Tomato Leaf Disease Classification using Transformer Variants

Waheed Moonwar[1], Mohammad Rakibul Hasan Mahin[1], Fahmid Bin Kibria[1], Aurnab Biswas[1], Fatiha Ishrar Chowdhury[1], and Annajiat Alim Rasel[1]

BRAC University, 66 Mohakhali, Dhaka 1212, Bangladesh
waheed.moonwar@g.bracu.ac.bd
mohammad.rakibul.hasan.mahin@g.bracu.ac.bd
fahmid.bin.kibria@g.bracu.ac.bd
aurnab.biswas@g.bracu.ac.bd
fatiha.ishrar.chowdhury@g.bracu.ac.bd
annajiat@bracu.ac.bd

Abstract. Agriculture has played a significant role for many years. Its growing significance is attributable to the money it has generated. The full advantages of crop cultivation is, however, prevented by a number of circumstances. Organic plant diseases have a role. For an agriculture dependent country like Bangladesh, extreme weather and heavy pesticide use are accountable for its economic crisis. This work aims to offer farmers visual information to facilitate the implementation of preventive measures beforehand. We proposed 5 different transformer models for tomato leaf disease classification, which includes Vision Transformer(ViT), Swin Transformer(SwT) and Compact Convolutional Transformer(CCT). In addition to that, two distinct variations of the ViT algorithm were incorporated into the categorization process. The models have been trained and validated on the PlantVillage dataset which resulted in test accuracy of 95.22%, 82.61%, 71.14%, 93.90% and 92.78% for ViT, SwT, CCT, ViT with Shifted Patch Tokenization and ViT with Locality Self Attention respectively.

Keywords: Image Classification · Vision Transformer · Plant Leaf · Tomato Disease · Compact Convolutional Neural Network · ConvMixer

1 Introduction

It is vital to have timely diagnosis of plant diseases as it plays an important role for the production of healthy crops. As plants are a crucial source of food for us humans, it is necessary that there should be a reduced negative impact of the harmful effects in plant leaf diseases. However, as the size of the symptoms are comparatively small and the local farmers do not have the knowledge and expertise to identify these diseases, it becomes a difficult task to accurately diagnose diseases in most cases. Artificial Intelligence (AI) has been used to aid us with the notable convolutional neural network (CNN), disease recognition

and the latest Vision Transformer (ViT) structure[1]. ViT variations have attention mechanisms that work with patches, and leaf diseases have spots that are location-specific, an attribute that is essential to identifying the disease. For this reason, ViT variations are a strong possibility for identifying plant leaf diseases. Moreover, ViT detects the object based on how a human perceives an image. As a general rule, a person will concentrate on a certain part of the area of attraction; hence, the ViT structure classifies pictures in the same manner.

While pre-designed models such as ViT and ResNet are widely used in image processing tasks, they consist of a large number of trainable parameters that require a large dataset in order to find a suitable output. Usually, models like these are trained on ImageNet and then are refined to implement transfer learning. Pre-designed architectures like these may struggle to obtain a fair result when trained on tiny datasets only if the domain of the second dataset differs from the first dataset. Moreover, pre-designed architectures are computationally sensitive as they are big models which can result in the computer underperforming and as a result, can cause issues for real-time applications. Furthermore, it is necessary to decide if transfer learning is always advantageous or whether lighter architectures are able to provide a decent output.

- Five cutting-edge Vision Transformer models, CCT, ViT, ViT with Shifted Patch Tokenization, ViT with Local Self Attention, and Swin Transformers, have been implemented. Comparisons are made between their different parameters and results are produced.

- The models were trained and tested on the dataset, and the produced results were compared with each other to find out the most accurate Vision Transformer that will give the best results for Tomato Leaf Disease Identification.

2 Related Works

The images of the plant have been surveyed to find out about the diseases of the specific plants. The images that make up the dataset are photographed under suitable conditions. If we consider analyzing diseases in wheat rusk in [1], both aerial and non aerial pictures were recorded and labeled for object detection. Using object detection networks to localize the wheat leaf by recognizing the corresponding bounding box, was the idea put forward by the researcher. Then, in order to determine the actual class, the cropped box was treated as classification network's input. For this research, VGG16 cite{RW2, ResNet-50 [3], Inception [4], MobileNetV3 [5] and EfcientNet-B0 [6] are 5 Computer vision models developed for categorization. From comparitive analysis, the most accurate classification model was the EfficientNet-B0 whilst having less computational complexity and cost.

In [7], various models are utilized which include the C-DenseNet structure. The dataset in this literature was thought to be balanced according to the allocation of pictures in each class. All the images in the dataset were cropped for

preprocessing because of which the images did not have a complex background. In addition to this, six class divisions are made for the magnitude of the disease intensity. The C-DenseNet was implemented to address the minute differences in the features of images in some levels. Dense layers were integrated into the convolutional layers, and the convolutional layers, comprising the channel-specific and location-specific attention methods, being applied. From feature map in the channel attention module, features are retrieved using Fully connected layers. Later on, with the help of sigmoid they found out which section of the input channel needed better attention. The output of the CBAM is achieved by multiplying the output of both the models (channel and spatial attention) with the input feature map. The main feature finding of the images is more precise, around 97.99% because of implementing the attention mechanism. M-bCNN structure[8], which is a matrix based convolutional neural network, was also used especially for small wheat leaf disease features. This structure could highlight the key characteristics of the images and overpowered the need of AlexNet[9] and VGG-16[2] networks.

From analyzing[10], we found images of contaminated rice leaves in India. In order to diagnose the disease in rice leaves, primarily they performed some preprocessing of the images which include removing the background of complicated pictures and then picture separation strategies such as K-means clustering and Ostu's method were employed for finding problematic areas of the leaf. The label of the pictures in [10] were found from the segmented images -which was data fed into a Support Vector Machine classifier.

The plant village dataset [11] consists of colored images of numerous leaves and their associated diseases. However, the grayscale and segmented set can also be found. In [12], each version was subjected to individual testing with various multiple training and distinct configurations. The base models were AlexNet[9] and GoogLeNet[13]. The dataset versions were used for training various train-test split ratios, which are 1:4, 2:3, 3:2 and 4:1. The training was held twice, firstly with arbitrary weight parameters and then with transfer learning using pre-trained AlexaNet and GoogLeNet on ImageNet. The models with transfer learning demonstrated greater productivity and functionality compared to those trained from zero, as thought of. The researchers in [14] used the tomato leaf images from the Plant village dataset and the tomato leaves were diagnosed with the implementation of a CNN. For their network, LVQ was used as the classifier and resulted in around 86% accuracy rate.

In [15] AlexNet and VGG16 architectures were implemented where segmented version of the tomato leaf images (from the Plant village dataset) were used. The background pixels have a value of zero and this paper resulted in a precision of 97.49%. Despite the fact that segmented images make it simpler for a neural network to classify objects, they are rarely available in real-world situations. In such situations, either the pictures must be categorized manually expensing time or a neural network must be used. A neural network that is capable of segmentation is also capable of classification. Additionally, a classification-designed convolutional neural network can independently process images with non-zero

backgrounds. Halil et al. attempted to attain a framework which can be implemented in immediate application using that class from the same dataset (plant village). They attained 95.6% and 94.3% accuracy by training with AlexNet and SqueezeNet[16]. For AlexNet and SqueezeNe[t17] 150ms and 50ms, respectively, were noted as the prediction time.

The papers [18] and [19] highlight the use of original variants of ViT (ViT-B16 having 16 and ViT-B32 having 32 attention heads) except modifying the classification in the agriculture sector. In [20], two ViT models are used parallelly for managing pictures having different resolutions. The paper [19] and [20] addressed a specific plant disease while [18] is centered on the classification of the plant leaf. With increasing quantities of plants and their diseases, the classification process turns out to be more complex. In addition to that, the speed of forecast, a deciding factor for instantaneous categorization, was not taken into consideration. Implementing a complex framework for unhealthy leaf categorization may be rigorous, and rather simpler structures may suffice in some cases. This paper presents low-parameter frameworks for immediate unhealthy crop categorization using Vision Transfomer. These frameworks will be run on 3 differnt pictorial data, and compared with CNN-based frameworks with similar complexity in terms of accuracy and prediction speed. Additionally, combinations of CNN and ViT will be looked into, to improve accuracy, and the impact of image size on the outcomes will be examined.

3 Dataset

For this research project we used the publicly available dataset known as Plant Village[21] dataset which contains images of 38 different classes combining both healthy and diseased images. However, we used only the classes that belong to the tomato leaf. In total there are 9 diseased classes and 1 class containing healthy images. Combining diseased and healthy images the dataset contains around 18835 images. In Figure 1 we can see one image from each class. Moreover in Table 1 we can see a detailed image distribution of our dataset.

For training our model with the dataset we resized our image to 100 by 100. We also augmented the dataset, for augmentation we flipped random images horizontally, randomly rotated images by 2 degrees and finally, randomly zoomed images by 20% in both width and height during our training. This augmentation was done to reduce the overfitting problem in our model.
Moreover, we splitted our dataset into two parts into training and testing sets. Our training set contains 16951 images and the testing set contains 1884 images. Furthermore, we splitted our training dataset by 10% to create a validation set which contains 1694 images.

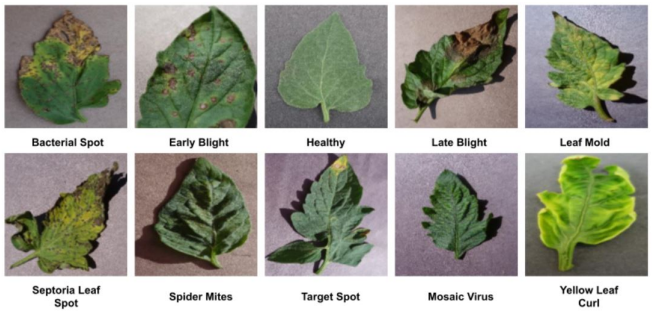

Fig. 1. Sample Images of dataset

Table 1. Image Distribution of Plant Village Dataset

Name	Number of Images
Bacteria Spot	2127
Early Blight	1000
Healthy	1591
Last Blight	1909
Leaf Mold	1000
Septoria Leaf Spot	1771
Spider Mites	1676
Target Spot	1404
Mosaic Virus	1000
Yellow Leaf Curl	5357
Total	18835

4 Methodology

4.1 Vision Transformer (ViT)

To classify images, the Vision Transformer (ViT) paradigm applies an architecture like a Transformer to different parts of the image. The typical Trans-

Fig. 2. Normal Image

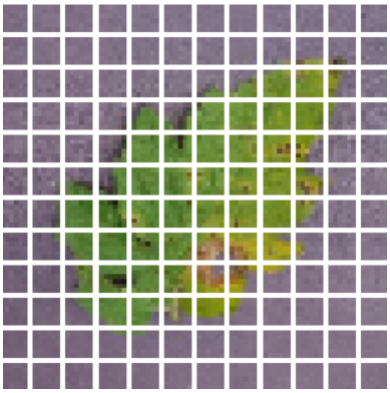

Fig. 3. Patches of Image

former encoder is given a sequence of vectors generated by slicing an image into patches of a certain size, linearly embedding each patch, and finally adding position embeddings as illustrated in Fig.2 and Fig.3. Introducing an innovative "classification token" that can be learned is a tried and effective way for doing categorization.

Typically, a 1D series of token embedded data is sent into the conventional Transformer as input. The input image $\alpha \epsilon R^{(H\times W\times C)}$ is altered into a collection of two-dimensional patches $\alpha\_p \epsilon R^{(N\times(P^{2\dot{C}})}$. Here, the picture resolution is denoted by (H,W), C refer to the channel number, and patches are of size P. The number of generated patches is calculated by the equation $\beta = \frac{H\times W}{P^2}$. We utilize a trainable linear projection to flatten the patches and map to D dimensions, as the Transformer maintains a latent vector size of D throughout all of its layers. The results of this projection are known as patch embeddings. The state of the embedded patches at the output of the Transformer encoder is the picture representation γ; similarly to BERT's [class] token, and we append a

learnable embedding to this sequence. An initial classification head is connected to the state, and a final classification head is added during tuning. A Mtultilayer Perceptron (MLP) with one hidden layer during the pre-training phase and one linear layer during the fine-tuning phase is used as the classification head. To keep track of where things are in a scene, we augment the patch embeddings with position embeddings. Since we have not seen any appreciable performance advantages from utilizing more sophisticated 2D-aware position embeddings, we stick with the tried-and-true approach of employing ordinary learnable 1D position embeddings. The encoder takes the generated vector sequence as input.

4.2 ViT with Shifted Patch Tokenization

We have implemented one variant of the Vision Transformer, which is ViT with Shifted Patch Tokenizer. The architecture of the model is illustrated in Fig.4. First, the input image is spatially transformed, and the combination of transformed images are either left-up, right-up, left-down and right-down. Then, concatenation of the transformed images are performed to produce a concatenated feature, which is then forwarded to the patch partition layer to extract the patch features. Afterwards, we generate a 1D flattened patch, which is then propagated to the normalization and projection layers to generate the tokens for visual features.

4.3 ViT with Locality Self Attention

Another variation of the Vision Transformer implemented for our study is the amalgamation of Locality Self Attention in ViT, which is based on an attention mechanism similar to that of BERT. The model representatation is described in Fig.6 Initially, a dot product of the query and key is computed which emphasizes the attention score of the token. Then, it is propagated to a softmax activation layer, before being scaled and masked, to introduce non-linearity. Finally, the attention score is generated using dot product of the input value and the scaled score.

4.4 Compact Convolutional Transformer (CCT)

Compact Convolutional Transform (CCT), which incorporates convolutions into transformer models, is the idea put out by Hassani et al[22]. The CCT technique improves inductive bias and eliminates the need for positional embeddings by using sequence pooling and patch embedding in place of convolutional embedding. One of the numerous benefits of the CCT is that input parameter flexibility is proliferated along with precision. Fig.5 shows the architectural design of CCT.

4.5 Swin Transformer (SwT)

The concept of Swin Transformer was originally presented in [23], which functions like a universal framework in the field of computer vision. The term Swin

Transformer (the word Swin stands for **S**hifted **win**dow). The representation of this layered Transformer is built from moving windows.For increased output, the shifted window method uses non-overlapping local windows for self-attention computation. Concurrently, it makes it possible to link different windows with one another.

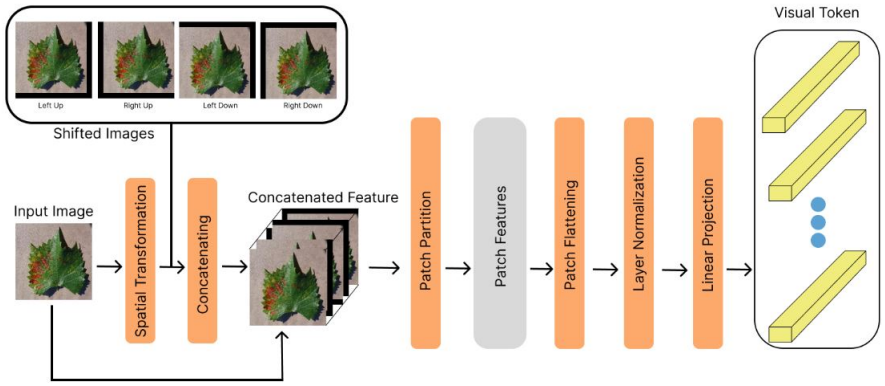

Fig. 4. Vision Transformer (ViT) with Shifted Patch Tokenization

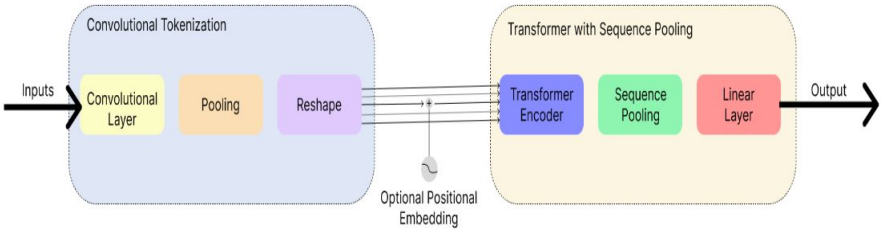

Fig. 5. Compact Convolutional Transformer(CCT)

5 Experiments and Results

To keep our experiments fair for comparison between models we used same environment for all the models. All models were trained with Intel i5 7500, 24 GB of RAM and RTX 3060 GPU.

We trained all the models for 30 epochs. We kept learning rate of 0.001 and weight decay of 0.001. Moreover, we used batch size of 32. We, trained

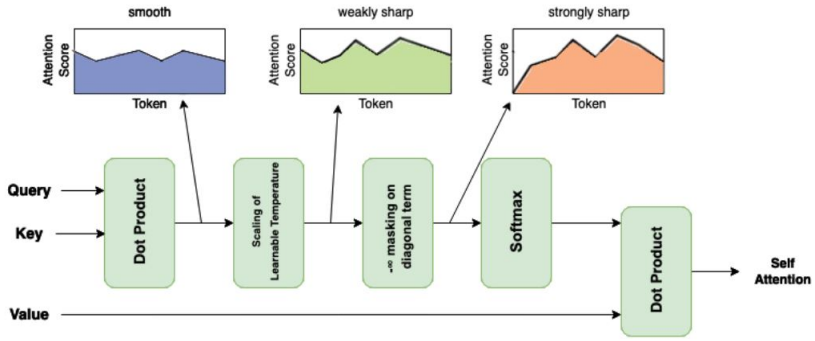

Fig. 6. Vision Transformer(ViT) with Locality Self Attention

our model with image size of 100 by 100. Now for patch size we kept image size of 72 and patch size of 6, then to determine number of patches we used the formula $(\frac{image\_size}{patch\_size})^2$. For loss function we used Sparse Categorical Crossentropy. Last but not the least we used Model Checkpoint to save the best weights found during our training process, so when we tested we load our best weights and evaluated our model with the test set.

Using the Plant Village dataset, the evaluation of the performance of five different models were made. The dataset consists of 18,835 plant leaf images of 10 classes of tomato leaves affected by various types of diseases. The five models we have implemented here are Vision Transformers (ViT), Swin Transformers (SwT), CCT, ViT with Shifted Patch Tokenization and ViT with Locality Self Attention. The results shown in Fig.7 shows the training and testing accuracy of each model. The Vision Transformer (ViT) shows the highest training and testing accuracy of 95.51% and 95.22% respectively(Fig.8). Following this, ViT with Shifted Patch Tokenization (Fig.11) had a training accuracy of 95.44% and a testing accuracy of 93.90%. Then, ViT with Local Self Attention had a training accuracy of 94.50% and a testing accuracy of 92.78%. Afterwards, comes the Swin Transformer (SwT) with a training and testing accuracy of 81.19% and 82.61% accordingly(Fig.9). Finally comes CCT, the model with the lowest accuracy of 73.04% in training and 71.14% in testing (Fig.10).

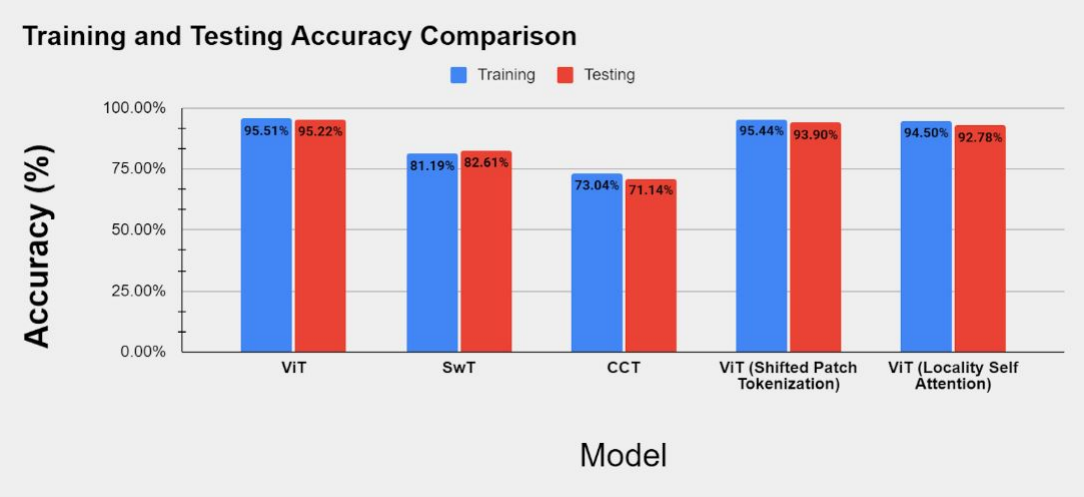

Fig. 7. Accuracy Comparison of 5 Models

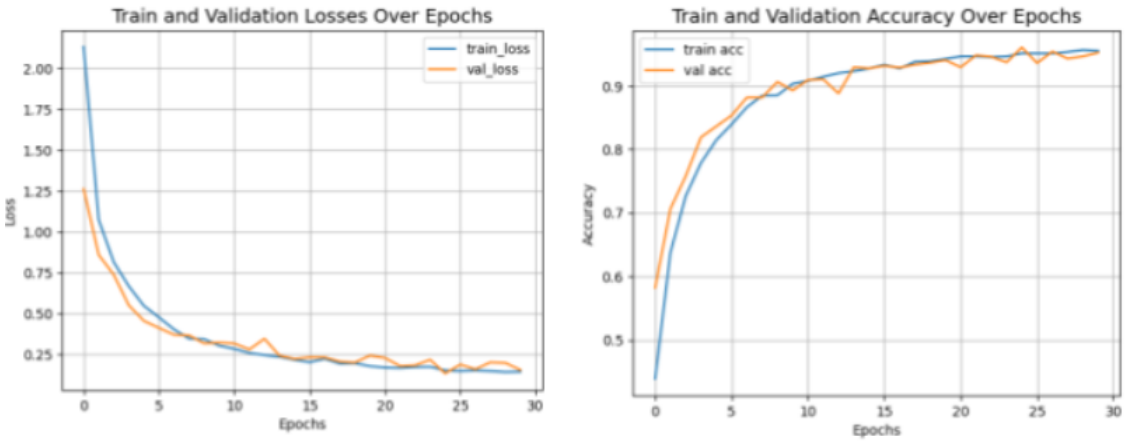

Fig. 8. Accuracy and Loss curves of ViT

6 Future Works and Improvements

To improve our accuracy and precision, we can use ensemble learning. We want to extend our dataset also in future which can help to further validate the efficacy of the models. Moreover, we can do web implementation, can add more ViT model implementations such as the recently introduced Mobile ViT architecture. Using more ViT Models could lead to more practical and accessible solutions for farm-

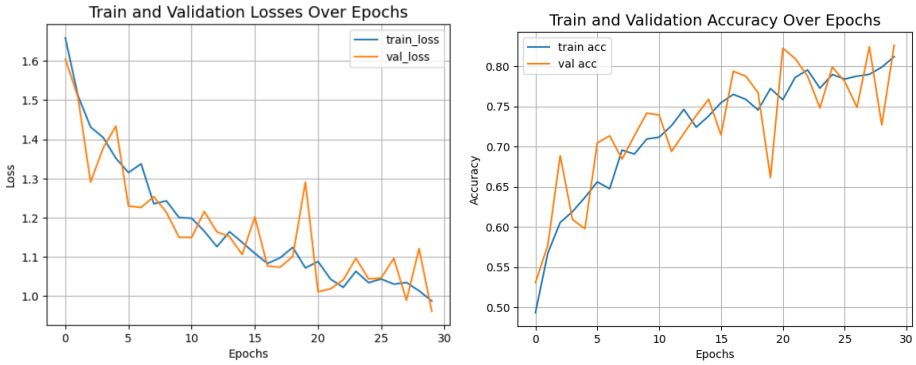

Fig. 9. Accuracy and Loss curves of SwT

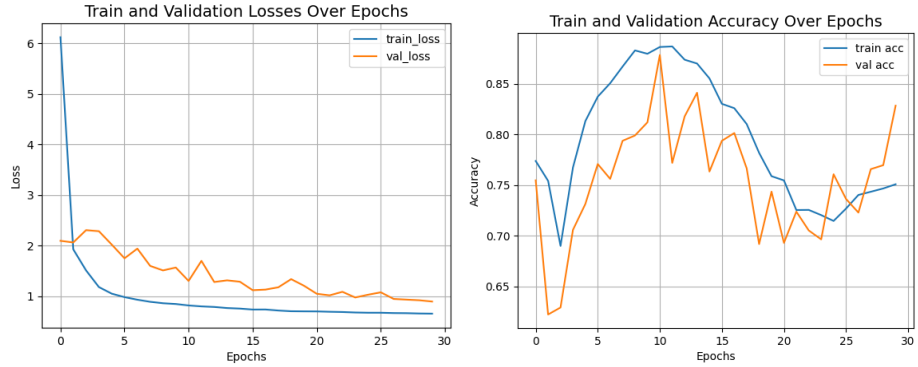

Fig. 10. Accuracy and Loss curves of CCT

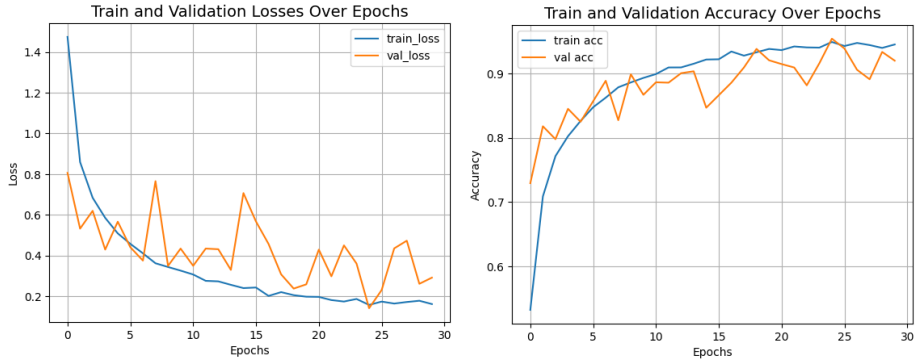

Fig. 11. Accuracy and Loss curves of ViT with Shifted Patch Tokenization

ers and plant researchers. Web Implementation would allow more accessibility of the models. The farmers and researchers can be enabled to get access easily and utilize these models for real life applications.

7 Conclusion

Plants are one of the most important elements of the environment. Saving plants by identifying diseases is one of the most significant responsibilities for everyone right now. In this work, the Vision Transformer (ViT) has been used. For which, the result of this work can be gained quickly and the cost will be reduced . It is also important for farmers to understand the disease of the plant. This research can help to play a huge economic benefit while farmers can recognize the disease through it. By getting visual information through this project, the farmers can have a clear concept about preventions.

References

1. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, http://arxiv.org/abs/2010.11929, last accessed 2023/05/04
2. Automated Wheat Disease Detection using a ROS-based Autonomous Guided UAV, http://arxiv.org/abs/2206.15042, last accessed 2023/05/05
3. Very Deep Convolutional Networks for Large-Scale Image Recognition, http://arxiv.org/abs/1409.1556, last accessed 2023/05/05
4. Deep Residual Learning for Image Recognition,https://doi.ieeecomputersociety.org/10.1109/CVPR.2016.90, last accessed 2023/05/05
5. Rethinking the Inception Architecture for Computer Vision,http://arxiv.org/abs/1512.00567, last accessed 2023/05/05.
6. Searching for MobileNetV3, http://arxiv.org/abs/1905.02244, last accessed 2023/05/05.
7. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks, http://arxiv.org/abs/1905.11946, last accessed 2023/05/05.
8. Wheat Stripe Rust Grading by Deep Learning With Attention Mechanism and Images From Mobile Devices, https://www.frontiersin.org/articles/10.3389/fpls.2020.558126, last accessed 2023/05/05.
9. A Unified Matrix-Based Convolutional Neural Network for Fine-Grained Image Classification of Wheat Leaf Diseases, https://ieeexplore.ieee.org/document/8606914, last accessed 2023/05/05.
10. ImageNet Classification with Deep Convolutional Neural Networks, https://content.iospress.com/articles/intelligent-decision-technologies/idt301 ,last accessed 2023/05/05.
11. Detection and classification of rice plant diseases, Intelligent Decision Technologies, vol. 11, no. 3, pp. 357-373, 2017
12. An open access repository of images on plant health to enable the development of mobile disease diagnostics, http://arxiv.org/abs/1511.08060, last accessed 2023/05/05.
13. Using Deep Learning for Image-Based Plant Disease Detection, http://arxiv.org/abs/1604.03169, last accessed 2023/05/05.
14. Going Deeper with Convolutions, http://arxiv.org/abs/1409.4842, last accessed 2023/05/05.
15. Plant Leaf Disease Detection and Classification Based on CNN with LVQ Algorithm, https://ieeexplore.ieee.org/document/8566635, last accessed 2023/05/05.
16. Tomato crop disease classification using pre-trained deep learning algorithm, https://www.sciencedirect.com/science/article/pii/S1877050918310159, last accessed 2023/05/05.
17. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and ¡0.5MB model size, http://arxiv.org/abs/1602.07360, last accessed 2023/05/05.
18. Transformer Neural Network for Weed and Crop Classification of High Resolution UAV Images, https://www.mdpi.com/2072-4292/14/3/592, last accessed 2023/05/05.
19. Multi-granularity Feature Extraction Based on Vision Transformer for Tomato Leaf Disease Recognition, https://ieeexplore.ieee.org/document/9695688.
20. Artificial Cognition for Early Leaf Disease Detection using Vision Transformers, https://ieeexplore.ieee.org/document/9598303.

21. PlantVillage Dataset, https://www.kaggle.com/datasets/emmarex/plantdisease.
22. Escaping the Big Data Paradigm with Compact Transformers, https://arxiv.org/abs/2104.05704.
23. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows, https://arxiv.org/abs/2103.14030.