**Development of a model for the detection and classification of tomato plant diseases using Swin Transformer**

Oscar David Valencia - 20192182074

Maria José Capera Firigua – 20201188262

Facultad de Ingeniería, Universidad Surcolombiana

Electiva Ciencias de la computación II - Visión Artificial

Ferley Medina Rojas

Agosto de 2023

**TABLE OF CONTENTS**

[**INTRODUCTION 3**](#_tycrd2ytbpn5)

[**PROBLEM 4**](#_76696e9e9q2l)

[**OBJECTIVES 5**](#_5wmsknvme73h)

[General objective 5](#_f6dlx77kymqv)

[Specific objectives 5](#_d6r42owmwevb)

[**LITERATURE REVIEW 6**](#_r2lsqc1cu1it)

[**THEORETICAL FRAMEWORK 17**](#_osxtvooujh3s)

[**REFERENCES 22**](#_3irfze7s0g8d)

# 

# INTRODUCTION

Early and accurate detection of tomato diseases has become an imperative to mitigate their devastating impact. Although advances in computer vision and machine learning have provided potential solutions, these methods have been insufficient in terms of speed, cost, and accuracy. Promising Deep Learning techniques, particularly convolutional neural networks (CNNs), have not re-proposed new architectures or even a paradigm shift, calling for innovative rethinking.

This study aims to explore and validate an innovative approach by applying the "Swin Transformer", a novel architecture in the computer vision domain. Through a systematic analysis and rigorous methodology, it will seek to develop a model that overcomes existing limitations in tomato disease detection and classification.

It will focus on answering critical questions related to the development and validation of this new model, as well as its comparison with CNN-based methods. The research will address essential aspects such as data augmentation, completeness of results, optimal configuration and training of the Swin Transformer architecture, and visual interpretation of model decisions.

The proposed research has significant potential to transform current practice in tomato disease detection and classification by providing a more accurate and efficient model.

The paper is structured in introduction, problem, objectives, state of the art, theoretical framework, methodology, materials and methods, results, conclusions and future work.

# PROBLEM

Tomato cultivation is essential for food and the global agricultural economy because it is one of the most widely produced vegetables worldwide. According to FAO (2021), world production exceeded 189 million tons, covering more than 5 million hectares under cultivation. Nutritionally, it is a rich source of phytochemicals, nutrients, antioxidants and phenolic compounds (Collins et al., 2022). Its consumption has been found to provide anti-cancer benefits and protection against cardiovascular and neurodegenerative diseases (Collins et al., 2022).

However, tomato is vulnerable to a wide variety of diseases caused by fungi, bacteria, phytoplasmas, viruses and viroids (Panno et al., 2021) Table 1. This vulnerability is increased by factors such as its limited genetic diversity resulting from intensive selection, the tendency to monoculture, international exchange of infected material, and climate change (Panno et al., 2021). These diseases represent economic losses exceeding US$30 billion per year (Caruso et al., 2022). Such losses impact agricultural economics, public health, and environmental sustainability by affecting crop yield and quality, destabilizing ecosystems, and raising production costs due to the intensive use of pesticides and other control measures (Panno et al., 2021).

Early and accurate detection of diseases in tomato is essential to mitigate their impact, but despite significant advances in computer vision over the last decade, there have been limitations. In machine learning, methods often rely on manual feature extraction and heuristic algorithms, which can be slow, laborious, costly and error prone (Zahangir Alom et al., 2018). These approaches have been overtaken by Deep Learning techniques such as convolutional neural networks (CNNs) (Rawat et al. 2022; Thangaraj et al. 2022). However, the latter seem to have reached a stagnation point following the development of architectures such as ResNet, DenseNet and EfficientNet (Li, 2020).

Therefore, the present research suggests an approach based on the "Swin Transformer" architecture, which has proven to be highly effective in computer vision tasks but has never been applied to tomato disease classification.

In this context, the following questions are raised:

* How to develop and validate a tomato disease detection and classification model using Swin Transformer architecture that surpasses the accuracy of convolutional neural networks (CNN)?
* How to build balanced dataset across the different tomato disease classes to ensure the integrity of the model predictions?
* How to optimally configure and train the Swin Transformer architecture?
* How to evaluate and interpret the Swin Transformer model predictions for each class?
* How does the accuracy of the Swin Transformer model compare to models based on convolutional neural networks?

# OBJECTIVES

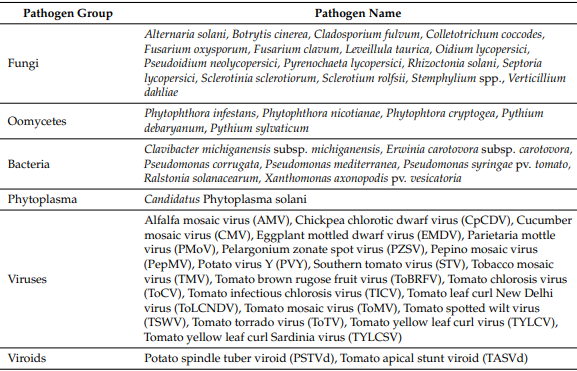
## General Objective

To develop and validate a tomato plant disease detection and classification model using the Swin Transformer architecture, with the purpose of achieving superior accuracy compared to convolutional neural networks.

## Specific Objectives

1. Build a balanced dataset on the different tomato disease classes to ensure the integrity of the model predictions.
2. Optimally configure and train the Swin Transformer architecture.
3. Evaluate and interpret the predictions of the Swin Transformer model for each class.
4. Compare and analyze the accuracy of the Swin Transformer model with models based on convolutional neural networks.

**Table 1.** List of tomato plant pathogens (Panno et al., 2021).



**LITERATURE REVIEW**

In the paper ***"Machine Learning Approach towards Tomato Leaf Disease Classification"***, Gadade and Kirange, (2020) present a machine learning approach for the identification and classification of tomato leaf diseases. The research process started with the collection of 9,000 tomato leaf images from the PlantVillage dataset, focusing on seven classes of diseased plants and one class of healthy plant.

In the preprocessing phase, image noise was minimized using a median filter, then feature extraction was performed using GLCM (Gray-Level Co-Occurrence Matrix), Gabor and SURF (Speeded Up Robust Features) techniques. To classify the images into normal or diseased classes, several classification algorithms such as support vector machine (SVM), k-Nearest Neighbors (KNN), Naïve Bayes (NB) and decision trees were employed.

The results showed that the combination of SVM and Gabor obtained superior metrics to all others with an 'accuracy' of 0.7339, 'precision' of 0.2525, 'recall' of 0.9492 and 'f1-score' of 0.3989. Despite these results, the authors indicate that more advanced techniques, such as Adaptive Neuro Fuzzy, Neural Networks and genetic algorithms, could improve the classification.

Nevertheless, this work evidence that traditional machine learning methods show inferior metrics compared to modern deep learning techniques based on convolutional neural networks. This observation points to a possible obsolescence of traditional methods in the field of tomato leaf disease classification through computer vision techniques.

A study entitled ***"ToLeD: Tomato Leaf Disease Detection using Convolution Neural Network"*** by Mohit Agarwala et al. (2020) proposes a Convolutional Neural Network (CNN) based approach for tomato leaf disease detection and classification. The study uses the Plant Village dataset containing images of tomato leaves affected by nine types of diseases in addition to a class of healthy leaves. To balance the dataset, data augmentation techniques were applied resulting in 10,000 images for training, 7,000 for validation and 500 for testing.

The CNN architecture developed in the study consists of 3 convolutional layers, 3 max-pooling layers and 2 fully connected layers. Pre-trained CNN models were also evaluated through transfer learning, using the same dataset, in order to test the effectiveness of the proposed model. The resulting metrics showed accuracy ranging from 76% to 100% for various classes, with an average accuracy of 91.2%. This performance outperforms pre-trained models such as VGG16, with an accuracy of 77.2%, InceptionV3, with 63.4%, and MobileNet, with 63.7%.

The proposed model has a smaller number of trainable parameters compared to that of the pre-trained models, suggesting potential advantages in storage and computation. However, variability in accuracy between classes and comparison with more current work indicate that the model, while effective, is outperformed by more recent approaches.

The paper ***"Tomato plant disease detection using transfer learning with C-GAN synthetic images"*** by Abbas et al. (2021) addresses a deep learning approach for tomato plant disease detection and classification. The methodology implements two main phases, synthetic image generation using Conditional Generative Antagonistic Networks (C-GAN) to extend the dataset and disease classification using a previously trained DenseNet121 model.

The performance of the method was evaluated with the PlantVillage dataset, which includes nine classes of tomato leaves with different diseases in addition to a class of healthy leaves. The synthetic images generated with C-GAN were added to the existing dataset. Then, the DenseNet121 model was trained with the extended dataset.

Two sets of experiments were performed, the first focused on C-GAN training and synthetic image generation, the second on training DenseNet121 model using the extended dataset. A comparative analysis with pre-trained models such as VGG19, ResNet50, Inception-V3, Xception and MobileNet showed that the proposed model achieved the highest level of accuracy, with 97.11% in classifying the 10 classes of the study.

However, although the proposed model demonstrated high efficiency in the classification of tomato plant diseases and the application of C-GAN helped to avoid overfitting, it is important to evaluate the fidelity of the generated synthetic images, as they might not have the level of detail presented by the real images and, therefore, negatively affect the model performance in generalization tasks.

The study ***"Early Detection and Classification of Tomato Leaf Disease Using High-Performance Deep Neural Network"*** by Trivedi et al. (2021) addresses the identification and classification of tomato leaf diseases using Convolutional Neural Networks (CNN). A dataset composed of 3000 images, distributed into nine classes of affected leaves and one class of healthy leaves, was used. In the preprocessing phase, the images were normalized and resized.

The process comprises preprocessing, segmentation, and classification stages through CNN. The specified CNN architecture consists of eight convolutional layers, eight max pooling layers, three fully connected layers, with a dropout rate of 0.5 and Relu activation function. Experiments were conducted by varying parameters such as learning rate and number of epochs.

The results of the accuracy metric were 98.49%. Further experiments showed that the accuracy varies slightly with different learning rates and epochs, generally ranging around 98.4% to 98.58%. Comparison with standard models such as MobileNet, VGG16 and InceptionV3 indicated that the proposed model outperforms these CNN models in accuracy.

The study highlights the effectiveness of the proposed model, especially in comparison with other models. However, it omits discussions on limitations, future applicability or the rationale for the choice of the specific architecture employed.

The paper entitled ***"Tomato leaf disease classification using supervised learning techniques: contrasting analysis"***, presented by Rawat et al., (2022), offers a contribution to the field of tomato leaf disease classification using supervised learning techniques.

The study is built around four main components: selection of the tomato image dataset from the "Plant Village" archive, the preprocessing of these images for denoising using a median filter, the application of various machine learning algorithms such as SVM, K-Nearest Neighbour, Naïve Bayes, Decision Tree, Feed Forward Neural Network, Back Propagation Neural Network, Deep Neural Network, Convolutional Neural Network and Multi Kernel SVM, and finally the evaluation of the effectiveness of these algorithms using metrics such as Accuracy, F1-Score and Recall.

The study performs a comparative analysis of the effectiveness of the different algorithms. In this analysis, the Convolutional Neural Network (CNN) achieves an accuracy of 98.5%, surpassing other algorithms such as SVM, which achieves 90% accuracy, and KNN with 83.6%. The research concludes that CNNs offer superior performance in tomato leaf disease classification compared to traditional machine learning methods.

The study entitled ***"AlexNet Convolutional Neural Network for Disease Detection and Classification of Tomato Leaf"*** (Chen et al., 2022) addresses tomato leaf disease detection and classification using a convolutional neural network (CNN) based on the AlexNet architecture. The authors used a dataset of 18,345 training images and 4,585 test images, distributed into nine classes corresponding to various tomato leaf diseases and one healthy leaf class.

The modification made to the AlexNet network architecture involves three convolution layers, three fully connected layers and one output layer. The Adam optimizer was used with a learning rate of 0.0005, 75 epochs and a batch size of 128. The loss function was cross-entropy. The model showed an accuracy of 96%, a precision rate of 98%, a recall value of 95% and an F1-score of 97%.

The conclusions highlight the effectiveness of the model for tomato leaf disease classification. It is noted that the model is light enough to be implemented on mobile devices, which is relevant given the memory capacity limitations of these devices.

The study ***"BotanicX-AI: Identification of Tomato Leaf Diseases Using an Explanation-Driven Deep-Learning Model"*** by Bhandari et al., (2023), addresses the classification of nine classes of tomato leaf diseases and one class for healthy leaves using a deep learning model based on EfficientNetB5. The dataset is built by preprocessing images through resizing and data augmentation resulting in 11,000 tomato leaf images distributed in 10 classes.

The methodology consists of applying transfer learning using an EfficientNetB5 architecture to which additional layers such as batch normalization, dense layers, and regularization techniques such as dropout were added. To increase the interpretability of the model, techniques such as GradCAM and LIME are integrated, which help to identify the regions of the image that contribute most to the classification.

In terms of metrics the proposed model achieved an accuracy of 99.07%, outperforming other pre-trained models, such as MobileNet (94.00%), Xception (95.32%), VGG16 (93.35%), ResNet50 (96.03%) and DenseNet121 (96.30%). GradCAM and LIME techniques were able to identify the most important regions in the leaf images for classification. Although it was noted that GradCAM had limitations in identifying certain regions in some disease classes.

Finally, the authors recommend consideration of other interpretability techniques, such as HiResCAM and SHAP, for future work, arguing that these may increase confidence in the applicability of the model.

The reviewed studies address the problem of tomato leaf disease detection and classification using various techniques ranging from traditional machine learning to deep learning. Machine learning algorithms, such as SVM and KNN, show lower performance metrics than deep learning based models. Among deep learning models, Convolutional Neural Networks (CNNs) and their variations, such as DenseNet121 and EfficientNetB5, show superior metrics in terms of precision, accuracy and F1-score.

Methods incorporating synthetic images generated using Conditional Generative Antagonistic Networks (C-GAN) and transfer learning techniques seem to offer advantages in terms of performance and overfitting prevention. However, the quality of the generated synthetic images is a variable to consider for generalization tasks.

Applicability on mobile devices and model interpretability are aspects that are highlighted in some works, evidencing the importance of these factors in practical applications. Explainability techniques such as GradCAM and LIME offer possibilities for understanding model decisions, although it is suggested that additional methods can improve model confidence.

Deep learning-based methods show superior performance and stand out as the most effective method in tomato plant disease classification. However, they have not presented innovations in their architecture recently. Furthermore, the application of transformer vision architectures for this specific problem has not been investigated, which represents an unexplored area that could offer solutions to limitations such as interpretability or improved accuracy and therefore warrants further research.

# THEORETICAL FRAMEWORK

**Tomato Diseases:** To address this challenge, it is critical to understand tomato diseases and their effects on production. These diseases can negatively affect tomato production, reducing yield and crop quality. In addition, they are spread in a variety of ways, either by direct contact, through vectors such as insects, or even by adverse weather conditions. The long-term economic impact of these diseases is significant and can influence the economic stability and prices of tomato-related products in the global market.

**Plant Pathology and Plant Epidemiology:** Plant pathology focuses on the study of plant diseases, and plant epidemiology investigates how these diseases spread in plant populations. These fields are fundamental to understanding the nature and spread of diseases affecting crops such as tomatoes.

**Food Supply:** Tomatoes and other vegetable crops are fundamental sources of food in the human diet. They provide essential nutrients, vitamins, and minerals. Any threat to the production of these crops can have a direct impact on the food supply.

**Global Economy:** The production and marketing of crops such as tomatoes represent a significant part of the global agricultural economy. Food security is linked to economic stability, as food shortages or price fluctuations can trigger economic and political crises.

**Climate Variability:** Climate change is causing extreme weather events, such as droughts and floods, which can damage crops. Early detection of tomato diseases, which can be aggravated by adverse weather conditions, is critical for climate risk management and food security.

**Plant Genetic Diversity and Plant Disease Resistance**: Genetic diversity within plants plays an important role in their ability to resist disease. Some tomato varieties may be naturally more resistant than others, highlighting the importance of genetic diversity in food security.

**Disease Losses:** Tomato diseases can cause significant yield losses if not detected and controlled in a timely manner. These losses not only affect food availability but can also increase food product prices.

**Food Quality:** Food safety is not only about the quantity of food available, but also about quality. Diseases can affect the quality of tomatoes, which in turn can affect the health of the people who eat them.

**Pesticide Use:** Early and accurate detection of tomato diseases allows for more efficient and targeted use of pesticides. This is important for reducing the amount of chemicals used in agriculture, which can have benefits for human health and the environment.

**Machine Vision and Machine Learning** are essential tools in tomato crop disease detection. Computer vision allows computers to process and understand images, while machine learning enables machines to learn patterns and make decisions based on training data.

**Convolutional Neural Networks (CNNs):** Within computer vision, convolutional neural networks (CNNs) have been widely used for disease detection in plants, including tomatoes. These networks use convolutional layers to detect patterns and features in images. In addition, transfer learning has been successfully applied using pre-trained CNN models on large data sets.

**Swin Transformer Architecture**: Swin Transformer, which stands for "Shifted Window Transformer," introduces an innovative architecture that addresses the limitations of traditional ViTs and CNNs. It starts by dividing the image into patches and uses sliding window attention instead of global attention, which significantly reduces computational complexity. In addition, it implements a pyramid structure to process images at different resolutions, allowing features to be captured at different scales.

**Transfer Learning and Neural Network Architecture**: Transferring knowledge from pre-trained models into neural networks is a powerful technique that can accelerate training and improve performance in disease detection.

**PlantVillage Dataset:** To train tomato disease detection models, a high-quality dataset is essential. The PlantVillage dataset provides a variety of tomato disease classes and reflects the complexity of detection in the field. The quality of the labeled data is critical to the success of model training.

**Classes of the PlantVillage Dataset:**

Tomato Yellow Leaf Curl Virus (YLCV)

Tomato Bacterial Spot (Bctsp)

Tomato Late Blight (TLB)

Tomato Septoria leaf spot (SptL)

Tomato Two Spotted Spider Mite (SpdM)

Tomato Target Spot (TISS)

Tomato Early Blight (TEB)

Tomato Leaf Mold (LMld)

Tomato Mosaic Virus (MscV)

Tomato healthy (Hlth)

**Image Resolution and Data Augmentation:** In tomato disease detection using computer vision, image resolution and data augmentation techniques can improve accuracy by providing high quality images and increasing the amount of training data.

**Attention Layers:** In the Swin Transformer model, attention layers are employed to prioritize certain image features during the analysis and classification process. These layers make it possible to assign different weights to different regions of the image, thus optimizing the model's ability to perform specific tasks.

**Interpretability in Machine Vision Models:** Interpretability in machine vision models is crucial to assess the quality of model decisions in tomato disease classification. This interpretability allows understanding how each decision is arrived at and can guide corrective actions.

**Supervised Learning:** is an approach in machine learning where an algorithm is trained using a dataset containing output labels. The goal is to learn a function that maps inputs to outputs. This method is commonly applied in tasks such as classification and regression.

**Image Classification:** This task belongs to the field of computer vision. It consists of assigning a label to an input image selected from a set of predefined categories. The process generally involves the use of deep learning models such as convolutional neural networks.

**Confusion Matrix:** A table that presents the performance of a classification algorithm. The rows and columns represent the actual classes and the predicted classes, respectively. The elements of the matrix are:

- **True Positives (TP):** correctly identified positive cases.

- **True Negatives (TN):** Correctly identified negative cases.

- **False Positives (FP):** Negative cases incorrectly identified as positive.

- **False Negatives (FN):** Positive cases incorrectly identified as negative.

**Evaluation Metrics:** Parameters that quantify the quality of a model in specific tasks. They are defined mathematically as:

* 
* 
* 
* 

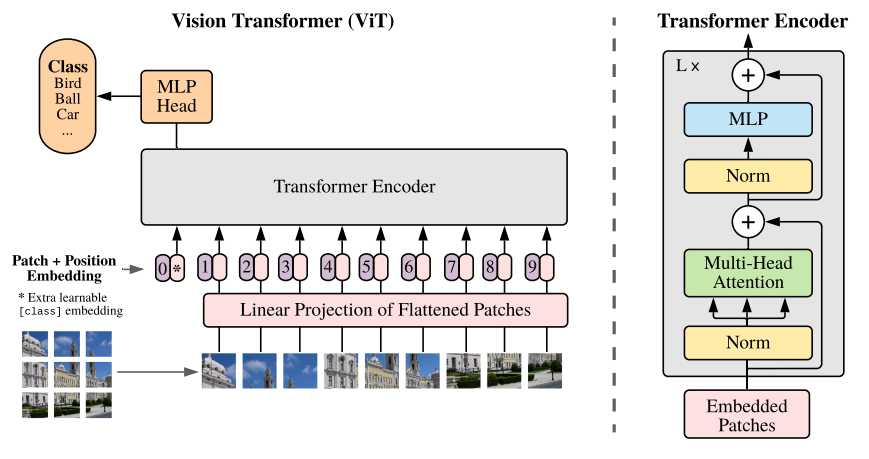
**Overfitting and Underfitting:** Two common problems in machine learning. Overfitting occurs when the model captures noise instead of the underlying pattern of the training set, resulting in poor performance on unseen data. Underfitting occurs when the model does not sufficiently capture the underlying trend of the data set.

**Cross Validation:** This is a method that involves splitting the dataset into multiple training and testing subsets. The model is trained and evaluated multiple times, changing the subsets used. Model performance is estimated by averaging the performance metrics for each iteration.

**Hyperparameters:** Parameters that are not learned from the training set but are preconfigured for the learning process. They include learning rate, number of epochs, and batch size in algorithms such as neural networks.

**Introduction to the Vision Transformers (ViTs)**

Before getting into the Swin Transformer, it is essential to understand the concept of Vision Transformers (ViTs). These models adapt the transformer architecture, originally designed for natural language processing, for computer vision tasks. The key idea is to divide an image into patches, flatten them and feed them to a series of transformer layers, allowing the model to capture long-range relationships between different parts of the image (Alexey Dosovitskiy et al., 2020).



**Figure 1:** Model overview of the first visual transformer (Alexey Dosovitskiy et al., 2020)

**Limitations of Traditional ViTs**

Although ViTs have proven to be effective, they have certain limitations. Primarily, they require enormous amounts of data and computational power to train from scratch. In addition, the global attention used in traditional ViTs may not be optimal in terms of computational efficiency, especially for high-resolution images (Alexey Dosovitskiy et al., 2020; Touvron et al., 2021).

**General Architecture of the Swin Transformer: A Paradigm Shift**

The Swin Transformer, which stands for "Shifted Window Transformer," addresses the limitations of ViTs by introducing two key concepts: sliding windows and pyramid structure.

The Swin Transformer, like other Vision Transformers (ViTs), starts by dividing an image into non-overlapping patches using a splitting module. Each patch is considered a "token" and its feature is set by concatenating the RGB values of the raw pixels. In the Swin Transformer implementation, a patch size of 4x4 is used, resulting in a feature dimension of 48 (4 × 4 × 3). These features are subsequently projected to an arbitrary dimension (denoted as *C*) by means of a linear embedding layer.

The term "Stage" refers to the different stages of image processing through the architecture. In the Swin Transformer, the image resolution is successively reduced, from "Stage 1" to "Stage 4". This hierarchical structure is similar to traditional convolutional neural networks such as VGG and ResNet, where features are processed at different resolutions.

**Some detailed explanations:**

**MSA (Multi-Head Self-Attention):** a type of attention mechanism that allows the model to pay attention to different parts of the input simultaneously. It is essential to the transformer's ability to capture long-range relationships.

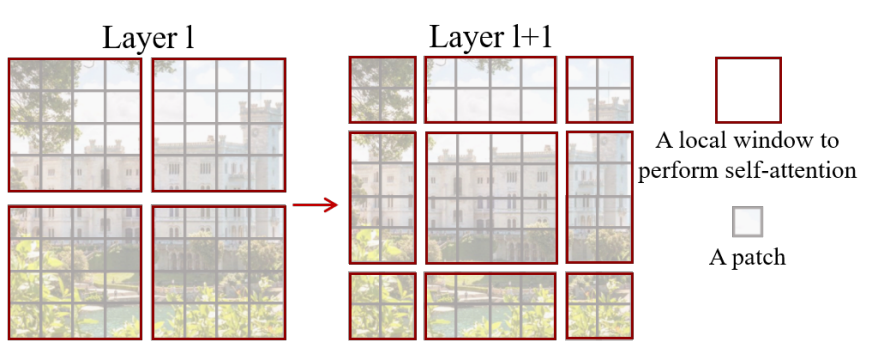
**MLP (Multi-Layer Perceptron):** Refers to a fully connected neural network. In the context of transformers, MLPs are used to transform features after layers of attention.

**GELU (Gaussian Error Linear Unit):** An activation function used in neural networks. It has been shown to have beneficial properties in transformer models.

**LN (LayerNorm or Layer Normalization):** It is a normalization technique used to stabilize and accelerate the training of neural networks.

**Ω(MSA) and Ω(W-MSA):** These are notations to represent the computational complexity of global attention and window-based attention, respectively.

**- Sliding Windows:** Instead of performing global attention over the entire image (which would have a quadratic complexity with respect to the number of tokens), the Swin Transformer introduces the concept of attention within non-overlapping local windows. This significantly reduces the computational complexity (Liu et al., 2021).



**Figure 2.** The shifted window approach to computing self-attention in the proposed Swin Transformer architecture.

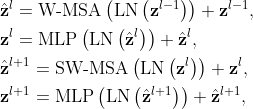
Equations (1) and (2) show the difference in complexity between global care and window-based care. Window-based attention is scalable and computationally more efficient (Liu et al., 2021).

(1) 

(2)

However, in order not to lose the ability to model relationships between patches in different windows, the Swin Transformer introduces the idea of windows shifted in successive blocks. This strategy alternates between two partitioning configurations in consecutive blocks of transformers, allowing each patch to relate to patches in neighboring windows.

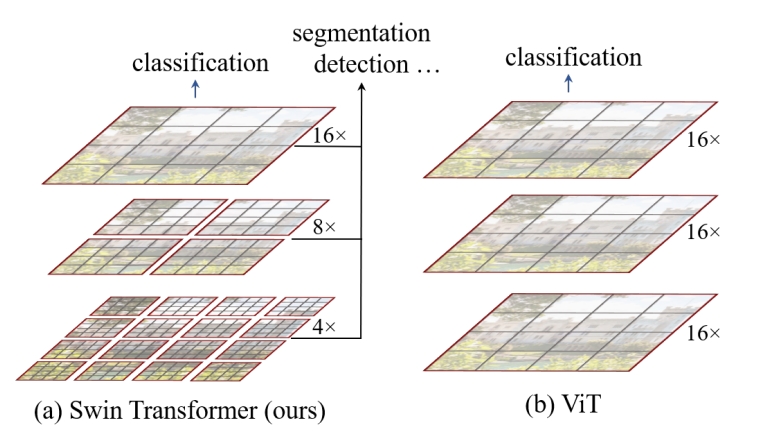
Equations (3) show how features are computed in successive blocks using different window-based attention configurations (Liu et al., 2021).

(3) 

In this context,  and  represent the resulting characteristics from the (S)W-MSA module and the MLP module corresponding to block . W-MSA and SW-MSA refer to window-based multi-head self-attention utilizing standard and shifted window partitioning setups, respectively.

Relative position bias is introduced to account for spatial relationships between patches within a window. This bias significantly improves model performance and is preferable to absolute position embeddings.

**- Pyramid Structure:** The Swin Transformer processes images in a series of resolutions, similar to a pyramid. It starts with small windows at a higher resolution and gradually clusters patches while reducing the resolution, allowing the model to capture features at different scales (Liu et al., 2021).



**Figure 3.** Comparison of the pyramidal structure method used by Swin tranformer and the method used by Vit.

**Architectural Details**

**- Tokenization and Embeddings:** Like other ViTs, the Swin Transformer starts by dividing the image into patches, which are then flattened and converted into embeddings through a linear layer (Liu et al., 2021).

**- Attention within Windows:** The attention layer in Swin Transformer operates only within sliding windows, using standard attention mechanisms. After each attention layer, the windows are scrolled to ensure coverage of the entire image (Liu et al., 2021).

**-Swin Transformer Block:** A Swin Transformer block is a fundamental unit in the architecture. It is where the shifted window-based attention occurs, which is essential for the efficiency and effectiveness of the model.

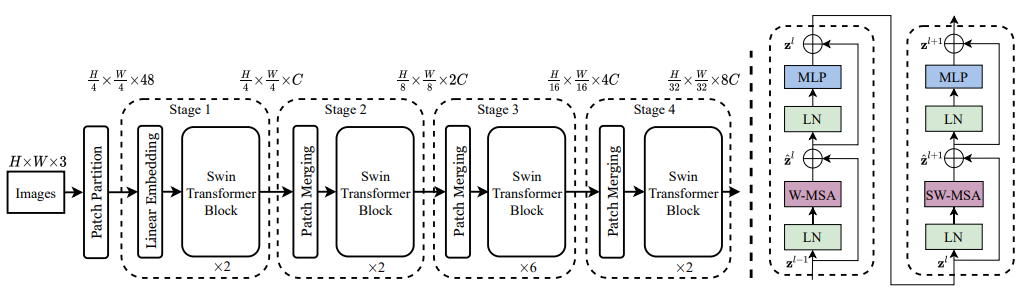
**-GELU:** GELU, or Gaussian Error Linear Unit, is a nonlinear activation function. It is used in neural networks to introduce nonlinearities into the model. In the Swin Transformer block, the GELU is used as an activation function in the 2-layer multilayer perceptron (MLP).

**-LN (LayerNorm):** LN, or Layer Normalization, is a normalization technique that is applied to features in a specific layer, rather than to a specific mini-batch. In Swin Transformer, LayerNorm is applied before each auto-tuned multi-head attention module (MSA) and each MLP. Normalization helps stabilize and accelerate training.

**-Residual Connection:** In deep networks, residual connections help avoid the gradient fading problem by allowing activations to skip one or more layers. In the Swin Transformer, a residual connection is applied after each attention module and MLP.

**- Resolution Reduction and Upscaling:** After processing the image at one resolution, the Swin Transformer reduces the resolution by grouping tokens and processing them in larger windows. This process is repeated several times, forming the pyramidal structure of the model (Liu et al., 2021).

**- Patch Merging:** "Patch Merging" is a technique used in the Swin Transformer to reduce the spatial resolution of the feature map and simultaneously increase the feature dimension, allowing the model to capture more abstract representations as it goes deeper. It works by aggregating information from neighboring patches and creating a new patch with a higher feature dimension.



**Figure 4.** On the left is the architecture of a Swin transformer (Swin-T) and on the right are two successive blocks of the Swin transformer W-MSA and SW-MSA which are multihead self-attention modules with regular and offset window configurations, respectively.

**Advantages of the Swin Transformer**

**- Computational Efficiency:** By avoiding global focus on the entire image, the Swin Transformer is significantly more efficient than traditional ViTs, especially for high-resolution images.

**- Modeling Capability:** The pyramidal structure allows the Swin Transformer to capture features at different scales, which can be crucial for tasks such as plant disease detection where patterns can vary in size.

**- Benchmarking:** Tests have shown that Swin Transformer performs very well and better on various data sets, outperforming other ViTs and CNNs in several benchmarks.

# REFERENCES

Collins, E.J.; Bowyer, C.; Tsouza, A.; Chopra, M. Tomatoes: An Extensive Review of the Associated Health Impacts of Tomatoes and Factors That Can Affect Their Cultivation. Biology 2022, 11, 239.

Caruso, A. G., Bertacca, S., Parrella, G., Rizzo, R., Davino, S., & Panno, S. (2022). Tomato brown rugose fruit virus: A pathogen that is changing the tomato production worldwide. Annals of Applied Biology, 181(3), 258–274.

Panno, S., Davino, S., Caruso, A. G., Bertacca, S., Crnogorac, A., Mandić, A., Noris, E., & Matić, S. (2021). A Review of the Most Common and Economically Important Diseases That Undermine the Cultivation of Tomato Crop in the Mediterranean Basin. Agronomy, 11(11), 2188. https://doi.org/10.3390/agronomy11112188

Singh, V. K., Singh, A. K., & Kumar, A. (2017). Disease management of tomato through PGPB: current trends and future perspective. 3 Biotech, 7(4). https://doi.org/10.1007/s13205-017-0896-1

Zahangir Alom, Taha, T. M., Yakopcic, C. G., Westberg, S., Sidike, P., Nasrin, Mst Shamima, Esesn, V., Abdul, & Asari, V. K. (2018). The History Began from AlexNet: A Comprehensive Survey on Deep Learning Approaches. ArXiv E-Prints, 2. https://doi.org/10.48550/arxiv.1803.01164

Boulent, J., Foucher, S., Théau, J., & St-Charles, P.-L. (2019). Convolutional Neural Networks for the Automatic Identification of Plant Diseases. Frontiers in Plant Science, 10. https://doi.org/10.3389/fpls.2019.00941

Li, E. Y. (2020, August 1). 10 Papers You Should Read to Understand Image Classification in the Deep Learning Era. Medium. https://towardsdatascience.com/10-papers-you-should-read-to-understand-image-classification-in-the-deep-learning-era-4b9d792f45a7

Rawat, V., Singh, N., Kaur, B., & Bora, S. (2022). Tomato Leaf Disease Classification Using Supervised Learning Techniques: Contrasting Analysis. 2022 International Conference on Advances in Computing, Communication and Materials (ICACCM), 1(2642-7354). https://doi.org/10.1109/icaccm56405.2022.10009617

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2. https://doi.org/10.1109/iccv48922.2021.00986

Rawat, Vandana & Singh, Neelam & Kaur, Bhavleen & Bora, Saksham. (2022). Tomato Leaf Disease Classification Using Supervised Learning Techniques: Contrasting Analysis. 1-7. 10.1109/ICACCM56405.2022.10009617.

Z. Liu et al., "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021, pp. 9992-10002, doi: 10.1109/ICCV48922.2021.00986.

Alom, M. Z., Taha, T. M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M. S., ... & Asari, V. K. (2018). The history began from alexnet: A comprehensive survey on deep learning approaches.

Boulent, J., Foucher, S., Théau, J., & St-Charles, P. L. (2019). Convolutional neural networks for the automatic identification of plant diseases. Frontiers in plant science, 10, 941.

Gadade, H. D., & Kirange, D. D. (2020). Machine learning approach towards tomato leaf disease classification. International Journal of Advanced Trends in Computer Science and Engineering, 9(1), 490-495.

Abbas, A., Jain, S., Gour, M., & Vankudothu, S. (2021). Tomato plant disease detection using transfer learning with C-GAN synthetic images. Computers and Electronics in Agriculture, 187, 106279.

Rawat, V., Singh, N., Kaur, B., & Bora, S. (2022, November). Tomato Leaf Disease Classification Using Supervised Learning Techniques: Contrasting Analysis. In 2022 International Conference on Advances in Computing, Communication and Materials (ICACCM) (pp. 1-7). IEEE.

Thangaraj, R., Anandamurugan, S., Pandiyan, P., & Kaliappan, V. K. (2022). Artificial intelligence in tomato leaf disease detection: a comprehensive review and discussion. Journal of Plant Diseases and Protection, 129(3), 469-488.

Uppada, R., & Kumar, D. R. (2023). Computer-aided fusion-based neural network in application to categorize tomato plants.Springer-Verlag London Ltd., part of Springer Nature 2023, 1-9.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021, July). Training data-efficient image transformers & distillation through attention. In International conference on machine learning (pp. 10347-10357). PMLR.