Published by
NIGERIAN SOCIETY OF PHYSICAL SCIENCES
Available online @ https://journal.nsps.org.ng/index.php/jnsps

J. Nig. Soc. Phys. Sci. 7 (2025) 2940

Journal of the Nigerian Society of Physical Sciences

Efficient-ViT B0Net: A high-performance lightweight transformer for rice leaf disease recognition and classification

Santosh Kumar Upadhyay[a], Rajesh Prasad[a,b,*]

[a]*Department of CSE, Ajay Kumar Garg Engineering College, Ghaziabad, Uttar Pradesh, India*
[b]*Department of Computer Science, African University of Science and Technology, Abuja, Nigeria*

Abstract

Plant disease detection has become a demanding and challenging task in today's environment because many different types of plants exist worldwide, and very varied infections are found in them. The proposed work introduced a hybrid architecture to perform plant disease recognition and classification accurately and efficiently. The proposed model utilizes the strengths of CNN and Vision Transformer, where CNN successfully extracts local fine-grained texture features quickly. At the same time, ViT plays a vital role in extracting global and deep features from the leaf images. The suggested model was evaluated on a rice leaf dataset for paddy disease recognition and classification. The dataset consists of images representing four different types of rice leaves, with each class containing 4,000 samples. It includes healthy and diseased leaves, where the diseased category is further divided into three specific classes: Brown Spot, Bacterial Leaf Blight, and Leaf Smut. The suggested model worked well on the input dataset and achieved a testing accuracy of 99.13%. The Precision, recall, and F1 score of the proposed model were recorded as 99.13%, 99.13%, and 99.13%, respectively. The proposed method achieves a classification accuracy of 99.13%, outperforming SOTA models such as ViT-small, DenseNet121, ResNet50, EfficientNet B0 and SqueezeNet by 2–9% on the same dataset. The proposed method was compared with other approaches on the same experimental environment. These results demonstrate the effectiveness of our EfficientNet-ViT-based pipeline in capturing both local and global features for accurate rice disease classification.

DOI:10.46481/jnsps.2025.2940

Keywords: Plant disease, Deep learning, Efficient net B0, Vision transformer

Article History :
Received: 17 May 2025
Received in revised form: 02 July 2025
Accepted for publication: 04 July 2025
Available online: 15 July 2025

Communicated by: O. Akande

1. Introduction

Agriculture is the foundation of civilized world, comprising a vital part of the world economy. Farming provides livelihood, food, and shelter to millions of human beings across the globe. However, this key domain faces many challenges, encompassing the continuing effects of environmental change, the progression of crop disease-causing organisms, and the ongoing need to adopt sustainable agricultural practices. Plant diseases are becoming a significant obstacle to the world food supply and safety. Risk analysis and estimation of plant infection become very difficult due to their changing nature, depending on the types of diseases, the types of plants, the climatic conditions, and the soil conditions. Due to the similarity in symptoms of certain diseases or minimal differences, farmers often struggle to diagnose the diseases accurately.

Further, the symptoms of infection in the early stage are very tiny and it is difficult to see them with the naked eye, preventing the farmer from accurately detecting the disease early. Moreover, since many farmers lack the expertise to recognize infections, AI can assist them in making more accurate diag-

*Corresponding author Tel. No.: +234-902-402-1958.
Email address: rprasad@aust.edu.ng (Rajesh Prasad)

noses [1]. Much contribution is expected from the evolution of AI to help agriculture specialists and farmers make worldwide crop production and the agro ecosystem sustainable. Plant diseases, caused by viruses, fungi, bacteria, or other factors, have the potential to damage the crops, undermining the hard work of farm workers. Uncontrolled disease growth can destroy entire crops and lead to food shortages, environmental degradation, and economic losses. Therefore, there is a critical demand for automatic disease diagnosis that can accurately detect diseases in the early stages and thus save crops, money, and the environment.

ML and its breakthrough Deep Learning (DL) use in crop infection diagnosis have increased, demonstrating promising accuracy in identifying diseases from digital images [2]. Some important machine learning algorithms used in disease detection are ANN [3], SVM [4–6], and Naive Bayes [7] algorithms. Significant evolution and breakthroughs are making deep learning, particularly CNN, the first choice for researchers in the computer vision and object detection research fields. Barbedo [8] and Mohanty *et al.* [9] conducted advanced research in this field, utilizing CNN models such as VGG16, AlexNet, GoogleNet, and ResNet while incorporating the transfer learning approaches.

Additionally, computer scientists are developing tailored CNN techniques that integrate convolutional layers from pretrained architectures such as VGG16 and Inception Nets. This architecture uses convolution, Relu, max-pooling/average-pooling, and fully connected layers. The Inception block includes a max-pooling (3×3) layer and different kernel sizes of parallel convolutional layers. The VGG model remains a foundational architecture for many other models due to its effectiveness and simplicity, securing second position in the ILSVRC 2014 challenge [10–12].

Mohanty *et al.* [9] introduced a prediction method with a promising average accuracy of 99.30% through cross-validation. Likewise, many crop disease detection methods have been introduced for real crop field conditions. Kawasaki *et al.* [13] conducted ground breaking research by examining a 3-convolution layer network to identify three types of infections found in cucumber crop based on authentic crop leaf images, obtaining a 94.90% validation accuracy. Additional research on cucumbers/rice [14–16] and tomatoes [17] further validated the effectiveness of CNN and its variations in the farming domain. Recent advancements in the agriculture domain suggest that integrating CNN with a vision transformer greatly improves plant disease detection. Vision Transformers (ViT) have demonstrated outstanding potential in creating models for accurate crop disease detection. A prominent example is the Inception Convolution Vision Transformer (ICVT) [18] which is validated on three different datasets, including PlantVillage, PlantDoc and AI2018, where these datasets have given 99.94%, 77.74%, and 86.89% accuracy, respectively. The problem with Vision Transformer is that it is trained on many computational parameters, ranging from 25 to 86 million. Pure Vision Transformer-based models become very challenging for low-resource/ edge computing devices. Therefore, striking the right balance between model size and computational requirements is crucial.

Crop disease detection and effective control are critical to sustaining agricultural output, guaranteeing food security, and fostering economic expansion. Researchers favour deep learning models, particularly CNNs, as they can automatically extract meaningful and pertinent characteristics from incoming data, improving performance. However, as CNN models mainly concentrate on correlations between spatially neighboring pixels within the filter size-determined receptive field, they cannot capture links between distant pixels. Recently, researchers have investigated the use of attention mechanisms to overcome the limitations of CNN models in capturing relationships between distant pixels.

Our research presents a novel approach that utilizes the well-established EfficientNet-B0 Network for effective feature extraction from leaf images. This framework trains the EfficientNet-B0 using hyperparameter tuning to receive local patterns efficiently. EfficientNet-B0 was selected as the backbone feature extractor due to its optimal trade-off between accuracy and model size. Compared to other architectures such as ResNet, SqueezeNet or Inception, EfficientNet-B0 offers comparable or better result with significantly lower computational cost and lower number of parameters, making it well-suited for our application. Then, a vision transformer variant, ViT-small model, captures global relationships in disease patterns. Finally, the MLP head (a multi-layer perceptron) equipped on the vision transformer is utilised to classify the diseases. Our contributions can be succinctly outlined as follows:

- We integrated EfficientNet-B0 with Vision Transformer (ViT) to create a hybrid model that leverages local feature extraction and global attention mechanisms.
- The model design significantly reduces computational complexity by feeding ViT with compact, high-level features extracted from EfficientNet, minimizing the number of input tokens.
- This approach enhances accuracy while maintaining efficiency, making the model suitable for deployment in real-time and resource-constrained environments.
- Our model demonstrates strong generalization capabilities, especially on smaller datasets, thanks to the inductive bias introduced by the EfficientNet backbone.
- The hybrid architecture outperforms standard CNNs and ViTs in performance and efficiency for plant disease detection tasks.

The remaining sections of this work are as follows: Section 2 reviews recent studies on plant disease identification. Section 3 provides a detailed explanation of the materials and methodology used. Section 4 presents the experimental results along with a thorough discussion. Finally, Section 5 concludes the paper.

2. Literature review

This section discusses various methodologies, advancements, and techniques in the field. This section provides a summarised view of surveyed works with their main features and limitations. This survey work helps to find the gaps in the domain and accordingly motivates setting the objectives for the research work.

2.1. Related works

It provides an overview of the subject matter and presents past research on disease classification in plant leaves using computer vision, colour transformation, segmentation, and deep learning techniques. This section presents related works by various academicians and researchers from 2019 to 2025. The following are the reviews of various literature published in reputed journals and conferences.

Albattah *et al.* [19] introduced a drone-based disease detection system utilizing EfficientNetV2-B4 with additional dense layers. The model effectively deals with diverse image nature while reducing computational overhead, achieving 99.99% accuracy on the PlantVillage repository. However, its reliance on drone-captured imagery makes it less accessible to small-scale farmers, and its high computational demands pose difficulties for installation on low-power edge devices.

Zeng & Li [20] presented a CNN model with a self-attention mechanism for agricultural leaf disease detection. Several plant species from the MK-D2 and AES-CD9214 datasets were utilised to assess the suggested model. The model's classification accuracy on the MK-D2 dataset was 98.0%, but on the more realistic AES-CD9214 dataset, which included substantial variability in each sample, it was 95.33%. Their method increased accuracy but could only be tested on specific crop datasets.

Chen *et al.* [21] developed a rice disease recognition model using leaf images, integrating a lightweight attention mechanism with a pre-trained MobileNet-V2 for efficient and accurate identification, even in complex backgrounds. Tested on rice plants, the model achieved 99.67% accuracy on a public dataset and 98.48% on challenging images. Its lightweight design, however, could compromise detail on high-resolution or diversified datasets.

Wu Huang [22] developed a Vision Transformer-based model for detecting tomato leaf infections, demonstrating strong capability in distinguishing visually similar diseases. Tested on tomato plants, the model achieved 88.1% accuracy, though its high computational requirements may limit real-time applicability.

Reedha *et al.* [23] proposed a Vision Transformer with self-attention for crop and weed classification, demonstrating remarkable classification results even with small labelled training datasets. The model achieved 99.28% accuracy when tested on weeds, spinach, parsley, off-type green leaf beets, and beets images captured by a UAV. However, its high computational requirements may limit real-time applicability.

Thakur *et al.* [24] created a hybrid model based on a vision transformer and a CNN for plant leaf disease detection. Its lightweight design makes it perfect for Internet of Things-based smart agriculture. The model's accuracy was 98.33% for rice, 92.59% for maize, and 93.55% for apple when tested on the Embrapa and PlantVillage datasets for rice, maize, and apples. However, its performance can deteriorate under real-world circumstances with different environmental elements.

Yang *et al.* [25] created a modified GoogLeNet model that offers improved robustness in natural scene images while maintaining reduced model complexity for rice crop disease classification. The model was tested on a dataset of rice leaves gathered from crop fields and the Kaggle platform, and it achieved 99.58% accuracy. Although it works well for single-kind infected leaves, it might not work well for leaves affected by multiple diseases.

Sharma *et al.* [26] created a lightweight CNN for crop disease diagnosis ideal for real-time agricultural applications with just 6.4 million trainable parameters. The model's accuracy was 96.56%, 99.50%, 92.34%, and 93.56% on the tomato, grape, cucumber, and citrus datasets. However, the quality of the training samples may impact their performance.

Hosny *et al.* [27] developed a lightweight deep CNN model for plant leaf disease recognition. They coupled the CNN with Local Binary Pattern features to achieve high accuracy across multiple datasets. Apple, Grape, and Tomato datasets from the PlantVillage repository have given 98.3%, 96.5%, and 98.8% accuracy, respectively. Its dependence on LBP feature extraction may limit its ability to capture complex illness patterns. Similarly, Ref. [28] proposed a lightweight CNN structure for classifying leaf diseases. This model minimises the computation by generating fewer parameters. The model is validated on nightshade plant leaves. Validation achieved accuracy ranged from 95% to 100%. However, because it was trained solely on nightshade plant leaves, it might not be able to adapt to a range of real-world scenarios and generalise to other crops.

Thai *et al.* [29] applied Sparse Matrix-Matrix Multiplication (SPMM) in vision transfer architecture to propose a deep framework for cassava leaf disease identification with reduced computational complexity and improved attention pruning. Cassava leaf image samples were used to validate the model, and it succeeded in achieving 95% accuracy. However, it might be unable to record complex illness information due to its reliance on attention pruning.

Taji *et al.* [30] applied meta-heuristic algorithms to propose an optimized hybrid CNN structure to detect and classify plant leaf diseases. When evaluated on the PlantVillage dataset, the model attained 99.8% accuracy. However, the complex tuning of metaheuristics leads to high computational costs and time-intensive parameter optimisation.

In order to overcome class imbalance and use fewer parameters than current CNN and Transformer-based models, researchers in [31] created a lightweight MobileViT model with GAN-based augmentation for plant disease recognition using leaf pictures. It had 99.92% accuracy on PlantVillage and 75.72% on the PlantDoc dataset. However, its lower accuracy on PlantDoc suggests difficulties extrapolating to actual situations.

Aboelenin *et al.* [32] developed a hybrid deep learning

framework combining CNNs and ViT for detecting apple and corn leaves, effectively identifying and classifying multiclass plant diseases. Evaluated on apple and corn plants, the model achieved 99.24% accuracy and 98% for corn; however, combining multiple CNNS and Vit results in high computational complexity.

Baek [33] developed a Multi-Vision Transformer (Multi-ViT) model that effectively extracted spatially distributed illness symptoms. The model obtained 99% accuracy when tested on the Apple, Tomato, and Grape datasets. However, using multiple ViTs makes things more computationally complicated, making deploying in real time on devices with low resources challenging.

2.2. Research gaps

- CNNs are excellent in capturing small-scale patterns because they use local receptive fields, but are less adept at recognising global links in images.
- In certain plant illnesses, CNNs could overlook minute variations in color or texture. ViTs are excellent at distinguishing tiny changes in colour and texture, which boosts the classification accuracy of visually identical diseases. However, Vision Transformers (ViT) increase computing costs and challenge real-time deployment on edge devices.

2.3. Research objectives

- To build a Light weight hybrid deep framework equipped with Vision Transformer (Vit) that improves the recognition of minor colour and texture differences in plant infections.
- The suggested architecture uses CNNs first to generate compact feature maps, which lowers the number of patches needed by the self-attention mechanism, then effectively captures different features, allowing accurate detection and lowering the computational cost of self-attention.

3. Materials and methods

The methodology proposed for rice leaf disease classification is presented in Section 3. It outlines a workflow comprising distinct stages: dataset collection and its description, Baseline CNN architecture for automatic local feature extraction, global features extraction by utilising a lightweight vision transformer, and finally disease recognition and classification by integrating the baseline CNN and vision transformer into the hybrid framework, each method detailed in subsequent subsections.

3.1. Dataset description

The rice leaf dataset, obtained from Kaggle [34], was used to carry out the experiments in this study. It comprises images of four different types of rice leaves, with each class containing 4,000 images, as outlined in Table 1. The dataset includes

Table 1: Dataset classes and their size.

Leaves class	Number of images
Bacterial leaf blight	4,000
Brown spot	4,000
Health	4,000
Leaf smut	4,000

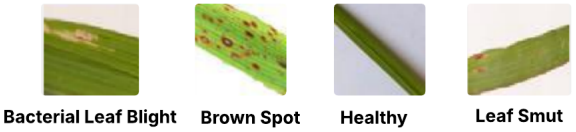


Figure 1: Samples of rice leaf diseases and healthy leaves from the dataset.

both healthy and unhealthy rice leaf images. The unhealthy category is divided into three distinct disease classes: Brown Spot, Bacterial Leaf Blight, and Leaf Smut. Therefore, the dataset effectively represents one healthy class and three disease-affected classes.

Figure 1 presents visual examples of healthy and diseased rice leaves, clearly illustrating the images utilised in this research.

The symptoms and causal organisms of the three kinds of rice diseases stored in the dataset are briefly described as follows:

1. Bacterial leaf blight: Bacterial Leaf Blight primarily affects the plant leaves. The disease is characterised by elongated lesions that can extend several inches. As the infection progresses, these lesions may spread across the entire leaf. Initially white, the lesions turn yellow due to bacterial activity. The disease is mainly transmitted through irrigation water and wind. It is caused by the bacterium *Xanthomonas oryzae*.
2. Brown spot: Brown Spot primarily affects the leaves of the plant. It is characterized by small, round to oval-shaped dark brown spots. The disease also leads to leaf shrinkage. While a fungus usually causes it, bacteria can also occasionally be responsible. The main causal organism is the fungus *Helminthosporium oryzae*.
3. Leaf smut: Leaf Smut primarily affects the leaves of the plant. The disease is identified by angular, slightly raised dark spots and dull patches with reddish-brown edges appearing on both sides of the leaves. It is caused by the fungus *Entyloma oryzae*. Leaf Smut can also lead to the development of other diseases in rice plants. The infection spreads through contaminated plant debris present in water and soil.

The utilised dataset consists of real, field-captured rice leaf images with diverse backgrounds, varying levels of disease severity (from mild to severe). The images show close-up views of the leaf surfaces, allowing for visual identification of disease symptoms such as lesions, spots, or discolouration. The deliberately low quality of the input images introduces natural

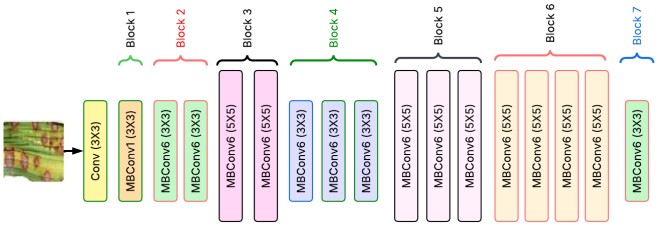


Figure 2: EfficientNet B0 architecture.

Table 2: Comparison of the utilized ViT-Small architecture with the ViT-Base model.

Model	Embedding Dimension	Layers	Heads	MLP Hidden Dimension	Parameters (Millions)
ViT-Small	384	12	6	1536	22M
ViT-Base	768	12	12	3072	86M

noise and variability, challenging the model to extract only the most critical features. As a result, the model becomes more robust and capable of generalising effectively to new, unseen data, rather than relying on superficial details. This dataset is valuable for training machine learning models for plant disease detection and classification, enabling automated diagnosis and contributing to precision agriculture efforts. Batch-wise data augmentation was applied during training to boost variability further and improve model generalisation. The plan involves building a more comprehensive dataset by capturing rice leaf images from actual crop fields under varying lighting and environmental conditions across multiple regions in our country, this will further enhance the model robustness.

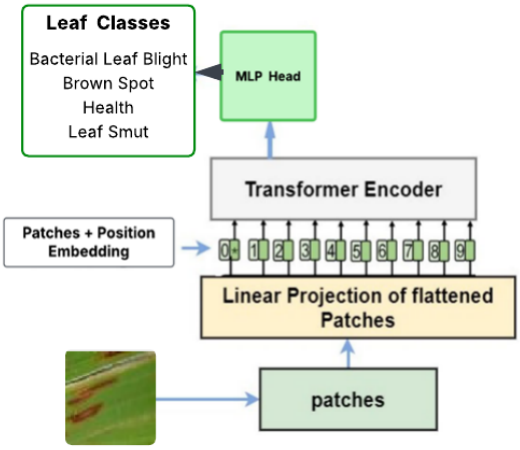
3.2. Baseline CNN architecture (EfficientNet-B0)

The developed model incorporates EfficientNet-B0 to enable transfer learning capabilities. Images from the balanced dataset are fed into the EfficientNet-B0 architecture (Figure 2). This architecture consists of a series of MBConv blocks, through which each image passes through for feature extraction. As the image moves through different layers, it undergoes transformations based on the filter size, enhancing its representation for further processing.

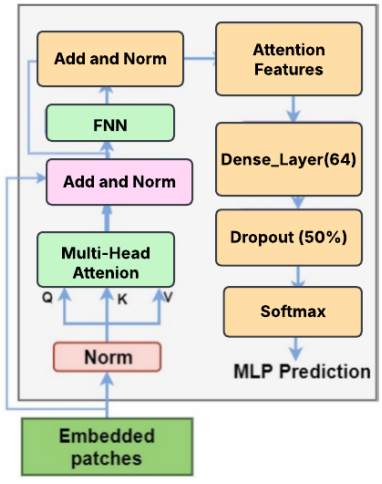
Initially, the input image has a size of 224×224. The choice of 224×224 as the input image resolution is driven by the architectural constraint of EfficientNet-B0, which is designed to accept images of this specific size. It is first processed by a convolutional layer with a 3×3 filter, transforming its dimensions to 224×224×32. This transformation helps retain essential features while making the image more suitable for further processing in deeper network layers. The feature map output from MBConv1 is then passed through two successive MBConv6 layers. A 3×3 filter is applied in these layers, further transforming the feature map to a resolution of 112×112×24. This step enhances feature extraction while maintaining spatial information for deeper network processing.

Next, the output is fed into two MBConv6 layers with a 5×5 filter. These layers further reduce the feature map size to 56×56×40, allowing for more refined feature extraction while preserving important spatial information. The output is then processed through three consecutive MBConv6 layers with a 3×3 filter, reducing the feature map size to 28×28×80. This step helps in capturing deeper features while progressively decreasing the spatial dimensions.

The resulting feature map is then passed through seven MBConv6 layers with a 5×5 filter, further reducing its spatial dimensions to 14×14 while increasing the depth to 192 channels. This step enhances feature extraction by capturing more complex patterns in the image. Finally, the feature map is passed through the last MBConv6 layer with a 3×3 filter, reducing its dimensions to 7×7×320. This final transformation prepares the extracted features for further processing, such as classification or detection, in the subsequent layers of the model.



(a) VIT model suggested by Dosovitskiy *et al.* [35].



(b) The transformer encoder.

Figure 3: Structure of ViT.

3.3. Vision transformer

Dosovitskiy *et al.* [35] developed the Vision Transformer (ViT) architecture by modifying the original transformer encoder, as shown in Figure 3, which effectively addresses NLP

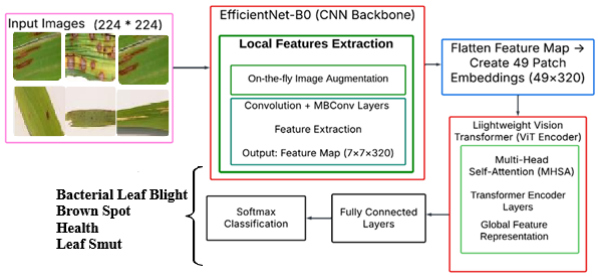


Figure 4: Proposed efficient-ViT B0Net architecture.

challenges. The ViT is a deep learning framework that replaces CNNs with transformer encoder enriched with self-attention mechanisms for image classification. ViTs have been more familiar since their proposal because of their ability to extract intricate spatial linkages and long-range dependencies in images. The ViT comprises positional embedding, MHA blocks, and networks of MLP with linear projection. The structure of the Vision transformer is shown in Figure 3 (a) and Figure 3 (b).

In proposed approach, we have used ViT-Small Architecture vision transformer model pretrained on ImageNet. This lightweight architecture has only 22M parameters in contrast to 86M parameters of the VIT-Base model. Comparison of ViT-Small architecture with VIT-Base model is given in Table 2.

3.4. Proposed hybrid architecture (Efficient-ViT B0Net)

EfficientNet-B0 pre-trained CNN architecture and light weight vision transformer (ViT) are combined in a hybrid model (Efficient-ViT B0Net) to achieve quicker processing than a pure ViT. Pretrained CNN architecture is used for automatic feature extraction for small-scale patterns, and the VIT model is used for capturing the minor colour or texture variation in the symptoms, which enables the model to identify the diseases even with similar symptoms correctly. In features extraction phase, On-the-fly image augmentation is used to enrich the dataset and generalize the model. On-the-fly image augmentation refers to real-time data augmentation that is applied dynamically to each mini-batch during training, rather than preprocessing and storing augmented images beforehand. The specific details of On-the-fly image augmentation is illustrated in Table 4. A hybrid approach lowers the computational cost of self-attention by first employing CNNs to create compact feature maps, which reduces the number of patches required by transformer encoder. Using CNNs for down-sampling and early feature extraction, a hybrid CNN-ViT model is quicker than a pure Vit while retaining high accuracy. It also speeds up training and increases real-time inference efficiency. Proposed Efficient-ViT B0Net model is shown in Figure 4. Training and testing algorithm of proposed framework (Efficient-ViT B0Net) is illustrated in Algorithm 1.

4. Experimental results and discussions

This part offers a thorough analysis of our study, beginning with the experimental configuration that details the conditions

Algorithm 1 Training and testing of efficient-ViT B0Net.

Given
Training dataset: $D = \{\{(I_i, y_i)\}\}_{i=1}^{N}$, where $I_i \in \mathbb{R}^{224\times224\times3}$, and y_i is the label.
Model parameters:
  EfficientNet-B0 parameters: θ_{EffNet}
  ViT-Small parameters: θ_{ViT}
  Fully connected classification head: θ_{FC}
Loss function: Cross-Entropy Loss, $\mathcal{L}$
Optimizer: O (AdamW)
Training Process
Step 1: Initialization
for each epoch $e = 1$ to E **do**
  Initialize total loss: $\mathcal{L}_{epoch} = 0$.
  Step 2: For each mini-batch B of size M sampled from D:
  Step 2.1: Preprocessing
  a. Apply data augmentation $\mathcal{T}$(random rotation within range of 10-20 degree, random translation within range of 3-6 pixels both horizontally and vertically)
  $I'_i = \mathcal{T}(I_i), \forall i \in B$.
  b. Normalize I'_i to [0,1].
  Step 2.2: Feature Extraction using EfficientNet-B0
  a. Pass each image through EfficientNet-B0:
  $F = f_{EffNet}(I'; \theta_{EffNet}) \in \mathbb{R}^{M\times7\times7\times320}$
  b. Flatten into 49 patches:
  $P = Flatten(F) \in \mathbb{R}^{M\times49\times320}$
  Step 2.3: Linear Projection for ViT-Small
  a. Project patches into ViT token dimension $d = 384$:
  $Z = PW_P + b_P, Z \in \mathbb{R}^{M\times49\times384}$
  Step 2.4: Feature Learning using Vision Transformer
  for each transformer block, $t = 1$ to 12 layers **do**
    $Q_t = XW^t_Q, K_t = XW^t_K, V_t = XW^t_V$
    b. Compute Multi-Head Self-Attention (MHSA):
    $MHSA(X) = \sum_{h=1}^{H} Softmax\left(\frac{Q_h K_h^T}{\sqrt{d}}\right) V_h$
    c. Apply Residual Connection + LayerNorm:
    $X' = LN(X + MHSA(X))$
    d. Apply an MLP feedforward network:
    $X'' = LN(X' + MLP(X'))$
    e. Update token representations:
    $X \leftarrow X''$
  end for
  Step 2.5: Classification
  a. Compute global representation:
  $X_{global} = \frac{1}{49}\left(\sum_{k=1}^{49} X_k\right)$
  b. Pass through fully connected layers:
  $X_{fc1} = \sigma(X_{global} W_{fc1} + b_{fc1})$
  $X_{fc2} = \sigma(X_{fc1} W_{fc2} + b_{fc2})$
  c. Compute Softmax output:
  $\hat{y} = Softmax(X_{fc2})$
  Step 2.6: Compute Loss and Backpropagation
  a. Compute Cross-Entropy Loss:
  $\mathcal{L}_{batch} = -\frac{1}{M}\left(\sum_{i=1}^{M} y_i \log(\hat{y}_i)\right)$
  b. Accumulate total loss:
  $\mathcal{L}_{epoch} += \mathcal{L}_{batch}$
  c. Compute gradients:
  $\nabla_\theta \mathcal{L} = \frac{\partial \mathcal{L}}{\partial \theta_{EffNet}, \theta_{ViT}, \theta_{FC}}$
  d. Update model parameters using optimizer O:
  $\theta_{EffNet} \leftarrow \theta_{EffNet} - \eta \nabla_{\theta_{EffNet}} \mathcal{L}$
  $\theta_{ViT} \leftarrow \theta_{ViT} - \eta \nabla_{\theta_{ViT}} \mathcal{L}$
  $\theta_{FC} \leftarrow \theta_{FC} - \eta \nabla_{\theta_{FC}} \mathcal{L}$
end for
Step 3. Compute Average Epoch Loss
$\mathcal{L}_{avg.} = \frac{\mathcal{L}_{epoch}}{Number\ of\ batches}$
Testing Process
Step 1: Set the model to evaluation mode.
for each batch in the test dataset **do**
  Step 2.1: Compute predictions $\hat{y}$
  Step 2.2: Compare with ground truth y.
  Step 2.3: Compute accuracy:
  $Accuracy = \frac{Correct\ Predictions}{Total\ Samples} \times 100$.
end for
Step 3: Compute average accuracy:
$Avg.\ Accuracy = \frac{Sum\ of\ batchwise\ Accuracy}{Number\ of\ batches}$.

Table 3: Train and test split of the dataset.

Leaves class	Number of images	Number of training samples	Number of testing samples
Bacterial leaf blight (BLB)	4,000	3200	800
Brown Spot (BS)	4,000	3200	800
Health	4,000	3200	800
Leaf Smut (LS)	4,000	3200	800

under which our models were developed and evaluated. Next, we describe the assessment measures utilized to evaluate our models' performance. The experimental results evaluating the effectiveness and efficiency of the suggested hybrid deep framework for recognizing and classifying plant leaf diseases are presented here. It also performs the ablation study and analysis of the model's performance. Lastly, we perform a comparative analysis of the suggested approach with the state-of-the-art (SOTA) models. The effectiveness of different model architectures was analyzed using the complete dataset provided.

4.1. Experimental environment

The suggested deep learning framework, Efficient-Vit B0Net, was implemented utilising MATLAB 2019a platform and Deep Learning Toolbox for vision transformer and efficient Netb0 Network support package. Adamw and categorical cross-entropy are utilised as optimisation and loss functions to perform the training process on the given samples. Both the models (Efficient Netb0 and VIT -small) used a learning rate of 0.001 with eight epochs to ensure early and effective convergence. The computational setup consists of a Dell laptop with a Core i7 processor operating at 3.60 GHz and a powerful NVIDIA RTX 3080 GPU, providing significant processing and graphical capabilities. This hardware is supported by 16 GB of DDRAM for system memory and features a 1TB storage drive with a 512 GB SSD component for fast data access and system responsiveness. The software environment runs on the Windows 10 operating system, with implementation primarily carried out using MATLAB 2019a, incorporating the TensorFlow library for relevant tasks. The proposed model was trained on 80% of the dataset samples and tested on the remaining 20%. The training and testing split of the dataset is shown in Table 3.

4.2. Model assessment metrics

The performance assessment of the proposed deep architecture and the baseline architecture was done using a comprehensive set of metrics calculated on the held-out test set:

4.2.1. Classification metrics

1. Accuracy: The overall proportion of correctly classified images. Accuracy = True Classification/ Total Classification (True Classification = TP + TN, Total Classification = TP + TN + FP + FN).
2. Precision: The system can recognise only relevant samples for each class. Precision = TP / (TP + FP). Reported as macro-average (unweighted mean across classes) and/or per-class.
3. Recall (Sensitivity): Recall is the ability of the classifier to recognise all relevant samples for each class. Recall = TP / (TP + FN). Reported as macro-average and/or per-class.
4. F1-Score: A single parameter to balance Recall and Precision is a harmonic mean of both. F1-Score = 2*(Precision*Recall) / (Precision + Recall). Reported as macro-average and/or per-class.
5. Confusion Matrix: A table visualising the performance, indicating the counts of TP, TN, FP, and FN for each class.

4.3. Experimental results and discussions

This subsection demonstrates the performance analysis of the proposed hybrid deep framework, integrating pretrained EfficientNetB0 and pretrained Vision Transformer (ViT)-Small in sequence, for plant disease recognition. The assessment was done on the sourced dataset consisting of four distinct leaf classes: Healthy, Brown-spot, Bacterial-leaf-blight, and Leaf-smut.

On-the-fly image augmentation was applied using random translation, scaling, and rotation. The specific values for image augmentation are detailed in Table 4. Deep learning models were developed using a rice leaf dataset containing 16,000 images for training and validation, in which 12,800 samples were utilised for training and 3,200 samples were utilised for testing. Class-wise division of training and testing samples is shown in Table 3. The training phase utilised a batch size of 32 samples, with 100 iterations in each epoch. The model was evaluated on different numbers of epochs, initially starting from 6 epochs. The suggested model, Efficient-ViT B0Net, was most effective and efficient in 8 epochs.

The model achieved a high overall accuracy of 99.13%, with all classes showing strong, accurate, favourable rates. Classification results of Proposed Efficient-ViT B0Net model in form of Confusion Matrix is shown in Figure 5. The Healthy class had the highest accuracy with 798 correct predictions and only 2 misclassifications. At the same time, Brown Spot, though still performing well, had the most confusion, primarily misclassified as BLB in 5 instances. The Leaf smut and Bacterial leaf blight classes also showed excellent results with minimal misclassifications. These class-wise misclassifications are shown in Table 5. Visualization of these mis-classification along with correct classification is shown in Figure 6.

The true positive and accurate negative counts are high across all classes, and the false positives and false negatives remain very low, as listed in Table 6, indicating the model's ability to distinguish between similar leaf conditions with great Precision. This analysis confirms that the hybrid model is highly reliable for practical use in disease detection and classification tasks in agricultural applications. The class-wise performance metrics are illustrated in Table 6, whereas the proposed model's

Table 4: Augmentation details.

Augmentation	Description	Typical Parameters	Purpose
Random Translation	Shifts the image horizontally and/or vertically by a random amount.	Horizontal/Vertical shift range (by pixel distances of -30 and 30, respectively)	Helps the model become invariant to object positioning.
Random Scaling	Randomly zooms in or out by resizing the image, optionally keeping the original size with padding/cropping.	Scale range (0.8× of original size to 1.2× of original size)	Trains the model to recognize objects at different sizes and distances.
Random Rotation	Rotates the image by a random degree within a given range.	Angle range (−20° to +20°)	Improves rotational robustness and reduces overfitting to specific orientations.

Table 5: Class-wise misclassification results.

Class	Total Samples	True Positive	Mis-classification	True positive rate (Sensitivity)
Bacterial Leaf Blight	800	792	8 (4 as Brown Spot, 2 as Health, 2 as Leaf Smut)	99.00%
Brown Spot	800	789	11 (5 as Bacterial Leaf Blight, 2 as Health, 4 as Leaf Smut)	98.63%
Health	800	798	2 (1 as Bacterial Leaf Blight, 1 as Brown Spot)	99.75%
Leaf Smut	800	793	7 (2 as Bacterial Leaf Blight, 5 as Brown Spot)	99.13%

Table 6: Class-wise performance results.

Class	TP	TN	FP	FN	Accuracy	Recall	Precision	F1-score
Bacterial Leaf Blight	792	2392	8	8	99.5%	99.00%	99.00%	99.00%
Brown Spot	789	2390	10	11	99.34%	98.63%	98.75%	98.69%
Health	798	2396	4	2	99.81%	99.75%	99.50%	99.62%
Leaf Smut	793	2394	6	7	99.59%	99.13%	99.25%	99.19%

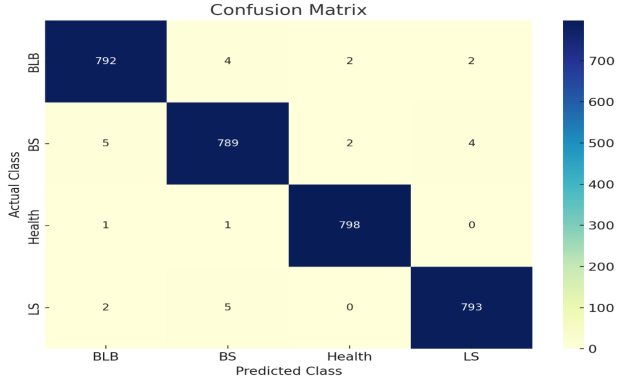


Figure 5: Classification results of proposed Efficient-ViT B0Net model (confusion matrix).

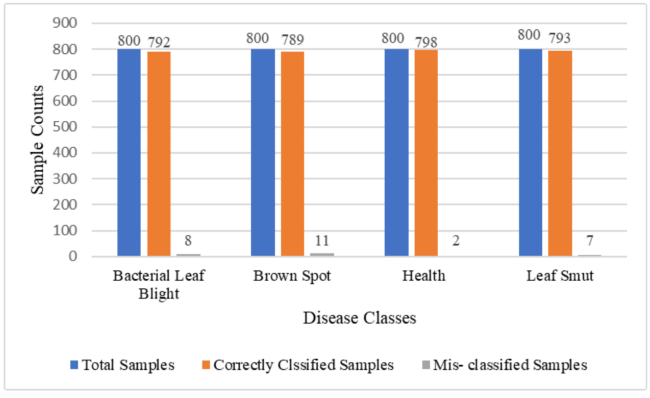


Figure 6: Disease wise sample counts of correct and incorrect classification.

performance parameters are listed in Table 7. Disease wise performance comparison of proposed model is illustrated in bar chart shown in Figure 7. All the illustrated performance metrics were computed based on testing dataset. Overall accuracy of the proposed model was recorded as 99.13%, which is an ac-

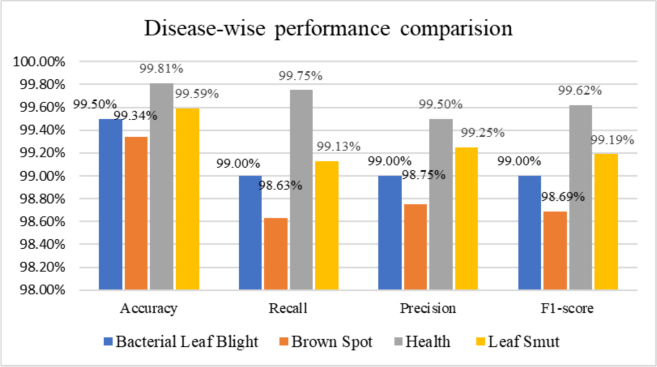


Figure 7: Disease wise performance comparison of proposed model.

Table 7: Proposed model's performance results.

Performance parameters	Formula	Value
Overall Accuracy (OA)	$OA = \frac{Total\ TP\ across\ all\ classes}{Total\ number\ of\ Samples}$	99.13%
Macro Recall (MR)	$MR = \frac{Total\ Recall\ across\ all\ classes}{Total\ number\ of\ classes}$	99.13%
Macro Precision (MP)	$MP = \frac{Total\ Precision\ across\ all\ classes}{Total\ number\ of\ classes}$	99.13%
Macro F1-Score(MF)	$M = \frac{Total\ F1-Score\ across\ all\ classes}{Total\ number\ of\ classes}$	99.13%

Table 8: Performance comparison with SOTA models.

SOTA Model	Architecture	Accuracy
Efficient Net Bo	Lightweight CNN	91.3%
Squeeze Net	Lightweight CNN with Fire modules	90.7%
ResNet50	Deep residual network	94.5%
DenseNet121	Dense connections with feature reuse	96.1%
VIT-Small	Vision Transformer-Small Variant	97.0%
Proposed Efficient-ViT B0Net	EfficientNet-B0 + Vision Transformer + Self-Attention	99.1%

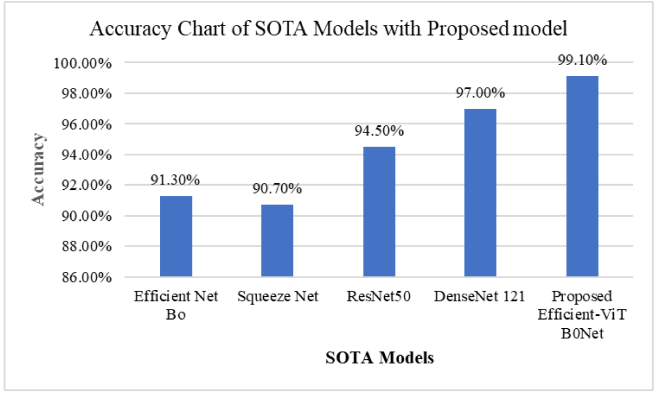


Figure 8: Accuracy comparison of SOTA model with proposed model.

curacy showing excellent classification behaviour of the model.

Recall, Precision and F1-Score of the model are computed by taking class-wise macro average of Recall, Precision and F1-Score, respectively. The proposed model achieved a Macro Recall of 99.13%, a Macro Precision of 99.13%, and a Macro F1-Score of 99.13%.

4.4. Performance comparison with state-of-the-art (SOTA) models

We have compared the performance of SOTA models by keeping the same experimental setup. Results are analysed over eight epochs with the same batch size of 32 samples. On-the-fly augmentation is used with all the models. The comparison of the performance of the SOTA models with a proposed network is summarised in Table 8. The comparison of state-of-the-art models highlights the effectiveness of the proposed Efficient-ViT B0Net, which outperforms all other architectures with an impressive 99.13% accuracy. While lightweight models like SqueezeNet and EfficientNet-B0 offer speed and efficiency, they fail to handle visually similar disease patterns due to limited contextual understanding. Deeper models like ResNet50 and DenseNet121 improve accuracy but cannot model global relationships. Vit-Small addresses this with better context awareness, achieving 97%. However, the hybrid approach of combining EfficientNet-B0 with a Vision Transformer and self-attention delivers the best of both worlds-efficient local feature extraction and robust global attention, resulting in superior class separation and minimal misclassification. Illustration of accuracy comparison of SOTA models with proposed model is shown in Figure 8.

4.5. Limitations and challenges

While highly accurate, the proposed Efficient-ViT B0Net hybrid model presents several limitations and challenges. It shows sensitivity to environmental factors such as lighting variations, background noise, and seasonal changes, which can affect its generalizability in real-world agricultural conditions. Occasional misclassifications occur, particularly between diseases with similar symptoms but differing visual traits, indicating a need for more refined feature extraction. Additionally, the proposed model fails to recognise and classify multiple disease infections found on a single leaf. Future work will focus on improving environmental robustness and reducing false detections.

5. Conclusion and future scope

The proposed Efficient-ViT B0Net hybrid model integrates the lightweight and efficient EfficientNet-B0 with the powerful Vision Transformer and self-attention mechanisms, and has proven highly effective in plant leaf disease classification. By combining local feature extraction and global context awareness, the model achieves a remarkable accuracy of 99.13%, significantly outperforming other state-of-the-art (SOTA) models such as ResNet50, DenseNet121, Vit-Small, and SqueezeNet.

The class-wise high Precision, recall, and F1-score scores collectively suggest that the model is accurate overall. It maintains a strong balance between correctly identifying positive instances (high recall) and ensuring that the instances it identifies as positive are correct (high Precision). This level of performance strongly supports the hypothesis that combining EfficientNetV1-B0 and ViT-Small provides a synergistic advantage for this task. The model can distinguish between visually similar disease classes and shows minimal misclassification, demonstrating its strong generalisation in controlled conditions. However, challenges remain in adapting the model to real-world agricultural environments, particularly in dealing with diverse lighting, seasonal changes, and background variations.

In future work, the proposed Efficient-ViT B0Net model will be further enhanced to improve its adaptability and performance in real-world agricultural environments. A key focus will be expanding the dataset to include images captured under diverse environmental conditions such as varying lighting, seasonal changes, crop growth stages, and different backgrounds. This will help improve the model's robustness and generalisation across different field scenarios. Additionally, advanced data augmentation techniques and multispectral or hyperspectral imaging integration will be explored to simulate real-world variability better and enhance feature representation. Efforts will also be directed toward refining the model's feature extraction and attention mechanisms to minimise false positives and negatives, especially in visually similar diseases.

Data availability

The data that support the findings of this research are openly available at: https://www.kaggle.com/datasets/bahribahri/riceleaf.

References

[1] Y. Borhani, J. Khoramdel & E. Najafi, "A deep learning based approach for automated plant disease classification using vision transformer", Sci. Rep. **12** (2022) 1. https://doi.org/10.1038/s41598-022-15163-0.

[2] M. Shoaib, B. Shah, S. EI-Sappagh, A. Ali, A. Ullah, F. Alenezi, T. Gechev, T. Hussain, F. Ali, "An advanced deep learning models-based plant disease detection: A review of recent research", Frontiers in Plant Science **14** (2023) 1. https://doi.org/10.3389/fpls.2023.1158933.

[3] H. Hamdani, A. Septiarini, A. Sunyoto, S. Suyanto & F. Utaminingrum, "Detection of oil palm leaf disease based on color histogram and supervised classifier", Optik (Stuttg). **245** (2021) 1. https://doi.org/10.1016/j.ijleo.2021.167753.

[4] C. Hou, J. Zhuang, Y. Tang, Y. He, A. Miao, H. Huang, S. Luo, "Recognition of early blight and late blight diseases on potato leaves based on graph cut segmentation", J. Agric. Food Res. **5** (2021) 1. https://doi.org/10.1016/j.jafr.2021.100154.

[5] Y. Sun, Z. Jiang, L. Zhang, W. Dong & Y. Rao, "SLIC_SVM based leaf diseases saliency map extraction of tea plant", Comput. Electron. Agric. **157** (2019) 1. https://doi.org/10.1016/j.compag.2018.12.042.

[6] S. Zhang & Z. Wang, "Cucumber disease recognition based on global-local singular value decomposition", Neurocomputing **205** (2016) 1. https://doi.org/10.1016/j.neucom.2016.04.034.

[7] A. Johannes, A. Picon, A. Alvarez-Gila, J. Echazarra, S. Rodriguez-Vaamonde, A. D. Navajas, A. Ortiz-Barredo, "Automatic plant disease diagnosis using mobile capture devices, applied on a wheat use case", Comput. Electron. Agric. **138** (2017) 1. https://doi.org/10.1016/j.compag.2017.04.013.

[8] J. G. A. Barbedo, "Impact of dataset size and variety on the effectiveness of deep learning and transfer learning for plant disease classification", Comput. Electron. Agric. **153** (2018) 1. https://doi.org/10.1016/j.compag.2018.08.013.

[9] S. P. Mohanty, D. P. Hughes & M. Salathé, "Using deep learning for image-based plant disease detection", Front. Plant Sci. **7** (2016) 1. https://doi.org/10.3389/fpls.2016.01419.

[10] J. Chen, J. Chen, D. Zhang, Y. Sun & Y. A. Nanehkaran, "Using deep transfer learning for image-based plant disease identification", Comput. Electron. Agric. **173** (2020) 1. https://doi.org/10.1016/j.compag.2020.105393.

[11] P. S. Thakur, T. Sheorey & A. Ojha, "VGG-ICNN: a lightweight CNN model for crop disease identification", Multimed. Tools Appl. **82** (2023) 1. https://doi.org/10.1007/s11042-022-13144-z.

[12] S. R. Shah, S. Qadri, H. Bibi, S. M. W. Shah, M. I. Sharif & F. Marinello, "Comparing inception V3, VGG 16, VGG 19, CNN, and ResNet 50: a case study on early detection of a rice disease", Agronomy **13** (2023) 1. https://doi.org/10.3390/agronomy13061633.

[13] Y. Kawasaki, H. Uga, S. Kagiwada & H. Iyatomi, "Basic study of automated diagnosis of viral plant diseases using convolutional neural networks", in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2015, pp. 1. [Online]. https://doi.org/10.1007/978-3-319-27863-6_59.

[14] N. Upadhyay, S. K. Singh, R. Kumar, A. Sagar, "Rice leaves disease detection mechanism using VGG16 deep learning architecture", in *Innovations in data analytics*, ICIDA 2023, D. Bhattacharya, A., Dutta, S., Dutta, P., Samanta (Ed.), Springer, 2024, p. 1005. https://doi.org/10.1007/978-981-97-4928-7_17.

[15] H. Mavi, S. K. Upadhyay, N. Srivastava, R. Sharma and R. Bhargava, "Crop recommendation system based on soil quality and environmental factors using machine learning", in 2024 International Conference on Emerging Innovations and Advanced Computing (INNOCOMP), 2024, p. 1. https://doi.org/10.1109/INNOCOMP63224.2024.00089.

[16] B. Rashmi, S. K. Upadhyay, R. Sharma, N. Srivastava, H. Mavi, "Plant disease detection using Machine learning and CNN on leaf images", in 2024 1st International Conference on Advanced Computing and Emerging Technologies (ACET), 2024, p. 1. https://doi.org/10.1109/ACET61898.2024.10730720.

[17] A. Fuentes, S. Yoon, S. C. Kim & D. S. Park, "A robust deep-learning-based detector for real-time tomato plant diseases and pests recognition", Sensors (Switzerland) **17** (2017) 1. https://doi.org/10.3390/s17092022.

[18] S. Yu, L. Xie & Q. Huang, "Inception convolutional vision transformers for plant disease identification", Internet of Things (Netherlands) **21** (2023) 1. https://doi.org/10.1016/j.iot.2022.100650.

[19] W. Albattah, A. Javed, M. Nawaz, M. Masood & S. Albahli, "Artificial Intelligence-Based Drone System for Multiclass Plant Disease Detection Using an Improved Efficient Convolutional Neural Network", Front. Plant Sci. **13** (2022) 1. https://doi.org/10.3389/fpls.2022.808380.

[20] W. Zeng & M. Li, "Crop leaf disease recognition based on Self-Attention convolutional neural network", Comput. Electron. Agric. **172** (2020) 1. https://doi.org/10.1016/j.compag.2020.105341.

[21] J. Chen, D. Zhang, A. Zeb & Y. A. Nanehkaran, "Identification of rice plant diseases using lightweight attention networks", Expert Syst. Appl. **169** (2021) 1. https://doi.org/10.1016/j.eswa.2020.114514.

[22] S. Wu, Y. Sun & H. Huang, "Multi-granularity feature extraction based on vision transformer for tomato leaf disease recognition", in 2021 3rd International Academic Exchange Conference on Science and Technology Innovation, IAECST 2021, 2021, p. 1. https://doi.org/10.1109/IAECST54258.2021.9695688.

[23] R. Reedha, E. Dericquebourg, R. Canals & A. Hafiane, "Transformer neural network for weed and crop classification of high resolution UAV images", Remote Sens. **14** (2022) 1. https://doi.org/10.3390/rs14030592.

[24] P. S. Thakur, P. Khannam, T. Sheorey, A. Ojha, "Explainable vision transformer enabled convolutional neural network for plant disease identification: PlantXViT", 2022. ArXiv [Online]. https://doi.org/10.48550/arXiv.2207.07919.

[25] L. Yang, X. Yu, S. Zhang, H. Long, H. Zhang, S. Xu, Y. Liao, "GoogLeNet based on residual network and attention mechanism identification of rice leaf diseases", Comput. Electron. Agric. **204** (2023) 1. https://doi.org/10.1016/j.compag.2022.107543.

[26] V. Sharma, A. K. Tripathi & H. Mittal, "DLMC-Net: deeper lightweight multi-class classification model for plant leaf disease detection", Ecol. Inform. **75** (2023) 1. https://doi.org/10.1016/J.ECOINF.2023.102025.

[27] K. M. Hosny, W. M. El-Hady, F. M. Samy, E. Vrochidou & G. A. Papakostas, "Multi-class classification of plant leaf diseases using feature fusion of deep convolutional neural network and local binary pattern", IEEE Access **11** (2023) 1. https://doi.org/10.1109/ACCESS.2023.3286730.

[28] B. M. Joshi & H. Bhavsar, "A nightshade crop leaf disease detection using enhance-nightshade-CNN for ground truth data", Vis. Comput. **40** (2024) 5639. https://doi.org/10.1007/s00371-023-03127-y.

[29] H. T. Thai, K. H. Le & N. L. T. Nguyen, "FormerLeaf: an efficient vision transformer for Cassava Leaf Disease detection", Comput. Electron. Agric. **204** (2023) 1. https://doi.org/10.1016/j.compag.2022.107518.

[30] K. Taji, A. Sohail, T. Shahzad, B. S. Khan, M. A. Khan, K. Ouahada, "An ensemble hybrid framework: a comparative analysis of metaheuristic algorithms for ensemble hybrid CNN features for plants disease classification", IEEE Access **12** (2024) 61886. https://doi.org/10.1109/ACCESS.2024.3389648.

[31] A. K. Singh, A. Rao, P. Chattopadhyay, R. Maurya & L. Singh, "Effective plant disease diagnosis using vision transformer trained with leafy-generative adversarial network-generated images", Expert Syst. Appl. **254** (2024) 1. https://doi.org/10.1016/J.ESWA.2024.124387.

[32] S. Aboelenin, F. A. Elbasheer, M. M. Eltoukhy, W. M. El-Hady & K. M. Hosny, "A hybrid framework for plant leaf disease detection and classification using convolutional neural networks and vision transformer", Complex Intell. Syst. **11** (2025) 1. https://doi.org/10.1007/s40747-024-01764-x.

[33] E. T. Baek, "Attention score-based multi-vision transformer technique for plant disease classification", Sensors **25** (2025) 1. https://doi.org/10.3390/s25010270.

[34] "rice-leaf" Accessed: May 05, 2025. [Online]. Available: https://www.kaggle.com/datasets/bahribahri/riceleaf.

[35] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, "An image is worth 16×16 words: transformers for image recognition at scale", in ICLR 2021 - 9th International Conference on Learning Representations, 2021, p. 1. ArXiv. [Online]. https://doi.org/10.48550/arXiv.2010.11929.