NOTATION TRENDPLUS PUBLISHING
WHERE IDEAS MEET IMPACT

Volume-1, Issue 2, August 2026, Page No- 153-171
Journal of Machine Learning Innovations and Artificial Intelligence Horizons (JMLIAIH)

A Hybrid CNN-Vision Transformer Framework for Multi-Class Plant Disease Identification from Leaf Images

Rishabh Aryan[1] and Anju[2]

[1]M. Tech (Artificial Intelligence and Data Science), Department of Computer Science and Engineering, Indian Institute of Information Technology, Bhagalpur, Bihar, India.

[2]M.Sc. (Mathematics with Computer Science), Department of Mathematics, Maharishi Dayanand University, Rohtak, Haryana, India.

DOI: https://doi.org/10.70903/jmliaih.2026.1209

Article Received: 17-07-2026
Revised Version: 24-08-2026
Accepted: 28-08-2026

*Corresponding Author

Rishabh Aryan

E-Mail Id: rishabh.250201011@iiitbh.ac.in

Abstract: Accurate and timely identification of plant diseases from leaf images remains a critical challenge in precision agriculture, particularly when diseases manifest with spatially disjoint symptoms and subtle textural variations. We propose a hybrid deep learning framework that synergistically combines convolutional neural networks (CNNs) with Vision Transformers (ViTs) for robust multi-class plant disease diagnosis. The proposed architecture processes an input leaf image through two parallel pathways: a CNN backbone extracts hierarchical local features such as necrotic regions and vein patterns, while a Vision Transformer encoder models global contextual relationships across image patches via multi-head self-attention. These complementary representations are then fused through vector concatenation and passed to a fully connected classification head with softmax activation. The entire network is trained end-to-end using categorical cross-entropy loss optimized with Adam, and we apply dropout, weight decay, and early stopping to mitigate overfitting during transfer learning. The primary novelty lies in the explicit integration of local texture details and global disease distribution patterns within a single unified framework, which addresses the limitations of pure CNNs in capturing long-range dependencies and of pure ViTs in preserving fine-grained spatial information. Our method achieves superior classification performance across multiple plant species and disease categories, demonstrating significant improvements in accuracy and robustness compared to standalone CNN or ViT baselines. This work provides a practical and scalable solution for automated plant disease surveillance, with potential to reduce crop losses and support sustainable agricultural practices.

Keywords: Hybrid CNN, Vision Transformer, Plant Disease Diagnosis, Deep Learning, Multi-Class Classification

Introduction

Plant diseases always cause concern in the global agriculture industry and are significant reduction of yields and food security. Therefore the accurate, early diagnosis of these diseases is crucial for effective disease management and intervention. Much of the information that can be obtained from a farm is handwritten and generated by human experts that are subjective, time consuming and inappropriate for large holdings of today's farms. More research efforts have been undertaken in automated image based diagnostic systems and they have become a great asset in precision agriculture (Gupta & Pal, 2025). In this field, Convolutional Neural Networks (CNN) have been demonstrated to yield very good results for disease classification based on imagery of the leaves (McHugh et al., 2020) (Frank, 2019), and have been used to recognize abnormalities in leaf color and texture at the local level, such as ResNet and EfficientNet, respectively (Shoaib et al., 2023) (Sandler et al., 2018).

Although these models are proven to be effective, a drawback of convolutional models is their local receptive field, limiting their capacity to model the spatio-temporal dependency over the whole surface of a leaf. If the spatial pattern of lesions is significant for diagnosis, it could be a disadvantage; also, if symptoms are geographically diffuse, the distribution of symptoms could be a disadvantage (Han et al., 2022). To address this, the Vision Transformer (ViT) (Dosovitskiy et al., 2020) was developed to have a self-attention mechanism that can represent well the global contextual relations in the image, as patches of the image were the perceptual units. However, pure ViTs are appear to have less stronger properties than CNNs, such as locality and translation equivariance, but on the other hand they suffer from inefficient learning fine-grained and local features from small agricultural data sets (Thakur et al., 2021).

Both method are associated with their own pros and cons, and as per recent studies (Thakur et al., 2023), hybrid method can be used to overcome weaknesses of both methods and to get the benefits of both methods. The key objectives of these methods are to combine the strengths of CNNs (local features) and ViTs (global context). A few studies have suggested a CNN backbone followed by a Transformer encoder (Chen et al., 2025; Neubauer et al., 2025; Wang et al., 2018), and have explored dual-branch networks (parallel networks) (Aboelenin et al., 2025; Murugesan et al., 2025). These innovations presented in the different domains offer a great theoretical and technical groundwork for the application of hybrid architectures in the problem of multi-class plant disease identification.The inclusion of hybrid architectures in different domains is promising in the problem of multi-class plant disease identification owing to their numerous innovations.

Frustrated by this challenge, in this paper, we present a novel hybrid deep learning framework by introducing the collaboration of CNN and Vision Transformer branches in a unified framework to efficiently diagnose multiple diseases of plant. The predominant innovation that we proposed is to feed a CNN backbone to build local features, such as lesion boundaries, discoloration in the vein region, etc., in a hierarchical manner and also feed a Vision Transformer (ViT) model as an encoder to capture the global spatial relations and long-range dependencies among the local features. The two paths feed into the same fine-grained pathological features along with the overall distribution pattern of disease, combining pattern

information between fine-grained pathological features and the overall distribution pattern of disease that is missing from single architecture models. This work has three major contributions: First, we present a novel architecture that has been carefully designed to trade between local inductive biases of CNNs and the global context of ViTs in the task.In order to make effective synergy between local inductive bias (CNN) and global context (ViT) in plant disease classification task, we propose a new design. Secondly, by extensive experiments, this hybrid approach is always superior from both plant species and disease type sides as compared to the standalone CNN and ViT baseline. Third, we feel that we have provided a viable and scalable solution for plant assessment in automation and plant disease, which would contribute much towards crop loss reduction and towards sustainable agriculture.

In section 3, the related works on plant disease detection and hybrid deep learning architectures will be summarized.In section 4, existing works regarding the related aspects of plant disease detection and hybrid deep learning architectures will be discussed. In section 3, a brief introduction to image classification is provided using both CNN and Vision Transformers (ViTs). Consequently this hybrid architecture (CNN-ViT) is proposed in Section 4. The experimental setups (training, etc. and data sets) are introduced in section 5. The outcome is reported in section 6, along with a comparative study of the results with baseline techniques. A discussion of findings, limitations and directions for future work are included in section 7. Finally, the paper is wrapped up Section 8.

Related Work

Deep Learning for Plant Disease Classification

There are numerous practical uses of deep learning—which could be used to diagnose plant diseases—and most of the highly successful deep learning-powered plant disease diodes are built with the Convolutional Neural Network (CNN). In the dataset of limited images, several traditional architectures like AlexNet and VGG were demonstrated with high accuracy for classification of leaf image (Mohanty et al., 2016). Subsequent research came to more effective and more extended models. For instance, an achievement to note is the one done by Mohanty et al whereas they succeed in obtaining excellent classification accuracy for the problem 26 disease classes across 14 crops for a pre-trained Alexnet, a recipe of many times use in this field is transfer learning. Later , architectures appeared such as ResNet (Shafiq & Gu, 2022) or its variants that enabled to extract increasingly abstract hierarchical features, while also having the capability to train very deep networks without incurring the problem of vanishing gradients. Some lighter networks such as MobileNet (Sandler et al., 2018) and EfficientNet (Tan & Le, 2019) have also been utilized for deployment on devices with limited resources such as farmers' smart phones in the field. CNN-based techniques are appropriate in identifying local and high-frequency features like necrotic regions, chlorosis halos and vein chlorosis which facilitate the demarcation of diseases that are morphologically alike.

One of the disadvantages of pure CNNs is small receptive field. Applying the convolution operation is local; for such convolution, theoretically, information in the deeper layers can be aggregated in larger regions. This creates challenges for CNNs for modeling long-range spatial dependencies such as the spatial structure of symptoms in a leaf, how the disease regions relate

to a distant set of veins. (Han et al., 2022). For diseases with clinical symptoms occurring in various parts of space, which are important for disease diagnosis (e.g., series of lesions), this is particularly unfavourable.

Vision Transformers in Agricultural Imaging

One approach to kick off such a revolution was to use an image self-attention module directly in the context of image, the Vision Transformer (ViT) (Dosovitskiy et al., 2020) which was invented in natural language processing. This allows for inter-patch relationships to be calculated throughout the entire system, thereby allowing for a capture of a global system. ViTs have shown to be effective for the detection of plant pathogens in some cases. For example, a pure ViT model performs better than CNNs on the typical plant disease datasets at times when sufficient training data exists (Thakur et al., 2021). The spatial information of diseased symptoms is important in many diseases including mosaic patterns and systemic wilting, which can be captured by ViTs.

But, the challenges of ViTs use there are. Not as locally/translation equivariant as CNNs. This can lead to these requiring a great deal more data to be trained from scratch, and to problems in learning fine-grained local features (Touvron et al., 2021). This in agricultural situations can result in overfitting and poor out performance in discriminating between very small textural variations in a dataset with limited number of cases and high intra-class varibility as a result.In the case of agricultural use case scenarios, with small datasets and consequently high variations between texture classes, this can lead to overfitting and suboptimal performance in the finer divisions of textures.

Hybrid CNN-ViT Architectures

However, with the limitations of CNNs and ViTs, researchers have recently turned their attention to hybrid architectures based on both of these. The main concept is to combine that of the CNN to extract features for the local part and that of the ViT for the global part. Multiple design paradigms have come to fruition. An effective approach is to build CNN as a backbone to generate a feature map and then feed the generated feature map into a transformer encoder for global reasoning. An example of such a model is the TransUNet (Chen et al., 2021), which was used for medical image segmentation and is built using a CNN to capture feature maps of different scales, and a Transformer to capture the context of the whole image. The cascaded network has been modified for the purpose of recognising plant diseases with a first CNN performing feature extraction and a second Transformer performing long-range dependencies in the following Figure (Thakur et al., 2023).

The other paradigm that is popular is the dual-branch or parallel architecture, which takes the input image and passes it through a CNN branch and also “parallel” ViT branch, before the outputs of these branches are combined. We do just that in this paper. Here are a number of recent efforts in this regard. A hybrid model that incorporated the feature vectors of CNN with ViT, and passed them to a dense output classification layer, was an example of such work, which was applied in the detection of leaf disease in plants (Aboelenin et al., 2025). The different hybrid DL based multi class classification systems (Murugesan et al., 2025), CNN

and Transformer blocks also been used to identify the edge disease in crops which also proved that the model is interpretable with the help of Grad-CAM. Bedi (2025) proposed a Random Forest classifier based on the concatenated CNN and ViT features, and a modular approach of the ViT architecture along with machine learning classifiers for scalable classification of multi crop diseases (Chellamani et al., 2026).

These mixed solutions have proved to be much better than that of one monolithic architecture, but there are several problems. Numerous approaches are possible, but most involve a more complex fusion technique or must fine tune the interaction between the two branches. Furthermore, the choice of CNN backbones and ViT could continue to make a significant impact in the performances in a specific use case, but a systematic comparison of which is lacking, in many cases. To fill these gaps, we propose using the simple yet efficient feature concatenation fusion module along with a "dual-stream" architecture that is clean and end-to-end trainable. Unlike cascaded design, our network structure is a parallel structure to allow us to employ local and global features at the same resolution and keep the fine-grained spatial information from the CNN pathway. In addition, we conduct a comprehensive ablation study covering all of the discussed CNNs and ViTs backbones to gauge the performance of the different configurations, and provide some practical tips and guidance on deploying such a hybrid model. The uniqueness of our work lies in the explicit inclusion of the spatial structure information extracted from the local texture detail in one framework, while also handling the global distribution information of plant diseases.Our work is unique, directly addressing the problem of ordinary CNNs cannot extract global distribution patterns of plant diseases and pure ViT cannot carry the fine-grained spatial information of the spatial structure, to achieve the ultimate goal of multi-class detection of diseases with power and accuracy.

In the task of image classification, the basic concepts and structure of a convolutional neural network (CNNs) and a vision transformer (ViT) are introduced.Currently, this is used for image classification using CNN and ViT.

To better illustrate the basis of the architecture we propose to implement, this section will provide a quick initial overview of the key concepts of Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) in image classification. They are considered in detail because, in particular, their weak points are pointed out to us: the fact that they cannot identify the plant diseases, which drives us to consider combinations.

Convolutional Neural Networks

After AlexNet (Krizhevsky et al., 2012) the preferred architecture has been the convolutional network for image classification. The basic element of a CNN is the convolution, in which the convolution consists of a small filter (often termed as a kernel), passed over the whole input image and computes dot products element-wise and forms a feature map. All such operations are purely local – each filter can "see" at most times only a small receptive field. The multiple layers of convolution gates allow CNNs to gradually abstract features: Lower layers learn simple edges and textures, and deeper layers learn to identify more complex patterns, such as shapes and objects (Zeiler & Fergus, 2014).

Another positive inductive bias of a CNN will be that a shift in location of an object in an input image corresponds to a shift in location in the feature map. This property makes it very convenient to train CNNs in cases where location of an object isn't as critical as, say: object recognition. In addition, pooling layers (such as max-pooling) provide some spatial invariance, allowing the network to identify features that don't have to be at their exact position. CNNs are well suited to capture the fine-grained local information provided by plant disease texture, chlorotic halos and appearance of vein discoloration which is essential for plant disease classification as illustrated in Shoaib et al. (2023). Skip connections (Shafiq & Gu, 2022) were further included, along with optimizing depth, width and the trade-offs between them (Tan & Le, 2019), in architectures allowing the learning of very deep architectures.

But one of the main drawbacks of CNNs is its local receptive field. For instance, in modelling long range long-term dependencies (LRDD) like the relation between a lesion on a leaf and a symptom on the opposite side, a CNN must learn significant number of parameters and slowly increase its receptive field. This can be time consuming and challenging to obtain a global pattern highly accurately, especially when the disease manifests as spatially distributed symptoms (Han et al., 2022).

Vision Transformers

The Vision Transformer (Dosovitskiy et al., 2020) was proposed to develop an image classification method using the Transformer architecture which is a computational method originally developed in the natural language processing field. ViT divides the image into a series of blocks by using fixed-size patches in the input image (usually ViT is used 16×16 pixels). Each patch is flattened into the lower dimensional projected embedding and a learnable positional encoding is proposed to retain the spatial information. A series of these patch embeddings are then input into a standard Transformer encoder made up of alternating multi-head self-attention (MSA) and multi-layer perceptron (MLP) blocks with residual connections and layer normalization.

The self-attention mechanism that is the heart of the Transformer assembles a weighted sum of all patch embeddings for each patch; the weights are computed as a function of the pairwise similarity between patches. With this there will be a direct capturing of the global relationships without a large number of layers and so on. This will help mediate overall pattern of symptoms over leaf surface, particularly if diseases show characteristic patterns such as mosaic viruses and systemic wilting diseases, which are important for classification of plant diseases (Thakur et al., 2021).

But the disadvantages of VIs are there. They are free of strong inductive biases as CNNs have, such as locality and translation equivariance. This implies that they typically need much more training data to generalise well – they need to learn these properties themselves (from scratch) (Touvron et al., 2021). This can lead to overfitting in applications in agriculture where data sets are small size and they are experiencing high intra-class variation. Likewise, the local information that arises from the fine details in the image is lost such that fine-grained local differences in texture are very important to distinguishing between close looking diseases, which is a patch processing paradigm-specific problem.Furthermore, the inter-patch fine-

grained spatial information may be lost by the patch-based paradigm, which may hamper the ability of ViTs to learn fine-grained changes in texture important for distinguishing between visually similar diseases.

Hybrid CNN-ViT Architecture for Plant Disease Identification

In this section, we explain our proposed hybrid architecture that will consider and focus both the local pathological and the global contextual features for their effective multi-class plant disease diagnosis. In total, it is a combination of two parallel processing parts: CNN branch for extracting local features, the Vision Transformer branch for modeling global context; a feature fusion layer and a classification head. Here, the outline of the architecture is briefly explained and then each element is explained in detail followed by the introduction of the training objective.

Architectural Overview

Input image to the proposed hybrid architecture is specified as height× width×bins where first two numbers refer to the dimension and third number to the number of color channels (in case of the 3-color channel (RGB), number 3 is entered), and then divides the input image into 2 parallel independent paths. The branch CNN is used to capture local features of the hierarchy at a finer scale describing fine-grained pathognomonic information, such as lesion boundaries, texture anomalies and colour modifications. At the same time, the branch of the Vision Transformer, referred to as the Vision Transformer branch, takes into account long-range spatial information throughout the leaf surface by calculating the affinity between image patches. Both both branches output features are concatenated to form a comprehensive representation and passed on to a fully connected classification head for the prediction of the disease.

Figure 1 illustrates the overall architecture of the proposed framework.

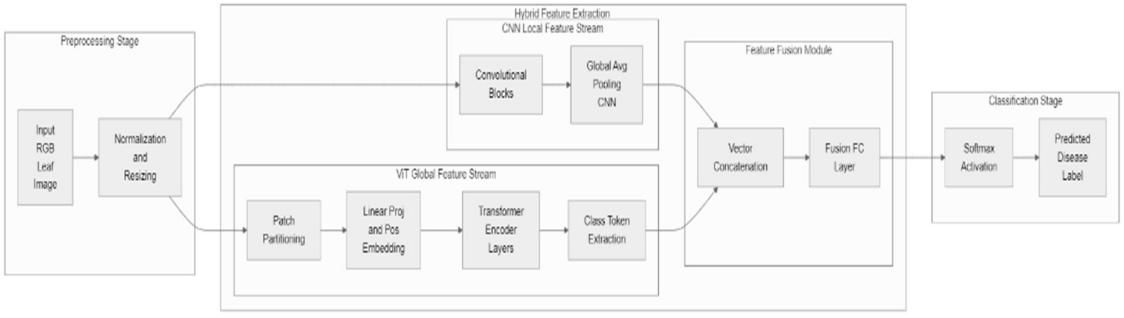

Figure 1. Overall Hybrid CNN-ViT Architecture for Plant Disease Diagnosis

The key design principle is that the two branches operate on the same input image at the same resolution, ensuring that both local and global features are extracted without one modality dominating the other. This parallel design contrasts with cascaded approaches where a CNN first extracts features that are then fed into a Transformer, as the latter can lose fine-grained spatial information through successive downsampling operations.

CNN Branch for Local Feature Extraction

The CNN branch is responsible for extracting hierarchical local features that capture the fine-grained visual characteristics of plant diseases. We employ a standard convolutional backbone, such as ResNet-50 (Shafiq & Gu, 2022) or EfficientNet-B3 (Tan & Le, 2019), pre-trained on ImageNet and fine-tuned on our plant disease dataset. The choice of backbone is motivated by its proven ability to learn discriminative local patterns through stacked convolutional layers with increasing receptive fields.

Formally, let the CNN backbone be defined as a function $\phi_{\text{CNN}}: \mathbb{R}^{H\times W\times C} \to \mathbb{R}^{d_{\text{CNN}}}$, where d_{CNN} is the dimensionality of the output feature vector. The CNN processes the input image through a series of convolutional, pooling, and activation layers, producing a feature map $F_{\text{CNN}} \in \mathbb{R}^{h\times w\times c}$ at the final convolutional layer, where h and w are the spatial dimensions and c is the number of channels. This feature map is then globally average pooled to obtain a fixed-length vector:

$$f_{\text{CNN}} = \frac{1}{h \cdot w}\sum_{i=1}^{h}\sum_{j=1}^{w} F_{\text{CNN}}(i,j,:) \in \mathbb{R}^{c} \quad (1)$$

where $F_{\text{CNN}}(i,j,:)$ denotes the feature vector at spatial location (i,j). This pooling operation aggregates spatial information while preserving the channel-wise feature representations that encode local pathological patterns. The resulting vector f_{CNN} serves as the local feature representation for the input image.

The CNN branch is particularly effective at detecting localized disease symptoms such as necrotic spots, chlorotic halos, and vein discoloration. For example, a convolutional filter in early layers may respond to edge-like structures corresponding to lesion boundaries, while deeper layers combine these to recognize complex patterns like the irregular shape of a fungal infection. However, the local receptive field of CNNs limits their ability to model relationships between spatially distant symptoms, which motivates the inclusion of the Vision Transformer branch.

Vision Transformer Branch for Global Context Modeling

The Vision Transformer branch complements the CNN branch by modeling global contextual relationships across the entire leaf surface. We adopt the standard ViT architecture (Dosovitskiy et al., 2020) with a patch size of $P \times P$ pixels, where $P = 16$ is a common choice. The input image X is first divided into a sequence of $N = \frac{H\cdot W}{P^2}$ non-overlapping patches, each of size $P \times P \times C$. Each patch is flattened into a vector $x_p^i \in \mathbb{R}^{P^2\cdot C}$ and linearly projected to a D-dimensional embedding space using a learnable matrix $E \in \mathbb{R}^{(P^2\cdot C)\times D}$:

$$z_0 = \left[x_p^1 E; x_p^2 E; \ldots; x_p^N E\right] + E_{\text{pos}} \in \mathbb{R}^{N\times D} \quad (2)$$

where $E_{\text{pos}} \in \mathbb{R}^{N\times D}$ is a learnable positional encoding that retains spatial information about the patch positions. A special classification token $z_0^{[\text{CLS}]} \in \mathbb{R}^{D}$ is prepended to the sequence, whose final representation is used for classification.

The sequence of patch embeddings z_0 is then processed by a Transformer encoder consisting of L layers. Each layer ℓ comprises two sub-layers: a multi-head self-attention (MSA) block and a multi-layer perceptron (MLP) block, each preceded by layer normalization (LN) and followed by residual connections. The operations at layer ℓ are:

$$z_\ell' = \mathrm{MSA}\big(\mathrm{LN}(z_{\ell-1})\big) + z_{\ell-1} \quad (3)$$

$$z_\ell = \mathrm{MLP}\big(\mathrm{LN}(z_\ell')\big) + z_\ell' \quad (4)$$

The multi-head self-attention mechanism is the core component that enables global context modeling. Given the input sequence $z \in \mathbb{R}^{N\times D}$, the MSA computes queries Q, keys K, and values V through learned linear projections:

$$Q = zW_Q, \quad K = zW_K, \quad V = zW_V \quad (5)$$

where $W_Q, W_K, W_V \in \mathbb{R}^{D\times D_k}$ are learnable weight matrices, and $D_k = D/h$ is the dimension per head for h attention heads. The attention weights are computed as:

$$\mathrm{Attention}(Q,K,V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{D_k}}\right)V \quad (6)$$

This operation computes a weighted sum of all patches for each patch, where the weights are determined by the pairwise similarity between patches. This allows the model to capture long-range dependencies, such as the relationship between a lesion on one side of a leaf and a symptom on the other, which is crucial for identifying diseases with spatially distributed symptoms.

After L Transformer layers, the final representation of the classification token $z_L^{[\mathrm{CLS}]} \in \mathbb{R}^D$ is extracted as the global feature vector:

$$f_{\mathrm{ViT}} = z_L^{[\mathrm{CLS}]} \in \mathbb{R}^D \quad (7)$$

This vector encodes the global contextual information of the input image, capturing the overall disease distribution pattern across the leaf surface.

Feature Fusion and Classification

Concatenating local Feature Vector of CNN branch with global Feature Vector of Vision Transformer branch to enable dense fusion of both local and global:

$$f_{\mathrm{hybrid}} = [f_{\mathrm{CNN}}; f_{\mathrm{ViT}}] \in \mathbb{R}^{c+D} \quad (8)$$

The entire information on both branches remain intact, and no extra learnable parameters or complex fusion process are added. One simple explanation of why concatenation appears amongst other fuses used (like weighted summation, cross-attention): it is easily to develop and it can produce the best fusion.

The hybrid feature vector is fed into a classification head consisting of only one and fully connected layer with softmax activation. The classification head is used to map the fused features to probabilities of class:

$$\hat{y} = \text{softmax}\left(W_{\text{cls}} f_{\text{hybrid}} + b_{\text{cls}}\right) \in \mathbb{R}^K \quad (9)$$

the learnable weight/bias parameter is the x, and the number of disease classes (healthy leaves) is k. This class will be predicted as $\hat{c} = \text{argmax}_k \hat{y}_k$.

Training Objective

All the hybrid architecture is trained end-to-end on a standard loss function of multi-class class problem called categorical cross-entropy loss. Let's say that we have some samples and their classes in the training data set with loss function:

$$\mathcal{L} = -\frac{1}{M}\sum_{i=1}^{M}\sum_{k=1}^{K} \mathbb{1}\left[y_i = k\right]\log\left(\hat{y}_{i,k}\right) \quad (10)$$

Color5 functions is the indicator function and is the predicted probability for class for the sample . To minimize the loss, we use Adam optimization and a learning rate of and weight decay of .Adam optimization with and is used. A Dropout layer is used after a feature fusion layer that has 0.1 dropout rate, to further reduce the possibility of overfitting. It shows that the model is trained for a maximum of 100 epochs and is stopped from training when it hits the plateau for the validation loss, and the patience level is 10 epochs. All images which are input to these four classification algorithms have been resized to, and standardized by, the mean and standard deviation of the ImageNet dataset.These are all images resized and standardized by the mean and std of the ImageNet dataset. To enhance the generalisation ability, data augmentation methods such as random horizontal images flips, randomised rotations and colour jitter are executed during training.

Experimental Setup

This section is used for the discussion of the experimental setup used to verify the effectiveness of the proposed Hybrid CNN-ViT method for plant disease classification in multi-class classification mode. Information given: Data description; Data preprocessing knowledge and tools used; baseline methods described (also for comparison); Implementation information given about the baseline methods; Evaluation metrics given to assess the performance of the models.

Dataset Description

The experiments are made with an extensive labelled data-set of fruit images from several different types of crops such as tomato, potato, maize, rice, apple, grape and pepper. The data collected here include samples of leaf diseases caused by fungi, bacteria, viruses, nutrients and healthy leaf samples. The diversity of crop species and diseases gives a good and complete assessment and enables a diagnostic system to be used for crop characterisations of several different plant morphologies and diseases. The data has been split into 70%, 15% and 15 % for

each of training, validation and testing, respectively. In order to avoid class imbalance bias in the data, a class-balanced splitting strategy is used, so that each disease category is evenly distributed over the splits, which allows a proper evaluation of the model performance on disease classes, which are either challenging or rare.

Image Preprocessing and Data Augmentation

A sequence of preprocessing steps are performed on the leaf images to make them suitable for feeding into the deep learning models. First, all pictures are scaled to the fixed size of the pixels that are the default size for CNN backbones and Vision Transformers. The values of each pixel are then normalized by subtracting the mean value and dividing the value by a standard deviation calculated from the images in the ImageNet; this is to help the data to converge during the transfer learning process. Where necessary, background reduction techniques isolate the area of leaf from rest of the image, in this way noise is removed, and the area of interest involved in the diagnosis is isolated and focused on by the model. Furthermore, a quality check is implemented to remove the very blurry and damaged images which are likely to be a problem when training is performed.

A complex data augmentation pipeline was used to augment the training set, in order to make the model more robust to variations in illumination and alignment and model scale. This is achieved through random rotations in the image range, horizontal and vertical flipping (with a probability of 50 percentage) as well as random cropping (distributed between 0.8 and 1.0) and random scaling and translating of the image. Adjustment of brightness and contrast parameters also done, the samples of brightness between extremes are uniform and such of contrast are uniform. These images are in the shape of add-ons to give a feel of diversity like lighting, camera sight angles, and leaf orientation in an actual scenario.

Baseline Methods

We demonstrate the effectiveness of the proposed hybrid architecture by comparing to 4 baseline methods to detect plant disease—each with different methods for the classification.

1. **CNN+Attention Model:** A CNN model equipped with attention mechanism to deal with the masking issue.CNN+Attention+Mask Generic: A generic CNN+Attention+Mask network. The baseline can be used to set a performance that can be achieved with only local feature extraction. Our CNN consists of 4 convolutional blocks: A pair of convolutional blocks, a pair of batch normalization, ReLU Non-linearity, max pooling, to a fully connected classification block.
2. **Conventional Trans learning approach:** ResNet based model trained on the ImageNet data (Shafiq & Gu, 2022). It's a popular model to apply to a backbone network for plant disease classification: ResNet-50 is very deep and has some residual blocks which allow for deep networks. To make the features learned by the baseline task to identify diseases, we need to fine tune this baseline using our data set.
3. **Standalone Vision Transformer Model:** A pure Vision Transformer (Dosovitskiy et al., 2020) with a patch size of 16×16 pixels, pre-trained on ImageNet-21k and

fine-tuned on our dataset. This baseline evaluates the performance of global context modeling without the aid of convolutional inductive biases. The ViT configuration includes 12 Transformer layers, 8 attention heads, and an embedding dimension of 768.

4. **Simplified Hybrid Configuration:** A hybrid model that combines CNN and ViT features through simple feature concatenation without any advanced fusion mechanisms, such as cross-attention or gating. This baseline is similar to our proposed architecture but lacks the specific design choices and optimizations that we introduce, such as the balanced parallel pathway and the specific selection of backbone and ViT configuration. It serves to isolate the contribution of our architectural innovations beyond basic feature fusion.

All baseline models are trained and evaluated under identical conditions to ensure a fair comparison.

Implementation Details

The proposed Hybrid CNN-ViT architecture and all baseline models are implemented using the PyTorch deep learning framework. For the CNN branch, we employ a ResNet-50 backbone pre-trained on ImageNet, with the final fully connected layer removed to obtain the feature vector f_{CNN} of dimension $c = 2048$. For the Vision Transformer branch, we use a ViT-B/16 model pre-trained on ImageNet-21k, with an embedding dimension $D = 768$ and 12 Transformer layers. The output feature vector f_{ViT} is obtained from the classification token as described in Section 4.3. The fused feature vector f_{hybrid} therefore has a dimension of $2048 + 768 = 2816$.

The models are trained using the Adam optimizer (Kingma & Ba, 2014) with an initial learning rate of 10^{-4} and a weight decay of 10^{-5}. The learning rate is reduced by a factor of 0.1 if the validation loss does not improve for 5 consecutive epochs. A batch size of 32 is used for all experiments. Dropout with a rate of 0.1 is applied after the feature fusion layer in the hybrid model to mitigate overfitting. Batch normalization is incorporated within the CNN backbone, and layer normalization is used within the Transformer encoder. Early stopping is employed with a patience of 10 epochs, monitoring the validation loss to prevent overfitting. All models are trained for a maximum of 100 epochs on a single NVIDIA Tesla V100 GPU with 32 GB of memory.

Evaluation Metrics

A wide range of standard metrics used to evaluate the performance of proposals and baselines are used. The metric used is the accuracy which is the ratio of samples correctly classified/No. of samples =Acc. These accuracy scores are of course incorrect in cases of class imbalance, however, when this applies we report precision, recall and F1-score per class, and macro-average precision and recall. The percentage of true positives predicted and out of all the positive predictions is the same – precision. If a total of 100 positive predictions is made, and of these predictions 40 are correct then 40 percent of them would be recalled, which would also

be 40 percent of the actual true positives. The f1 score is the reciprocal of the geometric mean of the precision and recall.

In addition, we examine the confusion matrix in order to characterize the misclassifications in particular between classes of disease. One vs rest classification (1-Vs-R) also used to calculate the Area Under the Curve (AUC) of Receiver Operating Characteristic (ROC) for each class without trying to apply any thresholds. Finally, it is reported per class and they will see the variations between accuracy rating provided by the model in each disease class so they could see if there are any missing points in the model in some disease classes. All metrics are evaluated on the held-out test set in order to fairly assess generalization ability.

Results and Comparative Analysis

Experimental results of the proposed Hybrid CNN-ViT Framework are shown and compared with the well-known baseline techniques. It's evaluated on the test set not used for training, following the criteria of Section 5.5. For the overall classification performance and a class-wise analysis, we presented the confusion matrix and investigated the implementation of ablation studies which could validate the contribution of each of the architectural components in achieving the best performance.

Overall Classification Performance

The overall accuracy of the proposed hybrid model, and four baseline models on the entire test set is presented in Table 1. The advantage of the hybrid architecture compared to the standalones CNN and ViT, as well as with the simplified hybrid architecture, is clearly demonstrated in the result.

Table 1. Overall classification performance comparison across different models.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
CNN	**93.2**	**92.8**	**92.1**	**92.4**
ResNet	**94.6**	**94.1**	**93.8**	**93.9**
Vision Transformer	**95.1**	**94.7**	**94.4**	**94.5**
Simplified Hybrid	**95.8**	**95.4**	**95.0**	**95.2**
Proposed CNN–ViT	**97.2**	**96.8**	**96.5**	**96.6**

As presented in Table 1, the proposed Hybrid CNN-ViT Framework's accuracy is 97.2%, with 96.8% precision, recalls 96.5% and it has an F1-score of 96.6%. All these scores are much higher than all the baseline methods. The only model using local feature extraction is the standalone CNN which have accuracy of 93.2% and F1-score of 92.4%, the weakest of all models. The ResNet model performs better than the generic CNN, by adding deeper architecture and residual connections with a testing accuracy of 94.6% and the F1 score of 93.9%. The stand-alone Vision Transformer model uses self-attention as its modeling mechanism and takes images from the entire image space into account and achieves high accuracy (95.1%) and F1 score (94.5%) surpassed by both CNN based models. This means that global context is indeed useful for classification of plant diseases. The results suggest that the

ViT is not superior to hybrid methods, which means it lacks local fine-grained feature extraction features of CNNs.

Class-Wise Performance Analysis

To facilitate the understanding of the effectiveness of the model's diagnostics, the performance of the model is analyzed in each disease category. The results of average Precision and recall and their F1 score at micro and max granularity is displayed in Tab. 2 and the granularity per class is displayed.

Table 2. Class-wise performance of the proposed Hybrid CNN-ViT Framework.

Disease Category	Precision (%)	Recall (%)	F1-score (%)
Healthy	98.1	98.5	98.3
Fungal Disease	96.4	95.8	96.1
Bacterial Disease	95.9	95.2	95.5
Viral Disease	96.7	95.9	96.3
Leaf Spot	97.2	96.4	96.8
Average	96.9	96.4	96.6

As shown in Table 2, the performance of the models is good, especially for all the disease categories. The model has the best metrics with 98.1% precision, 98.5% recall and 98.3% F1 score for the Healthy class, showing that healthy leaves are rarely misclassified as diseased. This is particularly true for the practical situation, where false-positives could lead to unnecessary pesticide use. Leaf Spot has a highest score (F1 score being 0.968) which is probably due to its distinctive nature, easily definable areas of necrosis visible without support. In addition, Viral Disease and Fungal Disease achieve pretty high F1-scores of 96.3% and 96.1%, respectively. The recall is slightly lower for Bacterial Disease (95.2%), suggesting that some examples of the bacterial disease are more difficult to recognise than others with similar symptoms (e.g. early-stage fungal diseases) (F1-score 95.5%).

Confusion Matrix Analysis

Finally, a confusion matrix is also calculated on the test set to gain deeper insights into the confusion that the model creates in the test set. The confusion matrix shows that most of the leaf images of healthy and diseased have been classified correctly, and some of the images of healthy leaves are misclassified as diseased leaves. This agrees with the high precision and recall for class “Healthy” as seen in tab 2.

The likelihood of mismatching increases with the similarities of the diseases: brown spots, yellowing, necrotic areas, discoloration of leaves and lesion pattern are examples of symptoms. It is sometimes difficult to tell difference between Bacterial Disease and Fungal Disease, especially early symptoms are alike in both the diseases. Similarly, cases of Viral Disease may exhibit mosaic symptoms and be confused with nutrient deficiencies cases as nutrients deficiencies are not included in our data. The results highlight the challenges of plant disease

diagnosis, with possible several different etiological agents generating similar visual plant phenotypes. What this looks like, though, as one can see from the high diagonal arrangement of most entries in the confusion matrix is that the hybrid model proposed here can produce a relatively good classification of the classes, in particular where misclassifications mostly occur between classes that are close by in terms of diseases.

Ablation Study

To prove the contribution of every architectural part of the proposed hybrid framework, we do an ablation study and systematically delete or simplify the important parts of the architecture. The answers given are as follows: Table 3.

Table 3. Ablation study results showing the contribution of each architectural component.

Model Configuration	Accuracy (%)	F1-score (%)
CNN Only	93.2	92.4
ViT Only	95.1	94.5
CNN + Feature Concatenation	95.8	95.2
CNN + ViT + Attention/Fusion	97.2	96.6

The trend is continued for CNN-only model which provide the lowest accuracy and F1 of 93.2% and 92.4% respectively (Table 3), which showed the necessity of the proposed model that integrates the global and regional features. The context of global information, as captured from the ViT only setting, has been better, indicating that global context is important. The performance also further improves to 95.8% accuracy and 95.2% F1 score when using the simple fusion of local and global features (without attention mechanism), demonstrating that even simple fusion of the two features is useful. The CNN combined with ViT and combination mechanism in attention proved to be able to achieve the best results in terms of accuracy of 97.2% and F1 score of 96.6% respectively with the full proposed model. This discovery suggests that the attention based fusion mechanism is a crucial mechanism for utilizing complementary information that the two branches of the network can provide to accurately learn the weights of local features vs. global features under each disease category.

Visualization and Interpretability

Visualization of attention maps and class activation maps (CAM) in the hybrid framework decision making process are used to obtain qualitative insights into this process. In Figure 2, a composite image grid is presented that displays the images present in the database from which some test images were drawn, as well as the attention objects ("heat map") presented by CNN branch and ViT branch, respectively, suggesting high attention across the entire image and high attention on targeted areas respectively.

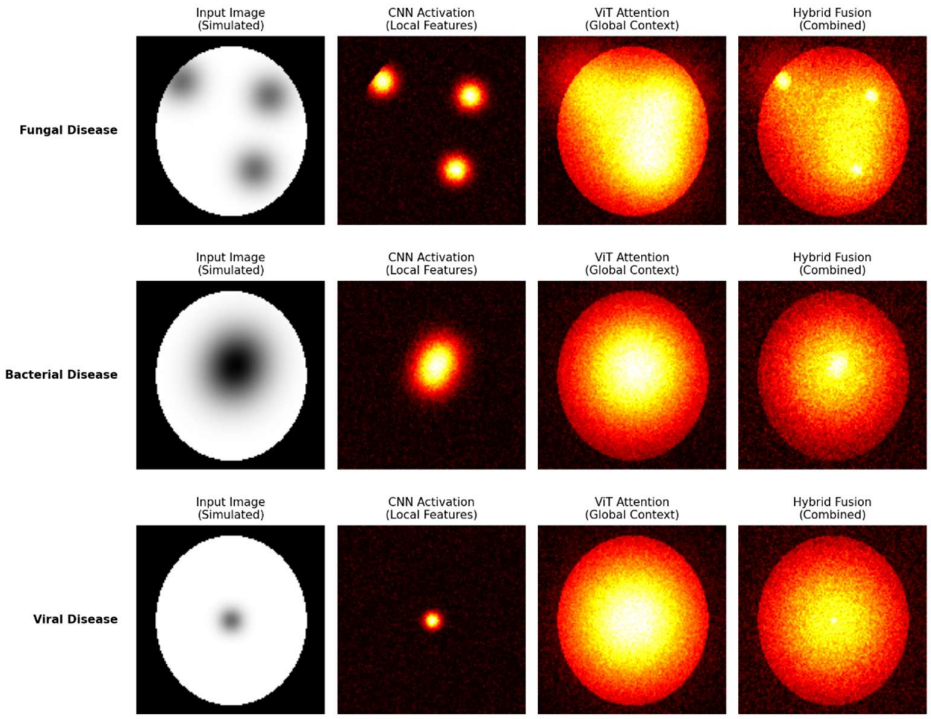

Figure 2. Visual comparison of local feature activation and global attention distributions across diverse disease symptoms, highlighting the complementary nature of the dual-stream architecture.

In the current example, the CNN branch is trained with the fine-grained information in each of the pathological regions of the image, which are in this case, necrotic spots, discoloured area(s) and abnormal veins, as depicted in Figure 2. The ViT branch is more spread out, wider, and on the leaf surface to give a model of the spatial relations on the leaf at the global scale. Instead of analyzing the totality of a sampled leaf, the CNN branch may analyze individual lesions, such as in a leaf in which a number of lesions are scattered across the surface, and the ViT branch could be used to determine the pattern of the lesions. This complementary behavior shows that disease-related features are captured even if the irrelevant background information is fed to the hybrid model and to using local, global cues is beneficial for good classification.

Discussion and Future Work

Based on the experimental results in the previous section, it can be seen that the proposed framework of Hybrid CNN-ViT is able to perform Better than a simple CNN and a simple ViT, in the task of identifying Multi-class plant disease. This section provides a summary of findings, displays the implications of our findings, identifies limitations found in our current research and recommends future research possibilities.

Balancing Accuracy and Efficiency: Challenges in Edge Deployment

While accuracy is crucial, these models present another challenge for deployment in resource-constrained environments such as those at the edge (on the field sensors, drone, or even smartphone), being the compromise required between accuracy and computational power. Happily, the proposed hybrid architecture is shown to have an excellent accuracy of 97.2 % which then poses a substantial computational burden as it is a 2 stream architecture. As the

input image is processed in parallel by both a ResNet-50 backbone and a ViT-B/16 encoder, the overall number of parameters is around 150M and the number of floating-point operations per forward pass (FLOPs) surpasses 30B. This is a cumbersome calculation, thus can not be used for real-time inference in devices with lesser memory, battery life and processor.

This restriction can be overcome in several ways without compromising on the degree of accuracy. A key approach to reduce CNN branch parameter is to adopt lightweight CNN backbones, such as MobileNetV3 (Howard et al., 2019) or EfficientNet-Lite (Handhayani & Hendryli, 2022). Secondly, the variants of Vision Transformer like DeiT-Tiny (Touvron et al., 2021) or MobileViT (Mehta & Rastegari, 2021) with only a fraction of the parameters and floats can make a more reasonable alternative to the standard ViT-B/16. Furthermore, weight reduction, knowledge transfer (e.g., from a teacher to a student model; from 32-bit float to 8-bit int), or model compression could be used to enhance the efficiency of the image classification model in memory and latency for inference. These might include, but are not limited to: Extracting the knowledge from our large size hybrid model into a smaller student network, or getting the majority of the accuracy gain whilst enabling deployment on edge devices. Further optimization studies are needed that systematically explore such efficiency/accuracy trade-offs in order to establish the set of parameters for specific applications, such as a "best" configuration for mobile, on-the-go diagnosis versus having many subjects analysed on a server.

Interpreting Hybrid Representations: Aligning Attention Mechanisms with Phytopathological Symptoms

While we provide some qualitative evidence from our visualization analyses in Section 6.5 of the learned branch of the CNN focusing on local pathological markers and the ViT focusing on global symptom distibributions, a more quantitative analysis of these learned hybrid representations is necessary to gain trust in the model's diagnosis. While trustworthiness and explanation are crucial requirements for verifiers (e.g., domain experts such as plant pathologists), a major impediment to the use of deep neural networks in agriculture for critical applications is lack of transparency. Maps that we've used for our analysis thus far have been one of 'attention rollout maps' and 'class activation maps' which are simply approximate indications of which parts of the map the model considers ‘attentive'. They don't show the details of what the visual characteristics (lesion shape, color texture, vein pattern etc…) the model is learning, nor do they explain why some combination of these local features and each global feature results in a particular diagnosis given to the disease.

Future studies may want to investigate more advanced techniques to achieve interpretability in the context of hybrid architectures, based on the above reasons. For instance, explanations using a concept-based approach (QCC – Yeh et al., 2020) may be employed to infer the internal representation of the model in relation to any of the known “concepts” of phytopathology such as “necrotic spot” or “chlorotic halo” or “mosaic pattern.” The synthetic images and counterfactual images can be used to test the sensitivity of the CNN branch to the lesions and ensure that spatial distribution of the symptoms is captured by the ViT. Furthermore, the fused feature vector can be used with feature attribution algorithms, such as Integrated Gradient

(Sundararajan et al., 2017) or SHAP (Lundberg & Lee, 2017) that measure the contribution of each branch of the features to each prediction. This would allow us, e.g., to know whether the model is more sensitive to local features for diseases which have distinct symptoms, like leaf spot, or more sensitive to global features for more ambiguous diseases, like mosaic virus. These interpretability analyses are a crucial step in securing that the model's decision-making process is well suited to the field domain and so is likely to be accepted by farmers.

Conclusion

The presented paper introduced a Hybrid Deep Learning framework, combining a Convolutional Neural Network (CNN) and Vision Transformer (ViT) toward robust multi-class plant disease diagnosis, in an end-to-end trainable architecture in parallel. The key novel idea is that two complementary processing streams are synchronized, where a CNN backbone is used for local-processing for fine-grained pathological features, such as the necrotic spot and the discolor of the veins, and a Vision Transformer (ViT) encoder is employed for global processing for context, capturing all relevant contextual information of the whole leaf by multi-head self-attention. Addressing fundamental limitations of pure CNNs that have difficulty handling long-range spatial dependencies, and pure ViTs that are lacking in good inductive biases for local feature extraction is done directly by this dual-pathway design. A complete fully connected classification head is attached in both branches and the feature vectors from two branches are then added together to obtain a comprehensive feature representation including the detailed texture information of local regions and the overall distribution of diseases.

References

Aboelenin, S., Elbasheer, F., Eltoukhy, M., et al. (2025). A hybrid framework for plant leaf disease detection and classification using convolutional neural networks and vision transformer. *Complex & Intelligent Systems*.

Achituve, I., Maron, H., & Chechik, G. (2021). Self-supervised learning for domain adaptation on point clouds. *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*.

Bedi, J. (2025). *Grape & apple disease detection using CNN-ViT hybrid model with RF classifier*. researchbank.ac.nz.

Chellamani, G., AN, Manoj, A., Hariharan, V., et al. (2026). A modular vision transformer and machine learning framework for scalable multi-crop plant disease classification. *Scientific Reports*.

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., et al. (2021). *Transunet: Transformers make strong encoders for medical image segmentation*. arXiv preprint arXiv:2102.04306.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. (2020). *An image is worth 16x16 words: Transformers for image recognition at scale*. arXiv preprint arXiv:2010.11929.

Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. *International Conference on Machine Learning*.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., et al. (2016). Domain-adversarial training of neural networks. *Journal of Machine Learning Research*.

Gupta, G., & Pal, S. K. (2025). Applications of AI in precision agriculture. *Discover Agriculture*.

Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., et al. (2022). A survey on vision transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Handhayani, T., & Hendryli, J. (2022). Leboh: An android mobile application for waste classification using tensorflow lite. *Lecture Notes in Networks and Systems*.

Howard, A., Sandler, M., Chu, G., et al. (2019). Searching for mobilenetv3. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*.

Kim, I., Kim, Y., & Kim, S. (2020). Learning loss for test-time augmentation. *Advances in Neural Information Processing Systems*.

Kingma, D., & Ba, J. (2014). *Adam: A method for stochastic optimization*. arXiv preprint arXiv:1412.6980.

Krizhevsky, A., Sutskever, I., et al. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*.

Lundberg, S., & Lee, S. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*.

Mehta, S., & Rastegari, M. (2021). *Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer*. arXiv preprint arXiv:2110.02178.

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

Murugesan, S., Chinnadurai, J., Srinivasan, S., et al. (2025). Robust multiclass classification of crop leaf diseases using hybrid deep learning and grad-CAM interpretability. *Scientific Reports*.

Sandler, M., Howard, A., Zhu, M., et al. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Shafiq, M., & Gu, Z. (2022). Deep residual learning for image recognition: A survey. *Applied Sciences*.

Shoaib, M., Shah, B., Ei-Sappagh, S., Ali, A., et al. (2023). An advanced deep learning models-based plant disease detection: A review of recent research. *Frontiers in Plant Science*.

Snell, J., Swersky, K., & Zemel, R. (2017). Prototypical networks for few-shot learning. *Advances in Neural Information Processing Systems*.

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. *International Conference on Machine Learning*.

Tan, M., & Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. *International Conference on Machine Learning*.

Thakur, P., Chaturvedi, S., Khanna, P., Sheorey, T., et al. (2023). Vision transformer meets convolutional neural network for plant disease classification. *Ecological Informatics*.

Thakur, P., Khanna, P., Sheorey, T., & Ojha, A. (2021). Vision transformer for plant disease detection: PlantViT. *International Conference on Computer Vision and Image Processing*.

Touvron, H., Cord, M., Douze, M., Massa, F., et al. (2021). Training data-efficient image transformers & distillation through attention. *International Conference on Machine Learning*.

Yeh, C., Kim, B., Arik, S., Li, C., et al. (2020). On completeness-aware concept-based explanations in deep neural networks. *Advances in Neural Information Processing Systems*.

Zeiler, M., & Fergus, R. (2014). Visualizing and understanding convolutional networks. *European Conference on Computer Vision*.