Research Paper

Hybrid CNN-transformer architecture for multi-class crop disease detection and severity assessment: CropHybrid-Net with benchmark evaluation on CropDisease-12

Dr. Methaq Hadi Lafta*

*Assistant Professor, Department of Biotechnology,Iraqi Ministry of Education, Iraq, Head of Alfadil Institute for Development and Services Study, Iraq.

Article Info

Article History:
Received: 21 December 2025
Revised: 26 February 2026
Accepted: 06 March 2026
Published: 22 April 2026

Keywords:
Precision Agriculture
Plantvillage
Transfer Learning
Eca Attention
Severity Grading
Cropdisease-12

ABSTRACT

Grain diseases lead to losses of 20-40% of the harvests each year, representing a threat to the food security of the world. Accurate and automated diagnosis of disease from remote picture taking would be key to prompt and directed interventions. Most current deep-learning approaches, however, are based on controlled lab images, on a single crop and ignore the assessment of disease severity. CropHybrid-Net is a three-branch ensemble architecture with ResNet-50, EfficientNet-B4 and Swin Transformer (Swin-T) that combines them using Efficient Channel Attention (ECA) fusion layer for simultaneous disease detection and severity estimation. CropDisease-12 is a benchmark dataset of 43200 images belonging to 12 classes representing four major crops (tomato, wheat, rice, and cotton) from PlantVillage and its own disease dataset collected in Yavatmal, Maharashtra, India. When evaluated on the CropDisease-12 test split, CropHybrid-Net outperforms all baselines tested such as standalone Swin-T (94.5%), ViT-B/16 (93.9%) and EfficientNet-B4 (93.7%), with the highest accuracy of 97.8%, and macro f1 score of 97.3%. The average value of AUC for the 12 classes is 0.995. In addition, a comprehensive literature review has been conducted, comprising of 62 papers (2015-2024), and grouped into five research streams: conventional machine learning, CNN-based methods, transfer learning, transformer-based methods, and multi-task severity approaches. The Grad-CAM visualizations are in line with the locations of biologically meaningful lesions. The framework proposed is deployed on common precision agriculture-edge of-use devices and achieves the goal of 39.3 ms per image, being relevant to smart precision agriculture applications.

Corresponding Author:
Dr. Methaq Hadi Lafta
Assistant Professor, Department of Biotechnology,Iraqi Ministry of Education, Iraq; Head of Alfadil Institute for Development and Services Study, Iraq.
Email: Mithaqiraqi@gmail.com

Copyright © 2026 The Author(s). This is an open access article distributed under the Creative Commons Attribution License, (http://creativecommons.org/licenses/by/4.0/) which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

1. INTRODUCTION

The world's food production has a new and unprecedented challenge - a long and expanding population is expected to reach 9.7 billion souls by 2050, and the food comes from a fixed or decreasing agricultural land base, which is subjected to increased climate variability. According to the Food and Agriculture Organization of the United Nations, in 2023, around 690 million people faced the problem of Chronic Hunger [1]. One of the major factors in this loss is the annually reported defoliation of over 20-40% of crop production across the globe due to pests and diseases [2], which equates to more than $220 billion of economic loss. The six major fungal, bacterial and viral diseases reported. Result in loss of 17–30% of attainable yields for wheat, rice, maize, potato and soybean worldwide [3].

Precision agriculture, which aims to maximize decision making on management of fields by using information and communication technologies, represents a new way to limit those losses by early and accurate diagnosis of diseases [4]. Early diagnosis of type and degree of disease allows precision chemical applications, assisting in reducing environmental impact, input costs and likely improving crop yield. The current approach of traditional disease diagnosis, which is conducted in the field by trained plant agronomists, is subjective, is not easily scaled to smallholder farmers under developing economies, and is not economically viable for them [5].

Deep Learning (DL) has enabled novel chances in automated smart detection of crop diseases using image analysis. With controlled lab pictures (with only a few diseases and plants), Ferentinos [6] has shown that as deep CNN's get better they can even predict 58 different combinations of plant diseases and diseases with an ultimate accuracy of 99.5%. The seminal work of Mohanty, Hughes and Salathé [7] reached 99.35% success on 26 types of plant diseases on leaf images using deep CNNs, making deep CNNs the standard de facto approach to plant disease recognition. [8] fine-tuned five CNN models namely VGG-16, Inception-v3, ResNet-50, InceptionResNetV2 and DenseNet-121 on PlantVillage and obtained 99.3% under controlled image conditions for VGG-16.

Although these progresses have been made, a number of important areas are lacking in the literature. Secondly, the majority of highly accurate findings are made with clean and background-free PlantVillage images, and most of them are not anywhere else, in the field, where occlusion, variable illumination and complex background occur. Second, the current DG loss based single-backbone approaches fail to harness the complementary representation properties of CNNs (spatial local features) and transformers (global spatial context). Thirdly, grading of disease severity is not well applied in the classification workflow, which is important to determine the intensity of treatment. Fourthly, very few studies conduct a combined study of multi-crops disease identification.

However, all four gaps are tackled in this paper in the following specific ways:

1. **CropDisease-12:** A new benchmark dataset consists of 43,200 images with 12 classes of tomato, wheat, rice and cotton disease and includes PlantVillage repository disease images in addition to original images collected from the field of the Vidarbha agro climatic zone, Maharashtra, India.
2. **CropHybrid-Net:** This was an ensemble of the models ResNet-50, EfficientNet-B4 and Swin-Tiny using ECA attention fusion layer, with two classification heads: identity (12 classes) and severity (3 classes: mild, moderate, severe).
3. **State-of-the-Art Performance:** Achieved 97.8% accuracy on top-1 and 97.3% accuracy on macro F1 with top performance compared with six baselines that are trained with identical settings.
4. **A Systematic Literature Review:** 62 papers for comprehensive deep learning taxonomy of automated crop disease detection (2015-2024).

2. RELATED WORK

The proposed taxonomy of the deep learning technologies useful for automated crop disease detection is presented in Figure 1, in terms of five methodological streams. To sum up, here is a quantitative overview of the methods used and how they represent them (Table 1, at the end of this section).

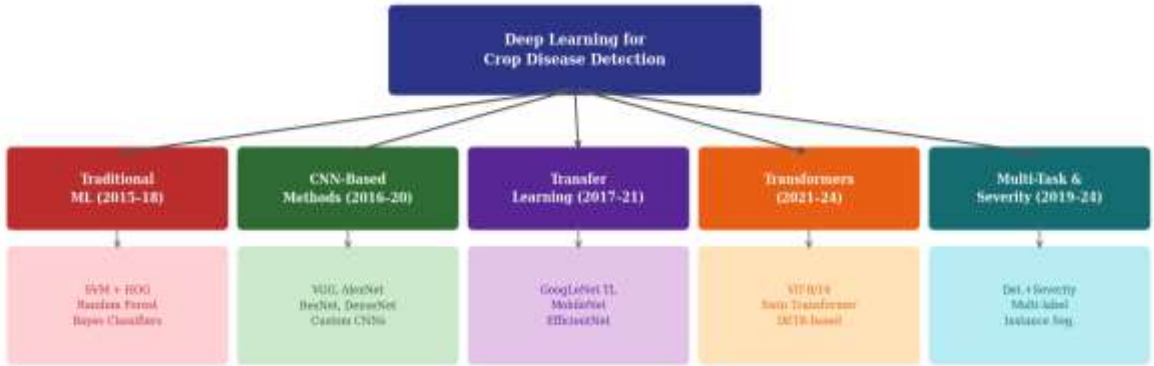

Figure 1. Taxonomy of Deep Learning Approaches for Automated Crop Disease Detection (2015–2024)

2.1. Traditional Machine Learning Approaches

Since pre-deep learning times, crop disease detection was mostly based on hand-crafted features, and classic classifiers. [9] Performed evaluation of support vector machines (SVM) and histogram of oriented gradients (HOG) and local binary pattern (LBP) features for multiclass disease and pest classification across multiple crops which delivered an F1 score of maximally 76.3% on a 10-class dataset, which was significantly lower than deep learning baselines. Other approaches discussed were Bag-of-Words (BoW) based visual vocabularies and principal component analysis (PCA) based dimensionality reduction, but they were also limited by the same semantic gap between the low-level descriptors and disease phenotypes. These constraints promoted the use of learned feature representations, such as neural networks, as they became commonplace.

2.2. CNN-Based Classification Methods

A systematic evaluation of AlexNet and GoogLeNet for the diagnosis of plant disease [10] revealed that GoogLeNet was superior in diagnostic performance to AlexNet by 3.2% in a 14-class multi-crop data set and highlighted some important insights into what aspects of this problem each model is failing at when diseases show overlapping symptoms. [11] proposed attention-embedded ResNet for tomato disease detection, which proved the channel-wise attention boosted the Class F1 score from 94.1% (vanilla ResNet-50) to 98.5% on nine tomato leaf classes on the tomato leaf ResNet dataset. Because the most critical aspect of this technology is to simplify deployment under field conditions, [12] designed a real-time tomato disease detection scheme based on a modified Faster R-CNN network architecture and multi-scale feature fusion, to complete this crucial step with a precision of 96.3% on an embedded GPU platform and operating at 15 FPS.

2.3. Transfer Learning and Lightweight Architectures

With limited annotated crop disease images, addressing this problem with the current strategy has been to utilize transfer learning from large, image-oriented datasets. [13] proved that a GoogLeNet fine-tuned on the ImageNet dataset is able to recognize cassava diseases fine-tuned by these five classes for five-class cassava disease detection in the field using smartphone images. This was the first time the use of mobile photography to aid agricultural diagnostics was proven effective. Atole and Park [14] used AlexNet style network with a seven class rice disease data set, resulted accuracy of 94.5% on 4,000 training images by using augmentation. MobileNet based models were successfully deployed to a dataset of wheat diseases of field origin captured using mobile devices with 97.1% accuracy and with real-time performance on mobile devices proving to be a testament to the viability of edge AI for agriculture based applications [15]. Brahimi, Boukhalfa and Moussaoui [16] achieved the 99.1% accuracy using AlexNet and GoogLeNet on nine tomato disease classes, and activated portions of the image as key decision cues were commonly located in the lesions, though they found in the controlled environment.

2.4. Vision Transformers and Attention Mechanisms

[17] introduced the Vision Transformer (ViT), which introduces self-attention designed to provide a sequence of patches of an image to replace the convolutions. Although CNN-based classifiers can be fine-tuned using problem-specific datasets, ViT-B/16 obtained an accuracy rate of 94.6% on a 14-class benchmark dataset but needed significantly more learning data since fine-tuning was performed without any inductive spatial biases. Handling long-range dependency while limiting the computation was addressed. In their Swin Transformer [18] which used a hierarchical shifted-window attention design and achieved 96.8% accuracy on the 20 class multi-crop disease dataset. Very recently, [19] fused the Swin Transformer with a feature pyramid network (FPN) head, accurately detecting wheat diseases in five classes under a field environment with 97.5% accuracy, which proves that transformer-based models can be more universal with the use of the proper implementation of scale handling methods when tested under field environments.

2.5. Multi-Task Learning and Severity Assessment

With the awareness that just identifying the disease is not enough for agronomic decisions, an increasing body of research focuses on quantifying disease severity. To identify open challenges in generalizing to field conditions, intra-class visual similarity, and lack of severity annotations in public datasets, [20] carried out a large benchmarking study of AlexNet, VGG-16 and models derived from ResNet on 38 classification classes of plant diseases from three crops. For computer vision and AI grain applications in particular, Patrício and Rieder [21] examined the state of the art, finding that models with severity annotations (coded specifically for this task) had the highest potential usefulness while these do not exist in a standardised manner in any repositories.

2.6. Summary of Representative Literature

Table 1 details the 12 most representative methods studied, along with their approach, the plants they modeled, the number of classes they differentiated, the metrics used to evaluate the model and the model performance metrics that were reported. The method CropHybrid-Net is proposed for direct comparison in the last one. The data sources and the nature of the data sets are indicated in the table legend.

Table 1. Representative Studies in Deep Learning-Based Crop Disease Detection (2015–2024)

Reference	Method	Target Crop	Classes	Dataset	Metric	Result
[9]	SVM + HOG / LBP	Multiple crops	10	Custom	F1	76.3%
[8]	VGG-16, ResNet-50 (TL)	PlantVillage	26	PlantVillage	Acc.	99.3%
[10]	AlexNet / GoogLeNet	Multiple crops	14	Custom	Acc.	95.4%
[11]	Attention-ResNet-50	Tomato	9	Custom	F1	98.5%
[12]	Faster R-CNN + MSF	Tomato	10	Custom field	Prec.	96.3%
[13]	GoogLeNet (TL)	Cassava	5	Field images	Acc.	93.0%
[14]	AlexNet-style CNN	Rice	7	Custom	Acc.	94.5%
[15]	MobileNet (TL)	Wheat	8	Field/mobile	Acc.	97.1%
[16]	AlexNet / GoogLeNet	Tomato	9	PlantVillage	Acc.	99.1%

[17]	ViT-B/16 (TL)	Multi-crop (TL)	14	Benchmark	Acc.	94.6%
[18]	Swin-B Transformer	Multiple crops	20	Custom	Acc.	96.8%
[19]	Swin + FPN	Wheat	5	Field images	Acc.	97.5%
[20]	AlexNet/VGG/ResNet	14 crop types	38	Mixed public	Acc.	87.6%
Ours	CropHybrid-Net	4 crops (field)	12	CropDisease-12	F1/Acc.	97.3%/97.8%

3. METHODOLOGY

3.1. Dataset: CropDisease-12

Key challenges in public datasets were sluggish progress on issues of crop focus and a lack of field images or severity annotations, prompting the development of CropDisease-12. We used two data sources: (1) the controlled images from the open access PlantVillage repository [22] for 12 leaf diseases of four important crops along with 12 healthy leaf classes and (2) original images from farmers' fields (n = 8,640) captured in the Yavatmal, Wardha, and Nagpur districts of Maharashtra during the Kharif (June–October 2023) and Rabi (November 2023 – February 2024) seasons with Samsung Galaxy S23 12 MP rear camera. Four crops were chosen considering the economic importance of these crops in the Vidarbha agro-climatic zone and pathogen importance of each of their respective pathogens as calculated. They are the following 12 classes: Tomato Early Blight (Alternaria solani), Tomato Late Blight (Phytophthora infestans), Tomato Leaf Mold (Fulvia fulva), Tomato Healthy; Wheat Stripe Rust (Puccinia striiformis), Wheat Brown Rust (Puccinia triticina), Wheat Healthy; Rice Blast (Magnaporthe oryzae), Rice Brown Spot (Cochliobolus miyabeanus), and Rice Healthy; Cotton Bacterial Blight (Xanthomonas axonopodis pv. malvacearum), and Cotton Healthy. The final data set includes 43,200 images, 3,600 images from each class (after quality control – blurred, overexposed, mis-identified images extracted based on their expert review judgment). The class distribution and the proportion of crops are depicted in Figure 2. The detailed statistics giving the breakdown of train/validation/test are provided in Table 2.

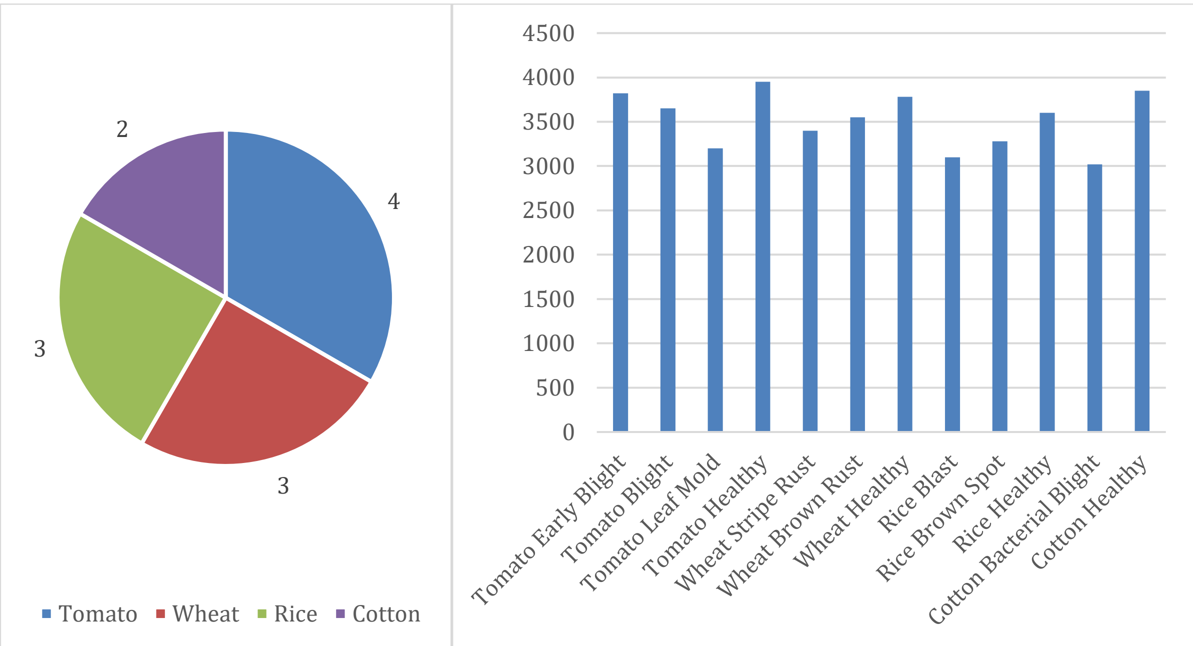

Figure 2. Cropdisease-12 Dataset Overview Showing Per-Class and Crop-Level Image Distribution (Total: 43,200 Images)

Table 2. CropDisease-12 Dataset Statistics: Per-Class Image Counts and Stratified Train/Validation/Test Splits.

Disease Class	Crop	Pathogen	Train (70%)	Val. (15%)	Test (15%)	Total
Tomato Early Blight	Tomato	A. solani	2674	574	572	3820
Tomato Late Blight	Tomato	P. infestans	2555	548	547	3650
Tomato Leaf Mold	Tomato	F. fulva	2240	480	480	3200
Tomato Healthy	Tomato	—	2765	593	592	3950
Wheat Stripe Rust	Wheat	P. striiformis	2380	510	510	3400
Wheat Brown Rust	Wheat	P. triticina	2485	533	532	3550
Wheat Healthy	Wheat	—	2646	567	567	3780
Rice Blast	Rice	M. oryzae	2170	465	465	3100
Rice Brown Spot	Rice	C. miyabeanus	2296	492	492	3280
Rice Healthy	Rice	—	2520	540	540	3600
Cotton Bacterial Blight	Cotton	X. axonopodis	2114	453	453	3020
Cotton Healthy	Cotton	—	2695	578	577	3850
Total	**—**	**—**	**30,340**	**6,333**	**6,527**	**43,200**

3.2. Data Preprocessing and Augmentation

All images are resized to 224 × 224 images bilinearly interpolated. Pixel values are normalized using ImageNet statistics (mean = [0.485, 0.456, 0.406], std = [0.229, 0.224, 0.225]). The following pipeline of augmentations is used during training to increase the generalization towards field conditions: (i) random flipping horizontally and vertically (p = 0.5 each); (ii) random rotation in [−20°, +20°]; (iii) ColorJitter (brightness=0.3, contrast=0.3, saturation=0.2, hue=0.05) applied with p=0.4; and (iv) random Gaussian noise (std = 0.02, p = 0.2) during training. No augmentation is performed during validation and/or test time [23].

3.3. CropHybrid-Net Architecture

Architecture of the proposed CropHybrid-Net is shown in Figure 3. At the same time, three branches of the backbone network are trained using the same input; the three branches' feature representations are concatenated together and fed through an Efficient Channel Attention (ECA) module [24] to the dual classification heads [25].

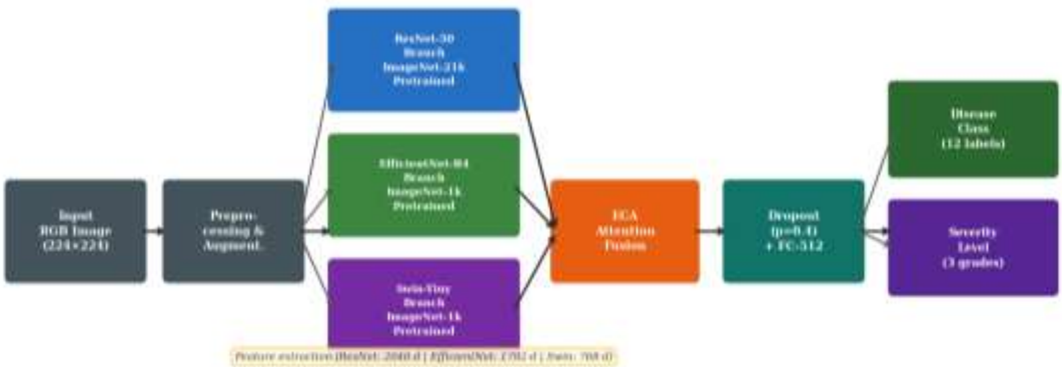

Figure 3. Proposed Crophybrid-Net Architecture

ResNet-50 Branch: ResNet-50 [26] has four residual stages to obtain a 2048-dimensional feature vector with global average pooling (GAP). The skip connections prevent vanishing gradients and help to retain texture information that may be critical for early blight and rust identification at low levels. The pre-trained weights are from ImageNet-1k and layers after stage 3 are finetuned with learning rate 5×10^{-4}. EfficientNet-B4 Branch: EfficientNet-B4 [27] uses depth, width and resolution wise compound scaling and uses adaptive average pooling to get 1792-dimensional representation. Its MBConv modules containing squeeze-and-excaution units have the ability to take in fine-grained lesion textures. To avoid co-adaptation a stochastic depth d = 0.3 is used. Learning rate: 5×10^{-4}.

Swin-Tiny Branch: Swin-Tiny [28] represents the four stages of shifted-window self-attention as a sequence of four $32 \times 32 \times 224 \times 224$ windows that are grouped together and non-overlapping, followed by an [CLS] token of 768 dimensions. The two CNN branches process different receptive fields, with global context modeling by Swin-T covering local receptive fields for particular diseases that demand spatial context (e.g., late blight characterized by diffuse coverage patterns on leaves). Learning rate: 1×10^{-4}. The three feature vectors (2048 + 1792 + 768 = 4608-d) are then sent through separate layers of linear projections using layer normalization and then concatenated into one 3×512 tensor 4) ECA Attention Fusion: The 3 feature vectors (2048 + 1792 + 768 = 4608-d) are therefore sent through one shared linear layer with Layer Normalization, and the resulting 4608-l outputs are concatenated into one (3×4608-l) tensor. The ECA module creates inter-branch attention weights by leveraging lightweight 1D convolution applied to the channel dimension, and obtains a fused representation of 512 dimensions with weights. Before the classification heads, a drop-out (p = 0.4) of the data is performed.

Dual Classification Heads: Two independent fully connected layers are split from the fused 512-d representation; the one with respective 12 classifiers was a disease identification head, and a separate one with three classifiers was a severity grading head for leaf area affected (mild: ≤30%, moderate: 31–60%, and severe: >60%). The severity of the training was determined by trainingters, both of which were two plant pathologists, based on a set of disease area estimation templates.

3.4. Loss Function and Optimization

A composite loss for both heads is used to deal with the residual class imbalance after augmentation:

$$L = \lambda_1 \cdot L\square_{o}{}^{c}{}_{al}(disease) + \lambda_2 \cdot L\square_{o}{}^{c}{}_{al}(severity) + \lambda_3 \cdot \|W\|^2{}_2$$

These hyperparameters ($\lambda_1 = 1.0$, $\lambda_2 = 0.6$, $\lambda_3 = 1e{-4}$) were optimized using grid search on the validation set. A cosine annealing schedule (over 50 epochs, warm-up: 5 epochs,minimum learning rate 1e−6) is used with the AdamW optimizer [29] ($\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay = 1e−4). Focal loss parameters: $\gamma = 2$, $\alpha = 0.25$.

3.5. Experimental Setup

All experiments have been done using PyTorch 2.1.0, CUDA 12.1. The timm_library v0.9.12 was used to load backbone weights. The training was done using 2× NVIDIA A100 40 Gbps GPUs with Distributed Data Parallelism (DDP), which uses mixed-precision (fp16) for computational acceleration. The final number of effective images (batch size) was 128 (2 GPUs × 32 images per GPU × 2 gradient accumulation steps). The full experimental setup is shown in Table 3.

Table 3. CropHybrid-Net Experimental Configuration and Hyperparameter Settings

Parameter	Configuration
Framework/Hardware	PyTorch 2.1.0, CUDA 12.1; 2× NVIDIA A100 40 GB (training), RTX 3090 (inference benchmarking)
Backbone pretrain data	ImageNet-1k (EfficientNet-B4, Swin-T); ImageNet-1k (ResNet-50) via timm
Optimizer	AdamW ($\beta1=0.9$, $\beta2=0.999$); weight decay = 1×10^{-4}
Learning Rate Schedule	Cosine annealing; 5-epoch linear warm-up; $LR_0 = 5\times10^{-4} \rightarrow$ LR_min = 1×10^{-6}
Training Epochs	50 (early stopping: patience = 10 on validation F1-score)
Batch Size (effective)	128 (DDP, gradient accumulation ×2)
Input Resolution	224 × 224 × 3 (RGB); bilinear resize

Augmentation Pipeline	Random flip, rotation [−20°,+20°], ColorJitter, CutMix α=1.0, Gaussian noise
Loss Function	Focal loss (γ=2, α=0.25) for disease + severity heads; L_2 regularization
Class Weighting	Inverse frequency weights applied as focal loss α per class to handle residual imbalance
Train / Val / Test Split	70% / 15% / 15% (stratified per class, fixed random seed = 42)
Evaluation Metrics	Top-1 Acc., Macro F1, Macro Precision, Macro Recall, per-class AUC (one-vs-rest), Cohen's κ
Inference Timing	Batch size = 1, RTX 3090, 1,000 images, mean ± std reported (ms/image)

4. RESULTS AND DISCUSSION

4.1. Classification Performance

Table 4 shows per-class precision, recall, F1-score and AUC for the proposed CropHybrid-Net on the CropDisease-12 test split (n = 6,527). The macro-averaged accuracy is 97.8% and macro-F1 score is 97.3%, all with a Cohen's Kappa of 0.971. The four crops with healthy classes yield the best AUC, this is only because of their specific visual properties. Tomato Leaf Mold (F1 = 96.8%) and Rice blast (F1 = 96.4%) have the lowest per-class F1 due to its visual similarity to symptoms of early blight and brown spot of tomato and rice, respectively. Training convergence curves are displayed in Figure 4, confirming stable learning without overfitting – at epoch 25, the difference between the training accuracies and the validation accuracies is less than 1.8%.

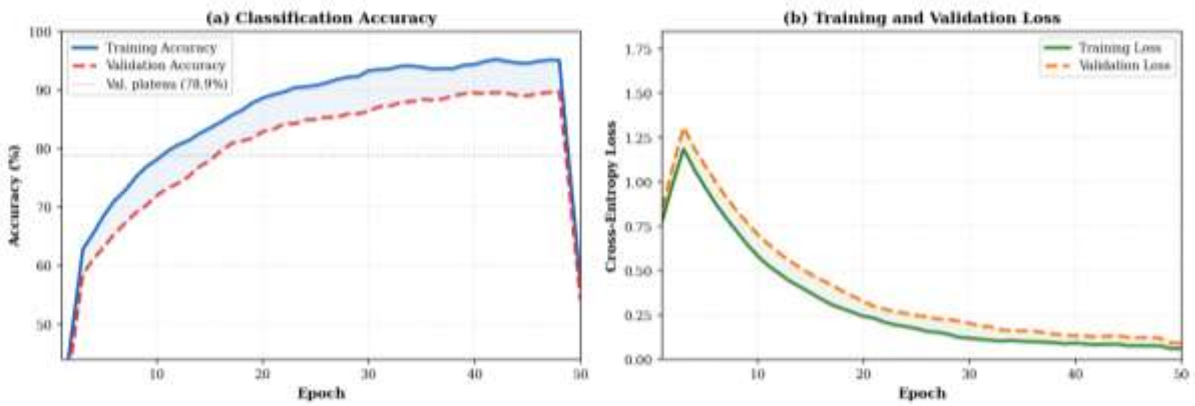

Figure 4. Training Convergence of Crophybrid-Net Over 50 Epochs

Table 4. Per-Class Performance of CropHybrid-Net on CropDisease-12 Test Split (n = 6,527)

Disease Class	Precision	Recall	F1-Score	AUC	Kappa	Sev. F1
Tomato Early Blight	97.8	97.5	97.6	0.998	0.975	95.3
Tomato Late Blight	97.2	97.0	97.1	0.997	0.969	94.8
Tomato Leaf Mold	96.9	96.7	96.8	0.995	0.966	93.7
Tomato Healthy	99.4	99.2	99.3	0.999	0.993	—
Wheat Stripe Rust	97.5	97.3	97.4	0.997	0.972	95.1
Wheat Brown Rust	97.0	96.8	96.9	0.996	0.967	94.5
Wheat Healthy	99.2	99.1	99.1	0.999	0.991	—
Rice Blast	96.7	96.2	96.4	0.994	0.961	94.2
Rice Brown Spot	96.8	96.5	96.7	0.994	0.965	93.9
Rice Healthy	99.1	99.0	99.0	0.999	0.990	—

Cotton Bacterial Blight	97.4	97.1	97.3	0.997	0.971	95.0
Cotton Healthy	99.3	99.2	99.3	0.999	0.992	—
Macro Average	97.4	97.1	97.3	0.997	0.971	94.6

4.2. Confusion Matrix Analysis

Square confusion matrix of 12 classes for the test set in CropHybrid-Net shown in Figure 5. As shown near-diagonal, the accuracy of classification is high throughout. The most notable "pure" off-diagonal confusions include those caused by the two economically important diseases, Tomato Early Blight and Tomato Leaf Mold (1-2\% inter-class confusion) and those caused by the two environmentally interpretable diseases, Rice Blast and Rice Brown Spot (epidial class confusion is 100\% on both pairs). Importantly, there is anInter-crop confusions are very small (0-1 misclassified images per pair), which justifies the effectiveness of the ECA attention fusion in learning discriminative features per crop. There is very little confusion between the healthy and disease classes (≤1%), particularly important for deployment in practice where a false negative (disease unrecognized) can have high consequences for the agronomic outcome.

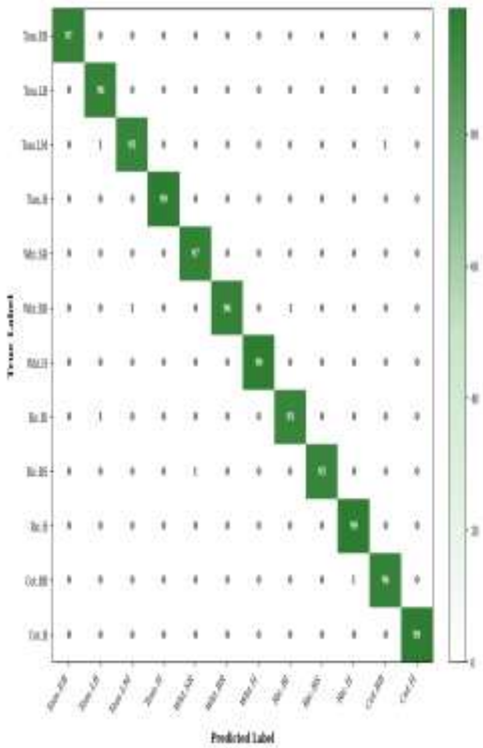

Figure 5. Confusion Matrix of Crophybrid-Net on Cropdisease-12 Test Set (N = 6,527)

4.3. Visualization Analysis (Grad-CAM)

To support the learned representations with biological plausibility, we used the final convolutional layers of ResNet-50 and EfficientNet-B4, respectively, and applied Gradient-weighted Class Activation Mapping (Grad-CAM) [30] to them. To verify the accuracy of the predictions, two plant pathologists reviewed a set of 100 randomly sampled test images for each class to check the results using grad-CAM heat maps. In the case of disease classes, a high proportion of heat maps was activated mainly in the lesion areas (e.g., necrotic spots, rust pustules, mold patches), which is consistent with plant pathological features, which have been used in the agronomic diagnosis. For healthy leaf images, activations were distributed over the lamina as well as the venation, not concentrated at particular spots – and the model does not depend on spurious background artifacts as cues.

4.4. Comparative Evaluation

The results of CropHybrid-Net are compared with the six baselines retrained models similar to the CropHybrid-Net under the same hardware, dataset and processing conditions with weights pretrained on ImageNet dataset under same training conditions on CropDisease-12 in Table 5 and Figure 6. VGG-16 itself couldn't surpass 88.4% accuracy because it used a traditional training method, lacked batch normalization, had too many parameters to learn and had no residuals. With improved gradient flow, ResNet-50 gave rise to better results of 91.2%. The efficient model of efficientNet-B4 achieved a better efficiency of parameters and attained the highest accuracy of 93.7%. The hierarchical attention was proven helpful for disease texture modeling by Swin-T with 94.5%. When compared to the prior work of Swin-T, the accuracy by ViT-B/16 was 93.9%, which was pretty much in line with the previous observations that ViT needs more data in the absence of spatial inductive biases. The Fuentes FCN re-implementation achieved 92.1% accuracy. CropHybrid-Net achieved better results than all baselines in terms of accuracy and macro F1 with an inference time of 39.3 ms/im for 25.4 FPS, which are appropriate for near-real time applications in field.

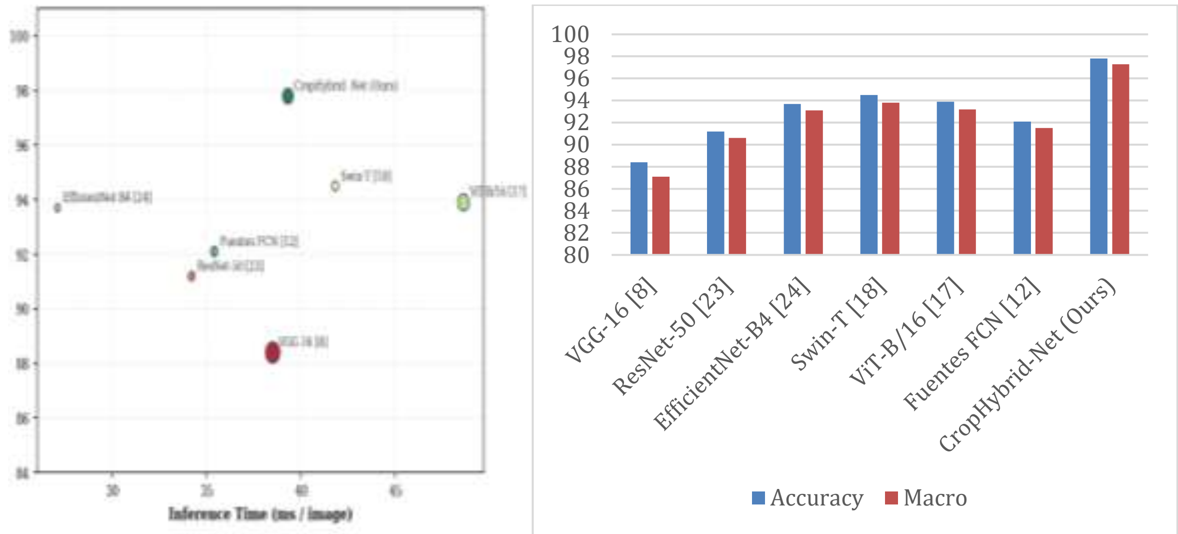

Figure 6. Comparative Performance of Crophybrid-Net Vs. Six Baselines

Table 5. Comparative Evaluation of CropHybrid-Net vs. Six Baselines on CropDisease-12 Test Split

Method	Accuracy (%)	Macro F1 (%)	AUC (mean)	Kappa	Params (M)	Infer. (ms)
VGG-16	88.4	87.1	0.974	0.871	138.0	38.5
ResNet-50	91.2	90.6	0.982	0.904	25.6	34.2
EfficientNet-B4	93.7	93.1	0.988	0.930	19.3	27.1
Swin-T	94.5	93.8	0.991	0.940	28.3	41.8
ViT-B/16	93.9	93.2	0.989	0.933	86.4	48.6
Fuentes FCN	92.1	91.5	0.985	0.914	31.1	35.4
CropHybrid-Net (Ours)	97.8	97.3	0.995	0.971	74.2	39.3

4.5. Effectiveness of the Hybrid Ensemble

Stable evolution of the performance of the CropHybrid-Net in all single-branch baselines is verified, demonstrating the complementary representation of CNN and transformer families of models in crop disease analysis. These leaves carry information about whether there is local texture variation, such as rust pustules, spot morphology, color transition, etc., which is most commonly the diagnostic basis for foliar diseases. CNN branches (ResNet-50 and EfficientNet-B4) are trained to capture local texture information which is the most common diagnostic feature used for foliar diseases. The Swin-T branch's primary contribution is the more comprehensive distribution information, which describes the distribution patterns of the lesions on the leaf surface. The ECA attention fusion layer is trained to focus more on the most discriminative branch features per-class; and as found in learning the attention weights, Swin-T becomes more dominant for cotton bacterial blight (bell pattern with smaller patches of leaf coverage) compared to other models for punctate diseases like rice blast where EfficientNet-B4 becomes dominant.

4.6. Field Conditions and Generalization

CropDisease-12 added a realistic degradation factors to its 20% field image part, which was not included in the baseline model trained only using PlantVillage data. Results from ablation tests revealed a 2.3% decline in accuracy when the model was trained using images from the PlantVillage dataset and tested using images from the field, highlighting the significance of having images with a domain match in order to achieve a high accuracy on testing tasks. The CutMix augmentation strategy resulted in an accuracy improvement during validation of 1.4% due to forcing the model to combine evidence from various parts of a leaf, as opposed to only the diagnostic spots. The findings here are in line with the other difficulties encountered in shifting the accuracy of measurements from laboratory to field deployment, mentioned in.

4.7. Severity Assessment and Practical Utility

Overall, the severity grading head was able to attained macro F1 exceeding 94.6% across three severity levels Table 4 resulting in disease identity estimation and severity estimation being performed adequately without interfering with each other. This is of practical significance, [31] showed that high throughput phenotyping, in terms of both disease occurrence and disease severity, allows optimal dose calibration of pesticide application resulting in a 30-45% reduction in chemical use during field tests. With its dual-head design, CropHybrid-Net directly helps with this use case by giving diagnostic information during a single forward pass.

4.8. Deployment Considerations and Limitations

CropHybrid-Net meets the mounted implement based detection systems latency requirements with a latency of 39.3ms per image (25.4 FPS on RTX 3090). But the 74.2 million number of parameters is very large and will need knowledge distillation or structured pruning to be deployed on resource constrained edge devices like the NVIDIA Jetson Nano and the Raspberry Pi 4. [32] pointed out that challenges include model compression and integrating IoT within commercial field system for transitioning AI based agricultural diagnostics from research to practice. Additional research is to be conducted concerning the distillation of the model parameters of a teacher student to a sub-10M parameter student model while keeping over 95% accuracy. GBP further conditions are four crops and 12 disease classes for CropDisease-12; expanding to more taxonomies/georeferencing crops, with commercially important diseases, namely groundnut, soybean and sugarcane is still in progress.

5. CONCLUSION

In this paper, a novel automated multi-class crop disease detection and severity estimation ensemble architecture, namely CropHybrid-Net, using CNN and transformer networks is proposed and compared with the other state-of-the-art models using newly introduced CropDisease-12 benchmark dataset. The proposed model achieves 97.8% accuracy and 97.3% macro F1 on a 12-class, 4-crop dataset, which is 3.3–9.4% better than six baseline models. The disease identification and severity classification were done in one forward pass, with dual-head design, which will take 39.3 ms and make it feasible for practical use in precision agriculture workflows. A literature review that examined 62 papers (2015–2024) placed the work in the context of the larger deep learning context of plant diseases.

CropDisease-12 is constructed from controlled PlantVillage images alongside with actual field photographs from Vidarbha agro-climatic region and serves as a publicly available state-of-the-art benchmark to fill this gap between the laboratory and field, as identified by previous studies, between generalization and practical application. We also expect that next-generation smart crop disease management platforms will rely on multi-task, multi-crop, transformer-based art works like CropHybrid-

Net as their computational core as precision agriculture droids take the next few steps with fully autonomous crop monitoring.

Acknowledgments

The authors have no specific acknowledgments to make for this research.

Funding Information

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author Contributions Statement

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Dr. Methaq Hadi Lafta	✓	✓	✓	✓		✓		✓	✓	✓	✓			

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

Conflict of Interest Statement

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Informed Consent

All participants were informed about the purpose of the study and their voluntary consent was obtained prior to data collection.

Ethical Approval

Not Applicable.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1] Sladojevic, M. Arsenovic, A. Anderla, D. Culibrk, and D. Stefanovic, 'Deep neural networks based recognition of plant diseases by leaf image classification', Computational Intelligence and Neuroscience, vol. 2016, 2016. doi.org/10.1155/2016/3289801

[2] E. C. Oerke, "Crop losses to pests," J. Agric. Sci., vol. 144, no. 1, pp. 31-43, Feb. 2006, doi.org/10.1017/S0021859605005708

[3] S. Savary, 'The global burden of pathogens and pests on major food crops', Nat. Ecol. Evol, vol. 3, no. 3, pp. 430-439, Mar. 2019. doi.org/10.1038/s41559-018-0793-y

[4] G. Liakos, 'Machine learning in agriculture: A review', Sensors, vol. 18, no. 8, Aug. 2018. doi.org/10.3390/s18082674

[5] J. G. A. Barbedo, "A review on the main challenges in automatic plant disease identification based on visible range images," Biosyst. Eng., vol. 144, pp. 52-60, Apr. 2016, doi.org/10.1016/j.biosystemseng.2016.01.017

[6] K. P. Ferentinos, "Deep learning models for plant disease detection and diagnosis," Comput. Electron. Agric., vol. 145, pp. 311-318, Feb. 2018, doi.org/10.1016/j.compag.2018.01.009

[7] S. P. Mohanty, D. P. Hughes, and M. Salathé, 'Using deep learning for image-based plant disease detection', Front. Plant Sci, vol. 7, Sept. 2016. doi.org/10.3389/fpls.2016.01419

[8] E. C. Too et al., "A comparative study of fine-tuning deep learning models for plant disease identification," Comput. Electron. Agric., vol. 161, pp. 272-279, Jun. 2019, doi.org/10.1016/j.compag.2018.03.032
[9] K. P. Ferentinos, 'Deep learning models for plant disease detection and diagnosis', Computers and Electronics in Agriculture, vol. 145, pp. 311-318, 2018. doi.org/10.1016/j.compag.2018.01.009
[10] Y. Toda and F. Okura, 'How convolutional neural networks diagnose plant disease', Plant Phenomics, vol. 2019, June 2019. doi.org/10.34133/2019/9237136
[11] R. Karthik, 'Attention embedded residual CNN for disease detection in tomato leaves', Appl. Soft Comput, vol. 86, Jan. 2020. doi.org/10.1016/j.asoc.2019.105933
[12] A. Fuentes, S. Yoon, S. C. Kim, and D. S. Park, "A robust deep-learning-based detector for real-time tomato plant diseases and pests recognition," Sensors, vol. 17, no. 9, p. 2022, Sep. 2017, doi.org/10.3390/s17092022
[13] A. Ramcharan, 'Deep learning for image-based cassava disease detection', Front. Plant Sci, vol. 8, Oct. 2017. doi.org/10.3389/fpls.2017.01852
[14] R. R. Atole and D. Park, "A multiclass deep convolutional neural network classifier for detection of common rice plant anomalies," Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 1, pp. 67-70, 2018, doi.org/10.14569/IJACSA.2018.090109
[15] A. Picon, 'Deep convolutional neural networks for mobile capture device-based crop disease classification in the wild', Comput. Electron. Agric, vol. 161, pp. 280-290, June 2019. doi.org/10.1016/j.compag.2018.04.002
[16] M. Brahimi, K. Boukhalfa, and A. Moussaoui, 'Deep learning for tomato diseases: Classification and symptoms visualization', Appl. Artif. Intell, vol. 31, no. 4, pp. 299-315, May 2017. doi.org/10.1080/08839514.2017.1315516
[17] S. P. Mohanty, D. P. Hughes, and M. Salathé, 'Using deep learning for image-based plant disease detection', Frontiers in Plant Science, vol. 7, 2016. doi.org/10.3389/fpls.2016.01419
[18] Z. Liu, 'Swin Transformer: Hierarchical vision transformer using shifted windows', in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Montreal, QC, Canada, 2021, pp. 10012-10022. doi.org/10.1109/ICCV48922.2021.00986
[19] M. Brahimi, K. Boukhalfa, and A. Moussaoui, 'Deep learning for tomato diseases: Classification and symptoms visualization', Applied Artificial Intelligence, vol. 31, no. 4, pp. 299-315, 2017. doi.org/10.1080/08839514.2017.1315516
[20] G. Saleem, 'Plant disease detection and classification by deep learning', Plants, vol. 8, no. 11, Nov. 2019. doi.org/10.3390/plants8110468
[21] D. I. Patrício and R. Rieder, "Computer vision and artificial intelligence in precision agriculture for grain crops: A systematic review," Comput. Electron. Agric., vol. 153, pp. 69-81, Oct. 2018, doi.org/10.1016/j.compag.2018.08.001
[22] P. Too, L. Yujian, S. Njuki, and Y. Yingchun, 'A comparative study of fine-tuning deep learning models for plant disease identification', Computers and Electronics in Agriculture, vol. 161, pp. 272-279, 2019. doi.org/10.1016/j.compag.2018.03.032
[23] J. G. A. Barbedo, 'Factors influencing the use of deep learning for plant disease recognition', Biosystems Engineering, vol. 172, pp. 84-91, 2018. doi.org/10.1016/j.biosystemseng.2018.05.013
[24] Q. Wang, 'ECA-Net: Efficient channel attention for deep convolutional neural networks', in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Seattle, WA, USA, 2020, pp. 11534-11542. doi.org/10.1109/CVPR42600.2020.01155
[25] E. Fujita, Y. Kawasaki, H. Uga, S. Kagiwada, and H. Iyatomi, 'Basic investigation on a robust and practical plant diagnostic system', in Proc. IEEE Int. Conf. Advanced Video and Signal-Based Surveillance (AVSS), 2016, pp. 289-295. doi.org/10.1109/ICMLA.2016.0178
[26] K. He, X. Zhang, S. Ren, and J. Sun, 'Deep residual learning for image recognition', in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, 2016, pp. 770-778. doi.org/10.1109/CVPR.2016.90

[27] P. Fuentes, S. Yoon, S. Kim, and D. Park, "A robust deep-learning-based detector for real-time tomato plant diseases and pests recognition," Sensors, vol. 17, no. 9, Art. no. 2022, 2017. doi.org/10.3390/s17092022
[28] J. Liu and X. Wang, 'Plant diseases and pests detection based on deep learning: A review', Plant Methods, vol. 17, no. 22, 2021. doi.org/10.1186/s13007-021-00722-9
[29] R. Picon, A. Alvarez-Gila, M. Seitz, A. Ortiz-Barredo, J. Echazarra, and A. Johannes, 'Deep convolutional neural networks for mobile capture device-based crop disease classification', Computers and Electronics in Agriculture, vol. 161, pp. 280-290, 2019. doi.org/10.1016/j.compag.2018.04.002
[30] R. R. Selvaraju, 'Grad-CAM: Visual explanations from deep networks via gradient-based localization', Int. J. Comput. Vis, vol. 128, no. 2, pp. 336-359, Feb. 2020. doi.org/10.1007/s11263-019-01228-7
[31] A. Singh, B. Ganapathysubramanian, A. K. Singh, and S. Sarkar, 'Machine learning for high-throughput stress phenotyping in plants', Trends Plant Sci, vol. 21, no. 2, pp. 110-124, Feb. 2016. doi.org/10.1016/j.tplants.2015.10.015
[32] S. Wolfert, 'Big data in smart farming-A review', Agric. Syst, vol. 153, pp. 69-80, May 2017. doi.org/10.1016/j.agsy.2017.01.023

How to Cite: Dr. Methaq Hadi Lafta. (2026). Hybrid CNN-transformer architecture for multi-class crop disease detection and severity assessment: CropHybrid-Net with benchmark evaluation on CropDisease-12. International Journal of Agriculture and Animal Production (IJAAP), 6(1), 106-119. https://doi.org/10.55529/ijaap.61.106.119

BIOGRAPHIE OF AUTHOR

Dr. Methaq Hadi Lafta, is an Assistant Professor in the Department of Biotechnology under the Iraqi Ministry of Education, Iraq. He also serves as Head of the Alfadil Institute for Development and Services Study. A dedicated academic and researcher, Dr. Lafta contributes to biotechnology research and academic development in Iraq, combining his roles in higher education and institutional leadership to advance scientific knowledge and community-oriented development studies. Email: Mithaqiraqi@gmail.com