

Swin-HViT: A Hybrid Transformer Approach for Accurate Early-Stage Crop Disease Diagnosis

hemalatha Gunasekaran

hemalatha.david@utas.edu.au

University of Technology and Applied Sciences, Ibri <https://orcid.org/0000-0001-5768-6072>

Wilfred Blessing N. R

University of Technology and Applied Sciences, Ibri

Naveen VijayaKumar Watson

naveen.kumar@utas.edu.au

University of Technology and Applied Sciences, Ibri

Hariharan B

SRM Institute of Science and Technology

Angelin Gladys Jesudoss

University of Technology and Applied Sciences, Ibri

Anupama C. G

SRM Institute of Science and Technology

Research Article

Keywords: Smart Agriculture, ViT, Swin, Hybrid Transformer, Crop Disease Prediction

Posted Date: January 14th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-8588269/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Swin-HViT: A Hybrid Transformer Approach for Accurate Early-Stage Crop Disease Diagnosis

Hemalatha Gunasekaran* ¹, Wilfred Blessing N. R¹, Naveen VijayaKumar Watson ¹, Hariharan B²
Angelina Gladys Jesudoss¹, Anupama C. G²

¹ College of Computing and Information Sciences, University of Technology and Applied Sciences,
Ibri 516, Oman

²Department of Computational Intelligence, School of Computing, SRM Institute of Science and Technology,
Kattankulathur, Chennai-603203, India

*Corresponding Author: hemalatha.david@utas.edu.om, naveen.kumar@utas.edu.om

Abstract

Agriculture plays a pivotal role in global economic growth, yet it faces significant challenges from pests and crop diseases. Early detection is crucial for preventing large-scale crop losses and ensuring food security. This study introduces a hybrid transformer model, Swin-HViT, which integrates the strengths of Vision Transformer (ViT) and Swin Transformer to accurately predict crop diseases. While ViT captures global image features, Swin Transformer excels at extracting fine-grained local details. Evaluated on two benchmark datasets, Corn and PlantDoc, our model achieved accuracy of 98.81% and 81.81%, respectively, surpassing recent works. Here, we demonstrate the effectiveness of combining complementary transformer architectures to enhance disease identification in diverse agricultural settings. The code, data and the hybrid model are available at <https://github.com/hema2107/Swin-HViT>.

Keywords: Smart Agriculture; ViT; Swin; Hybrid Transformer; Crop Disease Prediction

1. Introduction

The global population reached nearly 8 billion as of 2025, thereby significantly increasing the demand for food production. The global crop yield suffers on an average of 10% to 28% of loss annually due to pests and plant diseases [1,2]. Therefore, early identification of crop diseases is crucial to prevent large-scale crop loss and to maintain a stable and sustainable food supply for the growing population. Continuous monitoring of crops and timely predictions of diseases are essential to prevent crop damage [3]. Physically inspecting the crops for monitoring crop health is time consuming and tedious process. Therefore, there is a growing demand for an accurate and timely disease prediction system. There has been a notable rise in the adoption of AI in agriculture over recent years, especially for crop health monitoring and management.

Convolution Neural Network (CNN) was the widely used deep learning technique for crop disease prediction. In CNN, convolutional layers are used to extract low level features such as edges, corners, textures and color gradients from the images. More convolution layers are stacked to learn the complex features in the deeper layers. Several pre-trained models such as AlexNet, VGGNet, InceptionV3, ResNet, DenseNet, and MobileNet have been used by many researchers to identify and classify crop images [4]. These models can capture local spatial features and edges effectively through hierarchical convolutional operations. Despite their success, CNN-based architectures possess inherent limitations, they cannot capture long-range dependencies and global contextual relationships between spatially distant regions of an image. In addition, CNNs has limitation such as computational inefficiency, vanishing gradients, and overfitting issues when applied to real world dataset.

Transformer models, which are mainly used in Natural Language Processing (NLP), have recently been adapted for a variety of tasks in computer vision and other domains [5]. The Vision Transformer (ViT) model uses self-attention mechanism to capture long range global contextual dependencies in the image without using the convolutional layers [6, 7]. This global perspective makes the transformer models perform better than the deep learning models thus improving their performance in complex visual recognition tasks. Swin Transformer, on the other hand, introduces hierarchical representation and local window attention that shifts between layers [8]. This design allows the model to capture both the local and global dependencies efficiently while reducing computational cost. There are other variants of transformer models such as DeiT (Data-efficient Image Transformer) [9], and PVT (Pyramid Vision Transformer) have further improved upon ViT by introducing

hierarchical and efficient attention mechanisms. In the agricultural domain, these transformer-based models have shown great performance improvement in handling complex backgrounds, varying illumination, and high intra-class similarity among diseased leaves. However, their performance is often limited by quantity and the quality of the dataset.

Many hybrid models [10, 11, 12, 13] are available for plant disease classification and it mostly uses CNN as the base model for feature extraction and transformer models for classification for example, ViT-ResNet, Dense-ViT, and ConvNeXt-ViT. In spite, of significant progress made through the deep learning models and transformer models there are critical research gaps persist in the context of plant disease classification.

Firstly, CNN lacks generalization when applied to real-world datasets with complex backgrounds, varied lighting, and diverse disease stages, leading to overfitting. Secondly, the use of a single deep learning model or transformer model limits the model capability in capturing the features for instance, CNN focus locally, missing global context, while ViTs captures global dependencies but often misses fine-grained details. This leads to one of the critical features of spaces remaining unexplored. Thirdly, CNN and Transformer hybrid model employs shallow fusion strategies that combine features at the final layer, resulting in loss of important features.

To bridge these gaps, the current research focuses on building a hybrid architecture, a multi-transformer integration combining the global attention mechanism of Vision Transformer (ViT) and hierarchical and localized feature extraction capability of Swin Transformers, thereby exploring the strengths of both models to achieve more accurate and efficient feature representation.

The objective of this research is to develop an efficient hybrid transformer model for accurate classification of crop diseases. The primary contribution of this study is summarized as follows:

- The primary innovation of the Swin-HViT model lies in the integration of ViT and Swin Transformer within a unified framework. In the dual architecture, ViT captures the long-range dependencies and the global features while the Swin Transformer uses window-based self-attention with shifted windows to capture the local features.
- The model performs late feature fusion by concatenating the global [CLS] embedding from ViT with the pooled hierarchical feature embedding from Swin Transformer. This fusion combines long-range semantic context with local multi-scale spatial information, leading to richer representations, better generalization, and improved classification performance.

2. Related Works

There are several works that exist on crop disease prediction and classification, ranging from deep learning models, transformer-based models, and ensemble learning models. The first group of researchers used CNN and pre-trained models for efficient classification of crop diseases. Pal et al. [14] developed a lightweight mobile application Mob-Res for precise detection of crop disease. The author used MobileNetV2 for feature extraction and ResNet for classification. The model achieved an accuracy of 97.73% on plant disease dataset and 99.47% on plant village dataset. A. Y. Ashurov et al. [15] proposed a novel customized CNN deep-learning model integrating squeeze-and-excitation block with a improved residual skip connection and achieved an accuracy of 98% on plant village dataset. Oni et al. [16] developed a custom CNN deep-learning model for the tomato leaf dataset collected from a local farm and achieved an accuracy of 95.2% when compared to the pre-trained models such as YOLOv5, MobileNetV2, ResNet18 with accuracy 77%, 89.38% and 71.88% respectively. Rezaei et al.'s [17] novel few-shot learning (FSL) technique is suitable when the dataset is limited to as few as five images per class. The speciality of the model is it uses pre-training coupled with meta-learning and feature attention (FA) mechanism. ResNet50 and ViT are used as feature learners. Krishna et. al [18] developed series of deep learning models such as EfficientNet-B0, EfficientNet-B3, DenseNet201 and ResNet50 on PlantDoc dataset and achieved an accuracy of 76.77% using EfficientNetB3 model.

The second group of researchers used vision transformer models for crop disease prediction. Barman et. al [7] proposed a smartphone – based solution for plant disease prediction using ViT. They also compared the performance of ViT with InceptionV3 and found ViT performs much better than deep learning models. Li et. al [19] proposed a real-time light-weight plant-based MobileViT(PMVT) in which the convolution blocks were replaced with residual structure that effectively captures long-distance dependencies between different images. The author also included convolutional block attention module to focus on essential features. The model was evaluated on wheat dataset and achieved an accuracy of 93.6% which is 1.6% higher than MobileNetV3.

The third group of researchers used hybrid models for crop disease prediction. Vallabhajosyula et al. [20] proposed a Hierarchical Residual Vision Transformer (HRVT) that combines an Improved Vision Transformer (IVT) for feature extraction and a ResNet-9 network for classification. This hybrid Transformer–CNN model effectively captures both global contextual and local spatial features from leaf images. It achieves an accuracy of 68.53% compared to large models like ResNet50 or InceptionV3. The model performs exceptionally well across three datasets: Local Crop Dataset (13 classes), Plant Village Dataset (38 classes), Extended Plant Village Dataset (51 classes). Tunio et al. [12] proposed a hybrid framework that combines CNN for local feature extraction and MViT for global feature extraction. Further, the author used feature-fusion method to align channel dimensions and enhance the local details of global representation. The author evaluated the model with three datasets: plant village, plantDoc and AC-plantDoc and obtained a performance improvement of 13.67% compared with SOTA and baseline model. Aboelenin et al. [10] proposed a hybrid model using VGG16, Inception-V3 and DenseNet for global feature extraction and ViT model for classification. The author evaluated the model on two datasets: Apple and Corn dataset and obtained an accuracy of 99.24% and 98% respectively. Yang et al. [21] proposes a novel leaf disease identification network (LDI-NET) that uses a multi-label method to simultaneously identify plant type, leaf disease, and severity in a single branch model. LDI-NET architecture leverages the strengths of both Convolutional Neural Networks (CNNs) and transformers for extracting local and long-range global features. Zeng et al. [11] introduces a novel deep – learning framework for plant disease prediction. In this model, CNN is used for local feature extraction, ViT with self-attention mechanism is used to capture global dependencies and the classification is done by the fully connected layer. The model is evaluated on three different datasets: cucumber, banana and plant village dataset. The model achieved 97% accuracy with increased brightness and achieved 93% accuracy with 30% decreased brightness. The model achieved a peak accuracy of 97% with a 30% increase in image brightness and maintaining an accuracy of 93% even with a 30% decrease in brightness. Beak et al. [22] introduces attention-score based multi-ViT model to detect crop disease. The author has used multiple pre-trained ViT to learn diverse representation of the input image. The model is tested on apple leaf and grape leaf dataset. Zhang et al. [23] proposed a hybrid model for identifying plant leaf disease in a distributed agriculture data source. The author used Swin transformer for feature extraction with federated learning to enable privacy-preserving model training

The model performance many vary based on the quality of the dataset used in real-world scenarios. Plant leaf images often contain significant noise, background objects, and other distortions. Traditional deep learning (DL) models primarily capture local features, which limits their ability to handle such complex variations. In contrast, transformer-based models can capture global dependencies across the image. However, each transformer architecture has its own limitations. To address this limitation, we propose a hybrid transformer model that combines the strengths of the ViT and the Swin transformer. The Swin transformer is well-suited for real-world applications, as it is designed to perform effectively on images captured in natural, uncontrolled environments rather than in lab settings.

3. Materials and Methods

This research introduces a hybrid transformer model Swin-HViT to efficiently identify the plant disease. The hybrid transformer model combines two transformer models: ViT and Swin transformer as shown in Fig 1. ViT with the self-attention mechanism captures the global context of an image thus, it has a strong global feature modeling and simpler architecture than the CNN. Unlike ViT, the Swin transformer follows a hierarchical feature representation. Swin transformer uses window-based and shifted window self-attention mechanisms to capture both local and global features effectively.

3.1 ViT Transformer

The images from the dataset are resized to 224×224 pixels to match the ViT input requirements. The resized images are normalized to ensure numerical stability, balance color channels, and to improve generalization. The normalization is explained using the following formula:

$$x' = \frac{x - \mu}{\sigma} \quad (1)$$

where, x is the original pixel value (0-255), μ is the mean pixel value (per channel R, G, B), σ is the standard deviation (per channel), x' is the normalized pixel value.

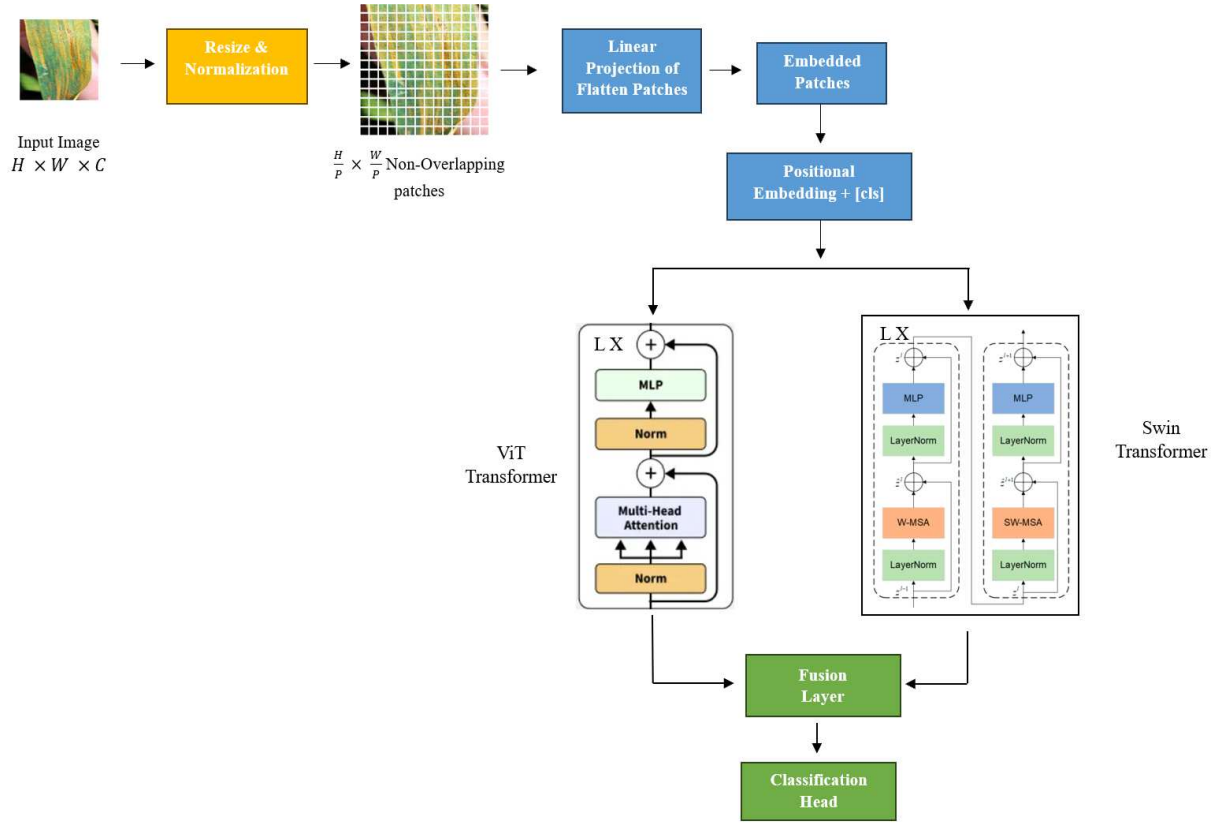


Fig 1 Proposed Hierarchical Vision Transformer (Swin-HViT) Architecture

The normalized images are then converted into sequences of non-overlapping patches of size $P \times P$ (in ViT $P=16$). If the input image size is $H \times W \times C$ (in ViT it is commonly $224 \times 224 \times 3$) and patch size is 16×16 , the number of patches $N = \frac{H}{P} \times \frac{W}{P}$. Therefore, the input images are converted into 196 ($14 \times 14 = 196$) non-overlapping patches of size (16×16). Each patch is flattened into 1D vector length of ($C \times P \times P$) and projected into vector embedding of dimension D ($3 \times 16 \times 16 = 768$).

A special learnable classification token denoted as $[CLS]$ is appended to the input sequence to represent a global summary of the entire image. After passing through all transformer layers, the final state of this token serves as the aggregated representation used for the image classification task. Positional encoding is attached to the patch embedding to maintain spatial dependencies. The final sequence is fed into the transformer's encoder stack, with L identical layers.

The Transformer Encoder block consist of L identical transformer encoder layer. Each layer has two core components:

- Multi-Head Self-Attention (MSA) Block
- Multi-Layer Perceptron (MLP) Block

MSA allows each token (patch or $[CLS]$) to attend to all other tokens in the sequence, computing a weighted sum of their feature vectors. This captures global context and relationships between distant image patches. The input is linearly projected into Query (Q), Key (K), and Value (V) matrices. Attention weights are calculated as:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \times V \quad (2)$$

where, $\sqrt{d_k}$ is the scaling factor. The output of the token from the final encoder layer is extracted and passed to an MLP head for the final classification prediction.

3.2 Swin Transformer

The Swin (Shifted Window) transformer introduces hierarchical representations and local-window attention to improve efficiency on large images. In the Swin transformer, the input image has dimensions ($H \times W \times C$), where H represents height, W width, and C number of color channels (typically RGB). The input images are generally resized to 224×224 pixels and normalized to a range of $[0, 1]$ before being processed by the model.

The input image is divided into sequence of non-overlapping patches of size 4×4 . The number of patches per image is given as $N = \frac{H}{P} \times \frac{W}{P}$. Patch is a small tensor of shape $C \times P \times P$. Therefore, the flattened length $L = C \times P \times P$. We learn a weight matrix W and bias b that project length- L patch vectors into an embedding of dimensionality D as given in the Formula:

$$e = Wx + b \quad (3)$$

Each patch maps to a 96-dim feature vector. Swin block consists of two stages.

- Window Multi-Head Self-Attention (W-MSA): Self-attention computed within local windows.
- Shifted Window Multi-Head Self-Attention (SW-MSA): Windows are shifted by half their size in the next block to enable cross-window connections. This shift alternation allows global context learning without excessive computation.

After a few Swin blocks, adjacent patches are merged (e.g., 2×2) to reduce spatial resolution and increase channel depth. This creates a hierarchical feature pyramid, similar to CNNs. In four stages the channel depth becomes (96 $\rightarrow$ 192 $\rightarrow$ 384 $\rightarrow$ 768 channels) thus helping the model to capture both local and global features. Finally, the feature maps are globally pooled and fed into a fully connected layer for classification.

3.3 Fusion Strategy

The final token $[CLS]$ contains the global image representation of the ViT model. In Swin transform, the final hidden stage contains the flattened global feature of the input image. Both feature vectors are concatenated along the feature dimension. By this way we use all the features learnt from the both the models without taking averaging or mixing. ViT learns the global features, Swin learns the spatial features, by concatenating both we get a deeper understanding of the image semantics. Algorithm 1 provides a detailed description of the hybrid model.

Algorithm 1: Hybrid Model

Input:

Image Dataset $D = \{(x_i, y_i)\}_{i=1}^N$, where x_i is a leaf image of a crop and $y_i \in \{1, 2, \dots, C\}$ is the disease class.

Pre-trained ViT (M_{ViT}) and Swin Transformer (M_{Swin}) model

Output:

Predicted crop diseases class y'

Step 1: Data Pre-processing

Input images are resized to 224×224 pixel and are normalized using ViT image processors.

Step 2: Split the dataset into training, validation, and testing sets.

Step 3: Model Initialization

Load the pre-trained ViT and Swin transformer and remove their original classification heads.

Step 4: Extract:

ViT hidden representation from the last transformer layer using the $[CLS]$ token. $f_{ViT} = M_{ViT}(x)$

Swin Transformer hidden representation from the final layer and apply mean pooling across spatial tokens.

$f_{Swin} = M_{Swin}(x)$

Step 5: Concatenate ViT and Swin feature vectors. $f_{hybrid} = [f_{ViT} || f_{Swin}]$

Step 6: Add a fully connected layer for final classification. Pass fused features through the classifier to obtain logits and apply softmax to obtain class probabilities.

Step 7: Model Training and evaluation

End Algorithm

3.4 Datasets

The first dataset used for hybrid model evaluation is available on Kaggle at <https://www.kaggle.com/datasets/smaranjitghose/corn-or-maize-leaf-disease-dataset> (accessed on August 2025). The dataset consists of 4188 images that belong to four different classes: “Healthy”, “Common_Rust”, “Gray_Leaf_Spot”, and “Blight”, as shown in Fig 2. It contains high-quality leaf images from the Corn crop captured under controlled laboratory conditions. The dataset distribution is shown in Fig 3.

The second dataset is also from Kaggle available at the link <https://www.kaggle.com/datasets/abdulhasibuddin/plant-doc-dataset> (accessed on August 2025) [24]. The dataset consists of 2551 images that belong to 27 crops as shown in Fig 4. PlantDoc contains images of multiple crops (such as tomato, maize, grape, apple, etc.) and covers several disease categories as well as healthy samples. The dataset distribution is shown in Fig 5.

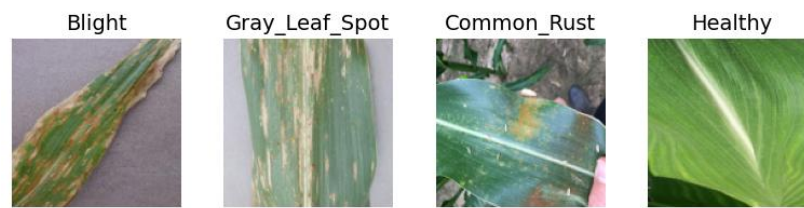


Fig 2 Sample Images of 4 classes from the dataset

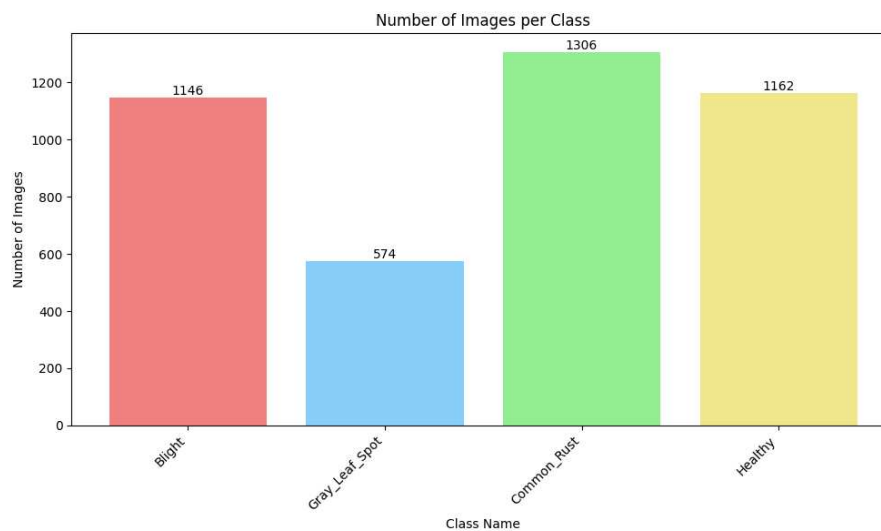


Fig 3 Class Distribution of Cron Dataset with 4 Classes

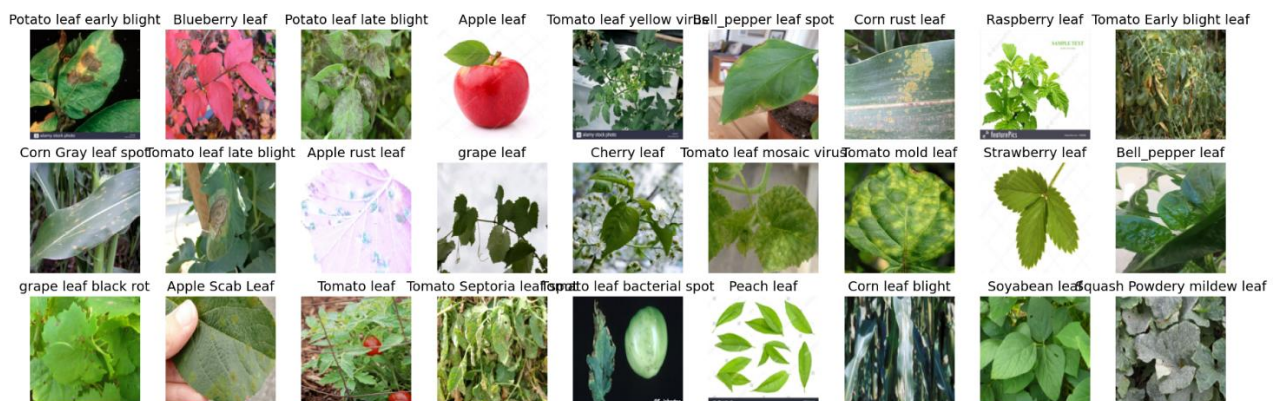


Fig 4 Sample Images of 27 classes from PlantDoc Dataset

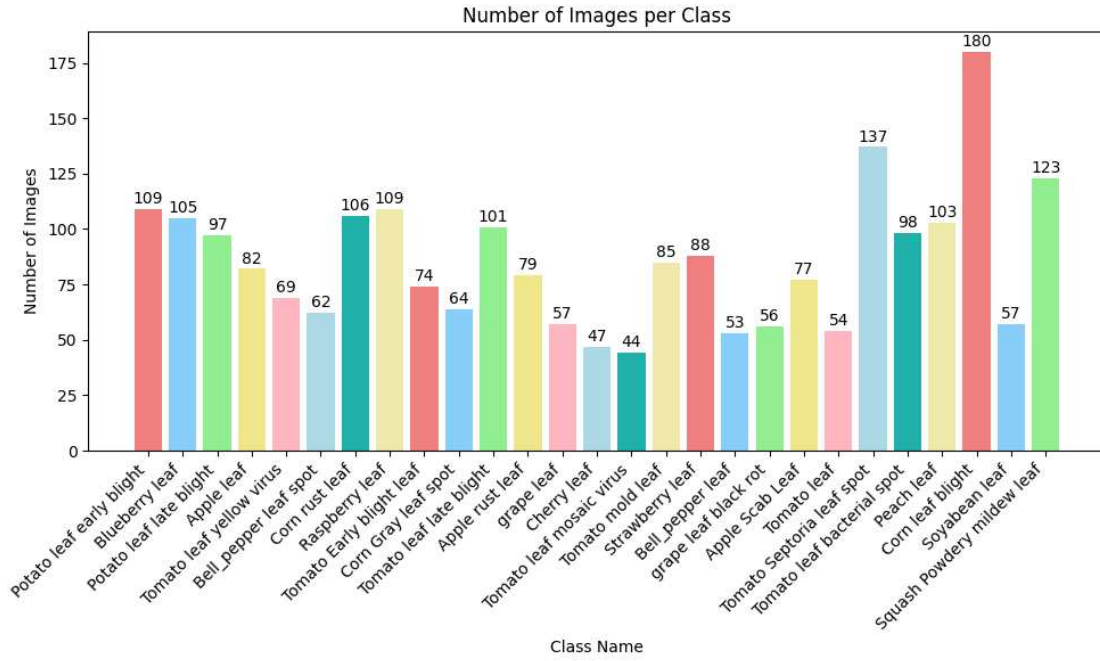


Fig 5 Class Distribution of PlantDoc Dataset with 27 Classes

4. Experimental Setup

The model was developed and tested in Google Colab Pro environment with python version 3.10. We used T4 GPU with high RAM session. The dataset was partitioned into training, validation, and testing subsets with proportions of 80%, 10% and 10% respectively, as summarized in Table 1.

Table 1 Data Distribution for the Corn Dataset and PlantDoc Dataset

Dataset	Training Samples	Testing Samples	Validation Samples
Cron	3392	377	419
Plant Doc	1875	209	232

Model performance is assessed using standard evaluation metrics, including accuracy (4), precision (5), recall (6) and F1-score (7) as defined below:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

$$F1\ Score = \frac{2TP}{2TP+FP+FN} \quad (7)$$

In this context, TP, TN, FP, and FN represent true positive, true negative, false positive, and false negative outcomes respectively. The model was trained for 15 epochs using the AdamW optimization algorithm with a categorical cross-entropy loss. A learning rate of 0.0001 and a batch size of 32 were employed as summarized in Table 2. The embedding dimensions were set to 768 for the Vision Transformer and 128 for the Swin Transformer.

Table 2 Hyper-Parameter Tuning

Function	Paramter	Value
Training Parameter	Optimizer	AdamW
	Learning Rate	0.0001
	Epochs	15
	Batch Size	32
ViT Parameters	Embedding Dimension	768
	Number of Attention Heads	12
	Patch Size	16
Swin Parameters	Embedding Dimension	128
	Number of Attention Heads	Varies in each stage [4, 8, 16, 32]
	Patch Size	4

5. Results and Discussion

The classification metrics for the Corn dataset with four classes are shown in Table 3. The proposed Swin-HViT Model demonstrates the best overall performance with a weighted average accuracy of 98.81%. It achieved the highest values for precision, recall, and F1-score in nearly every class, especially performing perfectly for the Healthy and Common Rust class and nearly perfect for Blight and Gray Leaf Spot class. The evaluation results on the more diverse, real-time PlantDoc Dataset, with 27 classes across 13 crop species, achieved an accuracy of 81.81%. There is a drop in accuracy due to the characteristics of the dataset. The PlantDoc dataset contains real-time images scraped from the internet, which include noise, diverse backgrounds, varying lighting conditions, and inconsistent image quality, all of which make the classification task more challenging.

Table 3 Classification results of Corn and PlantDoc Datasets

Dataset	Class	Evaluation Metrics			
		Accuracy	Precision	Recall	F1-Score
Corn	Blight	0.9881	0.9739	0.9825	0.9782
	Common_Rust		1.0000	0.9854	0.9926
	Gray_Leaf_spot		0.9615	0.9804	0.9709
	Healthy		1.0000	1.0000	1.0000
PlantDoc	Apple Scab Leaf	0.8181	0.8333	0.6250	0.7143
	Apple Leaf		1.0000	1.0000	1.0000
	Apple Rust Leaf		0.7000	0.8750	0.7778
	Bell_Pepper Leaf		0.8000	1.0000	0.8889
	Bell_pepper leaf spot		0.6667	0.4000	0.5000
	Blueberry Leaf		1.0000	0.9000	0.9474
	Cherry leaf		0.8571	1.0000	0.9231
	Cron Gray Leaf Spot		0.5000	0.1250	0.2000
	Corn Leaf Blight		0.6500	0.9286	0.7647
	Corn Rust Leaf		1.0000	1.0000	1.0000
	Peach Leaf		1.0000	1.0000	1.0000
	Potato Leaf Early Blight		0.6667	0.5455	0.6000
	Potato Leaf Late Blight		0.4545	0.5556	0.5000
	Raspberry Leaf		1.0000	1.0000	1.0000
	Soyabean Leaf		0.8333	0.8333	0.8333
	Squash Powdery Mildew Leaf		0.9167	0.7857	0.8462
	Strawberry		1.0000	1.0000	1.0000
	Tomato Early Blight Leaf		0.5000	0.2500	0.3333
	Tomato Septoria Leaf Spot		0.5455	0.7059	0.6154
	Tomato Leaf		0.6000	0.6000	0.6000
	Tomato Leaf Bacterial Spot		0.3500	0.4375	0.3889
	Tomato Leaf Late Blight		0.6923	0.5625	0.6207
	Tomato Leaf Mosaic Virus		0.5000	0.3333	0.4000
	Tomato Leaf Yellow Virus		0.7778	1.0000	0.8750
	Tomato Mold Virus		0.5714	0.6667	0.6154
	Grape Leaf		0.8000	1.0000	0.8889
	Grape Leaf Black Rot		1.0000	0.8333	0.9091

The graphs in Fig 6. a summarizes the training and validation loss of the hybrid model over 15 epochs on the Corn dataset. The validation loss drops steadily but has a slight increase in epoch 6 and after epoch 7 it stabilizes. This behavior suggests that the model generalizes well overall, with no severe overfitting. The training and the validation accuracy remains consistently high except in epoch 6 as shows in Fig 10. b indicates good generalization and stable learning across the epoch.

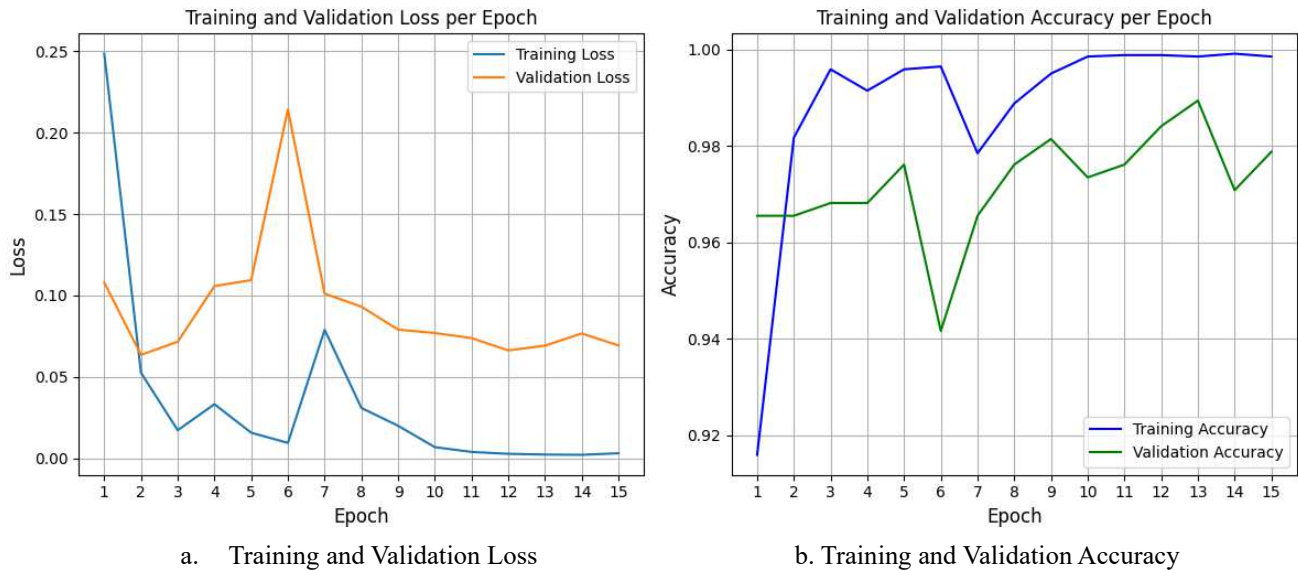


Fig 6 Training and Validation Loss and Accuracy for Corn Dataset

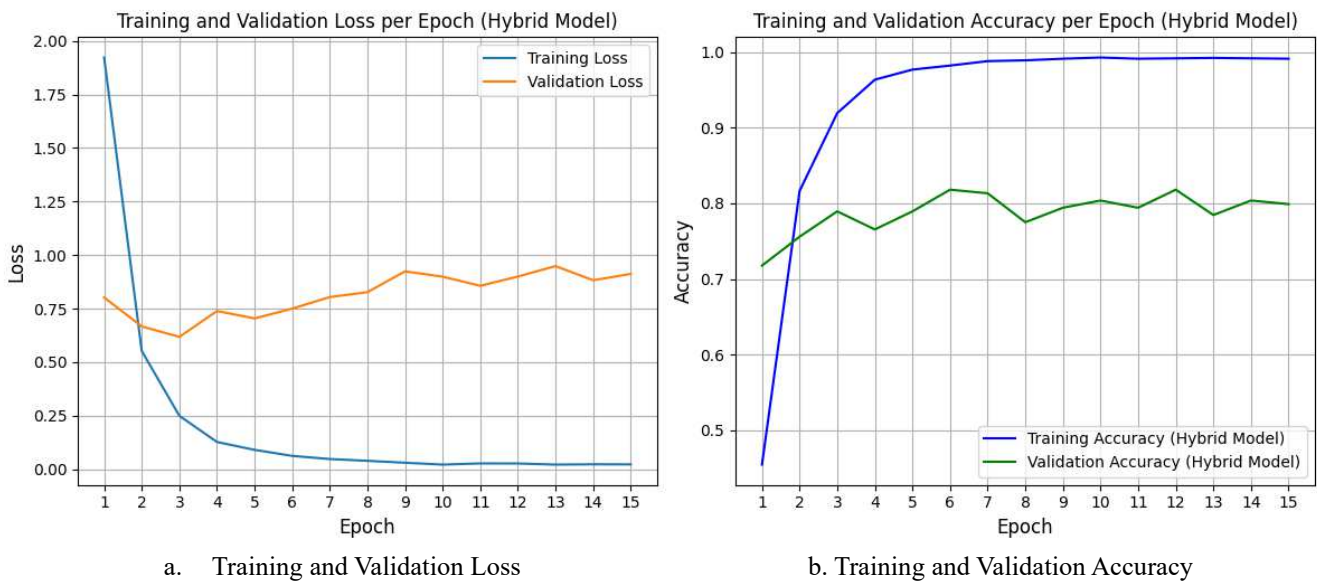


Fig 7 Training and Validation Loss and Accuracy for PlanDoc Dataset

The training loss as shown in Fig 7.a for the PlantDoc dataset shows a sharp reduction in the early epochs, and it converges effectively in the later epochs. In contrast, the validation loss decreases initially but gradually increases and fluctuates in the later epochs. The training accuracy as shown in Fig 7. b rises rapidly confirming excellent performance on training dataset. But the validation accuracy remains stable in the range 75% - 82% with some oscillations across the epochs.

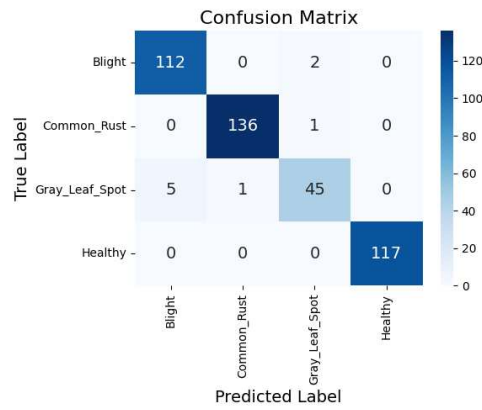


Fig 8 Confusion Matrix for Corn Dataset

The confusion matrix for the hybrid model on Corn dataset is shown in Fig 8. It demonstrates strong overall classification performance, with most predictions falling along the main diagonal, indicating correct classifications for each class. The three classes namely: "Blight", "Common_Rust", and "Healthy" samples are classified correctly, with only a minimal number of misclassifications. The primary confusion occurs with "Gray_Leaf_Spot," where a few samples are occasionally misidentified as "Blight" or "Common_Rust," though the number of such errors is minimal. The confusion matrix of the hybrid model for PlantDoc dataset is shown in Fig 9. Most of the predictions are correctly aligned across the diagonal with few misclassifications noted off diagonal in some classes, specially classes with few samples and with similar symptoms. The hybrid model misclassifies the class "squash powdery mildew leaf" as "tomato leaf bacterial spot", "tomato mold leaf" and "grape leaf". The AUC-ROC curve for the Corn Dataset is shown in Fig 10. The hybrid model Swin-HViT achieved an AUC of 1.00 for all the classes on the Corn dataset. This indicates a perfect or near-perfect discrimination capability of the model across all the classes. The Fig 11 presents class-wise ROC curves for PlantDoc dataset, where most classes achieve AUC values close to 1.0, indicating excellent discrimination performance. A few classes with slightly lower AUCs reflect higher visual similarity or data imbalance, but overall, the model demonstrates strong and reliable multi-class classification capability.

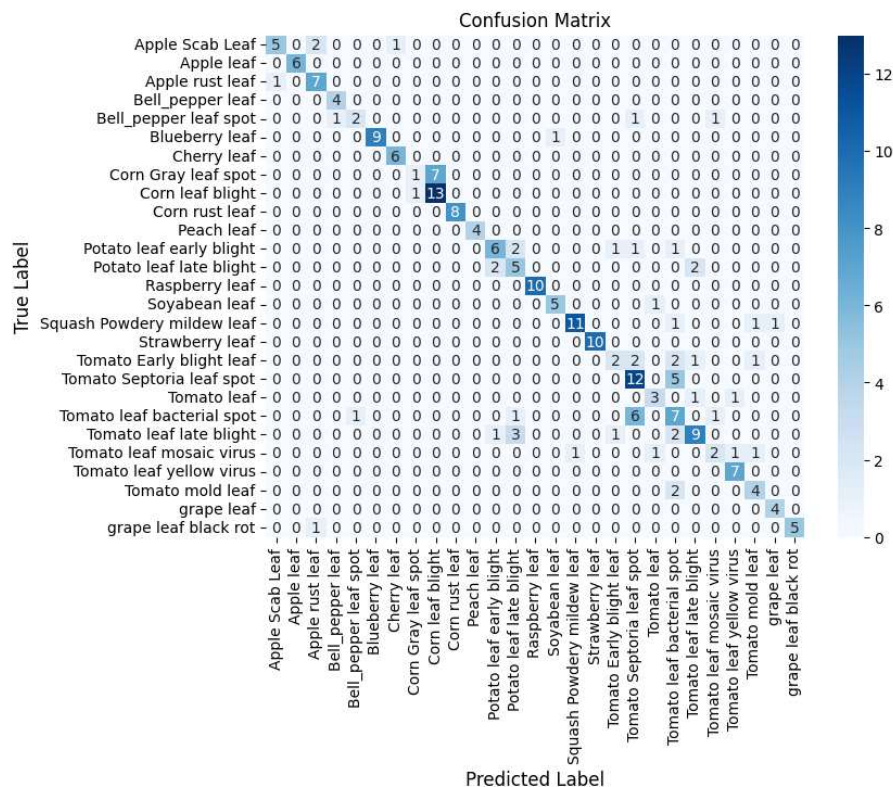


Fig 9 shows the confusion matrices for the PlantDoc dataset.

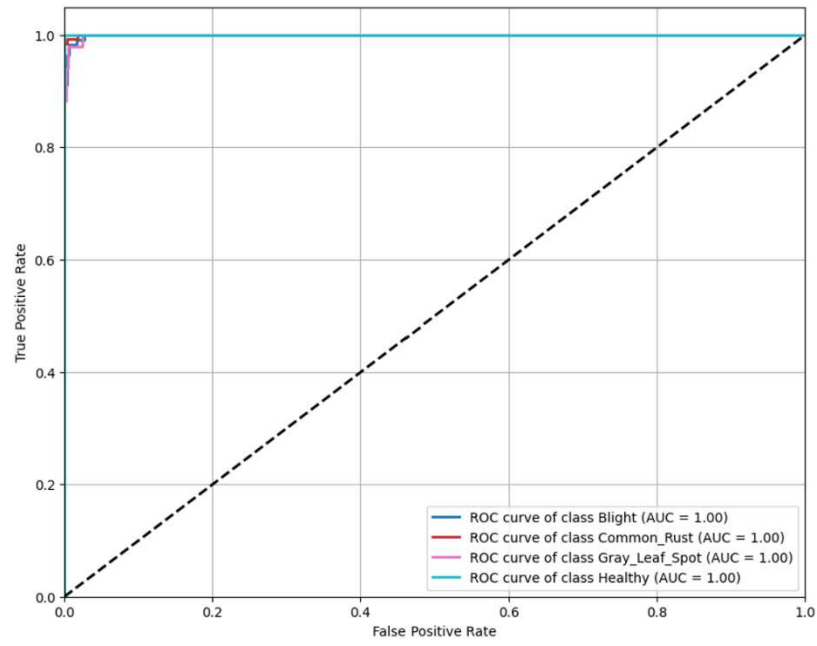


Fig 10 AUC ROC Curve for Corn Dataset

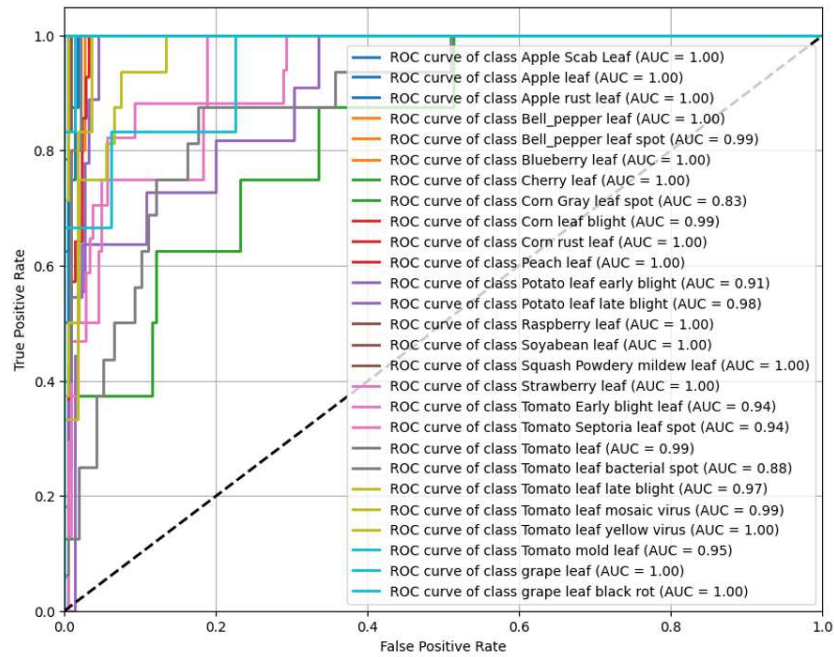


Fig 11 AUC – ROC Curve for PlantDoc Dataset

The proposed hybrid Swin-HViT model is compared with the existing SOTA models in Table 5. Swin-HViT outperforms existing methods on both the Corn and PlantDoc datasets, achieving the highest reported accuracies of 98.81% and 81.81% respectively. The existing deep learning models referred in [20, 27] and hybrid models in [17, 26] reported accuracies ranging from 94% to 98% with the Corn dataset. The proposed Swin-HViT reported 0.81% more than the existing model [20]. The existing deep learning models [19] and hybrid models [8, 18] reported accuracy ranging from 70% to 80.02% on PlantDoc dataset. The proposed Swin-HViT reported 81.8% which is 1.8% more than the existing model [8]. Thus, the proposed model Swin-HViT sets a new benchmark for crop disease classification.

Table 5 Comparison of Proposed Model with the Existing Models

References	Dataset	Methods	Accuracy
[13] (2024)	Cron Dataset	MobileNetV2 and Vision Transformer (ViT) Hybrid Model	96.73%
[10] (2025)	Cron Dataset	Hybrid CNN and ViT	98%
[25] (2020)	Cron Dataset	Optimized Dense CNN	98.06%
[26] (2024)	Cron Dataset	ViT-B/16 with SGD optimizer	94.51%
[12] (2025)	PlantDoc	Convolutional neural networks (CNNs) and MViTs	70%
[18] (2023)	PlantDoc	EfficientNet-B3	73.31%
[8] (2025)	PlantDoc	ST-CFI Swin Transformer with CNN	80.02%
Proposed (Swin-HViT)	Cron Dataset	Hybrid ViT-Swin	98.81%
Proposed (Swin-HViT)	PlantDoc	Hybrid ViT-Swin	81.8%

Conclusion

In this work, a hybrid deep learning model Swin-HViT combining the Vision Transformer (ViT) and Swin Transformer architectures was proposed for plant disease classification. By integrating the global contextual modelling capability of ViT with the hierarchical and localized feature learning of Swin Transformer, the hybrid model effectively captures both coarse and fine-grained disease characteristics. In this research, we evaluated the proposed hybrid transformer model on two datasets: Corn and PlantDoc. On the Corn dataset the proposed Swin-HViT achieved the classification accuracy of 98.81% and on PlantDoc it achieved an accuracy of 81.81% which is the highest among all the existing SOTA models. These findings highlight the effectiveness of combining complementary transformer architectures to improve disease identification across diverse, complex real-world agricultural datasets. Overall, the hybrid model provides an efficient and reliable solution for automated plant disease detection and can be extended to other agricultural and visual recognition tasks in the future.

Funding statement

This research work is financed by University of Technology and Applied Sciences, Ibri, Sultanate of Oman under the Internal Research Grant, Reference number: IRG-IBRI-25-37. The researchers show their gratitude to the university and its Research and Consultancy Department for the necessary motivation.

Declaration of Competing Interest

The authors of this research work declare that they do not have any known competing or personal interests that can appear to influence the work in this paper.

Author Contribution

Conceptualization, Writing HG, WB; Review and Editing, AG, NV; Methodology, HG ; Formal analysis HB, AC, Investigation and Supervision, NV, WB.

Reference

- [1] Maria Lodovica Gullino, Ramon Albajes, Francislene Angelotti, Subrata Chakraborty, Karen A. Garrett, and Brett Phillip Hurley, *Scientific review of the impact of climate change on plant pests*. Food and Agriculture Organization of the United Nations Rome, 2021. doi: 10.4060/cb4769en.
- [2] M. Aleme and G. Mengistu, "Impacts of Diseases and Pests on Forage Crop Production and Management Systems: A Review," *International Journal of Ecotoxicology and Ecobiology*, vol. 9, no. 3, pp. 104–111, Sep. 2024, doi: 10.11648/j.ijee.20240903.12.
- [3] H. Lu *et al.*, "A survey on deep learning-based object detection for crop monitoring: pest, yield, weed, and growth applications," *Visual Computer*, vol. 41, no. 12. Springer Science and Business Media Deutschland GmbH, pp. 10069–10094, Sep. 01, 2025. doi: 10.1007/s00371-025-04022-4.

- [4] X. Dong *et al.*, “PDDD-PreTrain: A Series of Commonly Used Pre-Trained Models Support Image-Based Plant Disease Diagnosis,” *Plant Phenomics*, vol. 5, 2023, doi: 10.34133/plantphenomics.0054.
- [5] R. Elghawth, W. Abbaoui, A. Ariss, and S. Ziti, “Deep Learning for Transformer-Based Plant Disease Detection: A Bibliometric Analysis,” in *ICATH 2025*, Basel Switzerland: MDPI, Oct. 2025, p. 29. doi: 10.3390/engproc2025112029
- [6] Y. Wang, Y. Deng, Y. Zheng, P. Chattopadhyay, and L. Wang, “Vision Transformers for Image Classification: A Comparative Survey,” *Technologies*, vol. 13, no. 1. Multidisciplinary Digital Publishing Institute (MDPI), Jan. 01, 2025. doi: 10.3390/technologies13010032.
- [7] U. Barman *et al.*, “ViT-SmartAgri: Vision Transformer and Smartphone-Based Plant Disease Detection for Smart Agriculture,” *Agronomy*, vol. 14, no. 2, Feb. 2024, doi: 10.3390/agronomy14020327.
- [8] S. Yu, L. Xie, and L. Dai, “ST-CFI: Swin Transformer with convolutional feature interactions for identifying plant diseases,” *Scientific Reports*, vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-08673-0.
- [9] Changxia Sun, Yong Li, Zhengdao Song, Qian Liu, Haiping Si, Yingjie Yang, Qing Cao, “Research on tomato disease image recognition method based on DeiT”, *European Journal of Agronomy*, vol. 162, 2025, <https://doi.org/10.1016/j.eja.2024.127400>.
- [10] S. Aboelenin, F. A. Elbasheer, M. M. Eltoukhy, W. M. El-Hady, and K. M. Hosny, “A hybrid Framework for plant leaf disease detection and classification using convolutional neural networks and vision transformer,” *Complex and Intelligent Systems*, vol. 11, no. 2, Feb. 2025, doi: 10.1007/s40747-024-01764-x.
- [11] Z. Zeng, T. Mahmood, Y. Wang, A. Rehman, and M. A. Mujahid, “AI-driven smart agriculture using hybrid transformer-CNN for real time disease detection in sustainable farming,” *Scientific Reports*, vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-10537-6.
- [12] M. H. Tunio *et al.*, “Advancing plant disease classification: A robust and generalized approach with transformer-fused convolution and Wasserstein domain adaptation,” *Computers and Electronics in Agriculture*, vol. 227, p. 109574, 2024, doi: <https://doi.org/10.1016/j.compag.2024.109574>.
- [13] Özüpak, Y., Alpsalaz, F., Aslan, E., & Uzel, H, Hybrid deep learning model for maize leaf disease classification with explainable AI. *New Zealand Journal of Crop and Horticultural Science*, vol. 53, no. 5, p. 2942–2964, 2025, <https://doi.org/10.1080/01140671.2025.2519570>.
- [14] C. Pal, S. Karmakar, I. Mukherjee, and P. P. Chakrabarti, “A lightweight and explainable CNN model for empowering plant disease diagnosis,” *Scientific Reports*, vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-94083-1.
- [15] A. Y. Ashurov *et al.*, “Enhancing plant disease detection through deep learning: a Depthwise CNN with squeeze and excitation integration and residual skip connections,” *Frontiers in Plant Science*, vol. 15, 2024, doi: 10.3389/fpls.2024.1505857.
- [16] M. K. Prama Tabia Tanzin and Oni, “Optimized Custom CNN for Real-Time Tomato Leaf Disease Detection,” in *Computational Science – ICCS 2025 Workshops*, A. S. and Z. Y. J. Paszynski Maciej and Barnard, Ed., Cham: Springer Nature Switzerland, 2025, pp. 81–88.
- [17] M. Rezaei, D. Diepeveen, H. Laga, M. G. K. Jones, and F. Sohel, “Plant disease recognition in a low data scenario using few-shot learning,” *Computers and Electronics in Agriculture*, vol. 219, Apr. 2024, doi: 10.1016/j.compag.2024.108812.
- [18] M. S. Krishna, P. Machado, R. I. Otuka, S. W. Yahaya, F. Neves dos Santos, and I. K. Ihianle, “Plant Leaf Disease Detection Using Deep Learning: A Multi-Dataset Approach,” *J*, vol. 8, no. 1, p. 4, Jan. 2025, doi: 10.3390/j8010004.
- [19] G. Li, Y. Wang, Q. Zhao, P. Yuan, and B. Chang, “PMVT: a lightweight vision transformer for plant disease identification on mobile devices,” *Frontiers in Plant Science*, vol. 14, 2023, doi: 10.3389/fpls.2023.1256773.

- [20] S. Vallabhajosyula, V. Sistla, and V. K. K. Kolli, "A novel hierarchical framework for plant leaf disease detection using residual vision transformer," *Heliyon*, vol. 10, no. 9, May 2024, doi: 10.1016/j.heliyon.2024.e29912.
- [21] B. Yang *et al.*, "A novel plant type, leaf disease and severity identification framework using CNN and transformer with multi-label method," *Scientific Reports*, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-62452-x.
- [22] E. T. Baek, "Attention Score-Based Multi-Vision Transformer Technique for Plant Disease Classification," *Sensors*, vol. 25, no. 1, Jan. 2025, doi: 10.3390/s25010270.
- [23] H. Zhang and G. Ren, "Intelligent leaf disease diagnosis: image algorithms using Swin Transformer and federated learning," *Visual Computer*, vol. 41, no. 7, pp. 4815–4838, May 2025, doi: 10.1007/s00371-024-03692-w.
- [24] D. Singh, N. Jain, P. Jain, P. Kayal, S. Kumawat, and N. Batra, "PlantDoc: A Dataset for Visual Plant Disease Detection," in *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD*, in CoDS COMAD 2020. New York, NY, USA: Association for Computing Machinery, 2020, pp. 249–253. doi: 10.1145/3371158.3371196.
- [25] A. Waheed, M. Goyal, D. Gupta, A. Khanna, A. E. Hassanien, and H. M. Pandey, "An optimized dense convolutional neural network model for disease recognition and classification in corn leaf," *Computers and Electronics in Agriculture*, vol. 175, p. 105456, 2020, doi: <https://doi.org/10.1016/j.compag.2020.105456>.
- [26] T. and J. R. and R. S. Ramadan Syed Taha Yeasin and Sakib, "Maize Leaf Disease Detection Using Vision Transformers (ViTs) and CNN-Based Classifiers: Comparative Analysis," in *Human-Centric Smart Computing*, J. S. and K. M. Bhattacharyya Siddhartha and Banerjee, Ed., Singapore: Springer Nature Singapore, 2024, pp. 513–524.