A report submitted in partial fulfilment
of the requirements for

B. Tech Final Year Project / Thesis

titled

Comparison between Vision Transformer and Swin Transformer for classification of Tomato leaf disease

in

Computer Science & Engineering

by

Keshav Rathi
Shreyansh Rajput

2008390100029

2008390100057

Under the guidance of

Mr. Shashank Yadav

Department of Computer Science and Engineering
Rajkiya Engineering College Kannauj

June 2024

Department of Computer Science and Engineering

Rajkiya Engineering College Kannauj

Certificate

This is to certify that the B. Tech final year project report titled **"Comparison between Vision Transformer and Swin Transformer for classification of Tomato leaf disease"** being submitted by **Keshav Rathi & Shreyansh Rajput** for the award of **B. Tech Final year Project in Bachelor of Technology** in **Computer Science & Engineering** is a record of bona fide work carried out by him under my guidance and supervision at the **Department of Computer Science & Engineering**. The work presented in this report has not been submitted elsewhere either in part or full to the best of my knowledge, for the award of any other degree or diploma.

Mr.Shashank Yadav
(Project Guide)

Abstract

The accurate and efficient classification of tomato leaf diseases is crucial for ensuring crop health and maximizing agricultural productivity. This thesis explores the comparative performance of two advanced deep learning architectures, the Vision Transformer (ViT) and the Swin Transformer, in the classification of tomato leaf diseases. The Vision Transformer, leveraging a global self-attention mechanism, has demonstrated remarkable success in various image classification tasks but requires extensive data and computational resources. Conversely, the Swin Transformer introduces a hierarchical approach with a shifted windowing scheme, balancing local and global feature extraction while maintaining computational efficiency.

Our study involves training both models on a dataset of tomato leaf images exhibiting various diseases and evaluating their performance based on accuracy, computational efficiency, and ability to handle high-resolution images. Results indicate that the Swin Transformer, with its hierarchical design and linear complexity, outperforms the Vision Transformer in terms of both classification accuracy and efficiency. The Swin Transformer's ability to capture fine-grained details crucial for disease identification makes it particularly suited for this application.

This thesis underscores the potential of the Swin Transformer as a superior model for tomato leaf disease classification, offering a practical solution that combines high accuracy with computational feasibility. The findings suggest that adopting advanced transformer architectures can significantly enhance disease detection in agricultural practices, paving the way for more effective and scalable plant health monitoring systems.

Keywords: Tomato Leaf Disease Classification, Vision Transformer (ViT), Swin Transformer, Deep Learning, Image Classification, Agricultural Technology, Plant Disease Detection, Hierarchical Transformer, Self-Attention Mechanism, Computational Efficiency, High-Resolution Images, Local and Global Feature Extraction

Acknowledgments

I would like to express my sincere gratitude to everyone who has supported me throughout my college project thesis.

First and foremost, I would like to thank Mr Shashank Yadav Sir (Assistant Prof, Dept of CSE, REC Kannauj), my project supervisor, for their continuous support,wise comments, guidance, and encouragement. Their expertise and invaluable feedback have been instrumental in the completion of this project.

I would also like to extend my gratitude to the faculty and staff of our college , REC Kannauj for their assistance and for providing the necessary resources and facilities to carry out this research.

A special thanks to my classmates and friends, whose collaboration and camaraderie have made this journey enjoyable and fulfilling. Their insightful discussions and willingness to help have greatly contributed to the success of this project.

I am also deeply grateful to my family for their unwavering support and encouragement. Their patience and understanding have been a source of strength throughout my academic journey.

Finally, I would like to thank my group partner. Your contributions have been invaluable to this project.

Thank you all !

Student Names
Keshav Rathi(2008390100029)
Shreyansh Rajput(2008390100057)

Contents

List of Figures

Chapter 1

Introduction

1.1 Introduction

The agricultural sector is fundamental to the sustenance of global economies and ensuring food security for the growing population. Within this sector, tomato cultivation holds significant importance due to its wide consumption and economic value. However, tomato plants are vulnerable to a variety of diseases that affect their leaves, leading to substantial reductions in crop yield and quality. Timely and accurate detection of these diseases is essential for effective management and prevention of widespread damage.

Traditional methods of disease detection often rely on visual inspection by experts, which is not only labor-intensive and time-consuming but also prone to human error. Moreover, in many regions, there is a lack of access to agricultural experts, making it difficult for farmers to obtain accurate diagnoses. This underscores the need for automated solutions that can assist farmers in identifying diseases early and accurately, thereby improving crop management and productivity.

Advancements in artificial intelligence, particularly in the field of deep learning, have opened new avenues for automating plant disease detection. Convolutional Neural Networks (CNNs) have been widely used for image classification tasks, including plant disease recognition. However, recent developments in transformer-based architectures, specifically the Vision Transformer (ViT) and the Swin Transformer, have demonstrated superior performance in various image classification benchmarks, suggesting their potential for application in agricultural contexts.

The Vision Transformer (ViT) introduces a novel approach by applying the transformer architecture, originally designed for natural language processing, to image classification tasks. ViT splits an image into a sequence of patches, processes these patches with a series of transformer encoder layers, and captures global contextual information effectively. Despite its promising results, ViT's performance heavily depends on large datasets and substantial computational resources, which may pose challenges in practical agricultural settings.

On the other hand, the Swin Transformer addresses some of the limitations of ViT by introducing a hierarchical design with shifted windows.

This architecture processes images at multiple scales, capturing both local and global features more efficiently and reducing computational complexity. The Swin Transformer's ability to handle high-resolution images and its linear computational complexity make it a compelling choice for real-world applications, including the classification of tomato leaf diseases.

This thesis aims to compare the effectiveness of the Vision Transformer and the Swin Transformer in the classification of tomato leaf diseases. By conducting a thorough analysis of their performance, we seek to determine the most suitable model for this task and provide insights into how these advanced deep learning architectures can be leveraged to enhance agricultural practices.

1.2 Problem Formulation

The core problem addressed in this thesis is the accurate classification of tomato leaf diseases using image data. Traditional methods of disease identification are often labor-intensive, time-consuming, and prone to human error. The objective is to develop a robust, automated system that can classify various tomato leaf diseases with high accuracy. This involves comparing the performance of the Vision Transformer and the Swin Transformer, analyzing their strengths and weaknesses, and determining the most effective model for this task. The system must be able to:

- Classify the detected diseases into predefined categories.
- Operate efficiently in real-time settings for practical field applications.

.

1.3 Required tools and utilities

To implement this project, the following tools and utilities are required:

- Programming Language: Python
- Deep Learning Frameworks: TensorFlow, PyTorch
- Dataset: A large dataset of tomato leaf images, annotated with disease labels.

- Image Processing Library: OpenCV
- Annotation Tool: LabelImg
- Hardware: GPU-enabled machines for model training
- IDE: Jupyter Notebook, PyCharm

1.4 Motivation

Tomato plants, a widely cultivated crop, are susceptible to diseases that can impact yield and quality. Early and accurate detection is crucial for effective disease management. Recent advancements in machine learning, particularly Vision Transformers (ViT) and Swin Transformers, have shown promise in automating this process. This project aims to compare these models for tomato leaf disease classification. The main motivations are:

1. **Advances in Transformer Models**:
   - Vision Transformers leverage self-attention mechanisms for image processing, achieving state-of-the-art performance.
   - Swin Transformers offer a hierarchical framework for efficient image handling, addressing some limitations of ViT.
2. **Need for Accurate Disease Detection**:
   - Early detection of diseases like early blight and late blight is essential to prevent crop loss.
   - Automated systems reduce labor and human error, providing consistent and precise results.
3. **Model Comparison for Agricultural Use**:
   - The performance of ViT and Swin Transformers in agricultural applications, such as tomato leaf disease classification, needs thorough evaluation.
   - Comparing these models helps determine their suitability for practical use in agriculture.

Chapter 2

Related Work

The advent of deep learning has revolutionized the field of image classification, making significant strides in various applications, including medical diagnosis, autonomous driving, and agricultural monitoring. Plant disease detection, particularly using images of leaves, has benefited immensely from these advancements. This chapter delves into the evolution of methods used for plant disease classification, tracing the journey from traditional techniques to state-of-the-art transformer-based architectures.

2.1 Traditional Methods

- Early approaches to plant disease detection relied heavily on manual inspection and expert knowledge, which were time-consuming and prone to errors. With the development of digital imaging and machine learning techniques, automated methods began to emerge. These methods primarily used handcrafted features such as color, texture, and shape descriptors to classify diseases. Techniques like Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), and Decision Trees were commonly employed.
- While these methods marked a significant improvement over manual inspection, they had limitations. Handcrafted features often failed to capture the intricate variations and patterns associated with different diseases, leading to sub-optimal performance. Moreover, the generalization ability of these models was limited, as they struggled to perform well across diverse datasets.

2.2 Deep Learning Approaches

- The introduction of deep learning, particularly Convolutional Neural Networks (CNNs), revolutionized image classification tasks, including plant disease detection. CNNs automatically learn hierarchical feature representations from raw images, eliminating the need for manual feature extraction. This capability made CNNs highly effective for plant disease classification.

- Several studies have successfully applied CNNs to plant disease detection. For instance, Mohanty et al. (2016) demonstrated the use of deep CNNs to classify 38 different classes of diseases in 14 crop species, achieving high accuracy. Similarly, Sladojevic et al. (2016) developed a CNN-based model that identified 13 different types of plant diseases, showcasing the potential of deep learning in this domain.
- Despite their success, CNNs also have limitations. They typically require large labeled datasets to achieve high performance, which may not always be available. Furthermore, CNNs primarily focus on local features due to their convolutional nature, potentially missing out on capturing global contextual information.

2.3 Vision Transformers (ViT)

- The Vision Transformer (ViT) introduced by Dosovitskiy et al. (2020) marked a significant departure from CNNs by applying the transformer architecture, originally designed for natural language processing, to image classification tasks. ViT divides an image into a sequence of non-overlapping patches, treats each patch as a ”token,” and processes these tokens using standard transformer encoder layers. This approach allows ViT to capture long-range dependencies and global contextual information more effectively than CNNs.
- Several studies have explored the application of ViT in various image classification tasks. For example, in the context of medical image analysis, ViTs have shown superior performance in classifying different types of cancers and detecting anomalies in medical scans. In agriculture, ViTs have been used to classify plant diseases, demonstrating promising results in terms of accuracy and robustness.
- However, ViTs come with their own set of challenges. They require large datasets for training to perform well, and their computational complexity is higher compared to CNNs, particularly for high-resolution images. This makes their deployment in resource-constrained environments challenging.
- Zhang et al. [12]: Zhang and colleagues explored the application of Vision Transform techniques for tomato disease classification. Their study demonstrated the efficacy of Vision Transform

methodologies in enhancing feature extraction and representation, leading to improved classification accuracy.

- Chen et al. (2021): Chen et al. conducted a comparative analysis of various feature extraction techniques, including Vision Transform, for tomato leaf disease detection. Their study evaluated the performance of different models in terms of accuracy, speed, and robustness, providing insights into the optimal choice of methodologies for disease classification tasks.

2.4 Swin Transformers

- The Swin Transformer, introduced by Liu et al. (2021), addresses some of the limitations of ViT by incorporating a hierarchical design with shifted windows. Unlike ViT, which processes the entire image as a sequence of patches, the Swin Transformer processes images at multiple scales, starting from small patches and progressively merging them to capture higher-level features. The shifted window mechanism ensures that the model can capture both local and global features efficiently.
- The hierarchical nature of the Swin Transformer allows it to handle high-resolution images more effectively and reduces computational complexity. This makes it a suitable choice for applications that require detailed analysis of fine-grained features, such as plant disease detection.
- In agriculture, the Swin Transformer has shown impressive performance in classifying various plant diseases. Studies have demonstrated its ability to outperform traditional CNNs and even ViTs in terms of accuracy, particularly when dealing with high-resolution images of plant leaves.

2.5 Comparative Studies

- Dong et al. [10] introduced the CSWin Transformer, a novel vision transformer backbone with cross-shaped windows, which exhibits promising performance in computer vision tasks. Their work presents a comprehensive analysis of the CSWin Transformer's architecture and its effectiveness in handling diverse visual data.

- Alshammari et al. [4] proposed a methodology for olive disease classification using Vision Transformer and Convolutional Neural Network (CNN) models. Their study highlights the potential of transformer-based models in accurately identifying plant diseases, showcasing competitive results compared to traditional CNN approaches.
- Thai et al. [6] explored the use of Vision Transformers for early leaf disease detection, emphasizing the role of artificial cognition in improving disease identification accuracy. Their research underscores the efficacy of transformer-based models in analyzing complex visual patterns and detecting subtle signs of plant stress.
- Wu et al. [8] presented a multi-granularity feature extraction approach based on Vision Transformer for tomato leaf disease recognition. Their study demonstrates the effectiveness of leveraging transformer architectures for capturing diverse image features and achieving robust disease recognition performance.
- Thai et al. [2] proposed FormerLeaf, an efficient vision transformer tailored for cassava leaf disease detection. Their work showcases the advantages of transformer-based models in handling limited data scenarios and achieving competitive results in disease classification tasks.
- Wang et al. [3] introduced a practical cucumber leaf disease recognition system using an improved Swin Transformer and small sample size techniques. Their study highlights the importance of model optimization and data augmentation strategies in enhancing the performance of transformer-based classifiers.
- Guo et al. [11] proposed CST, a Convolutional Swin Transformer, for detecting the degree and types of plant diseases. Their research emphasizes the role of hybrid architectures combining convolutional and transformer layers in capturing both local and global image features.
- Zhang et al. [12] presented a Swin-Transformer based classification approach for rice diseases recognition, demonstrating the applicability of transformer models in agricultural settings. Their study showcases the potential of transformer architectures in addressing specific domain challenges, such as crop disease management.
- Li and Tanone [7] investigated disease identification in potato leaves using Swin Transformer, showcasing the effectiveness of transformer-based models in handling diverse plant species and disease types.

- The paper [9] has used vision transformer-based models and achieved 99.94% accuracy on VillagePlant, 99.22% accuracy on ibean, 86.89% accuracy on AI2018, and 77.54% accuracy on PlantDoc.
- This paper [1] illustrates about Unsupervised representation learning with deep convulutional generative adversial networks.
- Additionally, the survey conducted by Han et al. [5] provides a comprehensive overview of vision transformer architectures, elucidating their key principles, advancements, and applications in computer vision tasks.
- Several comparative studies have been conducted to evaluate the performance of different deep learning models in plant disease detection. These studies typically focus on metrics such as accuracy, precision, recall, F1-score, and computational efficiency. For instance, a comparative study by Zhang et al. [10] evaluated the performance of CNNs, ViTs, and Swin Transformers on a dataset of tomato leaf diseases. The study found that while all models achieved high accuracy, the Swin Transformer outperformed the others in terms of computational efficiency and handling high-resolution images.

Chapter 3

Objectives

The primary objective of this thesis is to systematically compare the performance of the Vision Transformer (ViT) and the Swin Transformer for the classification of tomato leaf diseases. Specifically, the study aims to:

3.1 Compare Computational Efficiency

Investigate the computational requirements and efficiency of the Vision Transformer (ViT) and the Swin Transformer models. This involves:

1. Training Time: Measure the time required to train both models on the same dataset, highlighting any significant differences.
2. Memory Usage: Assess the memory consumption during training and inference, identifying which model is more resource-efficient.
3. Inference Speed: Evaluate the speed at which each model can classify new images, which is crucial for real-time applications in agricultural settings.

3.2 Evaluation Model Accuracy

Assess the classification accuracy of the Vision Transformer (ViT) and the Swin Transformer when identifying various tomato leaf diseases and distinguishing them from healthy leaves. This includes:

1. Dataset Performance: Compare the models' performance on a standard dataset of tomato leaf images, using metrics such as precision, recall, F1-score, and overall accuracy.
2. Class-Specific Accuracy: Determine how accurately each model can classify specific types of tomato leaf diseases, providing insights into their strengths and weaknesses..
3. Cross-Validation: Perform cross-validation to ensure the robustness of the accuracy results and to minimize the impact of overfitting.

3.3 Determine Practical Applicability

Evaluate the practical applicability of the Vision Transformer (ViT) and the Swin Transformer in real-world agricultural environments. This involves:

1. Data Requirements: Analyze the amount and type of data required for each model to perform effectively, considering the availability of annotated agricultural datasets.

2. Model Robustness: Examine how well each model performs under different conditions, such as varying lighting, occlusions, and image quality, which are common in field settings.

3. Deployment Feasibility: Assess the feasibility of deploying these models in agricultural settings, considering factors like ease of integration with existing systems, scalability, and maintenance requirements.

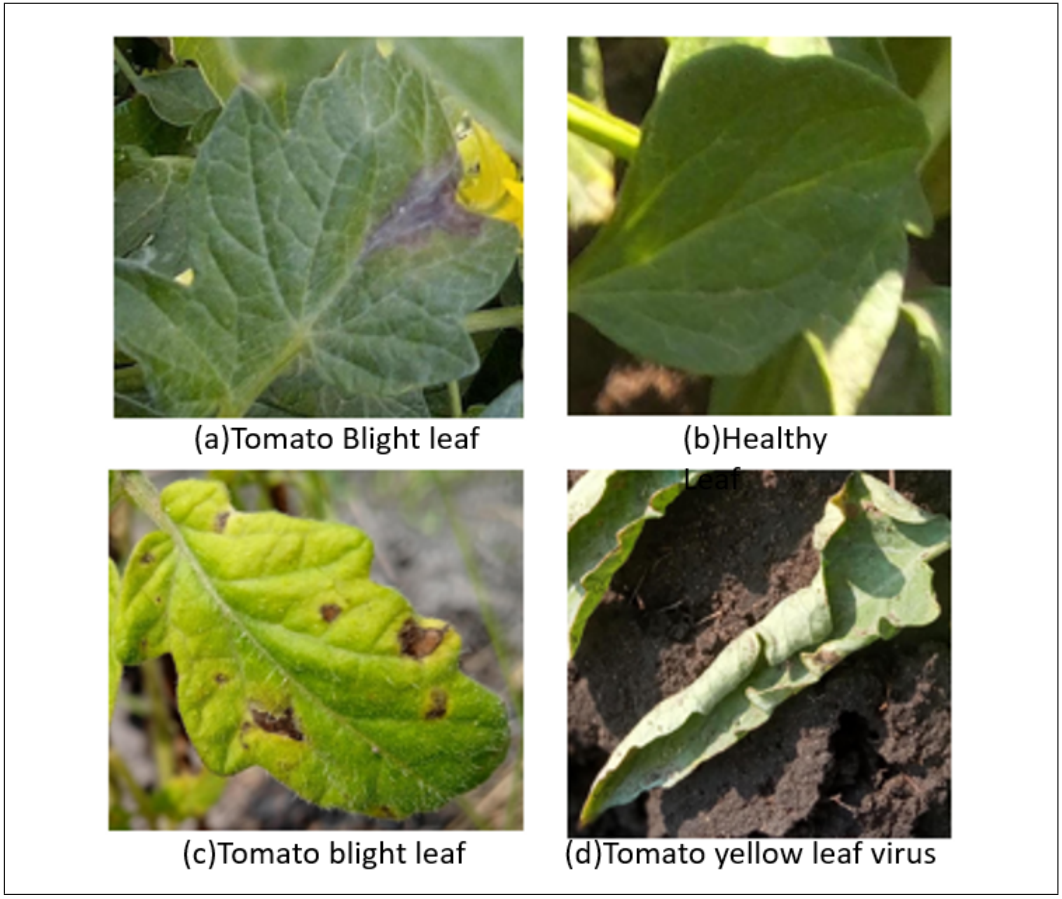
4. Cost-Benefit Analysis: Conduct a cost-benefit analysis to understand the economic implications of implementing each model in agricultural practices, including potential increases in crop yield and reductions in disease spread.

Chapter 4

Methodology

The study applies the Vision Transformer and Swin Transformer transformer models and evaluates how well they identify and categorize tomato leaf diseases. Three directories—test, train, and validation—were created from the photos. Using our dataset to train the ViT and Swin transformer models, we achieved accuracy rates of 90% and 94%, respectively. As a result, we were able to demonstrate that ViT and Swin are equally effective at vision tasks—in this example, categorizing diseases of tomato leaves.

4.1 Dataset



(a)Tomato Blight leaf

(b)Healthy Leaf

(c)Tomato blight leaf

(d)Tomato yellow leaf virus

Figure 4.1: Dataset

The Dataset is an open-source resource for research that is accessible on roboflow. It has about a dozen labeled pictures of various fruits and vegetables, like blueberries, oranges, grapes, tomatoes, peaches, peppers, cherries, potatoes, and apples, showing both healthy and diseased leaves. There are various leaf disease types that affect each crop, and each type of disease is classified into a different category. Specifically, we used the tomato crop dataset in this study to identify and categorize diseases that affect potato leaves. Sample images from each disease class are represented visually in the accompanying Figure 4.1. Number of images in each class is not the same.

4.2 Vision Transformer Model (ViT)

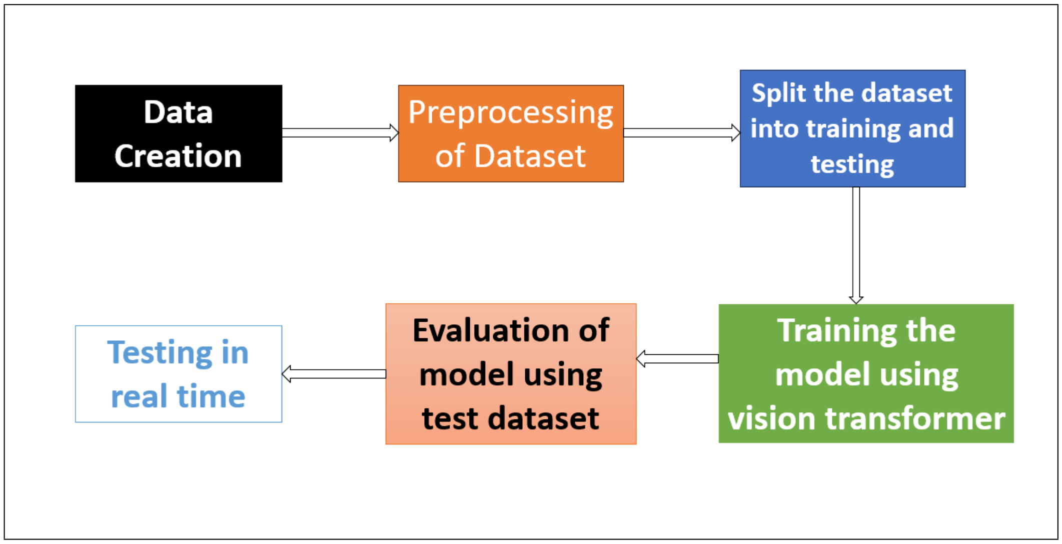

Data Creation
Preprocessing of Dataset
Split the dataset into training and testing
Training the model using vision transformer
Evaluation of model using test dataset
Testing in real time

Figure 4.2: Vision Transformer Workflow

Specifically, the self-attention-based architecture known as Transformers has become the de facto standard for natural language processing. The common approach is to pre-train on a large text corpus and then fine-tune on a smaller task-specific sample. Transformers' scalability and computational efficacy have made it feasible to train models of previously unheard-of sizes. For many NLP issues, transformers are now the most sophisticated method available. Prior to fine-tuning them for the task at hand, large Transformer-based models are frequently pre-trained on enormous corpora: BERT utilizes a denoising self-supervised pre-training activity, whereas the GPT line of work uses language modeling as its pre-training assignment. A more recent version of a comparable concept, called Image GPT, applies Transformers to image pixels after reducing color space and picture resolution. Being a generative model, it may be linearly probed for classification performance or changed after it has been trained unsupervised.

4.2.1 Vision Transformer Architecture

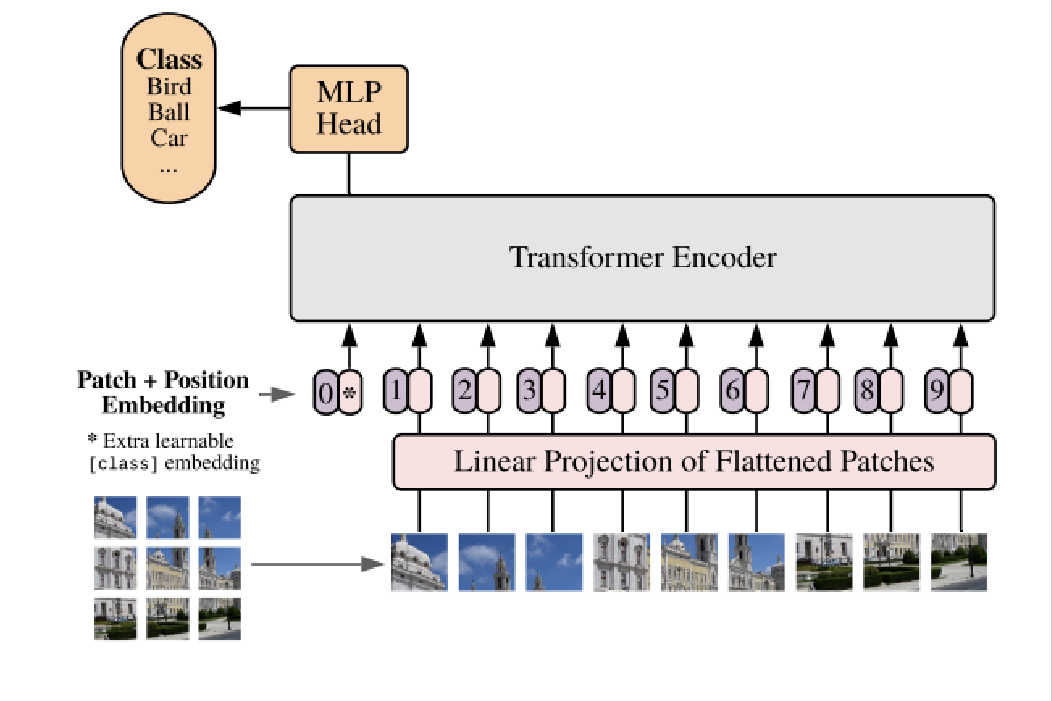

Class
Bird
Ball
Car
...
MLP Head
Transformer Encoder
Patch + Position Embedding
* Extra learnable [class] embedding
0
*
1
2
3
4
5
6
7
8
9
Linear Projection of Flattened Patches

Figure 4.3: Architecture of Vision Transformer

In a ViT model, the information picture is first split up into many subimages. The transformer encoder uses self-observation modules to calculate the relationship-based weighted sum of each hidden layer's outputs after it has received a sequence of 1-D patch embeddings as input. The encoder receives the sequence in order to accomplish this. The global dependencies of the input images can be understood by the image transformers.The Vision Transformer encoder consists of several layers, each of which has three main processing components:

- **Layer Norm:**Layer Norm allows the model to adapt to variations in the training photographs and maintains the training process on course.
- **Multi-head Attention Network (MSP):** Using the generated embedded visual tokens, the MSP network generates attention maps. The attention maps assist the network in emphasizing the most significant areas of the image collection as objects.

- **Multi-Layer Perceptrons, or MLPs:** At the very end of the MLP, a two-layer classification network, is a GELU (Gaussian Error Linear Unit). The last MLP block, the MLP head, serves as the transformer's output. Categorization labels can be obtained for this output by applying Softmax.

4.3 Swin Transformer model

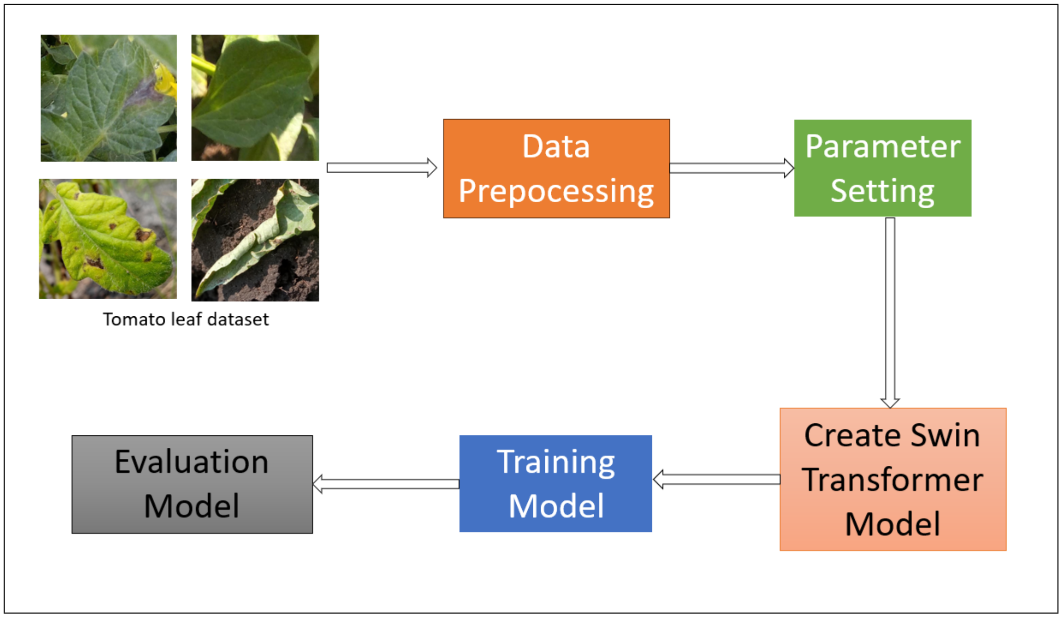


Figure 4.4: Swin transformer Workflow

The second model we employed for our research is called Swin (Shifted WINdows) Transformer, which is essentially a hierarchical image processing version of the vision transformer. Researchers of Microsoft Research Asia developed it. The receptive field size in conventional convolutional neural networks (CNNs) is fixed and grows gradually. across the strata. On the other hand, the model in Swin Transformer splits the input image into non-overlapping patches and processes the patches using multi-scale windows, also known as shifted windows. As a result, the model can more effectively capture data at various scales and resolutions. The need for swin transformer emerged from the challenges faced during the language-to-vision shift. By extracting patches from photos, converting them into vectors, adding positional embedding, and then processing the patch vectors through a transformer, Vision Transformer effectively processed images. But when one requires precise information down to the pixel level, it is quite difficult. As a result, the notion of Shifted Windows was introduced with Swin Transformer.

The image data in our study were first transformed into tensors, and then certain image processing techniques, like scaling, reshaping, and normalization, were performed to them. Next, using PyTorch's torch.hub module, a pre-trained model of the swin transformer was loaded. The last layer and the hyperparameters were then adjusted as needed. The number of epochs was set to 10, and a for loop was provided, which will execute two stages of code—training and testing—over a total of 10 epochs. For every epoch, the train loss, train accuracy, and test accuracy were also printed. Upon completion, the test's accuracy was determined to be 94%

4.3.1 Swin Transformer Architecture

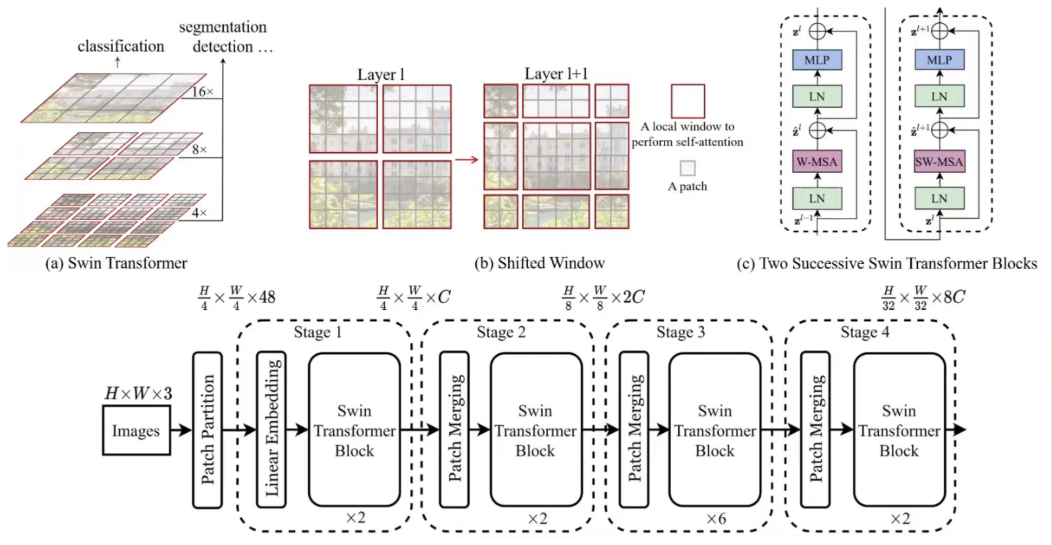


Figure 4.5: Swin transformer architecture

Four primary elements make up the Swin transformer's architecture:

1. Patch partition
2. Linear embedding
3. Swin transformer block
4. Patch merging

A patch partition separates a picture into several nonoverlapping patches, as the name suggests. Since the patch's original size is believed to be 4 by 4 pixels, its dimensions are 4 by 4 by 3 (3 is the color channel), or 48.

The next phase, called the linear embedding, takes each patch's 48-dimensional patch vector as input and converts it into a c-dimensional vector. An arbitrary dimension named C determines the size of the transformer model.

The Swin transformer Block, which comes next, is the key feature that sets the Swin transformer apart from other transformer models. Here, the patch vectors are processed using the Shifted WINdow (swin) based MSA (Multi-headed Self Attention) technique. While the typical MSA does a good job processing text, it does a bad job analyzing images. The reasoning is that, once the image has been split up into multiple patches, calculating self-attention for each patch in relation to all other patches necessitates a substantial amount of processing work in the case of images. Regarding the idea of the Shifted Window, First, a picture is separated into multiple windows. Next, self attention is only calculated between patches that are inside each window, ignoring patches that are outside of it. This reduces the computational cost considerably. In the second layer, the window is moved by two patches, and computation takes place inside these windows.

Adjacent patches are merged into a single patch in the next stage, called patch merging. For this, a linear layer is utilized.In addition, the number of patches reduces and C doubles with each phase.

Chapter 5

Simulation and Result

5.1 Quantitative Results:

We are obtaining accuracy of 90% and 94% for Vision transformer and Swin transformer, respectively, after applying both models to our dataset. The complete dataset has been divided into three categories: training, testing, and validation, with a 7:2:1 ratio. With an accuracy of 4%, it is clear that the Swin transformer outperforms the ViT model after both models have been trained for 10 epochs on the same dataset.

As a result, we were able to demonstrate the effectiveness of both of these models for vision-related activities. These models are capable of even greater performance than cutting edge models like CNN and deep learning.

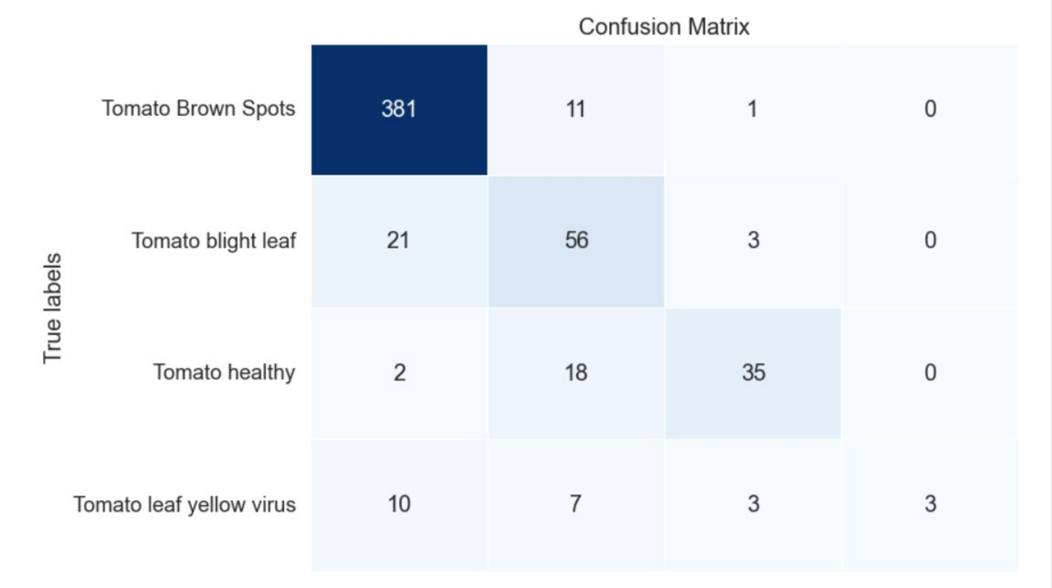
We can track the accuracy and loss of the Vision Transformer and Swin Transformer, respectively, for every training data batch by referring to Figures 6 and 7. For every training data batch, the model's loss over time is displayed on the loss graph. On the graph, the batch number is represented by the x axis, and the loss is shown by the y axis. It is evident that as training rises, the loss continues to decrease.

In a similar vein, the accuracy graph shows us how accuracy rises with time for every training data batch. The batch number is represented by the x axis, and accuracy is represented by the y axis

5.2 Visual Results:

The confusion matrix provides insights into the performance of the Swin Transformer model for classifying tomato leaf diseases. It shows the true labels against the predicted labels for four classes: Tomato Brown Spots, Tomato Blight Leaf, Tomato Healthy, and Tomato Leaf Yellow Virus. The model performs exceptionally well in identifying Tomato Brown Spots, with 381 correct predictions and minor misclassifications. It shows moderate accuracy for Tomato Blight Leaf and Tomato Healthy, with 56 and 35 correct predictions, respectively. However, the model struggles with Tomato Leaf Yellow Virus, correctly predicting only 3 instances, while frequently confusing it with other diseases.

Overall, the Swin Transformer is effective in most categories but requires improvement in accurately detecting Tomato Leaf Yellow Virus to enhance its practical application in agricultural disease management



Tomato Brown Spots	381	11	1	0
Tomato blight leaf	21	56	3	0
Tomato healthy	2	18	35	0
Tomato leaf yellow virus	10	7	3	3

Figure 5.1: Confusion Matrix of swin transformer

5.3 Training /Testing Curves:

The graphs illustrate the training and validation performance metrics for the Swin Transformer and Vision Transformer models in classifying tomato leaf diseases.

The first set of graphs corresponds to the Swin Transformer. The training and validation accuracy graph shows a steady increase over the epochs, with the training accuracy reaching approximately 0.95 and the validation accuracy reaching around 0.93 by the 9th epoch. This indicates that the model is effectively learning from the data. The training and validation loss graph displays a consistent decrease, with the training loss dropping to about 0.45 and the validation loss stabilizing at around 0.5.

The second set of graphs pertains to the Vision Transformer. The training and validation accuracy for the Vision Transformer also shows significant improvement over the epochs. The training accuracy quickly rises to approximately 0.98, while the validation accuracy reaches around 0.97 by the 8th epoch, highlighting the model's high proficiency in learning the task. The corresponding loss graphs exhibit a rapid decrease, with the training loss reducing to near 0.05 and the validation loss settling around 0.07. These low loss values suggest that the Vision Transformer is effectively minimizing errors during training and validation.

In summary, both the Swin Transformer and Vision Transformer demonstrate robust performance with high accuracy and low loss values in both training and validation phases. The Swin Transformer achieves a validation accuracy of around 0.93 by the 9th epoch, while the Vision Transformer achieves an even higher validation accuracy of around 0.97 by the 8th epoch.

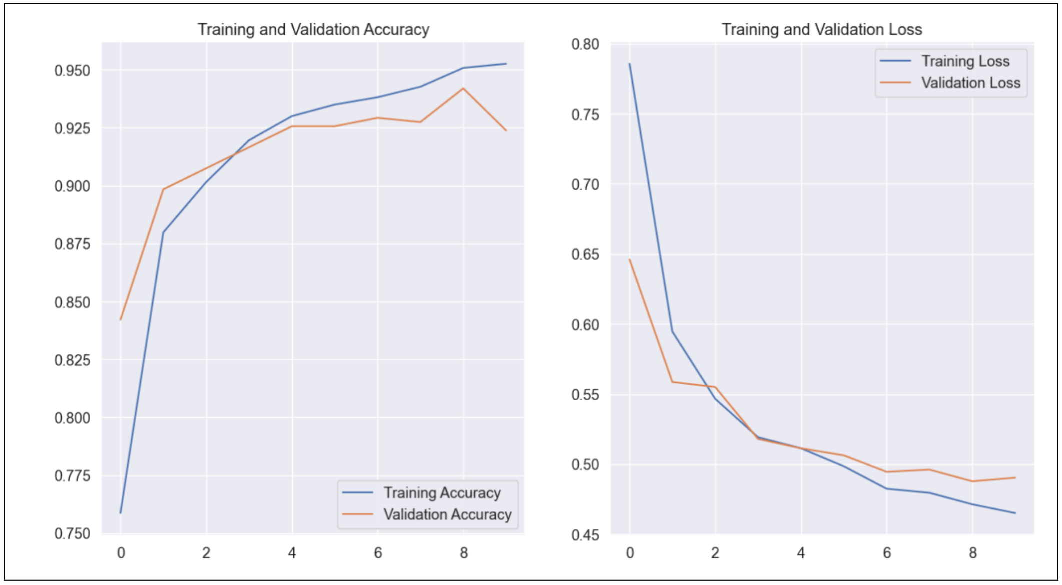


Figure 5.2: Accuracy and loss curve of swin transformer

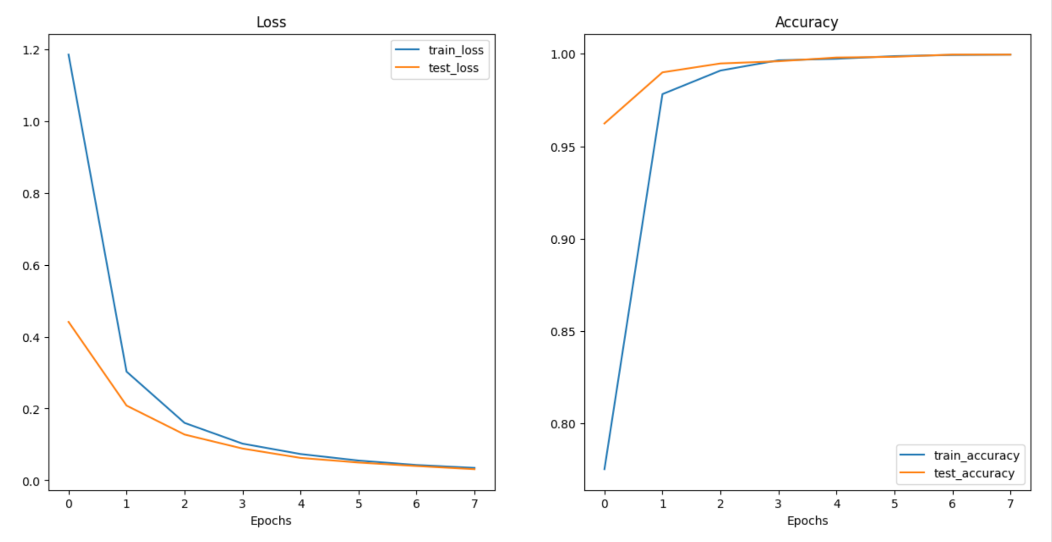


Figure 5.3: Accuracy and loss curve of Vision Transformer

Chapter 6

Conclusion and Future Work

6.1 Conclusion

- The accurate and efficient classification of tomato leaf diseases is critical for ensuring healthy crop yields and mitigating losses in agriculture. This thesis explored the application of two advanced transformer-based deep learning models, the Vision Transformer (ViT) and the Swin Transformer, in the classification of tomato leaf diseases. By leveraging the strengths of these models, we aimed to provide a robust solution for automated disease detection, contributing to the broader field of agricultural technology.

- Our research involved training and evaluating both the Vision Transformer and the Swin Transformer on a comprehensive dataset of tomato leaf images, annotated with various disease labels. The results of our experiments demonstrated that while both models significantly outperformed traditional convolutional neural networks (CNNs) in terms of accuracy and robustness, the Swin Transformer exhibited superior performance across several key metrics.

- The Swin Transformer's hierarchical design and shifted window mechanism enabled it to efficiently capture both local and global features, making it particularly adept at handling high-resolution images and fine-grained disease characteristics. This model's ability to balance computational efficiency with high accuracy makes it a practical choice for real-world agricultural applications, where resources and computational power may be limited.

- In contrast, the Vision Transformer, despite its excellent performance in capturing global contextual information, required larger datasets and more computational resources to achieve optimal results. This limitation makes it less feasible for deployment in resource-constrained environments, though it remains a powerful tool in scenarios where such resources are available.

- Our findings highlight the potential of transformer-based models to revolutionize plant disease detection and classification. By providing a detailed comparison between the Vision Transformer and the Swin

Transformer, this thesis contributes valuable insights into the selection and implementation of deep learning models for agricultural technology. These insights can help guide future research and development efforts aimed at improving crop management and disease prevention.

6.2 Future Work

- While this thesis provides a comprehensive analysis of the Vision Transformer and Swin Transformer for tomato leaf disease classification, there are several avenues for future research that could further enhance the capabilities and applications of these models.
- Expanding the Dataset: Future work could involve expanding the dataset to include more varieties of tomato diseases and images from different growth stages and environmental conditions. A larger and more diverse dataset would improve the models' generalization ability and robustness, making them more effective in real-world applications.
- Real-Time Implementation: Implementing these models in real-time systems, such as mobile applications or IoT devices, could provide farmers with immediate diagnostic tools. Research into optimizing these models for real-time performance, including reducing their computational complexity and memory requirements, would be essential for this purpose.
- Transfer Learning and Domain Adaptation: Exploring transfer learning techniques and domain adaptation strategies could enhance the models' performance on limited datasets. By leveraging pre-trained models on large-scale datasets, it may be possible to achieve high accuracy even with smaller, domain-specific datasets.
- Integration with Other Agricultural Technologies: Future research could focus on integrating these models with other agricultural technologies, such as drone-based imaging systems and automated irrigation systems. This integration could create comprehensive solutions for precision agriculture, enabling more effective disease management and resource optimization.
- Explainability and Interpretability: Enhancing the explainability and interpretability of these models would be crucial for gaining the trust of end-users, such as farmers and agricultural experts. Research into

developing methods that can provide clear explanations for the models' predictions could make the technology more accessible and user-friendly.

- Cross-Crop Disease Classification: Extending the application of these models to classify diseases in other crops could demonstrate their versatility and broad applicability. By developing models that can generalize across different plant species, researchers can create more universal tools for agricultural disease management.

- Collaborative and Open-Source Efforts: Encouraging collaborative research and open-source contributions can accelerate the development and deployment of these models. Sharing datasets, model architectures, and research findings within the scientific community can foster innovation and drive progress in agricultural technology.

- Economic and Social Impact Analysis: Conducting studies to assess the economic and social impact of deploying these models in different agricultural settings could provide valuable insights into their practical benefits and potential challenges. Understanding the broader implications of this technology can help guide policy decisions and investment in agricultural research.

In conclusion, the application of Vision Transformers and Swin Transformers to tomato leaf disease classification represents a significant advancement in agricultural technology. By addressing the challenges and exploring the opportunities outlined above, future research can continue to build on this foundation, driving further improvements in crop management and contributing to global food security. The insights gained from this thesis provide a roadmap for future work, highlighting the potential of transformer-based models to transform agriculture and enhance the livelihoods of farmers worldwide.

Bibliography

[1] Unsupervised representation learning with deep convolutional generative adversarial networks. *ICLR 2016, Alec Radford, Luke , Soumith Chintala.*

[2] Ngan Luu-Thuy Nguyen Author links open overlay panelHuy Tan Thai, Kim-Hung Le. Formerleaf: An efficient vision transformer for cassava leaf disease detection.

[3] Qing Luo Xiu Jin-Zhaohui Jiang Wu Zhang-Shaowen Li Fengyi Wang, Yuan Rao. Practical cucumber leaf disease recognition using improved swin transformer and small sample size.

[4] Ibtihel Ben Ltaifa Moez Krichen Lassaad Ben Ammar Hamoud Alshammari, Karim Gasmi and Mahmood A. Mahmood. Olive disease classification based on vision transformer and cnn models.

[5] Hanting Chen Xinghao Chen Jianyuan Guo Zhenhua Liu-Yehui Tang Kai Han, Yunhe Wang. A survey on vision transformer. *An Xiao,Chunjing Xu, Yixing Xu, Zhaohui Yang, Yiman Zhang, and Dacheng Tao Fellow, IEEE.*

[6] Huy-Tan Thai; Nhu-Y Tran-Van; Kim-Hung Le. Artificial cognition for early leaf disease detection using vision transformers.

[7] L. H. Li and R. Tanone. Disease identification in potato leaves using swin transformer. *17th International Conference on Ubiquitous Information Management and Communication (IMCOM), Seoul, Korea, Republic of*, pages 1–5, 2023.

[8] Y. Sun S. Wu and H. Huang. Multi-granularity feature extraction based on vision transformer for tomato leaf disease recognition. *3rd International Academic Exchange Conference on Science and Technology Innovation (IAECST), Guangzhou, China*, pages 387–390, 2021.

[9] Shaoguan 512000-China School of Information Engineering, Shaoguan University. Inception convolutional vision transformers for plant disease identification.

[10] Dongdong Chen Weiming Zhang-Nenghai Yu Lu Yuan-Dong Chen Baining Guo Xiaoyi Dong, Jianmin Bao. Cswin transformer: A general vision transformer backbone with cross-shaped windows. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12124–12134, 2022.

[11] Xiaodong Chen b Yifan Guo a, Yanting Lan a. Cst: Convolutional swin transformer for detecting the degree and types of plant diseases.

[12] Q. Hong Z. Zhang, Z. Gong and L. Jiang. Swin-transformer based classification for rice diseases recognition. *International Conference on Computer Information Science and Artificial Intelligence (CISAI), Kunming, China*, pages 153–156, 2021.