Title
An Application of Vision Transformer(ViT) for Image-Based Plant Disease Classification

Author
Kalaydjian, Carina Teolinda

Publication Date
2023

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

An Application of Vision Transformer(ViT) for Image-Based Plant Disease Classification

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Applied Statistics

by

Carina Teolinda Kalaydjian

2023

ABSTRACT OF THE THESIS

An Application of Vision Transformer(ViT)
for Image-Based Plant Disease Classification

by

Carina Teolinda Kalaydjian
Master of Applied Statistics
University of California, Los Angeles, 2023
Professor Ying Nian Wu, Chair

Crop loss due to plant diseases and pests poses a significant challenge for crop growers worldwide, affecting product quality, nutritional value, and overall crop yield [Dir22]. Accurate disease identification is crucial for implementing effective treatments and reducing crop losses [Sol21]. This paper explores the utilization of the Vision Transformer(ViT), for image-based plant disease classification. The study builds upon the work of Mohanty et al. who used Convolutional Neural Networks to classify diseases in the PlantVillage data set [MHS16]. However, instead of Convolutional Neural Networks, this research employs a ViT model pre-trained on the ImageNet-21k data set to leverage transfer learning to train a crop disease image classification model. By developing and leveraging such tools, the agricultural community can work towards minimizing crop loss and ensuring food security for the growing global population.

The thesis of Carina Teolinda Kalaydjian is approved.

Chad Hazlett

Michael Tsiang

Ying Nian Wu, Committee Chair

University of California, Los Angeles

2023

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

For the love of plants. Keep growing!!!

CHAPTER 1

Introduction

As the global population continues to expand, "it is estimated that food production will need to increase by 60% by 2050 to feed the estimated 10 billion people expected on Earth. An increase in production along with a reduction in food loss due to pests and pathogens and food waste will be needed to meet demand" [RAB21]. Crop loss resulting from plant diseases and pests poses a formidable challenge for crop growers worldwide. Plant diseases and pests lower the product quality or shelf-life of crops, decrease the nutritional value of vegetables and fruits, and reduce crop yield [Dir22]. Plant Diseases caused by fungal pathogens can cause crop losses of 10% to 20% each year [Ser22]. The Food and Agriculture Organization of the United Nations estimates that annually 20% to 40% of global crop production are lost to pests. Each year, plant diseases cost the global economy around $220 billion USD, and invasive insects around $70 billion USD [FU19].

A challenge crop growers face is accurately identifying the disease responsible for their crop losses. The identification process is particularly challenging as some plant diseases exhibit similar symptoms, particularly during the early stages of infection. Consequently, discerning the nuanced distinctions becomes a daunting task for the human eye. Often, crop growers can recognize the disease after it has significantly affected their crops or when the infection or infestation has persisted over a prolonged period of time, leading to observable alterations in leaf appearance or crop loss [Sol21]. It is crucial to emphasize the significance of proper disease identification, as employing the wrong treatment can be a waste of time, financial resources, and possibly cause further crop loss or damage [Sol21].

In order to facilitate the identification of plant diseases, Mohanty et al. proposed a novel approach in their scholarly work titled "Using Deep Learning for Image-Based Plant Disease Classification". The researchers explored the utilization of deep learning convolutional neural network models to effectively discern various types of plant diseases. The data set in their study was obtained from the PlantVillage project, encompassing a vast collection of 54,306 color images depicting 14 distinct crop species afflicted with 20 different disease types or healthy conditions [MHS16].

The authors conducted an extensive investigation, comparing the effectiveness of using color(RGB) images versus gray-scale and segmented images, exploring various training-validation-testing splits, comparing the outcomes of training models from scratch versus utilizing pre-trained models, and evaluating the performance of GoogLeNet and AlexNet, two different deep learning convolutional neural network architectures. Through the systematic exploration of these factors, they conducted a total of 60 experiments to ascertain the optimal combination of architectural configurations [MHS16].

While the study from authors Mohanty et al. primarily focused on deep convolutional neural networks, subsequent research has demonstrated that convolutional neural networks are not the sole approach to achieving excellent performance in image classification tasks [DBK21]. These claims come from the paper "An Image is worth 16X16 Words: Transformers for Image Recognition at Scale" by Alexey Dosovitskiy et al, [DBK21]. Their paper finds that employing a pure transformer applied directly to sequences of images, when pre-trained on substantial volumes of data and transferred to multiple mid-sized or small image recognition benchmarks such as ImageNet, CIFAR-100, or VTAB, can yield highly competitive outcomes [DBK21]. The Vision Transformer(ViT) architecture has showcased remarkable performance compared to state-of-the-art convolutional networks, while also significantly reducing the computational resources required for training [DBK21]. Consequently, the motivation for this project is to loosely follow the framework outlined in the study by Mohanty et al, [MHS16]. However, instead of employing a deep convolutional network architecture,

a Vision Transformer(ViT) model pre-trained on the ImageNet-21k data set will be utilized to implement transfer learning and to train a disease classification model with the project PlantVillage data set.

CHAPTER 2

Methods

2.1 Data Introduction

The data for this project is the project PlantVillage data which was found through the paper, "Using Deep Learning for Image-Based Plant Disease Detection" [MHS16]. The data consists of 54,306 color images of healthy and diseased crop leaves. In a machine learning sense, our data set of 54,306 images is considered small. Each image is the size of 256×256 pixels, has the three color channels RGB(red, green, blue), and is categorized under a Crop-Disease Classification Label. Each label has a crop species name and a plant disease name or healthy. There are 14 different crop species and 20 different crop diseases, which create 38 different crop-disease Classification Labels in this data set. See Table 2.1 for a detailed list of all Classification Labels. Table 2.1 shows the total number of images each Classification Label contains and how that amount is translated to be the overall percentage contribution to the data set.

Classification labels with over 5,000 images had the largest percent contributions to the data set. These labels are Orange-Haunglongbing with 10.1%, Tomato-Yellow Leaf Curl with 9.9%, and Soybean-Healthy with 9.4%. The classification labels with less than 500 images and having the least amount of percent contributions to the data set are Peach-Healthy, Raspberry-Healthy, and Tomato-Mosaic Virus, all with 0.7%, Apple-Cedar Apple Rust with 0.5%, and Potato-Healthy with 0.3%.

Table 2.1: Classification Labels Information Table

Classification Label		**Number of Images**	**Percent Contribution to Data Set**
1	Apple-Apple Scab	630	1.2%
2	Apple-Black Rot	621	1.1%
3	Apple-Cedar Apple Rust	275	0.5%
4	Apple-Healthy	1,645	3.0%
5	Bell Pepper-Bacterial Spot	997	1.8%
6	Bell Pepper-Healthy	1,478	2.7%
7	Blueberry-Healthy	1,502	2.8%
8	Cherry-Powdery Mildew	1,052	1.9%
9	Cherry-Healthy	854	1.6%
10	Grape-Black Rot	1,180	2.2%
11	Grape-Esca (Black Measles)	1,383	2.5%
12	Grape-Leaf Blight (Isariopsis Leaf Spot)	1,076	2.0%
13	Grape-Healthy	423	0.8%
14	Maize(Corn)-Cercospora Leaf Spot (Gray Leaf Spot)	513	0.9%
15	Maize(Corn)-Common Rust	1,192	2.2%
16	Maize(Corn)-Northern Leaf Blight	985	1.8%
17	Maize(Corn)-Healthy	1,162	2.1%
18	Orange-Haunglongbing	5,507	10.1%
19	Peach-Bacterial Spot	2,297	4.2%
20	Peach-Healthy	360	0.7%
21	Potato-Early Blight	1,000	1.8%

22	Potato-Late Blight	1,000	1.8%
23	Potato-Healthy	152	0.3%
24	Raspberry-Healthy	371	0.7%
25	Soybean-Healthy	5,090	9.4%
26	Squash-Powdery Mildew	1,835	3.4%
27	Strawberry-Leaf Scorch	1,109	2.0%
28	Strawberry-Healthy	456	0.8%
29	Tomato-Bacterial Spot	2,127	3.9%
30	Tomato-Early Blight	1,000	1.8%
31	Tomato-Late Blight	1,909	3.9%
32	Tomato-Leaf Mold	952	1.8%
33	Tomato-Mosaic Virus	373	0.7%
34	Tomato-Septoria Leaf Spot	1,771	3.3%
35	Tomato-Spider Mite(Two-Spotted)	1,676	3.1%
36	Tomato-Target Spot	1,404	2.6%
37	Tomato-Yellow Leaf Curl	5,357	9.9%
38	Tomato-Healthy	1,591	2.9%

Total Number of Images	**54,306**	**100.0%**

2.1.1 Crop Species

The 14 crop species in this data set are Apple, Bell Pepper, Blueberry, Cherry, Grape, Maize(Corn), Orange, Peach, Potato, Raspberry, Soybean, Squash, Strawberry, and Tomato. Each crop has healthy or diseased images. Some crops can also have different types of diseases as well. See Table 2.2 for more information on the summed amount of healthy and diseased images by crop species. Table 2.2 shows the summed amount images by crop species and

the percentage each crop represents in the data set overall.

The top three contributing crops to the PlantVillage data set are Tomato, Orange, and Soybean. There are 18,160 images of healthy and diseased Tomatoes, which contributes 33% to the data overall. There are 5,507 images of only diseased Oranges, which contributed 10.1% to the data set overall. There are also 5,090 images of only healthy Soybeans, which contributed 9.4% to the data set overall.

Table 2.2: Crop Species Information Table

	Crop Species	**Number of Images**	**Percentage Composition in Data Set**
1	Apple	3,171	5.8%
2	Bell Pepper	2,475	4.6%
3	Blueberry	1,502	2.8%
4	Cherry	1,906	3.5%
5	Grape	4,062	7.1%
6	Maize(Corn)	3,852	7.5%
7	Orange	5,507	10.1%
8	Peach	2,657	4.9%
9	Potato	2,152	4.0%
10	Raspberry	371	0.7%
11	Soybean	5,090	9.4%
12	Squash	1,835	3.4%
13	Strawberry	1,565	2.9%
14	Tomato	18,160	33.4%

2.1.2 Crop Diseases

Some crop species in this data set have healthy and diseased leaf images. However, some crops only have diseased images or only healthy images. Not all crop species in this data have contributed to the healthy crop images, and not all crop species have contributed to the diseased images. For a more detailed view of the 20 crop diseases included in this data set, see Table 2.3, which has the summed amount images for each disease in this data set and the percentage contribution to the overall data set.

Table 2.3: Crop Diseases Information Table

Crop Diseases		**Number of Images**	**Percent Contribution to Data Set**
1	Apple Scab	630	1.2%
2	Bacterial Spot	5,421	10.0%
3	Black Measles	1,383	2.5%
4	Black Rot	1,801	3.3%
5	Cedar Apple Rust	275	0.5%
6	Cercospora Leaf Spot(Gray Leaf Spot)	513	0.9%
7	Common Rust	1,192	2.2%
8	Early Leaf Blight	2,000	3.7%
9	Haunglongbing	5,507	10.1%
10	Isariopsis Leaf Spot	1,076	2.0%
11	Late Leaf Blight	2,909	5.4%
12	Leaf Mold	952	1.8%
13	Leaf Scorch	1,109	2.0%
14	Mosaic Virus	373	0.7%

15	Northern Leaf Blight	985	1.8%
16	Powdery Mildew	2,887	5.3%
17	Septoria Leaf Spot	1,771	3.3%
18	Spider Mite(Two-Spotted)	1,676	3.1%
19	Target Spot	1,404	2.6%
20	Yellow Leaf Curl	5,357	9.9%

Total Diseased Crop Images	**39,221**	**72.2%**

The top three diseases that compose this data set are Bacterial Spot, Haunglongbing, and Yellow Leaf Curl. Bacterial Spot, which composes 10% of the data, is a bacterial disease that affects many crops by causing their leaves to develop yellow spots that turn brown in the middle. It also causes crops to develop black or brown spots of rot on their fruits [SJG21]. Haunglongbing composes 10% of the data set. This bacterial disease affects citrus trees causing their fruits to stay green and fall to the ground early before becoming ripe [AU22]. This disease is common for citrus, but keep in mind that this data set only has images of this disease affecting Oranges. Yellow Leaf Curl composes 9.9% of the data set and is a Viral infection that only affects Tomatoes. "Yellow leaf curl virus is undoubtedly one of the most damaging pathogens of tomatoes, and it limits the production of tomatoes in many tropical and subtropical areas of the world. It is also a problem in many countries that have a Mediterranean climate, such as California. Thus, the spread of the virus throughout California must be considered a serious potential threat to the tomato industry" [CR13].

Note that the total percentage of diseased images contributing to this data set is 72.2% because the other 27.8% of this data set is healthy crop images. See Table 2.4 for a comparison of the amount of diseased and healthy crop images in this data set. The crops that contributed to the diseased images are Apple, Bell Pepper, Cherry, Grape, Maize(Corn), Orange, Peach, Potato, Strawberry, Squash, and Tomato. The crops that contributed to

the healthy images are Apple, Bell Pepper, Blueberry, Cherry, Grape, Maize(Corn), Peach, Potato, Raspberry, Strawberry, and Tomato.

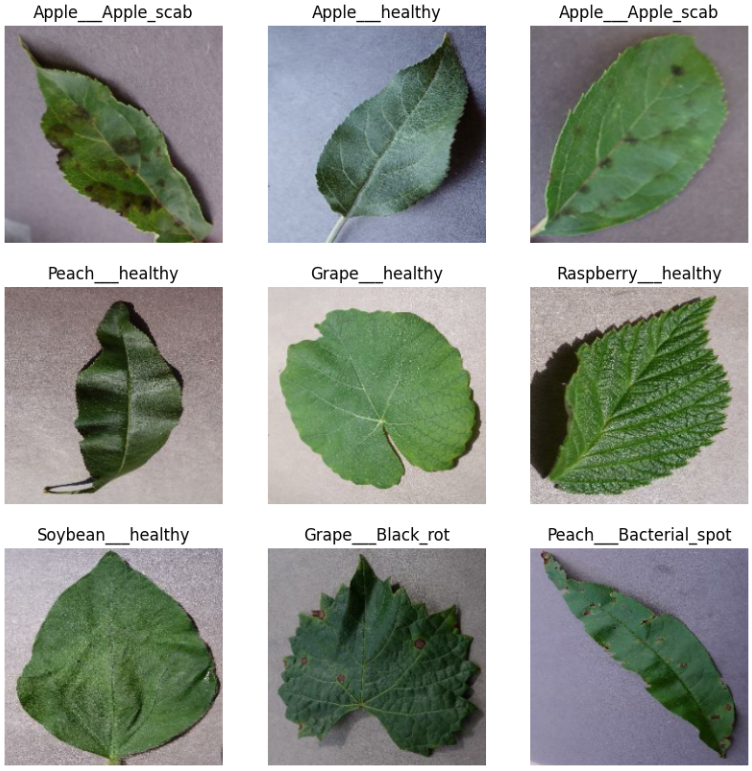
Table 2.4: Diseased versus Healthy Images Comparison

Total Diseased Crop Images	**39,221**	**72.2%**
Total Healthy Crop Images	**15,085**	**27.8%**

2.1.3 Visual of the Data

For a visual representation of what the diseased and healthy crop images look like, see Figure 2.1 for nine different crop images. The crop images have their classification labels above them to identify the crop name and disease or healthy. The images in Figure 2.1 are crop images of classification labels Apple-Apple Scab, Apple-Healthy, Peach-Healthy, Grape-Healthy, Raspberry-Healthy, Soybean-Healthy, Grape-Black Rot, and Peach-Bacterial Spot.

Figure 2.1: Visual Representation of Images and Classification Labels

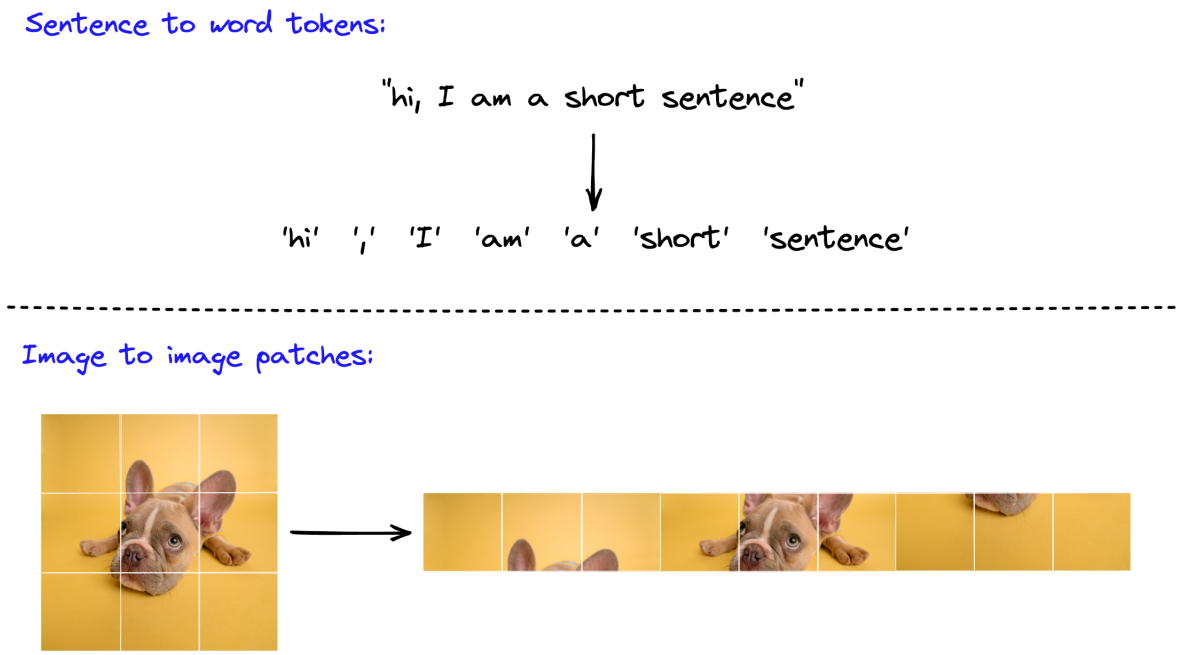


2.2 Vision Transformer: ViT Method

While the Transformer architecture has become the de-facto standard for natural language processing tasks, its applications to computer vision remain limited [DBK21]. This was the motivation for Dosovitskiy et al. to look into the implementations of the transformer model for image classification tasks, and the vision transformer was created. The transformer architecture for natural language processing tasks works similarly to a vision transformer [Bri22]. In natural language processing, sentences are broken down into words. Then each word is treated as a sub-token of the original sentence [Bri22]. Similarly, the vision transformer breaks down an image into smaller patches, each patch representing a small sub-section of the original image. To visually see how sentences are broken down into word tokens and images broken down into patches, see Figure 2.2 from [Bri22]. Keep in mind that the position

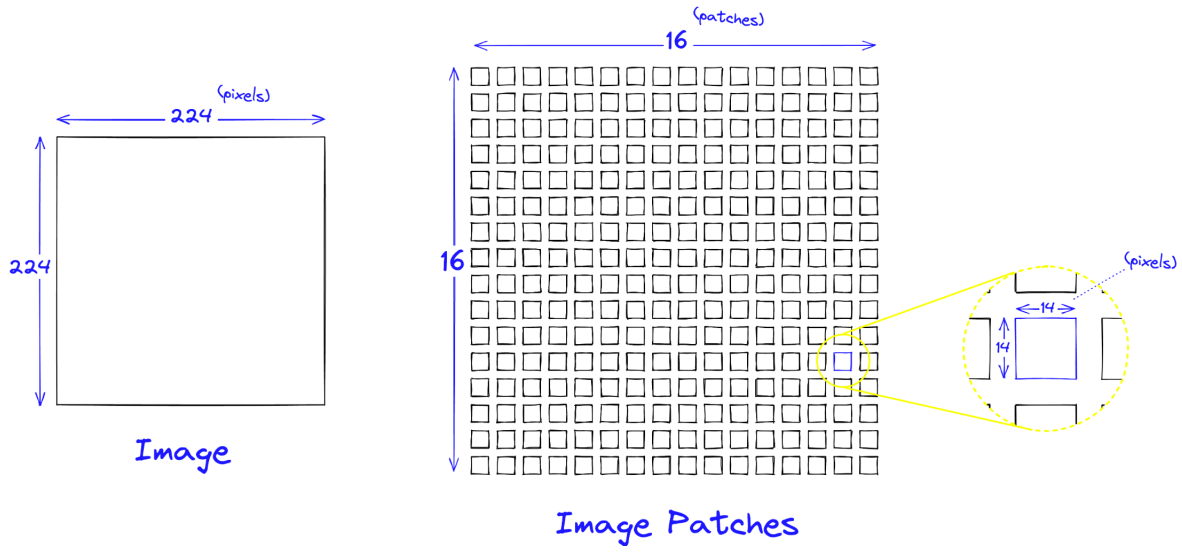
of the image patch is very important. If the image patches are out of order, then the original image will also be out of order.

Figure 2.2: How Sentences Become Word Tokens and Images Become Patches



The vision transformer model that Dosovitskiy et al. developed specifically works with only input images of size 224×224 [WXD20]. They decided on this size because 224×224 can be easily be divided into 256 (16×16) image patches. When there are 256 image patches then each image patch is 14×14 pixels. To visually see how this process is done, see Figure 2.3 from [Bri22]. It shows the transformation the image goes through to become the 256 image patches that are pixel size 14×14.

Figure 2.3: How an Image Becomes 256 Image Patches



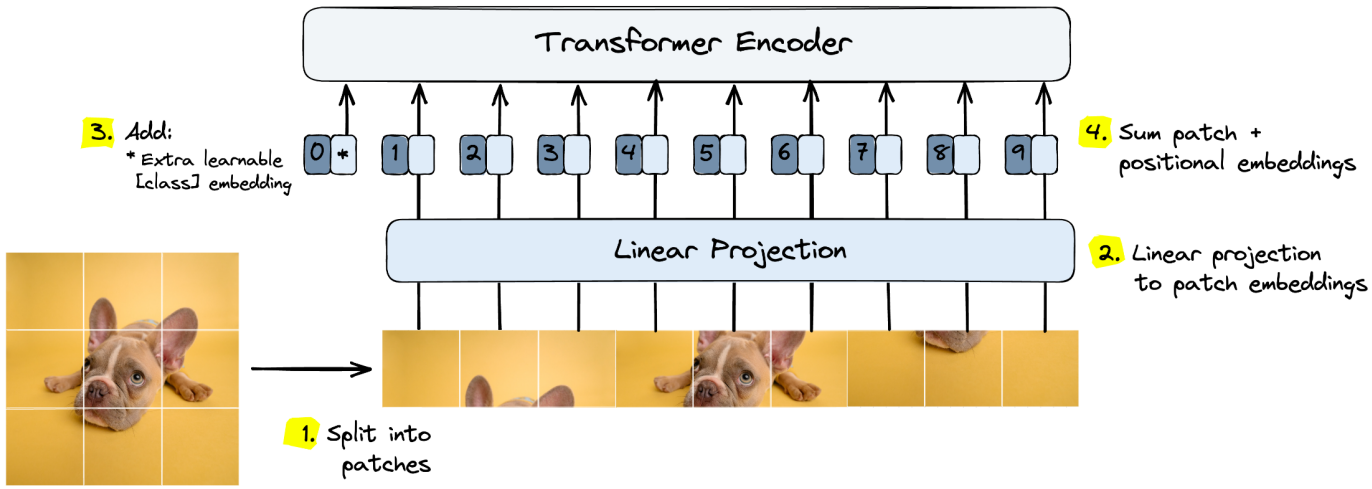
"After turning the image into a sequence of 256 image patches. A linear projection layer is used to map the image patch arrays to patch embedding vectors. The linear projection layer attempts to transform arrays into vectors while maintaining their physical dimensions. Meaning that similar image patches should be mapped to similar patch embeddings" [Bri22].

The vision transformer is designed to start with an extra learnable class embedding that is equivilent to 0 [DBK21]. Which represents the start of the image and sequence of patches to come [DBK21]. The extra learnable class embedding allows the model to learn embeddings specific to each classification label. "The pre-training function of vision transformer is based solely on the classification label given; therefore, the learnable class embedding is even more important to successfully pre-training the vision transformed model" [Bri22]. Without the learnable class embedding, the transformer will not understand the classification labels that are attached to each image.

To keep the order of the sequence of patches that make up the image, the patches are instilled with positional embeddings [Bri22]. "For the vision transformer, these positional embeddings are learned vectors with the same dimensionality as our patch embeddings.

Positional embeddings are learned during pre-training and sometimes during fine-tuning. After creating the patch embeddings and pre-pending the classification label embedding, it is then summed with the positional embeddings" [Bri22]. Finally, the summed embeddings are shown to the transformer encoder. After the entirety of the image is shown to the transformer encoder, the model has then learned that image under the given classification label [WXD20]. For a more understandable visual example, see Figure 2.4 from [Bri22]. It visually demonstrates the architecture of how the image patches are linearly projected and linearly embedded, how the patches receive a learnable class embedding, and then finally showing the image to the transformer encoder and the model learning the given image for the given classification label.

Figure 2.4: Visual Representation of the ViT Architecture



2.2.1 Transfer Learning

Dosovitskiy and et al. also found that "ViT attains excellent results when pre-trained at sufficient scale and transferred to tasks with fewer data points. When pre-trained on the public ImageNet-21k data set or the in-house JFT-300M data set, ViT approaches or beats

state of the art on multiple image recognition benchmarks" [DBK21]. A vision transformer model that has been pre-trained on large data sets such as ImageNet-21k are able to help in making a model more effective [DBK21].

2.3 Approach

The approach for this project is to use a pre-trained vision transformer(ViT) model to train an image classification model to recognize earlier mentioned plant diseases from PlantVillage data set.

2.3.1 ViT model Pre-Trained on ImageNet-21k

This project will implement the vision transformer developed by Dosovitskiy et al. and mentioned in their paper [DBK21]. The framework of their ViT model will be used and accessed through the Hugging Face platform and their package transforms using Python [WXD20]. The vision transformer model comes pre-train on the ImageNet-21k, a benchmark data set consisting of 14 million images and 21k classes [WXD20]. The vision transformer model has been pre-trained on images with pixel size 224×224. Therefore, any data that is to be further trained on this model must also be of pixel size 224×224 [WXD20].

2.3.2 Data Pre-processing

The data pre-processing needed for this project is to resize the images to the requirements of the pre-trained ViT Model which is pixel size 224×224[WXD20]. Images from the PlantVillage data set were originally pixel size 256×256 and resized to be 224×224. The images were normalized across the three RGB channels with means (0.5, 0.5, 0.5) and standard deviation (0.5, 0.5, 0.5) [WXD20].

2.3.3 Data Augmentation

"Data augmentation is the process of transforming images to create new ones for training machine learning models" [DC22]. "Data augmentation increases the number of examples in the training set while also introducing more variety in what the model sees and learns from. Both these aspects make it more difficult for the model to memorize mappings while also encouraging the model to learn general patterns. Data augmentation can be a good substitute when resources are constrained" because it artificially creates more of your data when it is not possible to get more data [DC22].

In the case of this project, the function being used to perform the data augmentations is set_transfrom() from the Hugging Face Datasets package in Python. This function performs data transformations only when the model training begins. Therefore, transformations can be done on the fly and save on computational resources [Con]. Then at each epoch, the transformations are applied to every image given to the model, so the amount of training data stays constant, but variation is added to the original data through transformations [Con]. This does not increase the number of training images as other data augmentation packages would, this artificially augments with transformations and variation [Con].

Data augmentation is an important step when training machine learning models because they can perform very powerfully if the given data sets for training are too small to train with [DC22]. These models can start to over-fit, which is a problem because then the model will memorize mappings between the inputs and expected outputs [DC22]. There are 54,306 images in this data set, which may seem like a lot of images, but for a machine learning model, it is not that much. That is why data augmentation is being implemented as a step to reduce possible model over fitting.

2.3.3.1 Types of Transformations

The data augmentation transformations used for this project are spatial and pixel transformations. The transformations were made using the PyTorch TorchVision package. More fine details for the transformations, how they work, and how to utilize them can be found on the TorchVision documentation [Con17].

Spatial Transformations

- Random Resize Crop: Where a random part of the image is resized and cropped to the size 224×224
- Random Horizontal Image Flips: Where the image is randomly flipped horizontally, left to right or right to left
- Random Vertical Image Flips: Where the image is randomly flipped vertically to be upside down or right side up
- Random Rotation: Where the image can be randomly rotated up to 30 degrees

Pixel level Transformations

- Color Jitter: Randomly changes the colors in the image. The settings used were, brightness = 0.4, contrast = 0.4, saturation = 0.4, and hue = 0.1
- Gaussian blur: Incorporates a purposeful random blur in the image, the kernel size used for the Gaussian Blur was 0.3

2.3.4 Measurements of Performance

The training-test split for this project is 85% training and 15% testing. The training data is further split to be 70% training and 15% validation for training. The classification model

was trained over 10 epochs with a batch size of 32. Hyper-parameters that were used for this model are a learning rate of 0.00005 and a warm-up ratio of 0.1.

F1 Score and Accuracy will be used in order to assess the training of the model over the 10 epochs. Accuracy will also be used to test the model. "Accuracy is the proportion of correct predictions among the total number of cases processed"[PVG11]. See Equation 2.1 for the formula of Accuracy [PVG11]. The model with the best F1 score and lowest training and validation loss will be chosen as the best model over the 10 epochs. The F1 score is the harmonic mean, the average, of precision and recall[PVG11], see Equation 2.4 for the F1 Score Formula [PVG11]. "Precision is the fraction of correctly labeled positive examples out of all of the examples that were labeled as positive", see Equation 2.2 for the Precision formula [PVG11]. "Recall is the fraction of the positive examples that were correctly labeled by the model as positive", see Equation 2.3 for the Recall formula [PVG11].

True positives are the examples correctly labeled as positive, and False positive examples are the examples incorrectly labeled as positive [PVG11].

Accuracy:

$$\text{Accuracy} = \frac{\text{Number of total Correct Predictions}}{\text{Total Number of Predictions}} \tag{2.1}$$

Precision:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{2.2}$$

Recall:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{2.3}$$

F1 Score:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2.4}$$

CHAPTER 3

Results

3.1 Model Training Results

See Table 3.1, for training results of the pre-trained ViT Image classification model trained on the PlantVillage data set. The table displays the model's Training Loss, Validation Loss, Precision, Recall, model F1 score, and model Accuracy over the 10 epochs. Model training loss indicates how well the model is fitting the training data. Model validation loss indicates how well the model fits new data. The best model that was chosen was epoch 10, which is bold in Table 3.1. The model from epoch 10 has a training loss of 0.088 with a validation loss of 0.073, the lowest pair out of all the epochs. For this model, Precision, Recall, the model F1 Score, and Accuracy all have the value of 1.00. These results indicate that the model has a high positive identification rate. The model seems to have reached a convergence in values between epochs 8 to 10. The training results for epoch 10 showed that the Evaluation of Samples per Second for the model is 50.29. This indicates that number of samples the machine learning model can process and make predictions of in one second is 50.29. Also, the Evaluation Steps per Second is 1.58, which indicates the number of iterations that the machine learning model can complete in one second.

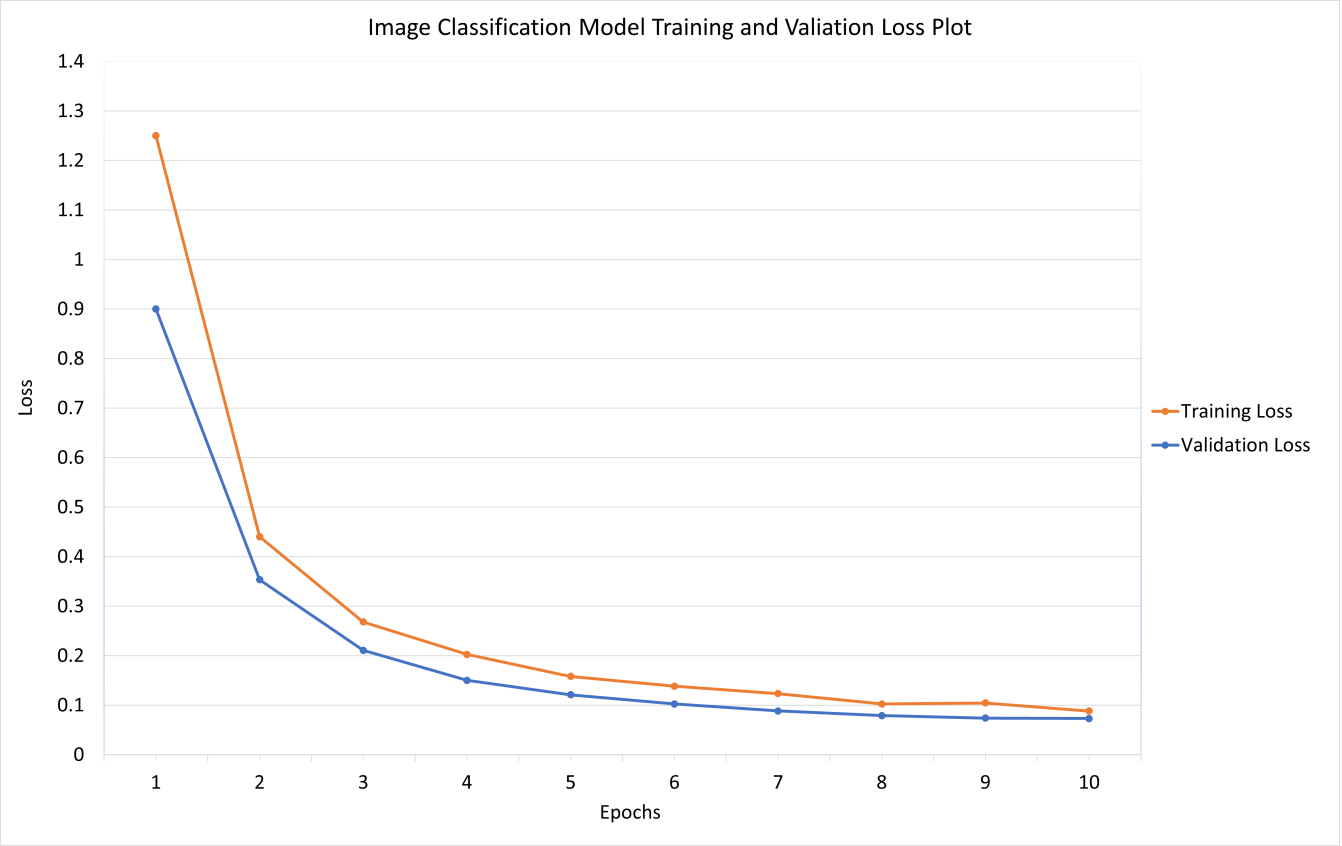
Table 3.1: Model Training Results

Epoch	**Training Loss**	**Validation Loss**	**Precision**	**Recall**	**F1 Score**	**Accuracy**
1	1.2499	0.9951	0.915733	0.915733	0.915733	0.915733

2	0.4403	0.3535	0.984533	0.984533	0.984533	0.984533
3	0.2681	0.2107	0.997333	0.997333	0.997333	0.997333
4	0.2026	0.1501	0.998400	0.998400	0.998400	0.998400
5	0.1580	0.1211	0.998400	0.998400	0.998400	0.998400
6	0.1382	0.1024	0.998400	0.998400	0.998400	0.998400
7	0.1233	0.0882	0.999467	0.999467	0.999467	0.999467
8	0.1026	0.0789	0.999467	0.999467	0.999467	0.999467
9	0.1046	0.0740	1.000000	1.000000	1.000000	1.000000
10	**0.0881**	**0.0730**	**1.000000**	**1.000000**	**1.000000**	**1.000000**

See Figure 3.1 for a comparison plot between the model's training loss and validation loss over the course of the 10 epochs. The plot shows a visual representation of the relationship between training and validation loss. Of the course of the 10 epochs, the training and validation loss are both decreasing. The lines do not cross each other. Since both lines are gradually decreasing and getting closer to 0, it indicates that over time the model is slowly learning the underlying patterns in the data.

Figure 3.1: Image Classification Model Training Loss and Validation Loss Plot over Epochs



3.2 Model Testing Results

Model testing was done with the 15% of the data that was reserved for testing and never shown to the model during training. See Table 3.2, in order to see the overall testing results of the model in a table. The model has an Accuracy of 99%. Meaning the model has the ability to guess the classification label 99% of the time correctly. Total Time in Seconds is the amount of time it took for the image classification model to be tested with the test data set, which is 1460.88 seconds also 24.34 minutes. Samples per Second is 1.15, which is the amount of data samples or instances that the image classification model can process and make predictions on in one second. Latency in Seconds is 0.86, which refers to the amount of time it takes for a single prediction to be made by the image classification model in one

second.

Table 3.2: Model Testing Result

Accuracy	0.9994
Total Time in Seconds	1460.88
Samples per Second	1.1554
Latency in Seconds	0.8654

3.2.1 Model in Action

In order to see how the image classification model can classify an image, see Figure 3.2 and Table 3.3. Figure 3.2 shows an example image of which the image classification model was tested on. The image has its true classification label above it, which is Apple-Apple Scab. Table 3.3 has the scores and labels of five different classification label predictions for what it thinks the image in Figure 3.2 could be. The score is on a scale from 0 to 1, with 1 meaning 100% confidence in the classification label that the model is predicting. The value of the score is split across the five predictions, meaning when we add up all of the score values for all five predictions, the value will add up to 1. The model's top prediction with a score of 0.91% is the classification label Apple-Apple Scab, which is the true classification label of the image in Figure 3.2.

Figure 3.2: Example Image the Image Classification Model was tested on

Table 3.3: Image Classification Model Predictions for the Image in Figure 3.2

Score	Label
0.9162	**Apple-Apple Scab**
0.0092	Apple-Cedar Apple Rust
0.0080	Apple-Healthy
0.0075	Peach-Bacterial Spot
0.0071	Potato-Early Blight

CHAPTER 4

Discussion

The overall performance of the pre-trained ViT image classification model with data augmentation shows good promise. The best epoch provided that F1, Accuracy, Precision, and Recall, were all equal to 1.0. This is not the best situation but shows there is room for improvement. Given the previous epochs from 1 to 8, has values for their F1 Score and Accuracy. It shows that there is possible room to improve the model with more fine-tuning and possibly an early stop in training to not have the model over-train to get the result of 1.0 for F1 Score and Accuracy. Another possible reason for the result could be the large imbalance of images between classification labels and that the Plant Village data set is considered to be a small data set. These types of results are good but not realistic for model prediction power. Even though these values are incredibly high, the training and validation loss is still fairly low and slowly converging to 0. This means that the model is learning over time, which is a good sign and shows room for improvement.

A possible future improvement for this project would be to find another data set with more specific plant disease information to train the model on. Efforts will be made to look for data that includes more plant pests. That variation would be beneficial because the data set for this project mainly contains crop disease images. The disease Spider Mite(Two-Spotted), is labeled as a disease in this data set but in reality, is not a disease. It is the only classification label that has pest-inflected crop images in this data set. Having more data on crop pests would be beneficial because crop pests also cause a significant amount of crop loss and damage as well.

Technical improvements for this project include developing a stronger data augmentation technique. Instead of using the set_transform() method which artificially augments the data set, using another package that can actually create the separate images and add them to the data set instead would be interesting to see how it would perform. More Hyper-parameter fine-tuning could be done such as exploring the learning rate and adding an optimizer. More Epochs should be tested in order to see how the model will perform over a greater period of time. Fewer epochs will also be tested to see how early-stopping the model from training will perform. Another idea is to explore more the Training-Validation-Testing splits chosen for the data. Exploring the performance between different splits could show how to better improve the classification model.

To achieve our expected agricultural need of feeding 10 billion people by 2050, we must prioritize minimizing crop loss wherever possible. More research is needed to help develop more tools to assist crop growers with preventing crop loss. The study "Mobile phone use is associated with higher smallholder agricultural productivity in Tanzania, East Africa" by Amy Quandt et al. looks into the relationship crop growers have with their cell phones as agricultural tools to help increase crop yields [QSN20]. "A key result is the positive association between phone use for agricultural activities and self-reported agricultural yields" [QSN20]. Cell phones are increasing accessibility to technological tools that help with agriculture. These technologies for assisting with crop loss will not only be utilized by commercial crop growers or the average hobbyist and enthusiast as well. Whether crop growers use a ViT image classification model or a convolutional neural network image classification model, or another type of machine learning architecture is used, more research is needed to help develop tools to assist crop growers worldwide in eradicating crop loss everywhere!

REFERENCES

[AU22] Animal and Plant Health Inspection Service of U.S. DEPARTMENT OF AGRICULTURE. "Citrus Greening.", Dec 2022. https://www.aphis.usda.gov/aphis/ourfocus/planthealth/plant-pest-and-disease-programs/pests-and-diseases/citrus/citrus-greening#:~:text=Huanglongbing%20(HLB)%2C%20also%20known,when%20feeding%20on%20new%20shoots.

[Bri22] James Briggs. "Vision-Transformers.", Nov 2022. https://www.pinecone.io/learn/vision-transformers/.

[Con] Hugging Face Contributors. "Hugging Face Datasets Package Reference: Main Classes." https://huggingface.co/docs/datasets/v2.13.0/en/package_reference/main_classes#datasets.Dataset, https://discuss.huggingface.co/t/image-data-augmentation-vit/20146.

[Con17] Torch Contributers. "Pytorch Documentation: Transforming and Augmenting Images.", 2017. https://pytorch.org/vision/stable/transforms.html#torchvision.transforms.ColorJitter.

[CR13] University of California Agriculture and Resources. "Agriculture: Tomato Pest Management Guidelines, Tomato Yellow Leaf Curl.", Dec 2013. https://ipm.ucanr.edu/agriculture/tomato/tomato-yellow-leaf-curl/.

[DBK21] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby, equal technical contribution, equal advising Google Research, and Brain Team. "AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE." *International Conference on Learning Representations*, 2021. https://arxiv.org/pdf/2010.11929v2.pdf, https://huggingface.co/docs/transformers/model_doc/vit#vision-transformer-vit.

[DC22] Ashwin D'Cruz. "What is data augmentation in deep learning?", May 2022. https://www.calipsa.io/blog/what-is-data-augmentation-in-deep-learning.

[Dir22] Caribbean Plant Heath Directors. "Plant Disease – Crop Loss.", May 2022. https://www.cphdforum.org/index.php/2022/05/26/plant-disease-crop-loss/.

[FU19] Food and Agriculture Organization of the United Nations. "New standards to curb the global spread of plant pests and diseases.", Apr 2019. https://www.fao.org/news/story/en/item/1187738/icode/.

[MHS16] Sharada P. Mohanty, David P. Hughes, and Marcel Salathé. "Using Deep Learning for Image-Based Plant Disease Detection." *Frontiers in Plant Science*, **7**:1419, Sep 2016. https://www.frontiersin.org/articles/10.3389/fpls.2016.01419/full.

[PVG11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research*, **12**:2825–2830, 2011. https://huggingface.co/spaces/evaluate-metric/precision, https://huggingface.co/spaces/evaluate-metric/recall, https://huggingface.co/spaces/evaluate-metric/f1, https://huggingface.co/spaces/evaluate-metric/accuracy.

[QSN20] Amy Quandt, Jonathan D. Salerno, Jason C. Neff, Timothy D. Baird, Jeffrey E. Herrick, J. Terrence McCabe, Emilie Xu, and Joel Hartter. "Mobile phone use is associated with higher smallholder agricultural productivity in Tanzania, East Africa." *PLOS ONE*, **15**(8), 2020. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7410319/.

[RAB21] Jean B. Ristaino, Pamela K. Anderson, Daniel P. Bebber, Kate A. Brauman, Nik J. Cunniffe, Nina V. Fedoroff, Cambria Finegold, Karen A. Garrett, Christopher A. Gilligan, Christopher M. Jones, Michael D. Martin, Graham K. MacDonald, Patricia Neenan, Angela Records, David G. Schmale, Laura Tateosian, and Qingshan Wei. "The persistent threat of emerging plant disease pandemics to global food security." *Proceedings of the National Academy of Sciences*, **118**(23):e2022239118, 2021. https://www.pnas.org/doi/abs/10.1073/pnas.2022239118.

[Ser22] USDA Agricultural Research Service. "Food Security: How Do Crop Plants Combat Pathogens?", Oct 2022. https://www.ars.usda.gov/oc/dof/food-security-how-do-crop-plants-combat-pathogens/#:~:text=The%20challenge%20is%20so%20great,%2D%24200%20billion)%20each%20year.

[SJG21] Marissa Schuh, Anna Johnson, Michelle Grabowski, and Angela Orshinsky. "Bacterial spot of tomato and pepper.", 2021. https://extension.umn.edu/disease-management/bacterial-spot-tomato-and-pepper.

[Sol21] Natalie Solares. "Why proper identification of plant diseases matters and how we can assist farmers.", Sep 2021. https://ucanr.edu/blogs/blogcore/postdetail.cfm?postnum=50453#:~:text=Pathogen%20identification%20is%20difficult%20to,training%20for%20an%20accurate%20diagnosis.

[WXD20] Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph Gonzalez, Kurt Keutzer, and Peter Vajda. "Visual Transformers: Token-based Image Representation and Processing for Computer Vision.", 2020. https://huggingface.co/google/vit-base-patch16-224-in21k.