SEMARAK ILMU PUBLISHING

Semarak International Journal of Machine Learning

Journal homepage:
https://semarakilmu.my/index.php/sijml/index
ISSN: 3030-5241

A Cross-Domain Transfer Learning Framework for Robust Plant Disease Identification from Leaf Imagery

Kamal Kumar Srivastava[1,*], Choo Wou Onn[2], Vivek Kumar[3], Sameerchand Pudaruth[4], Suman Kumar Mishra[5]

[1] Department of Information Technology, Babu Banarasi Das Northern India Institute of Technology, Lucknow, India
[2] Faculty of Data Science and Information Technology, INTI International University, Nilai, Malaysia
[3] School of Computer Science Engineering and Technology, Bennett University, Greater Noida, India
[4] ICT Department, University of Mauritius, Mauritius
[5] Department of Computer Science & Engineering, Khwaja Moinuddin Chishti language University, Lucknow, India

ARTICLE INFO

Article history:
Received 29 April 2025
Received in revised form 11 June 2025
Accepted 30 June 2025
Available online 15 July 2025

Keywords:

Food crops diseases; image recognition; vision transformer; attention mechanism; deep learning

ABSTRACT

Harvest efficiency depends on the plant health and climate variations. Identification of leaf disease at right time, using leaf imagery, certainly helps in treating the crop and getting good yield. Recent advancements in transformer-based models, particularly Vision Transformers (ViT), have revolutionized agricultural image analysis by capturing complex, non-linear patterns. Despite their effectiveness, ViTs require large, labelled datasets, posing challenges in plant disease identification due to similar symptoms and limited data. This study explores data augmentation and transfer learning with two ViT variants, ViT-Base and ViT-Large, using training-from-scratch and feature extraction techniques. The Robo-Flow augmentation combined with ViT-GZSL feature extraction achieved 96.62% accuracy. The work uses transfer learning in the background. ViT-Base excelled in classifying corn, chili, tea, and tomato diseases, while ViT-Large performed best on peanut and potato crops, reaching up to 98.85 % accuracy.

1. Introduction

Crop diseases rot the whole crop field and ends up in monetary loss of farmer. Leaf imagery may identify disease of the crop by inspecting the leaf. Deep learning applied to various tasks such as natural language processing, speech recognition, medical applications, and computer vision deep models may be particularly helpful in identifying the intricate and non-linear relationships hidden in imaging data [1,2]. Deep learning has undergone a revolution thanks to the recent invention of transformer-based deep models. Vision transformer (ViT) has shown its superiority across numerous studies. Vision transformer with self-attention-based architecture utilizes attention patterns in image recognition tasks like convolutional layers in CNN architecture. Vision transformer has demonstrated superior performance compared to state-of-the-art CNN ResNet and Efficient

* *Corresponding author.*
E-mail address: 2007.srivastava@gmail.com

https://doi.org/10.37934/sijml.6.1.1124

Net and requires significantly lower training computational resources for image classification tasks [20]. ViT employs self-attention mechanism as the main operation to learn spatial relationships among patches based on pixel representations from input images. Vision transformer exhibits advantages by retaining local image information within feature patches, without processes causing image resolution degradation, thus preventing loss of spatial information due to information skipping (such as max pooling, stride convolution, global average pooling) [5,7-9].

Therefore, the researchers decided to investigate deeply, especially to determine our current position in the agricultural field. The agricultural sector plays a crucial role in promoting employment, the national economy, poverty alleviation, food security, and competitiveness. The availability of food crop commodities plays a key role in maintaining the stability of food security amid increasing consumption rates and population growth. The recorded increase in consumption of potato is 0.83kg/capita/year, corn 0.21 kg/capita/year, cassava 1.9 kg/capita/year, and rice experience a decrease of 2.2 kg/capita/year. However, disruptions caused by plant diseases have led to a decrease in productivity and harvest quality. On average, there is a 22.6% loss in corn yield, 17.2% in potato yield 54% in bean yield up to 30% in tomato yield, 0.78% in tea yield and a 22% decrease in Indonesian chilli yield. The decrease in productivity of harvest results impacts food security instability and causes losses to farmers. Based on morphological symptoms and local symptoms, the identification of plant diseases may be conducted by observing changes in colour, shape, and texture in specific parts of the plant such as leaves [11,14,15,18]. Leaves are the most responsive part of the plant to environmental changes, thus allowing for the identification of diseases based on morphological symptoms on local symptom leaf objects. However, few diseases exhibit identical symptoms, making identification difficult and requiring more precision, Convolutional neural Network which are deep learning techniques in computer vision, have been widely used in image classification tasks, especially disease identification based on leaf image symptoms [6,13,16,21]. A similar work on precision agriculture using IoT technology is conducted for coffee beans production. CNNs, on the other hand, stack more convolution layers to progressively aggregate features from local to global. However, the vision transformer model employs the multi-headed self-attention mechanism, which allows the model to focus on each element in the input sequence, to capture long-range interactions [22].

The continuously growing huge architectures have enabled the ViT to achieve remarkable success. However, the vast number of parameters begins to demand hundreds of millions of labelled data which are often publicly inaccessible in imaging domain causing insufficient training data capacity to train large model architectures. Insufficient leaf disease training data is an unavoidable problem in generating models. A promising method to tackle this problem is the transfer learning, it tries to transfer the knowledge from the source domain to the target domain [4]. Ever-evolving large architecture has allowed ViT to achieve remarkable success. The vision transformer architecture trained on the large imagenet-21k dataset of 12 million images with 21,000 classes has provided good accuracy results when the model is transferred to tasks with fewer data points such as Image Net with 88.55% accuracy, Image Net-Real with 90.72% accuracy, CIFAR-100 with 94.55% accuracy, and the VTAB suite with 77.63% accuracy. Model performance is not only influenced by the adequacy of training data with the model architecture complexity but also the quality and diversity of the data, Augmentation is a technique that can increase data diversity but the performance of the ViT model using the ImageNet-21k dataset showed a decrease when using the AugReg augmentation technique [12].

The difficult task of identifying plant diseases from leaf pictures is addressed in this work by introducing a novel approach to cross-domain transfer learning. Our method leverages an existing ViT model that was adapted for the leaf imaging dataset after it was trained on the Image Net-21K

dataset. The following is a summary of the four works' main contributions: Initially, we tried a variety of model training approaches, including starting from scratch, extracting features from the ViT model, and employing multiple data augmentation methods including Tensor Flow's Image Data Generator, the Robo flow platform, and Keras API layer augmentation. This investigation aimed to assess how combining augmentation techniques with different model training strategies impacts overall model performance. Secondly, we evaluated the performance of pre-trained transfer learning models, specifically ViT-Base and ViT-Large variants, on a diverse dataset comprising leaf disease images from multiple plant species. This comparative analysis aimed to determine which model variants offer the most effective performance in identifying specific plant diseases. Lastly, we explored the effects of data reduction on model performance by systematically reducing the training dataset size and comparing the performance of the proposed model variants. This experiment aimed to understand how varying training data sizes affect the effectiveness of our proposed models in disease identification.

2 Materials and Methods

2.1 Dataset Description

The leaf diseases images dataset used to support the findings of this research is the Plants Diseases Dataset. The data is the result of data enrichment performed on the unbalanced data source from Agro AI Crop Deep from GitHub [19], which is integrated with other data sources, such as corn plant diseases data potato plant diseases, tomato plant diseases data from Tomato Leaves Dataset chili plant diseases data, tea plant diseases data from Tea Leaf Disease dataset in Figure 1 and Table 1 shown sample images of diseased and healthy leaf classes, and the amount of data per class of each plant.

(a)

(b)

(c)

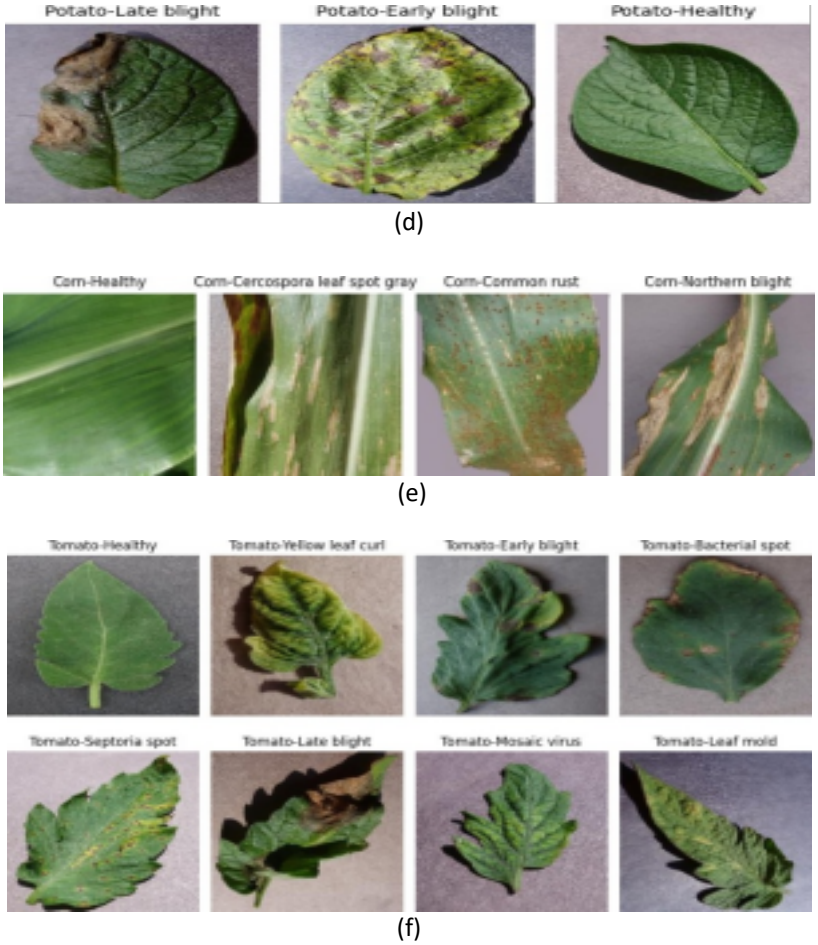

Fig.1. Sample images of diseased and healthy leaves of bean, (b) chili, (c) potato, (d) corn, (e) tea, (f) tomato

Table 1
Number data per class for each plan

Food Crop	Disease Class	Number data per class	Total Data
Corn	Cercospora leaf's pot grey, common rust, northern blight, healthy	711	2.844
Potato	Early blight, late blight, healthy	1582	4746
Bean	Anthracnose, rust, healthy	1200	3.600
Tomato	Bacterial spot, early blight, late blight, leaf mold, septoria spot, target spot, yellow curl, mosaic virus, healthy	745	5960
Chili	Bacterial spot, yellow leaf curl, healthy	465	1.395
Tea	Algal, brown blight, grey blight, red spot, healthy	835	4.175

Displays one leaf object that indicates a disease or healthy. The image validation process was also carried out with the help of a phytopathology who has knowledge in this field. Valid diseases data is ensured to have a balanced distribution of the amount of data per class in each crop. In this study, we split this data of each plant object into training/validation/testing subsets equal to 80%/10%/10%.

2.2 Vision Transformer

Utilizing knowledge from the associated domain (source domain) to enhance performance in the target domain or downstream ask is the goal of transfer learning vision transformer models. In addition, this approach can reduce training time and cost and overcome the need for extensive training datasets. ViT has been trained using the ImageNet-21k database. The pre-trained ViT is transferred and fine-tuned for smaller downstream tasks, thus removing the pre-trained classification head.

2.3 Feature Extraction ViT-GZSL

The feature extraction process is conducted using the generalized zero-shot learning (GZSL) technique, utilizing the attribute attention module (AAM) from pre-trained ViT. This technique takes patch features and image attributes as inputs, the next features related to those attributes. The process involves merging patch features with an attention mechanism to identify unseen classes from new data sources. We discuss the detailed computation process of AMM. Figure 2 shows the architecture of AAM. Suppose €ZRN×D denoted input, where Nand D are the row and column of a matrix. Were-shape Z and define query (Q), key(K) and value (V) as

$$Q=ZW^{q},\ K=ZW^{k},\ V=ZW^{v} \tag{1}$$

Where W^{q}, W^{k}, W^{v} are the train able weight matrices of

$$Attention(Q,K,V) = softmax\frac{QK^{T}}{\sqrt{d}v}V \tag{2}$$

Eq.2 is the Attention(Q,K,V) based on the pair wise similarity between the representation soft elements i.e. query qi and key kj and to calculate the similarity between each pair of tokens through the process of scale dot product, dV which means approximate normalization, which reduces by division against d. The soft max operation is applied to convert the self-attention score values into probabilities. Then calculating the weighted sum over all V values in the sequence produces a feature map. AAM as a feature extraction method is only trained with visible class images and attributes. The base pre-trained model is not fine-tuned, but rather utilizes the weight so flayers that have learned tasks from previous datasets.

2.4 Training ViT from Scratch

Training from scratch ViT models is the process of training the model from the initial layer with random parameter initialization and not using any pre-learned representations or knowledge from other models or pre-trained weights. This means that the model starts with zero knowledge of the task to be solved and must learn from the plant disease data provided during the training process. This process involves initializing weights and hyper parameters, regularization and tuning other parameters for model training.

2.5 Conditional Variational Auto Encoder (CVAE)

CVAE is used to synthesizing ViT features from image attributes. CVAE employs two networks, an encoder and a decoder, which are trained to maximize the conditional probability p(XViTa). For

the feature extraction process, the encoder maps the ViT feature (XViT) and image attribute (a) into a latent vector(Z).

The CVAE's encoder and decoder are built using a multi-layer perceptron (MLP) with a single hidden layer. Furthermore, the pre-trained classification head that is removed in Figure 3 upon transfer is replaced by the CVAE [5,9]. For unseen classes, we create synthetic ViT feature samples using a fully connected layer. In the fully connected layer, we employ a dropout layer to regularize the training process, a ViT feature classifier that uses the supervised learning model of the softmax classifier to forecast the class, and CVAE reconstruction by expanding the MLP depth to two hidden layers. Adjust the completely connected layer's weights. Weights will have to be adjusted from general feature maps to features during the training phase.

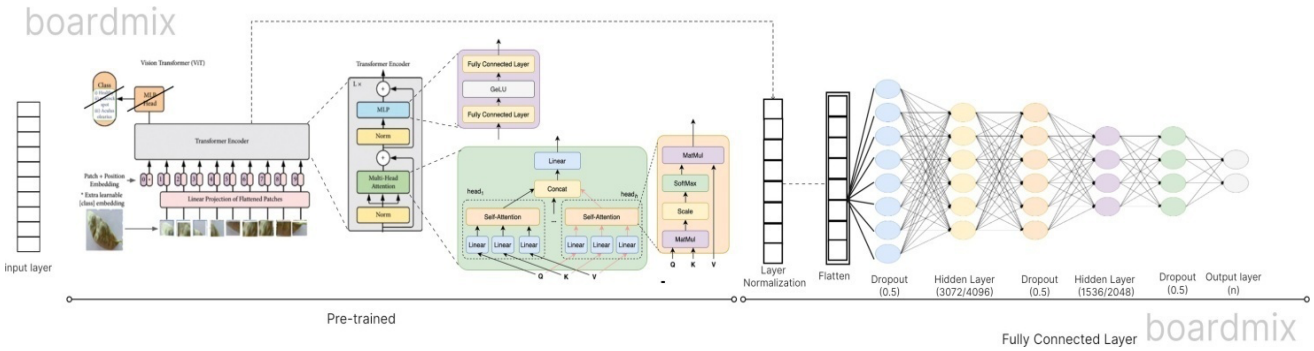

Fig.2. Architecture

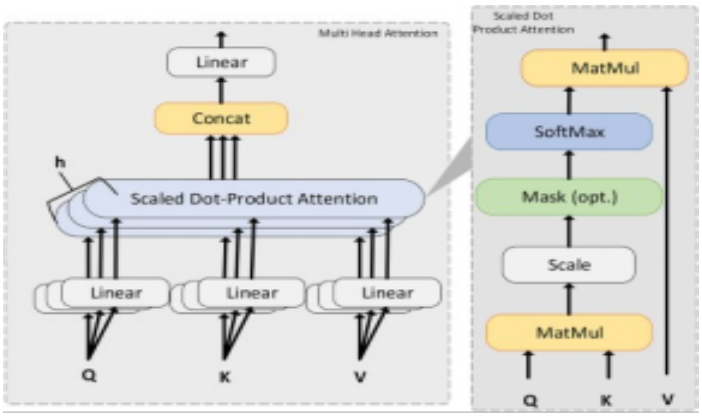

Fig.3. Attribute attention module

2.6 Hyper parameter

ViT-B/16 and ViT-L/16, the ViT-Base and ViT-Large model variations were trained using input patch sizes of 16×16 [3]. Adam outperforms SGD in crucial tasks including attention models. However, Loshchilov and Hutter also empirically reported that Adam with decoupled weight decay (AdamW) optimizer generalizes substantially better than Adam with L2 regularization optimizer [10]. All experiments in this study used the AdamW optimizer with parameter values including weight decay = 0.0001, learning rate = 0.001, β1=0.9, β2=0.999. Wefine-tune dat384 resolution, a batch size of 32, we else used the Categorical Cross Entropy function to compute loss, and GELU as the activation function.

2.7 Callback

Early Stopping callback is used to automatically halt training when the validation accuracy as the monitored metric, does not improve. The purpose of using the Early Stopping callback is to prevent over fitting and conserve time and computational resources by discontinuing further model training without clear benefits. All experiments in this study used Early Stopping with sets the patience to10, meaning the number of periods without improvement thereafter will halt the training process.

2.8 Evaluation Metrics

We estimated classification jobs using a portion of testing data from the dataset and evaluated the model's performance using evaluation measures like accuracy, precision, recall, and F-score. Furthermore, we quantify how well the model predicts the output or target. Accuracy to evaluate how closely the observations predicted by the model approach the magnitudes of the observations.

$$\mathbf{ACC} = \frac{TP+TN}{TP+TN+FP+FN} \tag{3}$$

Loss is used to evaluate the loss by calculating the difference between the predicted probability distribution and the actual probability distribution.

$$\mathrm{L}= \sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log(\hat{y}_{i,c})\frac{1}{N} \tag{4}$$

$$MAR = \frac{1}{n}\sum_{i=i}^{n}\frac{TPi}{TPi+FNi} \tag{5}$$

Macro Average Precision (MAP) evaluates the proportion of Predicted Positive cases that are correctly Real Positives.

$$MAp = \frac{1}{n}\sum_{i=i}^{n}\frac{TPi}{TPi+FPi}$$

(6)

Macro Average F1-score (MAF) evaluates model performance to identify positive and negative classes based on the balance between precision and recall values by Powers *et al* [16].

$$\mathrm{MAF}= 2 * \frac{MAR*MAP}{MAR+MAP} \tag{7}$$

Where FP stands for false positives, FN for false negatives, L for loss, TP for true positives, TN for true negatives, the acronym MAR stands for macro average recall. N stands for the number of testing data, MAP for macro average precision, and MAF for macro average f1-score and C refers to number of classes, yi, c refers to label or target, yˆi, c prediction probability value.

Applying the three proposed augmentation techniques consisting of API layer augmentation, image data generator and robo-flow platform. We also add by building a fully connected layer for

fine-tuning the downstream tasks Figure 2 which consist of Conditional Variation Auto encoder (CVAE) uses multi-layer perceptron (MLP) with two hidden layers, we set the number of first hidden units to 3071 and the second hidden units to 1536, dropout layer of 0.5 in each hidden layer. We trained our model for 25 epochs with parameter values including weight decay = 0.0001, learning rate = 0.001, β1=0.9, β2=0.999.

2.9 Different Pre-trained Model Variants and Plant Objects

This experiment adopts the best combination of Augmentation techniques and training strategies from previous experiments to be applied to two pre-trained variants including ViT-B/16 and ViT-L/16 for separate downstream tasks on six plants are shown inTable1. We also add the fully connected layer which consists of Conditional Variation Auto encoder (CVAE) uses multi-layer perceptron (MLP) and dropout layer are summarized in Figure 2 and Table 3.

2.10 Different Numbers of Training Data and Pre-trained Model Variants

Specifically, to understand how different training data counts affect the model variants' performance, we conducted two training rounds on potato and bean crops using different quantities of data on the two model variations stated in Table 4. Involved training the model for 50 epochs with hyper parameter as explained.

Table 3
Variants vision transformer

	Attribute	ViT-B	ViT-L
Pre-trained	Layers	12	24
	Hidden Size D	768	1024
	MLP Size	3072	4096
	Heads	12	16
	Parameters	86M	307M
FCL	1stHiddenUnits	3072	4096
	2^{nd} Hidden Units	1536	2046
	Dropout	0.5	0.5

Table 4
Number of training data

Plants	Versions	Train/Class	Validation/Class
Potato	V1	2.532	361
	V2	1.200	240
Bean	V1	1.920	240
	V2	1.200	150

3. Experiments and Results

We compared the combination of three augmentations Techniques (API keras layer sequential, image data generator using tensor flow framework and robo-flow platform) and two training strategies against model transfer (training from scratch and feature extraction ViT-GZSL). Other augmentation techniques (Mixup and Rand Augment) have been reported in previous studies to increase data diversity but did not help improve performance [17]. In this experiment, we utilized the augmentation parameters provided by each of the proposed augmentation techniques. The parameters are applied to subset of training and validation data resulting in different output

characteristics of each Technique. Table2 is the use of the provided parameters of each augmentation technique along with their values. The three augmentation techniques provide different parameters, characteristics and outputs. Augmentation technique of API keras layer sequential performs random augmentation on the image during the training process (real time), and the output produces images that have various effects without increasing the amount of data. Augmentation technique of image data generator generates data to be augmented before the training process, and the output produces images that have various effects without increasing the amount of data. Augmentation technique of tools robo-flow performs augmentation on data before the training process to generate new additional data with an increased number of images. Figure 4 and Figure 5 show the results of the three augmentation techniques on some sample data.

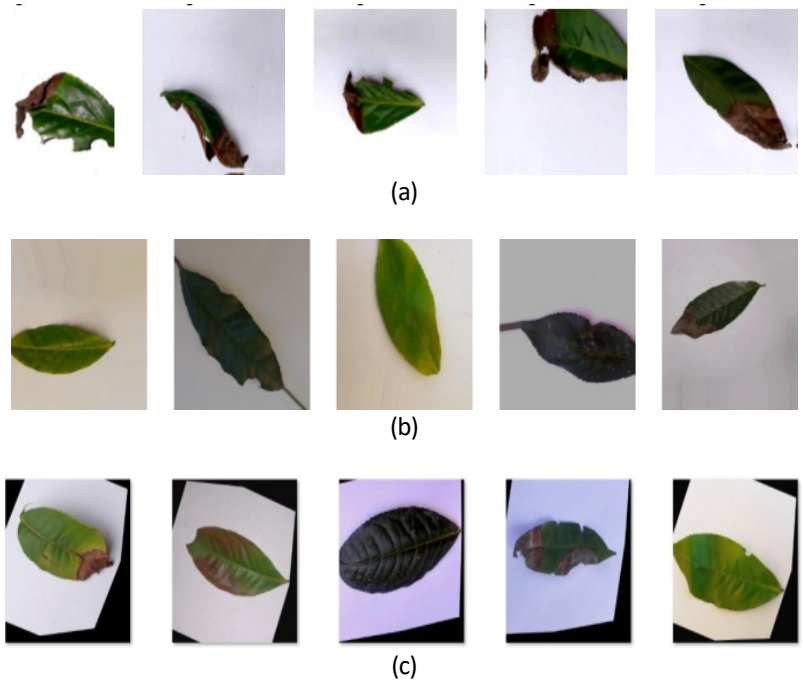

Fig. 4. Augmented sample images of (a) API keras layer sequential, (b) Image data generator tenfor flow, (c) Roboflow

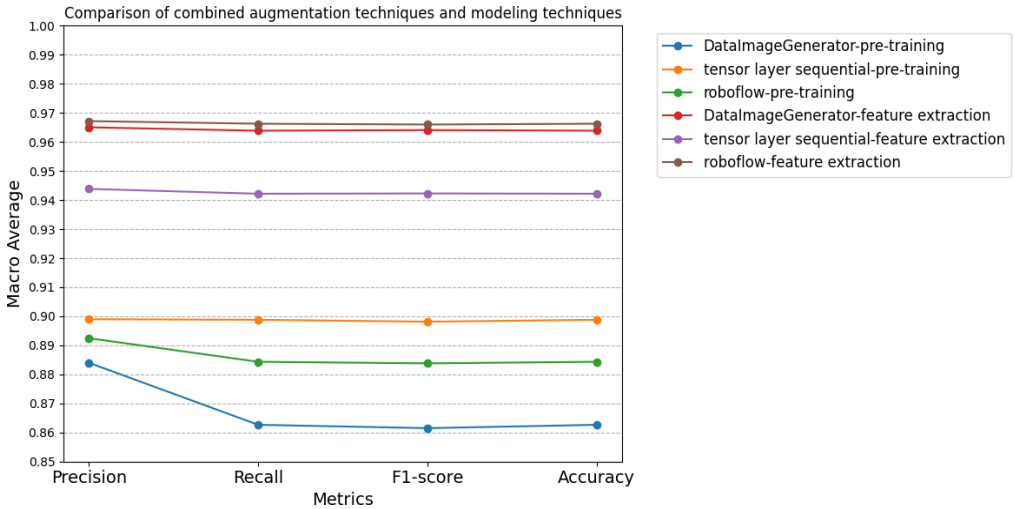

Fig.5. Comparison performance evaluation of combination augmentation techniques and model training strategies

Table 5
Performance of model variants on identifying diseases in each plant

	Training Strategies	Evaluation Matrix			
		Precision	Recall	F1-Score	Accuracy
Image Data	Training from scratch	88.4	86.2	86	86.26%
Generator	Feature extraction	96.6	96.2	96.6	96.38%
Layer Sequential	Training from scratch	90	89.8	89.6	89.87%
	Feature extraction	94.4	94.2	92.6	94.21%
Robo flow	Training from scratch	89.4	88.6	88.6	88.43%
	Feature extraction	96.8	96.8	96.6	96.62%

We evaluate and compare the effectiveness of different learning model variants, including ViT-B/16andViT-L/16 to determine the most effective classifier model for plant disease identification. The objective of this study was to evaluate the effectiveness of the variants of the vision transformer model on a subset of test data for each crop. Table6 shows the results. In the corn plant object scenario, the ViT-B/16 mode variant has the most effective performance obtaining the highest matrix evaluation in terms of accuracy to 98.94%, precision to 0.9894,recall to 0.9899, f1-score to 0.9894, in addition to ViT-B/16 also shows minimum loss value to 0.0349,in addition ViT-B/16 also obtained a higher correct percentage when evaluated using other test data samples. In the potato plant object scenario, the ViT-L/16 model variant has the most effective performance obtaining the highest matrix evaluation in terms of accuracy to 98.95%, precision to 0.9895, recall to 0.9895, f1-score to 0.9895, ViT-L/16 also shows the minimum loss value to 0.0458, in addition ViT-L/16 also obtained a higher correct percentage when evaluated using other test data samples. In the chili plant object scenario, the ViT-B/16 mode variant has the most effective performance obtaining the highest matrix evaluation in terms of accuracy to 99.28%, precision to 0.9928, recall to 0.9929, f1-score to 0.9928, ViT-B/16 also shows minimum loss value to 0.034, in addition ViT-B/16 also obtained a higher correct percentage when evaluated using other test data samples. In the tea plant object scenario, the ViT-B/16 mode variant has the most effective performance obtaining the highest matrix evaluation in terms of accuracy to 97.83%, precision to 0.9783, recall to 0.9792, f1-score to 0.9784, ViT-B/16also shows minimum loss value to 0.0536, in addition ViT-B/16 also obtained a higher correct percentage when evaluated using other test data samples. In the tomato plant object scenario, the ViT-B/16mode variant has the most effective performance obtaining the highest matrix evaluation in terms of accuracy to 97.97%, precision to0.9797, recall to0.9807, f1-score to 0.9797, ViT-B/16 also shows minimum loss value to 0.0536, in addition ViT-B/16 also obtained a higher correct percentage when evaluated using other test data samples. In the bean plant object scenario, the ViT-L/16 mode variant has the most effective performance obtaining the highest matrix evaluation in terms of accuracy to100%, precision to1.00, recall to1.00, f1-score to1.00,ViT-L/16 also shows minimum loss value to 1.9 e-07, in addition ViT-L/16 also obtained a higher correct percentage when evaluated using other test data samples .We found that vit-b/16 is predominantly effective for identifying diseases of 4plant objects. ViT-B/16 model has higher accuracy and ability to find true positive instances and lower loss value when predicting data than ViT-L/16on the disease identification task of jaung, chili, tea and tomato. While in the case of the other two plants, such as bean and potato, we found that ViT-L/16 is more effective in terms of precision and the ability to identify true positive instances is higher, with lower loss values when predicting data compared to ViT-B/16.”

Table 6
Performance of Model Variants on Identifying Diseases in Each Plant

Food Crops	Model Variants	Evaluation Matrix				F1 Score	True Percentage
		Accuracy	Loss	Precision	Recall		
Corn	ViT-B/16	98.94%	0.0349	0.9894	0.9899	0.9894	90%
	ViT-L/16	7.89%	0.0574	0.9789	0.9796	0.9789	85%
Potato	ViT-B/16	8.31%	0.0564	0.9831	0.9834	0.9831	86%
	ViT-L/16	8.95%	0.0458	0.9895	0.9895	0.9895	100%
Chili	ViT-B/16	9.28%	0.0340	0.9928	0.9929	0.9928	100%
	ViT-/16	98.5%	0.0352	0.9855	0.9858	0.9854	73%
Tea	ViT-B/16	7.83%	0.0536	0.9783	0.9792	0.9784	84%
	ViT-/16	7.11%	0.1470	0.9711	0.9721	0.9711	60%
Tomato	ViT-B/16	97.97%	0.0682	0.9797	0.9807	0.9797	95%
	ViT-/16	97.47%	0.0620	0.9747	0.9752	0.9745	80%
Bean	ViT-B/16	9.72%	0.0219	0.9972	0.9972	0.9972	87%
	ViT-L/16	1.00%	1.9e-07	1.00	1.00	1 .00	93%

3.1 Effect of Varied Training Data Sizes on Pre- Trained Model Performance

In this section, we report the influence of model variant performance using different sizes of training data described in sub subsection 2.1.2 based on the evaluation matrix described in subsection 1.5. Table7, in the potato plant object scenario, the ViT-L/16 variant model in both versions outperforms the ViT-B/16 of the same version based on the evaluation matrix. Therefore, reducing the training data from version 1 by 2532 to version 2 by 1200 does not affect the attainment of the best model variant. However, it affects the effectiveness of performance on each variant model. The performance of the ViT-B/16 variant model in version 1 is more effective compared to version 2, indicating that the effectiveness of the ViT-B/16 model decreases with data reduction. Meanwhile, the performance of the ViT-L/16 variant model in version 2 is more effective compared to version1, indicating that ViT-L/16improves with data reduction. Table in the bean plant object scenario, the ViT-L/16 variant model in version 1 outperforms the ViT-B/16 in the same version, while ViT-L/16 and ViT-B/16 from version 2, as well as ViT-B/16 from version1, show evaluation matrices that do not significantly differ. The process of reducing the training data version 1 by 1920 to version 2 to 1200 for bean plants does not greatly affect the effectiveness of the ViT-B/16variant model's performance, but it does impact the decline in ViT-L/16's performance From the experiments conducted we found that efficiency of the model performance may be affected not only by the augmentation techniques, model training strategies and model variants used, but also background images from the training data on the main imitations of ViT " the straight forward. Character datasets, including namely a potato dataset that displays a full leaf on the background as shown in Figure 1 and a bean dataset without a background whose image shows a full leaf with lines or edges that are truncated due to the enlarged image as shown in Figure 1. We hypothesize that, when the model learns the potato dataset, it is considered a challenge because it has a back ground and is unable to model the local structure of leaf edges and lines, but it should not be too difficult if the intensity of the background pixels is different from them a in object, on the other hand, when the model learns the bean dataset, it is considered to anticipate the weakness of ViT with data without visible background and leaf edges. We find that both variations of the model demonstrated nearly comparable performance effectiveness when trained with the

bean dataset, and that the evaluation matrix performance in both variants of the bean classification model was more optimal than the performance of the potato classification model.

Table 7
Comparison of model variants performance with varied sizes of potato training data

Food Crops	Model Variants	Evaluation Matrix				F1-Score	True Percentage
		Accuracy	Loss	Precision	Recall		
Version1 Potato	ViT-B/16	98.31%	0.0564	0.9834	0.9831	0.9831	87%
	ViT-L/16	98.95%	0.0458	0.9895	0.9895	0.9894	100%
Version2 Potato	ViT-B/16	97.89%	0.0626	0.9796	0.9789	0.9787	100%
	ViT-L/16	99.16%	0.0312	0.9916	0.9916	0.9916	100%
Version1 Bean	ViT-B/16	99.72%	0.0219	0.9972	0.9972	0.9972	87%
	ViT-L/16	1.00%	1.00	1.00	1.00	1.00	100%
Version2 Bean	ViT-B/16	99.72%	0.0148	0.9972	0.9972	0.9972	93%
	ViT-L/16	99.72%	0.0219	0.9972	0.9972	0.9972	93%

Nevertheless, the classification model performs better on data with a background when evaluated on a small dataset related to field circumstances. Such as edges and lines, because the input images are tokenized via hard splitting. VIT's inability to model the local structure of the image, such as the edges and lines of the leaf as the boundary of the leaf object with its background, results from the training process' feature extraction from the self-attention module, which learns the input image representation through simple tokenization that divides into patches. Through the experiment sub subsection 2.2.3, we have indirectly observed the weakness of ViT when training two variants of pre-trained ViT is unable to model the local structure of the image, model through feature extraction on two different.

4. Conclusion

In this study, we propose a learning model using a Model transfer approach using pre-trained vision transformer variants from the Image n ET-21K database for downstream tasks on identifying diseases in crop plants consisting of corn, potato, chili, tomato, tea and bean. Before evaluating the performance of the proposed pre-trained vision transformer variants to identifying diseases of these crops, we first evaluated the effectiveness of the combination of three augmentation techniques and two training strategies of the ViT pre-trained model in improving the model performance. We found that the combination of robo flow augmentation technique with feature extraction ViT-GZSL training strategy is most effective to improving the model performance with accuracy to 96.62%. Next, we adopted the best combination choice to be applied to both pre-trained ViT variants to identify the six plant diseases one by one. The goal is to find out the most effective performance model variant for diseases identifying of each crop. We compared the performance of each model based on matrix evaluation values using test datasets. We found that the ViT-B/16 model variant predominantly produced the optimal disease classification identification with accuracy of 98.85% and 100% respectively. In addition, we compared the performance of each model variant by training on reduced data, it was found that data reduction affected the effective performance of model variants in the case of potato crop disease identification but did not have a significant effect on bean crops model, including the identification of diseases of corn, chilli, tea, and tomato plants with accuracies of 98.94%, 99.28%, 97.83%, 97.97%respectively. While we obtained effective disease classification model performance from ViT-L/16 model variant for potato and bean disease.

Acknowledgement
The author wishes to thank all the authors for their encouragement in pursuing this line of research as well as for inspiring discussion.

References:

[1] Dar, Jawad Ahmad, Kamal Kr Srivastava, and Sajaad Ahmed Lone. "Design and development of hybrid optimization enabled deep learning model for COVID-19 detection with comparative analysis with DCNN, BIAT-GRU, XGBoost." *Computers in Biology and Medicine* 150 (2022): 106123. https://doi.org/10.1016/j.compbiomed.2022.106123

[2] Dar, Jawad Ahmad, Kamal Kr Srivastava, and Alok Mishra. "Lung anomaly detection from respiratory sound database (sound signals)." *Computers in Biology and Medicine* 164 (2023): 107311. https://doi.org/10.1016/j.compbiomed.2023.107311

[3] Gabeur, Valentin, Chen Sun, Karteek Alahari, and Cordelia Schmid. "Multi-modal transformer for video retrieval." In *European Conference on Computer Vision*, pp. 214-229. Cham: Springer International Publishing, 2020. https://doi.org/10.1007/978-3-030-58548-8_13

[4] Iman, Mohammadreza, Hamid Reza Arabnia, and Khaled Rasheed. "A review of deep transfer learning and recent advancements." *Technologies* 11, no. 2 (2023): 40. https://doi.org/10.3390/technologies11020040

[5] Kim, Jiseob, Kyuhong Shim, Junhan Kim, and Byonghyo Shim. "Vision transformer-based feature extraction for generalized zero-shot learning." In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1-5. IEEE, 2023. https://doi.org/10.1109/ICASSP49357.2023.10095217

[6] Kumari, Neeraj, Vivek Kumar, Neeraj Kumar Pandey, Amit Kumar Mishra, and Deepak Kholiya. "Tomato Leaf Disease Detection and Identification using Machine Learning." In *2023 3rd International Conference on Innovative Sustainable Computational Technologies (CISCT)*, pp. 1-5. IEEE, 2023. https://doi.org/10.1109/CISCT57197.2023.10351321

[7] Lee, Seung Hoon, Seunghyun Lee, and Byung Cheol Song. "Vision transformer for small-size datasets." *arXiv preprint arXiv:2112.13492* (2021).

[8] Leem, Saebom, and Hyunseok Seo. "Attention guided CAM: Visual explanations of vision transformer guided by self-attention." In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 4, pp. 2956-2964. 2024. https://doi.org/10.1609/aaai.v38i4.28077

[9] Liu, Yun, Yu-Huan Wu, Guolei Sun, Le Zhang, Ajad Chhatkuli, and Luc Van Gool. "Vision transformers with hierarchical attention." *Machine Intelligence Research* 21, no. 4 (2024): 670-683. https://doi.org/10.1007/s11633-024-1393-8

[10] Loshchilov, Ilya, and Frank Hutter. "Decoupled weight decay regularization." *arXiv preprint arXiv:1711.05101* (2017).

[11] Webb, Patrick, Tim G. Benton, John Beddington, Derek Flynn, Niamh M. Kelly, and Sandy M. Thomas. "The urgency of food system transformation is now irrefutable." *Nature Food* 1, no. 10 (2020): 584-585. https://doi.org/10.1038/s43016-020-00161-0

[12] Maharana, Kiran, Surajit Mondal, and Bhushankumar Nemade. "A review: Data pre-processing and data augmentation techniques." *Global Transitions Proceedings* 3, no. 1 (2022): 91-99. https://doi.org/10.1016/j.gltp.2022.04.020

[13] Pamela, P., D. Mawejje, and M. Ugen. "Severity of angular leaf spot and rust diseases on common beans in Central Uganda." *Uganda Journal of Agricultural Sciences* 15, no. 1 (2014): 63-72.

[14] Savary, Serge, Laetitia Willocquet, Sarah Jane Pethybridge, Paul Esker, Neil McRoberts, and Andy Nelson. "The global burden of pathogens and pests on major food crops." *Nature ecology & evolution* 3, no. 3 (2019): 430-439. https://doi.org/10.1038/s41559-018-0793-y

[15] Schreinemachers, Pepijn, Emmy B. Simmons, and Marco CS Wopereis. "Tapping the economic and nutritional power of vegetables." *Global food security* 16 (2018): 36-45. https://doi.org/10.1016/j.gfs.2017.09.005

[16] Singh, Vimal, Anuradha Chug, and Amit Prakash Singh. "Classification of beans leaf diseases using fine tuned cnn model." *Procedia Computer Science* 218 (2023): 348-356. https://doi.org/10.1016/j.procs.2023.01.017

[17] Steiner, Andreas, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. "How to train your vit? data, augmentation, and regularization in vision transformers." *arXiv preprint arXiv:2106.10270* (2021).

[18] Ulian, Tiziana, Mauricio Diazgranados, Samuel Pironon, Stefano Padulosi, Udayangani Liu, Lee Davies, Melanie-Jayne R. Howes et al. "Unlocking plant resources to support food security and promote sustainable agriculture." *Plants, People, Planet* 2, no. 5 (2020): 421-445. https://doi.org/10.1002/ppp3.10145

[19] Steininger, Daniel, Andreas Trondl, Gerardus Croonen, Julia Simon, and Verena Widhalm. "The cropandweed dataset: A multi-modal learning approach for efficient crop and weed manipulation." In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3729-3738. 2023. https://doi.org/10.1109/WACV56688.2023.00372

[20] Wei, Zhenchao, Xu Ji, Li Zhou, Yagu Dang, and Yiyang Dai. "A novel deep learning model based on target transformer for fault diagnosis of chemical process." *Process safety and environmental protection* 167 (2022): 480-492. https://doi.org/10.1016/j.psep.2022.09.039

[21] Wenxia, Bao, Huang Xuefeng, Hu Gensheng, and Liang Dong. "Identification of maize leaf diseases using improved convolutional neural network." *Transactions of the Chinese Society of Agricultural Engineering* 37, no. 6 (2021).

[22] S. Pawar, B. L. R. Samaga, K. Pandith, V. Nayak, V. Geetha, and R. Thinakaran, "Precision Agriculture Using IOT Technology with a Case Study of Coffee Beans Production." In *Machine Vision and Augmented Intelligence*, K. Kumar Singh, S. Singh, S. Srivastava, and M. K. Bajpai, Eds., Singapore: Springer Nature, pp. 255–263. 2025 https://doi.org/10.1007/978-981-97-4359-9_26