Smart Agricultural Technology 8 (2024) 100500

Contents lists available at ScienceDirect

Smart Agricultural Technology

journal homepage: www.journals.elsevier.com/smart-agricultural-technology

A customised vision transformer for accurate detection and classification of Java Plum leaf disease

Auvick Chandra Bhowmik [a], Md. Taimur Ahad [b,*], Yousuf Rayhan Emon [a], Faruk Ahmed [a], Bo Song [c], Yan Li [b]

[a] 4IR Research, Daffodil International University, Dhaka, Bangladesh
[b] School of Mathematics Physics and Computing, University of Southern Queensland, Toowoomba Campus, Toowoomba, Australia
[c] School of Engineering, University of Southern Queensland, Australia

ARTICLE INFO

Keywords:
Vision transformer
ViT
Deep learning
Java Plum leaf disease detection
Accuracy
Validation
Confusion matrix
Train
Primary dataset

ABSTRACT

Vision Transformer (ViT) has recently attracted significant attention for its performance in image classification. However, studies have yet to explore its potential in detecting and classifying plant leaf disease. Most existing research on diseased plant leaf detection has focused on non-transformer convolutional neural networks (CNN). Moreover, the studies that applied ViT narrowly experimented using hyperparameters such as image size, patch size, learning rate, attention head, epoch, and batch size. However, these hyperparameters significantly contribute to the model performance. Recognising the gap, this study applied ViT to Java Plum disease detection using optimised hyperparameters. To harness the performance of ViT, this study presents an experiment on Java Plum leaf disease detection. Java Plum leaf diseases significantly threaten agricultural productivity by negatively impacting yield and quality. Timely detection and diagnosis are essential for successful crop management. The primary dataset collected in Bangladesh includes six classes, '*Bacterial Spot*', '*Brown Blight*', '*Powdery Mildew*', and '*Sooty Mold*', '*healthy*', and '*dry*'. This experiment contributes to a thorough understanding of Java Plum leaf diseases. Following rigorous testing and refinement, our model demonstrated a significant accuracy rate of 97.51%. This achievement demonstrates the possibilities of using deep-learning tools in agriculture and inspires further research and application in this field. Our research offers a foundational model to ensure crop quality by precise detection, instilling confidence in the global Java Plum market.

1. Introduction

Vision Transformer (ViT) is a neural network architecture with deep learning technology that has gained significant attention in image detection and classification. Among the image classification tasks, ViT gained an auspicious position in plant leaf disease detection through its remarkable ability to simulate long-range dependencies using the self-attention mechanism (Alzahrani et al., 2023; [1–3]). While traditional CNNs depend on convolution-based architecture, ViT depends on transformer-based architecture. Transformer-based architecture collects information from image data patterns, making image processing more efficient (Alzahrani et al., 2023). ViT can automatically process image features and utilise these features for image classification and detection using pattern recognition [1]. The transformer's acyclic network structure is synchronised with parallel computing through an encoder and decoder. Moreover, training time reduction and performance enhancement in transformers are done by a self-attention mechanism [2, 3]. Thus, ViT has emerged as a promising methodology for disease detection on Java Plum leaves, as it enhances detection accuracy, demonstrating notable efficacy [4]. In the context of Java Plum production, accurate and timely disease identification is important (Alzahrani et al., 2023).

Java Plum is a delicious and nutrition-rich fruit, mainly grown in Southeast Asia. It is also known as jambolana, or black plum, and contains significant amounts of iron, calcium, and vitamins A and C [5]. Although this fruit has some significant health benefits, the production of this crop faces various challenges due to its extensive leaf diseases. Some of the primary diseases of Java Plum leaves are bacterial spot, brown blight, powdery mildew, and sooty mold. These diseases threaten the food supply considerably and hamper the overall production of Java Plum, which could lead to an economic loss for the Java Plum growing countries. This threat can be reduced by detecting these diseases

* Corresponding author.
E-mail address: MdTaimur.Ahad@unisq.edu.au (Md.T. Ahad).

https://doi.org/10.1016/j.atech.2024.100500
Received 13 May 2024; Received in revised form 1 July 2024; Accepted 1 July 2024
Available online 2 July 2024

accurately (De Silva et al., 2023). However, the manual detection of plant leaf disease is more time-consuming and often leads to mistakes [6]. Thus, effective and accurate identification of Java Plum leaf diseases is imperative for reducing fruit production loss. However, despite the advancements in precision agriculture, many existing technologies classify plant diseases with minimal effectiveness. This is because of the complex nature of Java Plum diseases in real-world scenarios, where conventional detection techniques often struggle to produce rapid and accurate results. To address the issue, this study aims to utilise a ViT-based model to improve the accuracy of Java Plum leaf disease detection from primary data. The main goal of the model is to detect the leaf diseases accurately, which can minimise the production loss of Java Plums.

Though ViT has received significant attention in the detection of plant leaf disease, It still needs to be explored in the detection of Java Plum disease, which is a significant gap in precision agriculture. Other limitations also exist. Among other limitations, the following are mentionable:

- Firstly, as Kamal et al. [7] highlighted, the computational intensity of ViTs compared to traditional CNNs necessitates customising ViT networks. Scholars such as Ahmed et al. (2023) and Mamun et al. [8] have responded to this need by customising ViTs to achieve optimal results in the agriculture domain. Mustofa et al. [9] suggested utilising ViT for plant leaf disease detection earlier.
- Secondly, while Vision Transformers provide a novel approach to image classification, their internal workings are often less interpretable than CNNs'. This lack of interpretability can be a barrier in agricultural applications, where understanding the model's decision-making process is crucial for gaining end-users trust, such as farmers and agronomists [10].
- Thirdly, the absence of well-defined models poses a significant challenge in accurately distinguishing between various leaf diseases in real-world scenarios. Moreover, most crop disease detection studies are conducted using secondary datasets. However, leaf disease varies from country to country due to differing environments. Therefore, it is necessary to create a comprehensive and well-sorted dataset of Java Plum leaf disease for training and testing the efficiency of the ViT model.
- The fourth problem is related to the accuracy of leaf disease detection. The classification of modalities is a concern because lower detection accuracy and large false-positive values will narrow the applicability and acceptability of ViT in precision agriculture.
- The fifth problem is associated with optimal ViT configuration. The ideal image size and patch size for achieving peak ViT performance in Java Plum leaf disease detection remains an open question. Understanding these parameters is critical to maximising accuracy.
- Lastly, the model complexity and computational intensity present another problem. ViT, while powerful, can be computationally demanding. Their architecture, consisting of self-attention modules, multilayer perceptron (MLP) networks, and associated embedding mechanisms, necessitates exploring optimisation strategies to enhance efficiency within the specific context of plant pathology.

To fill the gap, in this research, this study seeks to answer two research questions:

RQ1: How can Java Plum leaf disease detection accuracy be improved using ViT?
RQ2: Which patch size and image size combination can provide better accuracy in the Java Plum dataset?

This study aims to solve these problems and investigates the efficiency of ViT in Java Plum leaf disease detection. It is expected that the outcomes of this research will help farmers in timely Java Plum leaf disease detection by improving the precision of disease detection. Additionally, it aims to develop an automated and efficient tool for detecting and classifying Java Plum leaf diseases. To sum up, the contributions of this study are detailed below:

- This study introduces a customised ViT model that effectively classifies Java Plum leaf diseases based on ViT.
- This research collects a primary dataset to consider the scalability and practicality of deploying the ViT model in real-world Java Plum fields.
- This study proposes an algorithm to calculate the importance of attention heads in each layer of the ViT structure. Based on this criterion, we prune the less important attention heads to reduce the model size to 28%, boosting inference speed with a 2% F1 score improvement.
- This study exploits sparse matrix-matrix multiplication in self-attention blocks to improve the training efficiency by up to 10% with 15% less GPU consumption while keeping the corresponding F1 - score compared to the original model.
- This study introduces a model that effectively classifies leaf diseases based on ViT. It also comprehensively compares the experimental results with state-of-the-art studies [11,12]. The results of this study achieve better accuracy performance, with at least 1% more than the best of these models, instilling confidence in the robustness of our findings.

2. Related works

Several groups of researchers have used different techniques for plant leaf disease detection. However, Numerous researchers have employed diverse vision transformer techniques (See Table 1).

The initial group of researchers employed the conventional Vision Transformer (ViT) model for detecting plant leaf diseases. For instance, Salamai et al. (2023) and De Silva et al. (2023) utilised the traditional ViT model in their studies. Salamai et al. (2023) chose this model because of its ability to creatively learn visual representations of leaf diseases across spatial and channel dimensions. De Silva et al. (2023) opted for this model because it effectively identifies plant leaf diseases under real environmental conditions. In this line of research, De Silva et al. (2023) achieved a training accuracy of 93.71% and a test accuracy of 90.02%. However, it is noteworthy that both studies were conducted on publicly available datasets, and conducting experiments on primary datasets related to the actual environment could enhance the precision of these studies.

The second group of researchers employed various hybrid Vision Transformer (ViT) models to detect diseases in diverse plant leaves. For instance, Tabbakh et al. (2023) applied two different models, Transfer Learning and Vision Transformer (TLMViT), on a wheat dataset. Zhang et al. [3] implemented the Shuffle-convolution-based lightweight Vision Transformer (SLViT) model on sugarcane leaves and utilised the Improved Vision Transformer (ViT) for identifying agricultural pests. Sun et al. [15] also applied SE-VIT to diagnose diseases in sugarcane leaves. Zhan et al. [2] developed the IterationVIT model for diagnosing tea diseases. Thai et al. [16] used the Tiny-LeViT model for efficient leaf disease detection, and Zhang et al. [3] applied the hybrid IEM-ViT to quickly and accurately recognise tea diseases. In contrast, Parez et al. [17] used GreenViT. Zeng et al. [18] used the Squeeze-and-Excitation Vision Transformer (SEViT) for identifying large-scale and fine-grained diseases, and Li et al. [14] implemented the Plant-based MobileViT (PMVT) model for real-time plant disease detection. Some of these studies reported notable accuracies, such as Zhou et al. [13], Tabbakh et al. (2023), Sun et al. [15], Zhan et al. [2]; Thai et al. [16]; and Zhang et al., [3] achieved 92.00%, 98.81%, 97.26%, 98%, 97.25% accuracies respectively. However, it is worth noting that while Zhou et al. [13], Zhan et al. [2], Zhang et al. [3], and De Silva et al. (2023) conducted studies on primary datasets, other researchers implemented their models on publicly available datasets, introducing some

Table 1
Research Matrix.

Author	Model	Accuracy	Contribution
Salamai et al. (2023)	ViTs		Proposed a visual modulation network that can creatively acquire the visual representations of leaves.
De Silva et al. (2023)	ViT	Train 93.71% & Test 90.02%	Underscored the possibility of employing sophisticated imaging methods for precise and dependable identification of plant diseases.
Zhou et al. [13].	Pre-trained vision transformer	92.00%	Proposed a residual-distilled transformer architecture.
Li et al. [14]	SLViT	Accuracy bonus of 1.87% over MobileNetV3	Presented a hybrid model that was initially trained on the publicly available dataset.
Fu et al. [1].	Improved Vision Transformer (ViT)		This model is highly accurate in image processing and recognition technology.
Sun et al. [15]	SE-VIT	97.26%	Using limited datasets, the SE-VIT model smartly manages sugarcane plantations and addresses plant diseases.
Thai et al. [16]	Tiny-LeViT	97.25%	The Tiny-LeViT model, which relies on the transformer architecture, was suggested to classify leaf diseases efficiently.
Zhang et al. [3]	IEM-ViT	93.78%	Improved the recognition accuracy by nearly 20% compared to the ResNet18, VGG16, and VGG19,
De Silva et al. (2023)	ViT + CNN	Train 92.83% & Test 88.86%	Hybrid ViT combines the strengths of both CNN and ViT.
Parez et al. [17]	GreenViT		This proposed model outperforms state-of-the-art (SOTA) CNN for detecting plant diseases.
Zeng et al. [18].	Squeeze-and-Excitation Vision Transformer (SEViT)	Test 88.34%	SEViT has improved the classification accuracy by 5.15% compared to the baseline model.
Li et al. [14].	Plant-based MobileViT (PMVT)	93.6%, 85.4% & 93.1%	Created a plant disease diagnostic application utilising the PMVT model to detect plant diseases in various situations.
Rethik et al. (2023).	ViT1, ViT2 & pre- trained ViT_b16	85.87%, 89.16% & 94.16%.	Substituted CNN with a novel Vision Transformer technique to classify plant leaf diseases.
Hossain et al. [19]	EANet, MaxViT, CCT & PVT	89%, 97%, 91% & 93%	Analysed the effect of four different transformer-based models.

limitations to their studies.

The third group of researchers employed various models to compare accuracy and performance on the same or different datasets. Rethik et al. (2023) utilised ViT1, ViT2, and pre-trained ViT_b16 models to assess accuracy and performance for classifying plant leaf diseases. The experiment shows that the pre-trained ViT_b16 model performs better than other models and presents accuracy values of 85.87%, 89.16%, and 94.16% for ViT1, ViT2, and pre-trained ViT_b16 models. Hossain et al. [19] compared the accuracy of DenseNet169, ResNet50V2, and ViT models for early detection and recognition of tomato leaf diseases, concluding that DenseNet169 demonstrated the highest efficacy and presented 99.88% train and 99.00% test accuracy for the DenseNet121 model and 95.60% train and 98.00% test accuracy for the ResNet50V2 and ViT models. Alzahrani et al. (2023) evaluated the accuracy and performance of EANet, MaxViT, CCT, and PVT models for tomato leaf disease detection, with PVT emerging as the superior model and found 89%, 97%, 91%, and 93% accuracy rates for EANet, MaxViT, CCT, and PVT models, respectively. Öğrekçi et al. [20] implemented DenseNet121, ViT, and ViT & CNN models to compare accuracy and performance in identifying sugarcane leaf diseases, with ViT proving the most effective model. Among these studies, some researchers achieved notable accuracy for robust model comparisons and achieved 92.87%, 93.34%, and 87.37% for DenseNet121, ViT, and ViT + CNN models, respectively. Notably, the researchers used public datasets for their models, which might limit how much their studies add to knowledge enrichment compared to using first-hand datasets.

The fourth group of researchers employed multiple models to construct ensemble models, aiming for improved accuracy and performance. Kumar et al. [21] utilised EfficientNet, SEResNeXt, ViT, DeIT, and MobileNetV3 models to develop an ensemble model for detecting cassava leaf diseases and achieved an accuracy of 90.75% on their ensemble model. Ganguly et al. [22] introduced an ensemble model incorporating CNN, ResNeXt, and InceptionV3 to detect plant leaf diseases. Chang et al. [23] constructed an ensemble model comprising ViT, PVT, and Swin to enhance identification quality in plant leaf disease detection. Among these studies, only Kumar et al. [21] conducted their experiments on publicly available datasets.

The last group of researchers employed diverse models to enhance accuracy and performance in disease detection. Diana Andrushia et al. [24] Utilised convolutional capsule networks for identifying grape leaf diseases and achieved 99.12% accuracy in their study. Kumar et al. [21] introduced paddy leaf disease detection using a multi-scale feature fusion-based RDTNet and attained 99.55%, 99.54%, and 99.53% accuracy, f1-score, and precision, respectively. Hu et al. [25] proposed an Adaptive Fourier Neural Operators (AFNO)-based Transformer architecture named FOTCA, focusing on extracting global features in advance and reporting a 99.8% accuracy for the FOTCA model. Arshad et al. [26] presented a novel hybrid deep learning model, PLDPNet, designed to forecast potato leaf diseases automatically and achieved an overall accuracy of 98.66% with their proposed model. Zhang et al. [3] introduced a unique segmentation model for grape leaf diseases in natural scene photos, termed Locally Reversible Transformer (LRT). Thai et al. [16] developed a transformer-based leaf disease detection model, FormerLeaf. Devi et al. [11] created an EffectiveNetV2 model for pest identification and plant disease categorisation. Given the use of publicly available datasets by all researchers, there may be a potential deficiency in knowledge enrichment in their studies at some level.

3. Description of the experiment

The experiment of the study is described below:

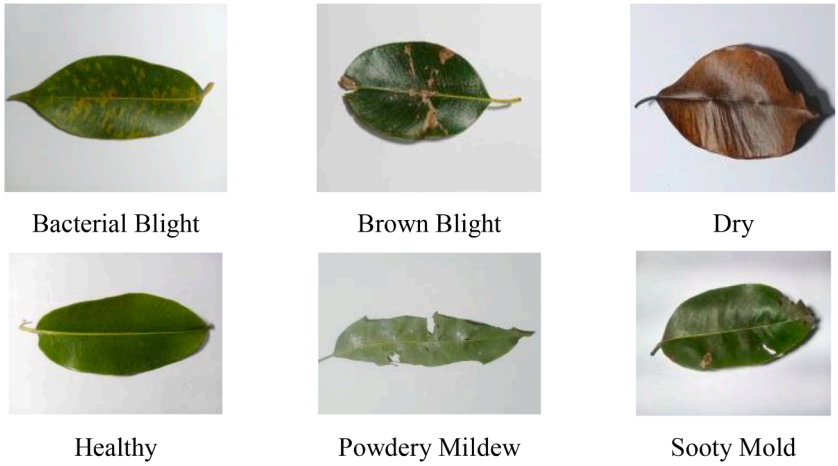
3.1. Dataset

The dataset used in this study was collected from two different areas in Bangladesh: "Titas" and "Barura", located in the "Cumilla" district. This dataset is designed to develop an accurate Java Plum leaf disease detection system using the ViT model. Different real-world Java Plum farm images were captured to build the Java Plum leaf disease dataset containing different classes. The dataset contains six distinct classes, each representing a specific category of Java Plum leaf (See Table 2). Four classes correspond to different Java Plum leaf diseases, including 'Bacterial Spot', 'Brown Blight' 'Powdery Mildew', 'Sooty Mold', 'healthy', and 'dry' Java Plum leaves.

3.2. Data splitting

The entire dataset is randomly partitioned into training, validation, and testing sets. This division enabled the training of the Vision Transformer model, validating it, and objectively evaluating its performance.

Table 2
Images of 6 classes from the dataset.



The dataset serves as the main supporting structure for this experiment and enables the construction and evaluation of a cutting-edge deep learning system for Java Plum leaf disease identification.

3.3. Experimental setup

This research experiment was carried out using the Keras library on Google CoLab. TensorFlow, a leading Python deep learning library for machine learning in Python, was employed. ViT was trained in this study using a Tesla graphics processing unit (GPU) on Google Colab. Google Collaboratory provides access to Tensor Processing Units (TPU). Initially, the Colab framework offers approximately 12 GB of random-access memory (RAM) and around 360 GB of GPU in the cloud for research purposes. The experimental setup for this study contains importing the required Python libraries (for instance, TensorFlow, Keras, etc. Configuring hyperparameters, enhancing the data, creating patches, using patch encoders, building custom ViT classifiers, running experiments, etc. The objective was to assess the performance of customised ViT architecture and determine the most effective model for identifying Java Plum leaf diseases. The architecture and algorithm of the customised ViT model are given in Sections 3.4 and 3.5, respectively.

3.4. Proposed model

The proposed model consists of a few prominent customisations. Firstly, patch size can be adjusted dynamically according to the input

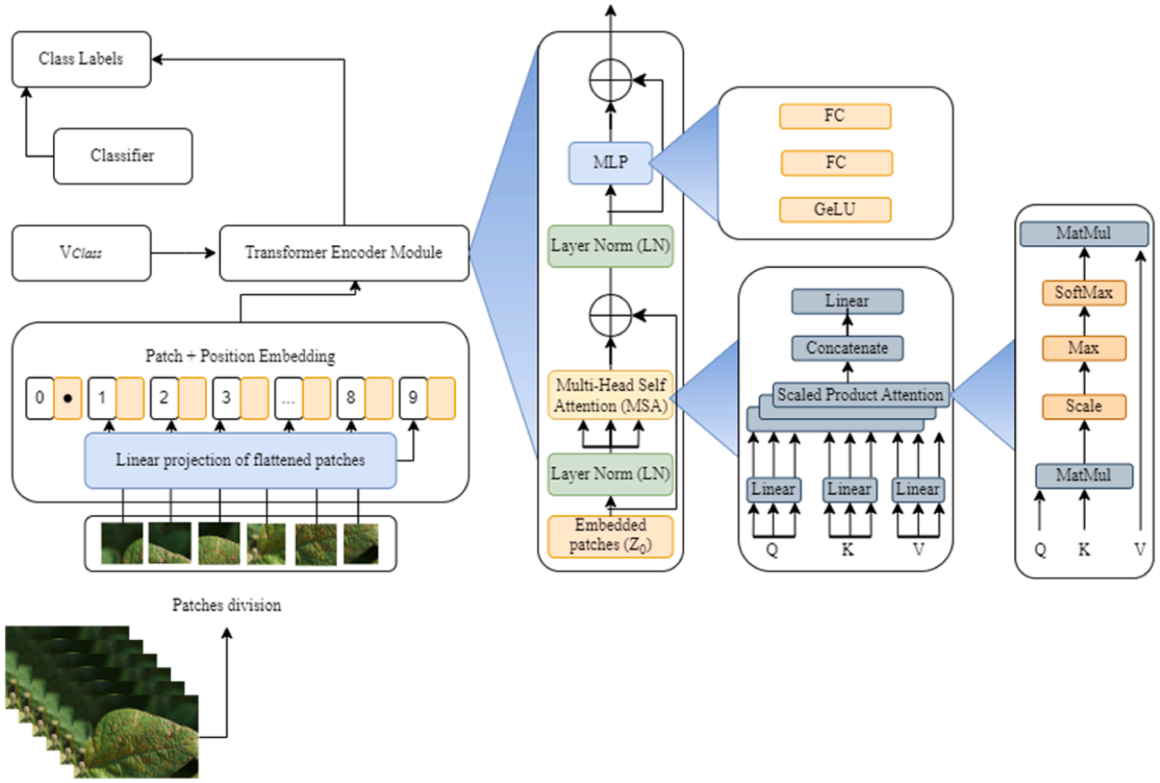


Fig. 1. Vision Transformer Architecture.

image size; after that, each patch size is linearly projected to a fixed projection dimension of 64. Secondly, positional encoding is added with project patch embedding to encode spatial information; the size of positional encoding is computed based on the input image size. Thirdly, a customisable encoder layer is built where the number of attention heads can be 2,4,6, etc. The architecture of the transformer encoder is built with one layer. Each layer contains a self-attention mechanism and a feed-forward network. After passing through transformer encoder layers, global average pooling aggregate information across all patches. Lastly, this model is trained using hyperparameter tuning by making a learning rate of 0.001, weight decay of 0.0001, and variable batch size. ViT shows exceptional performance in image classification and analysis by following this technique. Fig. 1 shows the model structure of the proposed ViT for Java Plum leaf diseases.

3.5. Algorithm of ViT

Algorithm of Customized ViT

1 **Input**: Image Data.
2 **Output**: Best Java Plum Leaf Detection Model (model.h5).
3 **Initialize** Vit models, list of image size, list of patch size, and evalList.
Four #Dataset preparation.
5 **Load** dataset (Folder inside image data).
6 #Model & Parameter Initialization.
7 **ImageSize** = [from 21 to 64].
8 **PatchSize** = [from 4 to 12].
Nine #Hyperparameter Optimization.
10 **For** each model in ViT:
For each size in **ImageSize**:
For each batch in **PatchSize**:
Model. Train (**kingsize, patch**).
Append evaluation result to **evalList**.
11 #Model Evaluation & Export.
12 **imgsize, patch** = Get optimal hyperparameter based on highest accuracy.
13 model. Evaluate (**imgsize, patch**).
14 Export model weight.

4. Experimental result

Different measures are utilised to assess the performance of machine learning classification models, providing insights into the effectiveness of vision transformer-based models within a specific application. The following metrics for performance evaluation are considered:

4.1. Accuracy

Accuracy is a measure of how well a detection model performs. In simpler terms, it is the proportion of predictions the model makes correctly. The computation of accuracy involves determining the ratio of correctly classified images to the overall number of images in the dataset. The following formula is used to calculate accuracy:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

4.2. Precision

Precision is an essential metric for evaluating classification models. It is calculated using the ratio of true positives to all positive values. Generally, it answers, "From all positive values, how many are true positives?" This metric has significant value in cases of false positives considered in references.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

4.3. Recall

Recall is also an essential metric for evaluating classification models. It is calculated using the ratio of true positives by summating actual positive and false negative values. Generally, it answers a specific question: "From all true positive and false negative values, how many are positive?" This metric has significant value, as false negatives are more costly than positives.

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

4.4. F1-Score

The F1-score uses both precision and recall, assessing detection accuracy. It is computed as the harmonic means of precision and recall. It achieves the highest value in terms of precision and recall, which are equal. The formula for the F1 score is as follows:

$$F1 - Score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{4}$$

4.5. Analysis of model accuracy and training times

Table 3 illustrates the evaluations of eight distinct image and patch size combinations at epoch 250, head 6, and a learning rate of 0.001. The combinations of image size and patch configurations are: (Image [64×64], patch 12), (Image [56×56], patch 7), (Image [48×48], patch 6), (Image [32×32], patch 8), (Image [32×32], patch 6), (Image [28×28], patch 7), (Image [28×28], patch 4) and (Image [21×21], patch 7). Table 5 presents the accuracy outcomes for these configurations, highlighting the highest accuracy of 97.51% in 245 s for (Image [64×64], patch 12) and the lowest accuracy of 93.36% in 240 s for (Image [21×21], patch 7).

4.6. Experimenting with vision transformer

Tables 4 and 5 illustrate the precision, recall, f1 score, and support for validation and testing for the image size and patch configurations (Image [64×64], patch 12) and (Image [56×56], patch 7). Of these, (Image [64×64], patch 12) has shown the best accuracy of 97.51%, and (Image [56×56], patch 7) has given 97.27% accuracy.

4.7. Confusion matrix

Figs. 2 and 3 illustrate the confusion matrix for validation and testing for the image size and patch configurations: (Image [64×64], patch 12) and (Image [56×56], patch 7). In the confusion matrix, there are six classes. The matrix shows that the number of true positives increases when the image size is more significant. It also shows that several False Positives depend on image size. However, the performance of this experiment increased when the image size was [64×64], and the patch size was 12.

Table 3
Evaluating eight image sizes and patch configurations for detection performance.

Image size	Patch size	Accuracy [%]		Training time [s]	Accuracy [%]
		Valid	Test		
64	12	97%, 98%		245	97.51%
56	7	97%, 97%		236	97.27%
48	6	98%, 97%		241	97.1%
32	8	92%, 95%		239	94.67%
32	6	97%, 96%		242	96.00%
28	7	96%, 96%		234	96.21%
28	4	97%, 95%		231	94.70%
21	7	95%, 93%		240	93.36%

Table 4
Performance matrices for image size [64×64] and patch 12.

Class Validation	precision	recall	f1-score	Support	Class Test	precision	recall	f1-score	Support
Bacterial Spot	0.96	0.96	0.96	67	Bacterial Spot	0.93	0.97	0.95	41
Brown Blight	0.97	0.94	0.95	64	Brown Blight	0.98	1.00	0.99	40
Dry	0.99	1.00	0.99	71	Dry	1.00	1.00	1.00	41
Healthy	0.92	0.93	0.92	58	Healthy	0.97	0.93	0.95	40
Powdery Mildew	0.98	0.98	0.98	61	Powdery Mildew	1.00	0.97	0.99	40
Sooty Mold	1.00	1.00	1.00	69	Sooty Mold	0.97	0.97	0.97	40
accuracy			0.97	390	accuracy			0.98	241
macro avg	0.97	0.97	0.97	390	macro avg	0.98	0.98	0.98	241
weighted avg	0.97	0.97	0.97	390	weighted avg	0.98	0.98	0.98	241

Table 5
Performance matrices for image size [56×56] and patch 7.

Class Validation	precision	recall	f1-score	Support	Class Test	precision	recall	f1-score	Support
Bacterial Spot	0.97	0.91	0.94	67	Bacterial Spot	0.93	0.95	0.94	40
Brown Blight	0.98	0.98	0.98	64	Brown Blight	0.95	1.00	0.98	40
Dry	1.00	1.00	1.00	71	Dry	1.00	1.00	1.00	41
Healthy	0.93	0.95	0.94	58	Healthy	0.97	0.90	0.94	40
Powdery Mildew	0.95	1.00	0.98	61	Powdery Mildew	0.98	1.00	0.99	40
Sooty Mold	1.00	1.00	1.00	69	Sooty Mold	1.00	0.97	0.99	40
accuracy			0.97	390	accuracy			0.97	241
macro avg	0.97	0.97	0.97	390	macro avg	0.97	0.97	0.97	241
weighted avg	0.97	0.97	0.97	390	weighted avg	0.97	0.97	0.97	241

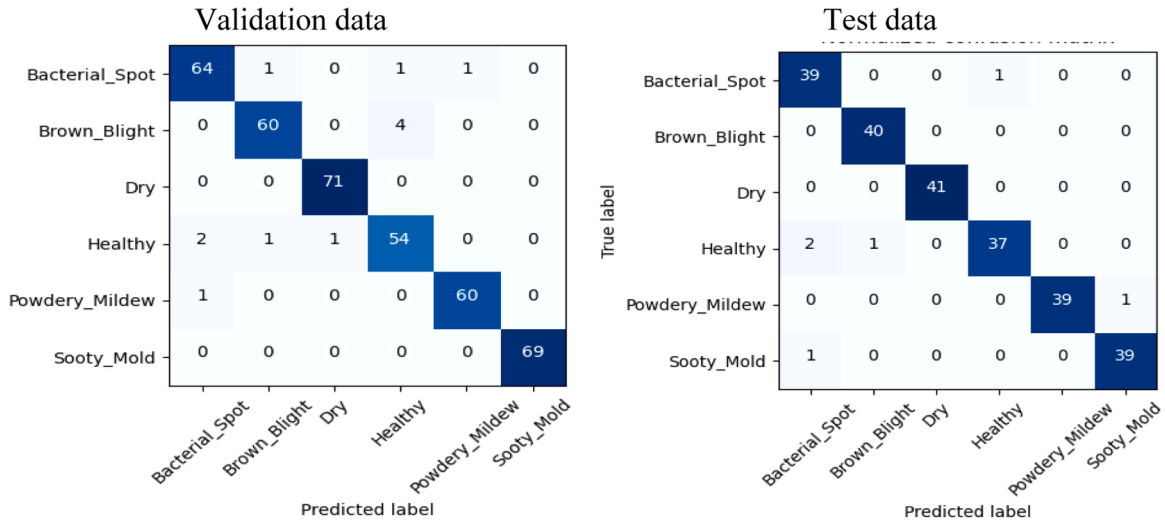


Fig. 2. Confusion matrices for image size [64×64] and patch 12.

4.8. Training loss and validation loss

It is essential to measure the training loss when designing a model. This metric evaluates the performance of fitting training data and compares its predicted output with target values. The goal is to decrease the loss, indicating that the model accurately represents the input-output relationship. It is important to note that the validation loss metric evaluates the model's performance on new data it has not seen before. Figs. 4 and 5 demonstrate the training loss and validation loss over epoch 250 and head six on the combinations of (Image [64×64], patch 12) and (Image [56×56], patch 7). The validation pattern for (Image [56×56], patch 7) is much noisier than (Image [64×64], patch 12). However, the accuracy is higher in the combination of (Image [64×64], patch 12), at 97.51%.

4. Discussion

This research built a customised model based on a vision transformer to diagnose Java Plum leaf disease. This research tested the number of attention heads to achieve the best accuracy and found that this model performed best when the number of attention heads was 6. Various combinations of image and patch sizes have been examined to determine this model's effectiveness. This study shows the performance of this model in different image and patch size combinations, as shown in Table 3. This model obtained the best performance for the image and patch size combination of 64/12. On the other hand, 28/4 took less time to train the dataset. Though the 64/12 combination took more time to train the dataset, it acquired 97.51% accuracy in testing, which is the highest among all combinations, and this accuracy answers RQ2. A comparison of testing accuracy in the experiment is shown in Fig. 6.

This customised ViT model surpasses some previous studies in terms

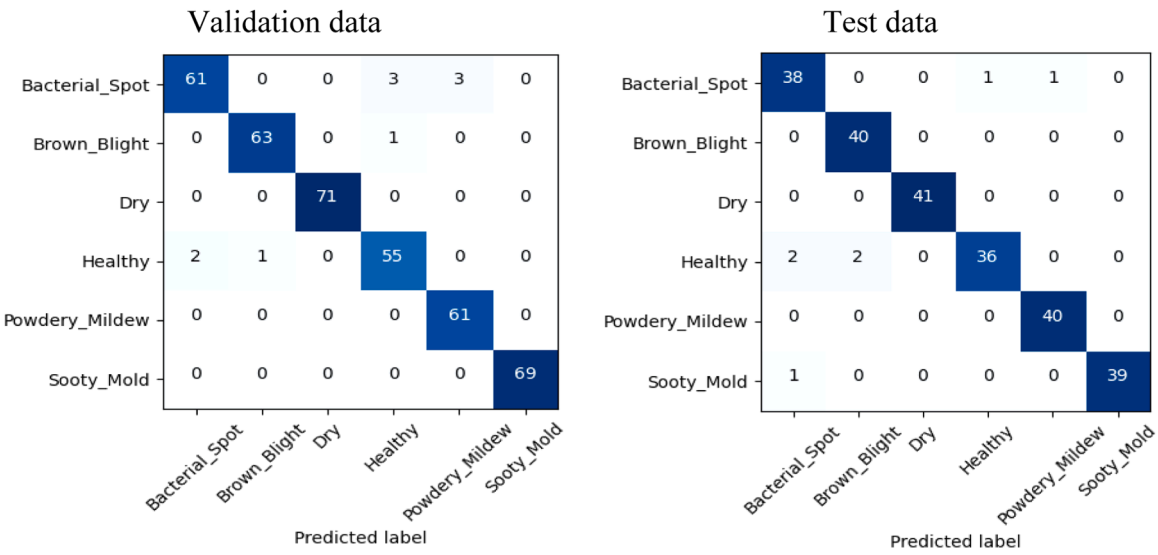


Fig. 3. Confusion matrices for image size [56×56] and patch 7.

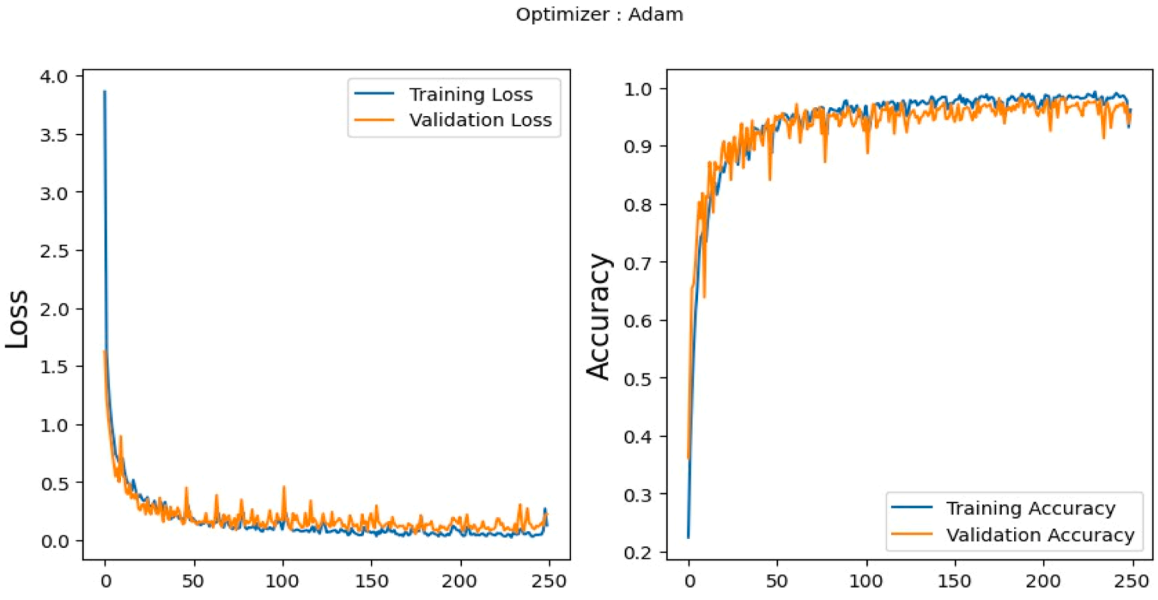


Fig. 4. Accuracy diagram for image size [64×64] and patch 12.

of accuracy. Several research studies have been conducted to detect plant leaf disease and its associated outcomes, as shown in Table 6. Using a Java Plum leaf disease dataset, Mehta et al. [12] developed an innovative approach to Java Plum Leaf Disease Identification technique using Federated Learning meets Convolutional Neural Networks, which achieved 96.35% accuracy. Research conducted by Zhang et al. [3], Zhou et al. [13], and Zeng et al. [18] achieved 93.78% and 88.34% test accuracy, respectively, in plant disease identification, whereas this model achieves 97.51% test accuracy in the field of Java Plum leaf disease. By using this technology, this study can contribute to a broader body of research in tropical crops. This study showcases the continuous advancement of artificial intelligence techniques in precision agriculture by utilising self-attention mechanisms.

5. Inference of the current study

This study investigates an optimised ViT model and its hyperparameters to detect and classify Java Plum leaf disease. The study experimented on the primary dataset using the image and patch size of 64/12, 56/7, 48/6, 32/8, 32/6, 28/7, 28/4, and 21/7. For image size 64 and patch size 12, the optimised ViT obtained the highest accuracy of 97.5% in the validation and test set. Although image size 64 and patch size 12 have the highest accuracy, they take more training time than all other image sizes and patch sizes. Attention head, according to the number of classes, provides the highest accuracy for classification on an image dataset. Epochs are known as 'Model and data-meeting strategies. Epochs need to be used, considering the overfitting and underfitting ratios of the model. Less epochs may cause underfitting, where the model needs to learn more from the data, and more epochs can cause overfitting, where training accuracy will be higher. However, for new data, it will not be able to detect diseases precisely.

6. Limitation

This study finds the fundamental limitations related to the research outcomes. It is crucial to emphasise that the Java Plum Leaf Disease Dataset was gathered during September, October, and November of 2023 in Cumilla (Bangladesh). Additionally, it is essential to recognise

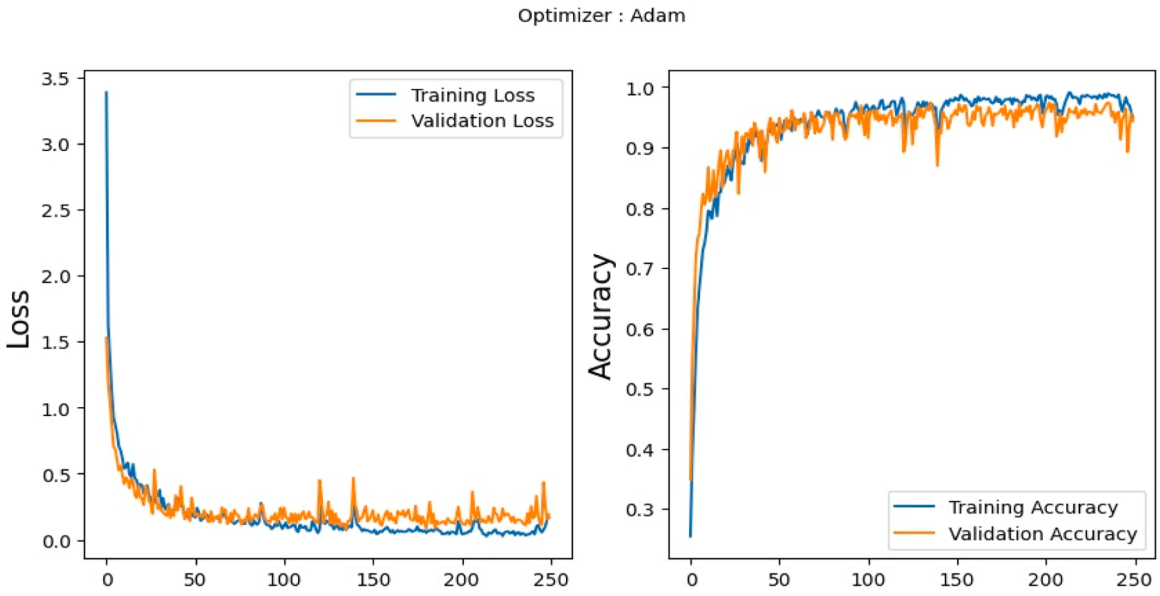


Fig. 5. Accuracy diagram for image size [56×56] and patch 7.

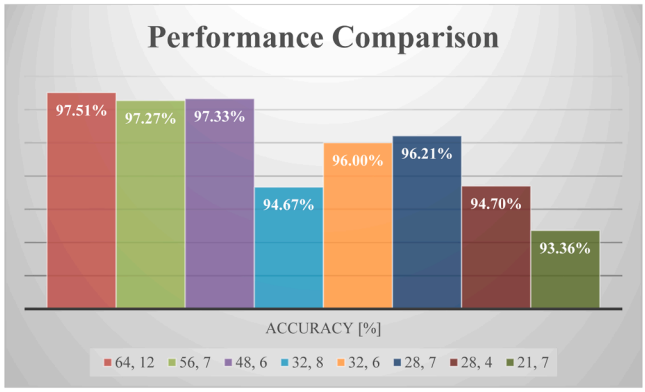


Fig. 6. Performance comparison of eight image sizes and patch configurations.

Table 6
Previous contribution of the authors on plant leaf disease.

Author	Model	Accuracy
Mehta et al. [12]	FL-CNN	96.35%
Zhang et al., [3]	IEM-ViT	93.78%
Zhou et al. [13].	Pre-trained vision transformer	92.00%
Zeng et al. [18].	Squeeze-and- Excitation Vision Transformer (SEViT)	Test 88.34%

that the success of deep learning approaches in disease classification depends on suitable datasets. Obtaining a diverse and comprehensive data set with accurately labelled instances of different diseases presents a significant challenge. Therefore, the proposed ViT models may perform differently in field conditions. Even though ViT shows promising performance in detecting Java Plum leaf disease, it has several limitations. One of the main limitations of ViT is that it uses a self-attention mechanism. However, the effectiveness of this approach depends on several factors, such as dataset size, model architecture, batch size, patch size, head numbers, etc. This model's lack of practical implementation is another limitation of this research work. Although this model demonstrates impressive recognition accuracy, it is crucial to understand that no model is free of errors, and it is also impossible to eliminate the chance of misclassification or false positives. This study does not go into depth to identify and classify other plant diseases, and it only focuses on some limited diseases that can affect Java Plum leaves. This study aims to assist future researchers in interpreting results more accurately by acknowledging these limitations.

7. Contribution

A handful of studies have advocated implementing machine learning models for the advancement of crop production ([6,8,9,27,28]a; [8]b), this research follows the path and presents ViT as an enabling ML tool for precision agriculture. Java plum is an economically significant agricultural product in Bangladesh. Unfortunately, a considerable research gap exists on Java Plum leaf disease in Bangladesh. Besides, the existing dataset is either poor or only consists of a few images. Moreover, very little research has been conducted on Java Plum. This study makes several contributions to lessening the gap. Firstly, this research applies a customised ViT model to detect Java Plum leaf disease. This study informs us about ViT's capabilities and highlights the use of this model in agriculture to aid in crop health management; this contribution answers RQ1. Secondly, although Zhang et al. [3] and Li et al. [14] nicely explained and implemented the ViT model, only a few researchers gave the algorithm of these models. For that reason, developing countries cannot use these models properly to secure the crop quantity. To solve this problem, this paper provides the algorithm of this model. Thirdly, rather than implementing a publicly available dataset, this study applied the method to a primary dataset collected from two Bangladeshi territories that contain six classes. Fourthly, as a developing country citizen, the principal investigator is familiar with these research constraints. This research is a step ahead in fulfilling the research constraints. Lastly, this research is being conducted at the right time to support the current Bangladeshi Government in implementing this model in all possible areas, especially agriculture and the economy. This could help rural development, employment opportunities, and modernisation in farming practices.

8. Conclusion & future scope

This study used the ViT model to learn and categorise five different Java Plum leaf disease classes and a healthy class. The findings of this research work provide a valuable direction for applying the ViT model to address agricultural challenges. Using the self-attention mechanism, the ViT model can learn patterns between infected and healthy images using

a standard transformer encoder as a sequence of patches and processes. This dataset is valuable for training other models and a valuable resource for future agriculture disease detection. This model achieves an impressive accuracy of 97.51%, proving its excellence in identifying smaller image targets and navigating complex scenarios. Future research aims to enhance the accuracy and efficiency of detecting Java Plum leaf diseases using machine learning, encompassing reinforcement learning, hybrid machine learning, and case-based reasoning. Reinforcement learning holds the potential for developing a Vision Transformer (ViT) model capable of autonomously identifying diseased Java Plum leaves, eliminating the need for human intervention. Hybrid machine learning involves integrating multiple models with a traditional machine learning algorithm to enhance the precision of Java Plum leaf disease detection. Case-based reasoning offers the opportunity to build a machine-learning model that learns from the errors of previous models, progressively improving its performance. To enhance the accuracy and performance of Java Plum leaf disease detection using ViT, several challenges must be solved with the help of these innovative research directions. These challenges mainly concern data scarcity, environmental variability, and real-time detection. Although many challenges exist in the research field, ViT emerges as a promising Java Plum leaf disease detection technology. After thoroughly examining different techniques and tackling the challenges, it is creditable to develop ViT models for efficiently identifying Java Plum leaf disease.

Data availability

This dataset was carefully collected from different areas of Bangladesh to make the primary repository of Java Plum leaf information. It has been deposited on Mendeley, a respected platform for academic cooperation and data dissemination. Those involved in research and academia can leverage this dataset to progress their studies, perform analyses, and contribute to the expanding knowledge base concerning Java Plum leaves. The intent behind sharing this data on Mendeley is to promote collaboration, openness, and continued scientific inquiry, enabling the research community to gain valuable insights from this extensive compilation of Java Plum leaf disease data.

Mendeley Dataset: https://data.mendeley.com/datasets/43d75vptz4/3

[Bhowmik, Auvick Chandra; Ahad, Taimur (2024), "Java Plum Leaf Disease Dataset", Mendeley Data, V3, 10.17632/43d75vptz4.3]

Funding statement

The research received no financial support, and none received funding for the work.

CRediT authorship contribution statement

Auvick Chandra Bhowmik: Writing – original draft. **Md. Taimur Ahad:** Writing – review & editing. **Yousuf Rayhan Emon:** Writing – original draft. **Faruk Ahmed:** Writing – review & editing. **Bo Song:** Writing – review & editing. **Yan Li:** Writing – review & editing.

Declaration of competing interest

The authors of this research work declare that they do not have any known competing or personal interests that can appear to influence the work in this paper.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.atech.2024.100500.

References

[1] X. Fu, Q. Ma, F. Yang, C. Zhang, X. Zhao, F. Chang, L. Han, Crop pest image recognition based on the improved ViT method, Inf. Process. Agric. (2023), https://doi.org/10.1016/j.inpa.2023.02.007.
[2] B. Zhan, M. Li, W. Luo, P. Li, X. Li, H. Zhang, Study on the Tea Pest Classification Model Using a Convolutional and Embedded Iterative Region of Interest Encoding Transformer, Biology (Basel) 12 (7) (2023) 1017, https://doi.org/10.3390/biology12071017.
[3] X. Zhang, F. Li, H. Jin, W. Mu, Local Reversible Transformer for semantic segmentation of grape leaf diseases, Appl. Soft Comput. 143 (2023) 110392, https://doi.org/10.1016/j.asoc.2023.110392.
[4] S. Yu, L. Xie, Q. Huang, Inception convolutional vision transformers for plant disease identification, IoT 21 (2023) 100650, https://doi.org/10.1016/j.iot.2022.100650.
[5] Bhowmik, A.C., Ahad, D.M.T., & Emon, Y.R. (2023). Machine learning-based soybean leaf disease detection: a comprehensive review. arXiv preprint arXiv:2311.15741.
[6] M.T. Ahad, Y. Li, B. Song, T. Bhuiyan, Comparison of CNN-based deep learning architectures for rice disease classification, Artif. Intell. Agric. 9 (2023) 22–35, https://doi.org/10.1016/j.aiia.2023.07.001.
[7] K.C. Kamal, Z. Yin, M. Wu, Z. Wu, Depthwise separable convolution architectures for plant disease classification, Comput. Electron. Agric. 165 (2019) 104948.
[8] S.B. Mamun, M.T. Ahad, M.M. Morshed, N. Hossain, Y.R. Emon, Scratch vision transformer model for diagnosis grape leaf disease, in: International Conference on Trends in Computational and Cognitive Engineering, Springer Nature Singapore, Singapore, 2023, pp. 101–118.
[9] Mustofa, S., Munna, M.M.H., Emon, Y.R., Rabbany, G., & Ahad, M.T. (2023). A comprehensive review on Plant Leaf Disease detection using Deep learning. arXiv preprint arXiv:2308.14087. doi: 10.48550/arXiv.2308.14087.
[10] Jeevan, P., & Sethi, A. (2021). Vision Xformers: efficient attention for image classification. arXiv preprint.
[11] R.S. Devi, V.R. Kumar, P. Sivakumar, EfficientNetV2 model for plant disease classification and pest recognition, Comput. Syst. Sci. Eng. 45 (2) (2023), https://doi.org/10.32604/csse.2023.032231.
[12] S. Mehta, V. Kukreja, S. Vats, Innovative approaches to Java Plum leaf disease identification: federated learning meets convolutional neural networks, in: 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), IEEE, 2023, pp. 1–6, https://doi.org/10.1109/ICCCNT56998.2023.10307120.
[13] 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.
[14] G. Li, Y. Wang, Q. Zhao, P. Yuan, B. Chang, PMVT: a lightweight vision transformer for plant disease identification on mobile devices, Front Plant Sci 14 (2023) 1256773, https://doi.org/10.3389/fpls.2023.1256773.
[15] 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.
[16] 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.
[17] S. Parez, N. Dilshad, N.S. Alghamdi, T.M. Alanazi, J.W. Lee, Visual intelligence in precision agriculture: exploring plant disease detection via efficient vision transformers, Sensors 23 (15) (2023) 6949, https://doi.org/10.3390/s23156949.
[18] Q. Zeng, L. Niu, S. Wang, W. Ni, SEViT: a large-scale and fine-grained plant disease classification model based on transformer and attention convolution, Multimedia Systems 29 (3) (2023) 1001–1010, https://doi.org/10.1007/s00530-022-01034-1.
[19] S. Hossain, M. Tanzim Reza, A. Chakrabarty, Y.J. Jung, Aggregating different scales of attention on feature variants for tomato leaf disease diagnosis from image data: a transformer driven study, Sensors 23 (7) (2023) 3751, https://doi.org/10.3390/s23073751.
[20] S. Öğrekçi, Y. Ünal, M.N. Dudak, A comparative study of vision transformers and convolutional neural networks: sugarcane leaf diseases identification, Eur. Food Res. Technol. 249 (7) (2023) 1833–1843, https://doi.org/10.1007/s00217-023-04258-1.
[21] H. Kumar, S. Velu, A. Lokesh, K. Suman, S. Chebrolu, Cassava leaf disease detection using Ensembling of EfficientNet, SEResNeXt, ViT, DeIT, and MobileNetV3 Models, in: Proceedings of the International Conference on Paradigms of Computing, Communication and Data Sciences: PCCDS 2022, Springer Nature Singapore, Singapore, 2023, pp. 183–193, https://doi.org/10.1007/978-981-19-8742-7_15.
[22] Ganguly, A., Tiwari, B., Reddy, G.P.K., & Chauhan, M. (2023). Ensemble learning for plant leaf disease detection: a novel approach for improved classification accuracy. doi: 10.21203/rs.3.rs-3257323/v1.
[23] B. Chang, Y. Wang, X. Zhao, G. Li, P. Yuan, A general-purpose edge-feature guidance module to enhance vision transformers for plant disease identification, Expert. Syst. Appl. 237 (2024) 121638, https://doi.org/10.1016/j.eswa.2023.121638.
[24] A. Diana Andrushia, T. Mary Neebha, A. Trephena Patricia, S. Umadevi, N. Anand, A Varshney, Image-based disease classification in grape leaves using convolutional capsule network, Soft Comput. 27 (3) (2023) 1457–1470, https://doi.org/10.1007/s00500-022-07446-5.
[25] B. Hu, W. Jiang, J. Zeng, C. Cheng, L. He, FOTCA: hybrid transformer-CNN architecture using AFNO for accurate plant leaf disease image recognition, Front. Plant Sci. (2023) 14, https://doi.org/10.3389/fpls.2023.1231903.

[26] F. Arshad, M. Mateen, S. Hayat, M. Wardah, Z. Al-Huda, Y.H. Gu, M.A. Al-antari, PLDPNet: end-to-end hybrid deep learning framework for potato leaf disease prediction, Alexandria Eng. J. 78 (2023) 406–418, https://doi.org/10.1016/j.aej.2023.07.076.

[27] F. Ahmed, Y.R. Emon, M.T. Ahad, M.H. Munna, S.B. Mamun, A fuzzy-based vision transformer model for tea leaf disease detection, in: International Conference on Trends in Computational and Cognitive Engineering, Springer Nature Singapore, Singapore, 2023, pp. 229–242.

[28] Ahmed, F., Ahad, M.T., & Emon, Y.R. (2023b). Machine learning-based tea leaf disease detection: a comprehensive review. arXiv preprint arXiv:2311.03240.