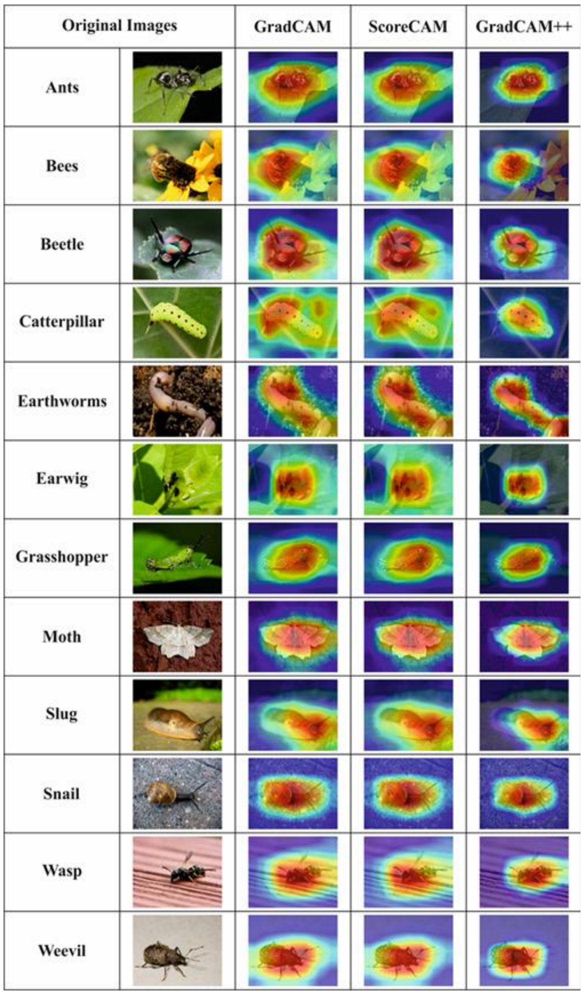




**Figure 2.** Comparative visual interpretation results of XAI models.

## Conclusion

In this paper, the MSAETN-PDPA methodology is introduced for the efficient detection and classification of 12 diverse classes of agricultural pests, including Ants, Moths, Bees, Beetles, Caterpillars, Earthworms, Snails, Earwigs, Wasps, Grasshoppers, Slugs, and Weevils. To achieve that, the proposed model utilizes MSA-MaxNet for effective feature representation.

**Table 3.** Ablation study of the AIDSA-PELTT method.

| Methodologies | Accuracy | Precision | Recall | F-measure |
|---|---|---|---|---|
| Multi-Scale Attention (MSA) + GCN | 95.49 | 88.72 | 88.65 | 88.65 |
| Multi-Axis Vision Transformer (MAViT) + GCN | 96.01 | 89.47 | 89.22 | 89.43 |
| MSA + MAViT + GCN | 96.56 | 90.06 | 89.94 | 90.06 |
| Multi-Scale Attention-Enhanced MAViT + GCN (Without Preprocessing) | 97.36 | 90.83 | 90.72 | 90.85 |
| Preprocessing + MAViT + GCN (Without Multi-Scale Attention) | 98.00 | 91.46 | 91.52 | 91.36 |
| MSAETN-PDPA (Preprocessing + Multi-Scale Attention-Enhanced MAViT + GCN) | 98.73 | 92.12 | 92.12 | 92.10 |

International Journal of
Medical Toxicology & Forensic Medicine

The extracted high-level features are then provided to the GCN for final classification. The comparative analysis showed that the MSAETN-PDPA system achieved 98.73% accuracy, outperforming existing methodologies. Although the MSAETN-PDPA method achieves high accuracy in pest detection, it can encounter challenges in real-world agricultural settings due to background noise, occlusions, and lighting fluctuations. Its higher computational cost also poses real-time utilization challenges on lower-resource devices. Future work can concentrate on emerging lightweight architectures for real-time deployments with UAVs and edge devices. Employing large datasets can improve generalization across diverse agricultural scenarios. Model compression methods, namely pruning and knowledge distillation, can decrease complexity. Also, adding temporal details from drone videos should improve detection performance in dynamic settings.

## Ethical Approval

Not applicable.

## Acknowledgment

None.

## Funding

No external funding was received for this research.

## Conflicts of Interest

The authors report there are no competing interests to declare.

## References

[1] KS G, P SK, A R, K G. Crop pest identification using deep network-based extracted features and MobileENet in smart agriculture. Land Degrad Dev. 2024;35(11):3642-52. [DOI: 10.1002/ldr.5157]

[2] Amulothu DVRT, Rodge RR, Hasan W, Gupta S. Machine learning for pest and disease detection in crops. In: Agriculture 4.0. CRC Press; 2024. p. 111-32. [DOI: 10.1201/9781003570219-6]

[3] Venkateswara SM, Padmanabhan J. Deep learning-based agricultural pest monitoring and classification. Sci Rep. 2025;15(1):8684. [DOI: 10.1038/s41598-025-92659-5]

[4] Amrani A, Diepeveen D, Murray D, Jones MG, Sohel F. Multi-task learning model for agricultural pest detection from crop-plant imagery: a Bayesian approach. Comput Electron Agric. 2024;218:108719. [DOI: 10.1016/j.compag.2024.108719]

[5] Li S, Yuan Z, Peng R, Leybourne D, Xue Q, Li Y, et al. An effective farmer-centered mobile intelligence solution using lightweight deep learning for integrated wheat pest management. J Ind Inf Integr. 2024;42:100705. [DOI: 10.1016/j.jii.2024.100705]

[6] Zhang M, Liu C, Li Z, Yin B. From convolutional networks to vision transformers: evolution of deep learning in agricultural pest and disease identification. Agronomy. 2025;15(5):1079. [DOI: 10.3390/agronomy15051079]

[7] Abinaya M, Daniel JA. Deploying YOLOv8m and deep Q-network for intelligent prediction of pests and soil-borne illness in sugarcane fields. Int J Comput Intell Syst. 2026. [DOI: 10.1007/s44196-025-01050-9]

[8] Yin S, Li H, Ivanović M, Chen T, Teng L. FNNMFF: crop pests and diseases detection based on fuzzy neural network and multilevel feature fusion in remote sensing images. Comput Sci Inf Syst. 2026;(00):12-12. [DOI: 10.2298/csis251109012y]

[9] Butt M. Deep learning for plant disease detection using image analysis [master's thesis]. University of Eastern Finland; 2026. [Link]

[10] Baeza-Moreno L, Blanco-Carmona P, Hidalgo-Fort E, Martín-Clemente R, González-Carvajal R. AI-driven detection of tiny pests in foliage: integrating image processing and deep learning. Mach Learn Appl. 2025:100834. [DOI: 10.1016/j.mlwa.2025.100834]

[11] Wu W, Huang J, Zhang M, Li Y, Yu Q, Zhao Q. MSA-MaxNet: multi-scale attention enhanced multi-axis vision transformer network for medical image segmentation. J Cell Mol Med. 2024;28(24):e70315. [DOI: 10.1111/jcmm.70315]

[12] Saha N, Dutta A. WCM-GCN: identifying influential spreaders using graph convolutional networks on weighted directed networks based on spreading properties. Knowl Inf Syst. 2026;68(1):33. [DOI: 10.1007/s10115-025-02675-0]

[13] Agricultural pests image dataset [dataset on the Internet]. Kaggle; [date unknown] [cited 2026 Sep 21]. [Link]

[14] de Oliveira Costa PL, de Oliveira Costa TM, Moreira LFR, Silva LHFP, Mari JF. Classification of agricultural pests through digital images using deep learning. Rev Inform Teor Apl. 2025;32(1):18-25. [DOI: 10.22456/2175-2745.143520]

[15] Cen X, Lu S, Qian T. YOLO-LCE: a lightweight YOLOv8 model for agricultural pest detection. Agronomy. 2025;15(9):2022. [DOI: 10.3390/agronomy15092022]

[16] Peng H, Xu H, Shen G, Liu H, Guan X, Li M. A lightweight crop pest classification method based on improved MobileNet-V2 model. Agronomy. 2024;14(6):1334. [DOI: 10.3390/agronomy14061334]