Optimized AI Solution for Plant Disease Diagnosis Using Enhanced Vision Transformer for Smarter Agriculture

B.Lanitha, Akil.M, Nishanth.M
Associate Professor Karpagam Academy of Higher education
UG scholar, Karpagam Academy of Higher education
lanitha.nandakumar@kahedu.edu.in

ABSTRACT

Detecting plant diseases quickly and accurately is essential to increasing agricultural output over the long run. The biggest threat to crop productivity is disease infection, which can result in financial losses. The roots, stems, and leaves of plant parts that can be impacted by fungi, viruses, bacteria, and other infecting organisms. The conventional methods for identifying plant diseases are labor intensive, resource intensive, and time consuming. As a result, the analysis and detection of automatic leaf disease detection using Artificial Intelligence (AI) and Vision Transformer (ViT) techniques is taken into consideration. This work suggests a method for differentiating between healthy and disease-ridden plants using the ViT model. By combining an enhanced multi-head self-attention mechanism with gate, the suggested method improves its capacity to extract crucial aspects of images. Deep Leaning Networks' (DNN) interpretive and classification skills are combined with ViT's ability to use a sequence of patches for processing images. An extensive set of leaf images classified into various disease categories to train and evaluate the model. The combination of deep neural networks and vision transformers provides a stable solution with high interpretability and accuracy, which helps to ensure the long-term production of this important crop. The experiment's results demonstrated that the proposed model outperformed other contemporary techniques and had an incredible accuracy rate. Through a comprehensive grid search, key hyper parameters such as batch size, learning rate, and loss function were optimized. Important hyper parameters like the number of batches, learning rate, and loss function were optimized through an extensive grid search. In summary, this work addresses the present agricultural demands for accurate and scalable methods by presenting a novel and effective approach to identifying leaf diseases. The study summarizes the experiment's results and provides recommendations and future directions for research into the use of ViT in the diagnosis of agricultural diseases.

Keywords: Image Classification, Multi-Head Self-Attention, Vision Transformer

1 INTRODUCTION

Diseases in plants can significantly affect the health of crops, resulting in considerable harm, lower production, and monetary deficits for farmers. Dealing with these diseases is crucial to ensure the availability of food and maintain sustainable agriculture. They include a wide range of diseases, infections, and anomalies that have an adverse effect on the growth and well-being of plants. Bacteria, nematodes, viruses, fungi or environmental elements like inadequate nutrition or unfavorable weather can all lead to them [1]. Effective disease management and prevention strategies depend heavily on the prompt and accurate plant disease detection. Traditional techniques for identifying plant diseases mainly rely on expert visual inspection carried out by manually. Nevertheless, these approaches have a number of shortcomings, such as being unreliable, exhausting, and prone to human error. Automation of plant disease detection systems have demonstrated significant potential in addressing the drawbacks of conventional techniques in the last several years [2]. Relying solely on the farmer's expertise to visually inspect and evaluate crops in traditional ways poses various challenges and limitations in agricultural research. If a crop infection goes unnoticed, it could lead to the entire crop failing and causing a decrease in yield.

In order to tackle the issues found in modern agriculture, computer-assisted automated research like Machine Learning(ML) and Deep Learning(DL) can play a key role in quickly and accurately identifying diseases. AI is growing in significance in agricultural studies, especially in the recognition and categorization of plant diseases. Implementing AI methods in farming can cut down on labor expenses, minimize inefficiencies in time, and improve both crop quality and overall production. Utilizing suitable management approaches can help implement disease control plans by utilizing early data on crop health and disease location. The initial step in this process is classification, which includes dividing data into categories. In this scenario, the focus is on identifying and categorizing plant leaves, with a specific aim to distinguish between healthy and diseased specimens. In order to execute, one must be aware of the categorization algorithms for deep learning and machine learning identification. Automated technologies are crucial for identifying plant diseases right now. They help reduce the occurrence of crop diseases and the resulting losses. The AI-powered disease detection system follows set steps automatically. The process includes multiple steps, such as placing different sensors in the farm to gather and save images of the plants.

The collected pictures are subsequently analyzed and divided into sections for utilization in algorithms in machine learning . The ML models subsequently determine if a leaf is in good condition or has a disease [3]. A framework is introduced in Figure 1, outlining predefined steps for predicting plant diseases.

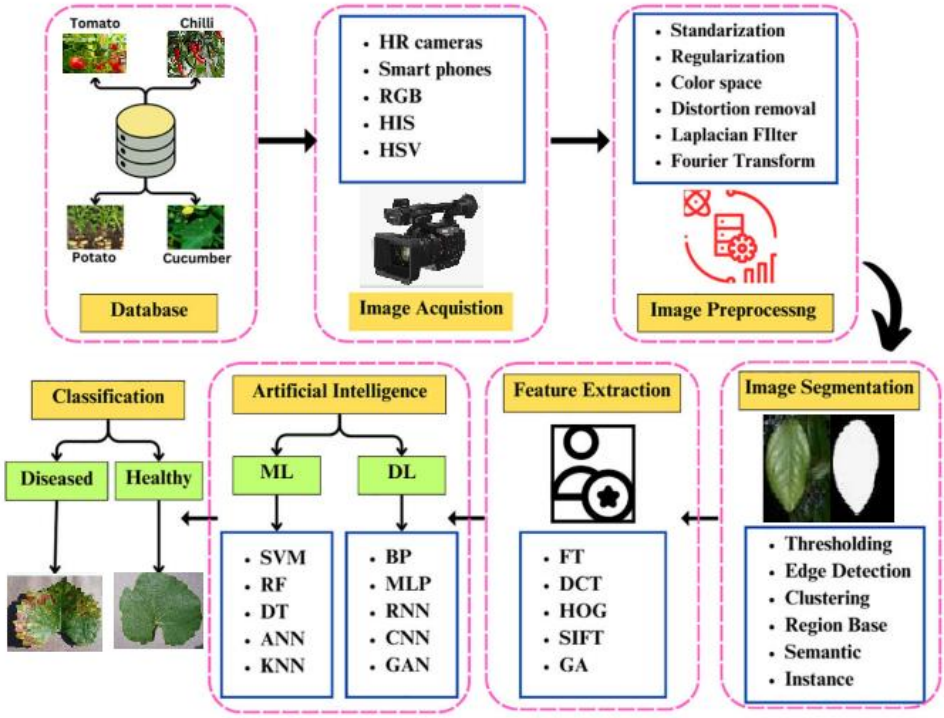


Figure1 Plant diseases prediction system with important steps

In general, the ViT model was executed based on Figure 2. The ViT model's performance was remarkable. sometimes surpassing traditional convolutional neural networks (CNNs).

It is capable of handling pictures with a fixed size patch set. allowing for scalability to larger image sizes. The basic ViT model operates by following these steps.

- Standardizing input image: A set of blocks with fixed sizes is created from the input image. flattened linearly to form a one-dimensional vector, which serve as input tokens for the model.
- Embedding: Each patch is transformed into a vector representation with fixed dimensional known as an embedding.

- Encoder of a transformer: The Transformer Encoder serves as the foundation of the ViT paradigm. It is made up of several layers, with feed forward neural network with self-attention building blocks
- Categorization: The transformer processes the data to form a feature vector, which is then forwarded to a fully connected CNN layer for classification.
- TrainingWeight changes in the transformer and embedding layers are fine-tuned through optimization methods such as the Adam optimizer and stochastic gradient descent during training.
- Interpretation: The trained model receives input, processes it through a forward pass, and produces an outcome based on its output probabilities

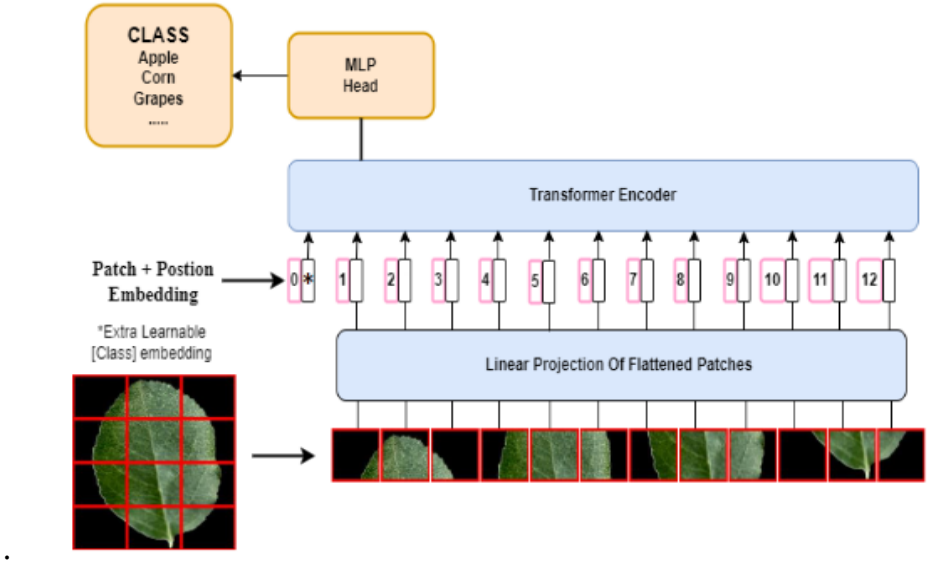


Figure 2 ViT model for detection of plant disease

A recent development in computer vision, ViTs use the Transformer architecture, originally developed for natural language processing to tackle image based tasks.[4]

2 RELATED WORKS

ViTs have benefits compared to multitask learning(MTL) frameworks for the detection of plant diseases. They gather information from both the surrounding area and the wider world, allowing for precise examination of images of plant . ViTs perform well with small amounts of labeled data by pre-training on large datasets such as ImageNet. Furthermore, they offer adaptability and scalability for various disease detection assignments using just one model, thus minimizing complexity and computational needs. Thakur[5] presented PlantViT, a combined model for automating identification of plant diseases merges the potential of a ViT and a CNN, including a multi head attention component. Testing was done on the model on two large databases for the detection of plant diseases specifically Plant Village. The model achieved detection accuracies of 87.87% on the Embrapa dataset and 98.61% on the PlantVillage dataset in the experimental results. PlantViT demonstrates superior performance compared to existing methods, showcasing its effectiveness in detecting plant diseases.

In [6], a method using the ViT architecture for quick automated classification of plant diseases is suggested. The research thoroughly analyzes different models such as ViT, conventional CNNs, and a blend of ViT and CNN. Thorough training and assessment on various datasets demonstrates how attention blocks increase precision but also cause delays in predictions. Nevertheless, combining CNN and attention

blocks achieves a perfect equilibrium between precision and efficiency, providing a reliable answer for instantaneous plant disease categorization.

Recent research has utilized deep learning models to recognize leaf diseases of cassava, replacing farmers' unreliable and traditional intuition-based analysis. Thai et al. [7] investigated utilizing the ViT model rather than a CNN to improve the precision of disease detection in these species. The ViT model has been shown in experiments to achieve high accuracy, outperforming well-known CNN models like Resnet50 and EfficientNet on the data set of disease of Cassava Leaves. These results emphasize the possible advantage of the ViT model in examining diseases in leaf.

Thakur et al. [8] conducted a thorough investigation into the use of ViTs in plant pathology, focusing on the development of PlantXViT, a CNN-based model for effectively detecting plant diseases in different crops, ideal for smart agriculture using IoT technology. Testing on five datasets demonstrates that PlantXViT surpasses the best existing models, with mean accuracies exceeding 98.33% , 92.59%, and 93.55%, on Rice datasets Maize and Apple respectively, even under difficult conditions. The model's interpretability is also evaluated through class activation maps that are gradient weighted, showcasing its capacity to offer understanding of the process for making decisionse.

Recent advances in deep learning have proven to be extremely successful in several computer vision tasks, such as identifying objects and classifying images. A significant advancement in technology is the ViT architecture [9], that has exhibited outstanding results in large scale classification assignments for images [10-12]. In contrast to CNNs frequently employed in tasks involving computer vision,ViT relies on self attentional mechanisms to capture overall connections between patches of image and effectively produce representations [13].

ViTs have been employed in carrying out various tasks at the same time by distributing data among different tasks in recent years. The UniT model is suggested for learning different tasks in various fields like object recognition, natural language processing, and multimodal processing simultaneously. The UniT uses an encoder to encode input modalities, then uses a shared decoder to generate task predictions based on the encoded input representations. Utilizing output heads specific[14] to each task, and the entire model has been trained collectively from start to finish with task losses. UniT stands out by utilizing common model parameters for every tasks, supporting a variety of tasks in different areas, and providing strong performance on 7 tasks across 8 datasets with minimized.

Likewise, the MulT framework suggests a Transformer based on end-to-end MTL for concurrent training of various complex visual tasks [15]. It surpasses single task Transformers and multitask CNNs , showing the importance of having shared focus on various tasks. Resilience is demonstrated by the MulT model and the ability to adapt to new areas in evaluations of multitasking. The research emphasizes the benefits of Transformer structures in multitask learning. These advantages include increased scalability, better representation of features and the capability to model intricate relationships between the tasks. This significant contribution to the field opens up avenues for future investigation and opportunities in domains that demand reliable multitask learning solutions.1[16].

Tian [16] presented the MultiTask ViT (MTViT), a technique for learning representations that utilizes ViTs. MTViT presents a new kind of transformer with multiple branches to process image patches for various tasks in a sequential manner, enabling communication between them via the attention module with Cross

Task. In contrast to previous models, MTViT improves memory and computational efficiency by extracting features using ViT's self-focus mechanism. Results from experiments on standards datasets show that MTViT performs either comparably or better than current CNN-based MTL techniques.

A combined network called the Lightweight based Shuffle Convolution Vision Transformer (SLViT)[17] was created to accurately diagnose diseases affecting sugarcane leaves. The study shows that SLViT outperforms the most advanced algorithms on the openly accessible Plant Village collection in terms of, weight, and peed, precision memory usage. SLViT integrates a lightweight CNN architecture with an adaptable Transformer encoder plug-inA ViT-based lightweight method called ConvViT [18] was created for the identification of apple leaf diseases. It enhances overall robustness and accuracy by integrating transformer and convolutional architectures to efficiently record both local and global features of disease spots on crops. To maintain edge insights and enable data exchange within the transformer, the patch embedding technique is enhanced. ConvViT achieves significant resource reduction by utilizing linear-complexity multiple-head attention operations and depth wise independent convolution. Significantly with FLOPs (21.7%) and fewer parameters (32.7%) , this model attains identification performance (96.85%) on a complicated apple leaf disease dataset. This demonstrates ConvViT's efficacy and usefulness as a disease identification model.

3 MATERIALS AND METHODS

Deep learning and machine learning have recently developed into effective methods for computerized detection of plant diseases. Deep neural networks, which are capable of deciphering complex patterns from large amounts of images, are employed in these methods. This study's suggested method for using ViT to identify diseases of plants involves several crucial steps. First, a collection of photos of leaves from plants is assembled, including examples of both healthy leaves and diseased. A plant village dataset was used in this study. The image data underwent rudimentary preprocessing, and the ViT models were trained to identify intricate patterns in the images. The model is able to classify recently acquired images into one of the disease classes or a healthy class soon after it has been trained. The optimal model was put into use for real time evaluation after the process.

3.1 Dataset

The information was gathered from the plant village data set. This data set [19] contains 54,306 samples of leaves that are either in good or poor health, s**o**rted into 38 distinct groups based on the type of plant and the presence of disease. The breakdown of the types of plants used for training and evaluation is detailed in Table 1. A selection of notable leaf images from the set is displayed in Figure 3.

3.2 Details of Implementation

The image processing and deep learning toolboxes in Matlab 2023a are used to implement this model. The i5 processor used for training and testing has GB of DDR4 RAM. This configuration guarantees precise plant disease identification and classification.

Apple : Scab

Grape:Black Measles

Corn: Common rust

Apple: Healthy

Grape :Healthy

Corn : Healthy

Figure 3: Leaf images from Plant Village dataset.

Table 1 Class distribution of data set in plant village

Leaf	**Training**	**Testing**	**Infection**	**Leaf**	**Training**	**Testing**	**Infection**
Apple	630	630	Scab	Potato	1000	1000	Early-Blight
	621	621	Black_rot		152	152	Healthy
	275	275	Rust		1000	1000	Late-Blight
	1645	1645	Healthy	Raspberry	371	371	Healthy
Blackberry	1502	1502	Healthy	Soybean	5090	5090	Healthy
Cherry	854	854	Healthy	Squash	1835	1835	Powdery mildew
	1052	1052	Powdery_ mildew	Strawberry	456	456	Healthy
Corn	513	513	Gray leaf spot		1109	1109	Leaf_scorch
	1192	1192	Common rust	Tomato	2127	2127	Bacterial Spot
	1162	1162	Healthy		1000	1000	Early Blight
	985	985	Leaf-Blight		1591	1591	Healthy
Grape	1180	1180	Black rot		1909	1909	Late-Blight
	1384	1384	Black Measles		952	952	Leaf-Mold
	423	423	Healthy		1771	1771	Leaf-Spot
	1076	1076	Leaf-Blight		1676	1676	Spider-Mite
Citrus	5507	5507	Greening		1404	1404	Target-Spot
Peach	2297	2297	Bacterial_Spot		373	373	Mosaic-Virus
	360	360	Healthy		5357	5357	Leaf-Curl-Virus
Pepper_Bell	997	997	Bacterial_Spot				
	1478	1478	Healthy	**Total**	**54,306**	**54,306**	

4 PROPOSED ENHANCED ViT MODEL

4.1 Formulation of the Problem

Suppose there is a dataset d= $(a_i, b_i)^N_{i=1}$, where a_i is the i^{th} image of input and b_i is the ground label. The issue of disease detection and categorization in images of plant leaves can be stated in the following way. Let X be an input tensor image in T^{hxwxc}, where channels are represented by c, width by w and height by h. The objective of the task localization of disease is to forecast mask that is binary BM $\in \{1,0\}^{hxw}$ that shows whether the disease is present or not in every pixel in the image. In the disease classification task, the label that predicts the class L must be predicted between 1 and K, where K is the total number of disease classes.

When image I as an input is fed into a splitter for patches, it is divided into a group of patches that do not overlap. P = $\{p_1, p_2, \ldots, p_n\}$. The data load layer records connections and interactions of varying spatial sizes among the patches and generates the result C_S = $\{a_{s1}, a_{s2}, ..., a_{sn}\}$, with each as_i representing the scaled version of patch a_i. The scaled version of every patch a_i captures interactions and dependencies between patches of different scales, allowing for a complete understanding of the image's content. The scalable representation C_S is used as an input for the co attention mechanism, which generates a result that illustrates the relationships in between the patches. The representation of the patches is enhanced by this process by focusing on the connections among various sections. The result represented by C_A = $\{c_1, c_2, \ldots, c_n\}$ can be used by following modules in the system to make classification predictions.

4.2 Architecture

The research outlines various steps depicted in the Figure 4. The process involves loading data, augmenting data, encoding with vision transformer, optimizing, and making predictions.

A. Data Load

ViT divides the 2D image $\boldsymbol{I} \in R^{\boldsymbol{h}\times\boldsymbol{w}\times}$ into separate patches of size $\boldsymbol{P}\times\boldsymbol{P}$ without overlap. Following the image's transformation, a patch sequence is produced, denoted as $\boldsymbol{I}_{\text{pe}} \in R^{\boldsymbol{N}^{\text{P}\times(\boldsymbol{P}^2\cdot\boldsymbol{C})}}$, where P is the patch size and NP is the number of patches determined by $\boldsymbol{N}\mathbf{P}=(\boldsymbol{h}\times\boldsymbol{w})/\boldsymbol{P}^2$. In order to embed, input patches must be taken, flattened, and mapping them into a space with dimensions of D with a linear projection $\boldsymbol{E} = [\boldsymbol{I}_{\boldsymbol{p}\text{e}}{}^{1}, \boldsymbol{I}_{\boldsymbol{p}\text{e}}{}^{\boldsymbol{2}}, \ldots, \boldsymbol{I}_{\boldsymbol{p}\text{e}}{}^{\boldsymbol{N}}]$. The patch flattering and patch embedding are combined. To preserve the positional information, position encoding is incorporated into the combined embedding. Each patch's spatial information is encoded using positional encoding since transformers are naturally permutation-invariant. This aids the model in determining each patch's location within the original image. Through the utilization of this encoding method, the model can ensure it can differentiate between patches that are close together or far apart.

The following equation represents the embedding process.

$$X_i = [x_c;_{\text{e}}{}^{1}, I_{p\text{e}}{}^{2}, \cdots, I_{p\text{e}}{}^{N}] + E_{PO} \qquad (1)$$

where $X_i \in R^{(\boldsymbol{N}+\mathbf{1})\times\boldsymbol{D}}$ denotes the final value of embedding for i-th patch and $E_{PO} \in R^{(\boldsymbol{N}+\mathbf{1})\times\boldsymbol{D}}$, D denotes a latent vector size of the transformer.

$$E_{PO} = \begin{cases} \sin\left(pos/1000^{2i/d}\right), & if\ i\ is\ even \\ \cos\left(pos/1000^{2i/d}\right), & if\ i\ is\ odd \end{cases} \qquad (2)$$

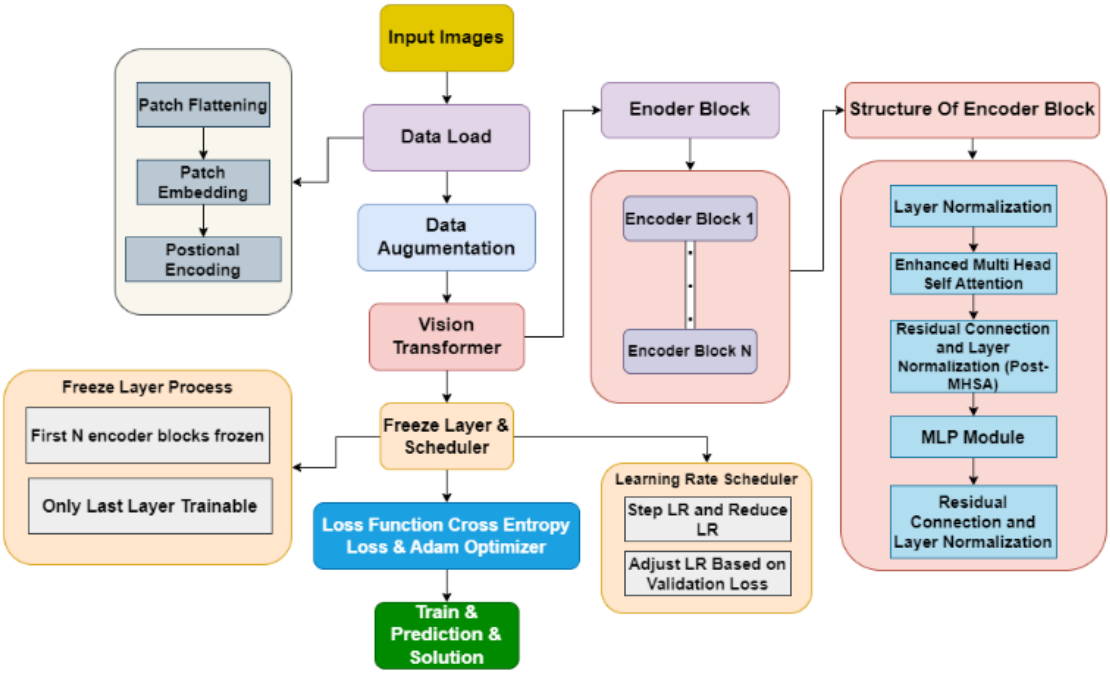


Figure 4. The proposed Enhanced ViT architecture

B. Data augmentation

Data augmentation techniques include flipping, rotation, random cropping, and color jitter are used to make the training dataset larger, creating diversity and preventing overfitting, thereby aiding in learning rotation or translation invariances. Random resizing and cropping, along with horizontal flipping, generate mirrored variations of images, enhancing the model's resilience against orientation alterations.

C. Enhanced Vision Transformer

The ViTs architecture, introduced by vision transformers (ViTs) [20], brought the idea of treating images patches as sequences, taking advantage of transformers' efficiency in natural language processing for computer vision task. Although ViT has demonstrated noteworthy results, there is potential for enhancement, specifically in improving the multi-head self-attention mechanism as depicted in Figure 5. In this section, the enhanced multi head self attention (eMHSA) with gate product, which includes several attention heads each head focuses on distinct aspects of the connections between patches. Attention is vital for understanding the connections between consecutive input patches. Values in the scaled dot-product attention mechanism have a dimensionality of d_{val}, while queries, keys, and values all have a dimensionality of d_{qk}. The calculation of the dot product between the query and all keys is done. A softmax algorithm is applied after the values are divided by the scale factor $\sqrt{d_{qk}}$. The three different embedding matrices—K(k), Q(q), and V(v)—are used in the attention calculation process. Assume that there is H attention heads. For every head,

$$q_h = X_i \cdot \mathcal{W}_h^q \; ; \; k_h = X_i \cdot \mathcal{W}_h^k ; \; v_h = X_i \cdot \mathcal{W}_h^v \qquad (3)$$

where each attention head's unique weight matrix is represented by $\mathcal{W}_h^q$, $\mathcal{W}_h^k$ and $\mathcal{W}_h^v$ .The probabilistic attention scores (PAS_h) for each head are then calculated by applying the softmax function to the scaled dot product of q_h and k_h.

$$PSA_h = SoftMax\left(\frac{q_h k_h^T}{\sqrt{d_k}}\right) \quad (4)$$

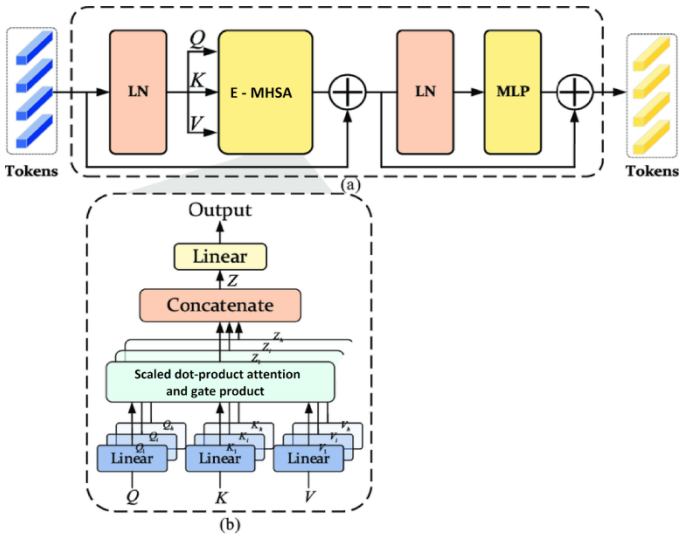


Figure 5 Proposed enhanced ViT model

The gated multi head self attention layer is an essential part of this model, created to improve the standard multi head self attention mechanism through the inclusion of a gating mechanism. The attention output is modulated by the gate values. This layer enables the model to regulate the information flow from the attention outputs in a selective manner, thereby enhancing the model's capacity for representation. The addition of a gating mechanism in the eMHSA layer permits the model to change the influence of the attention output based on the input context. An essential component of the model is the transformer encoder and used for identifying and categorizing plant diseases. The multi head attention's shared representation mechanism is taken and followed a series of layers in order to generate the features that are encoded. The encoder consists of an enhanced multi-head self attention, normalization layer, and a multi-player perceptron MLP. The input patches are processed by each encoder block, which consists of MLP modules and multi-head self-attention. The transformer encoder produces encoded features as their final outputs, which are then sent to other parts of the model for additional processing. The attention from the linear projection layer and gate is combined to create the enhanced multi-head self attention output (e$MHSA_o$). Training the model involves adjusting the parameters to reduce the loss during optimization.

$$\text{e}MHSA_o = Linear + \text{gate} \quad (5)$$

D. Layer normalization and multi-layer perceptron (MLP)

Deep learning uses a technique called layer normalization in order to normalize the activations in a neural network layer. The suggested approach is similar to batch normalization however, it normalizes the characteristics of a single data point rather than across a batch of data. By addressing the issue of internal covariate shift, batch normalization improves deep neural networks' capacity for training and generalization. In the end, the result is fed into a MLP layer, usually made up of two linear transformations. Nonlinearity is added in the following layers through the utilization of a MLP featuring the Gaussian-error linear unit (GeLU). Table 2 provides a summary of the ideal feature extractor parameters for the suggested model.

Table 2. Parameter for improved vision transformer

Parameter	Value
Image-Size	224x224
Patch-Size	16x16
Total number of patches	196
Dimension of projection	768
Number of Attention-heads	12
Number of Encoder Layers	12
Dimension of MLP unit	3072

5 EXPERIMENTAL SETUP

The model was specifically trained on images of Apple, grape, and corn leaves from the Plant Village dataset. The categorical cross-entropy loss was used for training, with weight updates performed using the adam optimizer for improving model performance via weight adjustments. The training used a scheduler to maximize the learning rate and comprised 50 epochs.. Table 3 presents a comprehensive summary of the hyper parameters.

Table 3 Hyper parameters

.

Hyper parameter	Values
Learning rate	1-e4
Pre trained model	Vit base patch 224
Epoch	50
Optimizer	Adam
Size of Batch	128
Scheduler	Learning rate
Loss Function	Cross Entropy

6 PERFORMANCE EVALUATIONS

Using an 80:10:1 0 ratio, the dataset was initially divided into three subsets: training set, the testing dataset and the validation set. The training set was then used to train the model, the validation set was used to evaluate and improve the model during training, and the testing set was used to check the model's predictions. The test set is solely utilized for final evaluation during inference and is not used during training. This study uses a confusion matrix in addition to measures like F1 score, accuracy, recall and precision, and to thoroughly evaluate the models' performance, as shown by the following equations:

$$\text{Accuracy} = (TP + TN)/(TN + TP + FP + FN)$$

$$\text{Precision} = TP/(TP + FP)$$

$$\text{Recall} = TP/(FN + TP)$$

$$F1 = (2 * \text{Precision} * \text{Recall})/(\text{Precision} + \text{Recall})$$

where:

- TP is a reference to samples that are true positives.
- TN is a reference to accurate negative samples.
- FN represents samples that are false negatives.
- FP represents samples that are false positives.

7 EXPERIMENTAL RESULT ANALYSES

This research utilized the baseline model ViTs to explore the most effective pairing of decoder blocks and attention heads for maximum accuracy.This research discovered that employing 12 attention heads in the encoder and decoder sections led to favorable outcomes. The choice was made to balance model complexity and efficiency for optimal results. These methods enhance the model's effectiveness without compromising the number of trainable parameters. The findings indicate that these methods enhance the model's performance positively.

Table 4 Classification report

Classification report			
	Precision	Recall	F1 Score
Apple black rot	97%	100%	99%
Apple cedar apple rust	100%	96%	97%
Apple scab	95%	100%	92%
Corn maize northern leaf blight	90%	97%	93%
Corn maize common rust	98%	100%	99%
Corn maize cercospora leaf spot gray leaf spot	95%	95%	90%
Grape leaf blight	98%	95%	97%
Grape esca	96%	100%	98%
Grape block rot	99%	99%	99%

The accuracy and loss for the Apple, Grape, and Corn disease dataset during training and validation are shown in Figure. 6.

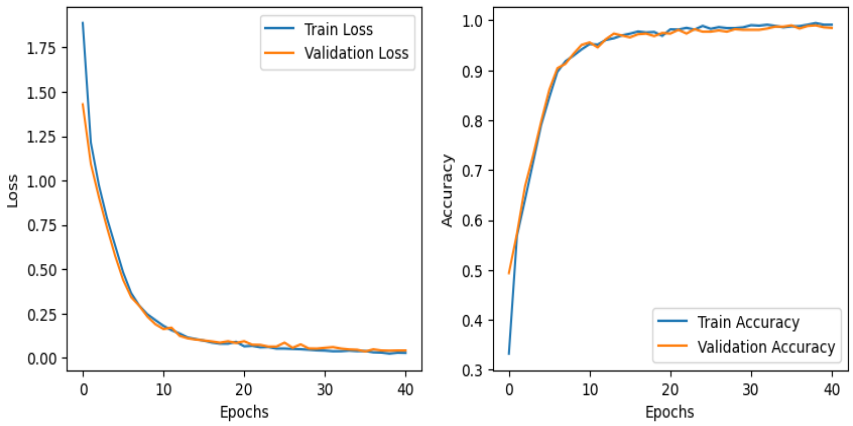


Figure 6 a) Apple leaf

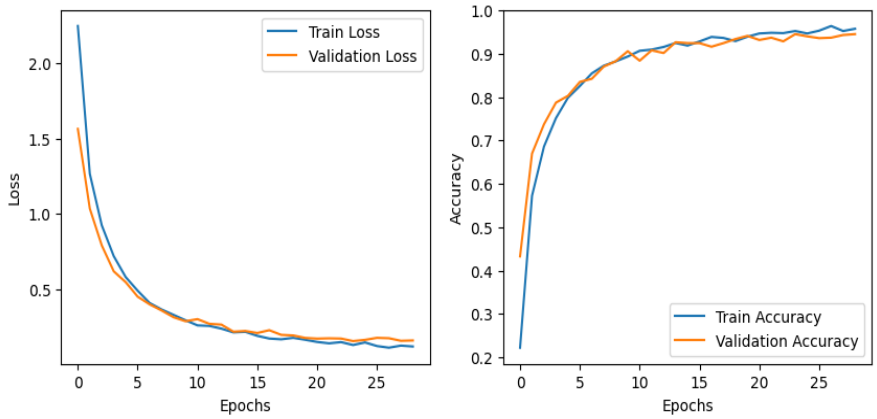


Figure 6 b) Corn leaf

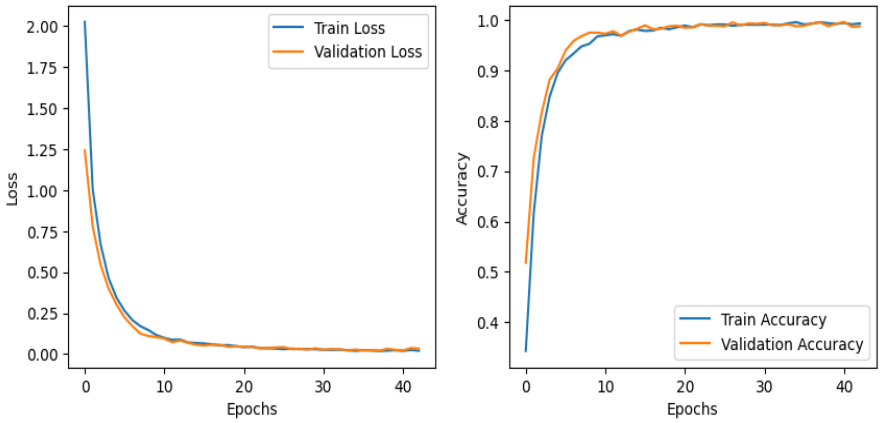
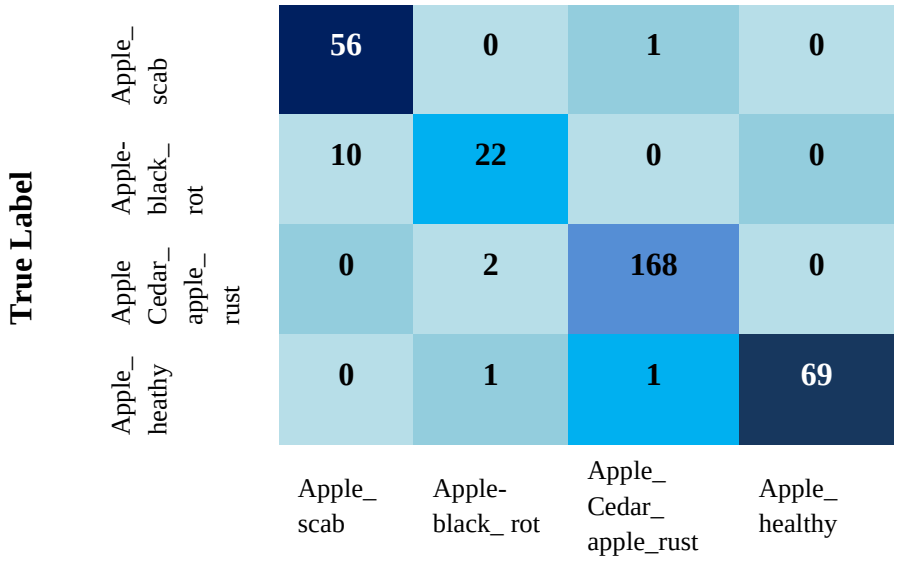
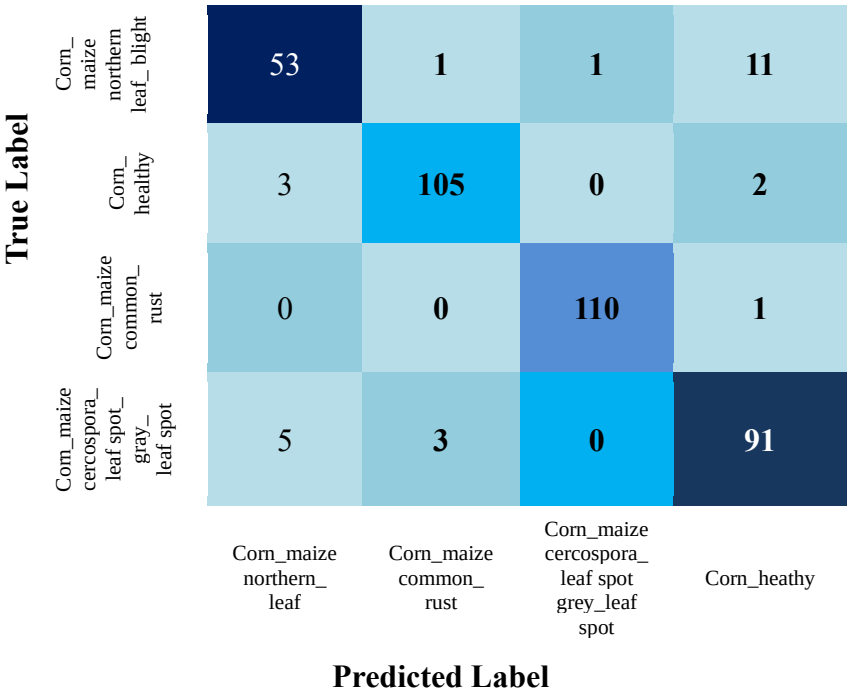


Figure 6 c) Grape leaf

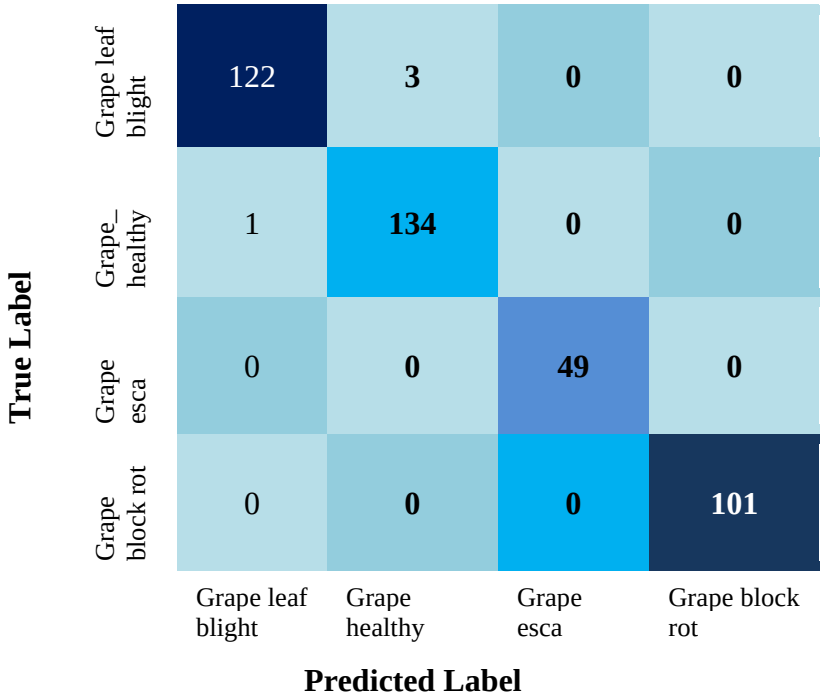
The outcomes of the suggested model on the test set are seen in the confusion matrix depicted in Figure 7. It consists of the count of accurate forecasts and mistakes for every category. The model shows impressive performance, making mostly correct predictions. The model shows some small inaccuracies, particularly in the form of incorrect positives and negatives in specific categories. The outcomes of the suggested model on the test set are shown in the confusion matrix in Figure 7.



7 a) Apple



7 b) Corn



7 c) Grape

Figure 7 Confusion matrix

For every class, it shows the proportion of accurate and inaccurate forecasts. The majority of the predictions made by the model are accurate, indicating its excellent performance. As shown in Table 4, the suggested model shows a strong performance in different categories, attaining nearly flawless F1 scores, recall, and precision. This finding suggests that the suggested model shows dependability and precision in its ability to make predictions. Certain classes, like mosaic virus, have the possibility for improved accuracy. The model using enhanced ViT with gate performs well, with a remarkable accuracy rate. Figure 8 represents the explainable predicted output with remedial solution.

8 a) apple 8 b) corn 8 c) Grape

Figure 8 Explainable predicted output

The authors' experiment with classifying leaf diseases using various algorithms and data sets is shown in Table 5.

Comparison of methods for identifying leaf diseases.

Sl No.	Author Tomato	Algorithm	Testing Accuracy
1	Abbas et al. [21]	DenseNet121	97.11%
2	Agarwal et al. [22]	VGG16	91.20%
3	Hossain et al. [23]	Multi-Axis Vision Transformer	93.00%
4	Zhou et al.[24]	Pre-trained vision transformer	92.00%
5	Sun et al[25]	SE-VIT	97.26%
6	Thai et al.[26]	Tiny-LeViT	97.25%
7	Proposed Work	Enhanced ViT	99.04%

These findings are displayed in Figure 9. It is evident that the enhanced vision transformer model produced the best outcomes. It was discovered during the experimental investigation that the proposed model performed better than the other models for the PlantVillage dataset.

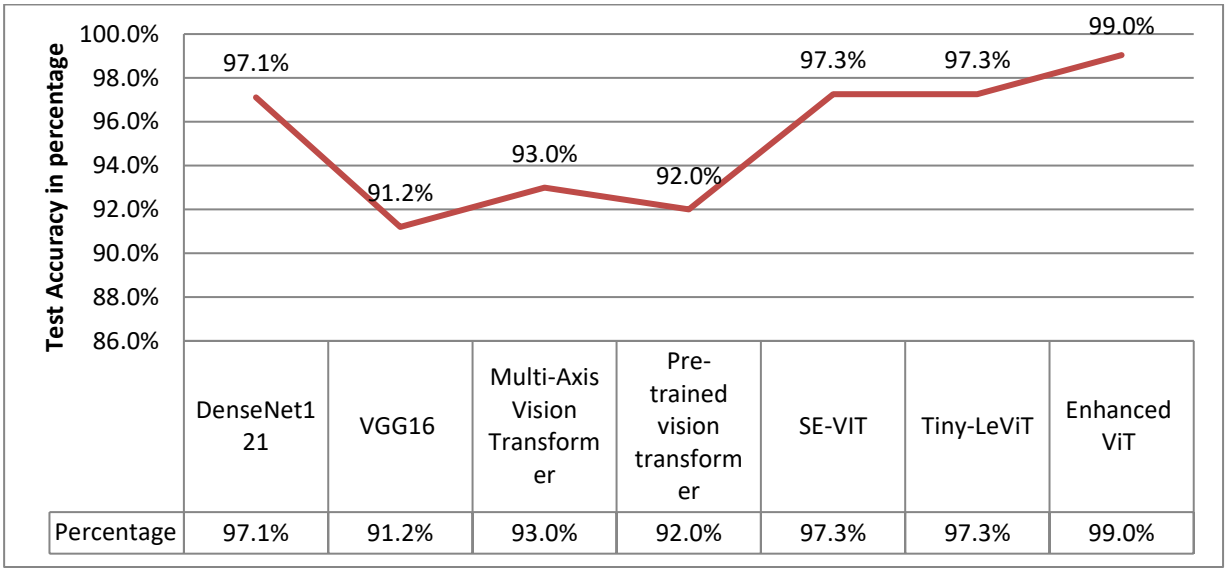


Figure 9 Comparison of methods for identifying leaf diseases

8 CONCLUSIONS

This research presents a new method for identifying diseases in plant leaves. The method involves a deep neural network framework with an advanced vision transformer. The results demonstrate that the model can accurately and efficiently classify various plant leaf diseases. The model demonstrates high levels of F1 scores , recall and precision for a range of classes that correspond to different plant diseases. Researchers and experts in the field, as well as pathologists and farmers looking to forecast plant diseases, might benefit from the approach described here.

DECLARATIONS:

Funding:

The work received financial support through seed money project grant by Karpagam Academy of Higher Education for the project No:KAHE/R-Acad/AI/Seed Money/015.

9 REFERENCES

1. Jeger, M., Beresford, R., Bock, C., Brown, N., Fox, A., Newton, A & Yuen, J.: Global challenges facing plant pathology: multidisciplinary approaches to meet the food security and environmental challenges in the mid-twenty-frst century. CABI Agriculture and Bioscience 2(1), 1–18 (2021)
2. Talaviya, T., Shah, D., Patel, N., Yagnik, H., Shah, M.: Implementation of artificial intelligence in agriculture for optimisation of irrigation and application of pesticides and herbicides. Artificial Intell Agricult 4, 58–73 (2020)
3. Ayaz, M., Ammad-Uddin, M., Sharif, Z., Mansour, A., and Aggoune., E.-H. M. (2019). Internet-of-things (IoT)-based smart agriculture: toward making the fields talk. IEEE Access 7, 129551–129583. doi: 10.1109/Access.6287639.
4. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Houlsby N An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010. 11929 (2020)

5. Thakur P S, Khanna P, Sheorey T, Ojha A Vision transformer for plant disease detection: PlantViT. In: International Conference on Computer Vision and Image Processing, 501–511. Cham: Springer International Publishing (2021, December)
6. Boukabouya RA, Moussaoui A, Berrimi M: Vision Transformer based models for plant disease detection and diagnosis. In: 2022, 5th international symposium on informatics and its applications (ISIA), 1–6. IEEE (2022, November). https://doi.org/10.1109/ ISIA.2022.9923797
7. Thai HT, Tran-Van NY, Le KH Artificial cognition for early leaf disease detection using vision transformers. In: 2021 International conference on advanced technologies for communications (ATC), 33–38. IEEE (2021, October). https://doi.org/10.1109/ATC53345. 2021.9631065
8. Thakur PS, Khanna P, Sheorey T, Ojha, A (2022) Explainable vision transformer enabled convolutional neural network for plant disease identification: PlantXViT. arXiv preprint arXiv:2207. 07919
9. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A. Weissenborn, D.; Zhai, X.; Unterthiner, T. Dehghani, M.; Minderer, M.; Heigold, G. Gelly, S.; et al. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv 2021, arXiv:2010.11929
10. Maurício, J.; Domingues, I.; Bernardino, J. Comparing Vision Transformers and Convolutional Neural Networks for Image Classification: A Literature Review. Appl. Sci. 2023, 13, 5521. [CrossRef]
11. Li, Z.; Tao, W.; Liu, J.; Zhu, F.; Du, G.; Ji, G. Tomato Leaf Disease Recognition via Optimizing Deep Learning Methods Considering Global Pixel Value Distribution. Horticulturae 2023, 9, 1034. [CrossRef]
12. Dhanya, V.G.; Subeesh, A.; Kushwaha, N.L.; Vishwakarma, D.K.; Kumar, T.N.; Ritika, G.; Singh, A.N. Deep Learning Based Computer Vision Approaches for Smart Agricultural Applications. Artif. Intell. Agric. 2022. [CrossRef]
13. Thakur, P.S.; Chaturvedi, S.; Khanna, P.; Sheorey, T.; Ojha, A. Vision Transformer Meets Convolutional Neural Network for Plant Disease Classification. Ecol. Inform. 2023, 77, 102245. [CrossRef]
14. Hu R, Singh A (2021) Unit: Multimodal multitask learning with a unifed transformer. In: Proceedings of the IEEE/CVF international conference on computer vision, 1439–1449. https://doi.org/10.1109/ICCV48922.2021.00146
15. Bhattacharjee D, Zhang T, Süsstrunk S, Salzmann M (2022) Multi: An end-to-end multitask learning transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 12031–12041 https://doi.org/10.1109/CVPR52688.2022.01173
16. Tian, Y., Bai, K.: End-to-end multitask learning with vision transformer. IEEE Trans Neural Netw Learn Syst (2023). https://doi. org/10.1109/TNNLS.2023.3278896
17. Li, X., Li, X., Zhang, S., Zhang, G., Zhang, M., Shang, H.: SLViT: Shufe-convolution-based lightweight Vision transformer for efective diagnosis of sugarcane leaf diseases. J King Saud University-Compute Inform Sci 35(6), 101401 (2023). https://doi.org/10.1016/j.jksuci.2023.01.014
18. Li, X., Li, S.: Transformer help CNN see better: a lightweight hybrid apple disease identification model based on transformers. Agriculture 12(6), 884 (2022). https://doi.org/10.3390/agricultur e12060884
19. Plant Village Dataset: https://data.mendeley.com/datasets/tywbt sjrjv/1 Accessed on 24 May 2023
20. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”, In: Proc. of 9th International Conference on Learning Representations (ICLR), pp.1-22, 2021.
21. Abbas, A.; Jain, S.; Gour, M.; Vankudothu, S. Tomato Plant Disease Detection Using Transfer Learning with C-GAN Synthetic Images. Comput. Electron. Agric. 2021, 187, 106279. [CrossRef]

22. Agarwal, M.; Singh, A.; Arjaria, S.; Sinha, A.; Gupta, S. ToLeD: Tomato Leaf Disease Detection Using Convolution Neural Network. Procedia Comput. Sci. 2020, 167, 293–301. [CrossRef].
23. Hossain, S.; Tanzim Reza, M.; Chakrabarty, A.; Jung, Y.J. Aggregating Different Scales of Attention on Feature Variants for Tomato Leaf Disease Diagnosis from Image Data: A Transformer Driven Study. Sensors 2023, 23, 3751. [CrossRef]
24. C. Zhou, Y. Zhong, S. Zhou, J. Song, W. Xiang, Rice leaf disease identification by residual-distilled transformer, Eng. Appl. Artif. Intell. 121 (2023) 106020, https:// doi.org/10.1016/j.engappai.2023.106020.
25. C. Sun, X. Zhou, M. Zhang, A. Qin, SEVisionTransformer: hybrid network for diagnosing sugarcane leaf diseases based on attention mechanism, Sensors 23 (20) (2023) 8529, https://doi.org/10.3390/s23208529.
26. H.T. Thai, K.H. Le, N.L.T. Nguyen, Towards sustainable agriculture: a lightweight hybrid model and cloud-based collection of datasets for efficient leaf disease detection, Future Generat. Comput. Syst. (2023), https://doi.org/10.1016/j. future.2023.06.016.