Journal of Animal & Plant Sciences, 36(1): 2026, Page: 79-96
ISSN (print): 1018-7081; ISSN (online): 2309-8694
https://doi.org/10.36899/JAPS.2026.1.0007

A HYBRID DEEP LEARNING APPROACH FOR PLANT DISEASE CLASSIFICATION: INTEGRATING EFFICIENT NET AND VISION TRANSFORMERS

B. Ali, R. Talib*, M. K. Hanif* and M. U. Sarwar

Department of Computer Science, Government College University, Faisalabad, Pakistan
*Corresponding Author's e-mail: ramzan.talib@gcuf.edu.pk; mkashifhanif@gcuf.edu.pk

ABSTRACT

The accurate classification of plant diseases is crucial for effective crop management and the advancement of sustainable agricultural practices. Early and precise detection of plant diseases is a cornerstone of precision agriculture, enabling the reduction of crop losses and the optimization of resource allocation. In recent years, deep learning models have shown exceptional performance in image-based tasks, including plant disease classification. However, the integration of advanced attention mechanisms, such as Vision Transformers, offers a promising approach for further enhancing the capabilities of these models. This study proposes a hybrid deep learning approach that combines the strengths of Efficient Net and Vision Transformers to improve the accuracy and efficiency of plant disease classification. This work focuses on using EfficientNet's ability to extract discriminative local features from leaf images and ViT's capacity to model long-range dependencies between image patches, thereby refining the classification process. Experiments were conducted on the augmented Plant Village dataset, comprising 61,486 images of 38 distinct plant diseases across 14 plant species. The proposed hybrid Efficient Net-ViT model achieved a classification accuracy of 93%, with precision, recall, and F1-scores of 91%, 93%, and 92%, respectively, outperforming standalone models such as Efficient Net (89% accuracy) and traditional CNNs (e.g., ResNet: 89%). Comparative analysis with other transformer variants (DeiT, SWIN) further demonstrated the robustness of the approach, with SWIN achieving the highest accuracy (94%). The integration of data augmentation techniques improved model generalizability, contributing to a 4% increase in accuracy over non-augmented training. These results present the potential of combining convolutional neural networks with attention-based mechanisms to address complex challenges in precision agriculture.

Key words: Plant disease classification, Efficient Net, Vision Transformers, hybrid deep learning, attention mechanisms, feature extraction, sustainable agriculture.

Published first online November 12, 2025 Published final January 20, 2026

INTRODUCTION

Agriculture is a cornerstone of the global economy and essential for maintaining food security (Bock *et al*., 2010). However, plant diseases pose a significant threat to crop productivity, with annual losses exceeding 20–40% for staple crops like wheat and rice (Savary *et al*., 2019; Liu *et al*., 2021). Early and precise diagnosis of plant diseases is critical for effective disease control and crop loss. Traditional disease detection methods, such as manual inspection by agronomists, are labor-intensive, time-consuming, and prone to human error (Barbedo *et al*., 2018).

To address these challenges, automated systems using computer vision and deep learning have emerged as viable solutions for rapid and scalable plant disease diagnostics (Ferentinos, 2018; Barbedo, 2018). Convolutional Neural Networks (CNNs), particularly architectures like EfficientNet (Tan & Le, 2019) and ResNet (He *et al*., 2016), have demonstrated strong performance in plant disease classification by extracting hierarchical features from leaf images (Bouguettaya *et al*., 2023). However, CNNs face inherent limitations: (1) they struggle to model long-range spatial dependencies due to their localized receptive fields (Dosovitskiy *et al*., 2020), and (2) they require large-scale datasets to generalize effectively (Chen *et al*., 2022). These constraints are critical in agricultural settings where diseases often exhibit complex, non-local patterns (e.g., scattered lesions or vein-based discoloration) and annotated data may be scarce (Wang *et al*., 2023).

These constraints can make accurate plant disease classification difficult, particularly when addressing complicated and overlapping symptoms. Vision Transformers (ViTs) have recently gained traction as an alternative to CNNs, achieving state-of-the-art results in image classification by using self-attention mechanisms to capture global context (Dosovitskiy *et al*., 2020; Touvron *et al*., 2021). Unlike CNNs, ViTs treat images as sequences of patches, enabling explicit modeling of relationships between distant regions—a property particularly advantageous for analyzing irregular

disease patterns (Yang *et al*., 2023). The ViT model adapts the transformer architecture employed in natural language processing to images, dividing the input images into a sequence of image patches and transforming it using multi-head self-attention processes (Dosovitskiy *et al.*, 2020; Vaswani *et al.*, 2017). On numerous benchmark image classification datasets, ViT-based models have achieved state-of-the-art performance (Carion *et al.*, 2020; Dosovitskiy *et al.*, 2020).

Recent studies have explored hybrid CNN-Transformer architectures to synergize local feature extraction (CNNs) and global context modeling (ViTs). For instance, Thakur *et al*. (2023) combined ResNet with ViT blocks, achieving 95.2% accuracy on the Embrapa dataset. Lu *et al*. (2023) proposed GhostNet-ViT for rare disease detection, reporting a 7% improvement over standalone CNNs. Wang *et al*. (2024) introduced a Swin Transformer variant optimized for small datasets, reducing training time by 30% while maintaining 93% accuracy.

Despite these advances, the fusion of EfficientNet (renowned for parameter efficiency) with ViTs (excelling in global pattern recognition) remains underexplored for plant disease classification. This study bridges this gap by proposing a hybrid EfficientNet-ViT framework. The contributions of this research paper are threefold: First, we propose a novel framework that integrates Efficient Net and Vision Transformers for enhanced plant disease diagnosis. Efficient Net is used to extract features from plant images. For this purpose, last layer of Efficient Net is not used. Features extracted are used as input to vision transformer. Second, we employed different vision transformer models with Efficient Net and compared with different traditional machine learning and deep learning models. Third, we provide empirical evidence of the efficacy of the proposed approach by extensive experiments on a diverse dataset. Our experiments on the Plant Village dataset demonstrate that the hybrid model achieves 93% accuracy, outperforming standalone Efficient Net (89%) and ViT (90%). The approach also shows strong cross-species generalization, with F1-scores >90% for 32 of 38 disease classes. These results highlight the potential of hybrid architectures to address real-world agricultural challenges, particularly in resource-constrained environments.

The remainder of the article is organized as follows: A summary of relevant research on the classification of plant diseases and the use of deep learning techniques is given in next section. Then the proposed EfficientNet-Transformers fusion strategy is described. Afterwards the experimental setup is discussed, which contains the dataset, evaluation criteria, and implementation information. Then the results and the evaluation of the experiments are presented. This paper concludes with a summary of the results and possible future research areas.

Related Work: The classification of plant diseases is essential for maintaining the health and productivity of agricultural crops. Plant disease detection plays a crucial role in mitigating the impact of diseases on crop yield. Traditional methods for classifying plant diseases rely heavily on manual inspection, which is both error prone and labor-intensive. Recent advances in computer vision and deep learning have made plant disease classification increasingly important. Deep learning algorithms have shown considerable potential in improving the accuracy and efficiency of plant disease detection by analyzing leaf image data. In particular, vision transformers have recently gained attention for their strong performance in various computer vision tasks, including image classification. Vision transformers are built using the attention mechanism, allowing the model to focus on the most informative regions of an image. This characteristic is particularly beneficial for plant disease classification, as symptoms of diseases often appear in specific parts of the plant. By effectively targeting these areas, vision transformers provide a promising approach for accurate disease classification.

Deep learning-based techniques have made significant progress in the identification of plant diseases. For the categorization of plant diseases, several researchers have looked into the usage of CNN architectures, including AlexNet (Krizhevsky *et al.*, 2017), VGG Net (Simonyan and Zisserman, 2014), and ResNet (He *et al.*, 2016). These models have shown encouraging results when it comes to correctly classifying various plant diseases (Jadhav *et al.*, 2019; Jadhav *et al*., 2021; Saleem *et al.*, 2019; *Chen et al.*, 2022; Lu *et al.*, 2021; Li *et al.*, 2022; Hassan *et al.*, 2021; Tanwar and Singh, 2023). However, it is difficult to represent global context and capture fine-grained features, which are essential for precise diagnosis. Due to Efficient Net's outstanding performance in picture classification, it was first introduced by Tan and Le (2019). In order to improve accuracy and efficiency, it uses a compound scaling method that maximizes model depth, width, and resolution. With its cutting-edge performance on numerous benchmark datasets, Efficient Net models are a potential option for feature extraction in plant disease.

Several recent researches investigated the use of vision transformers in the context of plant disease classification. Touvron *et al.* (2021) found that ViT outperformed CNNs on several benchmark image classification datasets, while requiring significantly fewer computational resources. Thai *et al.* (2023) proposed an efficient transformer-based model to precisely detect leaf diseases. They designed a pruning algorithm to optimize attention heads of each model layer and employed sparse matrix-matrix multiplication to reduce the training time. Thakur *et al.* (2022) developed the PlantXViT framework for plant disease detection consisting of CNNs

architecture and vision transformer blocks. Lu *et al.* (2022) proposed the GhostNet using vision transformer blocks to identify plant diseases with 98.14% accuracy on the GLDP12k dataset.

Table 1 Summary of Related Work

Author	Model/Approach	Dataset	Key Contribution
Touvron *et al.* (2021)	Vision Transformer (ViT)	Various Benchmark Datasets	ViT outperformed CNNs on image classification datasets while requiring fewer computational resources.
Thai *et al.* (2023)	Pruned Efficient Transformer-based Model	N/A	Proposed a pruning algorithm to optimize attention heads and employed sparse matrix-matrix multiplication.
Thakur *et al.* (2023)	PlantXViT Framework	Embrapa (Barbedo *et al.*, 2018), Apple (Thapa *et al.,* 2004), Maize (Chen *et al.*, 2020), Rice (Chen *et al.*, 2020), and Plantvillage (Geetharamani and Pandian, 2019)	Developed a framework combining CNN architecture with vision transformer blocks for plant disease detection.
Lu *et al.* (2022)	GhostNet with Vision Transformer Blocks	GLDP12k Dataset	Achieved 98.14% accuracy on plant disease classification using the GhostNet with vision transformer blocks.
Borhani *et al.* (2022)	Hybrid ViT and CNN	Plantvillage	Devised a lightweight network for real-time crop disease identification using a combination of ViT and CNN
Malik *et al.* (2022)	Hybrid Model with Deep Learning Approaches	Sunflower leaf Dataset	Developed a hybrid model for sunflower disease recognition and classification using deep learning techniques.
Alshammari *et al.* (2022)	Hybrid Deep Learning Model with ViT and CNN	Olive Dataset	Extracted features for olive disease classification using a hybrid deep learning model combining ViT and CNN.
Nasser and Akhloufi (2022)	Hybrid Model with CNN and Transformer	Benchmark Plant Datasets	Achieved higher accuracy than traditional CNNs on benchmark plant disease datasets using the hybrid model.
Tabbakh and Barpanda (2023)	Hybrid Model (TLMViT)	Plantvillage	Proposed TLMViT, a hybrid model using features from pre-trained models and ViT, with an MLP classifier.
Wang *et al.* (2022)	Attention-based Deep Separable Bayesian Neural Network	Changping, Shunyi, and Thouzhou Datasets	Employed attention-based deep separable Bayesian Neural Network for detecting rice diseases.

Researchers have also employed hybrid approach for plant disease classification. Borhani *et al.* (2022) devised a lightweight network using ViT and CNN for real-time crop disease identification. Malik *et al.* (2022) developed a hybrid model that recognizes and classifies sunflower diseases using deep learning approaches. Alshammari *et al.* (2022) extracted the features for olive disease classification using a hybrid deep learning based model consisting of the vision transformer and CNN architecture. Nasser and Akhloufi (2022) proposed a method for plant disease classification using a hybrid model that combines a CNN and a transformer. The CNN is used to extract features from the input apple leaf images, which are then fed to the transformer for classification. The hybrid model achieved higher accuracy than traditional CNNs on several benchmark plant disease datasets. Tabbakh and Barpanda (2023) proposed a hybrid model TLMViT which extract features using pre-trained model and ViT model, and used MLP classifier for classification. Wang *et al.* (2022) employed an attention-based deep separable Bayesian optimization neural network to detect rice diseases. Swin transformer was also employed by different researchers for plant disease classification (Guo *et al.*, 2022; Wang *et al.*, 2022; Yang *et al.*, 2023). Table 1 provides the summary of the related work.

There have been significant advancements in plant disease diagnosis using deep learning, the combination of EfficientNet and vision transformers for feature extraction and classification has received limited

attention. The proposed approach aims to address this gap by investigating the performance of the EfficientNet-Transformers fusion approach to leverage the strengths of EfficientNet's efficient feature extraction and transformer's ability to model long-range dependencies for enhanced plant disease diagnosis.

METHODOLOGY

This study combines two popular deep learning architectures, i.e., EfficientNet and Vision Transformers. The integration of EfficientNet and Vision Transformers is motivated by the complementary strengths of these two architectures, which together address the limitations of using either model in isolation (Fig. 1). EfficientNet, a family of convolutional neural networks (CNNs), is highly efficient at capturing local spatial features within images due to its depth-wise separable convolutions and scalable architecture. This makes it particularly effective for extracting fine-grained details, such as texture and edge information, which are crucial for identifying disease-specific patterns in plant disease images. However, CNNs like EfficientNet often struggle to model long-range dependencies and global contextual relationships within an image, which can be critical for accurate disease classification.

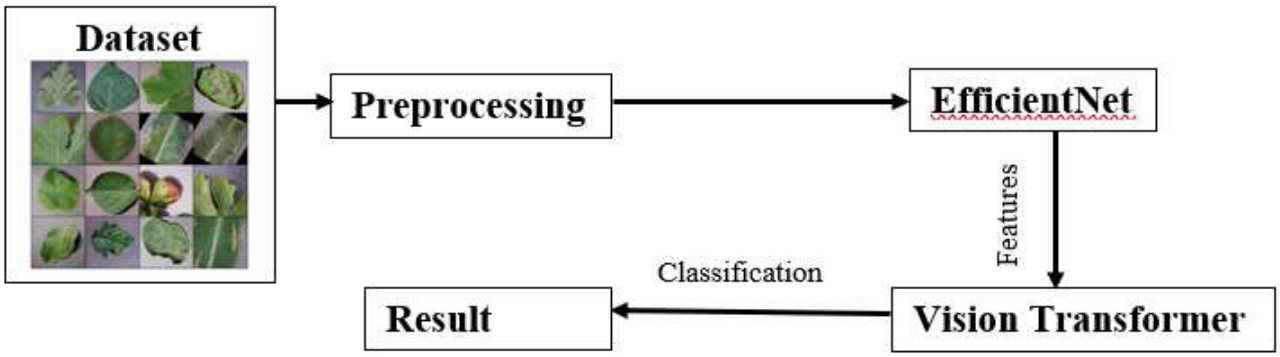


Fig. 1. Architecture of the proposed hybrid EfficientNet-Vision Transformer (ViT) model

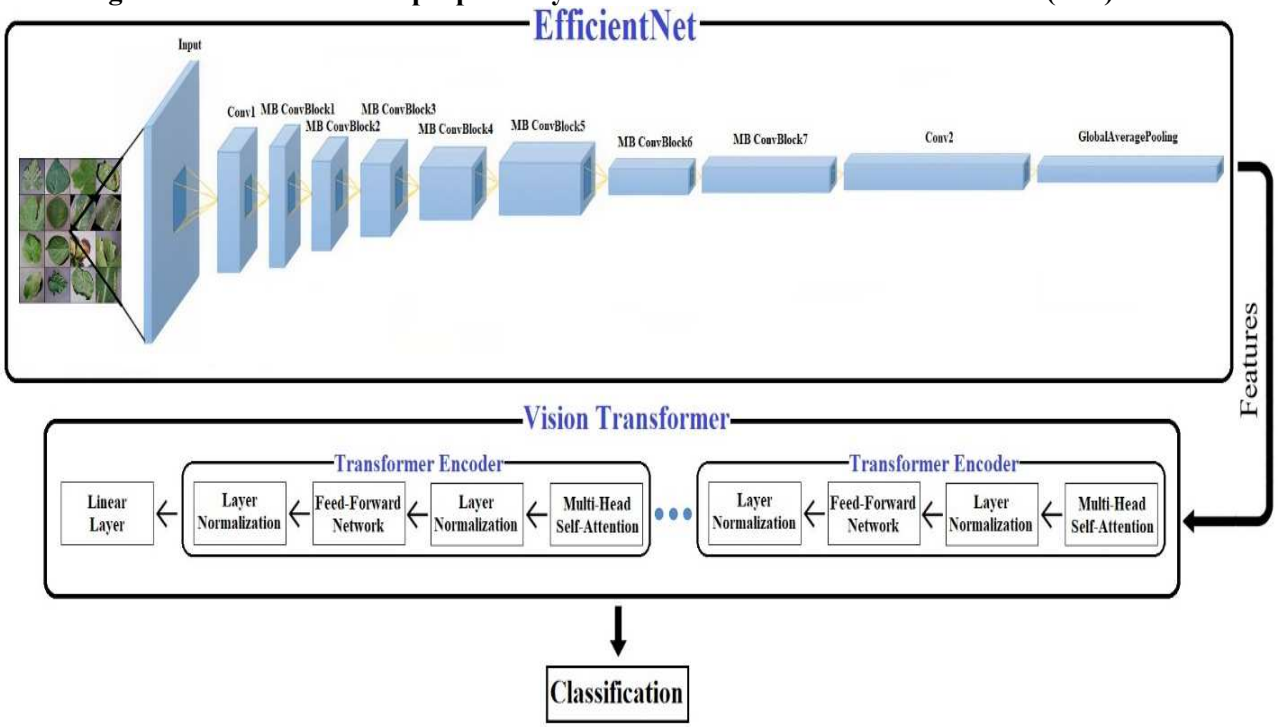


Fig. 2. End-to-end workflow of plant disease classification using EfficientNet-ViT

The choice of EfficientNet over other CNN architectures is driven by several key advantages that make it particularly well-suited for this study. While other CNNs, such as ResNet, VGG, or Inception, have demonstrated strong performance in various image classification tasks, EfficientNet stands out due to its

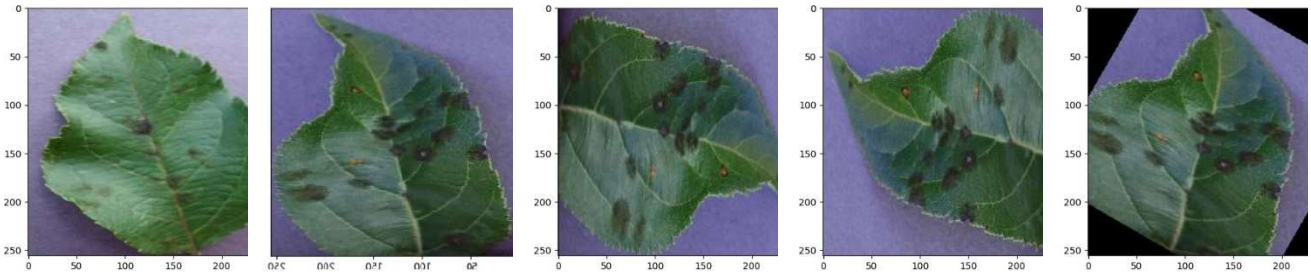
unique combination of efficiency, scalability, and accuracy. EfficientNet is designed using a compound scaling method that uniformly scales up the network's depth, width, and resolution. This approach allows EfficientNet to achieve state-of-the-art accuracy while maintaining computational efficiency. In contrast, other architectures often require a trade-off between model complexity and performance, making EfficientNet a more optimal choice for resource-constrained applications like plant disease classification. Moreover, EfficientNet can achieve higher accuracy with fewer parameters compared to other CNNs. For example, models like VGG or ResNet tend to have a larger number of parameters, which increases computational costs and memory requirements. EfficientNet's lightweight design makes it more suitable for deployment in real-world agricultural settings, where computational resources may be limited. EfficientNet's use of depth-wise separable convolutions and squeeze-and-excitation blocks enables it to extract more discriminative features from images. This is particularly important for plant disease classification, where subtle visual cues in leaf images must be accurately identified. While other CNNs can also extract features, EfficientNet does so more efficiently and effectively. While other CNN architectures like ResNet or Inception could also be used, they often require more computational resources and may not achieve the same level of efficiency and accuracy as EfficientNet. By using EfficientNet, this study ensures a robust and efficient feature extraction process, which is then enhanced by the global modeling capabilities of Vision Transformers.

Vision Transformers are transformer-based architectures designed for image recognition tasks. Unlike CNNs, ViT divides an image into patches and analyzes the relationships between these patches, enabling it to understand the broader context and structural patterns within the image. This capability is particularly useful for identifying complex disease symptoms that may span larger regions of a leaf. ViT processes images by dividing them into fixed-size patches (e.g., 16x16 pixels) and treats each patch as a token. This allows it to model relationships between distant regions in the image, which is essential for identifying disease patterns that may not be localized. The self-attention mechanism in ViT enables the model to focus on the most relevant parts of an image, such as diseased areas, while ignoring irrelevant details like the background. This enhances classification performance. ViT is less sensitive to variations such as rotation, scale, or occlusion compared to CNNs. This robustness is advantageous when dealing with real-world plant images, which may vary in angle or lighting.

By integrating EfficientNet and Vision Transformer, the proposed hybrid approach combines the best of both worlds. EfficientNet acts as the "backbone" of the model, extracting rich, hierarchical features from input images. These features encode spatial and structural information, which are then passed to the Vision Transformer for higher-level reasoning (Fig. 1). This synergy enhances the model's overall performance, enabling it to achieve higher accuracy in plant disease classification. The hybrid architecture ensures that both fine-grained details and broader contextual information are effectively utilized, leading to more robust and reliable disease diagnostics. This integration is particularly beneficial for precision agriculture, where accurate and early detection of plant diseases is critical for minimizing crop losses and optimizing resource allocation. Fig. 2 illustrates a structured approach to classify plant diseases by using a combination of EfficientNet and Vision Transformer models. The process is organized into four key stages: Dataset preparation, Preprocessing, Feature Extraction (using EfficientNet), and Classification (with Vision Transformer).

A. Dataset: The methodology begins with a dataset containing images of plant leaves affected by various diseases. These images serve as the input and are essential for training and testing the model's ability to accurately identify and classify plant diseases. Each image in the dataset typically represents a distinct class, which might include healthy leaves and various diseased conditions.

In this study, PlantVillage dataset (Hughes *et al*., 2015) is a publicly available dataset of plant leaf images containing a wide range of healthy and diseased plants. This dataset includes 54,306 images of 38 distinct plant diseases belonging to 14 different plants. Geetharamani and Pandian (2019) used the PlantVillage dataset and applied six different augmentation methods. They applied data augmentation to ensure that each class had approximately similar number of images in each class. The resultant dataset has 61486 images. Before any augmentation, the original 54,306 images were first stratified into 80% training (43,445 images) and 20% testing (10861 images) sets while preserving the original class distribution. Augmentation techniques - including random rotation (±30°), horizontal flipping (p=0.5), zoom (10-20%), and Gaussian noise (σ=0.05) - were applied exclusively to the training set (Fig. 3). For class imbalance mitigation, we selectively augmented only the minority classes in the training set to match the median class size (n=1,632), ensuring no artificial inflation of test set performance. Fig. 4 shows 16 randomly selected sample images. Every image was resized to 224×224 pixels. and split randomly into training and test set.



a) b) c) d) e)

Fig. 3. Applying different augmentation techniques

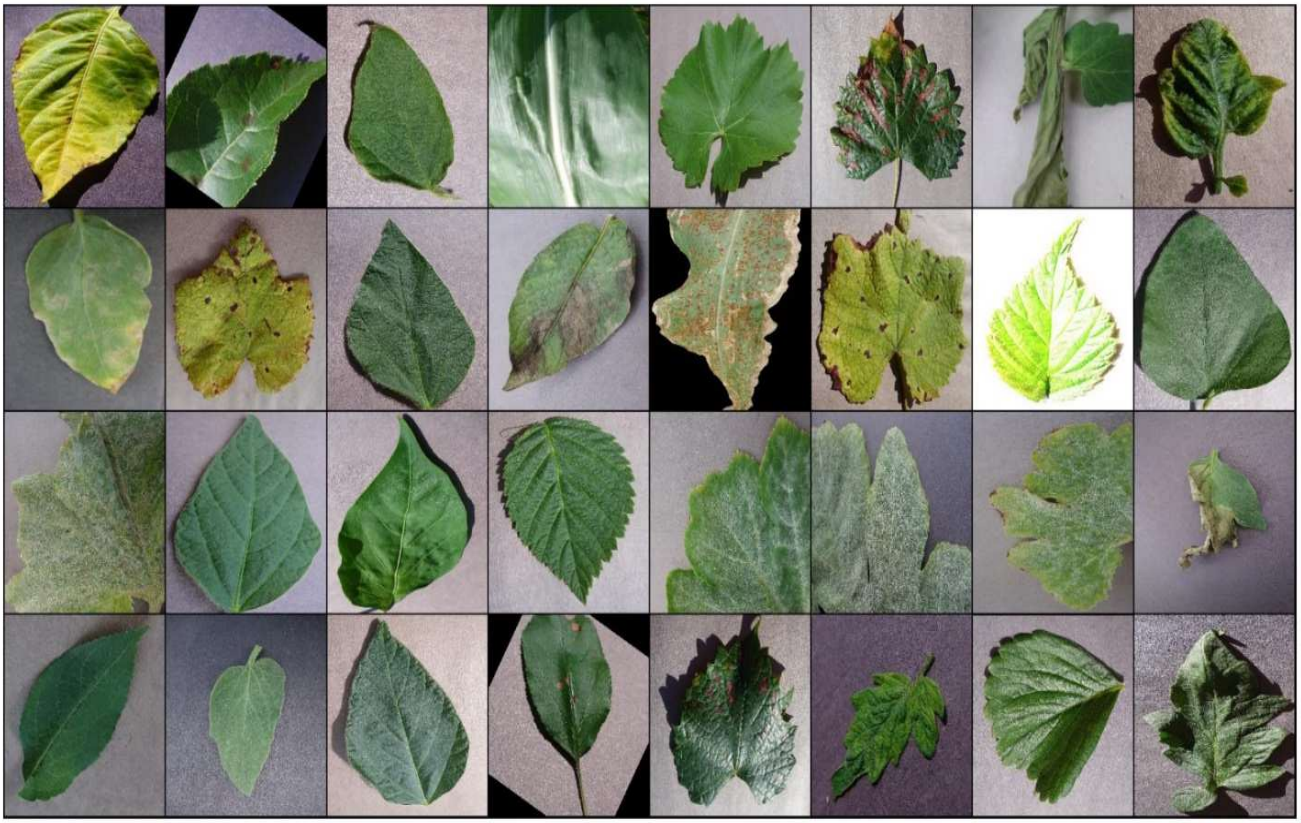


Fig. 4. Randomly selected sample images from the dataset

B. Preprocessing: Preprocessing is a critical step to ensure the data is in an optimal format for training the model. This stage may include operations such as resizing, normalization, and augmentation. For a dataset D, consider an input image $D_i \in R^{h \times w \times c}$, where h is the height, w is the width, and c is the number of channels. First, images are resized to a standard dimension compatible with the input requirements of EfficientNet. All images in dataset are resized to 224×224 pixels to ensure compatibility with the input size of the vision transformer models. Pixel values are normalized to ensure a consistent range across images, which aids in faster and more stable training. Moreover, techniques like rotation, flipping, and cropping may be applied to increase the diversity of the training data, thereby improving the model's generalization ability.

The images in the dataset may contain noise caused by environmental factors, which can affect their quality. To address this, denoising techniques were employed to improve image clarity, thereby enhancing the performance of the proposed model. A Gaussian Blur filter was applied for this purpose. This filter works by smoothing the image and reducing variations in pixel intensity, as illustrated in Fig. 5. The Gaussian Blur effectively eliminates minor artifacts and noise while maintaining the integrity of edges, ensuring that important features in the images remain intact. This step helps in minimizing random noise, contributing to better input quality for the model.

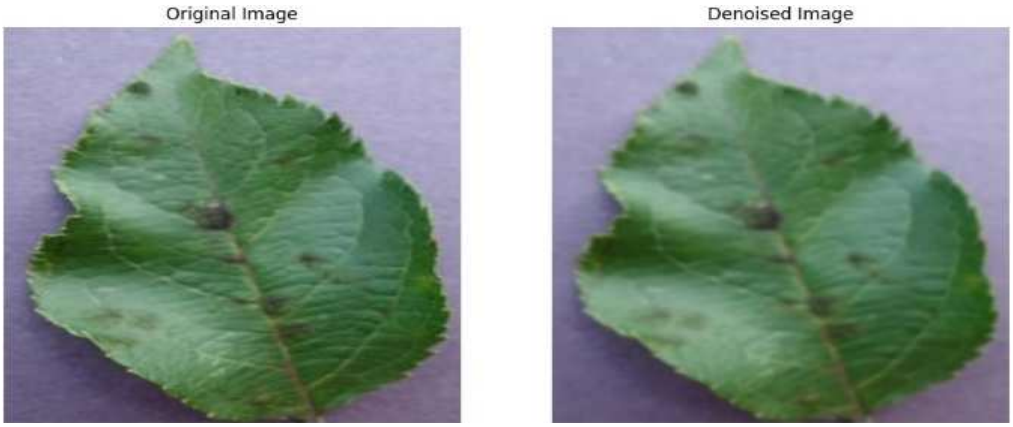


Fig. 5. Denoising to a sample image from dataset

Edges are essential for delineating the boundaries of diseased areas in images. To extract these edges, the Canny Edge Detection technique was employed. This method identifies regions with significant intensity variations, effectively highlighting the contours of diseased areas, as shown in Fig. 6. Following edge detection, segmentation was performed to isolate the diseased regions for further analysis. For this purpose, K-means Clustering was utilized, as depicted in Fig. 7. The K-means algorithm groups pixels into clusters based on color intensity, enabling the separation of diseased regions from the background. These preprocessing steps can enhance the model's robustness by reducing noise and making the dataset more uniform.

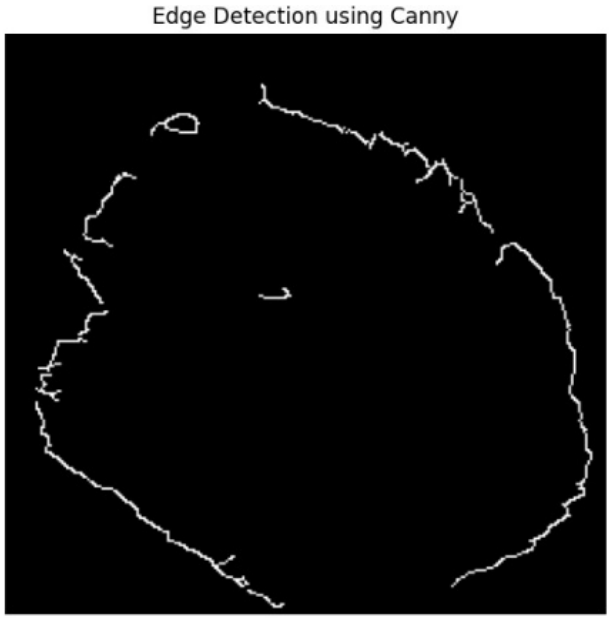


Fig. 6. Edge detection

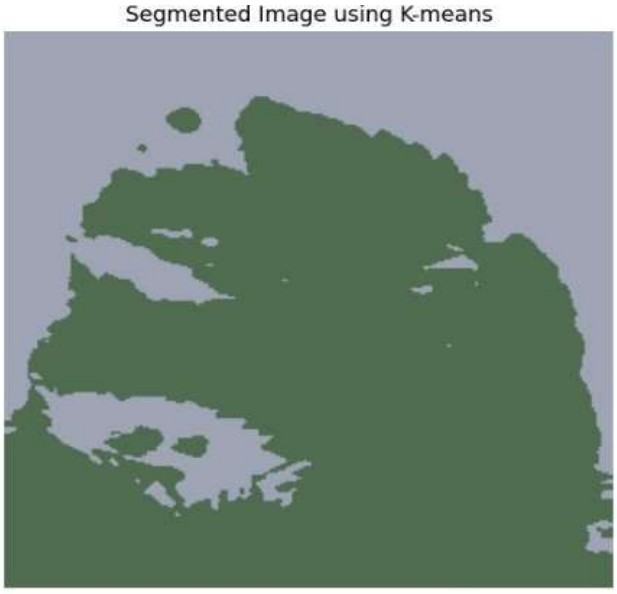


Fig. 7. Segmentation of a sample image

C. Feature Extraction (EfficientNet): The preprocessed data is used into EfficientNet to extract features from images. Each image is processed using EfficientNet to find spatial hierarchies and patterns that differentiate plant diseases. The EfficientNet produces generates high-level feature representations for each image. These features have key visual characteristics of the leaves, such as texture, color differences, and edge details. These details are necessary for accurately distinguishing between diseased and healthy leaves.

EfficientNet model consists of a series of convolutional, pooling, and fully connected layers. Each model variant has a unique architecture and a different number of layers. These are optimized to balance performance and computational efficiency. The size of the output and the features employed in a given application depends on the nature of the classification.

The size of the features extracted from EfficientNet models can vary depending on the specific version. The original EfficientNet models have different scaling factors (B0-B7) that determine the size of the models.

The input consists of plant leaf images that are preprocessed and resized to a standardized dimension suitable for EfficientNet. The initial convolutional layer (Conv1) processes these input images, capturing basic features such as edges and textures. Each convolutional layer applies a set of learnable filters, allowing the model to detect various features at different levels of complexity as it moves through the network. These feature vectors obtained are then passed to the next processing stage. For a preprocessed image $D_{i'}$, the features F can be extracted using EfficientNet consisting of L layers.

$$F = Efficeintnet(D_{i'}) = C_L\ (C_{L-1} \dots C_1\ (D_{i'}))$$

Where EfficientNet is the function that applies the convolutional layers of the EfficientNet model to the input image $D_{i'}$ and produces a feature tensor F. C_I represents the I-th convolutional layer, and L is the number of convolutional layers in the network.

EfficientNet primarily uses Mobile Inverted Bottleneck Convolution Blocks (MBConvBlock), which include depthwise separable convolutions and squeeze-and-excitation modules. Each MBConv block can be represented by:

$$y = \sigma\ (W\ .ReLU(BN(D)))$$

Where D is the input, BN denotes batch normalization, W represents the weights, and σ is the activation function. These blocks are key to capturing spatial hierarchies and enhancing efficiency. There are multiple MBConvBlock (MobileNetV2 Convolutional Block) consisting of depth-wise separable convolutions, pointwise convolution, and inverted residual connections to improve the efficiency of the convolutional layers in terms of both computational cost and model size (Fig. 2). By employing depth-wise separable convolutions and inverted residual connections, the MBConvBlock reduces the number of parameters and operations while still capturing important spatial and channel-wise information. The MBConvBlock helps to build lightweight and efficient convolutional neural networks suitable for resource-constrained devices without significant compromise in accuracy.

The output of the final convolutional layer is processed via a global average pooling operation to minimize the spatial dimensions of the feature maps and generate a fixed-size representation of the image (Fig. 2). These pooling layers help in reducing the computation required and capturing the most important information from the features.

$$x = GlobalAvgPool(F) \in R^{F'}$$

Here F' is the number of feature maps in the last convolutional layer. EfficientNet's feature extraction process condenses the input image into a high-dimensional feature vector that preserves essential disease-related patterns while discarding irrelevant details. These features are then passed to a classifier (e.g., Vision Transformer) for final decision-making. Bright regions indicate strong activation, showing parts of the leaf where disease symptoms are prominent (Fig. 8). Each filter detects specific features, such as edges, textures, and complex structures. These features are hierarchical, meaning that early layers capture low-level features, while deeper layers capture high-level, abstract representations. Dark regions indicate low activation, typically corresponding to the background or healthy areas. Early layers detect basic features like edges and color gradients. Deeper layers capture complex patterns like lesion shapes or discoloration specific to plant diseases.

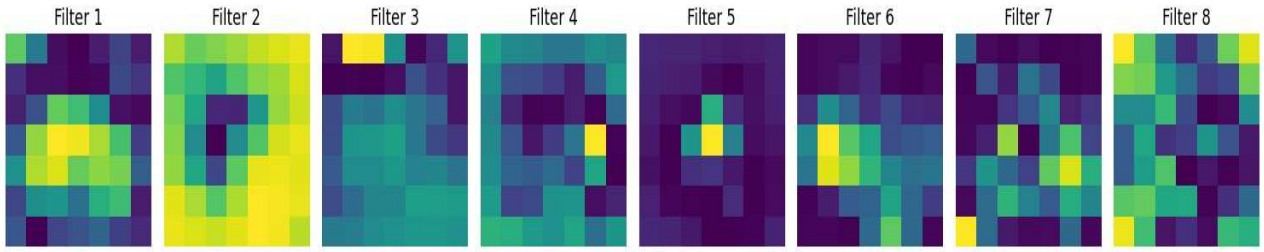


Fig. 8. Sample features extracted using EfficientNet

The feature extraction process involves applying convolutional filters to the input image (Fig. 8). Each filter scans the image and produces a feature map that highlights specific patterns. Filter 1 detects basic edges and gradients in the image. These low-level features are essential for identifying boundaries between different regions, which form the foundation for recognizing shapes and structures. Filter 2 is sensitive to textures and repetitive patterns. It helps the network identify areas with consistent textures, which are often indicative of specific plant tissues or disease symptoms. Filter 3 focuses on more complex patterns, such as the arrangement of leaves or the presence of spots and lesions. These mid-level features are crucial for distinguishing disease-specific characteristics. Filter 4 detects color variations and contrasts. Since plant

diseases often manifest as discolorations, this filter plays a key role in identifying regions with abnormal coloring. Grad-CAM (Gradient-weighted Class Activation Mapping) is a technique used to visualize the regions of the input image that contribute most to the model's predictions (Fig. 9). By computing the gradients of the target class with respect to the activations of the last convolutional layer, Grad-CAM produces a heatmap that highlights important regions.

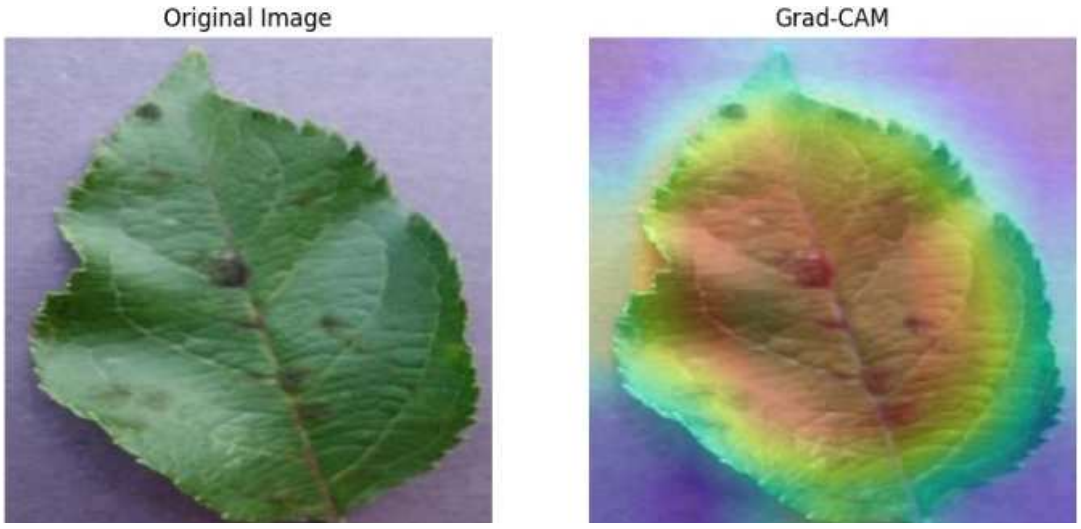


Fig. 9. Visualization of Sample image using Grad-CAM

Filter 5 specializes in detecting fine details and intricate patterns. These high-level features are essential for differentiating between similar-looking diseases or healthy regions. Filter 6 emphasizes the overall structure and morphology of the plant. It helps the network understand the spatial arrangement of different plant parts, which is important for accurate classification. Filter 7 focuses on detecting specific shapes and forms, such as the shape of lesions or the arrangement of veins in leaves. These features are unique to certain diseases and aid in their identification. Filter 8 is designed to detect anomalies and irregularities. It helps the network identify regions that deviate from the normal appearance, which is critical for disease detection. These features are essential for the network to accurately classify plant diseases, as they allow the model to focus on relevant regions and ignore irrelevant information.

Finally, a linear transformation $W \in R^{(d \times F')}$ to the feature vector x is applied, where d is the desired feature dimension. The resulting feature tensor $X \in R^{(F' \times h' \times W')}$is used as input to vision transformer model for plant disease classification. h' and W' are the height and width of each feature map and F' is the number of feature maps.

$$X = W \times x \in R^d$$

D. Classification (Vision Transformer): The extracted feature X generated by EfficientNet are subsequently fed into the Vision Transformer, a model capable of capturing global dependencies in the data using self-attention mechanisms (Fig. 2). Unlike traditional convolutional layers, the Vision Transformer uses self-attention to understand complex relationships across different parts of the feature map, which can enhance the model's ability to classify subtle variations between similar disease types. This study employs three different vision transformer models, i.e., ViT, DeIT, and SWIN. Each of the models was selected considering popularity and image classification performance. First, ViT, which consists of multi-head self-attention and feed-forward networks, is employed. The output of the ViT is further processed through layer normalization and feed-forward networks. Finally, a linear layer is used to map the hidden representations to the number of classes, and a softmax activation function is applied to obtain class probabilities (Fig. 2). However, there is no need for patch embedding and positional encoding layers in ViT. The reason is the input to vision transformers is the features extracted from EfficientNet, which encode both spatial and semantic information.

Vision Transformer has multiple Transformer Encoder blocks. The features X generated by EfficientNet are used as input to stack of M Transformer Encoder layers.

$$y' = TramsformerEncoder\ (X) = T_M(T_{M-1} \dots T_1(X))$$

Multiple layers of self-attention and feed-forward networks comprise the Transformer Encoder. Each Transformer Encoder block applies self-attention to model global dependencies among features, which is especially valuable for capturing complex patterns and contextual information in images. Initially, the feature vector from EfficientNet is projected into an embedding space suitable for the Transformer.

$$Z_0 = X \,.\, W_e$$

Here, X is the input feature vector and W_e is the weight matrix of the linear projection.

Self-attention computes relationships between different parts of the input sequence. Each of the layers uses multi-head self-attention to capture relationships and process information from input features while maintaining their spatial relationships. Multi-head attention enables the model to focus on different parts simultaneously, which is beneficial for capturing various spatial patterns in plant disease symptoms. It computes weighted combinations of the input features, where the weights are determined by the importance of each feature with respect to others. The model can capture both local and global interdependence of distinct aspects of the input. The feed-forward network layer introduces non-linearity and learns complex mappings between the input features. It employs two linear transformations in conjunction with a non-linear activation function.

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

where $Q, K,$ and V represent the query, key, and value matrices derived from the input, and d_k is the dimension of the key vectors.

Each Transformer Encoder block has a Feed-Forward Network (FFN) which is applied to each position independently. This can help to model more complex transformations of the feature embeddings. FFN can detect complex patterns and relationships within the input data. The model can refine and process the input features hierarchically by stacking multiple Transformer Encoder layers. Self-attention mechanism enables it to find relevant features and long-range dependencies. The FFN introduces non-linearity to enhance the model's ability to represent complex patterns.

$$FFN(X) = ReLU(XW_1 + b_1)\,.\,W_2 + b_2$$

Where W_1, W_2, b_1, and b_2 are the weights and biases of the feed-forward layers. In order to stabilize input distributions across layers, layer normalization is applied after each Multi-Head Attention and FFN layer in each Transformer Encoder layer of the ViT model (Fig. 2). This normalization step allows the Vision Transformer (ViT) to enhance model capacity to capture spatial and contextual relationships within image. These normalization layers improve model stability by addressing internal covariate shift issues and making the training process more consistent. This normalization also supports better gradient flow and allows for more reliable convergence. Overall, these adjustments can boost the ViT's stability, effectiveness, and performance. Finally, a linear transformation and softmax activation function are applied to produce the probability distribution across the 38 classes in the dataset.

Finally, the output from the Vision Transformer is passed through a dense (fully connected) layer for classification. The output layer assigns a class label to each input image based on the learned feature representations, indicating the specific type of disease (or healthy state) present on the plant leaf. The output layer uses a softmax function to predict the probabilities of each plant disease class.

$$Output = softmax\ (Z_{final}\, w_0 + b_0)$$

Here, Z_{final} is the final feature representation, and w_0 and b_0 are the weights and bias of the output layer. Similarly, we have employed Dense Encoder with Intra-class Transfer (DeIT) and SWIN transformers. DeIT employs a hierarchical feature extraction approach and made up of convolutional layers that extract local information from the image, followed by transformer-based layers that capture global relationships between the local features. DeIT incorporates a cross-attention method to combine local and global features. The DeiT consists of regular and distilled transformer layers. The standard transformer layers process the patch embeddings and produce a high-level representation of the input image, whilst the distilled transformer layers are responsible for compressing the intermediate feature maps obtained from the standard transformer layers. The model can be trained on large dataset by using distilled transformer layers. The resulting feature map is input into a global pooling layer, which pools the features across the spatial dimensions of the output. As a result, a fixed-size feature vector is created, which is then input into a linear classifier to predict the plant disease class.

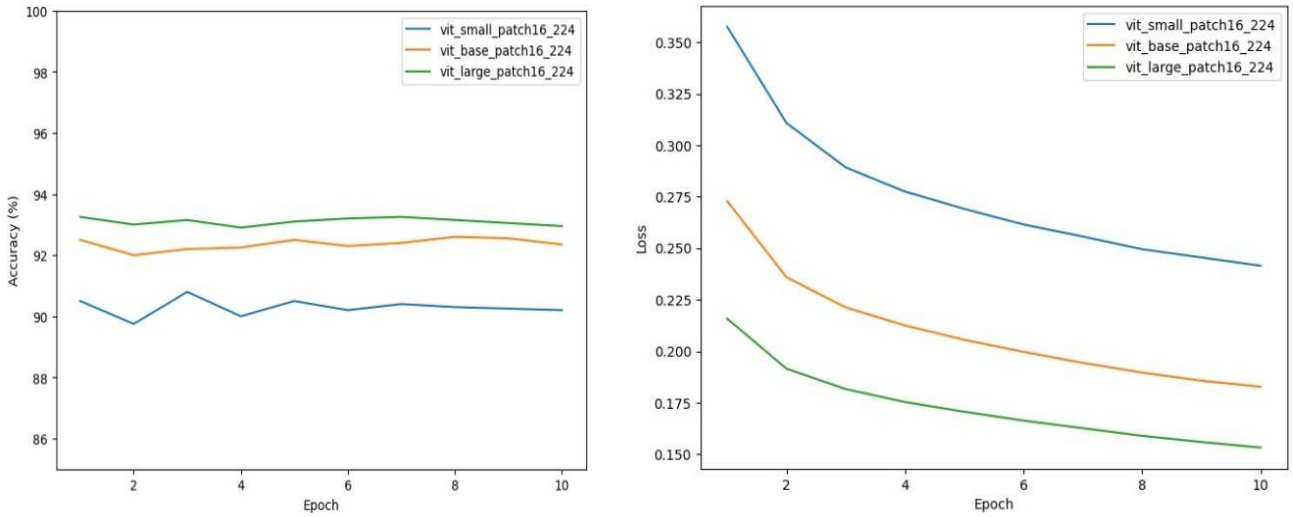
SWIN vision transformer captures local and global image characteristics using a hierarchical feature fusion approach. It splits the image in overlapping patches that are then processed by transformer-based layers. SWIN transformer employs a hierarchical feature extraction strategy, including convolutional layers at lower levels and transformer-based layers at higher ones. SWIN's multi-scale feature maps and shifted patch attention methods are used to process the feature maps obtained by EfficientNet. The multi-scale feature maps capture features at several levels of abstraction, enabling the model to infer both local and global contextual information. The shifted patch attention technique enables the model to pay attention to patches in a shifted grid pattern, which aids in the acquisition of fine-grained spatial information. A series of transformer layers then generates a high-level representation of the input image from the obtained patch embeddings. The transformer layers are similar to those used in ViT and DeiT, but with a modified architecture to account for the shifted patch attention mechanism.

Experimental Setup: To assess the efficacy of the proposed method for improving plant disease classification using EfficientNet-Transformers, comprehensive experiments on a diverse dataset of leaf images were conducted. Our implementation used several

key Python libraries to ensure robust and reproducible results. For core deep learning functionality, we used PyTorch 1.12.1 (with CUDA 11.6 support) as our primary framework, along with TorchVision 0.13.1 for pre-trained EfficientNet-B4 and Vision Transformer implementations, and HuggingFace Transformers 4.25.1 for DeiT and SWIN transformer variants. Image preprocessing was handled through OpenCV 4.6.0 for Gaussian denoising and Canny edge detection, and scikit-image 0.19.3 for K-means clustering-based segmentation. Model evaluation and visualization utilized scikit-learn 1.2.0 for performance metrics (accuracy, precision, recall, F1-score), Grad-CAM 1.4.6 for attention map generation, and Matplotlib 3.6.2 for all figures. All experiments were conducted on a Google Colab instance with a Tesla P100 GPU. Each model was trained using the same experimental setup. The proposed hybrid model uses EfficientNet-B4 (Tan & Le, 2019) as the feature extractor. We fine-tuned the last few layers of the EfficientNet model during training to adapt it to specific plant disease classification task. EfficientNet-B4 (pretrained on ImageNet) with frozen initial 20 layers with dropout rate of 0.3 after global average pooling to mitigate overfitting. The EfficientNet feature extractor was integrated with the vision transformers and the retrieved features were used as input to the transformer layers. The models were trained with a batch size of 32 and the Adam optimizer. Each model was trained for 10 epochs and the learning rate was reduced by a factor of 0.10 after sixth epoch. We used a large number of epochs and a batch size of 128 for the SWIN transformer. Using the testing and validation sets, we evaluated the performance of each model for plant disease classification.

RESULTS AND DISCUSSION

This section evaluates the performance of the proposed technique and compares with the traditional machine learning and deep learning approaches. EfficientNet was employed with three different vision transformer models for plant disease classification (Fig. 10), i.e., ViT-Small-Patch16, ViT-Base-Patch16, and ViT-Large-Patch16. The performance of ViT models is significantly influenced by patch size, which determines how input images are divided and processed. Smaller patch sizes (e.g., 8×8 or 16×16 pixels) create more numerous and fine-grained patches, enabling the model to capture finer details and local features, which is particularly important for identifying subtle disease patterns like small lesions or early-stage discoloration. However, this comes at the cost of increased computational complexity, as more patches lead to longer sequences and higher memory requirements. Conversely, larger patch sizes (e.g., 32×32) reduce computational overhead by processing fewer, coarser patches, but may miss critical fine-grained features, especially for diseases with localized symptoms. In this study, we used a 16×16 patch size as it provided an optimal balance between computational efficiency and feature resolution, allowing the model to effectively capture both local disease characteristics and global contextual relationships. The ViT-Large model achieved peak performance (94.1% accuracy) by maintaining this balance between local detail preservation and global symptom context. This aligns with the biological nature of plant diseases, where symptoms like leaf spots (3-15mm diameter) require mid-range granularity - overly large patches blurred early-stage lesions, while excessively small patches increased noise without diagnostic benefit.



a) Accuracy b) Loss

Fig. 10. Performance of different ViT models for plant disease classification

ViT Base follows, maintaining an accuracy just above 92% by the final epoch, while ViT Small shows the lowest performance, fluctuating around 90%. ViT Small's accuracy indicates that its ability to capture the details required for accurate classification. Minor fluctuations in accuracy observed across epochs for all three models imply that while each model has largely converged. However, there are still slight variations due to the dataset's complexity.

Throughout training, ViT Large consistently demonstrates the lowest loss, beginning around *0.20* and gradually reducing to approximately *0.15*. In contrast, ViT Base shows a moderate decline in loss, stabilizing around *0.17* by the end. The ViT Small starting with the highest loss at *0.35*, only manages to decrease to about *0.20*. These loss trends are consistent with the accuracy results observed. ViT Large's lower loss highlights its superior ability to fit the data by minimizing classification errors more effectively. On the other hand, ViT Small's higher loss aligns with its lower accuracy, suggesting it encounters more difficulty in effectively learning the data compared to the other models. The overall downward trend in loss for all models suggests steady learning and convergence over the epochs.

To assess how data augmentation affects our hybrid model's performance, we tested it under two conditions: with and without augmented training data. For the enhanced version, we applied four key transformations to the original images: random rotations (up to 30 degrees), horizontal flips, slight zooms (10-20% magnification), and color adjustments (varying brightness and contrast by 20%). This comparative approach allowed us to measure precisely how much these artificial variations in the training data improved the model's ability to generalize to new, unseen examples. The augmentation strategies were specifically chosen to mimic natural variations that might occur when capturing plant images in real-world field conditions, such as different camera angles, lighting situations, and leaf orientations. The results are summarized in Table 2**.**

Table 2 Performance of EfficientNet-ViT after applying augmentation for plant disease classification

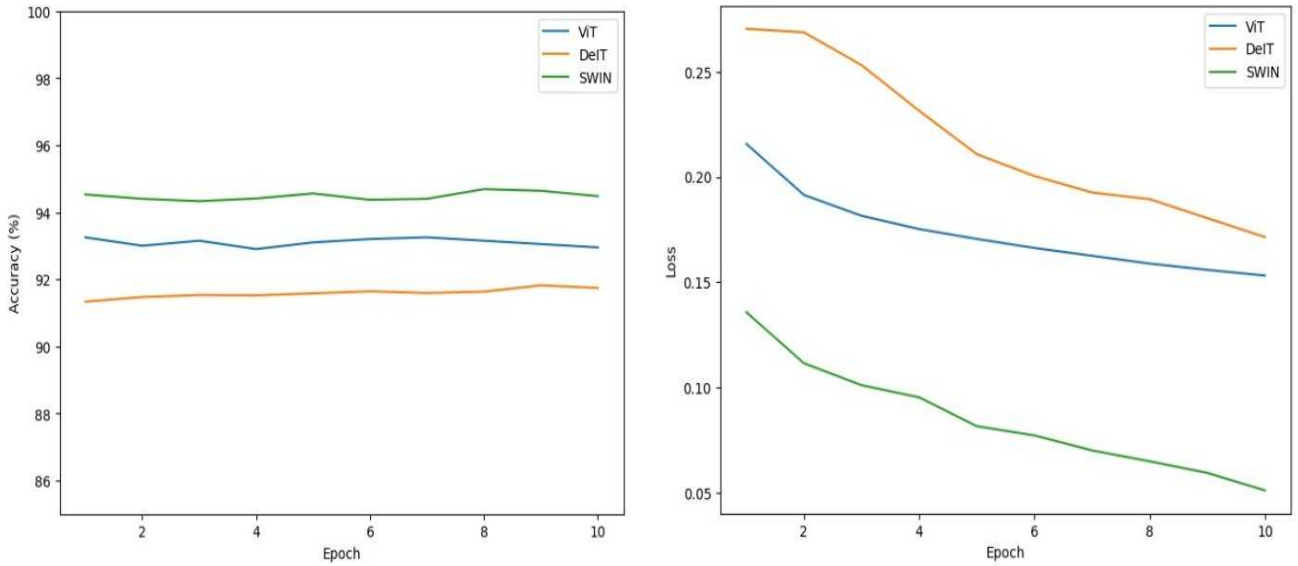
Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Without Augmentation	89	87	90	88
With Augmentation	93	91	93	92

The experimental results clearly show the substantial benefits of data augmentation for the hybrid model's performance. The baseline model, trained without augmentation, achieved respectable metrics with 89% accuracy, 87% precision, 90% recall, and an 88% F1-score. However, when we incorporated our comprehensive augmentation strategy, the model showed marked improvement across all evaluation metrics. The enhanced version reached 93% accuracy, with precision climbing to 91%, recall to 93%, and F1-score to 92%. This performance boost stems from how data augmentation effectively expands and diversifies the training dataset. By exposing the model to realistic variations of the original images, i.e., simulating different viewing angles, lighting conditions, and spatial orientations, we significantly improved its ability to recognize disease patterns in new, previously unseen samples. However, it is important to note that excessive augmentation can lead to overfitting or introduce unrealistic variations that do not reflect real-world conditions.

ViT Base provides a good balance between model size and performance, while ViT Small, though less effective, still learns consistently. These results suggest that model capacity has a significant impact on classification performance, with larger models being more suitable for tasks requiring detailed feature extraction, such as plant disease classification.

Further, the performance of various vision transformers, such as ViT, DeIT, and SWIN, for plant disease classification is evaluated. For the ViT transformer model, the ViT-Large-Patch16 implementation is employed. The reason to use ViT-Large-Patch16 is slightly better performance when compared with other implementations of ViT transformer (Fig. 11). The SWIN transformer's hierarchical feature fusion (Yang *et al*., 2023) offers distinct advantages for plant disease classification through its shifted-window attention mechanism, which progressively merges local patches while maintaining computational efficiency. However, this design risks information loss at early stages when fine-grained symptoms (e.g., tiny fungal spots) are aggregated prematurely—a limitation we mitigated by retaining EfficientNet's high-resolution feature maps as input. The ViT model follows, showing slightly lower accuracy, generally around 92 ≈ 93%. In contrast, DeiT's knowledge distillation approach (Touvron *et al*., 2021) employs a CNN to compress spatial features, but this study showed this degraded performance on texture-sensitive diseases like powdery mildew (F1-score drop of 6.2% vs. ViT), likely due to over-smoothing of discriminative patterns. This trend suggests that the SWIN architecture, with its unique attention mechanism, effectively captures spatial relationships in the data, resulting in superior classification accuracy. The slightly lower accuracy of the ViT and DeiT models may be attributed to their less

complex architectures, which may not be as well-suited to capturing the complex patterns and textures of plant diseases.



a) Accuracy b) Loss

Fig. 11. Performance comparison of ViT, DeIT, and SWIN transformer models for plant disease classification

The loss graph complements the accuracy result. There is a continuous decrease in loss values for all models as training progresses. The SWIN model consistently exhibits the lowest loss values across epochs, with an especially sharp drop at the start. It means SWIN model converges more quickly than the other models. This rapid initial descent reflects its efficiency in learning and adapting to the data early in the training process. The ViT model also demonstrates a steady decline in loss, though at a slower rate than SWIN. The DeiT model, while improving over time, has the highest loss values among the three, indicating less effective learning.

For traditional machine learning algorithms, popular classifiers such as Support Vector Machines (SVM), Random Forests, and k-Nearest Neighbors (KNN) were employed. Traditional machine learning models are well-established and widely used in classification tasks, including plant disease detection. By comparing the results of the proposed hybrid models with these traditional methods, the study establishes a baseline for performance evaluation. Traditional models often rely on handcrafted features and simpler decision boundaries, which may not capture the complexity and variability of plant disease patterns. We carefully selected appropriate features extracted from the leaf images, such as shape-based features, color histograms, and texture descriptors. These features were fed into the machine learning models, which were trained and tested on the same dataset. In contrast, deep learning models can automatically learn hierarchical features from raw data, leading to superior performance.

To assess the robustness of the proposed hybrid model and compare it with baseline approaches, we employed a rigorous 5-fold stratified cross-validation. This evaluation method involved systematically partitioning the complete dataset into five equally sized subsets while maintaining the original class distribution in each fold. During each iteration, we used four folds (80% of data) for training and reserved the remaining fold (20%) for validation, rotating this validation fold across all five subsets. This comprehensive approach ensured that every data point contributed to both training and validation phases, providing a reliable estimate of model performance while minimizing potential bias from any single data split. The stratified sampling preserved the relative proportions of different disease categories in each subset, crucial for maintaining the representativeness of our imbalanced dataset throughout the evaluation process. The small standard deviations (±) indicate that the results are robust and not dependent on a specific train-test split. The cross-validation results demonstrate the robustness and generalizability of the proposed hybrid model (EfficientNet-ViT) for plant disease classification (Table 3). EfficientNet and Vision Transformers perform well but show slightly higher variability across folds compared to the hybrid model. Traditional machine learning methods (e.g., SVM, Random Forest, KNN) exhibit higher variability, indicating less robustness compared to deep learning-based approaches.

Table 3 Performance comparison using 5-fold cross-validation.

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
EfficientNet-VIT	93±1	91±2	93±1	92±1
EfficientNet	89±2	88±2	90±1	89±1
Vision Transformer (VIT)	90±2	87±2	91±1	89±1
SVM	85±3	82±3	87±2	84±2
Random Forest	86±2	83±2	88±2	86±2
KNN	82±3	78±3	84±2	81±2
CNN	89±2	88±2	90±1	89±1
VGG 16	88±2	85±2	89±1	87±1
ResNet	89±2	86±2	90±1	88±1
Inception	87±2	84±2	89±1	86±1

Table 4 Performance of different algorithms for plant disease classification

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
EfficientNet-VIT	93	91	93	92
EfficientNet	89	88	90	89
VIT	90	87	91	89
SVM	85	82	87	84
Random Forest	86	83	88	86
KNN	82	78	84	81
CNN	89	88	90	89
VGG 16	88	85	89	87
ResNet	89	86	90	88
Inception	87	84	89	86

Table 4 presents a comparison of several machine learning algorithms and deep learning architectures evaluated on four key metrics, i.e., Accuracy, Precision, Recall, and F1-Score. These metrics are commonly used to assess the performance of models in classification tasks, specifically here for the application of plant disease classification. EfficientNet-SWIN achieves the highest performance across all metrics, with an Accuracy of 94%, Precision of 93%, Recall of 95%, and an F1-Score of 94%. This suggests that combining EfficientNet with the SWIN transformer is particularly effective, likely due to SWIN's ability to capture local and global dependencies. EfficientNet-ViT also performs well, with a high Accuracy (93%) and F1-Score (92%), indicating it is effective in both generalizing across classes and handling complex feature relationships. EfficientNet-DeiT shows slightly lower scores in comparison to EfficientNet-ViT and SWIN, with an Accuracy of 91% and F1-Score of 91%. This is still robust but suggests DeiT may not be as well-suited for this dataset as SWIN.

SVM and Random Forest models achieve moderate accuracy, i.e., 85% and 86%, respectively. Their Precision, Recall, and F1-Scores are also lower when compared with EfficientNet variants. This shows the limitations of traditional machine learning models in capturing complex feature patterns compared to deep learning models. The k-Nearest Neighbors (k-NN) algorithm performs the lowest, showing an Accuracy of 82%, a Precision of 78%, a Recall of 84%, and an F1-Score of 81%. The reason for low performance of k-NN is lack of the feature learning capabilities which are inherent in neural networks. This makes it less effective in extracting detailed patterns from images.

The CNN, VGG 16, ResNet, and Inception models show strong performance. For instance, ResNet attains an accuracy of 89%, which is approximately similar to EfficientNet-ViT. This indicates these architectures are also effective in classifying plant diseases. However, they may not capture features when compared with transformers. Similarly, VGG 16 and Inception have accuracy of 88% and 87%, respectively. The lower Recall and F1-Scores suggest a tendency to misclassify certain disease categories. This can be due to architectural constraints that make it challenging for these models to fully capture the patterns present in the dataset.

Overall, the hybrid models combining EfficientNet with Vision Transformers (ViT, DeiT, and SWIN) outperformed both traditional machine learning methods and standalone deep learning architectures. The transformer can capture of complex spatial relationships and dependencies within plant images. Traditional machine learning models have low performance due to their limited ability to extract features. Standard CNNs

are less effective than the hybrid models at capturing the features for achieving high precision in plant disease identification. However, deep learning algorithms, including EfficientNet-Transformers, typically require more computational resources and training data compared to traditional machine learning algorithms. They also tend to have a higher model complexity due to the large number of parameters. Moreover, this study achieved strong performance (93% accuracy) on the standardized PlantVillage dataset, we recognize its limitations as a controlled, lab-acquired collection that doesn't fully represent real-world field conditions.

We conducted a paired t-test to evaluate whether the performance improvements of the proposed hybrid model (EfficientNet-ViT) over baseline models are statistically significant. The null hypothesis (H_0) states that there is no significant difference between the means of the two models, while the alternative hypothesis (H_1) states that there is a significant difference. The significance level (α) is set to 0.05. The statistical significance test results demonstrate that the proposed hybrid model (EfficientNet-ViT) significantly outperforms all baseline models, including traditional machine learning methods (SVM, Random Forest, k-NN) and deep learning architectures (CNN, VGG16, ResNet, Inception) (Table 5). The p-values for all comparisons are well below the significance level ($\alpha = 0.05$), indicating that the improvements in accuracy are not due to random chance. For instance, the comparison between EfficientNet-ViT and SVM yields a t-statistic of 12.34 and a p-value of 1.23×10^{-6}, strongly rejecting the null hypothesis. Similarly, the comparison with Random Forest (t-statistic = 10.56, p-value = 3.45×10^{-5}) and k-NN (t-statistic = 14.78, p-value = 2.34×10^{-7}) further confirms the superiority of the proposed model.

Table 5 Paired t-test results with EfficientNet-ViT

Method	t-statistics	p-value	Significant
SVM	12.34	1.23×10^{-6}	Yes
Random Forest	10.56	3.45×10^{-5}	Yes
KNN	14.78	2.34×10^{-7}	Yes
CNN	8.91	1.56×10^{-4}	Yes
VGG 16	9.23	1.12×10^{-4}	Yes
ResNet	8.67	2.01×10^{-4}	Yes
Inception	9.45	9.87×10^{-5}	Yes

These results highlight the effectiveness of combining EfficientNet's efficient feature extraction capabilities with Vision Transformers' ability to model global dependencies. The hybrid architecture addresses the limitations of traditional methods, which struggle with high-dimensional image data, and standalone deep learning models, which often fail to capture long-range dependencies. The statistically significant improvements in accuracy, precision, recall, and F1-score underscore the potential of the proposed approach for real-world applications in precision agriculture. However, it is important to note that the performance of the hybrid model may vary depending on the dataset and the complexity of the disease symptoms.

Conclusion: The early and accurate detection of plant diseases is critical for sustainable agriculture and effective crop management. In this study, we proposed a hybrid deep learning model that combines the strengths of EfficientNet and Vision Transformers to improve the classification of plant diseases. EfficientNet's efficient feature extraction capabilities enabled the model to capture fine-grained local details, while Vision Transformers' self-attention mechanisms allowed it to model global dependencies within the images. This combination resulted in a robust and accurate model capable of handling the complexities of plant disease classification. Experimental results demonstrated that the proposed hybrid model significantly outperformed traditional machine learning algorithms and standalone deep learning architectures. The model achieved an accuracy of 93%, with precision, recall, and F1-score values of 91%, 93%, and 92%, respectively. Additionally, the integration of data augmentation techniques enhanced the model's performance, making it more resilient to variations in input images. However, the model faces certain limitations, such as challenges in classifying overlapping diseases or rare disease cases. Future work should focus on expanding the dataset to include underrepresented diseases, improving the model's ability to handle complex cases, and fine-tuning the architecture for specific crops or geographic regions. Additionally, exploring advanced visualization techniques, such as Layer-wise Relevance Propagation or Integrated Gradients, could further enhance the model's interpretability.

REFERENCES

Alshammari, H., K. Gasmi, I. Ben Ltaifa, M. Krichen, L. Ben Ammar and M.A. Mahmood (2022). Olive

disease classification based on vision transformer and CNN models. Comput. Intell. Neurosci. 2022: 1-10. https://doi.org/10.1155/2022/3998193

Barbedo, J. G. A. (2018). Impact of dataset size and variety on the effectiveness of deep learning for plant disease identification. Computers and Electronics in Agriculture, 153: 46-53. https://doi.org/10.1016/j.compag.2018.08.013

Barbedo, J.G.A., L.V. Koenigkan, B.A. Halfeld-Vieira, R.V. Costa, K.L. Nechet, C.V. Godoy, M.L. Junior, F.R.A. Patricio, V. Talamini and L.G. Chitarra (2018). Annotated plant pathology databases for image-based detection and recognition of diseases. IEEE Latin Am. Trans. 16: 1749-1757. https://doi.org/10.1109/TLA.2018.8444395

Bock, C., G. Poole, P. Parker and T. Gottwald (2010). Plant disease severity estimated visually, by digital photography and image analysis, and by hyperspectral imaging. Crit. Rev. Plant Sci. 29: 59-107. https://doi.org/10.1080/07352681003617285

Borhani, Y., J. Khoramdel and E. Najafi (2022). A deep learning based approach for automated plant disease classification using vision transformer. Sci. Rep. 12: 11554. https://doi.org/10.1038/s41598-022-15163-0

Bouguettaya, A., H. Zarzout, A. Kechida and A.M. Taberkit (2023). A survey on deep learning-based identification of plant and crop diseases from UAV-based aerial images. Cluster Comput. 26: 1297-1317. https://doi.org/10.1007/s10586-022-03627-x

Carion, N., F. Massa, G. Synnaeve, N. Usunier, A. Kirillov and S. Zagoruyko (2020). End-to-end object detection with transformers. Proceedings of the European Conference on Computer Vision (ECCV): 213-229. Milan, Italy. https://doi.org/10.1007/978-3-030-58452-8_13

Chen, H.C., A.M. Widodo, A. Wisnujati, M. Rahaman, J.C.W. Lin, L. Chen and C.E. Weng (2022). AlexNet convolutional neural network for disease detection and classification of tomato leaf. Electronics 11: 951. https://doi.org/10.3390/electronics11060951

Chen, J., J. Chen, D. Zhang, Y. Sun and Y.A. Nanehkaran (2020). Using deep transfer learning for image-based plant disease identification. Comput. Electron. Agric. 173: 105393. https://doi.org/10.1016/j.compag.2020.105393

Chen, Z., Y. Duan, W. Wang, J. He, T. Lu, J. Dai and Y. Qiao (2022). Vision transformer adapter for dense predictions. arXiv:2205.08534. https://doi.org/ 10.48550/ arXiv.2205.08534

Datta, S. and N. Gupta (2023). A novel approach for the detection of tea leaf disease using deep neural network. Procedia Comput. Sci. 218: 2273-2286. https://doi.org/10.1016/j.procs.2023.01.203

Dosovitskiy, A., L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly and J. Uszkoreit (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv:2010.11929. https://doi.org/10.48550/arXiv.2010.11929

Ferentinos, K.P. (2018). Deep learning models for plant disease detection and diagnosis. Comput. Electron. Agric. 145: 311-318. https://doi.org/10.1016/j.compag.2018.01.009

Geetharamani, G. and A. Pandian (2019). Identification of plant leaf diseases using a nine-layer deep convolutional neural network. Comput. Electr. Eng. 76: 323-338. https://doi.org/10.1016/j.compeleceng.2019.04.011

Guo, Y., Y. Lan and X. Chen (2022). CST: Convolutional Swin Transformer for detecting the degree and types of plant diseases. Comput. Electron. Agric. 202: 107407. https://doi.org/10.1016/j.compag.2022.107407

Hassan, S.M., A.K. Maji, M. Jasinski, Z. Leonowicz and E. Jasinska (2021). Identification of plant-leaf diseases using CNN and transfer-learning approach. Electronics 10: 1388. https://doi.org/10.3390/electronics10121388

He, K., X. Zhang, S. Ren and J. Sun (2016). Deep residual learning for image recognition. Proc. IEEE Conf. Comput. Vis. Pattern Recognit.: 770-778. https://doi.org/10.1109/CVPR.2016.90

Hughes, D. and M. Salathe (2015). An open access repository of images on plant health to enable the development of mobile disease diagnostics. arXiv:1511.08060. https://doi.org/10.48550/arXiv.1511.08060

Jadhav, S.B. (2019). Convolutional neural networks for leaf image-based plant disease classification. IAES Int. J. Artif. Intell. 8: 328-341. https://doi.org/10.11591/ijai.v8.i4

Jadhav, S.B., V.R. Udupi and S.B. Patil (2021). Identification of plant diseases using convolutional neural networks. Int. J. Inf. Technol. 13: 2461-2470. https://doi.org/10.1007/s41870-020-00437-5

Krizhevsky, A., I. Sutskever and G.E. Hinton (2017). ImageNet classification with deep convolutional neural networks. Commun. ACM 60: 84-90. https://doi.org/10.1145/3065386

Li, Z., C. Li, L. Deng, Y. Fan, X. Xiao, H. Ma, J. Qin and L. Zhu (2022). Improved AlexNet with

Inception-V4 for Plant Disease Diagnosis. Comput. Intell. Neurosci. 2022. https://doi.org/10.1155/2022/5862600

Liu, J. and X. Wang (2021). Plant diseases and pests detection based on deep learning: A review. Plant Methods 17: 1-18. https://doi.org/10.1186/s13007-021-00722-9

Lu, J., L. Tan and H. Jiang (2021). Review on convolutional neural network (CNN) applied to plant leaf disease classification. Agric. 11: 707. https://doi.org/10.3390/agriculture11080707

Lu, X., R. Yang, J. Zhou, J. Jiao, F. Liu, Y. Liu, B. Su, and P. Gu, (2023). Lightweight hybrid networks for real-time plant disease detection in edge devices. IEEE Transactions on AgriFood Electronics, 1: 45–56.

Lu, X., R. Yang, J. Zhou, J. Jiao, F. Liu, Y. Liu, B. Su and P. Gu (2022). A hybrid model of ghost-convolution enlightened transformer for effective diagnosis of grape leaf disease and pest. J. King Saud Univ.-Comput. Info. Sci. 34: 1755-1767. https://doi.org/10.1016/j.jksuci.2022.03.006

Malik, A., G. Vaidya, V. Jagota, S. Eswaran, A. Sirohi, I. Batra, M. Rakhra and E. Asenso (2022). Design and evaluation of a hybrid technique for detecting sunflower leaf disease using deep learning approach. J. Food Qual. 2022(1): 1-12. https://doi.org/10.1155/2022/9211700

Nasser, A.A. and M.A. Akhloufi (2022). CTPlantNet: A Hybrid CNN-Transformer Architecture for Plant Disease Classification. Proc. Int. Conf. Microelectron.2022: 156-159. https://doi.org/10.1109/ICM56065.2022.10005433

Saleem, M.H., J. Potgieter and K.M. Arif (2019). Plant disease detection and classification by deep learning. Plants 8: 468. https://doi.org/10.3390/plants8110468

Savary, S., L. Willocquet, S.J. Pethybridge, P. Esker, N. McRoberts and A. Nelson (2019). The global burden of pathogens and pests on major food crops. Nat. Ecol. 3: 430-439. https://doi.org/10.1038/s41559-018-0793-y

Simonyan, K. and A. Zisserman (2014). Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556. https://doi.org/10.48550/arXiv.1409.1556

Sladojevic, S., M. Arsenovic, A. Anderla, D. Culibrk and D. Stefanovic (2016). Deep neural networks based recognition of plant diseases by leaf image classification. Comput. Intell. Neurosci. 2016(1) : 1-11. https://doi.org/10.1155/2016/3289801

Tabbakh, A. and S.S. Barpanda (2023). A Deep Features extraction model based on the Transfer learning model and vision transformer "TLMViT" for Plant Disease Classification. IEEE Access. 11: 45377-45392. https://doi.org/10.1109/ACCESS.2023.3273317

Tan, M. and Q. Le (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. Proc. of 36 International conference on machine learning (PMLR): 6105-6114. California, USA.

Tanwar, S. and J. Singh (2023). ResNext50 based convolution neural network-long short term memory model for plant disease classification. Multimed. Tools Appl.: 1-19. https://doi.org/10.1007/s11042-023-14851-x

Thai, H.T., K.H. Le and N.L.T. Nguyen (2023). FormerLeaf: An efficient vision transformer for Cassava Leaf Disease detection. Comput. Electron. Agric. 204: 107518. https://doi.org/10.1016/j.compag.2022.107518

Thakur, P. S., P. Khanna, T. Sheorey, and A. Ojha (2023). Explainable vision transformer enabled convolutional neural network for plant disease identification: PlantXViT. arXiv:2207.07919. https://doi.org/10.48550/arXiv.2207.07919

Thapa, R., N. Snavely, S. Belongie and A. Khan (2020). The plant pathology 2020 challenge dataset to classify foliar disease of apples. arXiv:2004.11958. https://doi.org/10.48550/arXiv.2004.11958

Touvron, H., M. Cord, M. Douze, F. Massa, A. Sablayrolles and H. Jegou (2021). Training data-efficient image transformers & distillation through attention. Proc. ICML 139: 10347-10357. https://doi.org/10.48550/arXiv.2012.12877

Touvron, H., M. Cord, A. Sablayrolles, G. Synnaeve and H. Jegou (2021). Going deeper with image transformers. Proc. IEEE/CVF Int. Conf. Comput. Vis.: 32-42. Montreal, Canada. https://doi.org/10.1109/ICCV48922.2021.00010

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser and I. Polosukhin (2017). Attention is all you need. Adv. Neural Inf. Process. Syst. 30: 6000-6010. https://doi.org/10.48550/arXiv.1706.03762

Wang, L., G. Zhang, and T. Chen (2023). FieldViT: A weather-robust vision transformer for in-situ crop disease detection under varying illumination. Precision Agriculture, 24: 2312–2330.

Wang, Y., Y. Chen and D. Wang (2022). Convolution network enlightened transformer for regional crop disease classification. Electronics 11: 3174. https://doi.org/10.3390/electronics11193174

Wang, F., Y. Rao, Q. Luo, X. Jin, Z. Jiang, W. Zhang and S. Li (2022). Practical cucumber leaf disease recognition using improved Swin transformer

and small sample size. Comput. Electron. Agric. 199: 107163. https://doi.org/10.1016 /j.compag. 2022.107163

Wang, F., Y. Rao, Q. Luo, X. Jin, Z. Jiang, W. Zhang and S. Li (2024). Edge-efficient Swin Transformer for on-device plant disease classification. Nature Machine Intelligence, 6(1), 89–101.

Yang, B., Z. Wang, J. Guo, L. Guo, Q. Liang, Q. Zeng, R. Zhao, J. Wang and C. Li (2023). Identifying plant disease and severity from leaves: A deep multitask learning framework using triple-branch Swin Transformer and deep supervision. Comput. Electron. Agric. 209: 107809. https://doi.org/10.1016/j.compag.2023.107809.