

A Comprehensive Survey of Deep Learning for Pest and Damage Detection in Smart Agriculture: From YOLOv5 to Hybrid Vision-Transformer Models

Sakshi Bidve¹, Shreya Jadhav², Mansi Bhosale³, Sakshi Suryawanshi⁴, Prof. K. S. Bhagwat⁵

^{1,2,3,4,5}Vidya Pratishthan's Kamalnayan Bajaj Institute of Engineering and Technology, Baramati

¹sakshib1006@gmail.com ²shreya.jadhav203@gmail.com ³bhosale.mansi2004@gmail.com

⁴2003sakshi.i.s@gmail.com ⁵keshav.bhagwat@vpkbiet.org

Abstract—Pest and crop damage detection is a critical task in modern agriculture, helping farmers protect crops and improve productivity. In recent years, deep learning models have been widely used for this purpose. This survey paper reviews key techniques, from Convolutional Neural Networks (CNNs) to the YOLO (You Only Look Once) series and Vision Transformers (ViTs). YOLO has become popular for real-time pest detection because of its high speed and accuracy, mostly used to identify pests and lesions. Vision Transformers, on the other hand, are increasingly applied for detailed analysis, such as classifying the severity of crop damage by capturing global contextual features from images. Some studies also integrate these models with IoT devices to provide faster and more accessible detection for farmers. While both families of models are effective, optimized YOLO variants, especially for object detection tasks, often achieve a better balance between accuracy and computational efficiency for field use. This survey summarizes their strengths and limitations—such as limited datasets, need for powerful hardware, and challenges in real farm conditions—and highlights future directions to make pest detection systems more accurate, cost-efficient, and farmer-friendly.

Index Terms—Pest detection, YOLO, Vision Transformer, Image Segmentation, Severity Estimation, IoT, Smart Agriculture, Deep Learning, Hybrid Models, YOLOv5, YOLOv8, Dataset.

I. INTRODUCTION

Global food production faces extreme danger and severe threats from plant diseases and pests. These biological threats cause farmers to lose massive portions of their harvests every year, with pest damage alone destroying nearly 40% of global crop production, which severely strains the world's food security [1]. Early identification is very important for effective treatment, but it remains a major challenge.

For decades, farmers relied on manual field inspections to identify crop health issues. This traditional method is highly dependent on human experience, making it slow, laborious, and notoriously prone to error, especially when identifying subtle, early-stage symptoms or very small insects [5]. Furthermore, this manual approach often forced farmers into widespread, unnecessary use of chemical pesticides (known as broadcast spraying) [1], [16]. This practice significantly increases costs, damages biodiversity, and raises serious environmental and health risks, directly contradicting the goals of sustainable agriculture [16], [17]. To mitigate these harms, modern farming demands a fundamental shift from broad, reactive treatment to precise, preventative intervention.

To overcome these traditional limitations, modern agriculture is transforming into Precision Agriculture (PA). PA integrates advanced technologies like drones, sensors, and the Internet of Things (IoT) to monitor crops more efficiently and accurately [2], [17]. The core of this modernization is

Artificial Intelligence (AI), particularly Deep Learning (DL) and Computer Vision (CV) [17]. These AI-driven tools analyze digital images of leaves and plants to detect health issues non-destructively, enabling pinpoint accuracy for actions like targeted spraying, which drastically cuts down on the amount of chemicals used [1], [16].

This survey explores the evolution of AI models tailored for pest and disease detection, moving from initial Convolutional Neural Networks (CNNs) to current state-of-the-art architectures. We focus specifically on the two dominant approaches: the YOLO (You Only Look Once) model family, which excels at speed and real-time detection for mobile and drone-based systems [9], [15]; and Vision Transformers (ViTs), which are key for advanced diagnostic analysis. A prime example of YOLO's capability is a system achieving up to 99.5% accuracy in detecting pests in cruciferous vegetables [6]. ViTs, often used in hybrid systems, capture global contextual information, making them vital for accurately assessing disease severity, such as calculating the exact percentage of a leaf covered by a lesion [3], [14].

Despite these technological leaps, significant challenges remain, primarily relating to acquiring large, diverse, real-world datasets and optimizing complex models for resource-constrained edge devices [2], [11]. This paper reviews these advances, synthesizes the core findings, and outlines the critical future directions needed to make high-accuracy, AI-driven crop protection systems accessible and effective for farmers worldwide.

II. RELATED WORK

A. Evolution of Crop Monitoring

The shift in how we monitor crops is essential because constant threats from pests and diseases severely endanger global food security [2].

- The old, conventional method of manual field inspection is slow, laborious, and very likely to result in errors, especially when searching for tiny, early signs of problems [5].
- Because detection is so often late, farmers typically resort to widespread chemical pesticide application, which is costly, hurts the natural environment, and poses health risks [1], [17].

B. Rise of Deep Learning and CNNs

To modernize farming, Precision Agriculture (PA) requires integrating smart technology and powerful computing techniques [2], [17]. This necessity enabled the development of Deep Learning (DL).

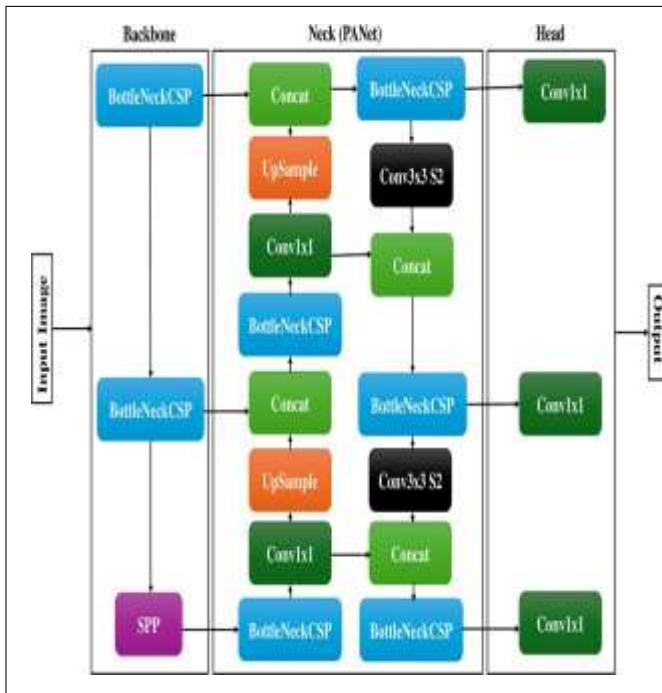


Fig. 1. The foundational architecture of YOLO, illustrating the efficient single-stage process (Backbone, Neck, Head) used for real-time object detection.

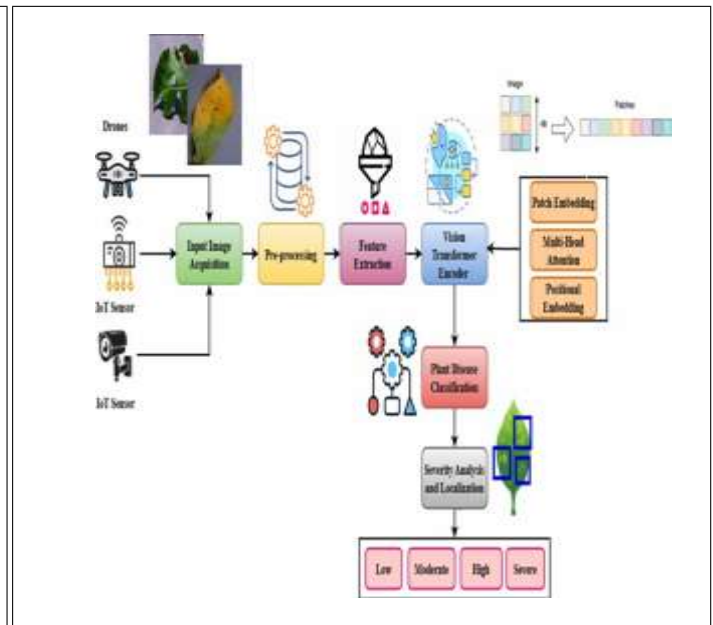


Fig. 2. Conceptual diagram of the Vision Transformer (ViT) mechanism, showing image patching and the global self-attention process used to capture context.

the entire plant [3], [14]. This global context processing is crucial for accurately assessing damage severity (see Figure 2).

- ViT Strength: This "global view" makes ViTs effective for tasks like accurately classifying disease severity [14].
- Hybrid Architectures: The best modern systems pair YOLO's rapid detection with the ViT's detailed contextual analysis in hybrid models [3], [11].
- Multimodal Detection: Combining images with inputs from acoustic sensors (listening for pest sounds) is a growing trend toward multimodal detection for improved robustness [3], [13].

C. YOLO: Revolutionizing Real-Time Detection

The absolute need for speed in field monitoring led to the dominance of the YOLO (You Only Look Once) model family [9].

- Core Mechanism: YOLO is extremely fast because it uses a single-stage network that identifies all objects in the image in just one pass [9], [15]. The efficient structure of these models is fundamental to modern detection (see Figure 1).
- Performance: This efficiency is reflected in results: a YOLOv5 system achieved an outstanding 99.5% average accuracy for detecting insects and their damage on cruciferous vegetables [6].
- Suitability: YOLO models are ideal for quickly checking crops from fast-moving drones or portable mobile applications [5].

D. Vision Transformers and Hybrid Models

Vision Transformers (ViTs) specialize in looking at the entire picture to understand complex situations, making them excellent for deeper analysis [3], [14].

- ViT Mechanism: ViTs use a powerful self-attention mechanism to capture global contextual information across

E. Deployment Challenges and Future Directions

Getting these sophisticated solutions into every farmer's hands still faces major hurdles.

- The reliability of DL models is challenged because limited data diversity means models trained on clean lab photos often fail with real-world clutter and lighting [2], [17].
- Highly accurate models carry a high computational cost (high FLOPs), making them too demanding for simple, low-cost equipment [1], [11].
- Therefore, successfully deploying these solutions requires optimizing them for cost-efficiency and accessibility [16], [17]. The ultimate goal is a functional application for farmers, as shown by detection results in Figure 3.

III. DATASETS

High-quality image data is the foundation of any successful deep learning system. The size, variety, and realism of the data determine whether a model will work reliably in a real-world field.



Fig. 3. Visual representation of real-time pest detection output, showing bounding boxes and confidence scores on plant leaves for immediate field assessment.

A. Standard Public Datasets (Benchmarks)

These common collections help researchers test and compare models, but they often have limitations.

- **Plant Village Dataset:** This data is often used for general plant disease research. However, a major limitation is that the photos are typically taken in perfect, simple lab settings, which prevents the models from performing well when they encounter real-world field conditions [2].
- **IP102 Dataset:** This is a large collection focused on 102 different types of insects [10]. It is especially useful for creating one general detection tool that simply flags the presence of any pest [1].
- **Pest24 Dataset:** This resource contains many labeled photos (25,378) [18] of agricultural pests. It is a very difficult benchmark because the pests are extremely small, often clustered together, and look very similar to each other [1], [18].

B. Custom and Specialized Datasets

Researchers often collect specialized data to ensure their models can handle specific crops or unique local problems.

- **Cruciferous Crop Dataset:** This custom data set started with 2,730 original images (expanded to 6,403). It covers 21 different categories, including pests, helpful insects, and damage marks, ensuring a full view of the plant ecosystem [6].
- **Apple Leaf Disease Datasets:** This collection has 14,066 images focusing on backgrounds that look messy and spots of varying sizes [8]. The goal is to prepare the models for the complexities of real farm conditions [8].

- **Coconut Disease/Pest Dataset:** This specific data helps models detect the serious Weligama Coconut Leaf Wilt Disease (WCLWD) and Coconut Caterpillar Infestation (CCI) [12].
- **Acoustic Data:** Specialized acoustic sensors are used to listen for insect sounds, providing another valuable detection method that uses hearing instead of just pictures [3], [13].

C. Main Challenges in Data Acquisition

Getting good data is one of the biggest challenges in agricultural AI.

- **Lack of Field Diversity:** This is a core problem: if models are trained only on clean lab photos, they often fail when dealing with real-world complexities like clutter, uneven lighting, and shadows [2], [17].
- **Annotation Cost and Class Imbalance:** Labeling images accurately is expensive and takes time [1], [17]. Furthermore, real-world data often shows imbalance, meaning rare pests do not have enough pictures for robust training [18].
- **Synthetic Data Generation:** Researchers are now using tools like Generative Adversarial Networks (GANs) to create fake photos and increase data size quickly [15]. However, questions remain about how realistic these generated images truly are [15].

IV. MODELS AND METHODOLOGIES

This section explains the technical foundation of modern crop protection, detailing the core mechanics and specialized tools used in deep learning.

A. Foundational Models and Optimization

• **Convolutional Neural Networks (CNNs)**

- **Core Mechanism:** Traditional CNNs automatically extract key visual characteristics (like shapes and textures) directly from image pixels [17].
- **Solution (Lightweight Architectures):** To manage model size, solutions like Depth-wise separable convolutions were developed to drastically reduce the total number of parameters and the computing power needed [19].

B. YOLO: The Standard for Real-Time Object Detection

The urgent need for speed led to the dominance of the YOLO (You Only Look Once) model family [9].

- **Core Mechanism:** YOLO is extremely fast because it uses a single-stage network that processes the image in one quick pass, identifying all objects simultaneously [9], [15].
- **High Performance:** This speed translates to reliable results: a YOLOv5 system achieved an outstanding 99.5% average accuracy for detecting insects on cruciferous crops [6].
- **Key Optimizations:** Models use Lightweight Backbones (like GhostNet) to reduce processing power (FLOPs) by up to 42% [18], and Multi-Scale Fusion to accurately find small, dense lesions [8].

C. ViTs and Hybrid Models for Advanced Analysis

Vision Transformers (ViTs) offer a new, deeper diagnostic view by analyzing the entire scene [14].

- **ViTs:** ViTs use a unique self-attention mechanism to capture global contextual information across the entire plant [3], [14]. This perspective makes them excellent for complex tasks like judging damage severity [14].
- **Hybrid Systems:** The best systems today use Hybrid Models to pair YOLO's speed with ViT's detailed contextual analysis [3], [11].
- **Multimodal Detection:** Combining image data with acoustic sensors is a growing trend for a robust, non-visual confirmation of pest presence [3], [13].

D. Synthesis of Findings

- **Accuracy vs. Efficiency:** Optimized YOLO variants consistently achieve the best balance between high accuracy and real-time efficiency for object detection tasks [1], [9].
- **Handling Complex Data:** Advanced techniques like Deformable Convolutional Networks (DCN) prove crucial for detecting dense, uneven lesions against complex backgrounds [8].
- **Beyond Visuals:** The inclusion of acoustic sensors demonstrates a shift toward multimodal detection, offering a non-visual method to confirm pest presence early [3], [13].
- **ViT for Depth:** ViTs excel at learning complex, global features, making them optimal for tasks like assessing disease severity [14].

V. RESULTS AND ANALYSIS

A. Performance Metrics of Key YOLO and ViT Models

TABLE I
PERFORMANCE METRICS OF KEY YOLO AND ViT MODELS

Model / Algorithm	Crop / Dataset	Metric	Result	Optimization / Context	Ref.
YOLOv5l	Cruciferous crops	sAccuracy	99.5%	Robust across 21 categories	[6]
YOLOv4	Peach / Strawberry	Accuracy / mAP	98% / 98%	Real-time efficiency (29s detection)	[15]
YOLOv8 (Improved)	Apple Leaf Lesions	mAP _{0.5}	58.5%	DCN + EMA helps dense lesion detection	[8]
BF-YOLO (YOLOv5 Hybrid)	Pest24	mAP / FPS	74.1% / 104.2	Edge-ready (42% reduction in FLOPs)	[18]
Vision Transformer (ViT)	Cotton Whitefly	Accuracy	99%	Reached accuracy within 10–20 epochs	[3]
AppViT (Hybrid ViT)	Apple Leaf Diseases	Accuracy	96.4%	Lightweight (1.3M params, 0.644B FLOPs)	[11]

B. Synthesis of Findings

- **Accuracy vs. Efficiency:** Optimized YOLO variants are currently the top choice because they strike the best

balance between being highly accurate and operating quickly in real-time. This efficient, single-stage approach ensures rapid identification of pests or damaged areas, which is essential for successful autonomous monitoring systems.

- **Handling Complex Data:** When dealing with messy field photos, detecting very dense or irregularly shaped disease spots is challenging[cite: 9]. Advanced methods like Deformable Convolutional Networks (DCN) and multi-scale fusion techniques are crucial for tackling this problem[cite: 9]. These techniques are specifically engineered to recognize complex, uneven lesions even against confusing natural backgrounds[cite: 9].
- **Beyond Visuals:** Detection is rapidly moving beyond just looking at pictures; the trend is toward gathering data from multiple sources (multimodal detection)[cite: 3, 14]. For instance, the inclusion of acoustic sensors allows the system to literally listen for the specific sounds of pests feeding or moving[cite: 3, 14]. This non-visual confirmation method is highly robust, especially for early pest detection when visual signs are still hidden[cite: 3, 14].
- **ViT for Depth:** ViTs excel at learning deep, complex features by analyzing the entire image at once (global features)[cite: 15]. This comprehensive, big-picture view makes them the optimal tool for demanding tasks like accurately judging how widespread or severe a disease is on a plant[cite: 15].

VI. CONCLUSION AND FUTURE DIRECTIONS

A. Conclusion

This survey confirms that deep learning has fundamentally changed how agricultural problems are diagnosed. Optimized YOLO models are the current standard for real-time object detection, excelling at high-speed identification of pests and damage. This achievement represents a massive step away from slow, human-centric visual inspection toward rapid, AI-driven diagnostics. We have witnessed a clear and strategic shift toward combining the unique power of different technologies: the blending of YOLO's fast localization capabilities with the profound understanding of visual context offered by Vision Transformers (ViTs). This successful integration marks the pathway toward highly sophisticated hybrid systems capable of both rapid localization and detailed severity analysis. The overall shift toward precise, efficient DL tools is a major victory for the future of sustainable crop management.

B. Future Directions

Despite the current technological leaps, several major challenges must be solved to make AI truly accessible and reliable for every farmer.

- **Data Quality and Diversity:** The biggest hurdle is still data. The primary need is for high-quality datasets that are actually captured under messy, real-world field conditions, not just in sterile labs. Gathering better, more varied data is essential to improve model generalization, ensuring models don't fail when faced with real shadows, uneven lighting, or plant clutter.

- Model Compression and Edge Deployment: Next, we need lighter, more efficient models. Future work must focus on optimizing high-accuracy models to significantly reduce their massive size and FLOPs (computing demands)[cite: 10, 12, 19]. Successfully shrinking these models is the only way to make advanced detection systems practical and affordable for the low-cost mobile and IoT devices that farmers use daily[cite: 10, 12, 19].
- Multimodal Integration: The future of warnings is multi-sensory. Further research should explore how to seamlessly integrate visual, acoustic (sound), and environmental sensor data (IoT) into single models to achieve much more robust, early warnings that work even when a pest is hidden from view[cite: 14, 3].
- Advanced Diagnostics: Finally, AI systems need to become smarter diagnosticians. Future models must tackle complex problems such as spotting several issues at once (mixed-disease detection) and delivering fine-grained severity estimation (calculating the precise percentage of leaf damage) to guide exact treatment plans[cite: 15, 13].

- [13] M. A. Ali, M. S. Sayeed, and S. F. A. Razak, "Pest Detection in Smart Farms: Integrating Sound, Deep Learning, and IoT," in *2025 International Conference on Information and Communication Technology (ICoICT)*, pp. 1-6, 2025.
- [14] M. S. Murugavalli and R. Gopi, "Plant leaf disease detection using vision transformers for precision agriculture," *Scientific Reports*, vol. 15, p. 22361, 2025.
- [15] Y. Alhwaiti, M. Khan, M. Asim, M. H. Siddiqi, M. Ishaq, and M. Alruwaili, "Leveraging YOLO deep learning models to enhance plant disease identification," *Scientific Reports*, vol. 15, p. 7969, 2025.
- [16] A. A. Shahid, A. Akter, and M. S. Hossain, "AI-CABB: An AI-Enabled Integrated Pest Detection System for Enhancing Cabbage Cultivation using Machine Learning," in *2023 5th International Conference on Sustainable Technologies for Industry 5.0 (STI)*, pp. 1-5, 2023.
- [17] D. Popescu, A. Dinca, L. Ichim, and N. Angelescu, "New trends in detection of harmful insects and pests in modern agriculture using artificial neural networks: a review," *Frontiers in Plant Science*, vol. 14, p. 1268167, 2023.
- [18] F. Qi, Y. Wang, Z. Tang, and S. Chen, "Real-time and effective detection of agricultural pest using an improved YOLOv5 network," *Journal of Real-Time Image Processing*, vol. 20, p. 33, 2023.
- [19] I. Kaushik, N. Prakash, and A. Jain, "Plant disease detection using a depth-wise separable-based adaptive deep neural network," *Multimedia Tools and Applications*, vol. 84, pp. 887-915, 2025.
- [20] M. S. Murugavalli and R. Gopi, "Precision leaf classification using convolutional neural networks," *Scientific Reports*, 2025.

VII. REFERENCES

REFERENCES

- [1] M. Vilar-Andreu, L. Garc'ia, A.-J. Garcia-Sanchez, R. Asorey-Cacheda, and J. Garcia-Haro, "Enhancing Precision Agriculture Pest Control: A Generalized Deep Learning Approach With YOLOv8-Based Insect Detection," *IEEE Access*, vol. 12, pp. 84420-84434, 2024.
- [2] E. Yilmaz, S. C. Bocekci, C. Safak, and K. Yildiz, "Advancements in smart agriculture: A systematic literature review on state-of-the-art plant disease detection with computer vision," *IET Computer Vision*, 2025.
- [3] R. S. T. Anjali, A. S. S. Ramasubbareddy, and Y. Cho, "The Power of Vision Transformers and Acoustic Sensors for Cotton Pest Detection," *IEEE Open Journal of the Computer Society*, vol. 5, pp. 356-367, 2024.
- [4] T. D. Salka, M. B. Hanafi, S. M. S. A. A. Rahman, D. B. M. Zulperi, and Z. Omar, "Plant leaf disease detection and classification using convolution neural networks model: a review," *Artificial Intelligence Review*, vol. 58, pp. 322-322, 2025.
- [5] M. H. Tanveer, Z. Fatima, M. A. Khan, M. F. Shahid, and R. C. Voicu, "Real-Time Plant Disease Detection Using YOLOv5 and Autonomous Robotic Platforms for Scalable Crop Monitoring," in *2024 9th International Conference on Control, Robotics and Cybernetics (CRC)*, pp. 1-4, 2024.
- [6] S. Chakrabarty, P. R. Shashank, C. K. Deb, M. A. Haque, P. Thakur, D. Kamil, S. Marwaha, and M. K. Dhillon, "Deep learning-based accurate detection of insects and damage in cruciferous crops using YOLOv5," *Smart Agricultural Technology*, vol. 9, p. 100663, 2024.
- [7] L. Du, J. Zhu, M. Liu, and L. Wang, "YOLOv7-PSAFP: Crop pest and disease detection based on improved YOLOv7," *IET Image Processing*, vol. 18, pp. 4913-4927, 2024.
- [8] S. Huo, N. Duan, and Z. Xu, "An improved multi-scale YOLOv8 for apple leaf dense lesion detection and recognition," *IET Image Processing*, vol. 18, pp. 4913-4927, 2024.
- [9] R. Sapkota, M. Flores-Calero, R. Qureshi, C. Badgular, U. Nepal, A. Poulouse, P. Zeno, U. B. P. Vaddevolu, S. Khan, M. Shoman, H. Yan, and M. Karkee, "YOLO advances to its genesis: a decadal and comprehensive review of the You Only Look Once (YOLO) series," *Artificial Intelligence Review*, vol. 58, pp. 274-274, 2025.
- [10] B. T. Vamsi, J. Yadla, and G. V. Kumar, "Application of IP102 Dataset in Pest Detection by using Deep Learning Techniques," in *2023 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI)*, pp. 1-6, 2023.
- [11] W. Ullah, K. Javed, M. A. Khan, F. Y. Alghayadh, M. W. Bhatt, I. S. Al Naimi, and I. Ofori, "Efficient identification and classification of apple leaf diseases using lightweight vision transformer (ViT)," *Discover Sustainability*, vol. 5, p. 116, 2024.
- [12] S. Vidhanaarachchi, J. L. Wijekoon, W. A. S. P. Abeysiriwardhana, and M. Wijesundara, "Early Diagnosis and Severity Assessment of Weligama Coconut Leaf Wilt Disease and Coconut Caterpillar Infestation Using Deep Learning-Based Image Processing Techniques," *IEEE Access*, vol. 13, pp. 24463-24477, 2025.