International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 05 Issue: 03 | March – 2026 DOI:10.55041/ISJEM05543
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Vision Transformer Based Automated Detection of Harmful Farm Insects

Ketan Kanjiya[1], Piyush Sonani[2], Upendrasinh Zala[3]

[1]*Chief Research Officer, Kshatrainfotech, Ahmedabad, Gujarat, India*
[2]*Chief Technology Officer, Kshatrainfotech, Ahmedabad, Gujarat, India*
[3]*Chief Executive Officer, Kshatrainfotech, Ahmedabad, Gujarat, India*

Abstract - Early and accurate detection of harmful farm insects is essential for ensuring agricultural productivity, food security, and sustainable farming practices. This study proposes a Vision Transformer based deep learning framework for automated multi class insect detection in agricultural environments. The model leverages global self-attention mechanisms and transfer learning from large scale pretraining to capture both local visual features and long range contextual relationships, enabling robust fine grained insect classification. A balanced learning strategy incorporating random oversampling and data augmentation is employed to address class imbalance and improve model generalization. The framework is evaluated on a multi-class insect image dataset containing 15 agriculturally significant species under realistic field conditions. Experimental results demonstrate stable convergence, strong generalization, and reliable classification performance across diverse insect categories. The proposed system provides a scalable and intelligent solution for precision agriculture, supporting early pest identification, targeted intervention, and data driven crop protection strategies. This work highlights the potential of transformer based architectures for advancing automated pest monitoring and sustainable agricultural management.

Key Words: Vision Transformer, Agricultural Pest Detection, Farm Insect Classification, Deep Learning, Computer Vision, Smart Farming

1. INTRODUCTION

Agriculture is a fundamental pillar of the global economy and a critical source of food and raw materials for the world's growing population. However, agricultural productivity is persistently threatened by harmful insect pests, which are responsible for an estimated 20-40% of global crop yield losses annually. According to the United Nations, the global population is projected to reach approximately 9.7 billion by 2050 [1], ensuring food security will require a significant increase in agricultural productivity. This challenge is further intensified by climate change and intensive farming practices, which contribute to more frequent and severe pest outbreaks. Beyond direct crop destruction, many insect species act as vectors for fungal, bacterial, and viral diseases, further degrading crop quality and threatening food safety.

Conventional pest monitoring methods primarily rely on manual field inspections and expert identification. These approaches are labor intensive, subjective, and reactive, often detecting infestations only after significant crop damage has occurred. Such delays typically result in indiscriminate pesticide application, which leads to environmental degradation, insecticide resistance, and serious human health concerns. Modern Integrated Pest Management strategies emphasize the importance of early, accurate, and targeted pest detection to enable sustainable and precision driven agricultural interventions.

Recent advances in Smart Pest Monitoring integrate the Internet of Things, unmanned aerial vehicles, and artificial intelligence to enable automated agricultural surveillance. Early computer vision approaches employed traditional machine learning techniques, such as Support Vector Machines and Random Forests, using handcrafted visual features. Although these methods improved automation, they lacked robustness in complex real-world environments characterized by variable lighting, cluttered backgrounds, and occlusions.

The emergence of deep learning, particularly Convolutional Neural Networks, significantly advanced pest recognition by enabling hierarchical feature learning directly from raw images. Despite their success, CNNs primarily model localized spatial patterns and often struggle to capture long range dependencies and global contextual relationships critical for distinguishing visually similar insect species in natural settings.

To address these limitations, Vision Transformers have introduced self-attention mechanisms that enable holistic modeling of global image context. By processing images as tokenized sequences, ViTs provide enhanced global

perception and contextual reasoning. This paper proposes a Vision Transformer based automated detection framework for harmful farm insects, designed to leverage global attention modeling and balanced learning strategies to deliver a robust, scalable, and intelligent pest detection system that supports precision agriculture, sustainable farming, and data driven pest management.

2. LITERATURE REVIEW

The evolution of automated insect detection has transitioned from traditional manual observation and handcrafted feature extraction to sophisticated deep learning architectures. Historically, pest monitoring relied heavily on human expertise and manual scouting, a process that is both subjective and labor intensive [2, 3]. Early technological interventions focused on traditional machine learning models such as Support Vector Machines, k-Nearest Neighbors, and Artificial Neural Networks [4-6]. These systems primarily utilized handcrafted descriptors based on an insect's color, shape, and texture to identify species [7]. However, these traditional methods often lacked robustness in unrestricted outdoor environments, as they were highly sensitive to varying lighting conditions, cluttered backgrounds, and natural occlusions [8].

The introduction of Convolutional Neural Networks marked a significant paradigm shift in the field. Unlike traditional ML, CNNs autonomously learn hierarchical feature representations directly from raw image pixels [9]. Researchers have extensively applied architectures such as VGG-16, ResNet, and InceptionV3 to achieve high classification accuracies in controlled settings [7, 10]. Despite their success, these models often require significant computational resources and may struggle with long range spatial correlations, as they primarily focus on localized feature extraction through convolution kernels [11].

To address real-time monitoring needs, object detection frameworks like the You Only Look Once series have become the industry standard. Version iterations from YOLOv3 to YOLOv8 have demonstrated exceptional performance in locating and identifying pests across various crops [12, 13]. Studies have utilized YOLOv5 for detecting insects in yellow sticky traps [14], while fine tuned YOLOv8 models have been optimized for rice and maize pests in field conditions [15]. These one-stage detectors are particularly valued for their speed and efficiency, making them suitable for deployment on edge devices and unmanned aerial vehicles [16, 17].

Recent advancements have integrated these AI models with the Internet of Things and remote sensing technologies. Drones and smart traps equipped with high resolution cameras now allow for autonomous data collection and real-time alerts through cloud platforms [18]. These systems often utilize cascaded classifiers to first detect an object and then filter it as an insect versus a non insect before final classification [19]. Furthermore, data augmentation techniques including flipping, rotation, and Mosaic augmentation have become essential for training robust models on relatively small datasets, preventing overfitting and improving generalization [10].

The recent evolution of computer vision in agriculture has seen growing interest in Vision Transformers as an alternative to convolution based architectures. ViTs replace traditional convolutional operations with self-attention mechanisms, enabling images to be processed as sequences of patches and allowing global contextual relationships to be modeled more effectively [20]. While CNNs remain effective for learning fine-grained local features, ViTs demonstrate a stronger capacity for capturing long range spatial dependencies and holistic scene representations [20]. Building on this paradigm shift, this study proposes a Vision Transformer based framework for automated harmful farm insect detection, aiming to enhance robustness, generalization, and scalability in real-world precision agriculture applications.

3. DATASET

This study utilizes the Dangerous Farm Insects Image Dataset, a publicly available image collection curated to support research in agricultural pest detection and computer vision. The dataset contains high quality images of 15 distinct insect species commonly found in agricultural environments, each represented by multiple samples that capture variations in appearance, color, texture, and morphological patterns. These images reflect realistic field conditions, including diverse backgrounds and visual complexity, making the dataset suitable for robust model training and evaluation.

The dataset is designed to support automated pest identification, early infestation detection, and intelligent crop protection systems. Its diversity enables the development and benchmarking of deep learning models for multi-class insect classification and precision agriculture applications. In this study, the dataset is used to train and evaluate a Vision Transformer based framework for harmful farm insect detection, supporting scalable, data driven pest monitoring and sustainable agricultural management. Figure 1 presents representative samples from the dataset.

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 05 Issue: 03 | March – 2026 DOI:10.55041/ISJEM05543
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Figure 1: Representative samples from the Dangerous Farm Insects Image Dataset

4. METHODOLOGY

This study proposes a Vision Transformer (ViT) based deep learning framework for automated harmful farm insect detection using a supervised multi-class image classification approach. The complete pipeline consists of dataset acquisition, preprocessing and balancing, data transformation, model fine-tuning, and performance evaluation.

4.1 DATA PREPROCESSING

Data preprocessing was performed to ensure balanced learning, robust feature extraction, and reliable model evaluation. To address class imbalance, random oversampling was applied using the RandomOverSampler technique, synthetically increasing minority class samples to achieve uniform class distribution across all 15 insect categories. This step reduced model bias toward dominant classes and improved training stability. The impact of the oversampling strategy on class balance is visualized in Figure 2, which presents the class-wise image distribution before and after oversampling.

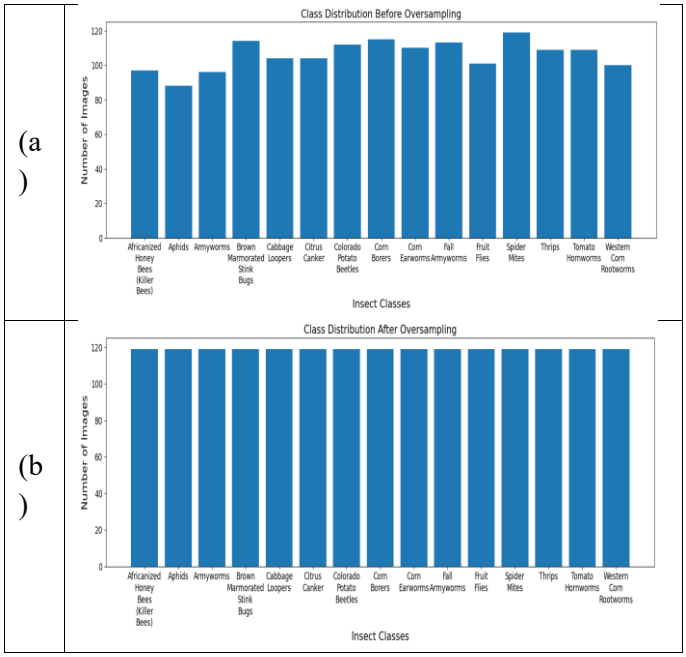


Figure 2: Class-wise distribution of insect images: (a) before oversampling, showing significant class imbalance across species; (b) after random oversampling, demonstrating balanced representation across all insect categories.

All images were resized to 224 × 224 pixels to match the input requirements of the Vision Transformer architecture. Data augmentation strategies, including random rotation and random sharpness adjustment, were applied to enhance model robustness and generalization. Images were further normalized using model specific mean and standard deviation values and converted into tensor format for deep learning compatibility.

Class labels were encoded into numerical indices using a consistent label mapping scheme. The dataset was then divided using a 60:40 stratified split to preserve class distribution across training and testing sets, ensuring fair and unbiased performance evaluation.

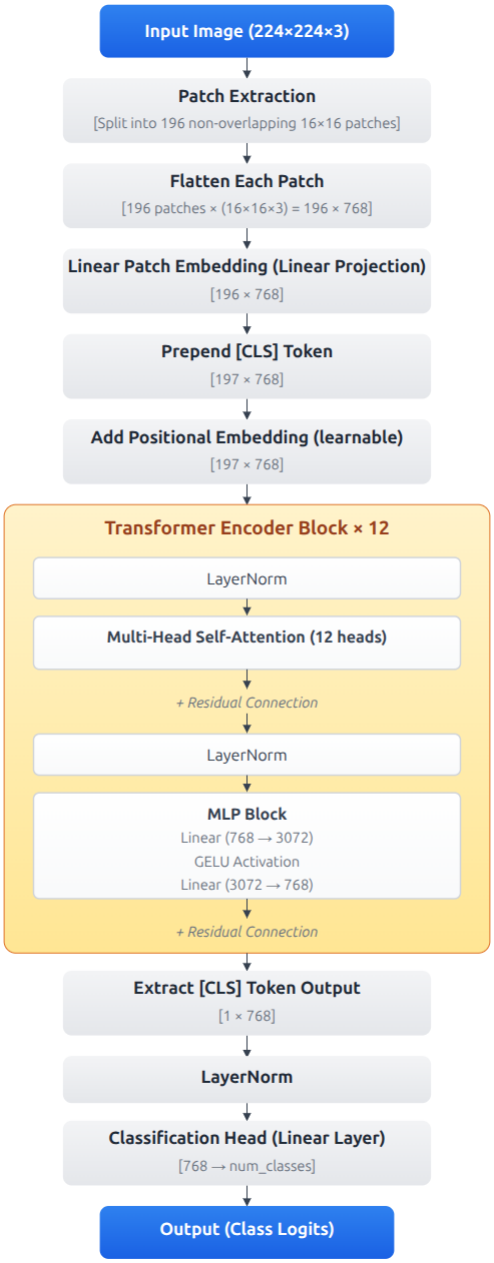
4.2 MODEL ARCHITECTURE

The proposed system employs the ViT-Base Patch16-224 architecture pre-trained on ImageNet-21k and fine-tuned for farm insect classification. As shown in Figure 3, input images are tokenized into 16×16 patches, embedded with positional encodings, and processed through 12 Transformer encoder layers with multi-head self attention and MLP blocks. The final [CLS] token is passed to a linear classification head for multi class prediction.

Fine tuning was performed using supervised learning with the AdamW optimizer, learning rate 2×10^{-6}, batch size 64, weight decay 0.02, warmup steps 50, and 500 training epochs. Epoch wise evaluation, best model checkpointing, and a custom training loop for accuracy and loss logging ensured stable convergence and robust generalization.

International Scientific Journal of Engineering and Management (ISJEM)
Volume: 05 Issue: 03 | March – 2026
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
DOI:10.55041/ISJEM05543

Figure 3: Vision Transformer (ViT-Base Patch16-224) architecture illustrating patch-based tokenization, global self-attention encoding, and the classification head for automated farm insect detection.



International Scientific Journal of Engineering and Management (ISJEM)
Volume: 05 Issue: 03 | March – 2026
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
DOI:10.55041/ISJEM05543

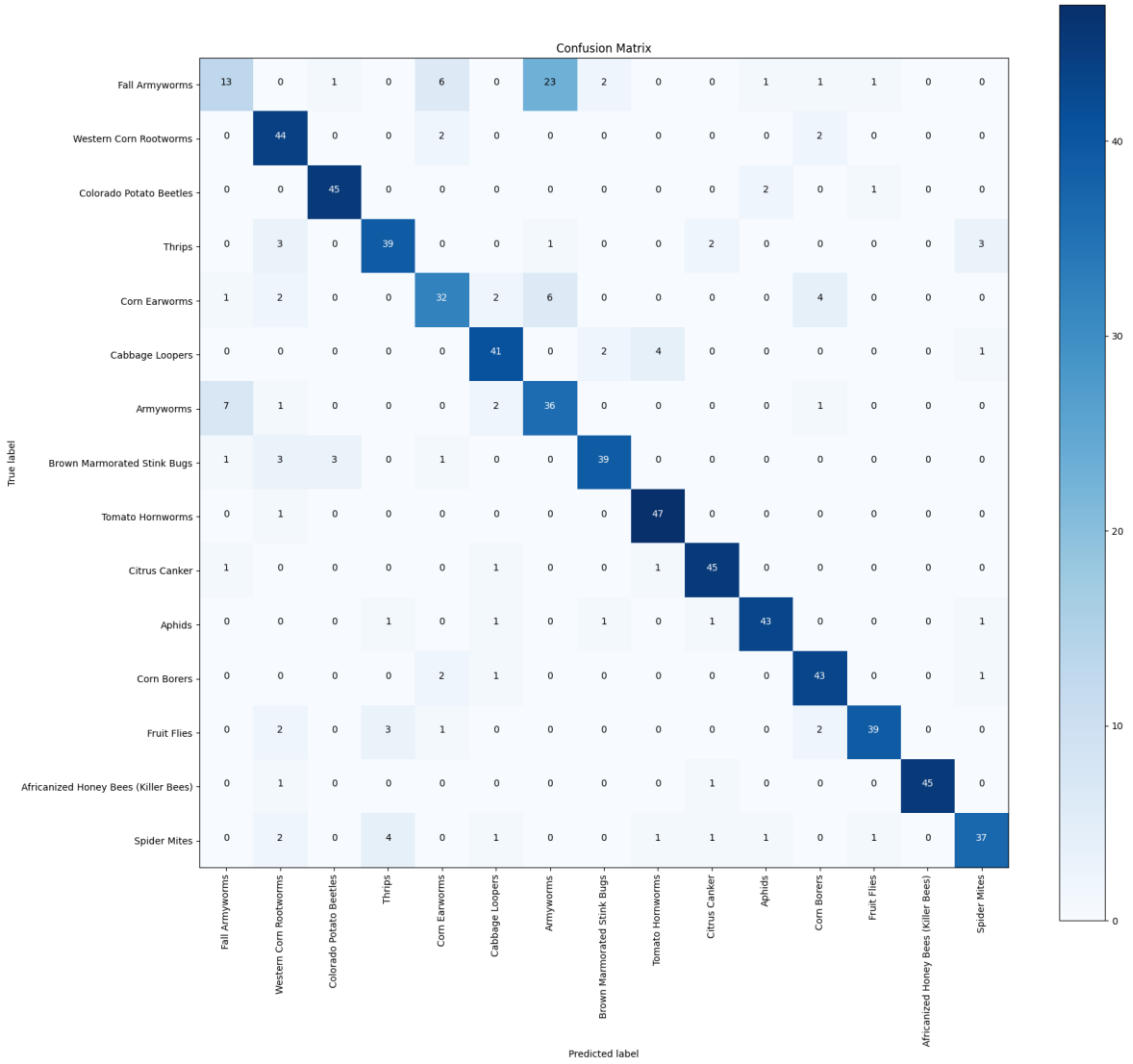


Figure 4: Confusion matrix of the multi class insect classification task, showing the distribution of correct predictions and misclassifications across all insect categories.

5. RESULTS

The overall quantitative performance of the proposed fine tuned Vision Transformer model is summarized in Table 1, which reports the global evaluation metrics on the test dataset.

The results demonstrate balanced classification performance across all insect categories, indicating the robustness and reliability of the proposed approach for multi class agricultural pest recognition.

The detailed class-wise evaluation is presented in Table 2, where performance variations across different insect species can be observed. The results reflect strong discriminative learning for visually distinctive classes, while relatively lower performance in certain categories highlights the inherent difficulty of fine grained insect classification due to visual similarity and inter class overlap, a common challenge in real world agricultural datasets.

Table 1: Performance Metrics

Metric	Value
Accuracy	0.8235
Macro F1-score	0.8175
Precision (Macro)	0.8238
Recall (Macro)	0.8236

The training behavior of the model is illustrated in Figure 5(a) and Figure 5(b), showing the training and validation accuracy and loss curves, respectively. These curves indicate stable convergence and consistent learning dynamics, demonstrating effective optimization and good generalization without severe overfitting.

Table 2: Class-wise Performance

Class	Precision	Recall	F1-score	Support
Fall Armyworms	0.5652	0.2708	0.3662	48
Western Corn Rootworms	0.7458	0.9167	0.8224	48
Colorado Potato Beetles	0.9184	0.9375	0.9278	48
Thrips	0.8298	0.8125	0.8211	48
Corn Earworms	0.7273	0.6809	0.7033	47
Cabbage Loopers	0.8367	0.8542	0.8454	48
Armyworms	0.5455	0.766	0.6372	47
Brown Marmorated Stink Bugs	0.8864	0.8298	0.8571	47
Tomato Hornworms	0.8868	0.9792	0.9307	48
Citrus Canker	0.9	0.9375	0.9184	48
Aphids	0.9149	0.8958	0.9053	48
Corn Borers	0.8113	0.9149	0.86	47
Fruit Flies	0.9286	0.8298	0.8764	47
Africanized Honey Bees (Killer Bees)	1	0.9574	0.9783	47
Spider Mites	0.8605	0.7708	0.8132	48

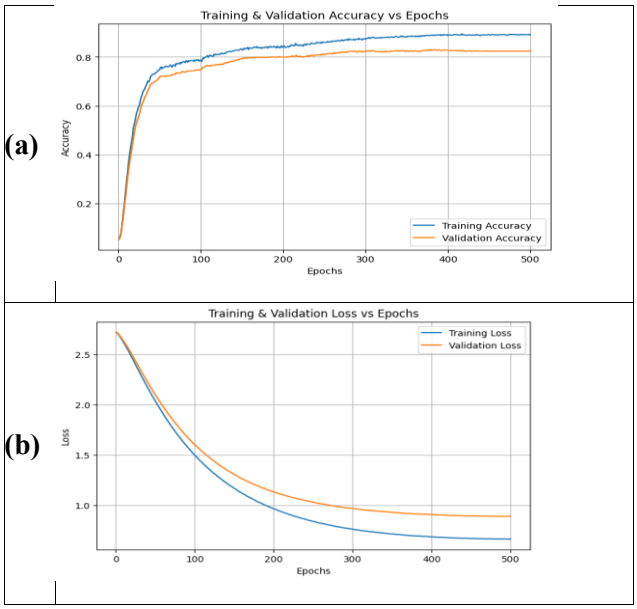


Figure 5: Training and validation performance curves: (a) Training and validation accuracy versus epochs. (b) Training and validation loss versus epochs.

Further analysis using the confusion matrix in Figure 4 reveals strong diagonal dominance, confirming reliable class separation, while the observed misclassification patterns primarily occur among morphologically similar insect species. Overall, the results validate the effectiveness and practical applicability of the proposed framework for intelligent agricultural pest monitoring systems.

CONCLUSION

This study presented a Vision Transformer based framework for the automated detection of harmful farm insects, demonstrating the effectiveness of transformer architectures for fine grained agricultural image classification. By leveraging global self attention mechanisms and transfer learning from large scale pretraining, the proposed model achieved robust performance across diverse insect categories while maintaining stable convergence and strong generalization. The experimental results validate the model's capability to learn discriminative visual representations under realistic field conditions, addressing key challenges such as visual similarity, background complexity, and inter class overlap.The integration of balanced data preprocessing, supervised fine tuning, and comprehensive evaluation further strengthens the reliability of the proposed system. This framework offers a scalable and intelligent solution for precision agriculture, supporting early pest detection, targeted intervention, and sustainable crop protection. Future work will focus on real-time deployment, multi object detection, edge device optimization, and integration with IoT and UAV based monitoring systems to enable large scale, autonomous agricultural pest surveillance.

REFERENCES

1. Domingues, T., Brandão, T., Ferreira, J.C.: Machine learning for detection and prediction of crop diseases and pests: A comprehensive survey. Agriculture 12(9) (2022) Article 1350. DOI: 10.3390/agriculture12091350

2. Chen, C.-J., Huang, Y.-Y., Li, Y.-S., Chang, C.-Y., Huang, Y.-M.: An AIoT based smart agricultural system for pests detection. IEEE Access 8 (2020) 180750–180761. DOI: 10.1109/ACCESS.2020.3024891

3. Domingues, T., Brandão, T., Ribeiro, R., Ferreira, J.C.: Insect detection in sticky trap images of tomato crops

using machine learning. Agriculture 12(11) (2022) Article 1967. DOI: 10.3390/agriculture12111967

4. Vilar-Andreu, M., García, L., García-Sánchez, A.-J., Asorey-Cacheda, R., Garcia-Haro, J.: Enhancing precision agriculture pest control: A generalized deep learning approach with YOLOv8-based insect detection. IEEE Access 12 (2024) 84420–84434. DOI: 10.1109/ACCESS.2024.3413979

5. Kasinathan, T., Singaraju, D., Uyyala, S.R.: Insect classification and detection in field crops using modern machine learning techniques. Inf. Process. Agric. 8(3) (2021) 446–457. DOI: 10.1016/j.inpa.2020.09.006

6. Marković, D., Vujičić, D., Tanasković, S., Đorđević, B., Ranđić, S., Stamenković, Z.: Prediction of pest insect appearance using sensors and machine learning. Sensors 21(14) (2021) Article 4846. DOI: 10.3390/s21144846

7. Rustia, D.J.A., Chao, J.-J., Chiu, L.-Y., Wu, Y.-F., Chung, J.-Y., Hsu, J.-C., Lin, T.-T.: Automatic greenhouse insect pest detection and recognition based on a cascaded deep learning classification method. J. Appl. Entomol. 145(3) (2021) 206–222. DOI: 10.1111/jen.12834

8. Deb, N., Rahman, T., Moniruzzaman, Md., Bin Obadi, A.S., Jizat, N.Md., Al-Bawri, S.S., Rahman, A.A.M.: Integrating feature selection and explainable CNN for identification and classification of pests and beneficial insects. Sci. Rep. 16(1) (2025) Article 2721. DOI: 10.1038/s41598-025-32520-x

9. Li, W., Zheng, T., Yang, Z., Li, M., Sun, C., Yang, X.: Classification and detection of insects from field images using deep learning for smart pest management: A systematic review. Ecol. Inform. 66 (2021) Article 101460. DOI: 10.1016/j.ecoinf.2021.101460

10. Sohel, A., Shakil, M.S., Siddiquee, S.Md.T., Marouf, A.A., Rokne, J.G., Alhajj, R.: Enhanced potato pest identification: A deep learning approach for identifying potato pests. IEEE Access 12 (2024) 172149–172161. DOI: 10.1109/ACCESS.2024.3488730

11. Haider, Z.A., Khan, F.M., Khan, I.U., Khan, M.A., Khan, R.: Early detection and prediction of pests in field crops using transfer learning. VFAST Trans. Softw. Eng. 12(3) (2024) 98–113. DOI: 10.21015/vtse.v12i3.1874

12. Mamdouh, N., Khattab, A.: YOLO-based deep learning framework for olive fruit fly detection and counting. IEEE Access 9 (2021) 84252–84262. DOI: 10.1109/ACCESS.2021.3088075

13. Liu, G., Di, J., Wang, Q., Zhao, Y., Yang, Y.: An enhanced and lightweight YOLOv8-based model for accurate rice pest detection. IEEE Access 13 (2025) 91046–91064. DOI: 10.1109/ACCESS.2025.3569819

14. Bjerge, K., Alison, J., Dyrmann, M., Frigaard, C.E., Mann, H.M.R., Høye, T.T.: Accurate detection and identification of insects from camera trap images with deep learning. PLOS Sustain. Transform. 2(3) (2023) Article e0000051. DOI: 10.1371/journal.pstr.0000051

15. Hakim, A., Srivastava, A.K., Hamza, A., Owais, M., Habib-ur-Rahman, M., Qadri, S., Qayyum, M.A., Ahmad Khan, F.Z., Mahmood, M.T., Gaiser, T.: Yolo-pest: An optimized YOLOv8x for detection of small insect pests using smart traps. Sci. Rep. 15(1) (2025) Article 14029. DOI: 10.1038/s41598-025-97825-3

16. Huang, J., Huang, Y., Huang, H., Zhu, W., Zhang, J., Zhou, X.: An improved YOLOX algorithm for forest insect pest detection. Comput. Intell. Neurosci. 2022(1) (2022) Article 5787554. DOI: 10.1155/2022/5787554

17. Katherine, S., Swain, S.K., Deori, C., Gouli, S.M., Pagire, K.S., Saikia, B., Bhosale, T.A., Chauhan, A., Choudhary, N., Choudhary, B.: Drones and AI in insect surveillance: Transforming pest forecasting systems. Int. J. Res. Agron. 8(8S) (2025) 366–376. DOI: 10.33545/2618060X.2025.v8.i8Se.3588

18. Park, Y.-H., Choi, S.H., Kwon, Y.-J., Kwon, S.-W., Kang, Y.J., Jun, T.-H.: Detection of soybean insect pest and a forecasting platform using deep learning with unmanned ground vehicles. Agronomy 13(2) (2023) Article 477. DOI: 10.3390/agronomy13020477

19. Kirkeby, C., Rydhmer, K., Cook, S.M., Strand, A., Torrance, M.T., Swain, J.L., Prangsma, J., Johnen, A., Jensen, M., Brydegaard, M., Græsbøll, K.: Advances in automatic identification of flying insects using optical sensors and machine learning. Sci. Rep. 11(1) (2021) Article 1555. DOI: 10.1038/s41598-021-81005-0

20. Utku, A., Kaya, M., Canbay, Y.: A new hybrid ConvViT model for dangerous farm insect detection. Appl. Sci. 15(5) (2025) Article 2518. DOI: 10.3390/app15052518