RESEARCH ARTICLE

INTERNATIONAL RESEARCH JOURNAL OF MULTIDISCIPLINARY TECHNOVATION

Asian Research Association

DOI: 10.54392/irjmt26610

Hybrid Vision Transformer for Plant Disease Detection from Remote Sensing Images

M. Sindhuja [a,*], I. Manimozhi [a]

[a] Department of Computer Science Engineering, Eastpoint College of Engineering & Technology, Bangalore, 560049, India.

* Corresponding Author Email: sindhujadsc@gmail.com

DOI: https://doi.org/10.54392/irjmt26610

Received: 06-04-2026; Revised: 29-10-2026; Accepted: 11-11-2026; Published: 26-11-2026

Abstract: Spread of plant diseases and the resulting crop damage/yield reduction affects agricultural productivity and threaten global food security. Thus timely and accurate disease detection is important to control the damage. Collecting leaf samples randomly from farm and analyzing the leaf sample images is the most used method for plant disease detection. But this method of manual sample collection and analysis is not effective for analyzing disease spread over large farms and assessing crop damage at scale. Analysis of remote sensing (RS) images collected via drones, satellites and other platforms can provide a broad perspective of crop health and spatial disease distribution. Existing deep learning techniques for disease detection used leaf sample images collected at closer range and not suitable for RS images. This work proposes a light weight hybrid vision transformer (LWHVT) which integrates the features collected from multiple crop RS images through drones to detect disease and assess the damage. The proposed technique combines convolutional layers for local feature extraction with transformer attention mechanism for global dependency modeling. In addition a feature fusion strategy is proposed to aggregate multi-temporal and multi-view RS data for robust disease detection. Furthermore, a feature fusion strategy is introduced to integrate multi-temporal and multi-view RS data, thereby improving the robustness and reliability of disease identification under varying imaging conditions. Through experimental analysis, the proposed LWHVT is found to provide about 3.1% higher accuracy compared to recent deep learning models for plant disease detection.

Keywords: Crop Disease, Remote Sensing Images, Convolutional Neural Network, Light weight Hybrid Vision Transformer

1. Introduction

Increasing the agricultural yield is important to ensure global food security and economic stability. Plant disease spread presents a significant challenge to this goal. They contribute to loss of about 220 billion USD annually [1]. Thus detecting the disease spread early and controlling it becomes important. Manual farm inspection and laboratory based analysis of randomly collected leaf samples were the conventional methods for detecting plant diseases. Various machine learning and deep learning techniques have been proposed to analyze the leaf samples for disease and nutritional deficiency detection. Though these methods provide micro level disease analysis, leaf sample collection is labor intensive, time consuming making them unsuitable for large scale agricultural environments [2, 3]. The disease detection accuracy depends on the leaf samples collected and recent studies [1, 4] have found a significant gap between laboratory-based performance and real world deployment. This limitation is caused by dependence on close-range leaf images, which fail to capture the spatial distribution of diseases across large agricultural fields. Without expert guidance on leaf sample selection, the disease detection effectiveness is reduced. Since close-range leaf imaging captures only a limited portion of the crop canopy, it often fails to represent the spatial variability of disease symptoms across extensive agricultural fields. In addition, manual sample selection requires expert knowledge, making large-scale disease surveillance both expensive and difficult to automate [5, 6]. Remote sensing (RS) techniques like drones, aerial vehicles have emerged as promising solution for sample collection. Analysis of RS images can provide a detailed view of crop health and disease spread over large farmland.

High-resolution UAV imagery enables rapid, non-destructive, and cost-effective assessment of crop health while providing comprehensive spatial coverage of agricultural fields [7–10]. Compared with conventional sampling-based approaches, UAV-based remote sensing facilitates continuous monitoring of disease development, supports precision farming practices, and enables timely management decisions through automated image analysis [8, 11]. Consequently, deep

learning techniques have been increasingly adopted for analysing UAV imagery because of their ability to automatically learn discriminative visual representations without manual feature engineering [9–12]. Recent studies have demonstrated the effectiveness of convolutional neural networks (CNNs), including ResNet-50, VGG16, EfficientNet, and MobileNet variants, for UAV-based crop disease detection [10–14]. Most of these approaches have used only single RS image without consideration for local and global spatial context spread across multiple RS images gathered for the same crop across the farm. [13–15]. Very few works has also used vision transformer (ViT) which can effectively model long-range dependencies using self-attention mechanisms. Compared to CNN variants, ViT provides improved detection effectiveness but they consume more resources and require large-scale datasets. Considering these two problems of lack of robust feature representation considering both local and global spatial features with modeling long range global dependencies and a lack of effective light weight ViT model, this work proposes a lightweight hybrid vision transformer framework for detecting plant diseases. The proposed model combines convolutional layers for efficient local feature extraction with transformer-based attention mechanisms to capture global spatial dependencies across large agricultural regions. Furthermore, a feature fusion strategy is introduced to combine multi-source RS data, enabling robust detection of disease patterns.

The proposed solution has following contributions

(i) A lightweight hybrid architecture that jointly exploits local spatial representations extracted using depthwise separable convolutions and global contextual representations learned through a Vision Transformer, specifically designed for sequential UAV crop imagery

(ii) An attention-guided multi-image feature aggregation mechanism that adaptively weights features extracted from consecutive UAV images before transformer encoding, allowing the model to exploit complementary spatial information across an image sequence.

(iii) A lightweight hierarchical feature fusion strategy that integrates shallow convolutional features with transformer representations while maintaining low computational complexity.

The paper is structured as follows. The existing deep learning works on detecting plant diseases and their issues are detailed in Section II. Section III details the proposed hybrid vision transformer for disease detection from RS images. Following it the results of the proposed hybrid model and its comparison to existing works are detailed in Section IV. Final section summarizes the work and provides concluding remarks.

2. Literature Survey

The survey is conducted in multiple categories which are presented below.

2.1 UAV Based Plant Disease Detection

UAV based remote sensing helps crop-health assessment over larger areas. Various works have been proposed for processing the RS images acquired by UAV and detect plant diseases. Yang *et al* [3] classified plant diseases using MobileNetV2 model. De Ocampo *et al* [4] estimated chlorophyll levels of plant using RS image. Lacerda *et al* [5] detected vegetation from RS images using generative adversarial networks (GAN). Nejad *et al* [6] predicted crop yield from RS images using vision transformers. Though most works considered both local and global features and fused them at multi scale level for increasing the detection accuracy, all the works were based on single image and did not consider jointly learning global features across multiple images.

2.2 Hybrid Deep Learning Architectures

Zhang *et al* [7] presented a review of various deep learning architectures of CNN, Vision transformers and Hybrid CNN-Vision transformers models for plant disease detection. Authors found that hybrid ViT architectures enhanced the global semantic understanding in complex backgrounds while maintaining the sensitivity of the local details. Farmonov *et al* [8] proposed a hybrid deep learning framework called HypsLiDNet. Authors augmented hyper spectral images with LiDAR detail to improve the structural details. Augmenting increased the detection accuracy for images acquired in closer view. Hajare *et al* [9] used a ensemble of deep belief network, recurrent neural network and CNN to increase disease detection accuracy. But all these architectures were tested for Smartphone acquired images and their performance for RS images need to be tested. Also these methods did not exploit disease characteristics in fine tuning the model for higher detection accuracy. Chang *et al* [11] proposed a hybrid neural network combining transformer and CNN for segmenting the RS images. But the effectiveness of this network in segmentation of diseases regions in plants was not tested. Zhu *et al.* [17] reviewed deep learning approaches for UAV-based crop disease and pest detection, highlighting the importance of image characteristics, disease symptom variability, and model selection in achieving higher disease detection accuracy. The same challenge has been identified in the works of Radočaj *et al.* [18] where in-depth review deep learning approaches for UAV based croop disease detection was presented. Mehta and Rastegari [20] introduced MobileViT combining CNN with transformer for mobile vision applications. Though the work demonstrated the feasibility of combining local spatial inductive biases with global attention in

computationally constrained settings, it was not tested for plant disease detection. Li *et al.* [21] proposed PMVT, a lightweight MobileViT-based architecture incorporating an inverted residual structure and convolutional block attention. The model was evaluated on wheat, coffee, and rice disease datasets, demonstrating that Transformer-based disease recognition can be designed for resource-constrained devices. But the model works for only single image and does not consider multiple ordered UAV images, with the objective of reducing dependence on the information available in an individual observation.

2.3 Attention Techniques

Zhang *et al* [12] detected diseased from crop images by exploiting the multi scale features with attention mechanism. Extracting features at multiple scales and integrating the features captures both the characteristics and contextual information of pests across different scales. But effectiveness of the approach for RS image is not tested. Soeb *et al* [13] detected tree leaf diseases using an improved YOLOv7 model. This improved model performed better compared to YOLOV5 model. The accuracy is highly dependent on shooting angles and thus suitability of the approach for RS images needs to be explored. Similarly Dong et al [14] extended YOLOv5 model for detecting crop diseases. YOLOv5 was added with Multi-Level Spatial Pyramid Pooling, channel Attention mechanisms and Content-Aware Reassembly of Features to enhance contextual understanding. Lee *et al* [15] used a modified Resnet50 model to detect pine-wilt diseased trees from time series UAV images. This improved model was able to recover the object boundaries effectively which is important functionality necessary while processing RS images. Temporal modeling was done only for region selection in this work. Zhang *et al* [19] proposed a adaptive-weighted attention VM-UNet framework for grape disease detection. Authors used multiscale feature aggregation and spatial-channel attention to improve disease-region representation. But this method works only for single image. The summary of the survey is given in Table 1.

Most of the existing works using CNN, ViT and hybrid transformers were designed for close-range leaf images rather than remote sensing (RS) data. Though some studies utilized drone or satellite imagery, their approaches are limited to single-image analysis, lack multi-image feature integration, or fail to capture disease spread across large agricultural areas. Though some multimodal and temporal models exist, they are primarily applied to yield prediction or classification tasks rather than disease detection from RS images. Thus there is a need for a lightweight hybrid model that can fuse multiple RS images, capture both local and global features. The model should facilitate scalable disease detection and damage assessment over large farms. Table 2 provides the functionalities realized in proposed LWHVT addressing the limitations in existing works.

3. Proposed Method

A lightweight hybrid vision transformer (LWHVT) for detecting plant diseases using remote sensing (RS) images acquired from drone platforms is proposed in this work. The proposed solution integrates local feature extraction, global contextual learning, and multi-image feature fusion in a computationally efficient manner for efficient disease detection and damage assessment. The proposed solution has following components (a) Image preprocessing (b) hybrid feature extraction and (c) disease detection. The interaction between the components is illustrated in Figure 1. The input RS images are preprocessed at first stage to remove noise and improve contrast. The preprocessed images are passed to hybrid feature extraction where a hybrid ViT is used to extract both local and global features. The features across multiple RS images are fused. The fused features are then classified by the softmax classifier to disease class.

Each of the components are detailed in below subsections. The notations used in the equations are given in Table 3.

3.1 Image preprocessing

RS images collected from drones covering large agricultural fields vary in resolution, illumination and perspective. Thus the RS images need to be preprocessed before passing it to next stage of hybrid feature extraction. The image is subjected to following preprocessing steps of: image resizing/normalization, Gaussian noise filtering and contrast enhancement for highlighting vegetation patterns. Let the RS images captured by drone be given as

$$I = \{I_1, I_2, \dots, I_N\},\ \ I_i \epsilon \mathcal{R}^{H \times W \times C} \quad (1)$$

Where H, W, C denote height, width and number of channels respectively. Each image is preprocessed and transformed as

$$\bar{\bar{I}}_i = T(i) \quad (2)$$

Where T is the image preprocessing operation.

Table 4 presents the summary of the configuration parameters used for image pre-processing operation.

3.2 Hybrid Feature Extraction

Hybrid architecture combining CNN and ViT is used for capturing both fine grained and global features. A light weight CNN model is used in this work for extracting textural features, leaf patterns and disease spots.

Table 1. Survey summary

Ref	Method employed	Dataset used	Principal findings	Technical limitations	Research gap
[3]	Improved MobileNetV2	PlantVillage dataset : has images of various crops	Developed a CNN-based system for crop-health assessment using drone imagery.	CNN-based feature extraction may not explicitly model long-range relationships across a sequence of UAV images.	Did not integrate efficient local feature extraction with adaptive multi-image fusion and global contextual learning.
[4]	Multi-vision framework for real-time UAV monitoring	UAV monitoring data not published	Combines UAV-based monitoring of agricultural activities and crop health.	Its stated focus is broad agricultural monitoring rather than the specific integration of disease classification and sequence-level feature fusion.	Did not consider a disease-focused model that combines local symptoms and contextual information across images.
[5]	AgI-GAN and AI-driven analytics for cost-efficient, high-resolution specialty-crop monitoring	Agroview dataset not published	Investigates high-resolution crop monitoring using AI-driven analytics.	It did not consider adaptive attention-based fusion across sequential images	Sequence-level feature aggregation for improved accuracy is not considered.
[6]	ConvLSTM, lightweight 3D-CNN, and Vision Transformer	Soybean yield sata, MODIS/terra surface reflectance, MODIS land cover, MODIS land surface temperature	Models spatiotemporal information for crop-yield prediction	Crop-yield prediction is a different task from direct plant-disease detection; the model's original objective may not transfer directly to disease classification.	Adapt temporal and global-context modeling to UAV-based disease detection while controlling computational cost.
[7]	Review of CNNs and Vision Transformers for agricultural pest and disease identification	–	Summarizes the evolution from convolutional models to transformer-based agricultural vision	Review does not itself provide a unified model or a controlled comparison under a common UAV sequence-level protocol.	Empirically evaluate a unified lightweight CNN–Transformer architecture on a consistent dataset and split.
[8]	HypsLiDNet: 3D–2D CNN with spatial–spectral morphological attention	DESIS hyperspectral imagery and LiDAR data. Not published	Combines spatial and spectral information for crop classification	Requires specialized spectral and LiDAR inputs; crop classification is not necessarily equivalent to disease classification from RGB UAV images.	Develop an efficient approach compatible with standard RGB imagery while learning contextual relationships across images.
[9]	Deep ensemble with optimal training for black gram disease classification	Black gram disease dataset	Investigates ensemble learning for disease classification	Ensemble methods may increase inference cost; the reference entry does not establish sequence-level UAV processing.	Explore a lightweight architecture that integrates multiple images without relying on a costly ensemble.
[10]	ECA-SegFormer for cucumber leaf disease-spot semantic segmentation	Cucumber leaf disease dataset	Uses efficient channel attention with a SegFormer-based segmentation model to identify disease spots.	Focuses on segmentation of leaf-level disease regions rather than adaptive fusion across UAV image sequences.	Combine disease-relevant local features with multi-image contextual representations.
[11]	UNeXt for efficient semantic segmentation of high-resolution	Vaihingen and GID5 dataset.	Addresses computational efficiency in high-resolution image segmentation	General remote-sensing segmentation does not by itself establish effectiveness for crop-disease recognition or	Adapt efficient feature extraction and contextual learning to agricultural disease recognition

	remote-sensing images			sequential UAV images.	
[12]	AM-MSFF: attention mechanism and multiscale feature fusion for pest recognition	IP102 dataset covers multiple crops of rice, corn, wheat, sugar beet, alfalfa, grapes, citrus, and mangoes. Not RS images.	Uses attention-guided multiscale feature fusion to improve pest recognition.	Multiscale fusion does not necessarily model long-range relationships across a sequence of UAV images using a Transformer encoder	Extend attention-based feature selection from multiscale features to sequence-level image representations.
[13]	YOLOv7-based tea-leaf disease detection and identification (YOLO-T)	Tea-leaf disease image dataset. Not published.	Applies an object-detection architecture to tea-leaf disease detection and identification.	The model is designed around leaf-level detection and may not explicitly aggregate contextual information across consecutive UAV images.	Investigate lightweight sequence-level feature fusion for field-scale disease recognition.
[14]	PestLite, a YOLO-based crop-pest detection method	Teddy Cup Pest Dataset + self collected images	Develops a pest-detection architecture with an emphasis on efficient detection.	Pest detection is not identical to plant-disease classification; the cited reference does not establish adaptive multi-image Transformer fusion.	contextual disease classification across UAV imagery is not considered.
[15]	Deep-learning semantic segmentation using time-series UAV imagery	Self collected Time-series UAV imagery for pine-wilt disease detection. Not published	Demonstrates the relevance of time-series UAV imagery for disease-related monitoring.	The task is semantic segmentation for pine-wilt disease; the reference does not establish the same lightweight CNN–Transformer and attention-guided fusion design as LWHVT.	Evaluate adaptive image-level weighting and global contextual modeling for multi-image disease classification
[17]	Review of deep learning for UAV-based crop disease and pest detection	_	Synthesizes UAV applications and deep-learning methods for agricultural disease and pest monitoring.	Review-level findings do not establish the performance of a particular unified architecture.	Validate a lightweight, multi-image architecture under a consistent experimental protocol.
[18]	review of UAV-based crop leaf disease detection	_	Examines the evolution of UAV-based crop leaf disease detection methods	Did not provide any discussion on integration of image-sequence fusion, lightweight feature extraction, and global context modeling.	Develop and empirically evaluate a unified architecture that addresses these components together
[19]	AWAVM-UNet with adaptive-weighted attention and multiscale feature aggregation	Self collected grape disease samples. Not published	Investigates adaptive attention and multiscale aggregation for grape-disease detection from UAV imagery.	Its U-Net-based segmentation approach and grape-specific task differ from LWHVT's stated multi-image classification framework.	Examine lightweight CNN–Transformer integration and sequence-level image weighting for disease classification.
[20]	MobileViT: lightweight mobile-friendly Vision Transformer	General-purpose vision benchmarks; not specifically a UAV crop-disease dataset	Combines convolutional operations with Transformer-based global representation learning for mobile vision.	It is a general architecture rather than a disease-specific UAV sequence-fusion method	Adapt lightweight Transformer design to agricultural sequences and explicitly fuse complementary information across images.
[21]	PMVT: MobileViT-based lightweight Transformer with	Wheat, coffee and Rice datasets.	Transformer-based disease recognition can be designed for	The reported evaluation uses agricultural image datasets, but does not establish the same	Extend lightweight attention-based feature extraction to sequence-level UAV fusion while

	a 7 × 7 convolution and CBAM attention		resource-constrained devices	attention-guided fusion of consecutive UAV images used by LWHVT.	measuring accuracy and computational cost
[22]	Review/meta-analysis of AI-based UAV pest and disease detection	–	Highlights the importance of robust evaluation and independent testing in UAV-based agricultural AI	Review findings do not validate LWHVT's own cross-location generalization	Test the proposed model on independent farms, growing seasons, and UAV acquisition conditions

Table 2. Research Gap Solutions

Research Gap	Limitations in existing studies	Proposed LWHVT solution
Lack of global contextual modelling	CNN-only architectures analyse local textures	Vision Transformer encoder captures long-range spatial dependencies
Limited exploitation of neighbouring UAV observations	Single-image classification ignores contextual information	Attention-guided multi-image sequence fusion
High computational complexity	ConvLSTM and 3D CNN substantially increase inference cost	Lightweight depthwise separable convolutions

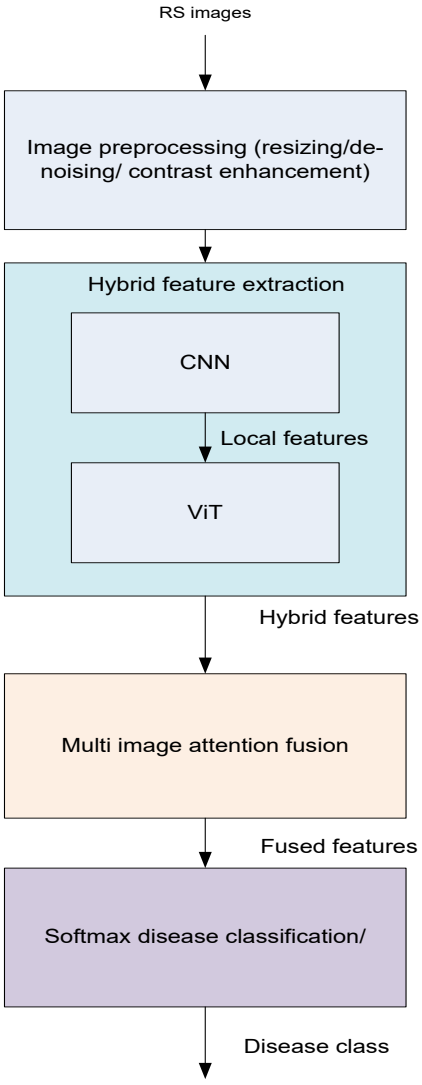


Figure 1. Component Interaction.

Table 3. Notations used in equations

Symbol	Description
N	Number of images in one UAV sequence
H, W	Image height and width
C	Number of image channels
F_i	CNN feature map of image i
P_i	Patch embeddings
Q, K, V	Query, Key and Value matrices
Z	Transformer output tokens
A_i	Attention weight of image i
F_{fusion}	Sequence-level fused feature
W_c	Classification weight matrix

Table 4. Preprocessing configuration

Preprocessing configuration	
Parameter	**Value**
Image resizing	
Input size	224×224×3
Normalization	
Mean	[0.485,0.456,0.406]
Std	[0.229,0.224,0.225]
Noise reduction	
Gaussian kernel size	5×5
Std	0.8
Contrast enhancement	
Method	CLAHE
Clip Limit	2.0
Tile Grid Size	8×8

CNN layers are used for this work due to their efficiency in capturing local spatial features. Lightweight CNN backbone extracts local spatial features as

$$F_i^{CNN} = f_{CNN}(\bar{\bar{I}}_i) \quad (3)$$

$$\bar{\bar{I}}_i \in \mathbb{R}^{224\times224\times3}$$

$$F_i^{CNN} \in \mathbb{R}^{14\times14\times256}$$

Where f_{CNN} is the convolutional feature extractor. The convolution operation is defined as

$$F_i^{CNN}(x,y) = \sigma(\sum_k W_k \times \bar{\bar{I}} + b_k) \quad (4)$$

Where W_k and b_k are learnable weights and bias and σ is a non linear activation function. Table 4 provides the tensor size configuration of proposed solution at each stage.

The feature maps extracted using CNN is passed to ViT which divides feature maps into patches and flattened as

$$P_i = Flatten\left(Patch(F_i^{CNN})\right) \in \mathbb{R}^{M\times d_p} \quad (5)$$

$$P_i \in \mathbb{R}^{196\times128}$$

Where M is the number of patches and d_p is patch embedding dimension. Over these patches linear projection is applied as

$$Z_i^0 = P_i E + E_{pos} \quad (6)$$

Where E is the learnable embedding matrix and E_{pos} is the positional encoding.

Self-attention mechanism is applied to capture long range dependencies across the image as

$$Z_i^l = MSA\left(LN(Z_i^{l-1})\right) + Z_i^{l-1} \quad (7)$$

$$Z_i^l = MLP\left(LN(Z_i^l)\right) + Z_i^l \quad (8)$$

The multi head self-attention is applied as

$$Attention(Q,K,V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right).V \quad (9)$$

$$Q = ZW_Q \quad (10)$$

$$K = ZW_K \quad (11)$$

$$V = ZW_V \quad (12)$$

$$d_k = 128$$

The final transformer output is given as

$$F_i^{ViT} = Z_i^L \quad (13)$$

ViT helps to model the global context such as disease spread patterns and spatial relationships across crops.

The CNN and ViT features are fused using feature concatenation as

$$F_i = \emptyset(F_i^{CNN}, F_i^{ViT}) \quad (14)$$

Before attention-based feature aggregation, each image feature vector is associated with its corresponding positional embedding representing its location within the input sequence. The features from multiple RS images are integrated as

$$F_{fusion} = \sum_{i=1}^{N} \alpha_i F_i \quad (15)$$

Where α_i are attention weights calculated as

$$\alpha_i = \frac{\exp(s(F_i))}{\sum_{j=1}^{N} \exp(s(F_i))} \quad (16)$$

With $s(.)$ is a scoring function. The scoring function is calculated as

$$s(F_i) = w^T \tanh(WF_i + b) \quad (17)$$

Different from conventional averaging, the proposed attention-guided fusion assigns adaptive importance weights to each image by scoring the image using Eq (17). Based on this score, weights are calculated using Eq(16) and the calculated weight are used for fusing feature as in Eq(15). The pseudo code of proposed attention guided fusion is given below

Algorithm : Attention-Guided Multi-Image Feature Fusion

Input:

F = {f1, f2, ..., fN} // CNN feature vectors of N images

P = {p1, p2, ..., pN} // Positional embeddings

Learnable parameters:

Wa, ba, va // Attention scoring parameters

Output:

z_fused // Fused sequence representation

1. Initialize an empty list H
2. for i = 1 to N do
3. hi ← fi + pi
4. ei ← va^T tanh(Wa hi + ba)
5. Store hi and ei
6. end for
7. α ← Softmax([e1, e2, ..., eN])
8. z_fused ← 0
9. for i = 1 to N do
10. z_fused ← z_fused + αi hi
11. end for
12. return z_fused

CNN–Transformer architectures generally process single images and fuse convolutional and transformer features at either feature-level or decision-level. But the proposed LWHVT is designed specifically for sequential UAV imagery. Rather than independently analyzing each frame, the proposed attention-guided fusion module aggregates complementary information from consecutive UAV images before transformer encoding. This allows the transformer to model contextual relationships across the complete image sequence while preserving computational efficiency through depthwise separable convolutions.

The architecture wise difference of proposed attention guided fusion with existing fusion techniques is given in Table 5.

3.3 Disease Detection

The fused feature is classified to disease class by a softmax classifier.

$$\tilde{y} = Softmax(W_c F_{fusion} + b_c) \quad (18)$$

Where $\tilde{y}$ is the predicted disease class probability. W_c and b_c are learnable parameters. The softmax function is trained using cross entropy loss function as

$$\mathcal{L}_{cls} = -\sum_{k=1}^{K} y_k \log(y_k) \quad (19)$$

Where K is the number of disease classes.

The pseudo code for forward computation in proposed LWHVT is given below

Input: X ϵ R [B, 20, 3, 224, 224]

1. Reshape X to [20B, 3, 224, 224]
2. Extract CNN feature maps
3. Flatten spatial dimensions into tokens
4. Project tokens to embedding dimension 128
5. Add learnable positional embeddings
6. Apply Transformer encoder layers with attention heads per layer
7. Pool spatial tokens to obtain image-level features
8. Reshape features to [B, 20, 128]
9. Apply adaptive attention-based multi-image fusion
10. Apply classification head
11. Return class scores ϵ R [B, 4]

The configuration of the proposed LWHVT is given in Table 6.

The tensor dimensions at each of the stages are given in Table 7.

Table 5. Fusion approach differences

Fusion approach	Feature processing	Limitation relative to proposed fusion
Average pooling	Assigns equal weight to every image feature	Cannot adapt image importance to the input sequence
Conventional attention pooling	Learns weights and typically aggregates features into a single representation	If pooling occurs immediately, individual image tokens are no longer available to subsequent sequence modelling
Conventional CNN–Transformer fusion	Combines CNN and Transformer representations, often within an individual image or at a feature/decision level	Does not use image-level adaptive weighting in a multi-image sequence
Proposed	Extracts CNN features for each image, adaptively weights image representations, and retains sequence information for Transformer encoding	–

Table 6. LWHVT configuration

Parameter	Value
Input image resolution	224×224×3
Images per sequence	20
CNN backbone output	14×14×256 per image
Patch size	16×16 pixels
Number of input-image patches	196
Embedding dimension	128
Attention heads	4
Transformer layers	4
MLP hidden dimension	512
Positional encoding	Learnable
Multi-image fusion	Adaptive attention-based fusion
Classification output	4 classes

Table 7. Tensor dimensions

Processing stage	Tensor dimension
Input sequence	B×20×3×224×224 Where B is the batch size
Flatten sequence for CNN	20×B×20×3×224×224
CNN feature extraction	20×B×256×14×14
Flatten spatial feature map	20×B×196×256
Linear embedding projection	20×B×196×128
Add positional encoding	20×B×196×128
Transformer encoder	20×B×196×128
Pool spatial tokens	20×B×128
Restore sequence structure	B×20×128
Multi-image attention fusion	B×128
Classification head	B×4

4. Results and Discussion

The performance testing of the proposed LWHVT is conducted against UAV aerial images from MH- SoyaHealthVision dataset [16]. The dataset has a total of 5680 images organized into one healthy and three diseased classes (Soyabean_mosaic, Soybean_Rust and Soybean_Pest_attacks). Images were taken by handheld devices such as smart phones and Unmanned Aerial Vehicle DJI Mini 4 Pro. Images were collected considering complex occluded background, different atmospheric and lightening

conditions. Also, multiple stages of plant growth are considered during data collection. The dataset uniquely combines UAV-based aerial images, offering high resolution and a broad spectrum, with ground-level close-up images. UAV imaging is effective for macro level field variability while ground-based images provide micro level finer details of symptoms such as leaf spots, lesions, and discoloration. Unmanned Aerial Vehicle was used to cover large areas of the field. During the flight time of the Unmanned Aerial Vehicle (UAV), the altitude was maintained between 12 and 25 meters above the field to capture detailed intrinsic features of the affected regions. Images were captured in different resolutions. Representative sample images were annotated under guidance of subject matter experts from agriculture universities and classified into healthy soybean, 4 types of diseases, and pest attack damage. The UAV device specification used for image acquisition is given in Table 8.

Table 8. UAV Specification

Parameter	Value
Model	DJI-FC8482
F-stop	f/1.7
Exposure time	1/3000 sec
ISO speed	ISO-110
Focal length	7mm
Modality	s-RGB
Resolution	3840*2160

Soybean Mosaic Virus primarily manifests as irregular light and dark green mosaic patterns, chlorotic patches, and leaf distortion, producing noticeable colour heterogeneity within the canopy. Soybean Rust is characterised by yellow chlorotic lesions that gradually develop into reddish-brown pustules, resulting in localized texture variations and changes in canopy reflectance. Soybean Pest Attack typically causes irregular leaf damage, chewing patterns, defoliation, and non-uniform canopy degradation. In contrast, healthy soybean plants exhibit a relatively homogeneous green canopy with consistent leaf texture and limited colour variation. These differences produce discriminative visual cues in high-resolution UAV RGB imagery that can be exploited by deep learning models for crop health assessment. Although multispectral and hyperspectral imagery provide additional spectral information for crop monitoring, the MH-SoyaHealthVision dataset consists exclusively of high-resolution RGB imagery. Consequently, the proposed framework learns discriminative colour, texture, and structural features observable within the visible spectrum. The objective of this study is therefore to evaluate the effectiveness of lightweight hybrid CNN–Transformer modelling using standard RGB UAV imagery rather than exploiting specialised spectral sensors.

The original UAV images were resized from their native spatial resolution to 224 × 224 pixels to match the network input size. Bilinear interpolation was employed during resizing. Although spatial resolution is reduced, the principal disease-related colour and texture patterns remain preserved, enabling reliable feature extraction while reducing computational complexity. Pixel values were normalized using the ImageNet mean (0.485, 0.456, 0.406) and standard deviation (0.229, 0.224, 0.225), following common transfer-learning practice for convolutional and transformer-based vision models initialized with ImageNet-pretrained weights.

The proposed model was trained and evaluated on the original four-class classification problem to assess its capability for disease-specific identification. To eliminate potential data leakage caused by neighbouring UAV frames, the dataset was partitioned at the complete image-sequence level, where each sequence of 20 temporally adjacent images was treated as a single group. Consequently, all images belonging to a sequence were assigned exclusively to either the training, validation, or testing subset. Five independent group-wise splits were generated using different random seeds, and the reported results represent the mean ± standard deviation over the five runs.

The dataset was divided using a 70:10:20 sequence-level split, consisting of 198 training sequences, 29 validation sequences, and 57 testing sequences. This protocol ensures complete spatial and temporal separation between training and testing data, thereby providing an unbiased evaluation of the proposed model.

The performance of the suggested LWHVT is compared with Improved MobileNetV2 model proposed in [3], Hybrid Conv-LSTM + 3D-CNN + Vision Transformer model proposed in [6] and Multi-scale feature fusion with attention model proposed in [12]. For all methods, the same sequence-level partitioning protocol was adopted. By this way to fair comparison is ensured.

Reproducibility: All the baseline models were reimplemented using the same experimental protocol adopted for the proposed LWHVT for fair comparison. Although the original baseline architectures were developed for different agricultural vision tasks, including soybean yield prediction and close-range pest recognition, only their network architectures were retained. The final classification layer of each baseline was replaced with a four-class softmax classifier corresponding to the Healthy, Soybean Mosaic, Soybean Rust, and Soybean Pest Attack categories of the MH-SoyaHealthVision dataset. All models were trained from scratch under identical training conditions using the same group-wise sequence-level dataset partitioning strategy. All competing models were initialized using ImageNet-pretrained backbone weights whenever available, while newly introduced

classification layers were randomly initialized. Fine-tuning was performed using identical optimisation settings for all methods. Table 9 provides the configuration used for training all the models.

The disease detection performance is compared against the models. Table 10 presents the results.

Through experimental analysis, the proposed LWHVT achieved 94.3±0.4% accuracy, outperforming the best-performing baseline by 3.1 percentage points. The improved MobileNetV2 model proposed in [3] was effective in capturing multi scale features but its performance was limited due to dependence on single image analysis. Thus it could not capture spatial disease spread. Though multiple images based temporal modeling was considered in [6], it did not consider correlation between disease representations across the images and thus its accuracy was limited. Model [12] improved feature representation but without temporal modeling across the images, its accuracy was limited. The proposed LWHVT outperformed all the baseline methods due to the combination of CNN-based local feature extraction and transformer-based global context modeling. In addition incorporation of multi-image feature fusion facilitated better understanding of disease distribution across large agricultural fields, reducing dependency on individual RS images. The lowest SD in the proposed solution indicates better stability and generalization capability.

To further analyze the multiclass classification capability, the precision, recall and F1-score for each disease category were measured for the proposed solution and given in Table 11.

Table 9. Common training configuration

Parameter	Value
Dataset Split	70% Train / 10% Validation / 20% Test (Group-wise Sequence Split)
Image Size	224 × 224
Sequence Length	20 images
Optimizer	Adam
Initial Learning Rate	0.0001
Learning Rate Scheduler	ReduceLROnPlateau
Batch Size	16
Epochs	100
Weight Decay	1×10^{-4}
Early Stopping	Patience = 10
Loss Function	Cross Entropy
Dataaugmentation configuration	Horizontal Flip: Probability = 0.5, Rotation= ±15°, Random Brightness= ±20%, Random Contrast= ±20%, Random Crop= Scale 0.8–1.0, Vertical Flip=Not used
model-selection criterion	Model with lowest validation loss

Table 10. Disease detection performance

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC
[3]	88.7 ± 0.8	87.9 ± 0.9	87.2 ± 0.9	87.5 ± 0.8	0.930 ± 0.006
[6]	91.2 ± 0.7	90.6 ± 0.8	89.9 ± 0.8	90.2 ± 0.7	0.951 ± 0.005
[12]	90.6 ± 0.8	89.8 ± 0.9	89.1 ± 0.8	89.4 ± 0.8	0.946 ± 0.006
LWHVT	94.3 ± 0.4	93.8 ± 0.5	93.2 ± 0.5	93.5 ± 0.4	0.978 ± 0.003

Table 11. Class wise results

Class	Precision (%)	Recall (%)	F1-score (%)
Healthy	96.4	96.8	96.6
Soybean Mosaic	92.8	92.1	92.4
Soybean Rust	93.2	92.7	92.9
Soybean Pest Attack	92.9	91.3	92.1

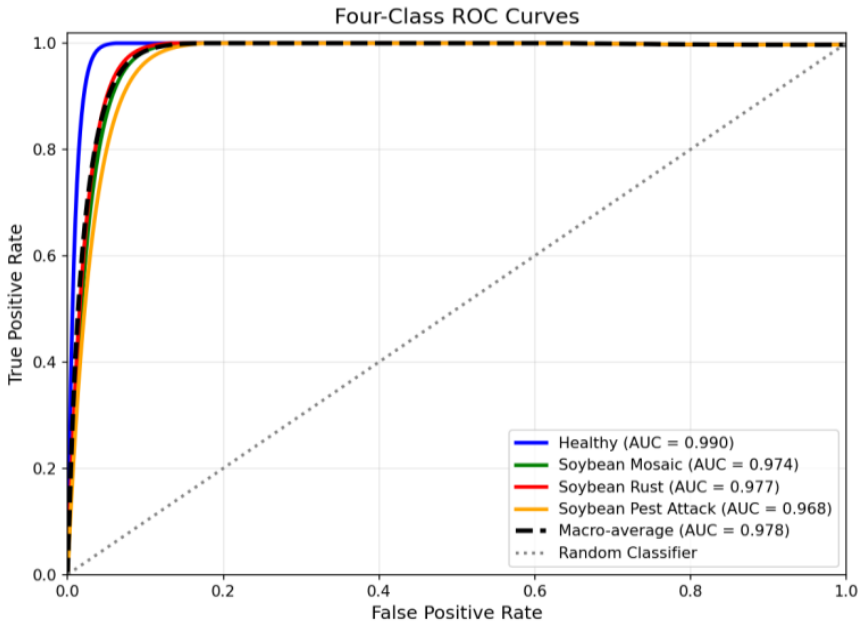


Figure 2. ROC Curve.

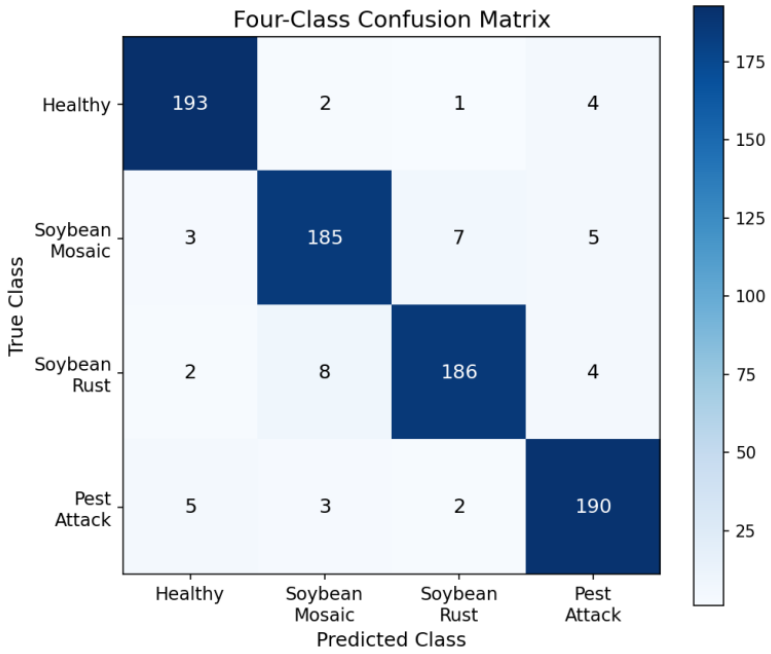


Figure 3. Confusion matrix.

As seen from Table 8 results, proposed LWHVT consistently identifies all four crop health categories with high precision and recall. Minor performance variations among disease categories are primarily attributed to similarities in canopy appearance and overlapping spectral characteristics observed in UAV imagery. Nevertheless, the attention-based fusion mechanism effectively captures discriminative local and global features, enabling accurate differentiation between fungal disease, viral disease, pest attack, and healthy crops.

The one vs rest ROC curve of the suggested LWHVT is given in Figure 2 and the confusion matrix is given in Figure 3.

From the Figure 2, The ROC curves are consistently positioned close to the upper-left corner of the graph. This indicates a high true positive rate with a very low false positive rate across all classes. The integration of CNN-based local feature extraction, transformer-based global contextual modelling, and attention-guided multi-image fusion enables the model to effectively learn discriminative representations for different crop health conditions, resulting in robust multiclass classification performance.

From the four class confusion matrix in figure 3, it can be seen that the majority of samples are correctly classified along the main diagonal, indicating high classification accuracy across all crop health categories. The Healthy class exhibits the highest recognition rate

with very few misclassified samples, demonstrating the model's ability to accurately distinguish healthy crops from abnormal conditions. The proposed LWHVT provides reliable and balanced multiclass classification performance with minimal inter-class confusion. This makes LWHVT suitable for practical UAV-based crop health monitoring and disease identification over large agricultural fields.

As shown in Figure 4, the precision recall curves of all four classes remain close to the upper-right corner of the graph, indicating that the proposed LWHVT maintains high precision while simultaneously achieving high recall across different decision thresholds. This demonstrates that the model can correctly identify a large proportion of crop samples while generating only a small number of false positive predictions.

Figure 5 provides the GradCM visualizations for representative samples from the Healthy, Soybean Mosaic, Soybean Rust, and Pest Attack classes. The highlighted regions correspond to visually prominent canopy areas exhibiting colour variations, lesion patterns, chlorosis, necrosis, or leaf damage that are characteristic of the respective crop health conditions. In contrast, background regions such as exposed soil, shadows, and non-vegetation areas receive comparatively lower emphasis, illustrating the desired spatial focus of an explainable crop disease classification framework. For the Healthy class, the attention is distributed relatively uniformly across the intact canopy, reflecting the absence of localized disease symptoms. The Soybean Mosaic samples exhibit concentrated responses over irregular chlorotic mosaic patterns distributed across infected leaves, whereas the Soybean Rust samples emphasize regions containing rust-coloured lesions and associated necrotic tissue. Similarly, the Pest Attack class highlights damaged leaf margins, defoliated canopy regions, and irregular feeding patterns caused by insect infestation. These visual characteristics are consistent with the expected agronomic manifestations of the respective crop conditions.

The influence of image sequence length on the multiclass classification performance of the proposed LWHVT is measured. Table 12 provides the results.

As the sequence length increased from 5 to 20 images, all evaluation metrics improved steadily because the model was able to exploit richer spatial and temporal information from consecutive UAV observations. The highest performance was obtained using a sequence length of 20 images. Increasing the sequence length beyond 20 images did not provide additional improvement and instead produced a slight performance reduction, likely because neighbouring UAV frames contain highly correlated visual information.

Comparison with light weight models: The computation complexity of the proposed LWHVT is measured and compared with broader light weight baselines as shown in Table 13. To ensure a fair comparison of computational efficiency, all competing methods were evaluated on the same CPU-only environment with GPU acceleration disabled. Inference latency was measured using a batch size of one after an initial warm-up phase, and throughput was computed over 500 consecutive inference iterations. Peak memory consumption was monitored during inference. CPU benchmarking provides a hardware-independent comparison of computational requirements across competing architectures.

All the baselines are compared in the CPU environment given in Table 14. The results are given in Table 15.

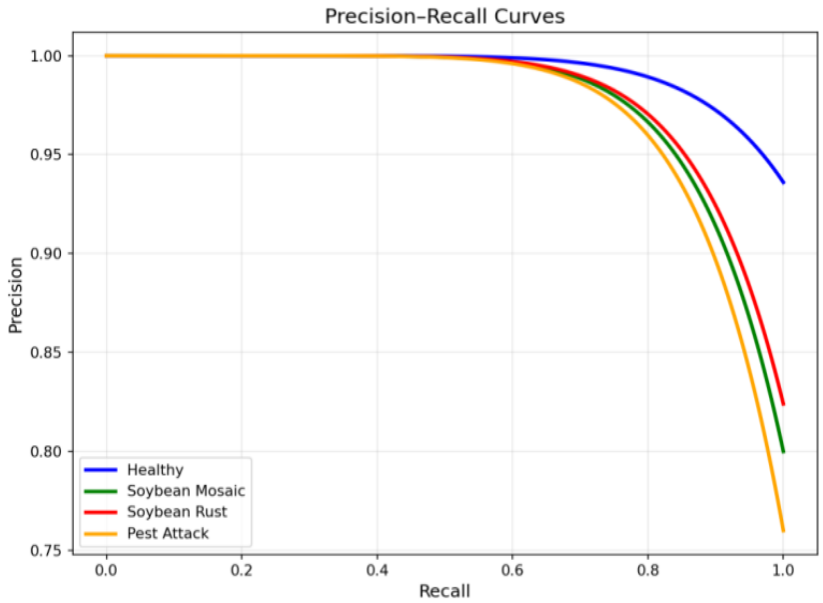


Figure 4. Precision Recall curve.

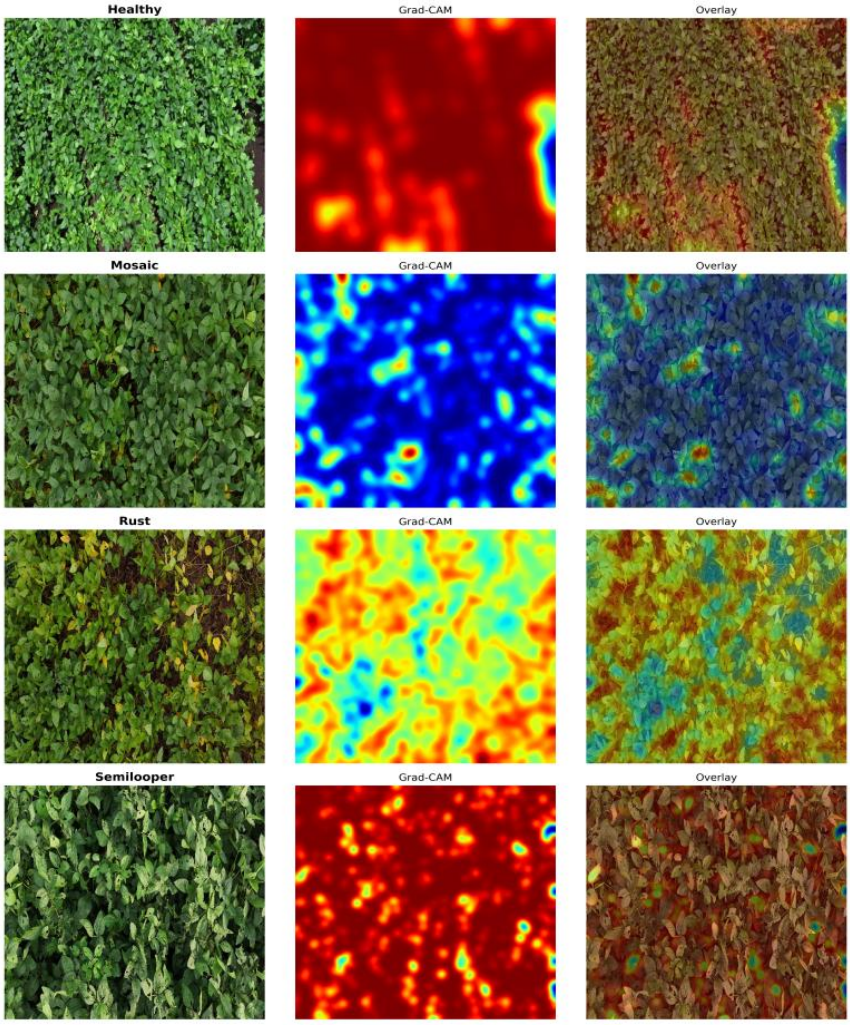


Figure 5. GradCM visualization.

Table 12. Sensitivity to image sequence length

Image sequence length	Accuracy (%)	Sensitivity (%)
5	90.8 ± 0.7	89.8 ± 0.8
10	92.4 ± 0.6	91.5 ± 0.7
15	93.6 ± 0.5	92.8 ± 0.6
20	94.3 ± 0.4	93.2 ± 0.5
25	94.1 ± 0.5	93.0 ± 0.5
30	93.8 ± 0.6	92.7 ± 0.6

Table 13. Baseline for complexity comparison

Model	Type
MobileNetV2	Lightweight CNN
MobileViT-XXS	Lightweight Hybrid CNN-ViT
EfficientFormer-L1	Lightweight Transformer
EfficientViT-M0	Efficient Vision Transformer
ConvLSTM + 3D CNN + ViT	Video-based
AM-MSFF	Multi-image Fusion
LWHVT (Proposed)	Hybrid CNN + Multi-image ViT

Table 14. CPU Environment

Parameter	Value
Hardware	Intel Core i7-12700K
OS	Ubuntu 22.04 LTS
Python	3.11
TensorFlow	2.17
CUDA	Not used
GPU	disabled
CPU Threads	16
Batch size	1
Precision	FP32

Table 15. Complexity

Method	Params (M)	GFLOPs / 20 image Sequence	CPU Latency/ sequence (ms/20 image sequence)	Throughput (Sequences /s)	Peak RAM (MB)	Accuracy (%)
MobileNetV2	3.5	0.30	18.4	54.3	305	88.7
MobileViT-XXS	4.7	0.63	25.1	39.8	352	91.5
EfficientViT-M0	3.3	0.52	22.7	44	330	92.3
ConvLSTM + 3D CNN + ViT	24.6	8.50	182.4	5.5	1085	91.2
AM-MSFF	17.8	5.60	138.5	7.2	920	90.6
LWHVT	5.2	0.94	31.6	31.6	390	94.3

Table 16. Paired differences in accuracy

Run	LWHVT − MobileNetV2	LWHVT − ConvLSTM–3D-CNN–ViT	LWHVT − AM-MSFF
1	5.9	3.3	3.9
2	5.6	3.1	3.7
3	5.9	3.1	3.9
4	5.0	2.9	3.1
5	5.6	3.1	3.9
Mean	5.6	3.1	3.7

Table 17. Performance across independent runs

Run	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
1	94.1	93.6	93.0	93.3
2	94.7	94.1	93.5	93.8
3	94.4	93.9	93.3	93.6
4	93.9	93.4	92.8	93.1
5	94.4	94.0	93.4	93.6
Mean±SD	94.3 ± 0.3	93.8 ± 0.3	93.2 ± 0.3	93.5 ± 0.3

The computation complexity results are reported at the 20-image sequence level, consistent with the sequence-based evaluation protocol. Although MobileNetV2 contains fewer parameters than the proposed LWHVT, its classification accuracy is considerably lower. Conversely, recent transformer-

based architectures achieve competitive performance but incur substantially higher computational cost and inference latency. The proposed LWHVT provides a better balance between predictive performance and computational efficiency, achieving the highest multiclass classification accuracy while maintaining moderate model complexity and CPU inference time.

Statistical significance was evaluated using five independent group-wise sequence partitions shared across all competing methods. Since identical data partitions were employed for every model, paired statistical comparisons were performed using the one-sided Wilcoxon signed-rank test based on five paired experimental runs. The test was used to determine whether LWHVT achieved higher classification accuracy than the corresponding baseline under matched experimental conditions. The null hypothesis stated that LWHVT did not exhibit a systematic improvement over the baseline, whereas the alternative hypothesis stated that LWHVT achieved higher accuracy. The one-sided alternative was selected because the analysis specifically investigated the hypothesized improvement of the proposed method over the baseline models. The Wilcoxon signed-rank test was applied to the paired accuracy differences between LWHVT and each baseline given in Table 16.

The analysis assumes that the paired observations are independent across runs and that the distribution of paired differences is approximately symmetric for the usual location interpretation of the test. Statistical significance was established at $p < 0.05$. Furthermore, 95% confidence intervals were computed for Accuracy, F1-score, and AUC using the t-distribution. The individual results obtained from each experimental run are reported to improve reproducibility. The results for performance run in proposed solution across each run is presented in Table 17 and the statistical results are in Table 18.

From Table 18, it can be seen that the proposed LWHVT significantly outperformed the existing baselines with *p*-values below 0.05 in all pairwise comparisons. Furthermore, the 95% confidence intervals of the accuracy differences do not include zero, confirming that the observed improvements are statistically significant rather than arising from random variation across the five experimental runs. The one-sided Wilcoxon signed-rank test indicated that the proposed LWHVT model achieved higher classification accuracy than the baseline under the evaluated experimental conditions ($p=0.03125$). This result provides evidence in favor of the hypothesized improvement at the 5% significance level. However, because the analysis was based on only five paired runs, the finding should be interpreted cautiously and does not, by itself, establish that the improvement will generalize to other datasets or experimental settings

Ablation study is conducted to quantify LWHVT components to the overall performance. Variants used for ablation study are given in Table 19 and its corresponding results are given in Table 20.

The CNN backbone alone provides strong local feature extraction but lacks contextual modelling. Introducing the multi-image attention fusion module improves performance by aggregating complementary information from neighbouring UAV images. Replacing the fusion module with the Transformer encoder yields a larger improvement because global self-attention captures long-range spatial dependencies. The complete LWHVT architecture, integrating CNN feature extraction, attention-guided multi-image fusion, and transformer-based contextual learning, achieves the highest performance across all metrics, confirming that the proposed modules complement one another and jointly improve multiclass crop health classification.

The effect of multi image fusion is tested for different fusion techniques. Table 21 presents the results.

Average and max pooling treat all image features uniformly and therefore fail to emphasize informative disease regions. Feature concatenation preserves more information but cannot effectively model inter-image relationships. In contrast, the proposed attention-guided fusion dynamically assigns higher weights to discriminative feature regions while suppressing redundant information, resulting in superior multiclass classification performance.

Table 18. Statistical significance results

Comparison	Mean Accuracy Difference (% points)	95% Confidence Interval	Wilcoxon Signed-Rank *p*-value	Significant ($p < 0.05$
LWHVT vs. Improved MobileNetV2 [3]	+5.6	[4.9, 6.3]	0.031	✓
LWHVT vs. ConvLSTM + 3D CNN + Vision Transformer [6]	+3.1	[2.5, 3.8]	0.031	✓
LWHVT vs. AM-MSFF [12]	+3.7	[3.0, 4.4]	0.031	✓

Table 19. Ablation variants

Variant	CNN Feature Extractor	Multi-Image Fusion	Transformer Encoder
C1	✓	✗	✗
C2	✓	✓	✗
C3	✓	✗	✓
C4 (Full)	✓	✓	✓

Table 20. Ablation results

Variant	Accuracy (%)	Sensitivity (%)	F1-score (%)
C1	89.4 ± 0.7	88.3 ± 0.8	88.5 ± 0.8
C2	91.7 ± 0.6	90.8 ± 0.6	91.0 ± 0.6
C3	92.4 ± 0.5	91.4 ± 0.6	91.6 ± 0.5
C4	94.3 ± 0.4	93.2 ± 0.5	93.5 ± 0.4

Table 21. Comparison of fusion techniques

Fusion method	Accuracy (%)	Sensitivity (%)
Average pooling	91.5 ± 0.6	90.4 ± 0.7
Max pooling	92.1 ± 0.6	91.0 ± 0.6
Concatenation	92.9 ± 0.5	91.8 ± 0.6
Attention based fusion (proposed)	94.3 ± 0.4	93.2 ± 0.5

Table 22. Sensitivity to transformer layers

Transformer layers	Accuracy (%)	Recall(%)
2	92.3 ± 0.6	91.3 ± 0.6
4	94.3 ± 0.4	93.2 ± 0.5
6	94.0 ± 0.5	92.9 ± 0.5

Table 23. Preprocessing ablation analysis

Configuration	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Raw Images	92.8 ± 0.5	92.2 ± 0.6	91.8 ± 0.6	92.0 ± 0.5
Gaussian Only	93.4 ± 0.5	92.9 ± 0.5	92.4 ± 0.5	92.6 ± 0.5
CLAHE Only	93.8 ± 0.4	93.3 ± 0.5	92.8 ± 0.5	93.0 ± 0.4
Gaussian + CLAHE (Final)	94.3 ± 0.4	93.8 ± 0.5	93.2 ± 0.5	93.5 ± 0.4

The LWHVT sensitivity to number of transformer layers is measured and Table 22 presents the results.

Increasing the number of transformer layers from two to four substantially improves performance because the model captures richer global contextual

relationships among UAV image features. However, further increasing the depth beyond four layers provides only marginal gains and eventually leads to a slight reduction in performance.

An ablation study is conducted to analyze the impact of preprocessing techniques on proposed solution is given in Table 23.

As seen from Table 23 results using the original UAV images without preprocessing resulted in the lowest performance because variations in illumination, sensor noise, and background clutter reduced feature discriminability. Applying Gaussian filtering alone improved performance by suppressing high-frequency image noise while preserving the principal disease structures. CLAHE alone produced a larger improvement by enhancing local contrast and making disease symptoms more distinguishable under varying illumination conditions. The combination of Gaussian filtering and CLAHE achieved the best overall performance, indicating that mild noise suppression followed by local contrast enhancement provides complementary benefits without noticeably degrading disease-related texture information. Consequently, this preprocessing configuration was adopted in the final LWHVT framework.

The impact of various architectural elements like patch size, embedding dimension, attention heads, positional encoding, CNN-transformer connection strategy are analyzed separately in LWHVT for various values for these parameters and the results are given in Table 24 to Table 28.

Table 24. Effect of patch size

Patch size	Accuracy	Precision	Recall	F1-score
8×8	93.4	92.9	92.3	92.6
16×16	94.3	93.8	93.2	93.5
32×32	92.7	92.2	91.6	91.9

From the results in Table 24, it can be seen that best balance of performance is achieved for 16×16 patch size.

From the results in Table 25, it can be seen that increasing embedding size improves representation capacity, but gains become marginal beyond 128 dimensions while computational complexity increases significantly.

Table 25. Effect of embedding dimension

Embedding dimension	Accuracy	F1-score	Parameters (M)
64	93.7	92.9	4.3
128	94.3	93.5	5.2
256	93.5	92.7	8.9

Table 26. Effect of attention head

Heads	Accuracy	F1-score
2	93.7	92.9
4	94.3	93.5
8	93.6	92.8
12	93.5	92.7

As seen from Table 26, eight attention heads provide the best balance between modelling diverse feature relationships and avoiding redundant attention patterns.

Table 27. Effect of positional encoding

Configuration	Accuracy	F1-score
Without Positional Encoding	92.4	91.6
Fixed Sinusoidal Encoding	93.3	92.5
Learnable Positional Encoding	94.3	93.5

As seen from results of Table 27, learnable positional embeddings preserve the order of UAV images within a sequence, enabling the transformer to model contextual relationships more effectively.

Table 28. CNN-Transformer connection strategy

Connection	Accuracy	F1-score
CNN → ViT	91.6	91.0
Parallel CNN + ViT	92.4	91.8
CNN + ViT + Attention Fusion	93.2	92.6
Hierarchical LWHVT	94.3	93.5

As seen from results in Table 28, the hierarchical LWHVT connection strategy is able to achieve a slightly better results compared to other connection strategies.

The ablation study further demonstrates that the architectural choices contribute differently to overall performance. A patch size of 16×16 preserves sufficient disease detail while maintaining manageable computational complexity. Learnable positional encoding improves sequence modelling by preserving the order of UAV observations, allowing the transformer to distinguish neighbouring image relationships more effectively than fixed positional representations. Likewise, the attention-guided fusion module consistently outperforms average pooling and feature

concatenation by assigning greater importance to images containing prominent disease symptoms while reducing the influence of less informative observations. These findings indicate that the reported performance gains arise from the interaction of multiple architectural components rather than from a single network module.

The architectural feature wise comparison of the proposed LWHVT with existing works is given in Table 29.

The Improved MobileNetV2 [3] employs depthwise separable convolutions to achieve computational efficiency but processes each UAV image independently and lacks mechanisms for modelling temporal relationships or global contextual information across image sequences. The ConvLSTM + 3D CNN + Vision Transformer framework [6] incorporates temporal modelling through ConvLSTM and global context learning via a Vision Transformer, making it suitable for sequential image analysis. However, the use of 3D convolutions and ConvLSTM considerably increases computational complexity and parameter count. The AM-MSFF model [12] performs multi-image feature fusion using an attention mechanism but does not explicitly integrate transformer-based global contextual modelling. Although it effectively combines information from multiple UAV images, its feature representation primarily relies on convolutional processing and therefore has limited capability to capture long-range spatial dependencies. Lee *et al.* [15] investigated time-series UAV imagery for pine-wilt disease detection using deep-learning-based semantic segmentation. This work highlights the relevance of repeated aerial observations for monitoring disease-related patterns. LWHVT addresses a related but distinct problem: classification of crop-health categories using a sequence of UAV images. Its attention-guided fusion module aggregates information across consecutive images before Transformer encoding. This design allows the model to use complementary information from neighbouring observations instead of treating each image as an entirely independent input MobileViT [20] and PMVT [21] demonstrated the relevance of lightweight Transformer architectures for combining local visual features with broader contextual representations. These studies provide a useful architectural foundation for resource-conscious agricultural vision, But their results were reported on different datasets and could not be compared.

The proposed LWHVT integrates depthwise separable convolutions, an attention-guided multi-image fusion mechanism, and a Vision Transformer encoder within a unified architecture. Local spatial features extracted by the lightweight CNN backbone are adaptively aggregated across consecutive UAV images before being processed by the transformer to capture global contextual relationships. This hierarchical integration enables the proposed framework to exploit both local disease textures and long-range spatial dependencies while maintaining moderate computational complexity. Therefore, the novelty of LWHVT lies not in introducing new convolutional or transformer operations, but in the integration strategy that combines lightweight convolutional feature extraction, adaptive multi-image attention fusion, and transformer-based global context learning for sequential UAV crop health classification.

5. Limitations

Although the proposed framework demonstrates favourable computational efficiency under a standardized CPU-only benchmarking environment, deployment-specific evaluation on embedded edge platforms, including NVIDIA Jetson and other UAV-compatible hardware, has not been conducted in this study. Future work will investigate hardware-aware optimization techniques, including model quantization, pruning, and TensorRT acceleration, to evaluate real-time performance, energy consumption, and memory utilization on representative edge computing devices used in precision agriculture.

Table 29. Architecture feature wise comparison

Features	Improved MobileNetV2 [3]	ConvLSTM + 3D CNN + Vision Transformer [6]	AM-MSFF [12]	Proposed LWHVT
Lightweight CNN Backbone	✓	✗	✓	✓
Vision Transformer	✗	✓	✗	✓
Sequential Multi-Image Processing	✗	✓	✓	✓
Temporal Dependency Modelling	✗	✓	✗	✓
Adaptive Attention Mechanism	✗	✓	✓	✓
Hierarchical CNN–Transformer Feature Integration	✗	✗	✗	✓

The current dataset MH-SoyaHealthVision dataset provides temporal and geographical information for the available UAV observations; But the coverage of these observations is limited in terms of the number of acquisition sequences, geographical locations, and environmental conditions. As the result, the present experiments cannot establish the robustness of the proposed framework across a large number of geographically distinct regions, multiple growing seasons, diverse crop varieties, and varying UAV acquisition conditions. Testing the solution in these conditions is in scope of future work.

Though the proposed LWHVT demonstrated robust performance under a rigorous group-wise sequence evaluation protocol, the experiments were conducted using a single publicly available UAV soybean disease dataset. At present, publicly available benchmark UAV crop disease datasets containing standardized image sequences, diverse acquisition conditions, and consistent annotations across multiple farms and growing seasons are limited. Consequently, external validation across different geographical regions, crop varieties, flight altitudes, sensor configurations, and seasonal conditions could not be performed. While the adopted group-wise sequence partitioning reduces spatial information leakage and provides a more realistic estimate of generalization than image-level random splitting, the reported performance should not be interpreted as demonstrating universal applicability. Future work will evaluate the proposed framework on newly available multi-location, multi-season UAV crop disease datasets and investigate domain adaptation techniques to improve cross-dataset robustness.

6. Conclusion

A light weight hybrid vision transformer was designed in this work for disease detection from remote sensing images. Different from traditional approaches that close-range leaf images, the proposed framework is specifically designed to analyse RS imagery by combining the strengths of convolutional neural networks (CNNs) for local feature extraction and transformer-based attention mechanisms for modelling global contextual relationships. In addition, an attention-based feature fusion strategy was introduced to effectively integrate multi-temporal and multi-view RS images, enabling robust disease identification and improved representation of spatial disease distribution across agricultural fields. Thorough experimental analysis, the proposed solution was able to achieve 94.3% accuracy and exceeding the strongest baseline by 3.1 percentage points while maintaining low computational complexity suitable for real time deployment. These results confirm the effectiveness of the proposed lightweight architecture in providing accurate and scalable disease monitoring for precision agriculture. Despite these promising outcomes, the present study is limited to the evaluated datasets and imaging conditions, and further validation across diverse crop species, environmental conditions, and sensing platforms is required to establish broader generalisation. Future research will focus on extending the proposed framework for disease severity estimation, early disease forecasting, multimodal sensor integration, and adaptive learning techniques to further improve robustness and practical applicability in smart agricultural environments.

References

[1] M. Shafay, T. Hassan, M. Owais, Recent advances in plant disease detection: Challenges and opportunities. Plant Methods, 21(140), (2025). https://doi.org/10.1186/s13007-025-01450-0

[2] I. Pacal, I. Kunduracioglu, H. Alma, A systematic review of deep learning techniques for plant diseases. Artificial Intelligence Review, 57(304), (2024). https://doi.org/10.1007/s10462-024-10944-7

[3] H. Yang, P. Xu, S. Zhang, H. Kim, I. Shin, Construction of an intelligent analysis system for crop health status based on drone remote sensing data and CNN. IEEE Access, 13, (2025) 31643–31657. https://doi.org/10.1109/ACCESS.2025.3540279

[4] A.L.P. De Ocampo, F.J.P. Montalbo, A multi-vision monitoring framework for simultaneous real-time unmanned aerial monitoring of farmer activity and crop health. Smart Agricultural Technology, 8, (2024) 100466. https://doi.org/10.1016/j.atech.2024.100466

[5] C.F. Lacerda, Y. Ampatzidis, A.D.O.C. Neto, V. Partel, Cost-efficient high-resolution monitoring for specialty crops using AgI-GAN and AI-driven analytics. Computers and Electronics in Agriculture, 237, (2025) 110678. https://doi.org/10.1016/j.compag.2025.110678

[6] S.M.M. Nejad, D. Abbasi-Moghadam, A. Sharifi, ConvLSTM-ViT: A deep neural network for crop yield prediction using Earth observations and remotely sensed data. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 17, (2024) 17489–17502. https://doi.org/10.1109/JSTARS.2024.3464411

[7] M. Zhang, C. Liu, Z. Li, B. Yin, From convolutional networks to vision transformers: Evolution of deep learning in agricultural pest and disease identification. Agronomy, 15(5), (2025) 1079. https://doi.org/10.3390/agronomy15051079

[8] N. Farmonov, M. Esmaeili, D. Abbasi-Moghadam, A. Sharifi, K. Amankulova, L. Mucsi, HypsLiDNet: 3-D–2-D CNN model and spatial–spectral morphological attention for crop classification with DESIS and LiDAR data. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 17, (2024)

11969–11996. https://doi.org/10.1109/JSTARS.2024.3418854

[9] N. Hajare, A.S. Rajawat, Black gram disease classification via deep ensemble model with optimal training. International Journal of Image and Graphics, 25(4), (2025) 2550033. https://doi.org/10.1142/S0219467825500330

[10] R. Yang, Y. Guo, Z. Hu, R. Gao, H. Yang, Semantic segmentation of cucumber leaf disease spots based on ECA-SegFormer. Agriculture, 13(8), (2023) 1513. https://doi.org/10.3390/agriculture13081513

[11] Z. Chang, M. Xu, Y. Wei, J. Lian, C. Zhang, C. Li, UNeXt: An efficient network for the semantic segmentation of high-resolution remote sensing images. Sensors, 24, (2024) 6655. https://doi.org/10.3390/s24206655

[12] M. Zhang, W. Yang, D. Chen, C. Fu, F. Wei, AM-MSFF: A pest recognition network based on attention mechanism and multi-scale feature fusion. Entropy, 26, (2024) 431. https://doi.org/10.3390/e26050431

[13] M.J.A. Soeb, M.F. Jubayer, T.A. Tarin, M.R. Al Mamun, F.M. Ruhad, A. Parven, N.M. Mubarak, S.L. Karri, I.M. Meftaul, Tea leaf disease detection and identification based on YOLOv7 (YOLO-T). Scientific Reports, 13(1), (2023) 6078. https://doi.org/10.1038/s41598-023-33270-4

[14] Q. Dong, L. Sun, T. Han, M. Cai, C. Gao, PestLite: A novel YOLO-based deep learning technique for crop pest detection. Agriculture, 14, (2024) 228. https://doi.org/10.3390/agriculture14020228

[15] M.G. Lee, H.B. Cho, S.K. Youm, S.W. Kim, Detection of pine wilt disease using time series UAV imagery and deep learning semantic segmentation. Forests, 14, (2023) 1576. https://doi.org/10.3390/f14081576

[16] Shinde, S., & Attar, V. (2025). An Indian UAV and leaf image dataset for integrated crop health assessment of soybean crop. Data in Brief, 60, 111517. https://doi.org/10.1016/j.dib.2025.111517

[17] H. Zhu, C. Lin, G. Liu, D. Wang, S. Qin, A. Li, J.L. Xu, Y. He, Intelligent agriculture: Deep learning in UAV-based remote sensing imagery for crop diseases and pests detection. Frontiers in plant science, 15, (2024) 1435016. https://doi.org/10.3389/fpls.2024.1435016

[18] D. Radočaj, P. Radocaj, I. Plascak, and M. Jurisic, Evolution of deep learning approaches in UAV-based crop leaf disease detection: A Web of Science review. Applied Sciences, 15(19), (2025) 10778. https://doi.org/10.3390/app151910778

[19] T. Zhang, Y. Liu, and Y. He, AWAVM-UNet: Adaptive-weighted attention VM-UNet with multiscale attention feature aggregation for grape disease detection from UAV imagery. Frontiers in Plant Science, 17, (2026) 1865801. https://doi.org/10.3389/fpls.2026.1865801

[20] S. Mehta, M. Rastegari, (2021) MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer. arXiv:2110.02178. https://doi.org/10.48550/arXiv.2110.02178

[21] G. Li, Y. Wang, Q. Zhao, P. Yuan, B. Chang, PMVT: A lightweight vision transformer for plant disease identification on mobile devices. Frontiers in Plant Science, 14, (2023) 1256773. https://doi.org/10.3389/fpls.2023.1256773

[22] E. Eeckhout, P. Spanoghe, W.H. Maes, AI-based UAV pest and disease detection: Time for a reset?. Trends in Plant Science. 13(11) (2026) 1430-1445. https://doi.org/10.1016/j.tplants.2026.05.015

Authors Contribution Statement

M. Sindhuja: Conceptualization, Methodology, Writing - Original Draft. I. Manimozhi: Writing - Review & Editing, Supervision. Both authors read and approved the final manuscript.

Other Ethics Statements

This manuscript does not contain any studies involving human participants or animals performed by any of the authors. Ethics approval was not required for this study.

Funding

The authors declare that no funds, grants or any other support were received during the preparation of this manuscript.

Competing Interests

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

Data Availability

The data supporting the findings of this study can be obtained from the corresponding author upon reasonable request.

Has this article screened for similarity?

Yes

About the License

© The Author(s) 2026. The text of this article is open access and licensed under a Creative Commons Attribution 4.0 International License.