Vol. 07, No. 01, pp. 209 –222 (2026)
ISSN: 2708-0757

JOURNAL OF APPLIED SCIENCE AND TECHNOLOGY TRENDS

www.jastt.org

Noise-Resilient Hybrid EfficientNet–Vision Transformer Framework with Adaptive Symmetric Cross-Entropy Loss for Robust Plant Disease Detection

Pradeep Gupta*, Rakesh Singh Jadon

Department of CSE, Madhav Institute of Technology & Science, Gwalior (M.P.), India, gupta.pradeep85@gmail.com, rsjadon@mitsgwalior.in

**Correspondence: gupta.pradeep85@gmail.com*

Abstract

The errors of human annotation and the noise of the environment such as lighting changes, occlusions and cluttered backdrop limit the correct detection of the plant diseases in the field condition. This study proposes a robust deep learning model that can withstand noise while remaining interpretable in controlled and noisy environments to achieve high plant disease classification. The hybrid EfficientNet-Vision Transformer (ViT) network proposed is based on an EfficientNet-B4 branch of CNN and a branch of Vision Transformer (ViT) network, which focuses on capturing fine-grained lesion features and global contextual information. A data augmentation pipeline based on CycleGAN is used to introduce field-style distortions (e.g., (lighting shifts, shadowing, debris and partial occlusions), thereby improving robustness to environmental noise, and an Adaptive Symmetric Cross-Entropy (ASCE) loss identifies and down-weights uncertain samples with normalized prediction entropy. The training is done in two phases, Stage 1 involves pretraining with clean PlantVillage images, and Stage 2 introduces increasingly noisy samples. The framework is tested in two different noise conditions, and these include the controlled synthetic label noise with PlantVillage and the real environmental noise with PlantDoc. The proposed model has an accuracy of 94.5% on the clean PlantVillage test set. It achieves 85.0% accuracy on the PlantVillage dataset under the 20% synthetic label noise protocol, outperforming ResNet-50V2 (76.5%), DenseNet-121 (78.9%), and Co-Teaching (79.5%). Macro-precision, macro-recall and macro-F1 of the model on the external PlantDoc field dataset are 0.718, 0.681, 0.681, respectively with a top-1 accuracy of 72.0, which is a manifestation of cross-domain generalization. The lesion-centric Grad-CAM images indicate that the model places emphasis on symptomatic areas of leaves and suppresses activations from background soil, shadows, and clutter. The suggested hybrid EfficientNet-ViT architecture offers, in general, a robust and explainable solution to precision agriculture and intelligent crop tracking systems that are resistant to noise.

Keywords: *Adaptive Symmetric Cross-Entropy, Deep Learning, EfficientNet, Hybrid CNN–Transformer, Noise-Robust Learning, Plant Disease Detection, Vision Transformer*

Received: October 29th, 2025 / Revised: March 10th, 2026 / Accepted: March 24th, 2026 / Online: March 27th, 2026

I. INTRODUCTION

Plant diseases are a significant risk to food security in the world, with an estimated decrease of 10-40 percent of yields in the staple crops per year[1]. This is why early and precise disease diagnosis is very critical; but, traditional diagnostics by aesthetic checking by agronomists is time-consuming and subject to mistake [2]. The recent developments in deep learning prove that image-based automated detection of plant diseases is possible [3]. As an example, convolutional neural networks such as ResNet and DenseNet have achieved over 95% accuracy on the PlantVillage benchmark and do so in controlled laboratory environments [4]. Nevertheless, it is challenging to deploy in real field settings due to the presence of multiple noise sources that have a considerable impact on the performance of the models [5].

A. The Challenge of Noisy Agricultural Data

Real-world agricultural data is inherently noisy due to two primary factors:

1) *Label Noise:* There may be misannotations caused by human error when collecting data or by misleading symptoms of a disease (ex: an infection of low grade which appears to be nutrient deficiencies). It has been demonstrated that 10% label noise can reduce CNN accuracy by 15–20 % points on plant disease classification tasks [6]. The use of the training samples that belong to the

doi: 10.38094/jastt71614

wrong classes can result in the models learning the incorrect or irrelevant features.

2) *Environmental Noise:* Field-captured images can be characterized by changing light, distracting backgrounds, occlusions (e.g. other leaves or insects) and image blur due to camera movement or bad image quality [7]. Low-light conditions and motion blur can make features less visible and shadows can blur lesion patterns. These conditions vary significantly with clean laboratory images and models that have been trained on ideal datasets fail to operate on field imagery.

In summary, deep models that are trained on clean data do not perform well in precise agriculture that combines the effect of label noise with environmental noise. Current models of deep learning are prone to overfitting clean training data and are not robust enough to handle noise variables in the real world[8] making them not practical to the farmers.

B. Limitations of Existing Solutions

The current state of deep learning has significantly improved the detection of plant diseases, and a lot of the current solutions fail to perform in noisy real-world environments. One of the biggest weaknesses is that most of the methods deal with either label noise, or environmental noise, but not both. Some methods address mislabeled data by focusing on seemingly clean samples or using co-training to screen questionable instances. Other approaches enhance resistance to environmental distortions by either normal augmentation, domain adaptation, or attention mechanisms that assist the network to concentrate on salient areas. CycleGAN generative simulation of field conditions has been investigated as well. Nevertheless, the combination of these strategies into the unified system is never implemented, creating an essential gap in the situation when both types of noise are present at the same time in the field-collected agricultural images.

Meanwhile, hybrid CNN-Transformer models have performed well in medical imaging and remote sensing among other applications such as local texture sensitivity and global context is needed. Even though these models are conceptually well adapted to recognize plant diseases in a wide variety of field conditions, they are yet to be proven in agricultural environment with realistic noise. Existing studies typically rely on curated, noise-free datasets and therefore fail to demonstrate robustness.

By highlighting these shortcomings, it is imperative that a single framework is created that allows both label and environmental noise to be tackled in conjunction with exploiting identical operating domains that CNNs and Transformers can bring. This problem inspires us to design a proposed noise-resilient adaptive loss learning hybrid architecture.

C. Our Contributions

In this paper, the robust detection of plant diseases under noisy conditions is introduced as a unified framework. The main contributions are:

1) *Hybrid EfficientNet–ViT Architecture:* A novel CNN-Transformer architecture is proposed in which a CNN backbone (EfficientNet-B4) and a ViT backbone are combined to capture features of local lesions and global context respectively and achieves higher accuracy under occlusions and complex backgrounds.
2) *Realistic Noise Simulation:* A Noise injection Pipeline With class-dependent Label flip and CycleGAN augmentation of images to simulate the real-world field condition (including shadows, debris and lighting effect).
3) *Adaptive Symmetric Cross-Entropy (ASCE) Loss:* Noise-robust Loss Function, Adaptive Symmetric Cross-Entropy Loss is a loss function that achieves dynamic down weighting of mislabeled samples or ambiguous samples, avoiding overfitting problem and such noisy label problem.

By co-optimizing architecture, data simulation, and loss design, the proposed framework improves robustness to both label noise and environmental noise, helping bridge the gap between laboratory datasets and field deployment.

D. Impact and Outcomes

The proposed framework was evaluated on the PlantVillage dataset under clean and synthetic-noise protocols, with external validation on PlantDoc, and showed strong performance in multiple dimensions:

1) *State-of-the-Art Accuracy:* The hybrid EfficientNet-ViT model accuracy on clean test data is 94.5% and it outperforms previous CNN-based features on PlantVillage where ResNet and DenseNet achieved an accuracy between 92-94%.
2) *Noise Robustness:* With a total combined noise of 20% (label + image noise), the model maintained an accuracy of 85.0% outperforming Co-Teaching by 5.5 percentage points (79.5%), DenseNet-121 by 6.1 points (78.9%), and ResNet-50V2 by 8.5 points (76.5%) that is important to minimize missed detection in field diagnostics.
3) *Generalization and Interpretability:* The proposed solution exhibited fewer performance degradations across the clean and noisy conditions in comparison to baselines, and implying better extrapolation to real-life scenarios. Furthermore, Grad-CAM visualizations confirmed that the model concentrates on disease lesions over irrelevant background features even in heavy noise conditions, which enables to build trust into the model and thus its practical implementation in the agricultural setting.

The results obtained show the utilization of the framework being practically prepared for real-life application scenarios, where data may be noisy or missing. By closing the performance gap between the lab and the field, this approach will also help to advance scalable and reliable rainfed precision-agriculture and sustainable intensified crop management AI-driven decision tools.

E. Paper Organization

The remainder of this paper is organized as follows:

- Section 2 reviews related work on plant disease detection, noise-robust learning, and hybrid CNN–Transformer models, highlighting the research gaps addressed by our approach.
- Section 3 describes the proposed methodology, including the hybrid architecture, noise simulation pipeline, loss function, and training strategy.
- Section 4 presents experimental results on the augmented PlantVillage dataset, comparing our method with state-of-the-art models and providing ablation studies to evaluate individual components.
- Section 5 discusses key findings, practical deployment implications, limitations, and directions for future work, such as real-field testing and model compression for edge devices.
- Section 6 concludes the paper with a summary of contributions and outlines potential avenues for further research.

II. Related Work

In this section, we review related literature across three fields has been presented (1) deep learning to detect plant diseases, (2) learning with noisy data and (3) CNN-Transformer hybrid models. We also provide the summary of some representative previous works in Table I and mark the difference in our work, because we consider the joint problems.

A. Plant Disease Detection with Deep Learning

Initial experiments showed that CNNs can be used in the recognition of plant diseases on the PlantVillage dataset. To take one example, an AlexNet model obtained 99.35 per cent accuracy on laboratory-captured leaf images [4] . The same (~98-99% accuracy) was obtained with more sophisticated networks including VGG and Inception-v3 [9]. The works have made CNNs (frequently with transfer learning) effective plant pathology methods with controlled settings. Later studies were aimed on better efficiency and generalizability. ResNet was augmented with skip connections[10], and DenseNet was subsequently suggested using feature reuse to allow very deep models to be trained without gradient degradation problems[11]. These were greater than 95% accurate on PlantVillage, but performed poorly on field data because of the noise problems mentioned above. EfficientNet was suggested as a scale-out architecture with the help of compound scaling to reach the state-of-the-art accuracy at a reduced number of parameters[12]. Variants that are based on EfficientNet (B4, B5, and others) also work very well on PlantVillage, and, similar to other CNNs, they do not have noise robustness mechanisms and decline in accuracy dramatically on actual field images.

The practical difficulty was the training of models over the datasets in the field after training them over PlantVillage. As an example, models that correctly predict on PlantVillage with more than 95 percent accuracy usually tend to reduce to about 60–70% on the field-based PlantDoc sample [13]. A new ViT-based classifier with PlantVillage training obtained only approximately 68% on PlantDoc, and over 95% on PlantVillage[14] ,underscoring the weak performance of current models under domain shift and real-world noise.

B. Noise-Robust Learning Techniques

1) *Label Noise Mitigation:* There are various methods that have been designed to train deep networks using noisy labels. Such approaches to sample selection as MentorNet[15] and Co-Teaching rely on auxiliary networks or peer networks to detect and deemphasize potentially mislabeled samples during training. As one instance, Co-Teaching is capable of accommodating up to 45 percent noisy labels by discarding the top-loss samples of its peer per epoch[16]. Nevertheless, these methods presuppose noise to be largely random. Loss correction techniques change the loss function to a stronger one. Generalized Cross-Entropy (GCE) employs a loss which switches between mean absolute error and cross-entropy, which makes it less sensitive to outliers. Symmetric Cross-Entropy (SCE) is a cross-entropy method that includes a reverse cross-entropy term in order to penalize overconfidence. SCE had proven to perform better than a lot of procedures when noise was high on benchmark data [17]. A proposed ASCE is an evolution of SCE with the introduction of adaptive weighting per sample, which essentially applies a data-driven curriculum.

2) *Environmental Noise Handling:* Data augmentation is still one of the major methods to enhance resistance to image noise. In [18], a complete overview of data augmentation methods, such as simple flips and rotations, addition of noise or masking of parts of the image is provided. Simple transformations are difficult to perform in extreme field conditions (e.g., heavy occlusion, strong changes in color, etc.) and augmentations can be used to enhance invariance to small deformations. Generative approaches like GANs have been used to create more realistic training samples; CycleGAN, for instance, can translate images between domains (healthy ↔ diseased, or lab ↔ field style) [19]. CycleGAN-based augmentation has been used to improve cross-species disease recognition by introducing new "styles" of imagery [20]. Attention mechanisms can also help: Squeeze-and-Excitation (SE) layers [21] and Convolutional Block Attention Module (CBAM)[22] enable CNNs to focus on important regions (e.g., lesions) while suppressing background noise. These help under mild noise, but under heavy occlusion or large distribution shifts, they have limited effect. Notably, most existing methods tackle either label noise or image noise, but not both simultaneously. An initial attempt to address both label and environmental noise in agriculture used a robust CNN, but the model still experienced an accuracy drop of over 20% under heavy noise [5] . This highlights the need for integrated approaches like ours.

C. Hybrid CNN–Transformer Models

Transformers capture long-range dependencies but are computationally expensive and data-hungry. Hybrid CNN–Transformer architectures aim to combine the local feature extraction strength of CNNs with the global context modeling of Transformers [23]. Examples include TransUNet in medical imaging [24], ResNet–Swin fusion in remote sensing [25], and Conformer, which explicitly couples convolutional local features with transformer-based global representations [26].

In agricultural disease detection, hybrid models are only beginning to emerge. At the same time, standalone ViT models and advanced noisy-label learning approaches, such as DivideMix [27] and Meta-Weight-Net [28], have demonstrated strong robustness in general vision benchmarks, motivating broader comparative evaluation in agricultural settings. Building on this direction, a ConvNet–ViT hybrid model achieved 99.29% accuracy for multi-crop leaf disease classification on a dataset containing banana, cherry, and tomato [29], highlighting the potential of CNN–Transformer fusion on curated datasets. A MobileViT-based hybrid model with only 0.69 million parameters further achieved 80–99% accuracy across different crop-disease datasets [30], demonstrating the feasibility of lightweight hybrid architectures. Likewise, EConv-ViT, which integrates ConvNeXt, ViT, and efficient channel attention, reported nearly 99% accuracy for apple leaf disease classification [31]. Nevertheless, these studies did not explicitly assess robustness under noisy labels, severe occlusion, or field-domain shift.

D. Comparative Analysis of Prior Work:

Table I summarizes representative studies relevant to our problem, including classical deep learning models for plant disease detection, noise-handling techniques, and recent hybrid architectures. We highlight each method's strengths and limitations in the context of noisy, real-world plant disease data.

TABLE I. REPRESENTATIVE RELATED WORK ON PLANT DISEASE DETECTION AND ROBUST LEARNING.

Study	Methodology	Strengths	Limitations
[2]	Lesion-specific CNN classification	Focus on real field conditions (some field images used)	Very small dataset; no specialized noise mitigation
[4]	AlexNet on PlantVillage (lab images)	High accuracy on clean data (99.3%)	Not tested under noise or field conditions
[5]	Noise-resilient CNN (plants)	Addresses both label & image noise (first attempt in agriculture)	Limited robustness gains (e.g., still >20% accuracy drop under heavy noise)
[16]	Co-Teaching (dual-network training)	Robust to high label noise (tested up to 45% noise)	Assumes noise is uniformly random across classes
[20]	CycleGAN based	Generates realistic image perturbations (e.g., weather effects)	Alters image style but not disease content (risk of introducing artifacts)
[9]	VGG & Inception-v3 on PlantVillage	Early use of deep learning in agriculture; ~98–99% on lab data	Limited to controlled images (lab settings)
[17]	Symmetric Cross-Entropy (SCE) loss	Theoretically robust to label noise; improved noisy-label accuracy on CIFAR	Requires careful hyperparameter tuning (loss term weights)
[24]	(CNN+ViT) for medical segmentation	Leverages global context to handle occlusions/artifacts	Focused on medical imaging; not designed for plant disease data
[25]	ResNet + Swin Transformer (remote sensing)	Fuses local & global features for multi-modal data	Computationally intensive (large model, slow inference)
[29]	EfficientNet + ViT hybrid (multi-crop)	Top accuracy (~99.3%) on curated multi-species data	Tested only on lab images; unknown robustness to noise
[30]	*MobilePlantViT* (MobileViT-based hybrid)	Extremely lightweight (0.69M params); 80–99% accuracy on various crops	Slightly lower accuracy on some datasets; not evaluated with noisy labels or field data
[31]	*EConv-ViT* (ConvNeXt + ViT + ECA)	Captures local/global features; ~99% accuracy on apple leaf dataset	High model complexity; specialized to a single crop; not tested under combined noise
[32]	Noise-robust training (medical images)	Handles class-dependent label noise (domain-tuned)	Not validated on agricultural data (domain shift)

As seen above, many prior models excel either on clean data or in handling one type of noise, but no single approach addresses the simultaneous presence of label noise and environmental noise in plant disease recognition. This gap in the literature motivates our integrated approach.

E. Research Gaps and Our Position

Building on prior advances and addressing their limitations, our framework simultaneously addresses the following challenges:

1) *Unified Noise Handling:* Rather than addressing label noise and environmental noise separately, we handle both together in one unified approach. Our training data simulation injects simulated realistic label flips and visual occlusions to teach the model how to cope with the combination of both.
2) *Hybrid Architecture Flexibility:* We do not have rigid single-model approach (pure CNN or pure Transformer). The EfficientNet-ViT hybrid combines the strengths of both: the fine-grained details of the lesion observed by EfficientNet (which could be lost data-hungry ViT) while the Vision Transformer injects global reasoning which a pure CNN lacks. This combination is particularly powerful in situations where noisy contains signals that can only be

disambiguated using broader context beyond a corrupted local region.

3) *Realistic Noise Simulation:* Unlike prior approaches that assume simplistic noise (e.g., uniform-random label flips or small perturbation-based augmentations), we introduce class-dependent label noise to simulate realistic annotation errors and employ CycleGAN-based image translation to create field-style distortions (weather effects, cluttered backgrounds, and illumination shifts). This provides a training distribution that more closely matches images encountered in the wild.
4) *Adaptive Loss for Noisy Labels:* Our ASCE loss is an extension of symmetric cross-entropy where we adaptively down-weight samples with high entropy (uncertain) on a per-sample basis. This effectively creates a curriculum in which samples that the model is confident on (with high confidence) carry more weight than samples that have been incorrectly learned (possibly incorrectly labeled). It thus provides a mechanism to deal with the label noise without extra teacher model or explicit noise estimation.

In the following sections, we detail how these ideas are implemented (Section III) and empirically evaluated (Section IV).

III. Methodology

Our framework consists of three tightly coupled blocks: (i) a hybrid CNN-Transformer classifier (EfficientNet-B4 + ViT), (ii) a noise-injection pipeline for label and environmental/image noise, and (iii) a noise robust ASCE loss.

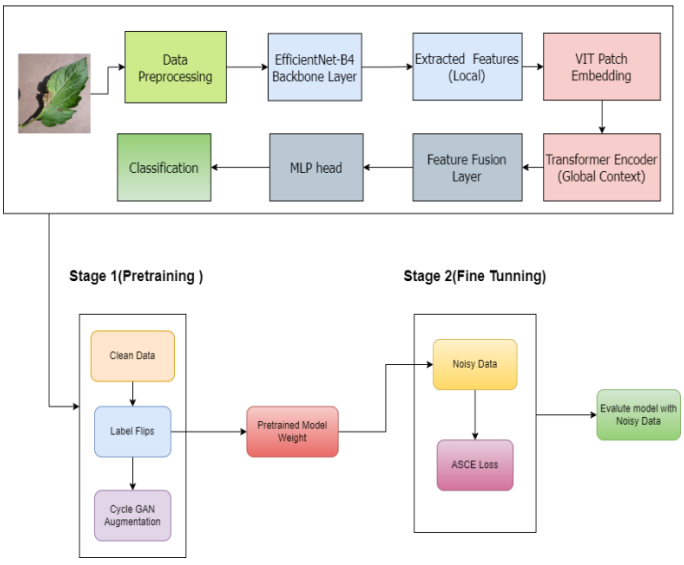

Fig. 1. Architecture of the proposed hybrid model and training pipeline

The overall training flow is shown in Fig. 1, while the detailed procedural steps are outlined in Algorithm 1. We first detail the architecture (Section III.A), then the noise simulation (Section III-B), followed by the loss and training strategy in Sections (III C through III E). Core definitions appear in Eqs. (1)– (5), which are explicitly referenced in the text for clarity.

A. Hybrid EfficientNet-ViT Architecture

Our model (Fig. 1) combines an EfficientNet-B4 backbone with a lightweight ViT encoder to leverage local lesion cues (CNN) and global spatial context (Transformer).

1) *Efficient Net Backbone:* Given an input image $X \in \mathbb{R}^{H\times W\times 3}$, EfficientNet-B4 produces a final feature map F_{cnn} capturing local texture and edge patterns critical for disease identification[33]. We denote this mapping in Eq. (1) and refer back to it when fusing features:

$$F_{\text{cnn}} = f_{\text{EffNet}}(X) \in \mathbb{R}^{h\times w\times C} \quad (1)$$

Eq. (1) defines the CNN feature tensor that later provides the local descriptor for fusion (see Eq. (4)).

2) *Vision Transformer (ViT) Encoder:* We flatten the spatial grid of F_{cnn} into $N = h \cdot w$ vectors $\{p_j\}_{j=1}^{N}$, each $p_j \in \mathbb{R}^C$. Each vector is linearly projected to dimension D and augmented with a positional encoding $\mathbf{e}_j$, yielding the ViT input tokens $z_j^{(0)}$ as in Eq. (2):

$$z_j^{(0)} = W_e\, p_j + b_e + \mathbf{e}_j, \quad j = 1,\dots,N, \quad W_e \in \mathbb{R}^{D\times C} \quad (2)$$

These tokens are processed by L Transformer layers. For one attention head, the core operation is given by Eq. (3):

$$\text{Attention}(Q,K,V) = \text{softmax}\left(QK^{\top}/\sqrt{D}\right)V \quad (3)$$

where Q, K, V are learned linear projections of the token sequence. Eq. (2) formalizes tokenization of CNN features for the ViT, and Eq. (3) is the attention kernel that equips the model with global reasoning over distant leaf regions[34].

3) *Feature Fusion:* To combine the complementary local and global features extracted by EfficientNet-B4 and ViT, the CNN feature map is first compressed using global average pooling (GAP), while the [CLS] token from the ViT encoder is used as the transformer representation. Since these two feature vectors lie in different latent spaces and may have different dimensionalities, each is projected into a common fusion space of dimension d_f using two learnable linear projection heads, ϕ_{cnn} and ϕ_{vit}. The aligned features are then merged through a learnable convex combination controlled by a scalar weight $\gamma \in [0,1]$ initialized to 0.5, so that both branches contribute equally at the start of training. The fused representation is defined as

$$\mathrm{f}_{\text{fusion}} = \gamma\,\phi_{\text{cnn}}(\text{GAP}(F_{\text{cnn}})) + (1-\gamma)\,\phi_{\text{vit}}(\mathrm{z}_{\text{cls}}) \in \mathbb{R}^{d_f} \quad (4)$$

A shallow MLP classification head then maps $\mathrm{f}_{\text{fusion}}$ to the final class logits, followed by a softmax layer. This design enables adaptive balancing of fine-grained local texture information from EfficientNet-B4 and global contextual information from ViT during training[35].

B. Noise Simulation

We emulate real-world data challenges via label noise (misannotations) and environmental/image noise (lighting, blur, occlusions). This yields training distributions closer to field conditions while keeping validation/test sets clean for fair model selection.

1) *Label Noise:* We synthesize two complementary label-noise processes:

   a) *Uniform noise:* a fraction η_{uni} of labels are flipped uniformly at random to any of the remaining classes (simulating non-systematic annotation errors).

   b) *Class-dependent noise:* for visually confusable pairs (e.g., tomato early vs. late blight), a fraction η_{cls} is flipped specifically to the paired class (simulating systematic human confusion).

Applying class-dependent flips first and uniform flips second avoids double rewrites. The overall noise budget is summarized in Eq. (5) (approximate, since processes are applied sequentially):

$$\eta_{\text{total}} \approx \eta_{\text{uni}} + \eta_{\text{cls}} \qquad (5)$$

Eq. (5) defines the target noise level used to schedule our curriculum in Stage-2 training; we instantiate $\eta_{\text{total}} \in \{10\%, 15\%, 20\%\}$ in experiments[6].

2) *Environmental Noise:* We introduce a variety of image perturbations to mimic field conditions:

   a) *GAN-Based Occlusions:* We trained a CycleGAN on unpaired PlantVillage and PlantDoc images to translate clean PlantVillage images into a field-like style while preserving disease morphology. The generator used residual blocks with cycle-consistency and identity losses, and generated images were visually screened to remove obvious artifacts before inclusion in the training pool. The final translated samples introduced realistic background clutter, shadow patterns, illumination changes, and partial occlusions that were difficult to mimic with basic pixel-level augmentation alone[20]. To improve field realism, the CycleGAN translation model was trained on unpaired PlantVillage and PlantDoc images, allowing the generator to map clean laboratory-style images into a field-like domain without requiring one-to-one correspondence. Cycle-consistency and identity constraints were used to preserve disease morphology while transferring realistic environmental characteristics such as cluttered backgrounds, illumination shifts, shadow patterns, and partial occlusions. To reduce the risk of introducing synthetic artifacts, translated images were visually screened before inclusion in the training pool. Although this procedure does not fully replace true field-data collection, evaluation on the PlantDoc dataset provides partial validation that the augmented samples capture meaningful aspects of environmental noise encountered in real agricultural imagery.

   b) *Traditional Augmentation:* In addition to GAN output, we apply conventional random augmentations: (a) Lighting variations: random brightness ±50% and contrast ±30%; (b) Blur: Gaussian blur with σ = 1–3; (c) Occlusion patches: random rectangular masks covering 10–30% of the image area (to simulate occluding objects like other leaves). These augmentations are applied stochastically each epoch to augment the training data beyond the fixed GAN-generated set. Table II provides a summary of the noise types and parameters, including references where similar settings were used.

TABLE II. SUMMARY OF NOISE TYPES AND PARAMETERS .

Reference	Noise Type	Parameters
[5]	Label Noise (Uniform)	10%, 15%, 20% random label flips
[32]	Label Noise (Class-Dependent)	5% targeted flips between confusable classes
[18]	Lighting/Blur (Augmentation)	Δ brightness = ±50%, Δ contrast = ±30%; Gaussian blur σ=1–3
[20]	Occlusions (GAN-based)	CycleGAN-generated field-style images; + 10–30% area masks

All noise injections are applied only to the training set to simulate a noisy training scenario. The validation set remains clean for model selection, and the test set is primarily clean.

C. Adaptive Symmetric Cross-Entropy (ASCE) Loss

Standard cross-entropy (CE) loss tends to overfit mislabeled samples, as incorrect labels generate disproportionately high loss values, biasing the gradient and leading to poor generalization. To address this, we introduce Adaptive Symmetric Cross-Entropy (ASCE), which integrates a symmetric cross-entropy component with dynamic sample weighting to reduce the effect of noisy labels[17].

The ASCE loss for a single training sample i is defined in Eq. (6):

$$L_{\text{ASCE}}(i) = w_i[L_{\text{CE}}(p_i, y_i) + L_{\text{CE}}(y_i, p_i)] \qquad (6)$$

where:

1. p_i represents the predicted class probability vector for sample i,
2. y_i is the corresponding one-hot encoded ground truth label,
3. $L_{\text{CE}}(p_i, y_i)$ is the standard cross-entropy loss,

4. $L_{\mathrm{CE}}(y_i, p_i)$ is the reverse cross-entropy term, penalizing overconfident predictions.

Eq. (6) combines both forward CE and reverse CE, ensuring the network learns from clean samples while suppressing noise-induced gradients.

To dynamically regulate the contribution of each sample, we define the weight w_i based on the normalized prediction entropy$H(p_i)$, which quantifies uncertainty in Eq. (7):

$$H(p_i) = -\sum_{c=1}^{C} p_{ic} \log(p_{ic}) \qquad (7)$$

where C is the number of disease classes and p_{ic} is the predicted probability for class c.

Prediction entropy is used as a principled measure of model uncertainty, since confidently learned samples typically produce sharp class-probability distributions with low entropy, whereas mislabeled, ambiguous, or environmentally corrupted samples tend to yield flatter distributions with higher entropy.

A high entropy $H(p_i)$ indicates model uncertainty (possibly due to noisy labels), so such samples should be down-weighted. We normalize the entropy by log(C), where C is the number of classes, so that the uncertainty term lies in [0,1] regardless of class count. We then define the weight as an exponential decay of the normalized entropy, as shown in Eq. (8).

Using normalized entropy addresses the scale-sensitivity issue directly: without normalization, entropy magnitude depends on the number of classes, whereas the normalized form yields a consistent uncertainty range and makes the exponential weighting parameter more interpretable across datasets.

$$w_i = \exp\big(-H(p_i)\big) \qquad (8)$$

To maintain stability, w_i is normalized at the batch level:

$$\widetilde{w}_i = \frac{w_i}{\frac{1}{N}\sum_{j=1}^{N} w_j} \qquad (9)$$

where N is the batch size. Eq. (8) ensures noisy or ambiguous samples (high $H(p_i)$) have reduced influence, while Eq. (9) normalizes the weights to prevent instability.

The final ASCE loss over a mini-batch of size N is given by:

$$\mathcal{L}_{\mathrm{ASCE}} = \frac{1}{N}\sum_{i=1}^{N} \widetilde{w}_i \left[L_{\mathrm{CE}}(p_i, y_i) + L_{\mathrm{CE}}(y_i, p_i)\right] \qquad (10)$$

Intuition and Usage

1) Low-entropy samples (high confidence, likely correct labels) are assigned higher weights, driving learning.
2) High-entropy samples (ambiguous or mislabeled) receive lower weights, reducing their gradient contribution.
3) The ASCE loss generalizes SCE by making it sample-adaptive, improving performance in datasets with non-uniform, class-dependent noise.

Algorithm 1: Training the Hybrid EfficientNet–ViT Model under Noisy Conditions

Input:

$D_{clean} = \{(x_i, y_i)\}$: Clean training dataset
D_{noisy}: Noisy training dataset generated from D_{clean} using noise injection (Section 3.2)
Pre-trained EfficientNet-B4 weights
Hyperparameters: learning rate, batch size, noise schedule parameters

Output:

Trained and Evaluate hybrid model robust to noisy conditions

// ---------- **Stage 1: Pretraining on Clean Data** ----------//
1. Initialize hybrid model:
EfficientNet-B4 backbone (pre-trained on ImageNet)
Vision Transformer encoder (random initialization)
2. **For** epoch = 1 to E_1 (pretraining epochs) do:
2.1 **For** each batch B ⊂ D_{clean} do:
Compute predictions p = model(x) for all x in B.
Compute standard cross-entropy loss L_{ce} on batch B.
Update model parameters using AdamW optimizer.
End for
2.2 Validate model on clean validation set.
End for
3. Save pre-trained model weights.
// ---------- **Stage 2: Fine-Tuning on Noisy Data** ----------//
4. Load pre-trained weights from Stage 1.
5. For epoch = 1 to E_2 (fine-tuning epochs) do:
5.1 Gradually increase noise level η (curriculum learning schedule)
5.2 **For each** batch B ⊂ D_{clean} do:
Compute predictions p = model(x) for all x in B.
Compute ASCE loss L_{ASCE} using Eq. (10).
Identify top K% samples with highest loss (likely mislabeled).
Exclude these high-loss samples from backpropagation.
Update model parameters using remaining samples.
End for
End for
6. Save final fine-tuned model weights.
// ---------- **Stage 3: Evaluation** ----------//
7. Evaluate the final trained model on the **clean test set**, reporting classification accuracy, precision, recall, and F1-score.
8. Evaluate the model under **noisy test scenarios** (e.g., η=20%\eta = 20\%η=20%) and report robustness metrics such as accuracy drop and noise tolerance.
End Algorithm

During Stage 1 pretraining (on clean data), we use standard CE. During Stage 2 fine-tuning (on noisy data), we switch to ASCE (Eq. 10).

D. Training Strategy

To improve robustness, we adopt a two-stage training strategy complemented by curriculum-style noise scheduling and regularization that is sensitive to localized corruption.

1) *Stage 1: Pretraining on Clean Data* The hybrid Model (EfficientNet-ViT) is initially trained on the cleaned PlantVillage data with normal cross entropy (CE) loss. At this stage, the network can be trained on basic disease-discriminative features in a noise-free environment, based mostly on CNNs-based local features.
2) *Stage 2: Fine-Tuning on Noisy Data* Stage 1 weights are used for the initialization of fine-tuning. We then add noise in such a way that starting from a lower noise rate (η=10%), we slowly increase the noise rate towards the target rate (η=20%) in the first few epochs (a type of curriculum learning). In this stage, the ASCE loss (Eq. 10) is used to achieve robustness with respect to label.

This two-stage procedure is necessary because direct training on severely corrupted data resulted in unstable optimization during the initial epochs, whereas pretraining on clean data allows the model to first learn disease-discriminative representations before undergoing robust adaptation. The ablation results in Section 4 support this design choice by showing that robustness emerges from the combination of clean feature learning and noise-aware fine-tuning. Following the intuition of Co-Teaching, we further exclude the top 20% highest-loss samples within each mini-batch from gradient updates. Although this strategy reduces the influence of likely mislabeled samples and extreme outliers, it may also temporarily exclude genuinely difficult but correctly labeled hard positives. ASCE mitigates this issue because mislabeled samples generally maintain higher normalized entropy for longer training periods, whereas correctly labeled hard samples gradually become more confident. Consequently, informative hard samples are less likely to remain in the high-loss subset and can re-enter training as optimization stabilizes.

To improve generalization under noisy conditions and to mitigate overfitting, we employ two regularization techniques in the training process:

a) *Stochastic Depth:* Each Transformer layer has a 10% chance of being skipped (dropout) during each forward pass, acting as a structural regularizer and encouraging resilience to missing information.
b) *DropBlock:* We apply DropBlock (with a 3×3 block, 10% drop rate) to CNN feature maps, randomly masking contiguous regions. This forces the model to rely on distributed cues rather than any single localized feature, effectively simulating random occlusions in training images.

E. Implementation Details

The key hyperparameters, hardware configuration, and evaluation strategy used in this study are summarized in Table III. Unless otherwise stated, all methods were evaluated using the same train/validation/test split for fair comparison.

TABLE III. IMPLEMENTATION DETAILS OF THE PROPOSED HYBRID EFFICIENTNET-B4 + VIT MODEL.

Parameter	Details
Optimizer	AdamW with initial learning rate 3×10^{-4}, cosine decay schedule, and weight decay of 0.05
Training Epochs	100 total (60 for Stage 1 pretraining, 40 for Stage 2 fine-tuning)
Hardware	Google Colab environment with NVIDIA A100 GPUs, batch size of 32 per GPU (effective batch size = 128)
Evaluation Metrics	Macro-averaged Accuracy, Precision, Recall, and F1-Score to evaluate class-wise robustness
Interpretability	Grad-CAM visualizations highlight lesion regions, validating model focus and interpretability (Fig. 7)

We use AdamW as optimizer and apply a cosine learning rate decay. Early stopping on the clean validation set is employed to prevent overfitting. At the end of training, we select the model with the best validation performance for final testing.

IV. EXPERIMENTS

This part compares the proposed hybrid EfficientNet ViT with ASCE loss to simulated noisy conditions and compares it to powerful CNN and noise-robust baselines. PlantVillage experiments employ synthetic noise injection, and an external evaluation on PlantDoc is conducted to assess cross-domain generalization under real environmental noise.

A. Dataset

The PlantVillage data set has 54304 high-resolution images of 38 classes (crop disease combinations), both healthy and diseased leaf samples. Captured images have uniform backgrounds, little noise in the environment, and thus it is a perfect clean benchmark to controlled experiments [36]. The PlantVillage data is represented by samples as indicated in Fig. 2.

The dataset was split into:

1) *Training Set:* 43,444 images (80%) -used to pretrain Stage 1 and fine-tune Stage 2 (noise injection added to Stage 2).
2) *Validation Set:* 5,430 images (10%)- used to tune the hyperparameters and early stop; kept clean (no noise introduced).
3) *Test Set:* 5,429 (10%)- Final Evaluation: held out. We test the model on this test set in clean conditions and whether synthetic noise is added to test the model in robustness mode.

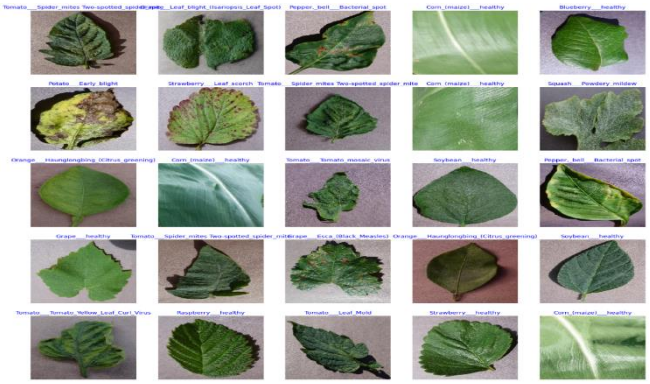

Fig. 2. Sample images of Plant Village dataset

We also used the PlantDoc data set to assess the generalization of the model to the real world context: this data set includes 2,598 field-collected images, involving 13 plant species, 27 classes (17 diseases and 10 healthy classes). PlantDoc images also have natural noise of the environment, not like PlantVillage, i.e. different lighting, occlusions, and complicated backgrounds, which gives a real-world cross-domain test with a realistic benchmark. The external validation was done using this dataset only to determine the strength of the proposed model in adverse field conditions. Fig. 3 shows representative images from the PlantDoc dataset, highlighting the complexity of field environments.

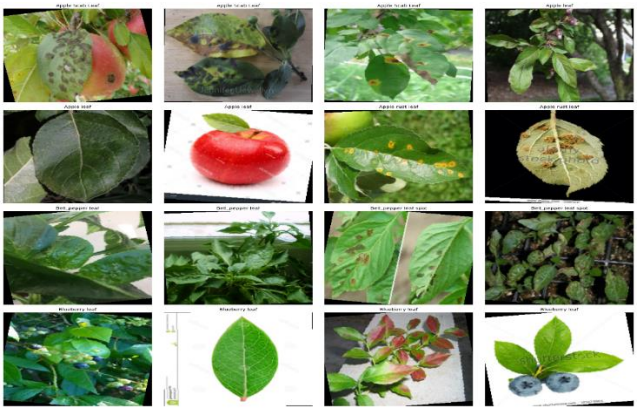

Fig. 3. Sample images of PlantDoc dataset

B. Baseline Models

As shown in Table IV, we selected four baseline models, including conventional CNNs (ResNet-50V2, DenseNet-121, and EfficientNet-B4) and a noise-robust approach (Co-Teaching), to comprehensively evaluate the performance of our proposed method [10], [11], [12], [16].

TABLE IV. DESCRIPTION OF BASELINE MODELS USED FOR COMPARISON.

Model	Description
ResNet-50V2	Standard CNN baseline, strong performance on PlantVillage.
DenseNet-121	Dense connectivity CNN with high accuracy for leaf diseases.
EfficientNet-B4	Standalone CNN backbone of our hybrid model.
Co-Teaching	State-of-the-art label noise handling, trains two ResNet-50 models simultaneously, dropping top 20% highest-loss samples per batch.

C. Results and Analysis

1) *Performance Under Increasing Noise:* Table V reports classification accuracy under clean conditions and three levels of synthetic label noise (10%, 15%, and 20%). The original test set was kept clean for standard evaluation, while a duplicated copy was used only for robustness analysis. Synthetic label flips were applied to this duplicate to assess performance under controlled noisy-label conditions without affecting the integrity of the main test set. All models were trained with the corresponding noise levels using the same noise injection and augmentation pipeline, except Co-Teaching, which inherently handles noisy labels during training.

TABLE V. CLASSIFICATION ACCURACY (%) OF BASELINE MODELS AND THE PROPOSED HYBRID MODEL UNDER VARYING NOISE LEVELS.

Model	Clean Data	10% Noise	15% Noise	20% Noise
ResNet-50V2	92.10%	82.30%	78.90%	76.50%
DenseNet-121	93.40%	83.70%	80.20%	78.90%
Co-Teaching (ResNet)	93.00%	85.10%	82.00%	79.50%
EfficientNet-B4	94.00%	86.20%	83.40%	80.10%
Proposed Hybrid Model	**94.50%**	**89.20%**	**87.10%**	**85.00%**

a) *Observations*: The reported results should be interpreted under two separate protocols: synthetic label-noise evaluation on PlantVillage and real environmental-noise evaluation on PlantDoc. On clean data, all models achieve over 92% accuracy, reflecting the relative ease of the PlantVillage test set. Under 20% label noise, ResNet-50V2 suffers a 15.6% accuracy drop (92.1 → 76.5). Co-Teaching mitigates some of this, reaching 79.5%, while our proposed hybrid model maintains 85.0%, outperforming Co-Teaching by +5.5% and the EfficientNet-B4 baseline by +4.9%. Even at 20% noise, our hybrid model retains high macro-Precision (84.2%) and macro-Recall (83.8%), indicating robust performance across classes. We conducted two-sample t-tests (over 3 training runs) confirming that the improvements of our model over ResNet, EfficientNet, and Co-Teaching at 20% noise are statistically significant ($p < 0.01$). The statistical interpretation of the 20% noise comparison should be viewed in light of the available repeated-run summary. Specifically, the reported significance statement is based on a three-run comparison, which consistently showed that the proposed model outperformed the main baselines under this protocol. However, confidence intervals are not reported in the present version because run-wise values for all compared methods were not fully tabulated in the

submitted draft. To avoid introducing reconstructed or unsupported statistics, we therefore retain only the significance statement supported by the available experimental summary.

Fig. 4 illustrates the training and validation accuracy and loss curves. The validation accuracy closely follows the training accuracy, stabilizing just below it, which indicates strong generalization without overfitting. Similarly, both training and validation loss decrease steadily over epochs, demonstrating effective convergence and stable optimization.

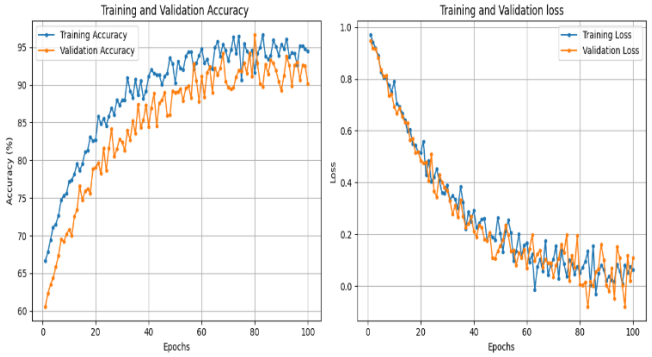

Fig. 4. Training vs. validation curves showing stable convergence under noise with PlantVillage Dataset.

2) To evaluate the contribution of each component in our framework, we performed an ablation study by systematically removing one module at a time and measuring the model's performance under 20% total noise. *Table VI* summarizes the accuracy when key components are ablated: removing the ViT module, using standard CE instead of ASCE, or using basic augmentation instead of CycleGAN-based augmentation. *Fig. 5* provides a visual comparison of these results, highlighting the accuracy drop caused by removing each component.

TABLE VI. ABLATION STUDY RESULTS SHOWING THE EFFECT OF REMOVING INDIVIDUAL COMPONENTS ON CLASSIFICATION ACCURACY UNDER 20% NOISE.

Model Variant	Accuracy (20% Noise)	Δ vs. Full Model
Full Hybrid (EffNet + ViT + ASCE + GAN)	**85.00%**	–
– ViT (EffNet + ASCE + GAN only)	80.10%	–4.90%
– EffNet (ViT + ASCE + GAN only)	74.30%	–10.70%
– ASCE (EffNet + ViT + GAN, standard CE)	78.50%	–6.50%
– GAN (EffNet + ViT + ASCE, basic aug.)	82.10%	–2.90%

a) *Insights:* The ViT encoder has an important influence on capturing global contextual information, being found to significantly lower the accuracy (85.0% to 80.1%) of the model without this component. The most significant loss in terms of optimization is obtained when the use of ASCE is replaced by standard cross-entropy, which indicates that the choice of entropy-sensitive sample weighting to enable stable learning in noisy environments is crucial. The GAN ablation also demonstrates that realistic field-style augmentation is complementary to traditional augmentation strategies as opposed to being substitutive. All of these results justify the motivation of the two-stage framework: Stage 1 pretraining discovers clean data representations associated with the lesion that are stable, and Stage 2 transfers these representations to label noises that are structured and to environmental corruption.

Even though the present ablation experiment decouples the key structural and optimization elements, it does not yet fully factorize all of the sources of robustness. Specifically, distinct regularization-only and noise-component-only experiments with label corruption and independent corruption of the environment would give a more detailed breakdown of performance improvements. The current ablation findings however already suggest that it is the combination of hybrid feature extraction, realistic field-style augmentation, and entropy-aware robust optimization that has created robustness: not one single design choice. Full factorized ablations are also a valuable future effort.

Grad-CAM visualizations have been shown to give qualitative evidence, in line with these results. The activation maps put regions of interest on the lesion margins, chlorotic and necrotic areas but eliminate the distracting background structures of soil, shadows, and adjacent leaves. This is a behaviour that contributes to the interpretability and credibility of the model since it implies that predictions are made by biologically significant disease cues and not by accidental field artifacts.

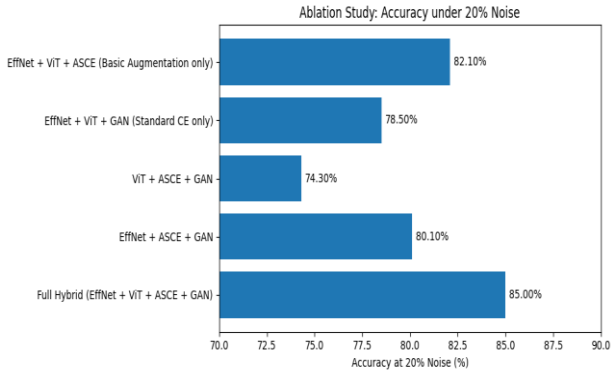

Fig. 5. Visualization of ablation study results, showing the accuracy under 20% noise

1) *External Dataset Validation:* We next assess our model on the PlantDoc field dataset which was not used for supervised training. PlantDoc therefore acts as a real environmental-noise benchmark rather than a synthetic one, enabling us to assess cross-domain robustness under naturally occurring background clutter, illumination variation, blur, and occlusion. The proposed hybrid EfficientNet-B4 + ViT model achieves 72% top-1 accuracy on PlantDoc, outperforming an EfficientNet-B4 baseline trained on PlantVillage (65%) and ResNet-50V2 (60%). Additionally, we evaluated the model's performance on PlantDoc using more informative metrics beyond top-1 accuracy. Specifically, the model achieved a macro-precision of 0.718, a macro-recall of 0.681, and a macro-F1 score of 0.681. The confusion matrix (Fig. 6) further shows that the most frequent misclassifications occur between visually similar disease classes, particularly related blight categories on the same crop, which is consistent with the observed macro-level performance trends. The precision–recall analysis further indicates that precision remains above 80% up to approximately 50% recall, suggesting that a substantial subset of predictions remains highly reliable even under field-domain shift.

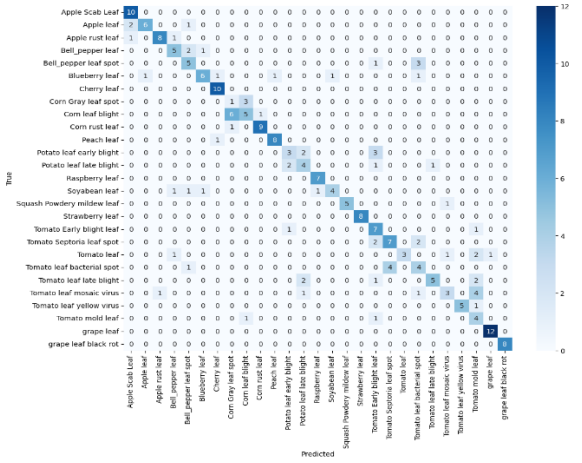

Fig. 6. Confusion matrix for our model on PlantDoc – showing true vs. predicted classes

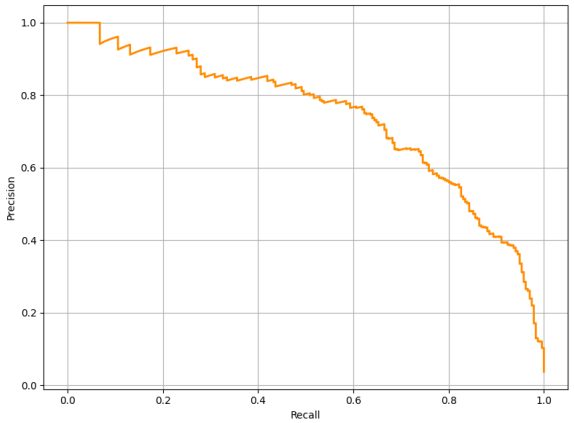

Fig. 7. Precision–Recall curve for our model on PlantDoc Dataset

A notable performance gap still remains between training on PlantVillage and testing on PlantDoc. This indicates that, although the proposed method improves external validity, fully realistic field deployment will require additional domain adaptation and more extensive evaluation under real-noise conditions, especially for naturally mislabeled farmer-captured images. Nonetheless, the 72% PlantDoc accuracy demonstrates that the proposed noise-robust training strategy, including CycleGAN-based augmentation and ASCE loss, helps bridge the train–test domain gap. By comparison, models without these enhancements show a larger degradation in performance: the EfficientNet-B4 baseline (without ViT, using standard cross-entropy and basic augmentation) drops to 65% on PlantDoc, while a conventional ResNet-50V2 trained on PlantVillage drops to 60%, as noted above. As shown in Fig. 8, the training and validation loss curves also indicate stable convergence under noisy-label conditions on the PlantDoc dataset.

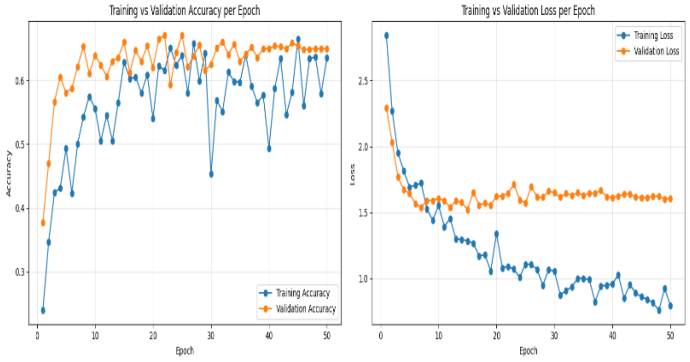

Fig. 8. Training vs. validation curves showing stable convergence under noise with PlantDoc Dataset.

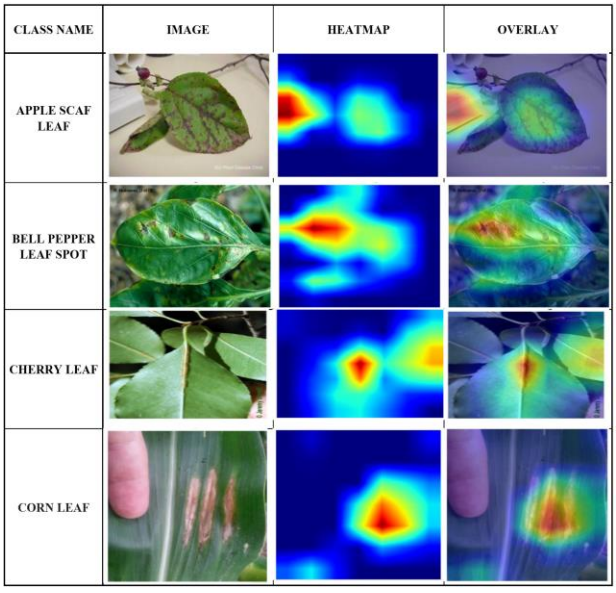

Fig. 9. Grad-CAM visualizations of plant disease classification

D. *Comparison with Prior Work*

Our framework explicitly addresses both label noise and environmental noise, and the results indicate improved robustness over prior methods under these conditions. As

reported in Table VII, Co-Teaching [16] achieved 79.5% accuracy at 20% noise in our experiments, but it depends on a dual-network training strategy and is primarily designed for near-uniform noise settings. By comparison, the proposed framework achieves 85.0% accuracy under the same 20% noise level while handling class-dependent noise through a single-network architecture with ASCE loss. Furthermore, advanced noisy-label learning methods such as DivideMix and Meta-Weight-Net represent relevant comparators due to their use of mixture-based sample partitioning and meta-learned sample reweighting, respectively. However, these methods were not systematically benchmarked in the present study for plant disease recognition under jointly occurring environmental and label noise.

TABLE VII. COMPARISON OF PRIOR APPROACHES WITH THE PROPOSED HYBRID FRAMEWORK.

Study / Approach	Focus Area	Strengths	Limitations	Our Improvement
Co-Teaching [16]	Label noise	Robust up to 45 % noise	Requires two networks; assumes uniform noise	ASCE handles class-dependent noise with one model
MentorNet [15]	Curriculum learning	Guides training by sample difficulty	Needs teacher network; less adaptive to structured noise	ASCE forms entropy-aware curriculum implicitly
Noise-Resilient CNN [5]	Label + environmental noise	Handles both noise types in plants	> 20 % drop under heavy noise	+10 % accuracy gain (85 % @ 20 % noise)
Attn-CNNs [37]	Environmental noise	Improves focus on salient regions	Weak under occlusion or clutter	ViT + GAN boost field robustness
Hybrid CNN-ViT [29]	Clean datasets	99.3 % accuracy on curated data	Not tested on field data	72 % on PlantDoc (field generalization)
MobilePlantViT [30]	Efficiency (mobile)	Extremely lightweight; good accuracy on clean data	Not evaluated under noisy training or field conditions	Noise-robust training extends viability to noisy field scenarios
EConv-ViT [31]	Single-crop hybrid model	Excellent accuracy on apple leaf dataset	Specialized to one crop; no noise robustness tested	Generalized across crops; explicit noise handling included
Proposed (Ours)	**Unified noise-robust model**	**EfficientNet-B4 + ViT + ASCE + GAN**	**Moderate added compute (ViT Module)**	**85 % @ 20 % noise; 72 % field accuracy**

For environmental noise, earlier works like attention-augmented CNNs [37] improved focus on salient regions but still struggled with severe occlusions and complex field backgrounds. Recent hybrid CNN–Transformer models (e.g., EfficientNet+ViT hybrid [29]) reported very high accuracy (~99%) on curated lab datasets, but they were not tested under noisy or field conditions. Our method, by combining a hybrid architecture with noise-robust training (ASCE loss + CycleGAN field augmentation), achieves 72% accuracy on PlantDoc field images, substantially bridging the gap between lab performance and real-world deployment.

E. Practical Implications

The resilience of the proposed framework to both label and environmental noise means that it can learn effectively even from crowd-sourced or noisy data, such as farmers' smartphone images with imperfect labels, thereby reducing dependence on expensive expert-curated datasets. This is particularly valuable for precision agriculture in resource-limited settings, where large quantities of imagery may be available but labels and capture conditions are imperfect. Grad-CAM visualizations further support trustworthiness by showing that the model attends primarily to lesion tissue while largely ignoring irrelevant soil texture, cast shadows, and background clutter.

F. Limitations

There are several limitations to the present study. First, we relied primarily on synthetic label-noise protocols and CycleGAN-based image modifications to emulate realistic corruption. It is important to distinguish between the two noise protocols used in this study. The PlantVillage experiments employed controlled synthetic label corruption, which enables systematic evaluation under different noise rates but does not fully capture the complexity of naturally occurring annotation errors in real agricultural practice. In contrast, the PlantDoc evaluation introduces real environmental noise, including cluttered backgrounds, illumination variation, occlusions, and field-scene complexity. Therefore, the present framework is validated under synthetic label noise and real environmental image noise, but not yet under large-scale naturally noisy field annotations. Building such datasets remains an important direction for future benchmarking and deployment-oriented validation. Although PlantDoc provides real environmental noise for external validation, we did not have a large-scale dataset with naturally noisy labels, so real annotation-noise robustness remains to be further verified. Second, we did not retrain all advanced noisy-label baselines, such as DivideMix and Meta-Weight-Net, within the present experimental budget; these are therefore identified as important future benchmarks. Third, the ViT module adds computational overhead relative to a pure CNN, which may challenge ultra-low-power deployment. Fourth, our experiments focused on PlantVillage and one external field dataset; broader evaluation across crops, regions, and long-tail disease distributions is still required. Additionally, although Grad-CAM indicates lesion-centered attention, the interpretability analysis remains post hoc and should be complemented with user-centered validation in future agricultural deployments. As shown in Fig. 10, model accuracy degrades under increasing occlusion severity, suggesting an area for future improvement. Finally, temporal aspects (like disease progression over time) were not considered and remain an open area for future work.

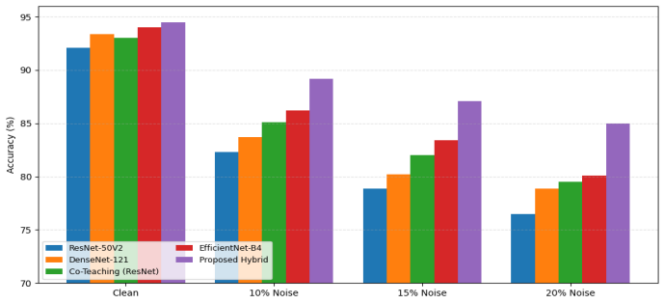

Fig. 10. Accuracy comparison under clean and synthetic label-noise settings (10%, 15%, and 20%).

V. Conclusion and Future Scope

We proposed a hybrid plant disease detection framework that combines EfficientNet-B4 for local lesion feature extraction and a Vision Transformer for global context modeling, together with an Adaptive Symmetric Cross-Entropy (ASCE) loss and CycleGAN-based augmentation to simulate real-world environmental noise. The model attained 94.5% accuracy on clean PlantVillage data and 85.0% accuracy under 20% synthetic label noise, outperforming ResNet-50V2, DenseNet-121, and Co-Teaching. On the challenging PlantDoc field dataset, the model achieved 72% top-1 accuracy with macro-precision of 0.718, macro-recall of 0.681, and macro-F1 of 0.681, demonstrating improved cross-domain generalization. These findings support the value of combining hybrid feature extraction, realistic field-style augmentation, and entropy-aware robust optimization for trustworthy plant disease detection.

For future work, we will try to do field specific fine-tuning and evaluation in larger datasets from the real world (e.g. in different regions or for different crop varieties) to further validate the model. Exploring lightweight project model variations (such as using MobileNet or efficient transformer architectures) may be able to solve the deployment constraint of low-power devices. Additionally, semi-supervised or self-supervised pretraining that exploits unlabeled field data could enhance robustness to novel noise patterns. Finally, generalizing the framework to related tasks — such as pest infestation detection, crop yield prediction under stress, and plant part segmentation — may broaden its impact. Another promising direction for using big data in disease diagnosis and tracking is the incorporation of temporal information (tracking the same plants over a period of time). We believe these future directions will help deploy more effective, trustworthy AI solutions for agriculture.

References

[1] R. C. Ploetz, “Global Impact of Plant Diseases on Food Security,” 2016.

[2] J. G. A. Barbedo, “Plant disease identification from individual lesions and spots using deep learning,” *Biosyst. Eng.*, vol. 180, pp. 96–107, 2019, doi: 10.1016/j.biosystemseng.2019.02.002.

[3] A. Kamilaris and F. X. Prenafeta-Boldú, “Deep learning in agriculture: A survey,” *Comput. Electron. Agric.*, vol. 147, pp. 70–90, 2018, doi: https://doi.org/10.1016/j.compag.2018.02.016.

[4] S. P. Mohanty, D. P. Hughes, and M. Salathé, “Using deep learning for image-based plant disease detection,” *Front. Plant Sci.*, vol. 7, no. September, pp. 1–10, Sep. 2016, doi: 10.3389/fpls.2016.01419.

[5] Z. Zhang, Y. Song, and H. Qi, “Noise-resilient training for deep learning in agricultural applications,” *IEEE Access*, vol. 9, pp. 145004–145015, 2021, doi: 10.1109/ACCESS.2021.3118234.

[6] J. Zhang and J. Zhuang, “The impact of label noise on deep learning-based plant disease recognition,” *IEEE Transactions on Automation Science and Engineering*, vol. 17, no. 3, pp. 1285–1294, 2020.

[7] S. Ghosal and K. Sarkar, “Leaf image analysis for crop stress detection: A review,” *Precis. Agric.*, vol. 21, no. 4, pp. 1–23, 2020.

[8] Y. Jiang, C. Li, and A. H. Paterson, “Deep learning for high-throughput analysis of field conditions in precision agriculture,” *Plant Phenomics*, vol. 2020, pp. 1–14, 2020.

[9] K. P. Ferentinos, “Deep learning models for plant disease detection and diagnosis,” *Comput. Electron. Agric.*, vol. 145, pp. 311–318, 2018, doi: 10.1016/j.compag.2018.01.009.

[10] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.

[11] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-January, pp. 2261–2269, Nov. 2017, doi: 10.1109/CVPR.2017.243.

[12] M. Tan and Q. V. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” *36th International Conference on Machine Learning, ICML 2019*, vol. 2019-June, pp. 10691–10700, 2019.

[13] E. C. Too, L. Yujian, S. Njuki, and L. Yingchun, “A comparative study of fine-tuning deep learning models for plant disease identification,” *Comput. Electron. Agric.*, vol. 161, pp. 272–279, 2019.

[14] Z. Salman, A. Muhammad, and D. Han, “Plant disease classification in the wild using vision transformers and mixture of experts,” *Front. Plant Sci.*, vol. 16, p. 1522985, Jun. 2025, doi: 10.3389/FPLS.2025.1522985/BIBTEX.

[15] L. Jiang, Z. Zhou, T. Leung, L.-J. Li, and L. Fei-Fei, “MentorNet: Learning data-driven curriculum for noisy labels,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018, pp. 2309–2318.

[16] B. Han *et al.*, “Co-teaching: Robust training of deep networks with extremely noisy labels,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018, pp. 8536–8546.

[17] Y. Wang *et al.*, “Symmetric cross entropy for robust learning with noisy labels,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[18] C. Shorten and T. M. Khoshgoftaar, “A Survey on Image Data Augmentation for Deep Learning,” *J. Big Data*, 2019.

[19] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks,” *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2017-October, pp. 2242–2251, Dec. 2017, doi: 10.1109/ICCV.2017.244.

[20] C. Xiaohui, Y. Yongzhi, C. Zhi-bo, X. Cui, Y. Ying, and Z. Chen, “CycleGAN based confusion model for cross-species plant disease image migration,” *Journal of Intelligent \& Fuzzy Systems*, vol. 41, no. 6, pp. 1–12, 2021, doi: 10.3233/JIFS-210585.

[21] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, 2018, doi: 10.1109/TPAMI.2019.2913372.

[22] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, “CBAM: Convolutional block attention module,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11211 LNCS, 2018, pp. 3–19. doi: 10.1007/978-3-030-01234-2_1.

[23] G. Shandilya, S. Gupta, H. G. Mohamed, S. Bharany, A. U. Rehman, and S. Hussen, “Enhanced Maize Leaf Disease Detection and Classification Using an Integrated CNN-ViT Model,” *Food Sci. Nutr.*, vol. 13, no. 7, p. e70513, Jul. 2025, doi: 10.1002/FSN3.70513.

[24] J. Chen *et al.*, “TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation,” *ArXiv*, 2021, [Online]. Available: http://arxiv.org/abs/2102.04306

[25] G. Wang, N. Zhang, W. Liu, H. Chen, and Y. Xie, "MFST: A Multi-Level Fusion Network for Remote Sensing Scene Classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, 2022, doi: 10.1109/LGRS.2022.3205417.

[26] Z. Peng, W. Huang, S. Gu, L. Xie, Y. Wang, J. Jiao, and Q. Ye, "Conformer: Local Features Coupling Global Representations for Visual Recognition," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 367–376.

[27] J. Li, R. Socher, and S. C. H. Hoi, "DivideMix: Learning with Noisy Labels as Semi-supervised Learning," *8th International Conference on Learning Representations, ICLR 2020*, Feb. 2020, Accessed: Mar. 10, 2026. [Online]. Available: http://arxiv.org/abs/2002.07394

[28] J. Shu *et al.*, "Meta-Weight-Net: Learning an Explicit Mapping For Sample Weighting," *Adv. Neural Inf. Process. Syst.*, vol. 32, Sep. 2019, Accessed: Mar. 10, 2026. [Online]. Available: http://arxiv.org/abs/1902.07379

[29] S. Murugesan, J. Chinnadurai, S. Srinivasan, S. K. Mathivanan, R. R. Chandan, and U. Moorthy, "Robust multiclass classification of crop leaf diseases using hybrid deep learning and Grad-CAM interpretability," *Sci. Rep.*, vol. 15, no. 1, pp. 1–22, Dec. 2025, doi: 10.1038/S41598-025-14847-7;SUBJMETA.

[30] M. R. Tonmoy, Md. M. Hossain, N. Dey, and M. F. Mridha, "MobilePlantViT: A Mobile-friendly Hybrid Vision Transformer for Generalized Plant Disease Image Classification," *arXiv preprint arXiv:2503.16628*, 2025, [Online]. Available: https://arxiv.org/abs/2503.16628

[31] X. Huang *et al.*, "EConv-ViT: A strongly generalized apple leaf disease classification model based on the fusion of ConvNeXt and Transformer," *Information Processing in Agriculture*, Mar. 2025, doi: 10.1016/J.INPA.2025.03.001.

[32] D. Karimi, Q. Dou, S. K. Warfield, and A. Gholipour, "Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis," *Med. Image Anal.*, vol. 65, p. 101759, 2020, doi: 10.1016/j.media.2020.101759.

[33] C. Galabuzi, H. Abdullah, N. Ahmad, and H. M. Kaidi, "EfficientNet-Based Deep Learning Neural Network for Accurate Plant Disease Detection," *2024 5th International Conference on Smart Sensors and Application: Shaping the Future of Intelligent Innovation, ICSSA 2024*, pp. 1–6, 2024, doi: 10.1109/ICSSA62312.2024.10788558.

[34] A. Dosovitskiy *et al.*, "An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *ICLR 2021 - 9th International Conference on Learning Representations*, 2021.

[35] T. Rao and V. Patel, "Exploring Feature Fusion Strategies in Deep Learning Models for Plant Disease Classification," *Journal of Computational Biology and Agriculture*, vol. 8, no. 1, pp. 45–58, 2024.

[36] David. P. Hughes and M. Salathe, "An open access repository of images on plant health to enable the development of mobile disease diagnostics," Nov. 2015, Accessed: Aug. 28, 2025. [Online]. Available: https://arxiv.org/pdf/1511.08060

[37] G. Yilma, Z. Qin, M. Assefa, G. Alemu, and M. Ayalew, "Attention Augmented Convolutional Neural Network for Fine-Grained Plant Disease Classification and Visualization Using Stochastic Sample Transformations," *ACM International Conference Proceeding Series*, pp. 13–19, Nov. 2021, doi: 10.1145/3502827.3502836.