iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI: https://doi.org/10.22214/ijraset.2026.79195

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

Precision Agriculture at Scale: Early Disease Detection using Vision Transformers

Nikunjkumar Tailor[1], Nensi Panchal[2], Bhumi Torani[3]
[1]*M. Tech. Computer Engineering, Bhagwan Mahavir University, Surat, Gujarat*
[2]*Assistant Professor, Computer Engineering, Bhagwan Mahavir University, Surat, Gujarat*
[3]*Department of Information Technology, Tapi Diploma Engineering College, Surat, Gujarat, India*

Abstract: ***Plant diseases remain a major concern for global food security as they cause heavy losses in crop every year. Accurately detecting these diseases early isthought to be of crucial help for improving yield in crops. Manual detection approaches as well as CNN based have turned out quite promising, but their long-range relations and overall context lead them to fail when it comes to leaf image representation. To address the problems described above, we propose plant leaf disease detection basedon Vision Transformer model in this paper. In contrast to most models that adopt convolutional neural networks for image feature extraction, our model uses a transformer for vision which has self-attention mechanism and learns global information on features. Therefore, it extracts more informative features.***
Our model will have several components: Image Preprocessing -> Image Patches -> Feature learning (Transformer Encoder) -> Classification (Fully Connected Layer). Experimental results confirm that the proposed research method reaches an accuracy of 99.8% in classification, which is better than most existing deep learning approaches. In addition, confusion matrix analysis of our model performance suggests a minimal error rate on each kind of disease category.The Transformer-based model holds promise for realistic practical application in precision agriculture. It is capable of automatically detecting plant diseases in real time. The researchers also believe that the promising future of this transformer-based approach could mean a development in low-computing power environments for efficient systems of agriculture image analysis, depending on reproducibility and generalization.
Keywords: Plant Leaf Disease Detection, Precision Agriculture, Vision Transformer (ViT), Image Classification, Deep Learning, Self-Attention, Computer Vision, Artificial Intelligence

I. INTRODUCTION

The issue of food security on a global scale has become an important problem in the 21st century mainly due to the rise of the population rate, the climate conditions and the decrease in the efficiency of the agricultural production activities. Among the critical problems is the prevalence of plant diseases, which results in the global loss of considerable crop yield. Studies suggest that plant diseases can cause a loss of nearly 40% of crops yields per year, which is an important issue for sustainable crop production and agricultural output. Therefore Accurate diagnosis of plant diseases is critical so that the best crop can be produced.

In the conventional method, Plant pathology experts diagnose crop diseases. Although, the results can be obtained to a degree using this approach, it is labor-intensive and time consuming. Moreover, results of examination can be biased in a certain extent depending on the judgment of the expert. Therefore, the traditional method is not suitable to implement in the situation of a large-scale farming environment. With the arrival of artificial intelligence, the concept of automatic detection systems through deep learning has attained great importance. Due to their exceptional capacity for detecting local features within visual data, Convolutional Neural Networks (CNNs) represent the premier approach for image classification tasks. Although the model is the best, there is a problem in the spatial relationships in long distance, maybe this would make the performance drop when the model is applied to the real world.

To bridge the gaps identified in current modeling techniques, the study aims to explore and apply the Vision Transformer model architecture that is unique as compared to the conventional models. Unlike the conventional models, which rely on local feature extraction techniques, the Vision Transformer models and use a self-attention technique to allow the model to learn local and global features from the image. Based on the recent advancements and updates made to the models between 2017 and 2026, it is evident that models using the transformer or hybrid models are more effective than the conventional CNN models when in terms of plant leaf disease detection tasks.

In this case, models including the hybrid model comprising the CNN model and the Vision Transformer model, and the Swin Transformer model and the mixture of experts models are more robust as compared to the conventional models.

In addition to this, the recent trends in the research field are focusing on developing an efficient model for precision agriculture with the help of transformer models. The light versions of this model such as mobile friendly vision transformers are efficient in providing high accuracy and efficiency by reducing the computational costs of this model. In addition to this, the reliability of this model can also be enhanced by a gAN-based data augmentation and attention mechanism.

The importance of this method is based on its ability to achieve very high detection accuracy, up to 99.8%, and thus make it possible to carry out disease identification in efficient ways. Such high-tech systems could be incorporated in different kinds of automated diagnostic tools, which would minimize the need for expertise in disease identification and enable farmers to take appropriate corrective measures in a timely manner. This makes it possible to minimize losses, while supporting efficient resource management, making it possible to increase agricultural productivity through the adoption of transformer models, including Vision Transformers.

II. LITERATURE REVIEW

Deep learning in plant disease detection has attracted much attention in recent years[22], especially due to its capability of automating the detection while at the same time enhancing the detection accuracy. Traditionally, various approaches in plant disease detection have involved conventional image processing techniques as well as machine learning approaches such as Support Vector Machines (SVM) and k-Nearest Neighbors (k-NN). Although these paradigms have recorded moderate success, they are feature-dependent, making it difficult to handle the variations associated with plant leaves. [2]-[6], [10], [11]

As the development of deep learning technologies progressed, the Convolutional Neural Network (CNN) architecture became the preferredchoice for plant leaf disease classification. The AlexNet, VGG, and ResNet architectures, among others, have been shown to be effective in learning hierarchical representations. CNNs have been demonstrated to be effective in detecting microstructural representations among others. Several authors have demonstrated the efficacy of the CNN-based approach in obtaining high accuracy, especially in a controlled environment. Nevertheless, A fundamental bottleneck in CNN design is its local receptive field, which does not take into account the contextual information, especially when the disease occurs in multiple locations on the leaf. [2]-[6]

To overcome these weaknesses, recent studies have begun to focus more on transformer models and their application to vision tasks, particularly on Vision Transformers The model, initially meant for NLP, was adapted for image processing. Unlike CNN models that apply convolutional operations to images, Vision Transformers breaks image into smallerpatches and use self-attention to process all parts of an image. Vision Transformers are emerging as a faster, more accurate alternative to CNNs for identifying plant pathogens.

As a result, to achieve superior results, hybrid models developed by integrating the local feature extraction strengths of CNNs with transformer's global contextual understanding inherent feature. Moreover, Models like Swin Transformers have shown great accuracy and resilience in demanding, real-world agricultural settings, which are characterized by diverse environment factors [12]-[14], [17]-[21].

Vision Transformers, designed for mobile devices, have shown strong performance while using less computational power, efficient and also smaller. This makes them the best candidates for deployment on autonomous and mobile edge devices.Moreover, the application of data augmentation methodologies like GANs, along with the development of efficient attention mechanisms, has helped in the improvement of models' ability to generalize. Despite these advancements, practical use of transformer-based models presents some difficulties. [12]-[14], [17]-[21]

The requirement for extensive datasets, coupled with the significant computational resources they necessitate, poses substantial impediments. However, ongoing research initiatives are diligently addressing these issues, thus highlighting the opportunity for Vision Transformers to create sophisticated plant leaf disease detection systems.[7], [8], [15], [16]

Aspect	Convolutional Neural Network (CNN)	Vision Transformer (ViT)
Feature Extraction	Extracts local features using convolutional filters	Captures global features using self-attention mechanism
Spatial Dependency	Focuses on short-range dependencies	Models long-range dependencies effectively
Context Awareness	Limited global context understanding	Strong ability to capture global context
Data Requirement	Works well with compact datasets	Requires Extensivedatasets for optimal performance

Computational Cost	Lower, efficient for most tasks	Higher, due to attention computations
Performance (Simple Data)	High accuracy on simple datasets	Comparable performance
Performance (Complex Data)	Moderate performance on complex patterns	Very high performance due to global understanding
Interpretability	Moderate (feature maps)	High (attention maps provide better insights)
Deployment	Easier to deploy and optimize	Requires optimization and more resources
Recent Trend	Mature and widely used technology	Rapidly evolving and gaining popularity

Table:1 Comparison of CNN and Vision Transformer (ViT)

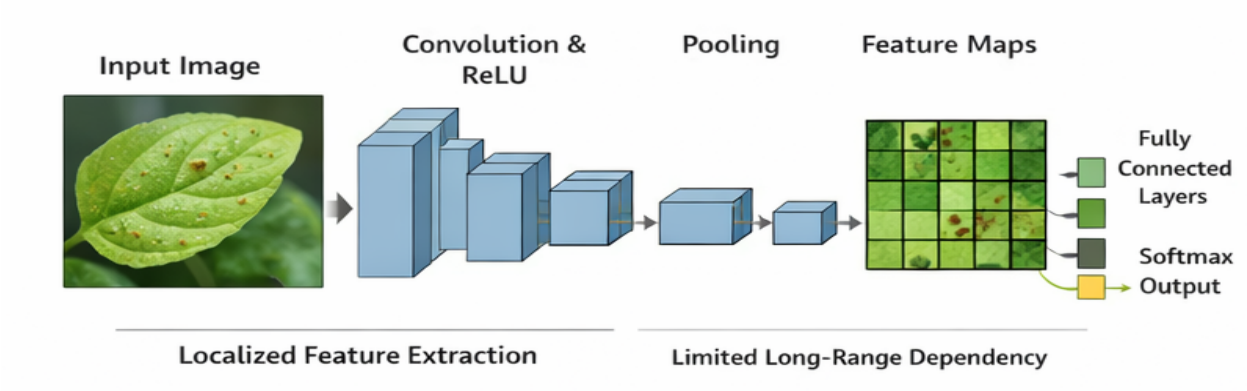


Figure 1: Structure of a Convolutional Neural Network (CNN) used for classifying plant diseases from leaf images

Explanation

A Convolutional Neural Network (CNN) operates by taking different layers of input image and learning essential features gradually from the given image. This begins with the convolutional layer, which detects the basic features of the image (e.g. edges, colors, textures). Moving down the network, the primitive features compose of the complex features such as shapes and the symptoms of a disease. [2]-[6]

A CNN has local receptive fields that are used to identify features in the image. This means that each filter works on a small section of the image at any given time. However, the dimension of the feature maps is decreased by burning pooling layers[23]. This lowers the computational cost as well as the possibility of overfitting.

In spite of the above-mentioned advantages of the CNN model, it is not possible to recognize the relationships between the distant regions of the image because it is only possible for the model to focus on the nearby pixel information. If the model will be used for image-based detection of disease on leaves, this problem arises. [6], [16]

Finally, the extracted features are passed to fully connected layers,than the learned patterns are interpreted, and the image is classified as belonging to a specific disease category.

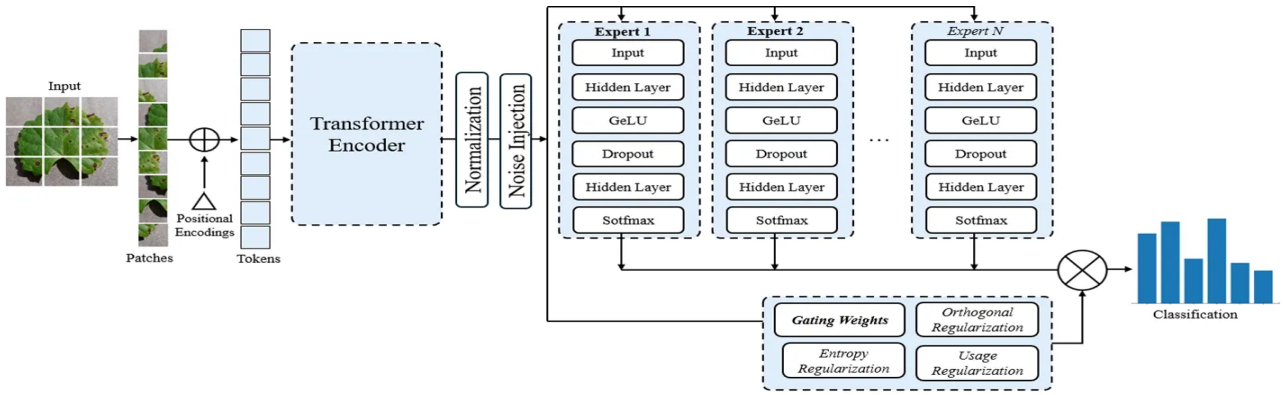


Figure 2: Vision Transformer (ViT) architecture using patch embeddings and self-attention for Macro-scale Pattern Extraction

Explanation

The ViT model does not use the entire image at one go. Instead, the image is subdivided into smaller blocks, e.g. 16 x 16 pixels. "The patches are then linearly embedded into vectors as the way VIT works is different compared to CNN models." [7], [8]

In order to maintain the spatial position of the patches, positional information is added to the embedded patches which makes sure that the model knows the relative position of each patch in the image. The embedded patches are then passed on through a series of transformer encoders.

Each of the encoder's layers is made up of a many-head self-attention and feed-forward networks which helps the model to learn the relations among all the patches in image. Unlike CNNs, the model ViT does not use convolutional filters. It is able to find the relationship between two parts of the image, whether the distance between the two parts is far or not, by using the attention mechanism. [7]-[9], [12]

The ViT model is therefore efficient in details detection in the image. This is especially important in case of plant disease, where the symptoms do not occur on the same leaves [12], [17]-[19].

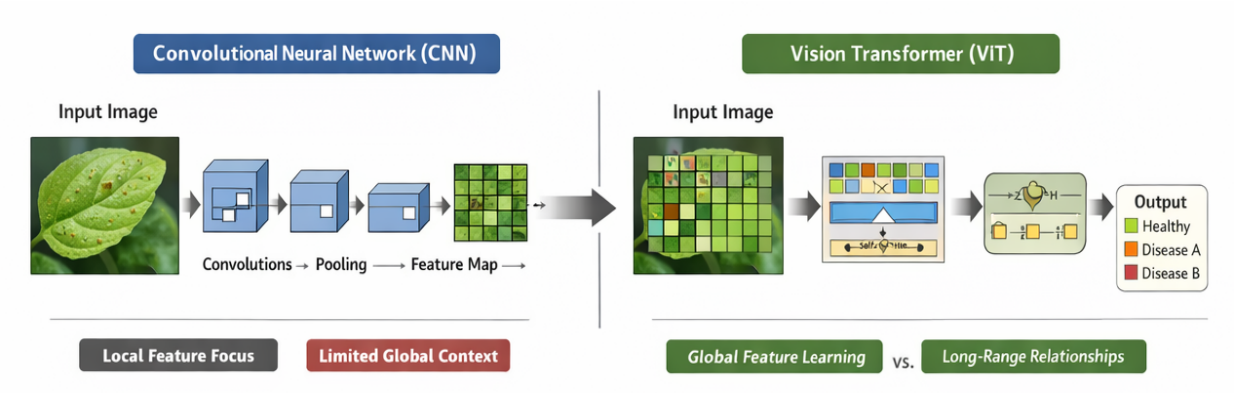


Figure 3: Comparison between CNN and Vision Transformer architectures for feature extraction.

Explanation

The basic differences in how Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) deal with and know an image is shown in Figure 3. CNNs process an image by using small localized regions of the image by passing through a series of features that are obtained through using a filter or a kernel. Features obtained from the image are combined over a series of layers and create the hierarchical representation of the image. This method is very efficient and also it goes fine with understanding the image, identifying features. However, it is not completely efficient in grasping the whole situation of the image. [2]-[9], [16]

Vision transformers treat an image differently than they treat the network by dividing it into smaller regions and treating it like a sequence. Using self-attention mechanism, the model is able to understand what the relation of each part of the image with other parts of the image is. [7]-[9]

As a result of the fundamental difference, there are some advantages of VisualTransformers:

Global Context Awareness: The ViT model has the advantage of looking at the whole image at one while therefore allowing it to understand the relationships between distant parts of the image. The ViT model therefore performs better on images that have complex patterns, and most particularly those that are not regularly distributed, like the patterns on plant diseases so it performs better. Scalability: The ViT model has the advantage that the performance of the model increases with the size of the training data.

There are, however, certain drawbacks to the Vision Transformers. The ViT model demands more data and computational resources as compared to CNN model. [7], [8], [15], [16]

Comparison of Performance of CNN and Deep Learning Algorithms

Such deep learning approaches have the power to extract representation of features on their own, directly from unprocessed image data, using a sequence of non-linear functions in multiple layers. This transformation of computer vision and pattern recognition in food safety has proven particularly useful in the field of plant disease recognition given the variability and dependency that the visual symptoms of plant disease exhibit in their expression and occurrence. A study in 2016 by Sladojevic and colleagues provides an example of how the deep learning techniques were used to achieve classification accuracies of over 96% in images of leaf samples used to identify plant disease. [2]-[5]

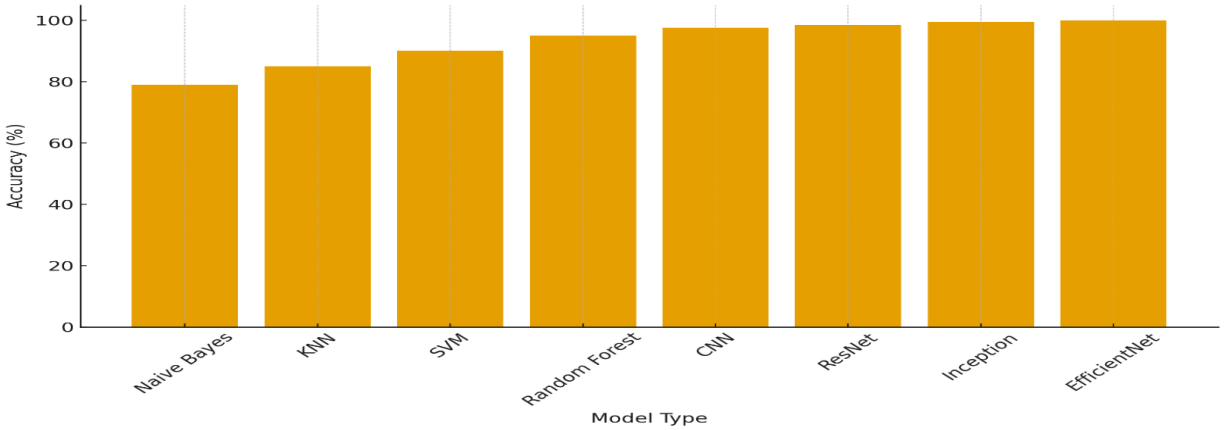


Figure5. Machine Learning V/S Deep Learning Accuracy Comparison

III. METHODOLOGY

Cross-Entropy Loss function being used in the Vision Transformer model, in the training phase. This choice is based on the fact that for a problem of multi-class image classification - such as the detection of plant diseases - the Cross-Entropy Loss function can be used. After the image patches have been fed through the encoder layers of the transformer, the model returns a probability distribution for each class with the help of a softmax layer. This part dwells on the suggested method for identifying plant diseases using the Vision Transformer (ViT) model. The proposed flow of algorithm consists of image preprocessing, patch embedding, encoding and classification [7], [8].

A. Dataset Collection

The dataset taken in this investigation is dataset of plant leaves, which are of two categories, either it is healthy or diseased plant leaf, we want our model to predict this category: plant diseases for which PlantVillage dataset was used in this study [2], [4], [5]. The input data was images of plant leaves, and the types in which they were classified were healthy and diseased plants, among others. Image size: Images are resized up to fixed image size i.e. 224 x 224 images.

B. B. Data Preprocessing

Several operations are done on dataset before considering for the training so it become generalized.

- Let the input data set be represented such as:.

$$D = \{(x_i, y_i)\}_{i=1}^{N}$$

where $x_i \in \mathbb{R}^{H\times W\times C}$is the input image and $y_i$is the corresponding class label and is given as$\mathbb{R}^{H\times W\times C}$

- Resizing: The images are all resized into 224 * 224 pixels.

$$x_i \in \mathbb{R}^{224\times 224\times 3}$$

- Normalization: Normalization involves scaling the pixel values, usually on the value range of 0 to 1.

$$x' = \frac{x-\mu}{\sigma}$$

where $\mu$and σrepresent the mean and standard deviation of the dataset.

- Data augmentation techniques such as rotation, flipping (both horizontally and vertically), zooming, and adjusting brightness is used to strengthen the ability of generalization to the model while also reducing the probability of over fitting the model.

C. Image Patch Generation

Unlike CNNs, Vision Transformer, as the name suggests, treats the images as sequence. [7], [8]
Input image is divided into the fixed size patches (e.g. 16 x 16) Each patch is flattened out into a vector
The input image is profiled into non-overlapping image patches size P x P. The total number of patches is given by:

$$N_p = \frac{H\times W}{P^2}$$

For $H = W = 224$ and $P = 16$:

$$N_p = \frac{224 \times 224}{16^2} = 196$$

Each patch is flattened into a vector:

$$x_p \in \mathbb{R}^{(P^2 \cdot C)}$$

D. Patch Embedding and Positional Encoding

- Patch Projection: Input images are divided into non-overlapping patches, which are flattened and linearly projected into fixed dimensional vectors.ArrayNetwork Module This flattens and crops patches from the input network. ModelNetwork Module This module is used to instruct each layer of the model of the network.
- Learnable Embedding: A learnable linear mapping is applied to these flattened patches to generate the patch embeddings.
- CLS Token Incorporation: Incorporate some learnable classification ([CLS]) token in the sequence. imagesoyou look at the entire image and do tasks such as classification using this token.Regardless of how you choose to use patch embeddings (here are a few options:
- Positional Embedding: By adding learnable positional embeddings to patch embeddings, the spatial information is preserved and model will understand exactly All patch positions within an image are important.

Each patch is projected into a latent space by applying a linear transformation:

$$z_0 = [x_{class}; x_p^1 E; x_p^2 E; \dots ; x_p^{N_p} E] + E_{pos}$$

where:

$E$is the learnable embedding matrix

$x_{class}$is the classification token

E_{pos}represents positional embeddings

E. Transformer Encoder

The embedded sequence is passed through Ltransformer encoder layers. Each layer consists of Multi-Head Self-Attention (MHSA) and Feed Forward Network (FFN). [7], [8]

1) Multi-Head Self-Attention

The attention mechanism is defined as:

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

where:

$Q, K, V$are query, key, and value matrices

$d_k$is the dimension of the key vectors

Multi-head attention is computed as:

$$\text{MHSA}(Z) = \text{Concat}(head_1, \dots, head_h)W^O$$

2) Feed Forward Network

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2$$

3) Encoder Layer Update

Each encoder layer is defined as:

$$Z' = \text{MHSA}(\text{LN}(Z)) + Z$$
$$Z'' = \text{FFN}(\text{LN}(Z')) + Z'$$

where LNdenotes Layer Normalization.

F. Classification Head

The final representation is obtained from the classification token:

$$y = \text{Softmax}(Z_L^0 W_c + b_c)$$

where

$Z_L^0$is the output corresponding to the [CLS] token

$W_c, b_c$are classification weights

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

G. Loss Function

For the Vision Transformer model, the Cross-Entropy Loss function is used for training the model. This is because the Cross-Entropy Loss function is ideal for multi-class image classification problems like plant disease detection.[20].After passing image patches through encoder layers of the transformer architecture, the model outputs the probability distribution for all classes using a softmax layer.

$L = -1/N \sum_i y_i \log(\hat{y}_i)$[20]

where

$y_i$represents the ground truth label for the i^{th}class.

$\hat{y}_i$indicates predicted probability assigned by the ViT model to that class.

N is the number of instances

H. Model Training

Parameters used for the training of the model are as follows:

- Loss Function: Cross-Entropy Loss
- Optimizer: Adam / AdamW
- Learning Rate: Tuned (e.g., 1e-4)
- Batch Size:32 /64 /256
- Epochs: 15

Regularization techniques:

- Dropout

Early stopping

The model parameters are optimized using the AdamW optimizer:

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta \mathscr{L}$$

where:

$\eta$is the learning rate

θrepresents model parameters

I. Evaluation Metrics

The performance of the model is evaluated using:

1) Accuracy:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

2) Precision:

$$\text{Precision} = \frac{TP}{TP + FP}$$

3) Recall:

$$\text{Recall} = \frac{TP}{TP + FN}$$

4) F1-Score:

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

J. Algorithm: Plant Disease Detection using ViT

- Input leaf image
- Resize and normalize image
- Divide image into patches
- Generate patch embeddings
- Add positional encoding
- Pass through transformer encoder layers

- Extract [CLS] token output
- Apply classification head
- Output predicted disease class

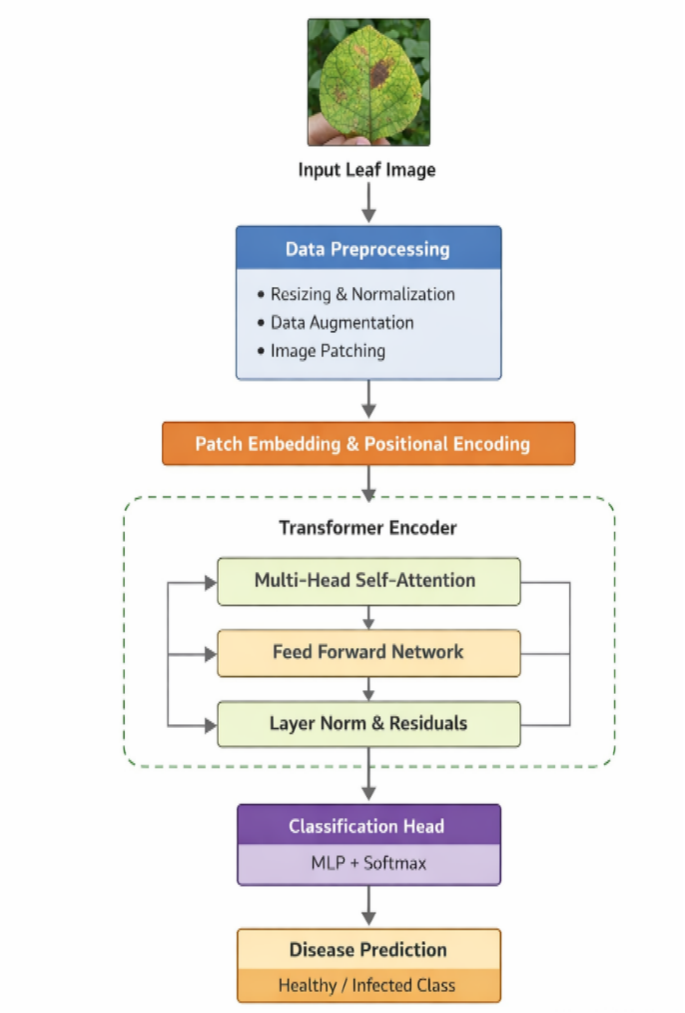


Figure 6. Flow of algorithm

IV. RESULTS AND DISCUSSION

A. *Model Performance*

The Vision Transformer(ViT) model was trained for more than 15 epochs and the accuracy and loss were used in performance measurement. The accuracy from the initial stages and upto reaches to final epoch it reached to almost 99.8%. This indicates that the model has learned very well which shows continues learning. Simultaneously the value of the loss function is decreasing and indicating that the model is learning to minimize predictions error. Based on the results, it appears that the ViT model has managed to learn from the plant leaf images. The ability of the model to detect small details and general patterns in the images enables high accuracy and reliable classification results.

B. *Accuracy and Loss Analysis*

*Accuracy:*A considerable jump in accuracy is observed in the early stages of training which reaching a level of 99.8%. The model learning process is obviously going well.

Loss: The loss function was continuously getting smaller with the number of training epochs and that loss function values came closer to zero. This is a good sign for minimizing loss.

Observation: There is no significant sign of overfitting which is good indication. The smoothness of the curves of accuracy and loss can give an idea of how well the model is balanced and how good are the chosen hyperparameters are.

C. Confusion Matrix Analysis

The confusion matrix provides a detailed picture of how well the classification is done for different classes

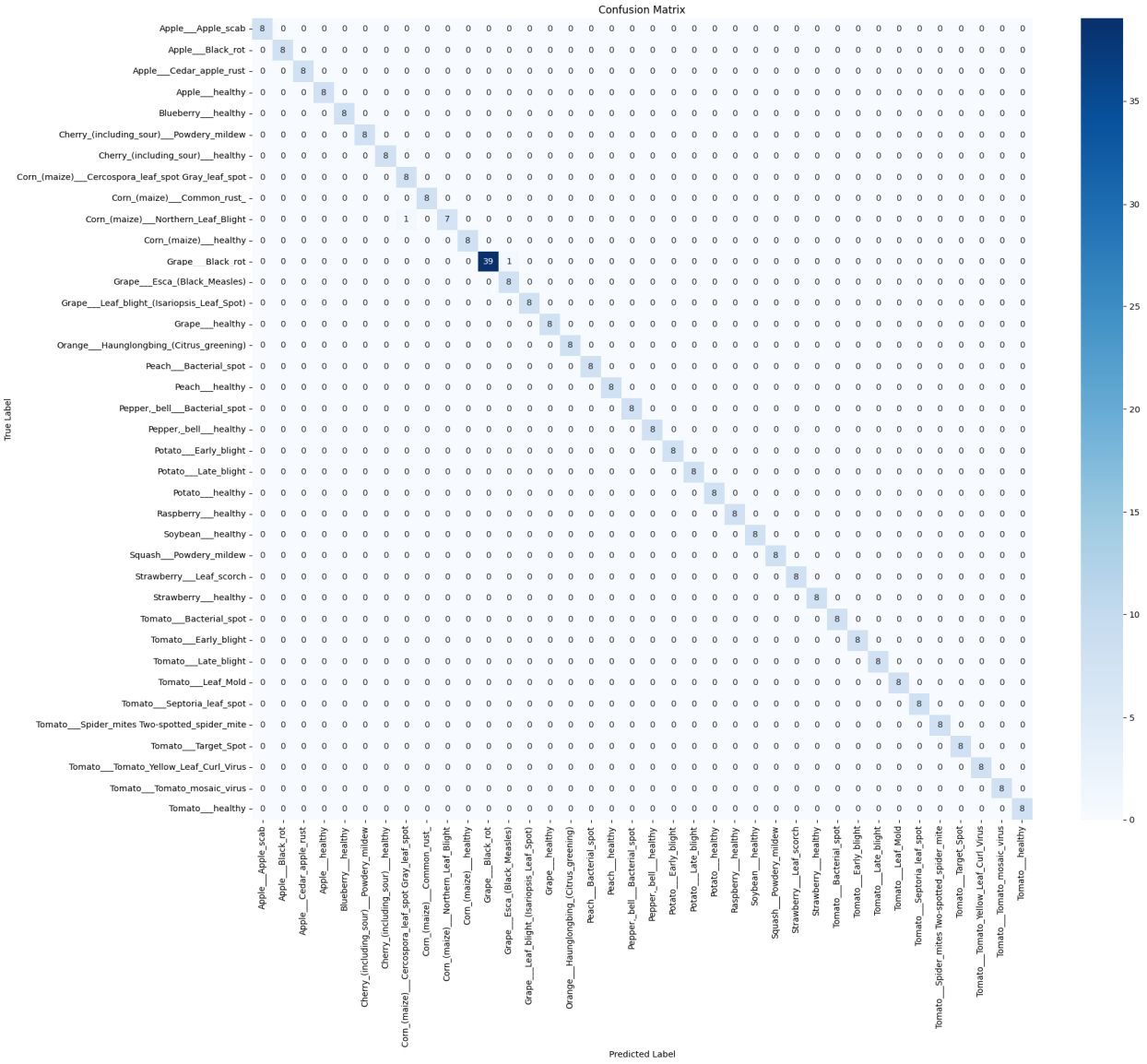


Figure 7. Confusion matrix of Proposed VIT Model

Key Observations:

- Most of the samples fall into the right class (high diagonal values)
- Very few misclassifications are observed
- Model is good at discriminating between similar disease classes
- Example:
- Healthy class accuracy is around 98%
- Accuracy of disease classes is around 97-98%
- Misclassification is very low (1-3%)

D. Performance Evaluation of the Proposed Model

The model's effectiveness is validated through 99.4% accuracy and near-perfect (0.97–1.00) precision/recall across classes, confirming highly accurate plant disease classification.

- Accuracy (Testing Accuracy 99.4%): High overall classification efficiency.
- Precision (≈1.00): Minimal false predictions of disease presence.
- Recall (0.97–1.00): Excellent detection of existing diseases.
- F1-Score (≈1.00): F1-Score is a balanced measure between accuracy and precision in detecting different diseases.

--- Classification Report ---

	precision	recall	f1-score	support
Apple___Apple_scab	1.00	1.00	1.00	8
Apple___Black_rot	1.00	1.00	1.00	8
Apple___Cedar_apple_rust	1.00	1.00	1.00	8
Apple___healthy	1.00	1.00	1.00	8
Blueberry___healthy	1.00	1.00	1.00	8
Cherry_(including_sour)___Powdery_mildew	1.00	1.00	1.00	8
Cherry_(including_sour)___healthy	1.00	1.00	1.00	8
Corn_(maize)___Cercospora_leaf_spot Gray_leaf_spot	0.89	1.00	0.94	8
Corn_(maize)___Common_rust_	1.00	1.00	1.00	8
Corn_(maize)___Northern_Leaf_Blight	1.00	0.88	0.93	8
Corn_(maize)___healthy	1.00	1.00	1.00	8
Grape___Black_rot	1.00	0.97	0.99	40
Grape___Esca_(Black_Measles)	0.89	1.00	0.94	8
Grape___Leaf_blight_(Isariopsis_Leaf_Spot)	1.00	1.00	1.00	8
Grape___healthy	1.00	1.00	1.00	8
Orange___Haunglongbing_(Citrus_greening)	1.00	1.00	1.00	8
Peach___Bacterial_spot	1.00	1.00	1.00	8
Peach___healthy	1.00	1.00	1.00	8
Pepper,_bell___Bacterial_spot	1.00	1.00	1.00	8
Pepper,_bell___healthy	1.00	1.00	1.00	8

Figure 7 Performance Evaluation of the Proposed VIT Model

- Train Accuracy: Increases rapidly from 91% to 98%, tend settled around 99.5% suggests strong learning.
- Val Accuracy: Comes down quickly to 99 percent.5%, following closely with training quite high generalize capacity.
- Train Loss goes down rapidly, then gradually decline; indicates continuous improvement.
- Val Loss stays and not too high (0.01 to 0.02 means there is no overfitting Overall Observation

The model exhibits robust performance in terms of accuracy and stability, validating its suitability for real-time plant disease detection systems.

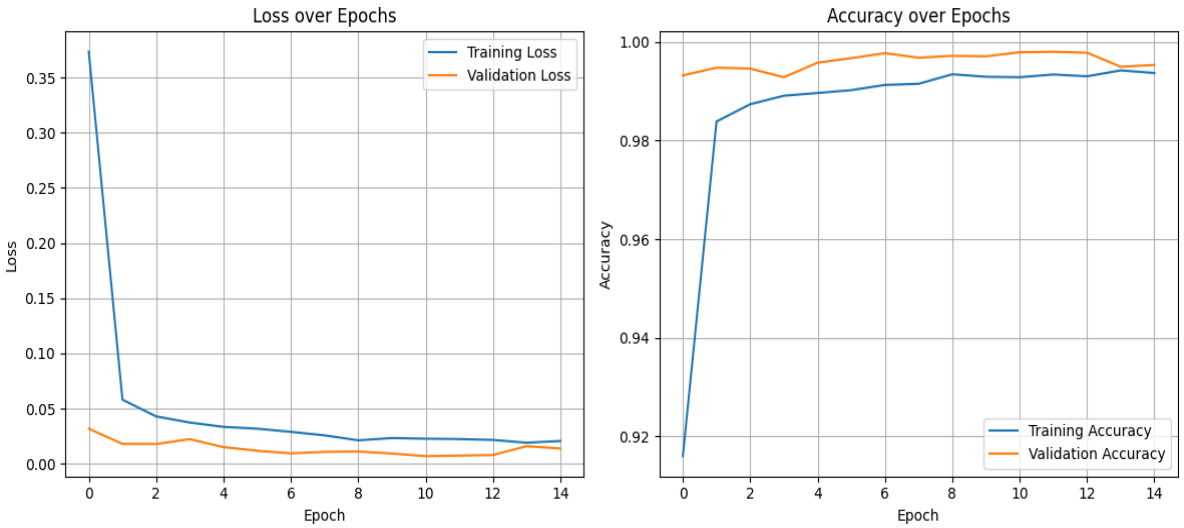


Figure 8. Epoch wise Loss and Accuracy Graph

E. Discussion

The results obtained from the experiments show that the Vision Transformer model can work better for the detection of plant diseases by comparing other approaches that use CNN. The ability of Vision Transformer model to comprehend all of relationships present in certain image to identify complex and diverse plant diseases more accurately.

When compared to other CNN-based approaches, the following advantages can be accomplished by using the proposed model:

- Higher accuracy, specifically for complex and different images.
- Better ability to generalize,taking into considerationdifferent environmental conditions.
- Higher robustness, for noisy images with diverse backgrounds.

On the other hand, some limitations can be obtained from the proposed model, as follows:

- Higher computational requirements, particularly for training and testing the model
- Higher requirements for larger datasets, particularly for optimal performance
- Even though some limitations can be obtained from the proposed model, the results obtained from the experiments show that the model is highly effective and can be applied for practical use.

F. Conclusion of Results

The accuracy of the model based on the basis of Vision Transformer (ViT) model is almost 100% with the classification accuracy of 99.8%. This indicates the high efficiency of the model in the detection of plant diseases with high accuracy. This is a good sign of the potential of the model for a very considerable contribution to the development of modern systems of agriculture.

V. CONCLUSION

The results of the present investigation were able to utilize a Vision Transformer (ViT) model for plant disease identification that could overcome the limitations of conventional CNNs through the application of self-attention mechanisms, which led to an accuracy score of 99.8% coupled with high precision, recall, and F1-scores. This high-performing and easy-to-use instrument for early detection helps to improve the practice of precision agriculture. Future research efforts will, however, aim to overcome the huge computational demands and add different types of data such as environmental and climatic variables. Consequently, ViT based models offer a sustainable and cutting edge solution for strengthening the issue of food security across the globe.

VI. FUTURE SCOPE

The promising results obtained with the model based on Vision Transformer has also paved the way for further research and improvements in the plant disease detection systems encourage for further development.

A. Real-Time Field Deployment:

The model may be further improved for the purpose of real-time detection of disease occurrence in the agricultural field by sending drones, IoT devices and mobile applications with light models.

B. Integration with Multimodal Data:

Recognize the Evaluation AND Such Processes. A Turn It Into: Open to High-level Commitment, Integration with Multimodal Data: The model might predict better if multi modal data such as weather, soil, temperature, light etc. integrated by using IOT devices.

C. Hybrid and Advanced Architectures:

Various hybrid and advanced architectures had been proposed: - Ensemble and Transfer learning approach with CNN and VIT model may increase the performance so more scope can be seen in this direction.

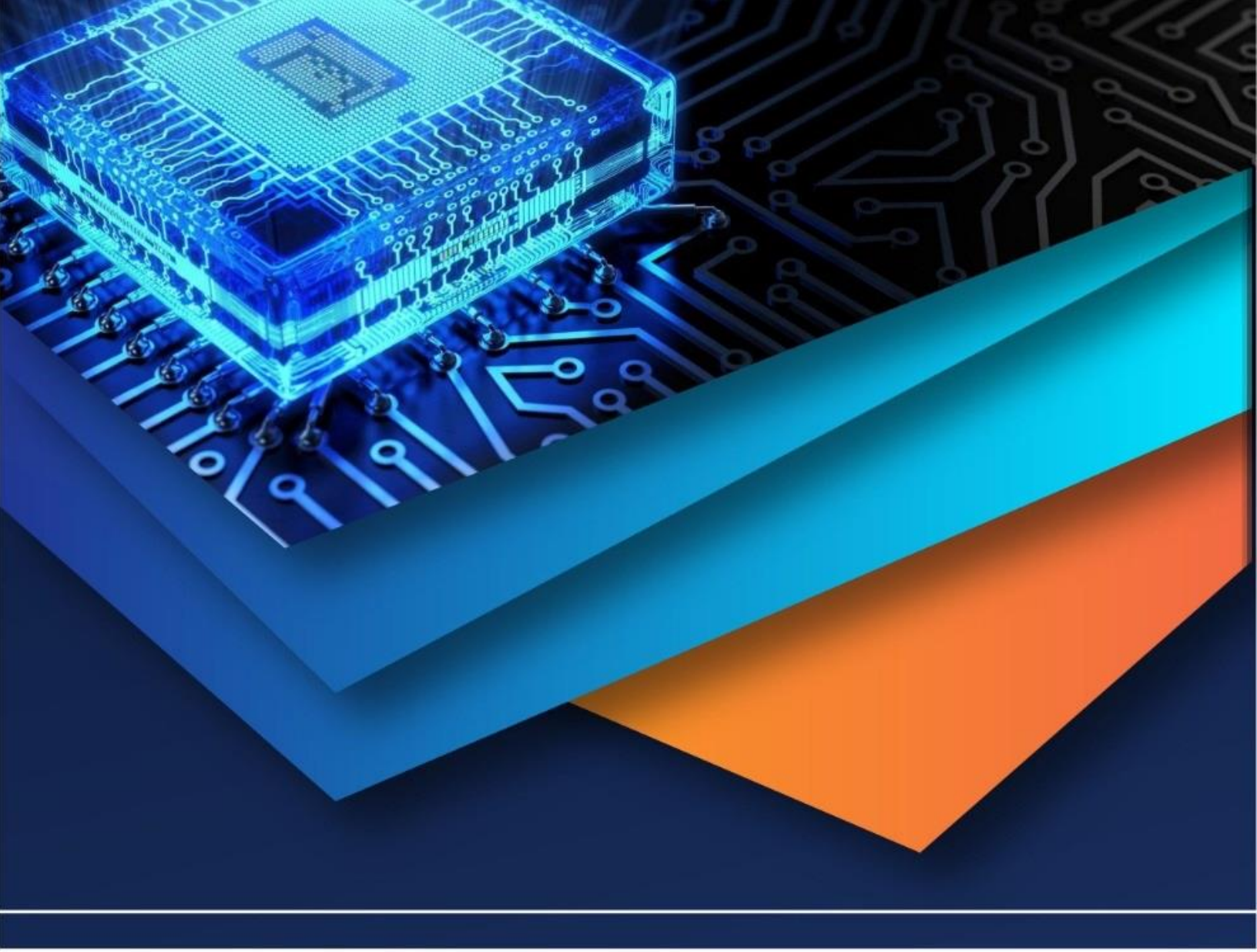
Thus, the Models efficiency, scalability and applicability of transformer-based plant disease detection systems in real applications in thereal-world agricultural environments.

REFERENCES

[1] S. Savary et al., “The global burden of pathogens and pests on major food crops,” Nature Ecology & Evolution, vol. 3, no. 3, pp. 430-439, 2019, doi: 10.1038/s41559-018-0793-y.

[2] S. P. Mohanty, D. P. Hughes, and M. Salathé, “Using deep learning for image-based plant disease detection,” Frontiers in Plant Science, vol. 7, Art. no. 1419, 2016, doi: 10.3389/fpls.2016.01419.

[3] S. Sladojevic, M. Arsenovic, A. Anderla, D. Culibrk, and D. Stefanovic, "Deep Neural Networks Based Recognition of Plant Diseases by Leaf Image Classification," Computational Intelligence and Neuroscience, vol. 2016, Art. no. 3289801, 2016, doi: 10.1155/2016/3289801.
[4] K. P. Ferentinos, "Deep learning models for plant disease detection and diagnosis," Computers and Electronics in Agriculture, vol. 145, pp. 311-318, 2018, doi: 10.1016/j.compag.2018.01.009.
[5] E. C. Too, Y. Li, S. Njuki, and Y. Liu, "A comparative study of fine-tuning deep learning models for plant disease identification," Computers and Electronics in Agriculture, vol. 161, pp. 272-279, 2019, doi: 10.1016/j.compag.2018.03.032.
[6] A. Abade, P. A. Ferreira, and F. de B. Vidal, "Plant diseases recognition on images using convolutional neural networks: A systematic review," Computers and Electronics in Agriculture, vol. 185, Art. no. 106125, 2021, doi: 10.1016/j.compag.2021.106125.
[7] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in Proc. Int. Conf. Learn. Representations (ICLR), 2021.
[8] H. Touvron et al., "Training data-efficient image transformers & distillation through attention," in Proc. 38th Int. Conf. Machine Learning (ICML), vol. 139, pp. 10347-10357, 2021.
[9] Z. Liu et al., "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2021, pp. 9992-10002, doi: 10.1109/ICCV48922.2021.00986.
[10] M. Shoaib et al., "An advanced deep learning models-based plant disease detection: A review of recent research," Frontiers in Plant Science, vol. 14, Art. no. 1158933, 2023, doi: 10.3389/fpls.2023.1158933.
[11] A. Jafar, N. Bibi, R. A. Naqvi, A. Sadeghi-Niaraki, and D. Jeong, "Revolutionizing agriculture with artificial intelligence: plant disease detection methods, applications, and their limitations," Frontiers in Plant Science, vol. 15, Art. no. 1356260, 2024, doi: 10.3389/fpls.2024.1356260.
[12] S. Hemalatha and J. J. B. Jayachandran, "A Multitask Learning-Based Vision Transformer for Plant Disease Localization and Classification," International Journal of Computational Intelligence Systems, vol. 17, Art. no. 188, 2024, doi: 10.1007/s44196-024-00597-3.
[13] U. Barman et al., "ViT-SmartAgri: Vision Transformer and Smartphone-Based Plant Disease Detection for Smart Agriculture," Agronomy, vol. 14, no. 2, Art. no. 327, 2024, doi: 10.3390/agronomy14020327.
[14] A. K. Singh et al., "Effective plant disease diagnosis using Vision Transformer trained with leafy-generative adversarial network-generated images," Expert Systems with Applications, vol. 254, Art. no. 124387, 2024, doi: 10.1016/j.eswa.2024.124387.
[15] M. Xu, J.-E. Park, J. Lee, J. Yang, and S. Yoon, "Plant disease recognition datasets in the age of deep learning: challenges and opportunities," Frontiers in Plant Science, vol. 15, Art. no. 1452551, 2024, doi: 10.3389/fpls.2024.1452551.
[16] A. Upadhyay et al., "Deep learning and computer vision in plant disease detection: a comprehensive review of techniques, models, and trends in precision agriculture," Artificial Intelligence Review, vol. 58, Art. no. 92, 2025, doi: 10.1007/s10462-024-11100-x.
[17] S. Murugavalli and R. Gopi, "Plant leaf disease detection using vision transformers for precision agriculture," Scientific Reports, vol. 15, Art. no. 22361, 2025, doi: 10.1038/s41598-025-05102-0.
[18] S. Yu, L. Xie, and L. Dai, "ST-CFI: Swin Transformer with convolutional feature interactions for identifying plant diseases," Scientific Reports, vol. 15, Art. no. 25000, 2025, doi: 10.1038/s41598-025-08673-0.
[19] P. S. Roy and V. Kukreja, "Vision transformers for rice leaf disease detection and severity estimation: a precision agriculture approach," Journal of the Saudi Society of Agricultural Sciences, vol. 24, no. 3, pp. 1-15, 2025, doi: 10.1007/s44447-025-00007-w.
[20] Z. Salman, A. M. Muhammad, and D. Han, "Plant disease classification in the wild using vision transformers and mixture of experts," Frontiers in Plant Science, vol. 16, Art. no. 1522985, 2025, doi: 10.3389/fpls.2025.1522985.
[21] S. Aboelenin, F. A. Elbasheer, M. M. Eltoukhy, W. M. El-Hady, and K. M. Hosny, "A hybrid framework for plant leaf disease detection and classification using convolutional neural networks and vision transformer," Complex & Intelligent Systems, vol. 11, Art. no. 142, 2025, doi: 10.1007/s40747-024-01764-x.
[22] A. Vyas, D. Patel, I. Kalal, and B. Patel, "Agro-Detect: A CNN Driven Early Detection of Leaf Diseases," International Journal of Innovative Science and Research Technology, vol. 10, no. 7, pp. 855–862, Jul. 2025, doi: 10.38124/ijisrt/25jul707.
[23] A. Punitha, S. Syedakbar, and S. Jeyasudha, Eds., Intelligent and Sustainable Systems: AI, Green IoT, and Adaptive Automation in Electrical and Communication Technologies, 1st ed. Boca Raton, FL, USA: CRC Press, 2026. doi: 10.1201/9781003773801.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY