Research Review of Deep Learning Algorithms for Agricultural Disease Image Classification

Shengjiu JIANG, Qian WANG

Shaoyang Industry Polytechnic College, Shaoyang 422000, China

Abstract In the context of rural revitalization and the development of smart agriculture, image classification technology based on deep learning has emerged as a crucial tool for digital monitoring and intelligent prevention and control of agricultural diseases. This paper provides a systematic review of the evolutionary development of algorithms within this field. Addressing challenges such as domain drift and limited global awareness in classical convolutional neural networks (CNNs) applied to complex agricultural environments, the paper focuses on the latest advancements in vision transformers (ViT) and their hybrid architectures to enhance cross-domain robustness and fine-grained recognition capabilities. In response to the challenges posed by scarce long-tail data and limited edge computing power in real-world scenarios, the paper explores solutions related to few-shot learning and ultra-lightweight network deployment. Finally, a forward-looking analysis is presented on the application paradigms of multimodal feature fusion, vision-based large models, and explainable artificial intelligence (AI) within smart plant protection. This analysis aims to offer theoretical insights for the development of efficient and transparent intelligent diagnostic systems for agricultural diseases, thereby supporting the advancement of digital agriculture and the construction of a robust agricultural nation.

Key words Agricultural disease image, Classification algorithm, Deep learning, Research Review

1 Introduction

In recent years, with the vigorous implementation of China's rural revitalization strategy and in accordance with the directives of No. 1 Central Document to develop new agricultural productive forces adapted to local conditions, the advancement of the comprehensive integration of artificial intelligence (AI) with agriculture, alongside the application of AI technologies to ensure food security and improve the digital capabilities for pest and disease management, has emerged as a significant national strategic priority[1]. Globally, crop diseases and pests remain common challenges that constrain agricultural productivity. According to statistical data, crop diseases and pests result in an annual loss of approximately 20% to 40% of global food production, leading to direct economic damages exceeding 220 billion US dollars. Traditional methods for disease diagnosis rely heavily on field visual inspections conducted by agronomists or plant protection specialists. These methods are not only time-consuming and labor-intensive but also highly subjective, thereby limiting their capacity to fulfill the real-time monitoring requirements of contemporary large-scale intensive farming operations.

The integration of AI for automated diagnosis has become a critical requirement to address existing gaps in the field. Early automation approaches primarily utilized algorithms such as support vector machines (SVM) and random forests, which depended on the manual extraction of color and texture features of pathogens. However, when applied to real-world environments characterized by significant variations in lighting and shading caused by branches, stems, and leaves, the performance of these models was substantially diminished. More recently, convolutional neural networks (CNNs) have emerged as the dominant methodology, employing multi-layer nonlinear transformations to automatically extract high-level semantic features, thereby achieving a significant improvement in classification accuracy on controlled datasets[2].

Although deep learning-based image classification demonstrates strong performance on laboratory datasets, its effective translation to practical field applications remains challenging. Currently, four primary issues hinder the deployment and implementation of deep learning models in real-world agricultural settings. First, a significant "domain shift" exists between data collected under controlled laboratory conditions and that obtained from actual farmland environments. Second, the fine-grained features of diseases, particularly during early pathological stages, are often obscured or blurred against complex natural backgrounds. Third, high-precision models typically involve a large number of parameters, while edge nodes within agricultural Internet of things (IoT) systems deployed in the field are constrained by limited computational resources and power capacity. Fourth, samples of sudden or rare diseases in real-world scenarios are exceedingly scarce, impeding the model's ability to acquire sufficient training data for these tail diseases.

2 Dataset evolution and real-world deployment challenges

The performance limits of deep learning models are fundamentally determined by the quality and distribution of the underlying data. Examining the progression of agricultural image datasets reveals a

Received: January 12, 2026 Accepted: March 5, 2026

Supported by School-level Project of Shaoyang Industry Polytechnic College (SKY24A06); Science and Technology Plan (Special Fund Subsidy) of Shaoyang City (2024PT4070); General Research Project of Hunan Provincial Department of Education in 2025 (25C1457).

Shengjiu JIANG, master's degree, teaching assistant, research fields: computer vision and its applications.

shift from controlled laboratory settings to complex farmland environments. Early publicly available datasets predominantly consisted of images captured against solid color backgrounds or petri dishes, with consistent and singular lighting conditions (Fig. 1).

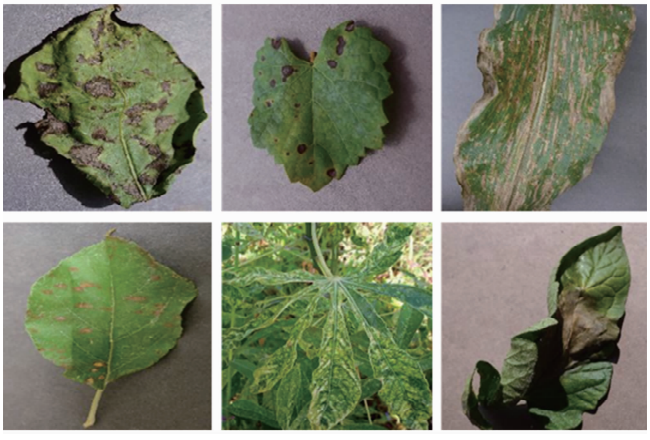

Fig. 1 Examples of dataset

Models trained exclusively on undisturbed background data frequently fail when deployed in real-world natural environments, primarily due to their limited capacity to adapt to noise. The complexity of actual farmland environments has necessitated a transformation in data collection standards. Recent studies have increasingly focused on the physical conditions present in real field scenarios. Newly developed datasets encompass challenges such as intense lighting variations, chaotic background disturbances (including soil and weeds), and significant obstructions caused by branches and leaves. Furthermore, these datasets have begun to incorporate multimodal features, including multispectral and depth maps.

2.1 Comparison of mainstream open-source datasets Table 1 provides a summary of the fundamental characteristics and primary challenges associated with the predominant open-source datasets currently utilized in the field of agricultural disease image classification. Notably, although the early PlantVillage dataset[3] was relatively large in scale, its use of solid color backgrounds likely contributed to model overfitting on a single background, thereby limiting its generalization capability to practical applications. Subsequent datasets, such as PlantDoc[4] and IP102[5], incorporated features including real farmland backgrounds, bounding box annotations, and extreme long-tail distributions, thereby better aligning with the recognition requirements encountered in natural environments. In recent years, the development of large-scale multimodal datasets, such as LeafNet[6] and CDDM[7], has facilitated data support for complex diagnostic reasoning tasks. These tasks include visual question answering (VQA), large-scale multimodal alignment, and generative AI fine-tuning, achieved through the integration of rich text annotations and question-answer (Q&A) pairs.

Table 1 Comparison of mainstream open-source agricultural disease image datasets

Dataset name	Scale (image) Sheet	Category coverage	Collection environment and main features	Main technical challenges
PlantVillage	54 305	14 types of crops, 38 categories	Controlled laboratory conditions, single-leaf solid color background	Susceptible to overfitting on a single background
PlantDoc	2 569	30 categories	Real farmland with bounding boxes	Large background noise, inconsistent lighting conditions, imbalanced data distribution
IP102	>75 000	102 species of pests	Covering different growth cycles of pests	Extreme long-tail distribution, high similarity between insect body and background
LeafNet	186 000	22 types of crops, 62 species of diseases	Multimodal data combined with rich text annotations	Supporting complex diagnostic reasoning of VQA
CDDM	137 000	A large number of diseases	Incorporating 1 million Q&A pairs on plant diseases	Large-scale multimodal alignment and generative AI fine-tuning

2.2 Fundamental realistic challenges Despite continuous enhancements in dataset richness, several challenges persist in the practical deployment of agricultural production systems. Image-based classification models primarily encounter two fundamental issues. The first issue is domain shift, which refers to the discrepancy in data distribution between the training environment and real-world scenarios, representing a significant constraint on the algorithm's implementation. Research indicates that models achieving accuracy rates as high as 99% on controlled laboratory datasets, such as PlantVillage, frequently experience a decline in test accuracy to between 70% and 85% when applied in real-world agricultural production environments affected by various factors, including changes in lighting and leaf shading. The second challenge pertains to fine-grained classification. The visual characteristics of pests and diseases are inherently complex, presenting as minimal inter-class differences and substantial intra-class variations in computer vision tasks. For instance, early blight and late blight exhibit nearly indistinguishable pathological spots during the initial stages of infection. Furthermore, the complexity of real-world backgrounds, including elements such as weeds and soil, can easily obscure the target regions, thereby exacerbating the difficulty for models to accurately extract and identify fine-grained pathological features.

3 Performance benchmarking of CNNs

To address the challenge of feature extraction in agricultural im-

age classification, the academic community initially widely adopted classical CNNs for architectural exploration and performance benchmarking. CNNs have demonstrated considerable effectiveness in capturing texture and edge details of small disease spots through the use of local receptive fields and weight-sharing mechanisms. Table 2 presents a comparative summary of the performance and application evaluation of typical CNN architectures on widely used public datasets. It is evident that different networks prioritize varying trade-offs between classification accuracy and computational resource consumption, including parameters such as the number of model parameters and memory usage.

Table 2 Benchmark benchmarking of typical CNN architectures in agricultural image classification

Network architecture	Accuracy rate (benchmark)//%	Parameter magnitude//MB	Evaluation in agricultural applications
VGG16[8]	93.75	56.79	The feature extraction process is precise and highly accurate, but it incurs substantial computational costs and presents challenges for deployment.
ResNet-50[9]	84.36	100.17	It addresses the issue of gradient vanishing by employing residual blocks, which enhance the strength of semantic features. However, it requires a substantial amount of memory.
GoogleNet[10]	88.00	85.51	The inception module demonstrates a robust capability for feature extraction, but its complex architecture presents challenges for fine-tuning.
EfficientNet-B0[11]	84.37	18.62	It offers relatively high accuracy while maintaining minimal resource consumption, rendering it suitable for deployment on mobile devices.

Although CNNs have demonstrated relatively high accuracy on common datasets, they exhibit certain limitations. Specifically, the locality inherent in standard convolution operations restricts the model's capacity to capture the overall structure of an image and to model global long-range dependencies. Consequently, when applied to scenarios such as the spread of multi-leaf diseases or real farmland images characterized by highly complex backgrounds, the model is particularly vulnerable to interference from local noise, which may result in misclassification.

4 Visual transformer (ViT) and hybrid architecture

In response to the limitations of traditional CNNs in global perception capabilities, the ViT[12], inspired by the self-attention mechanism from natural language processing, has been progressively adopted in the domain of agricultural disease detection. It has increasingly become a new standard model for enhancing the accuracy of disease recognition in complex environments.

4.1 ViT and hybrid model (CNN-ViT fusion) ViT segments the input image into fixed-size patches and employs a multi-head self-attention mechanism to dynamically model global interactions among features. This approach allows the model to integrate the overall morphology of the plant with contextual information from specific lesions, facilitating comprehensive inference. However, pure ViT models lack inductive biases inherent to CNNs, such as translation invariance and locality, which often results in a strong dependence on large-scale training datasets and substantially increased computational complexity.

To accurately capture the localized details of diseases and comprehend the overall pathological characteristics of crops, we conducted an extensive literature review and determined that the hybrid architecture combining CNN and ViT effectively addresses both aspects. This CNN-ViT fusion currently represents a highly efficient technical paradigm. For example, the PlantAIM model[13] integrates local edge features and global semantic features via a dual-path network, resulting in a marked enhancement in robustness when evaluated on real-world datasets. To address the issue of illumination variation in practical settings, MoE-ViT[14] introduced a mixture of experts system that dynamically routes input features to specialized expert classifiers through a gated network. Experimental results indicate that this architecture not only improves the overall accuracy by 20% but also achieves a generalization accuracy of 68% across different scenarios, thereby demonstrating substantial generalization capability.

4.2 Attention mechanism and data augmentation In the fine-grained disease identification task, researchers have further refined the attention mechanism. For example, by incorporating an efficient triple attention (ETA) module and employing a three-branch architecture to simultaneously compute attention weights across channel and spatial dimensions, the network's capacity to represent subtle and ambiguous pathological features is enhanced. At the data augmentation stage, conventional methods such as CutMix[15] exhibit certain limitations. Specifically, they often crop out critical lesion regions within the images, leading to "label mismatch" issues. To address this problem, attention-guided data augmentation techniques, exemplified by AttentionMix[16], have been developed. These approaches can adaptively locate and preserve key lesion areas during the mixing process.

5 Breaking through the challenges of few-shot learning

Although models based on ViT and hybrid architectures have demonstrated substantial improvements in classification accuracy within complex environments, their superior performance is heavily reliant on large volumes of well-labeled data. In contrast, real-world agricultural production settings frequently encounter data pertaining to emerging or rare diseases that exhibit severe "long-tail distributions" or extreme scarcity, resulting in a significant "cold start" problem for these models. To mitigate the limitations imposed by data dependency, few-shot learning has increasingly

emerged as a pivotal approach for addressing data scarcity in agricultural vision tasks. The primary goal of this paradigm is to enable models to attain high-precision generalization using only a minimal number of support samples (*e.g.*, 1 – 5 images)[17].

(i) Metric learning. Metric learning primarily focuses on extracting image feature spaces through the construction of Siamese networks with shared weights. By incorporating metric functions, such as cosine similarity, the network effectively reduces the feature distance between similar diseases while increasing the distance between heterogeneous features, thereby demonstrating strong robustness in fine-grained pathological identification.

(ii) Meta-learning. Meta-learning seeks to equip models with the capability to "learn how to learn". For example, the SWE-MAML[18] (sequential weighted ensemble model-agnostic meta-learning) algorithm integrates an ensemble learning mechanism within the MAML framework, effectively reducing gradient oscillations in models trained under extremely few-shot conditions. Empirical studies demonstrate that, in a real-world potato disease scenario where each category includes only 30 support samples, this method attains a classification accuracy of 75.71%.

(iii) Visual foundation model empowerment. Recent research has increasingly focused on leveraging prior knowledge from large general visual models. The PlantCaFo framework[19], introduced in 2025, achieved an accuracy of up to 93.53% under conditions of limited sample availability by freezing the backbone network weights of models such as CLIP or DINO and fine-tuning only lightweight, domain-specific adapters. Nevertheless, these approaches exhibit certain limitations. Specifically, the pre-training datasets for general foundational models predominantly consist of natural images, resulting in a significant domain gap between these data and the microscopic pathological features characteristic of agriculture. Consequently, the zero-shot generalization capabilities of such models remain constrained, particularly when addressing extremely long-tail agricultural diseases that have not been previously encountered.

6 Edge computing and lightweight network architecture

After addressing data scarcity through few-shot learning at the algorithmic level, a significant engineering challenge in implementing disease diagnosis systems remains. High-precision deep neural networks typically involve a large number of parameters and require substantial floating-point operations (FLOPs), which complicates their direct deployment on edge devices, such as unmanned aerial vehicles (UAVs) or agricultural IoT sensors, characterized by limited computational capacity and power constraints. For instance, lightweight models like MobileNet and GhostNet (Fig. 2), which utilize depthwise separable convolution, and offer a viable framework for the efficient deployment of such edge devices.

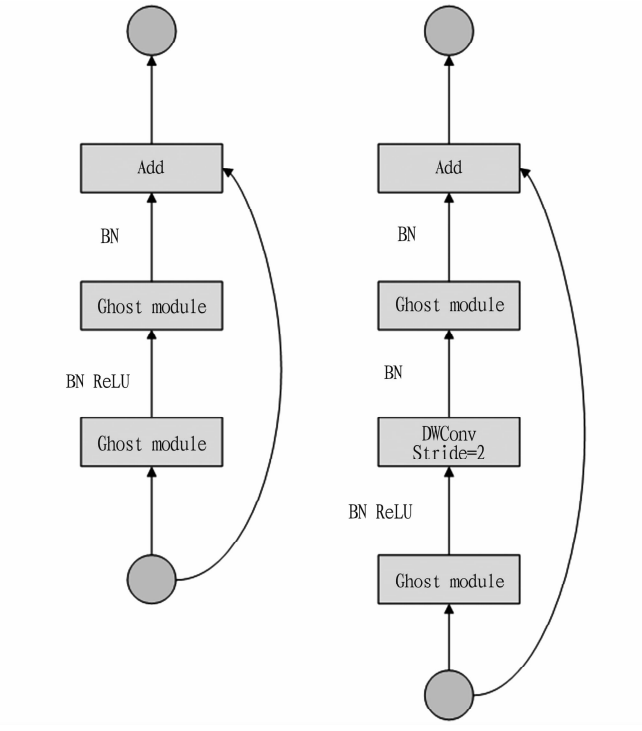

Fig. 2 GhostNet model

Table 3 presents a classification of lightweight network model architectures recently tailored for agricultural applications. Owing to variations in hardware testing platforms and quantization accuracy across different studies, latency metrics are not directly comparable. Nevertheless, these models provide valuable insights for parameter compression strategies.

Table 3 Technical features of lightweight models for agricultural edge devices

Model	Parameter quantity//M	Computational overhead/latency	Core technologies and application scenarios
ALNet[20]	0.17	151.98 MFLOPs	The pre-trained architecture, with a volume of only 677 KB, achieves an accuracy of 99.78% on the cross-fruit tree dataset, demonstrating its high compatibility with smartphone devices.
CropHealthyNet[21]	0.47	No memory bottleneck	The block self-attention mechanism is employed to reduce computational complexity. In the context of multi-scale disease diagnosis using UAVs, the model's parameters are 15 times fewer than those of DenseNet.
RepEfficientViT[22]	1.34	25.13 ms CPU latency	During training, a CNN-ViT hybrid model is employed to capture global information, while during inference, structural re-parameterization techniques are utilized to fold branches.
StrawberryDualNet[23]	0.04	83 FPS (16-bit quantization)	The model employs dynamic local-global gating fusion in conjunction with FP16 quantization and is compatible with autonomous reconnaissance vehicles possessing highly limited computational capabilities.

7 Summary and prospects: towards multimodal and interpretable intelligent plant protection systems

Image classification technology for agricultural disease detection has advanced beyond the era of basic CNNs. Models based on ViT and hybrid architectures have substantially improved diagnostic accuracy in complex agricultural environments. Additionally, approaches such as few-shot learning and ultra-lightweight networks have effectively mitigated the challenges posed by limited long-tail data and the constrained computational resources of edge devices.

With the in-depth advancement of technology, disease diagnosis is transitioning from traditional two-dimensional visual classification to multimodal fusion and enhanced decision-making transparency. At the perceptual level, reliance solely on RGB images is susceptible to misinterpretations arising from "different causes producing similar effects". A dual-modal system that integrates image data with IoT time series data can facilitate early warning during the "latent period" of diseases. Furthermore, leveraging the high-precision segmentation capabilities of general visual large models (*e. g.*, SAM) alongside the reasoning abilities of customized vision-language large models (VLMs) enables the system to generate professional agronomic recommendations. At the decision-making level, to address the "black box" nature of deep learning models, explainable artificial intelligence (XAI) techniques, including Grad-CAM and LIME, have been employed to produce heat maps and contribution weights. These visualizations offer intuitive verification for AI-based diagnoses and significantly reduce trust barriers among farmers.

Looking forward, in alignment with China's macro-level strategies for "rural revitalization" and "smart agriculture", the next-generation digital plant protection system will extensively integrate multi-modal early warning mechanisms, foundational large model ecosystems, and XAI for transparent decision-making. This system aims to develop a "digital plant protection generalist" characterized by anthropomorphic reasoning. Such advancements will not only overcome the "last mile" in applying algorithms to field-level decision-making but also serve as a pivotal driver in ensuring national food security and leading the digital transformation of agriculture.

References

[1] ZHAI ZY, CAO YF, XU HL, *et al.* Review of key techniques for crop disease and pest detection[J]. Transactions of the Chinese Society for Agricultural Machinery, 2021, 52(7): 1 – 18. (in Chinese).

[2] WEN YL, CHEN YP, WANG KQ, *et al.* An overview of plant diseases and insect pests detection based on machine vision[J]. Journal of the Chinese Cereals and Oils Association, 2022, 37(10): 271 – 279. (in Chinese).

[3] HUGHES D, SALATHÉ M. An open access repository of images on plant health to enable the development of mobile disease diagnostics[J]. arXiv preprint arXiv: 1511.08060, 2015.

[4] SINGH D, JAIN N, JAIN P, *et al.* PlantDoc: A dataset for visual plant disease detection[C]//Proceedings of the 7th ACM IKDD CoDS and 25th COMAD. New York: ACM, 2020: 249 – 253.

[5] WU X, ZHAN C, LAI YK, *et al.* Ip102: A large-scale benchmark dataset for insect pest recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2019: 8787 – 8796.

[6] BARRÉ P, STÖVER BC, MÜLLER KF, *et al.* LeafNet: A computer vision system for automatic plant species identification[J]. Ecological Informatics, 2017, 40: 50 – 56.

[7] WU T, CHEN Z, HE D, *et al.* CDDM: Channel denoising diffusion models for wireless semantic communications[J]. IEEE Transactions on Wireless Communications, 2024, 23(9): 11168 – 11183.

[8] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv: 1409.1556, 2014.

[9] HE K, ZHANG X, REN S, *et al.* Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 770 – 778.

[10] SZEGEDY C, LIU W, JIA Y, *et al.* Going deeper with convolutions [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2015: 1 – 9.

[11] TAN M, LE Q. Efficientnet: Rethinking model scaling for convolutional neural networks[C]//International Conference on Machine Learning. New York: PMLR, 2019: 6105 – 6114.

[12] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, *et al.* An image is worth 16 × 16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv: 2010.11929, 2020.

[13] CHAI AYH, LEE SH, TAY FS, *et al.* PlantAIM: A new baseline model integrating global attention and local features for enhanced plant disease identification[J]. Smart Agricultural Technology, 2025, 10: 100813.

[14] RIQUELME C, PUIGCEVER J, MUSTAFA B, *et al.* Scaling vision with sparse mixture of experts[J]. Advances in Neural Information Processing Systems, 2021, 34: 8583 – 8595.

[15] YUN S, HAN D, OH SJ, *et al.* Cutmix: Regularization strategy to train strong classifiers with localizable features [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2019: 6023 – 6032.

[16] ZHANG Y, ZHANG N, ZHU J, *et al.* Efficient triple attention and attentionMix: A novel network for fine-grained crop disease classification [J]. Agriculture, 2025, 15(3): 313.

[17] SUN J, CAO W, FU X, *et al.* Few-shot learning for plant disease recognition: A review[J]. Agronomy Journal, 2024, 116(3): 1204 – 1216.

[18] LI J, FENG Q, YANG J, *et al.* Few-shot crop disease recognition using sequence-weighted ensemble model-agnostic meta-learning[J]. Frontiers in Plant Science, 2025, 16: 1615873.

[19] JIANG X, WANG J, XIE K, *et al.* PlantCaFo: An efficient few-shot plant disease recognition method based on foundation models[J]. Plant Phenomics, 2025, 7(1): 100024.

[20] GAO H, DAI K, WANG K, *et al.* ALNet: An adaptive channel attention network with local discrepancy perception for accurate indoor visual localization[J]. Expert Systems with Applications, 2024, 250: 123792.

[21] WANG Y, GAO X, LIU J, *et al.* CropHealthyNet: A lightweight hybrid network for efficient crop disease detection [J]. Applied Sciences, 2026, 16(3): 1329.

[22] LIU T, WANG Y, YANG C, *et al.* A lightweight hybrid CNN-ViT network for weed recognition in paddy fields[J]. Mathematics, 2025, 13 (17): 2899.

[23] HAQIQ N, ZAIM M, SBIHI M, *et al.* Lightweight hybrid deep learning for strawberry disease recognition and edge deployment using dynamic multi-scale CNN-transformer fusion[J]. AgriEngineering, 2026, 8(2): 75.