A lightweight convolutional and vision transformer hybrid network for parameter efficient plant disease classification

Thanh-Hai Tong-Le

tonglethanhhai@tgu.edu.vn

College of Engineering and Technology, Tra Vinh University

Minh-Hai Le

College of Engineering and Technology, Tra Vinh University

Thanh-Nghi Doan

An Giang University, Vietnam National University

Research Article

Keywords: plant disease classification, lightweight CNN–Transformer, attention mechanisms, edge AI, explainable AI

Posted Date: April 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9235961/v1

Additional Declarations: No competing interests reported.

A lightweight convolutional and vision transformer hybrid network for parameter efficient plant disease classification

Thanh-Hai Tong-Le[1,2*], Minh-Hai Le[1] and Thanh-Nghi Doan[3*]

[1*]College of Engineering and Technology, Tra Vinh University, Vinh Long, Vietnam.
[2]Tien Giang University, Dong Thap, Vietnam.
[3*]An Giang University, Vietnam National University, Ho Chi Minh City, Vietnam.

*Corresponding author(s). E-mail(s): tonglethanhhai@tgu.edu.vn; dtnghi@agu.edu.vn;
Contributing authors: lmhai@tvu.edu.vn;

Abstract

Accurate plant disease classification for edge deployment requires models that are both precise and efficient. Convolutional neural networks (CNNs) learn local lesion patterns effectively. Pure Transformers capture global dependencies more explicitly, but they typically incur a higher computational cost. We propose CViTLw, a lightweight CNN-Transformer hybrid comprising a MobileNetV2 branch, a compact Vision Transformer branch, and an attention-enhanced cross-fusion module. We further evaluate two lightweight attention mechanisms, SE and CBAM, within the same framework. Experiments on the PlantVillage and Maize Leaf Disease datasets are conducted under both controlled and field-acquired conditions. On the Maize Leaf Disease dataset, CViTLw-SE attains 94.85% accuracy with only 0.33M parameters. On PlantVillage, both CViTLw-SE and CViTLw-CB achieve 99.74% accuracy (as well as precision, recall, and F1-score) and an AUC of 100%. The most compact variants operate in less than 2 ms per image and exceed 500 FPS. Overall, CViTLw achieves a strong balance of accuracy, efficiency, and practical deployability. The source code is publicly available at this link.

Keywords: plant disease classification; lightweight CNN–Transformer; attention mechanisms; edge AI; explainable AI.

1 Introduction

Plant diseases remain a major source of agricultural yield loss, with estimated reductions of 20% to 40% depending on crop and region [1]. This has increased the need for automated disease detection in precision agriculture. A practical system must be accurate, compact, and efficient enough for edge deployment. Manual inspection by field experts is difficult to scale because it depends on specialist experience, labor availability, and inspection time [2]. For

this reason, computer vision and deep learning have become central tools for image-based plant disease recognition.

CNNs have proved highly effective at extracting local patterns such as lesions, damaged boundaries, and textural changes. Weight sharing and progressively expanding receptive fields allow CNNs to learn recurring morphological motifs while retaining relatively strong data efficiency on moderately sized image datasets. Architectures such as VGG [3], ResNet [4], MobileNetV2 [5], and attention-augmented variants have reported strong results across a wide range of plant disease benchmarks [11, 20–22, 26, 27]. Nonetheless, the inherently local nature of convolution becomes a limitation when disease symptoms are spatially diffuse or depend on relationships among distant image regions.

Vision Transformers (ViTs) and related variants such as Swin Transformer capture global context more directly through self-attention [6, 7]. In principle, self-attention allows each token to interact with all others within every layer, so global receptive fields emerge earlier than in standard CNN pipelines. This property is especially useful when disease symptoms appear sparsely across the leaf surface. Yet many Transformer-based models remain parameter-heavy, memory-intensive, and computationally expensive at inference time. In addition, because their inductive bias is weaker than that of CNNs, Transformers often rely more heavily on large-scale pretraining or abundant training data to optimize stably. These factors reduce their suitability for resource-constrained agricultural systems.

This trade-off has motivated growing interest in CNN–Transformer hybrids. These models combine local feature extraction from CNNs with global context modeling from Transformers. They also reintroduce local inductive bias into the Transformer pathway, which can improve data efficiency and optimization stability on specialized datasets. Recent studies, including MobileViT [16], CoAtNet [17], and the work of Thakur et al. [29], show the promise of this direction. However, three gaps remain. First, many hybrid models are still too heavy for practical deployment. Second, many studies emphasize absolute accuracy without carefully analyzing the trade-off between predictive performance and architectural complexity. Third, the role of lightweight attention mechanisms such as SE and CBAM in compact hybrid architectures for noisy field data is still underexplored.

Addressing these notable gaps, we introduce CViTLw, a lightweight CNN-Transformer architecture engineered to strike an optimal balance between predictive accuracy and deployment feasibility. The proposed network integrates a MobileNetV2 branch explicitly for localized feature extraction, alongside a pared-down Vision Transformer branch meant to establish global context. These parallel streams finally merge via an attention-augmented cross-fusion module. We subsequently validate this configuration across the PlantVillage and Maize Leaf Disease datasets to encompass both pristine laboratory imagery and challenging, real-world field conditions.

Within our evaluation framework, PlantVillage acts as a standardized reference point to facilitate direct comparisons with historical studies. Conversely, the Maize Leaf Disease dataset functions as the primary testbed, as it accurately mirrors the unpredictable variables inherent to active field deployment. This dual-dataset approach intentionally isolates benchmark standardization from practical real-world validation.

The core contributions of this research include:

- The development of CViTLw, a highly compact CNN-Transformer hybrid tailored for plant disease classification. This architecture intelligently merges MobileNetV2-based local feature extraction, a lightweight ViT branch for global context, and a dedicated attention-enhanced cross-fusion module.

- A rigorous experimental trial spanning eight distinct network variants. This structural breakdown effectively isolates the individual impacts of Transformer depth, backbone scale, and specific lightweight attention mechanisms (SE versus CBAM).

- Empirical evidence demonstrating that ultralight hybrid networks can preserve exceptional predictive power with minimal computational overhead. Specifically, the CViTLw-SE configuration achieves 94.85% accuracy on the Maize dataset utilizing just 0.33M parameters, establishing a highly favorable balance between low latency and robust inference throughput.

Subsequent sections detail the related literature (Section 2), the architectural design of CViTLw (Section 3), and the exact experimental configuration (Section 4). Following this, Section 5 breaks down the quantitative and qualitative findings, while Section 6 contextualizes the results and outlines future deployment considerations. Section 7 provides closing remarks.

2 Related Work

2.1 CNN-Based Plant Disease Classification

Convolutional neural networks (CNNs) have long served as the backbone of plant-disease recognition because they excel at extracting local visual cues such as lesions, edge discontinuities, and texture variations. By sharing weights across spatial locations, convolutions generate repeated feature detectors that can be reused throughout an image, enabling efficient learning of morphological patterns. Early studies demonstrated the practicality of transfer learning with AlexNet and GoogLeNet on the PlantVillage repository [2], while subsequent work surveyed a broader set of classic CNNs and reported high accuracies under controlled imaging conditions [11]. These investigations established a solid baseline for leaf-image classification.

To further reduce computational demand, lightweight CNN families—including MobileNet [10], MobileNetV2 [5], and EfficientNet [13]—have been introduced. In agricultural contexts, attention modules are frequently appended to CNNs to highlight diagnostically salient regions [20–22, 26, 27]. Nevertheless, pure CNNs remain limited by their inherently local receptive fields, which can hinder performance when disease symptoms are spatially diffuse or occluded by background clutter in field images.

2.2 Vision Transformers for Image Classification

The Vision Transformer (ViT) paradigm treats an image as a sequence of patches and models inter-patch relationships through self-attention [6]. This design grants immediate access to long-range dependencies and global context, bypassing the gradual receptive-field growth required by stacked convolutions. Hierarchical variants such as Swin Transformer improve efficiency via shifted-window attention [7].

In plant-disease tasks, ViTs are attractive because pathological cues often span the entire leaf surface. Recent reports have shown that Transformer-based models can achieve competitive performance on several benchmarks [23–25]. However, this advantage typically incurs higher parameter counts, increased training and inference costs, and a stronger reliance on large-scale pretraining or abundant domain data. Consequently, pure Transformers may be unsuitable for resource-constrained agricultural deployments.

2.3 Hybrid CNN–ViT Architectures

Hybrid CNN–Transformer designs aim to combine the locality of convolutions with the global modeling capacity of self-attention. Representative works such as LeViT, MobileViT, and CoAtNet illustrate that integrating convolutional and attention pathways yields richer representations than either component alone [16–18]. For plant-disease analysis, this synergy is valuable because lesions provide fine-grained cues while disease patterns may also exhibit broader spatial structure. From a machine-learning perspective, hybrids reconcile the strong inductive bias of CNNs with the flexibility of attention, improving generalization on limited yet heterogeneous field data.

Several recent studies have applied hybrid designs to plant-disease datasets collected under diverse conditions. Notably, Thakur et al. introduced a parallel CNN–ViT architecture that achieved favorable results on multiple datasets, including field-collected maize images [29]. Yet, variations in dataset composition, label spaces, and augmentation protocols complicate direct cross-study comparisons. Moreover, systematic investigations of how compact hybrid models can be compressed without sacrificing performance remain scarce—an issue this work explicitly addresses.

2.4 Attention Mechanisms in Computer Vision

Attention mechanisms reweight feature maps to emphasize informative signals while suppressing irrelevant background. The Squeeze-and-Excitation (SE) block performs channel-wise recalibration by aggregating spatial information [8], whereas the Convolutional Block Attention Module (CBAM) adds a spatial-attention branch on top of channel attention [9].

In plant-disease recognition, such lightweight attention modules are especially beneficial because pathological signs are often small, unevenly distributed, and easily confused with healthy tissue or background. Nevertheless, rigorous comparisons between SE and CBAM within compact hybrid architectures are limited. Accordingly, this study not only proposes the CViTLw model but also evaluates the distinct contributions of these two attention mechanisms under a unified framework.

3 Methodology

3.1 Attention Modules

3.1.1 CBAM (Convolutional Block Attention Module)

CBAM [9] extends SE by adding spatial attention in addition to channel attention. This structure is particularly suitable for field data, where background clutter and context variation can weaken disease cues if feature recalibration is performed only along the channel dimension. CBAM consists of two sequential stages:

Channel attention (determining the most relevant channels):

$$M_c(F) = \sigma(\mathrm{Conv}_{1\times1}(\delta(\mathrm{Conv}_{1\times1}(\mathrm{GAP}(F))))) \tag{1}$$

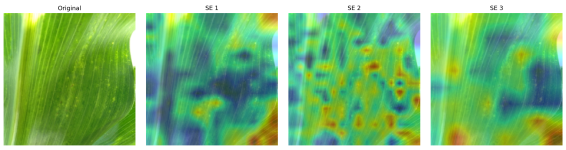
Spatial attention (determining the most relevant regions):

$$M_s(F') = \sigma(\mathrm{Conv}_{7\times7}([\mathrm{AvgPool}_{ch}(F'); \mathrm{MaxPool}_{ch}(F')])) \tag{2}$$

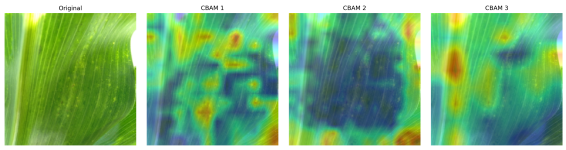
Here, AvgPool_{ch} and MaxPool_{ch} perform pooling across the channel dimension to produce maps in $\mathsf{R}^{H\times W\times 1}$; the concatenated output is then passed through a 7×7 convolution to generate a spatial attention map that highlights diagnostically meaningful regions.

$$F_{\text{out}} = M_s(F') \otimes F', \quad F' = M_c(F) \otimes F \tag{3}$$

CBAM is expected to be especially helpful when the model must determine which features matter most and where the diagnostically relevant cues are located. Figure 1 illustrates the qualitative difference between SE and CBAM on representative diseased maize leaves.



(a) SE attention focused on channel recalibration



(b) CBAM attention highlighting spatial regions

Figure 1: Visualization of attention behavior on the Maize dataset. SE mainly redistributes weights across channels, whereas CBAM adds spatial localization and highlights lesion regions more clearly. Warmer colors indicate stronger attention.

3.2 Cross-Fusion Module

The cross-fusion module is a key component of CViTLw because it is the stage at which local and global signals are brought into a shared representational space. The two branches produce feature maps with different spatial resolutions because their downsampling processes differ. Let F_T denote the output of the Transformer branch after backbone-specific alignment. For the custom-designed ViT, this tensor has size $16\times 16\times 48$; for Swin-Tiny and ViT-B/16, the original outputs are interpolated and projected with a 1×1 convolution to obtain a $16\times 16\times 128$ representation. This alignment allows feature mixing to occur on a common spatial scale and prevents resolution mismatch from distorting the fused signal during channel concatenation. After alignment, the tensors are concatenated along the channel dimension:

$$F_{\text{fused}} = [\text{Resize}_{14\to 16}(F_{\text{CNN}}); F_T] \in \mathsf{R}^{16\times 16\times (32+C_T)} \tag{4}$$

where Resize denotes bilinear interpolation that maps F_{CNN} from $14\times 14\times 32$ to $16\times 16\times 32$, $C_T = 48$ for the custom ViT, and $C_T = 128$ for Swin-Tiny/ViT-B/16. The fused tensor is then processed by an inverted residual block:

$$F_{\text{mixed}} = \text{InvRes}_{(32+C_T)\to 128}(F_{\text{fused}}) \in \mathsf{R}^{8\times 8\times 128} \tag{5}$$

$\text{InvRes}_{(32+C_T)\to 128}$ consists of: (i) a pointwise expansion convolution with expansion ratio 6, (ii) a 3×3 depthwise convolution with stride 2 that reduces the spatial resolution to 8×8, and (iii) a pointwise projection to 128 channels. This structure keeps parameter cost low while still mixing information from both feature branches before the final attention refinement stage. In effect, the post-fusion inverted residual block serves both as an efficient local mixer and as a compact compression layer for downstream classification:

$$\hat{y} = \text{Linear}_{128\to C}(\text{GAP}(\mathsf{A}(F_{\text{mixed}}))) \in \mathsf{R}^{C} \tag{6}$$

where A is the final attention module (SE or CBAM), global average pooling compresses the 8×8 spatial tensor, and the linear layer produces logits for C output classes, with $C = 38$ on PlantVillage and $C = 4$ on Maize. Thus, the cross-fusion module concatenates features while also aligning, reorganizing, and refining the representation before classification.

3.3 Architectural Variants of CViTLw

3.3.1 Overview

Table 1 summarizes the eight variants evaluated in this study. All models share the same CNN–Transformer fusion framework and differ only in three factors: the Transformer backbone, the depth of the custom ViT branch, and the final attention mechanism. This setup allows the effect of each architectural decision to be analyzed in isolation without altering the overall character of the model.

Table 1: Architectural comparison of the eight variants evaluated in this study.

Model	CNN Backbone	ViT Backbone	ViT Layers	Attention	Params
CViTSE	MobileNetV2 (8 blocks)	Custom ViT (7×7)	8	SE	0.41M
CViTCB	MobileNetV2 (8 blocks)	Custom ViT (7×7)	8	CBAM	0.41M
CViTLw-SE	MobileNetV2 (8 blocks)	Custom ViT (7×7)	4	SE	0.33M
CViTLw-CB	MobileNetV2 (8 blocks)	Custom ViT (7×7)	4	CBAM	0.33M
CViTSE-swin	MobileNetV2 (8 blocks)	Swin-Tiny	Pretrained	SE	28.32M
CViTCB-swin	MobileNetV2 (8 blocks)	Swin-Tiny	Pretrained	CBAM	28.32M
CViTSE-ViT	MobileNetV2 (8 blocks)	ViT-B/16	Pretrained	SE	86.60M
CViTCB-ViT	MobileNetV2 (8 blocks)	ViT-B/16	Pretrained	CBAM	86.60M

Note: The CNN branch always produces features of size $14 \times 14 \times 32$. The Transformer branch produces $16 \times 16 \times 48$ features for the custom ViT and $16 \times 16 \times 128$ features for Swin/ViT-B/16. The fusion head always yields a 128-dimensional representation.

3.3.2 Shared Components

CNN branch. All eight variants use the same CNN backbone: the first eight inverted residual blocks of MobileNetV2 pretrained on ImageNet [5]. Keeping the CNN branch fixed helps ensure that observed differences arise primarily from the Transformer pathway and the attention mechanism rather than from changes in the local backbone. This branch produces a $14 \times 14 \times 32$ feature map before fusion.

Classification head. The final prediction head is identical across all variants, consisting of global average pooling followed by a linear layer FC($128 \to 38$). This design keeps the classifier compact while preserving sufficient discriminative capacity for the 38-class PlantVillage benchmark; for Maize, the number of output neurons is adjusted accordingly to 4 classes.

3.3.3 ViT Branch Variants

The eight models are distinguished primarily by the Transformer branch design:

- Custom ViT family. These models use a lightweight, custom-designed Transformer with 7×7 patches, projection dimension $d = 48$, and 4 attention heads. CViTSE and CViTCB

use 8 Transformer layers, whereas CViTLw-SE and CViTLw-CB use only 4 layers to reduce complexity. This group directly reflects the central hypothesis of the study: a small Transformer branch, if designed appropriately, can still be sufficiently expressive for the target task.

- Swin Transformer (CViTSE-swin, CViTCB-swin). These variants replace the custom ViT with Swin-Tiny pretrained on ImageNet [7]. The hierarchical shifted-window attention design provides multi-scale feature extraction and serves as a strong reference point at moderate computational cost.

- ViT-B/16 (CViTSE-ViT, CViTCB-ViT). These heavy variants use ViT-B/16 pretrained on ImageNet [6], processing 16×16 patches with 12 Transformer layers, embedding dimension $d = 768$, and 12 attention heads. This branch represents the other end of the trade-off spectrum, where stronger global modeling comes with the highest parameter and inference cost.

3.3.4 Attention Mechanisms

The two attention mechanisms evaluated in this study are intended to test how the effectiveness of feature recalibration depends on data characteristics and backbone scale:

- SE (Squeeze-and-Excitation) [8]. SE performs channel-wise recalibration through global average pooling and two fully connected layers with reduction ratio $r = 4$.

- CBAM (Convolutional Block Attention Module) [9]. CBAM extends SE by adding spatial attention on top of channel attention. The spatial branch uses pooled feature maps followed by a 7×7 convolution to jointly model channel importance and spatial relevance.

3.3.5 Fusion Module

The fusion module concatenates features from the CNN and Transformer branches along the channel dimension, producing a tensor of size $16 \times 16 \times 80$ for the custom ViT variants ($32 + 48$ channels) and $16 \times 16 \times 160$ for the Swin/ViT variants ($32 + 128$ channels). An inverted residual block with expansion factor 6 and stride 2 then reduces the spatial resolution to 8×8 and compresses the representation to 128 channels. A final attention module (SE or CBAM) is applied before global average pooling to increase the selectivity of the fused representation.

3.3.6 Computational Cost Analysis

The custom ViT variants (CViTSE and CViTCB) are the most parameter-efficient standard models, each requiring only 0.41M parameters. The ultra-light variants (CViTLw-SE and CViTLw-CB) further reduce this to 0.33M by halving the number of Transformer layers. In contrast, the Swin-based models (28.32M) and ViT-B/16-based models (86.60M) provide greater representational capacity but at a much higher memory and computational cost. The difference between SE and CBAM contributes only marginally to the total parameter count, allowing for a fairer comparison between the two attention mechanisms.

4 Experimental Setup

4.1 Datasets and Data Characteristics

To evaluate the proposed architecture comprehensively, we employ two benchmark datasets that differ markedly in acquisition conditions, class cardinality, and visual characteristics.

Leveraging both PlantVillage and Maize datasets broadens the validation scope and assesses whether the lightweight design principle remains effective when transitioning from controlled laboratory images to field-acquired photographs.

4.1.1 PlantVillage Benchmark

PlantVillage [12] is a widely used benchmark for leaf-image plant disease recognition. The version used in this study contains 54,305 images from 14 plant species and 38 classes, including both diseased and healthy leaves. The images were collected under controlled laboratory conditions with relatively uniform backgrounds and stable lighting. PlantVillage is therefore suitable for evaluating inter-class discrimination in a low-noise setting. Here, it is used as a common reference benchmark for assessing the competitiveness of the proposed architecture on a standardized dataset. Table 2 presents the class distribution used in our experiments.

4.1.2 Maize Leaf Disease (Field Conditions)

Unlike PlantVillage, the Maize Leaf Disease dataset [30] represents field conditions and therefore serves as the primary benchmark in this study. It contains 15,350 images of size 256×256 captured under natural outdoor conditions. The dataset includes four classes: healthy maize (33.3%), maize lethal necrosis disease (25.9%), maize streak virus type 1 (20.6%), and maize streak virus type 2 (20.2%). The class distribution is shown in Figure 2.

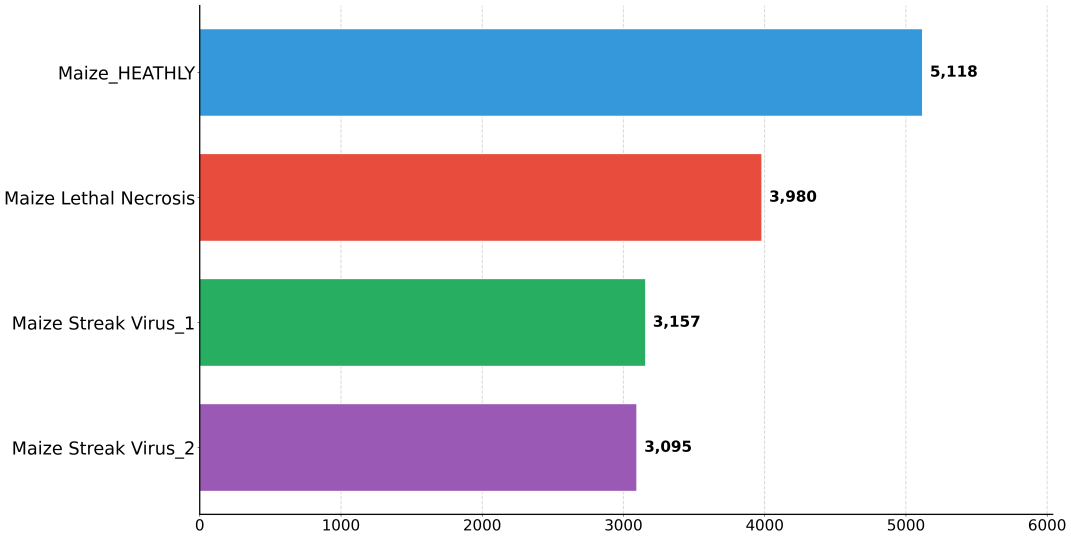


Figure 2: Class distribution of the Maize dataset. The healthy class accounts for roughly one-third of the samples, which helps limit severe class imbalance.

This dataset is substantially more challenging than PlantVillage because it contains non-uniform backgrounds, uncontrolled lighting, and high visual similarity between the two maize streak virus classes. It is therefore a more appropriate test bed for evaluating the robustness of the architecture under realistic deployment conditions, where the benefit of a lightweight but well-structured representation should become most evident.

Table 2: Class distribution of the PlantVillage benchmark used in this study.

No.	Crop	Condition (Class)	Samples
1	Apple	Apple_scab	2.472
2	Orange	Haunglongbing_(Citrus_greening)	5.507
3	Corn	Common_rust	1.192
4	Corn	Northern_Leaf_Blight	985
5	Grape	Black_rot	1.188
6	Grape	Esca_(Black_Measles)	1.385
7	Grape	Leaf_blight_(Isariopsis_Leaf_Spot)	1.072
8	Apple	Cedar_apple_rust	621
9	Tomato	Bacterial_spot	1.700
10	Strawberry	Healthy	456
11	Potato	Late_blight	1.000
12	Cherry	Powdery_mildew	1.052
13	Peach	Bacterial_spot	2.277
14	Pepper_bell	Bacterial_spot	997
15	Tomato	Early_blight	1.000
16	Tomato	Late_blight	1.859
17	Tomato	Leaf_Mold	1.877
18	Tomato	Septoria_leaf_spot	1.771
19	Tomato	Spider_mites_Two-spotted_spider_mite	1.676
20	Tomato	Target_Spot	1.440
21	Tomato	Tomato_YellowLeaf_Curl_Virus	5.356
22	Tomato	Tomato_mosaic_virus	458
23	Blueberry	Healthy	1.502
24	Cherry	Healthy	854
25	Corn	Cercospora_leaf_spot_Gray_leaf_spot	513
26	Grape	Healthy	423
27	Peach	Healthy	360
28	Pepper_bell	Healthy	1.478
29	Potato	Early_blight	1.000
30	Raspberry	Healthy	150
31	Soybean	Healthy	5.090
32	Squash	Powdery_mildew	1.835
33	Strawberry	Leaf_scorch	1.109
34	Tomato	Healthy	1.591
35	Apple	Black_rot	621
36	Apple	Healthy	2.005
37	Potato	Healthy	152
38	Corn	Healthy	934
		Total	**54,305**

4.2 Preprocessing and Data Augmentation

All input images are resized to $224 \times 224 \times 3$, ensuring a unified input size across all backbone configurations used in this study. To reduce overfitting, especially for variants with a Transformer branch, we apply a light but purposeful augmentation strategy focused on common variations encountered in leaf image acquisition:

- Spatial transforms: random horizontal flipping with probability 0.5 and random rotation within $\pm 20^\circ$ to simulate different leaf poses.
- Photometric transforms: color jitter with brightness ± 0.2 and contrast ± 0.2 to mimic illumination variation under field conditions.
- Normalization: ImageNet statistics with $\mu = [0.485, 0.456, 0.406]$ and $\sigma = [0.229, 0.224, 0.225]$, matching the pretrained backbones.

4.3 Training Configuration

Training and evaluation are conducted on a dedicated workstation. The hardware and software environment is summarized in Table 3. Reporting this configuration clarifies the experimental conditions and improves reproducibility.

Table 3: Hardware and software environment used for model development and evaluation.

Item	**Specification / Version**
Hardware Environment	
CPU	Intel Core i5-13400F (10 cores / 16 threads)
RAM	32 GB DDR4
GPU	NVIDIA GeForce RTX 4070 Ti SUPER (16 GB VRAM)
Storage	NVMe SSD 455 GB
Operating system	Ubuntu 24.04 (Linux Kernel 6.17)
Software Environment	
Base environment	Python 3.x
Deep learning framework	PyTorch 2.6.0 + CUDA 12.4 backend
Preprocessing & computer vision	torchvision 0.21.0, pillow 11.3.0
Numeric & data processing	numpy 1.26.4, pandas 2.3.1
Machine learning toolkit	scikit-learn 1.7.1
Visualization	matplotlib 3.10.3, seaborn 0.13.2

The main hyperparameters are set as follows:

- Loss function: cross-entropy for multi-class classification.
- Optimizer: Adam with an initial learning rate of 10^{-4}.
- Batch size and epochs: batch size of 32 and up to 20 training epochs.

- Early stopping: patience of 10 epochs based on validation loss.
- Data split: 80% for training and 20% for validation.

This training setup is kept consistent across all eight variants so that differences in the results mainly reflect architectural choices rather than changes in optimization hyperparameters.

4.4 Evaluation Metrics

Model performance is quantified using the following standard metrics:

$$\begin{aligned}
\text{Accuracy (AC)} &= \frac{\text{TP} + \text{TN}}{\text{Total}} \\
\text{Precision (PR)} &= \frac{\text{TP}}{\text{TP} + \text{FP}} \\
\text{Recall (RE)} &= \frac{\text{TP}}{\text{TP} + \text{FN}} \\
\text{F1-score (F1)} &= 2 \times \frac{\text{PR} \times \text{RE}}{\text{PR} + \text{RE}} \\
\text{Cohen's kappa } (\kappa) &= \frac{p_o - p_e}{1 - p_e}
\end{aligned}$$

where p_o denotes observed agreement and p_e expected agreement by chance.
Area under the ROC curve (AUC) is computed from the receiver-operating characteristic curve.

All metrics are computed per-class and macro-averaged unless otherwise noted.

4.5 Implementation Details

All experiments are implemented in Python 3.11 with PyTorch 2.6.0. The software pipeline follows the modular design described in Section 3, including data preprocessing, model construction, optimisation, validation, and testing stages. Experimental runs are controlled through a unified configuration scheme that specifies the dataset split, model variant, attention mechanism, optimiser settings, and a fixed random seed (42). Model checkpoints are stored after each epoch, and the checkpoint yielding the lowest validation loss is selected for final evaluation.

4.6 Reproducibility

To facilitate reproducibility, we provide the full source code, pretrained weights, configuration files, and the exact random seeds used in all reported experiments as supplementary material. Hardware specifications are summarised in Table 3, while the software dependencies and package versions are documented in the accompanying requirements file. In addition, all variants are evaluated under the same data partitioning, preprocessing protocol, optimisation strategy, and stopping criteria to ensure fair comparison across models.

5 Results

5.1 Overall Performance on the Two Benchmarks

We evaluate the eight variants using loss, accuracy (AC), precision (PR), recall (RE), F1-score (F1), area under the ROC curve (AUC), and Cohen's kappa (κ). The overall quantitative results are reported in Tables 4 and 5. In line with the study objective, the Maize results are interpreted first because this benchmark more directly reflects the deployment value of the lightweight design.

Table 4: Classification performance of the eight variants on the Maize Leaf Disease dataset under field conditions.

Group	Variant	Par. (M)	Loss	AC	PR	RE	F1	AUC	Kappa
Light (Custom)	**CViTLw-SE**	**0.33**	**0.14**	**94.85**	**94.91**	**94.85**	**94.84**	**99.31**	**0.93**
	CViTLw-CB	0.33	0.14	94.59	94.67	94.59	94.56	99.29	0.93
	CViTSE	0.41	0.14	94.50	94.61	94.50	94.46	99.29	0.93
	CViTCB	0.41	0.14	94.76	94.75	94.76	94.74	99.35	0.93
Medium (Swin)	CViTSE-swin	28.32	0.15	94.01	93.97	94.01	93.99	99.19	0.92
	CViTCB-swin	28.32	0.14	94.69	94.96	94.69	94.65	99.18	0.93
Large (ViT)	CViTSE-ViT	86.59	0.21	92.83	93.05	92.83	92.77	98.90	0.90
	CViTCB-ViT	86.59	0.19	93.16	93.11	93.16	93.09	99.08	0.91

On the challenging field Maize dataset (Table 4), CViTLw-SE achieves the highest overall performance, reaching 94.85% accuracy with $\kappa = 0.9304$ while requiring only 0.33M parameters. The compact variants within the custom ViT family remain highly competitive against larger architectures, often matching or outperforming models based on Swin-Tiny and ViT-B/16. This finding substantiates the central hypothesis of our study: for field-acquired data, a carefully compressed, task-specific hybrid architecture can prove more effective than indiscriminately scaling the backbone.

Conversely, for the PlantVillage dataset (Table 5), all models exceed 99.70% accuracy, suggesting the benchmark is nearing saturation. The highest accuracy is attained by CViTCB-swin at 99.89% with a loss of 0.0055. Nevertheless, the performance disparity among the tested variants is marginal. The custom ViT configurations sustain practically equivalent accuracy while operating with orders of magnitude fewer parameters. Thus, the PlantVillage results primarily demonstrate that the proposed architecture remains robust on a standardized corpus, whereas the Maize benchmark more fully illustrates the practical advantages of the lightweight design.

5.2 Training Convergence

The convergence curves provide useful insight into optimization stability and generalization behavior. Figure 5 plots representative learning curves on the Maize dataset. On this benchmark, the CViTLw-SE model maintains a stable training progression. The gap between training and validation loss restricts itself to a small margin throughout the optimization

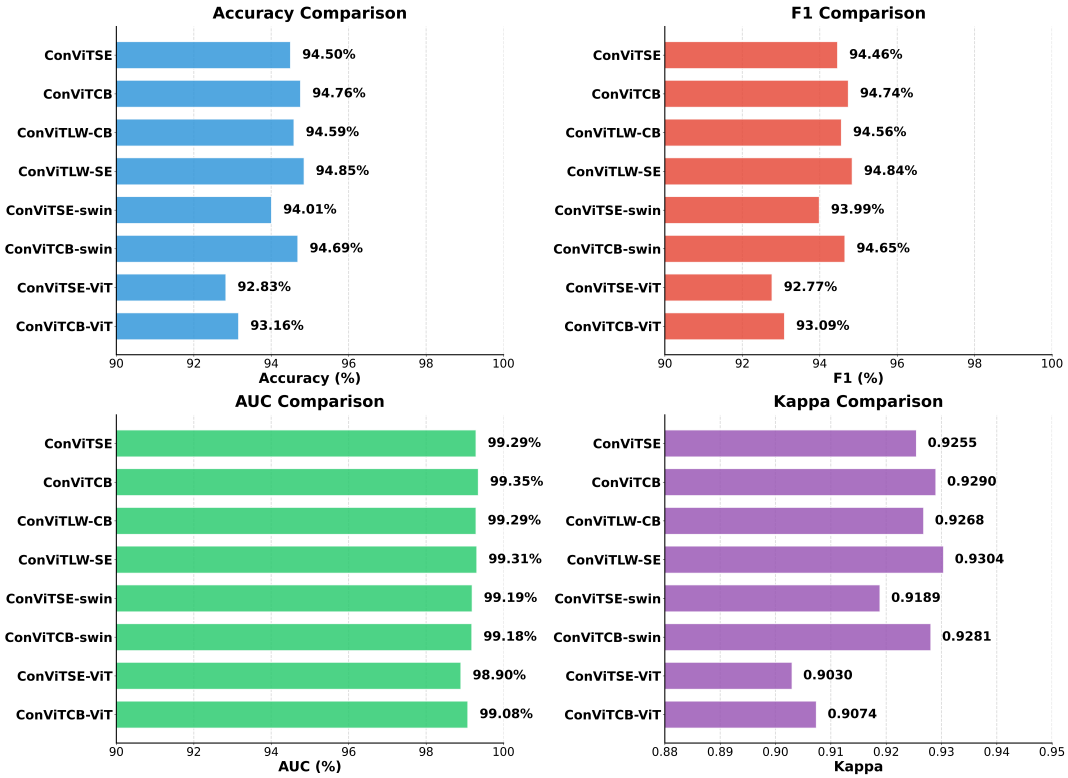


Figure 3: Comparison of performance metrics across the evaluated variants on the field Maize dataset.

Table 5: Classification performance of the eight variants on the PlantVillage benchmark under controlled conditions.

Group	Variant	Par. (M)	Loss	AC	PR	RE	F1	AUC	Kappa
Light (Custom)	CViTLw-SE	0.33	0.01	99.74	99.74	99.74	99.74	100.00	1.00
	CViTLw-CB	0.33	0.01	99.74	99.74	99.74	99.74	100.00	1.00
	CViTSE	0.41	0.01	99.70	99.70	99.70	99.70	100.00	1.00
	CViTCB	**0.41**	**0.01**	**99.79**	**99.79**	**99.79**	**99.79**	**100.00**	**1.00**
Medium (Swin)	CViTSE-swin	28.32	0.01	99.83	99.83	99.83	99.83	100.00	1.00
	CViTCB-swin	**28.32**	**0.01**	**99.89**	**99.89**	**99.89**	**99.89**	**100.00**	**1.00**
Large (ViT)	CViTSE-ViT	86.60	0.01	99.82	99.82	99.82	99.82	100.00	1.00
	CViTCB-ViT	86.60	0.01	99.83	99.83	99.83	99.83	100.00	1.00

process. Conversely, the larger CViTCB-ViT variant saturates earlier on the validation set and experiences stronger fluctuations. This indicates that heavy backbone architectures are not always optimal for high-variance field data.

For the PlantVillage benchmark (Figure 6), optimization is much smoother across all architecture configurations. This aligns with the controlled nature of laboratory-acquired data. Most variants reach a low validation loss within the initial five to seven epochs. Under

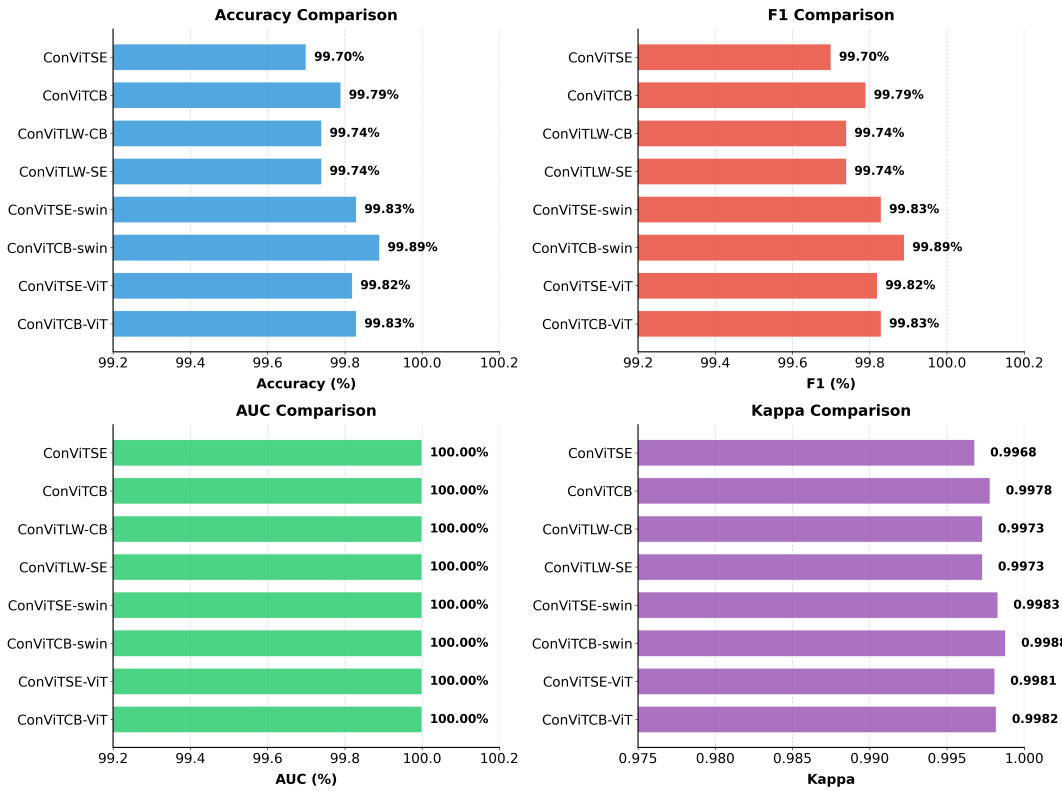
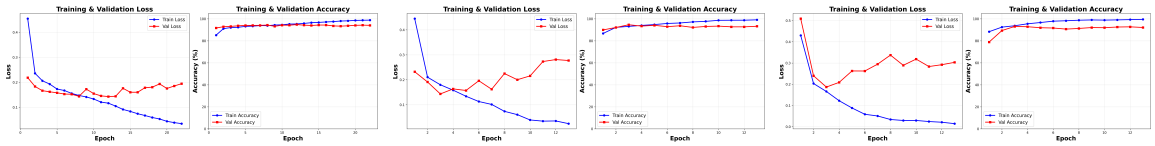


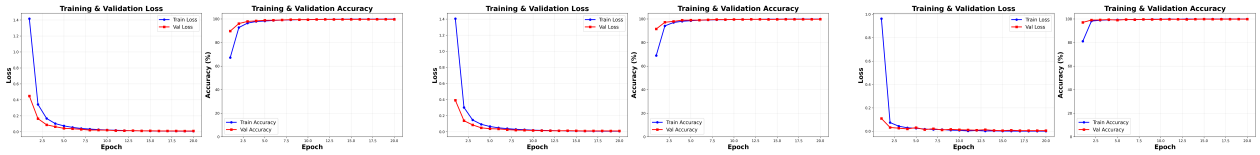
Figure 4: Comparison of performance metrics across the model variants on PlantVillage.



(a) CViTLw-SE (Maize) **(b)** CViTCB-swin (Maize) **(c)** CViTCB-ViT (Maize)

Figure 5: Training convergence on the Maize dataset. The lightweight model shows more stable optimization, whereas the ViT-B/16-based variant exhibits earlier validation instability.

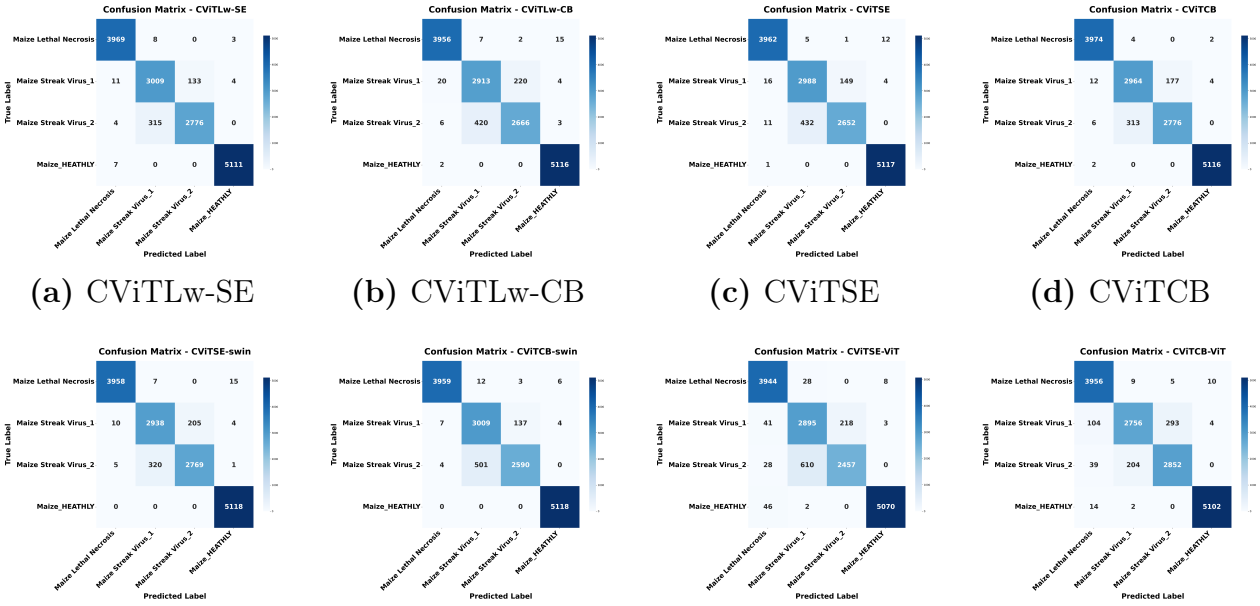
these controlled conditions, Swin-based models gain a slight benefit from their larger representational capacity. However, the performance gap among the different configurations is noticeably smaller than what was observed on the Maize dataset.



(a) CViTLw-SE (PlantVillage) **(b)** CViTCB (PlantVillage) **(c)** CViTCB-swin (PlantVillage)

Figure 6: Training convergence on PlantVillage. Because the benchmark is more controlled, optimization remains more stable across all model scales.

To examine prediction behavior on the Maize dataset in more detail, Figure 7 presents the confusion matrices of all eight variants. The most difficult distinction involves *Maize Streak Virus 1* and *Maize Streak Virus 2*. These two conditions share highly similar visual symptoms. Despite this challenge, the custom ViT variants like CViTLw-SE experience fewer cross-classification errors than the larger Transformer models. This observation reinforces the idea that an efficiently parameterized, domain-aligned design can establish stronger decision boundaries in practice.



(a) CViTLw-SE **(b)** CViTLw-CB **(c)** CViTSE **(d)** CViTCB

(e) CViTSE-swin **(f)** CViTCB-swin **(g)** CViTSE-ViT **(h)** CViTCB-ViT

Figure 7: Confusion matrices of the eight evaluated variants on the field Maize dataset.

5.3 Extended Ablation Study

An ablation analysis isolates the impact of individual design decisions on overall performance. Figure 8 summarizes this comparison within the precision-recall space. On the PlantVillage dataset, the more expressive architectures generally occupy the upper-right region. However, their advantage over the lightweight configurations remains relatively modest. This outcome suggests that while scaling up the model size provides some benefit on this controlled benchmark, it is not the decisive factor for success.

Effect of SE and CBAM on larger Transformer backbones. As detailed in Table 6, the CBAM module offers a small yet consistent advantage over SE for the Swin- and ViT-B/16-based variants, particularly on the Maize dataset. This finding supports the hypothesis that spatial attention proves more valuable when processing images with strong background clutter and heavy context variation.

Effect of Transformer depth. Regarding the custom-designed ViT branch, reducing the encoder depth from 8 layers to 4 achieves an optimal balance between accuracy and computational efficiency. The 4-layer SE variant secures the highest accuracy on Maize (94.85%) while maintaining a minimal footprint of only 0.33M parameters. In comparison, the 8-layer versions remain competitive but sacrifice parameter efficiency. This indicates that a deep Transformer branch is not strictly necessary to realize clear practical benefits for the present classification task.

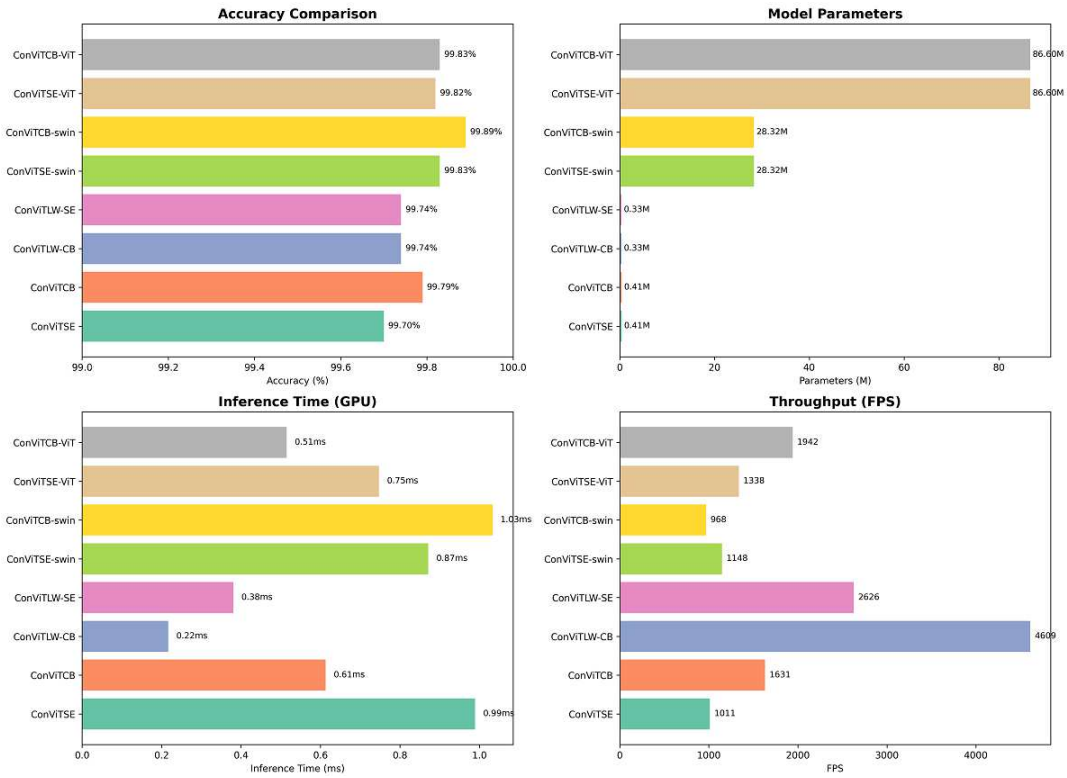


Figure 8: Summary of the ablation analysis on PlantVillage.

Table 6: Effect of CBAM versus SE on Transformer backbones for Maize and PlantVillage.

Backbone	**Maize**				**PlantVillage**			
	SE-Acc	**CB-Acc**	**SE-Kap**	**CB-Kap**	**SE-Acc**	**CB-Acc**	**SE-Kap**	**CB-Kap**
Custom (8L)	94.50	94.76	0.93	0.93	99.70	99.79	1.00	1.00
Swin-Tiny	94.01	94.69	0.92	0.93	99.83	99.89	1.00	1.00
ViT-B16	92.83	93.16	0.90	0.91	99.82	99.83	1.00	1.00

5.4 Inference Time and FPS Analysis

Beyond classification accuracy, inference latency and frames per second (FPS) serve as critical metrics, particularly when the ultimate goal involves deployment on embedded agricultural platforms like drones or automated monitoring systems. Table 7 details the runtime characteristics of all evaluated variants. This provides a clear assessment of the trade-off between predictive performance and practical deployment feasibility.

Table 7 emphasizes the practical advantages of the lightweight ViT architecture. The two CViTLw variants require only 0.33M parameters each, yet they comfortably exceed 500 FPS on the Maize dataset. On PlantVillage, their inference speed ranges from 861.0 to 1302.8 FPS with a latency consistently below 2 ms. Such results demonstrate that the proposed design successfully pairs robust predictive power with the speed required for real-time operations.

In contrast, the larger Swin-Tiny and ViT-B/16 variants suffer from noticeably higher latency, which is particularly evident on the Maize benchmark. This comparison confirms

Table 7: Inference time and FPS of the evaluated variants

(a) Inference analysis on Maize

Model	Time (ms)	FPS	Acc (%)	F1 (%)	Params (M)
CViTLw-SE	1.86	537.40	94.85	94.84	0.33
CViTLw-CB	1.98	504.60	94.59	94.56	0.33
CViTCB	2.45	408.80	94.76	94.74	0.41
CViTSE-ViT	4.06	246.10	92.83	92.77	86.59
CViTCB-ViT	4.11	243.60	93.16	93.09	86.59
CViTSE-swin	5.37	186.30	94.01	93.99	28.32
CViTSE	6.19	161.60	94.50	94.46	0.41
CViTCB-swin	6.69	149.40	94.69	94.65	28.32

(b) Inference analysis on PlantVillage

Model	Time (ms)	FPS	Acc (%)	F1 (%)	Params (M)
CViTLw-CB	0.77	1302.80	99.74	99.74	0.33
CViTLw-SE	1.16	861.00	99.74	99.74	0.33
CViTSE-swin	1.45	690.50	99.83	99.83	28.32
CViTCB-swin	1.64	611.20	99.89	99.89	28.32
CViTCB	1.81	553.90	99.79	99.79	0.41
CViTSE	2.02	495.70	99.70	99.70	0.41
CViTCB-ViT	3.14	318.60	99.83	99.83	86.60
CViTSE-ViT	3.29	304.30	99.82	99.82	86.60

that the primary strength of CViTLw lies in securing a highly favorable balance between accuracy, efficiency, and deployability, rather than simply maximizing the sheer scale of the model.

5.5 Comparison with State-of-the-Art Methods

5.5.1 Shared Benchmark and Contextual Comparison

Previous research on field maize disease has not yet converged on a single universal benchmark. Therefore, we split our comparative analysis into two distinct levels. Table 9 presents results on PlantVillage, which serves as a widely standardized benchmark in numerous earlier studies. In contrast, Table 8 functions strictly as a contextual reference for related works on field maize disease. This distinction is necessary because those studies vary significantly in their data sources, label definitions, augmentation protocols, and overall evaluation setups.

Within the PlantVillage evaluation (Table 9), where all compared studies utilize the same benchmark, the proposed variants exhibit highly competitive results. Specifically, both CViTLw-CB and CViTLw-SE attain 99.74% across accuracy, precision, recall, and F1 metrics. They also yield an AUC of 100.0% and match the highest Kappa score reported, all while operating on a compact architecture of just 0.33M parameters. When compared to Thakur et al. [29], who introduced a lightweight hybrid model with roughly 0.7M parameters, the

Table 8: Contextual comparison with related studies on field maize disease tasks.

Method	Loss	AC(%)	PR(%)	RE(%)	F1(%)	AUC(%)	Kappa
Li et al. [24]	0.62	86.05	86.05	86.05	86.05	94.34	0.81
Chen et al. [22]	0.49	83.72	85.71	83.72	84.70	97.17	0.78
Yang et al. [27]	0.39	86.05	86.05	86.05	86.05	99.22	0.81
Karthik et al. [20]	2.19	58.00	60.00	41.86	49.31	76.60	0.45
Chen et al. [26]	1.26	83.72	83.72	83.72	83.72	93.17	0.78
Chen et al. [21]	0.43	86.05	86.05	86.05	86.05	96.90	0.81
Thakur et al. [25]	0.67	86.05	86.05	86.05	86.05	95.65	0.81
Thai et al. [23]	0.62	72.09	73.68	65.12	69.14	92.82	0.63
Thakur et al. [28]	0.39	93.02	93.02	93.02	93.02	97.37	0.91
Thakur et al. [29]	0.19	93.02	93.02	93.02	93.02	99.47	0.91
CViTLw-CB (Proposed)	0.14	94.59	94.67	94.59	94.56	99.29	0.93
CViTLw-SE (Proposed)	0.14	94.85	94.91	94.85	94.84	99.31	0.93

Table 9: Comparison with representative state-of-the-art methods on the PlantVillage benchmark. The proposed models achieve highly competitive performance among the reported results.

Method	Loss	AC(%)	PR(%)	RE(%)	F1(%)	AUC(%)	Kappa
Li et al. [24]	0.46	86.51	88.85	85.08	86.92	98.90	0.86
Chen et al. [22]	0.17	96.68	97.49	95.83	96.64	99.26	0.97
Yang et al. [27]	0.27	96.99	97.35	96.71	97.03	99.93	0.97
Karthik et al. [20]	0.16	95.83	96.20	95.60	95.89	99.70	0.96
Chen et al. [26]	0.05	99.19	99.19	99.19	99.19	99.85	0.99
Chen et al. [21]	1.07	74.63	80.92	69.64	75.03	98.18	0.74
Thakur et al. [25]	0.07	98.61	98.24	98.33	98.28	99.01	0.98
Thai et al. [23]	0.28	96.46	99.33	92.44	95.76	99.65	0.96
Thakur et al. [28]	0.04	98.86	98.90	98.81	98.85	99.92	0.99
Thakur et al. [29]	0.02	99.63	99.65	99.63	99.64	99.96	0.99
CViTLw-CB (Proposed)	**0.01**	**99.74**	**99.74**	**99.74**	**99.74**	**100.0**	**0.99**
CViTLw-SE (Proposed)	**0.01**	**99.74**	**99.74**	**99.74**	**99.74**	**100.0**	**0.99**

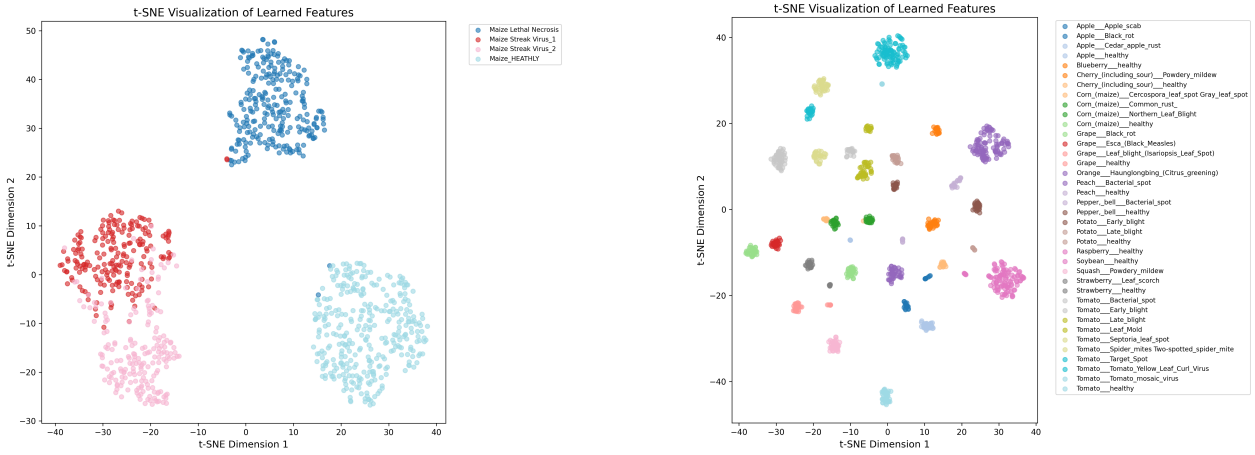
CViTLw formulations cut the parameter requirement in half. At the same time, they sustain comparable or slightly superior classification capabilities on the identical dataset. When paired with the earlier complexity analysis, these findings indicate that CViTLw provides an excellent balance between performance and parameter efficiency. It is important to note that these metrics do not imply absolute superiority over all existing research, as disparities in data splitting, preprocessing, and augmentation can influence direct comparisons. Rather, the PlantVillage outcomes confirm that our proposed architecture holds its ground on a widely adopted standard.

When examining the field Maize data (Table 8), literature-reported figures require a more cautious interpretation. Certain recent investigations, such as the work by Thakur et al. [29], rely on proprietary self-collected datasets, alternative label spaces, and occasionally synthetic imagery. Consequently, they do not offer a direct one-to-one comparison with the Mduma2023 benchmark utilized in our study. For this reason, Table 8 serves primarily as a contextual anchor to help position CViTLw against broader trends. Viewed through this lens, the 94.85% accuracy achieved by CViTLw-SE stands out. It demonstrates that a lightweight hybrid network with a mere 0.33M parameters can preserve robust discriminative power on noisy, real-world agricultural imagery. This size is significantly lower than the nearly 0.7M parameters detailed by Thakur et al. [29]. Thus, these findings validate the core argument that an ultra-compact hybrid design remains highly effective under field conditions. The cross-fusion mechanism enables local CNN features and global Transformer context to mutually reinforce one another, producing an optimal mix of high accuracy and practical deployability.

5.6 Feature Space Analysis and Explainability

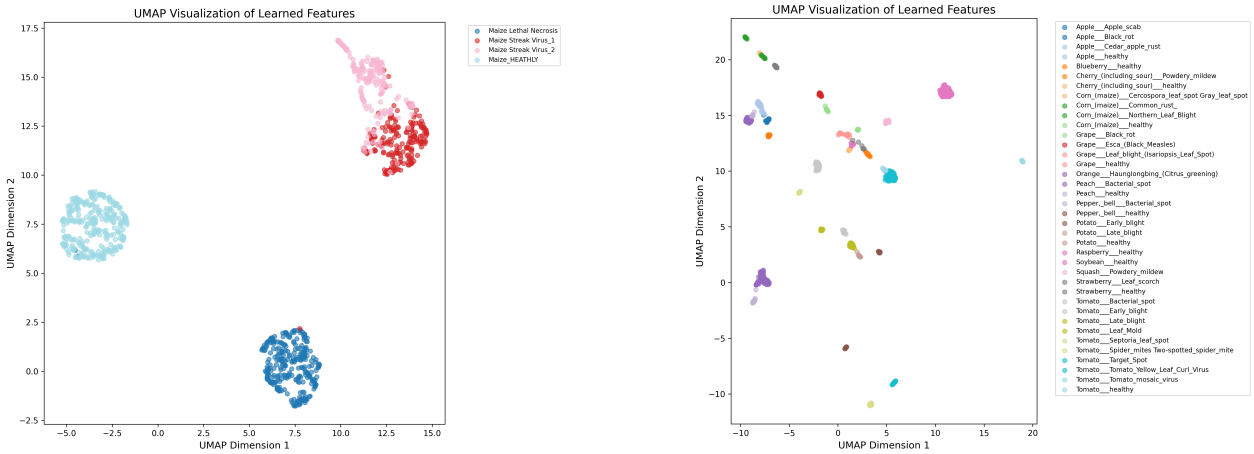
5.6.1 Class Clustering with t-SNE and UMAP

To evaluate the quality of the 128-dimensional feature space immediately prior to the classification layer, we visualize the learned representations utilizing t-SNE and UMAP (Figures 9 and 10).



(a) t-SNE (Maize) **(b)** t-SNE (PlantVillage)

Figure 9: Visualization of the learned feature space using t-SNE. The proposed representation forms clearly separated clusters even for disease classes with similar visual characteristics.



(a) UMAP (Maize)

(b) UMAP (PlantVillage)

Figure 10: Visualization of the learned representations using UMAP. The compact 128-dimensional representation still preserves meaningful manifold structure on both datasets.

5.6.2 Explainability Analysis with LIME and Grad-CAM

To investigate model interpretability, we scrutinize the decisions made by CViTLw-SE utilizing LIME, Grad-CAM, and Grad-CAM++. The gradient-based maps are extracted from the deepest convolutional layer within the attention-enhanced fusion block, as this layer acts as the primary spatial anchor prior to final classification. Figures 11 and 12 contrast superpixel-based LIME explanations with gradient-based activation maps across representative maize samples.

The visual evidence demonstrates strong alignment between the two explainer techniques. When processing diseased leaves, the highlighted attention regions consistently localize around lesion structures and symptomatic veins, successfully ignoring the surrounding background. Conversely, for healthy leaves, the model's responses appear appropriately broad and diffuse, showing no fixation on spurious local anomalies. These observations confirm that the network's predictive logic relies predominantly on biologically relevant visual cues rather than arbitrary background correlations.

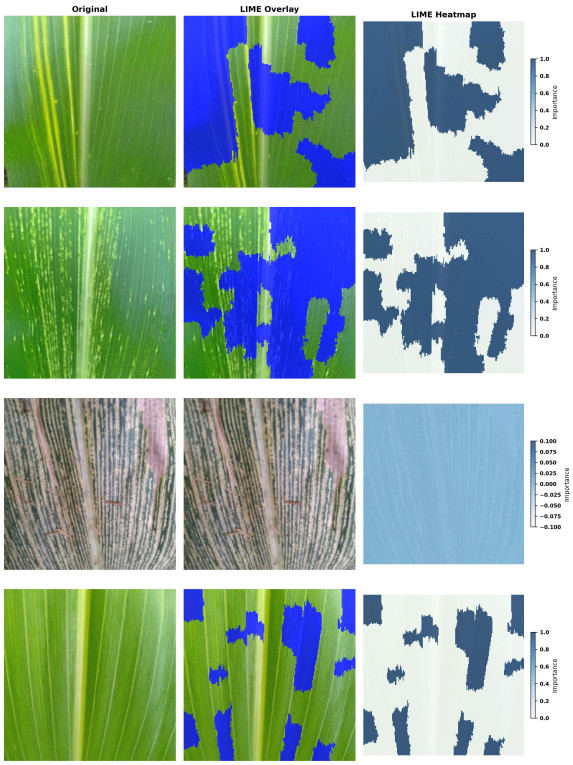


Figure 11: LIME explanations for representative samples from the four classes of the Maize dataset.

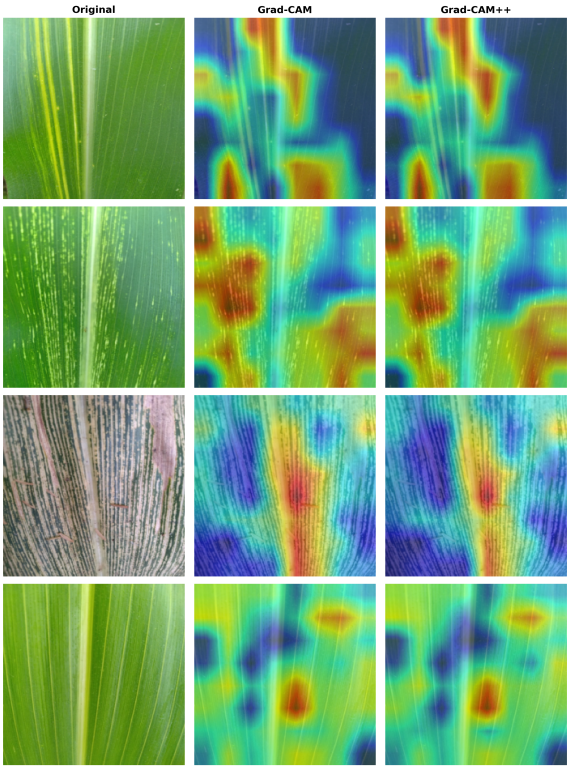


Figure 12: Grad-CAM visualization for representative disease groups in the Maize dataset.

6 Discussion

6.1 Complexity and Performance Trade-Off

The experimental data clearly illustrates that merely increasing model capacity does not yield monotonic performance improvements in plant disease classification, particularly when processing images captured directly in the field. Within the Maize dataset, where samples exhibit considerable background noise, unpredictable illumination, and high visual overlap among disease classes, heavy architectures like ViT-B/16 fail to deliver benefits that justify their computational overhead. Conversely, the CViTLw models secure equal or superior outcomes by employing a representational structure tailored to the specific nature of the task. This phenomenon aligns with established principles in deep learning: when dealing with constrained target-domain datasets, a well-calibrated inductive bias frequently outweighs raw parameter count.

Our quantitative runtime analysis corroborates this perspective. By achieving inference speeds surpassing 1300 FPS on the PlantVillage dataset and exceeding 500 FPS on the Maize benchmark, the CViTLw networks seamlessly satisfy the strict demands of real-time operation on resource-limited hardware. Consequently, the core contribution of this research is not an indiscriminate pursuit of maximum accuracy, but rather the realization of an optimal balance among predictive precision, execution latency, and architectural complexity.

6.2 Impact of Hybrid Architecture and Attention Mechanisms

The robust performance of CViTLw originates directly from the synergistic relationship between its two representational branches. The MobileNetV2 pathway excels at isolating

highly localized lesion patterns, irregular boundaries, and fine textural abnormalities. Simultaneously, the compact Transformer branch maps the broader spatial relationships linking disparate image regions. By routing both signal types through a dedicated cross-fusion module prior to final classification, the network constructs a far more balanced feature representation than systems relying heavily on a single paradigm. From the standpoint of representation learning, this fusion effectively anchors fine-grained symptom features within their wider global context. Such spatial grounding proves essential when disease diagnosis hinges on both the morphological traits of a lesion and its relative distribution across the leaf surface.

Furthermore, the ablation findings demonstrate that the utility of specific attention mechanisms is highly data-dependent. The CBAM module proves particularly beneficial under field conditions, where its spatial attention component actively suppresses background clutter and accentuates target lesions. On the other hand, within the controlled environment of PlantVillage (which features uniform backgrounds and clearly defined symptoms), the channel-wise recalibration provided by SE is generally sufficient. This behavioral split indicates that the choice of an attention mechanism should not be treated as a universal default, but must instead be carefully matched to the statistical realities of the intended application domain.

6.3 Limitations and Future Work

While the experimental outcomes are highly promising, several limitations warrant acknowledgment. Primarily, the existing framework focuses strictly on single-label classification, lacking an explicit mechanism to handle instances where a single leaf displays concurrently overlapping diseases. Additionally, the pronounced visual similarity between specific conditions, most notably MSV1 and MSV2, continues to generate challenging decision boundaries. Furthermore, although the Maize dataset constitutes a public field benchmark, previous literature targeting similar agricultural scenarios has not yet adopted a unified testing protocol. Consequently, the comparative metrics provided for the field maize literature serve mainly as contextual references rather than perfect one-to-one comparisons. Finally, the current suite of evaluations was conducted in an offline environment; subsequent validation involving live video streams directly on embedded mobile hardware will be crucial to fully solidify the deployment claims.

Future research avenues will explore advanced quantization, structural pruning, and hardware-specific compilation strategies to minimize computational demands even further on edge platforms. It will also be highly advantageous to adapt the network for multi-label disease detection, execute repeated-run statistical validations across varied initialization seeds, and rigorously test the model against a broader, more diverse collection of challenging field datasets.

7 Conclusion

This research introduces CViTLw, a lightweight CNN-Transformer architecture equipped with attention-enhanced cross-fusion designed specifically for plant disease classification. Rather than indiscriminately scaling up the parameter count of the backbone, CViTLw smartly merges highly localized feature maps extracted by MobileNetV2 with broader contextual representations formed by a compact Vision Transformer. The overarching objective revolves around attaining a superior balance among predictive reliability, structural simplicity, and practical deployment capability.

Comprehensive experiments validate this architectural strategy. Evaluated on the challenging Maize field dataset, the CViTLw-SE model secured an impressive 94.85% accuracy

utilizing a mere 0.33M parameters, simultaneously ensuring ultra-low latency alongside high inference speeds. Similarly, on the controlled PlantVillage corpus, the network achieved a remarkable 99.74% across accuracy, precision, recall, and F1 metrics without inflating model complexity. In a direct contextual evaluation against recent lightweight hybrid models featuring approximately 0.7M parameters, CViTLw demonstrated significantly superior parameter efficiency. Collectively, these empirical findings confirm that the architecture successfully navigates the critical performance-to-complexity trade-off.

Finally, independent visualizations generated via Grad-CAM++, LIME, t-SNE, and UMAP indicate that the learned feature space closely correlates with biologically relevant lesion zones while maintaining distinct and logical class clusters. In conclusion, the robust outcomes of this study firmly advocate for the integration of highly compact, deliberately structured hybrid neural networks within precision agriculture, particularly in scenarios demanding both exceptional diagnostic accuracy and seamless edge deployment.

Declarations

Ethics declaration: Not applicable.
Consent to Participate declaration: Not applicable.
Consent to Publish declaration: Not applicable.
Funding: This research received no external funding.
Data availability: The datasets used and/or analyzed during the current study are cited in the manuscript.
Competing interests: The authors declare that they have no competing interests.
Acknowledgements: We acknowledge the support of time and facilities from Tra Vinh University, Tien Giang University, An Giang University for this study.
Author contributions: Thanh-Hai Tong-Le: Conceptualization, original draft preparation, and analysis. Minh-Hai Le: Writing, review, and editing. Thanh-Nghi Doan: Review, editing, and supervision.
Use of Generative AI in writing: The authors acknowledge the use of AI-based tools to assist with editing, grammar improvement, and spelling checks during the preparation of this manuscript. The authors take full responsibility for the content of this manuscript.

References

[1] Savary, S., Willocquet, L., Pethybridge, S. J., Esker, P., McRoberts, N., & Nelson, A. (2019). The global burden of pathogens and pests on major food crops. *Nature Ecology & Evolution*, 3(3), 430–439.

[2] Mohanty, S. P., Hughes, D. P., & Salathé, M. (2016). Using deep learning for image-based plant disease detection. *Frontiers in Plant Science*, 7, 1419.

[3] Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks. *ICLR*.

[4] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *CVPR*, 770–778.

[5] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. *CVPR*, 4510–4520.

[6] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words. *ICLR*.

[7] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer. *ICCV*, 10012–10022.

[8] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. *CVPR*, 7132–7141.

[9] Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. (2018). CBAM: Convolutional block attention module. *ECCV*, 3–19.

[10] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). MobileNets. *arXiv:1704.04861*.

[11] Ferentinos, K. P. (2018). Deep learning models for plant disease detection. *Comp. Electron. Agric.*, 145, 311–318.

[12] Hughes, D. P., & Salathé, M. (2015). An open access repository of images on plant health. *arXiv:1511.08060*.

[13] Tan, M., & Le, Q. V. (2019). EfficientNet. *ICML*, 6105–6114.

[14] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. *ICML*, 10347–10357.

[15] Wang, W., Xie, E., Li, X., Fan, D. P., Song, K., Liang, D., Lu, T., & Luo, P. (2021). Pyramid Vision Transformer. *ICCV*, 568–578.

[16] Mehta, S., & Rastegari, M. (2022). MobileViT. *ICLR*.

[17] Dai, Z., Liu, H., Le, Q. V., & Tan, M. (2021). CoAtNet. *NeurIPS*, 34, 3965–3977.

[18] Graham, B., El-Nouby, A., Touvron, H., Stock, P., Sablayrolles, A., Grangier, G., & Jégou, H. (2021). LeViT. *ICCV*, 12259–12269.

[19] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *NeurIPS*, 30, 5998–6008.

[20] Karthik, R., Hariharan, M., Anand, S., Mathikshara, P., Johnson, A., & Menaka, R. (2020). Attention embedded residual CNN for disease detection in tomato leaves. *Appl. Soft Comput.*, 86, 105933.

[21] Chen, J., Zhang, D., Suzauddola, M., & Nanehkaran, Y. A. (2021). Identifying crop diseases using attention embedded MobileNet-V2 model. *Appl. Soft Comput.*, 113, 107901.

[22] Chen, J., Zhang, D., Zeb, A., & Nanehkaran, Y. A. (2021). Identification of rice plant diseases using lightweight attention networks. *Expert Syst. Appl.*, 169, 114514.

[23] Thai, H. T., Tran-Van, N. Y., & Le, K. H. (2021). Artificial cognition for early leaf disease detection using vision transformers. *IEEE Int. Conf. Adv. Technol. Commun.*, 33–38.

[24] Li, H., Li, S., Yu, J., Han, Y., & Dong, A. (2022). Plant disease and insect pest identification based on vision transformer. *Int. Conf. Internet Things Mach. Learn.*, 194–201.

[25] Thakur, P. S., Khanna, P., & Ojha, A. (2022). Vision transformer for plant disease detection. *Int. Conf. Comput. Vis. Image Process.*, 501–511.

[26] Chen, J., Chen, W., Zeb, A., Yang, S., & Zhang, D. (2022). Lightweight inception networks for the recognition and detection of rice plant diseases. *IEEE Sensors J.*, 22(14), 14628–14638.

[27] Yang, L., Yu, X., Zhang, S., Long, H., Zhang, H., Xu, S., & Liao, Y. (2023). GoogLeNet based on residual network and attention mechanism identification of rice leaf diseases. *Comput. Electron. Agric.*, 204, 107543.

[28] Thakur, P. S., Chaturvedi, S., Khanna, P., & Sheorey, T. (2023). Vision transformer meets convolutional neural network for plant disease classification. *Ecological Informatics*, 77, 102245.

[29] Thakur, P. S., Chaturvedi, S., Seal, A., Khanna, P., Sheorey, T., & Ojha, A. (2025). An ultra lightweight interpretable convolution-vision transformer fusion model for plant disease identification: ConViTX. *IEEE Transactions on Computational Biology and Bioinformatics*, 22(1), 310–321.

[30] Mduma, N., & Laizer, H. (2023). Machine learning imagery dataset for maize crop: A case of Tanzania. *Data in Brief*, 48, 109108.