**Desarrollo de una aplicación móvil para la detección y clasificación de enfermedades en la planta del Tomate utilizando 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

**TABLA DE CONTENIDOS**

[**INTRODUCCIÓN 3**](#_tycrd2ytbpn5)

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

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

[Objetivo general 5](#_f6dlx77kymqv)

[Objetivos especificos 5](#_d6r42owmwevb)

[**REVISIÓN DE LITERATURA 6**](#_r2lsqc1cu1it)

[Machine Learning Approach towards Tomato Leaf Disease Classification. International Journal of Advanced Trends in Computer Science and Engineering, 9(1). 6](#_xa2sjoqh2ody)

[Tomato plant disease detection using transfer learning with C-GAN synthetic images 7](#_nhdykxx5j871)

[Tomato leaf disease classification using supervised learning techniques contrasting. 9](#_93hzfrq8tjbd)

[Artificial intelligence in tomato leaf disease detection 11](#_1pm0ksx7pk1k)

[Computer-aided fusion-based neural network in application to categorize tomato plants. Springer-Verlag London Ltd., part of Springer Nature 2023 14](#_obxgds67bnog)

[**MARCO TEÓRICO 17**](#_osxtvooujh3s)

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

# INTRODUCCIÓN

La detección temprana y precisa de enfermedades en el tomate se ha convertido en una necesidad imperante para mitigar su impacto devastador. Aunque los avances en la visión artificial y el aprendizaje automático han proporcionado soluciones potenciales, estos métodos han sido insuficientes en términos de velocidad, costo, y precisión. Las prometedoras técnicas de Deep Learning, particularmente las redes neuronales convolucionales (CNN), no han vuelto a plantear nuevas arquitecturas ni tampoco un cambio de paradigma, lo que exige un replanteamiento innovador.

Este estudio se propone explorar y validar un enfoque innovador mediante la aplicación del "Swin Transformer", una arquitectura novedosa en el dominio de la visión por computadora. A través de un análisis sistemático y una metodología rigurosa, se buscará desarrollar un modelo que supere las limitaciones existentes en la detección y clasificación de enfermedades del tomate.

Se enfocará en responder preguntas críticas relacionadas con el desarrollo y validación de este nuevo modelo, así como en su comparación con los métodos basados en CNN. La investigación abordará aspectos esenciales como el aumento de datos, la integridad de los resultados, la configuración y entrenamiento óptimo de la arquitectura Swin Transformer, y la interpretación visual de las decisiones del modelo.

La investigación propuesta tiene un potencial significativo para transformar la práctica actual en la detección y clasificación de enfermedades del tomate, ofreciendo un modelo más preciso y eficiente.

El trabajo se estructura en introducción, problema, objetivos, estado del arte, marco teórico, metodología, materiales y métodos, resultados, conclusiones y trabajo futuro.

# PROBLEMA

El cultivo del tomate es esencial para la alimentación y la economía agrícola global debido a que es una de las hortalizas más producidas mundialmente. De acuerdo con la FAO (2021), la producción mundial superó los 189 millones de toneladas, abarcando más de 5 millones de hectáreas cultivadas. Nutricionalmente es una fuente rica de fitoquímicos, nutrientes, antioxidantes y compuestos fenólicos (Collins et al., 2022). Se ha determinado que su consumo aporta beneficios anticancerígenos y protección contra enfermedades cardiovasculares y neurodegenerativas (Collins et al., 2022).

No obstante, el tomate es vulnerable a una amplia variedad de enfermedades causadas por hongos, bacterias, fitoplasmas, virus y viroides (Panno et al., 2021) Tabla 1. Esta vulnerabilidad se ve incrementada por factores como su limitada diversidad genética resultado de la selección intensiva, la tendencia al monocultivo, el intercambio de material infectado a nivel internacional y el cambio climático (Panno et al., 2021). Estas enfermedades representan pérdidas económicas que superan los 30.000 millones de dólares al año (Caruso et al., 2022). Dichas pérdidas repercuten en la economía agrícola, la salud pública y la sostenibilidad ambiental, al afectar el rendimiento y la calidad del cultivo, desestabilizar ecosistemas y elevar los costos de producción por el uso intensivo de pesticidas y otras medidas de control (Panno et al., 2021).

La detección temprana y precisa de enfermedades en el tomate es primordial para mitigar su impacto, pero a pesar de los avances significativos en visión artificial en la última década, estos han presentado limitaciones. En el ámbito del machine learning, los métodos suelen recurrir a la extracción manual de características y algoritmos heurísticos, lo que puede ser lento, laborioso, costoso y susceptible a errores (Zahangir Alom et al., 2018). Estos enfoques han sido superados por técnicas de Deep Learning como las redes neuronales convolucionales (CNN) (Rawat et al. 2022; Thangaraj et al. 2022). Sin embargo, estas últimas parecen haber llegado a un punto de estancamiento tras el desarrollo de arquitecturas como ResNet, DenseNet y EfficientNet (Li, 2020).

Es por ello, que la presente investigación sugiere un enfoque innovador basado la arquitectura "Swin Transformer", el cual ha demostrado ser altamente efectivo en tareas de visión por computador pero que nunca ha sido aplicado en la clasificación de enfermedades del tomate. El objetivo es desarrollar y validar un modelo más preciso que las soluciones anteriores utilizando el dataset de PlantVillage, el cual contiene nueve clases de enfermedades del tomate y una clase de la planta sana. Posteriormente, se comparará su rendimiento con estudios previos basados en CNN.

En este contexto, se plantea las siguientes preguntas:

* ¿Cómo desarrollar y validar un modelo de detección y clasificación de enfermedades del tomate utilizando la arquitectura Swin Transformer que supere la precisión de las redes neuronales convolucionales (CNN)?
* ¿Cómo se puede construir y preparar un dataset balanceado en las diferentes clases de enfermedades del tomate que garantice la integridad de las predicciones del modelo?
* ¿Cómo configurar y entrenar de manera óptima la arquitectura Swin Transformer?
* ¿Cómo evaluar e interpretar las predicciones del modelo Swin Transformer para cada clase?
* ¿Cómo se compara la exactitud (accuracy) del modelo Swin Transformer con los modelos basados en redes neuronales convolucionales?

# OBJETIVOS

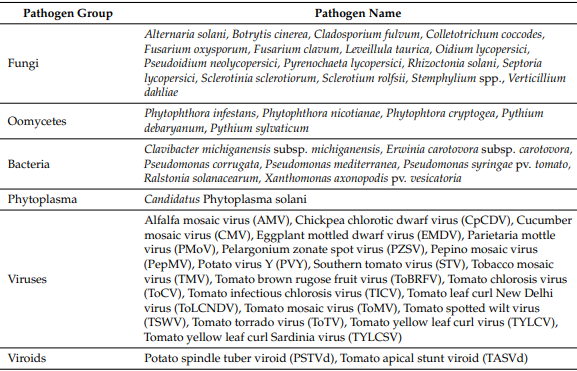
## Objetivo general

Desarrollar y validar un modelo de detección y clasificación de enfermedades de plantas de tomate utilizando la arquitectura Swin Transformer, con el propósito de conseguir una precisión superior en comparación con las redes neuronales convolucionales.

## Objetivos específicos

1. Construir y preparar un dataset balanceado en las diferentes clases de enfermedades del tomate que asegure la integridad de las predicciones del modelo.
2. Configurar y entrenar de manera óptima la arquitectura Swin Transformer.
3. Evaluar e interpretar las predicciones del modelo Swin Transformer para cada clase.
4. Comparar y analizar la exactitud (accuracy) del modelo Swin Transformer con modelos basados en redes neuronales convolucionales.

**Tabla 1.** Lista de patógenos de plantas de tomate (Panno et al., 2021).



**ESTADO DEL ARTE**

En el artículo ***"Machine Learning Approach towards Tomato Leaf Disease Classification"***, Gadade y Kirange, (2020) presentan un enfoque exhaustivo para la identificación y clasificación de las enfermedades de las hojas del tomate. El estudio sigue meticulosamente un proceso que comienza con la recopilación de datos, donde aprovecharon 9.000 imágenes de hojas de tomate del dataset PlantVillage. Esta recopilación se centró especialmente en siete clases de imágenes enfermas y una clase de imagen sana.

A continuación, los autores llevaron a cabo una etapa de preprocesamiento, en la que se minimizó el ruido de las imágenes de la enfermedad de la hoja de tomate utilizando un filtro de mediana. Después se realizó la extracción de características mediante tres técnicas: GLCM (Gray-Level Co-Occurrence Matrix), Gabor y SURF (Speeded Up Robust Features). Para clasificar estas imágenes de hojas de tomate en categorías normales o enfermas, los investigadores emplearon una serie algoritmos de clasificación. Entre ellos, la máquina de soporte vectorial (SVM), k-Nearest Neighbors (KNN), Naïve Bayes (NB) y árboles de decisión.

Los resultados de la investigación se centraron en métricas clave como accuracy, precision, recall y f1-score. Al emplear árboles de decisión con características extraídas mediante Gabor, se registraron las siguientes métricas: 'accuracy' de 0,6497, 'precision' de 0,2047, 'recall' de 0,861 y 'f1-score' de 0,2772. Sin embargo, al combinar el clasificador SVM con Gabor, los resultados mejoraron considerablemente: 'accuracy' de 0,7339, 'precision' de 0,2525, 'recall' de 0,9492 y 'f1-score' de 0,3989. Utilizando el clasificador KNN y Gabor, las métricas alcanzaron 'accuracy' de 0,732, 'precision' de 0,2555, 'recall' de 0,9831 y 'f1-score' de 0,4056. Finalmente, con Naïve Bayes y Gabor, se observaron valores de 'accuracy' de 0,675, 'precision' de 0,2187, 'recall' de 0,9695 y 'f1-score' de 0,3568.

Acorde a los resultados obtenidos, el estudio destaca que Gabor fue el método de extracción más eficaz, y que SVM resultó ser el algoritmo de clasificación más competente. Consecuentemente, Gadade y Kirange (2020) respaldan la combinación de SVM y Gabor como una herramienta robusta para la clasificación de enfermedades en hojas de tomate, subrayando su aplicabilidad en escenarios en tiempo real. Aun así, sugieren que técnicas más avanzadas, como Adaptive neuro fuzzy, Neural Networks y algoritmos genéticos, podrían ampliar las capacidades de clasificación.

No obstante, este trabajo evidencia que los métodos tradicionales de aprendizaje automático presentados aquí muestran métricas inferiores en comparación con las técnicas modernas de Deep learning basadas en redes neuronales convolucionales. Esta observación subraya una posible obsolescencia de los métodos tradicionales en el ámbito de la clasificación de enfermedades en hojas de tomate a través de técnicas de visión por computadora.

En el estudio titulado ***"ToLeD: Tomato Leaf Disease Detection using Convolution Neural Network"*** de Mohit Agarwala et al. (2020) se propone un enfoque basado en una Red Neuronal Convolucional (CNN) para la detección y clasificación de enfermedades en las hojas del tomate. El trabajo parte de la importancia de los tomates como cultivo global, abordando el problema de la disminución en la calidad y cantidad debido a diversas enfermedades.

El estudio utiliza el dataset de Plant Village que contiene imágenes de hojas de tomate afectadas por nueve tipos de enfermedades, además de una clase de hojas sanas. Para equilibrar el dataset, se aplicaron técnicas de aumento de datos resultando en 10,000 imágenes para el entrenamiento, 7,000 para la validación y 500 para las pruebas.

Se diseñó una arquitectura CNN específica que incorpora 3 capas convolucionales, 3 capas de max-pooling y 2 capas completamente conectadas. Además, se utilizaron modelos CNN preentrenados mediante transferencia de aprendizaje y utilizando el mismo dataset, con el fin de comparar su eficiencia con el modelo propuesto.

Los resultados de las métricas evaluadas varían en un rango de precisión del 76% al 100% en función de las clases, con una precisión promedio del 91,2%, superando a los modelos pre-entrenados, como VGG16 con un accuracy promedio del 77.2%, InceptionV3 del 63.4% y MobileNet del 63.7%.

El modelo propuesto presenta un número menor de parámetros entrenables en comparación con los modelos pre-entrenados, lo que sugiere que podría ser más eficiente en términos de almacenamiento y computación. Sin embargo, la variabilidad en el ‘accuracy’ por clase y su bajo puntaje promedio comparado con el de estudios más recientes son un indicador claro de que, aunque el modelo propuesto es en general eficaz, todavía hay margen para mejorar la predicción por clases y plantea como trabajo futuro el ajuste de la arquitectura para incrementar la precisión.

## Title: Tomato plant disease detection using transfer learning with C-GAN synthetic images

**Authors:** Amreen Abbas, Sweta Jain, Mahesh Gour

**Methodology**

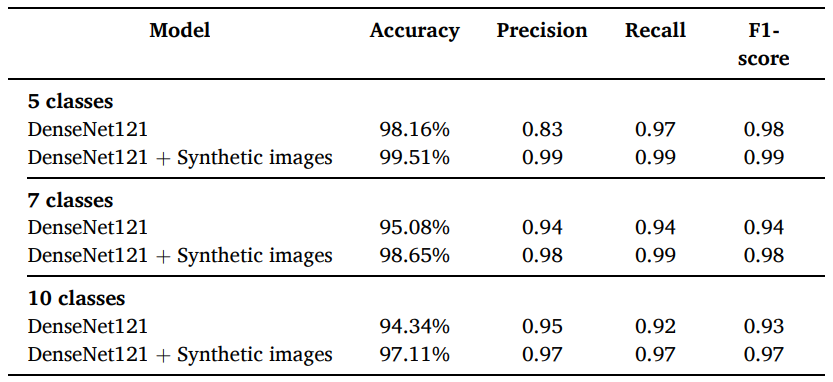
A method based on deep learning has been presented to detect diseases in tomato plants. This technique makes use of a Conditional Adversarial Generative Network (C-GAN) to generate synthetic visual representations of tomato leaves. Subsequently, training of a model called DenseNet121 is carried out, which is fed with both real and synthetic images. This training process is performed through knowledge transfer, allowing the model to accurately classify tomato leaf images into up to ten different disease categories.

The proposed model has undergone an exhaustive training and evaluation phase using the PlantVillage public dataset, which focuses on plant-related issues. The results obtained are highly promising: an accuracy of 99.51% in classifying images in five disease categories, 98.65% in seven categories and 97.11% in ten categories has been achieved. These results highlight the effectiveness of the approach in accurately detecting tomato leaf diseases, which could potentially have a significant impact on improving the health and yield of agricultural crops.

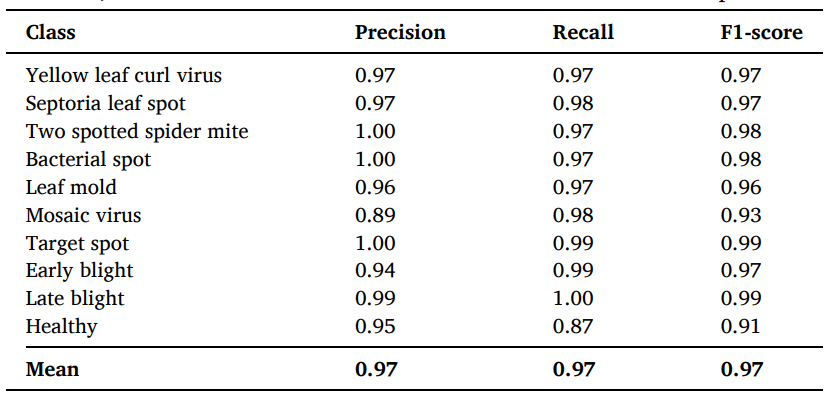
**Results**

Table represents the classification performance of the proposed method on the PlantVillage dataset and augmented dataset (PlantVillage + Synthetic images dataset). The proposed method achieved a classification accuracy of 99.51%, 98.65%, and 97.11% for 5-class

**Table 2. Accuracy, Precision, Recall, and F1-score for the model with and without augmentation. [10]**

****

**Table 3. Precision, Recall and F1-score for different disease classes of tomato plant. [10]**

****

**Conclusion**

The proposed data augmentation technique effectively enhances the neural network's generalization capacity, thereby mitigating the risk of overfitting. The achieved results with the proposed model are highly encouraging, achieving precision rates of 98.16%, 95.08%, and 94.34% on the original PlantVillage dataset for 5, 7, and 10-class classification tasks, respectively. Moreover, when utilizing the expanded dataset that includes synthetic PlantVillage+ images, precision rates increase to 99.51%, 98.65%, and 97.11% for the same classification tasks.

The experimental outcomes unequivocally underscore the superiority of the proposed method compared to existing approaches. In future work, the intention is to extend this methodology to encompass disease identification and classification across various parts of the plant, including fruits, stems, and branches. Additionally, there is a plan to undertake the identification of distinct stages of disease development in plants.

## 

## Title: Tomato leaf disease classification using supervised learning techniques: contrasting analysis. International Conference on Advances in Computing, Communication and Materials (ICACCM).

**Authors:** Vandana Rawat, Neelam Singh, Bhavleen Kaur, Saksham Bora

**Methodology**

The methodology of the study consisted of the following components:

1. **Dataset**: They utilized all the images of tomatoes from the "Plant Village" collection, which contained 10 categories, among them pictures of healthy tomatoes.
2. **Preprocessing**: Obtaining and preprocessing images of tomato leaf disease, including noise removal.
3. **Machine Learning Techniques**: Applied various algorithms including SVM, K-Nearest Neighbor, Naïve Bayes, Decision Tree, Feed Forward Neural Network, Back Propagation Neural Network, Deep Neural Network, Conventional Neural Network, and Multi Kernel SVM.
4. **Evaluation Metrics**: Utilized Accuracy, Precision, F1-Score, and Recall for evaluating the performance of the learning algorithms.

**Results**

**Table 4. Comparison of ML Algorithms [11]**

|  |  |  |  |
| --- | --- | --- | --- |
| **Author and Year** | **Dataset** | **Algorithm** | **Accuracy (%)** |
| H. D. Gadade [2020] | PlantVillage dataset | Decision Tree | 67,00 |
| Jayanthi M.G [2020] | PlantVillage dataset | SVM | 90,00 |
| Nishant Vijay [2021] | PlantVillage dataset | KNN | 83.6 |
| Dr. D.K. Kirange [2020] | PlantVillage dataset | Naive Bayes | 67,00 |
| Dr. Sreelatha P [2021] | PlantVillage dataset | FFNN | 56.89 |
| Sridhar Udaya kumar [2021] | PlantVillage dataset | BPNN | 57.19 |
| K. Ch. Sri Kavya [2021] | PlantVillage dataset | Linear Kernel SVM | 59.28 |
| S. Karthick [2021] | PlantVillage dataset | RNN | 59.54 |
| Nishant Vijay [2021] | PlantVillage dataset | CNN | 98.5 |
| Shashikumar D.R. [2020] | PlantVillage dataset | MKSVM | 97.34 |

The results were presented as a comparative analysis of various machine learning techniques applied to the hydroponics dataset. The Convolution Neural Network (CNN) achieved the highest accuracy of 98.5%. Other algorithms' accuracies were also presented, highlighting the effectiveness of different methods in the tomato leaf disease classification. [11]

**Conclusión**

The study concluded that machine learning technologies are frequently utilized for tomato leaf disease prediction and prognosis. Among the various machine learning techniques compared, the Convolution Neural Network (CNN) method was identified as the best for detecting tomato leaf disease early on, with an accuracy of 98.5%. The study also emphasized the potential of hydroponics and the application of data-driven techniques in modern agriculture.

## Title: Artificial intelligence in tomato leaf disease detection: a comprehensive review and discussion

**Authors:**  Rajasekaran Thangaraj, S. Anandamurugan

**Methodology**

In the study, the researchers primarily searched for articles on tomato leaf disease from major electronic databases like IEEE Xplore, ScienceDirect, Google Scholar, and ACM library. These sources were selected because they contain a significant number of studies on tomato leaf disease detection using image processing, including machine learning (ML) and deep learning (DL) techniques.

The focus has been on publications from 2015 onward due to the growing interest in tomato plant disease detection using AI techniques in recent years. The search was initiated using specific keywords related to tomato leaf disease and AI techniques.

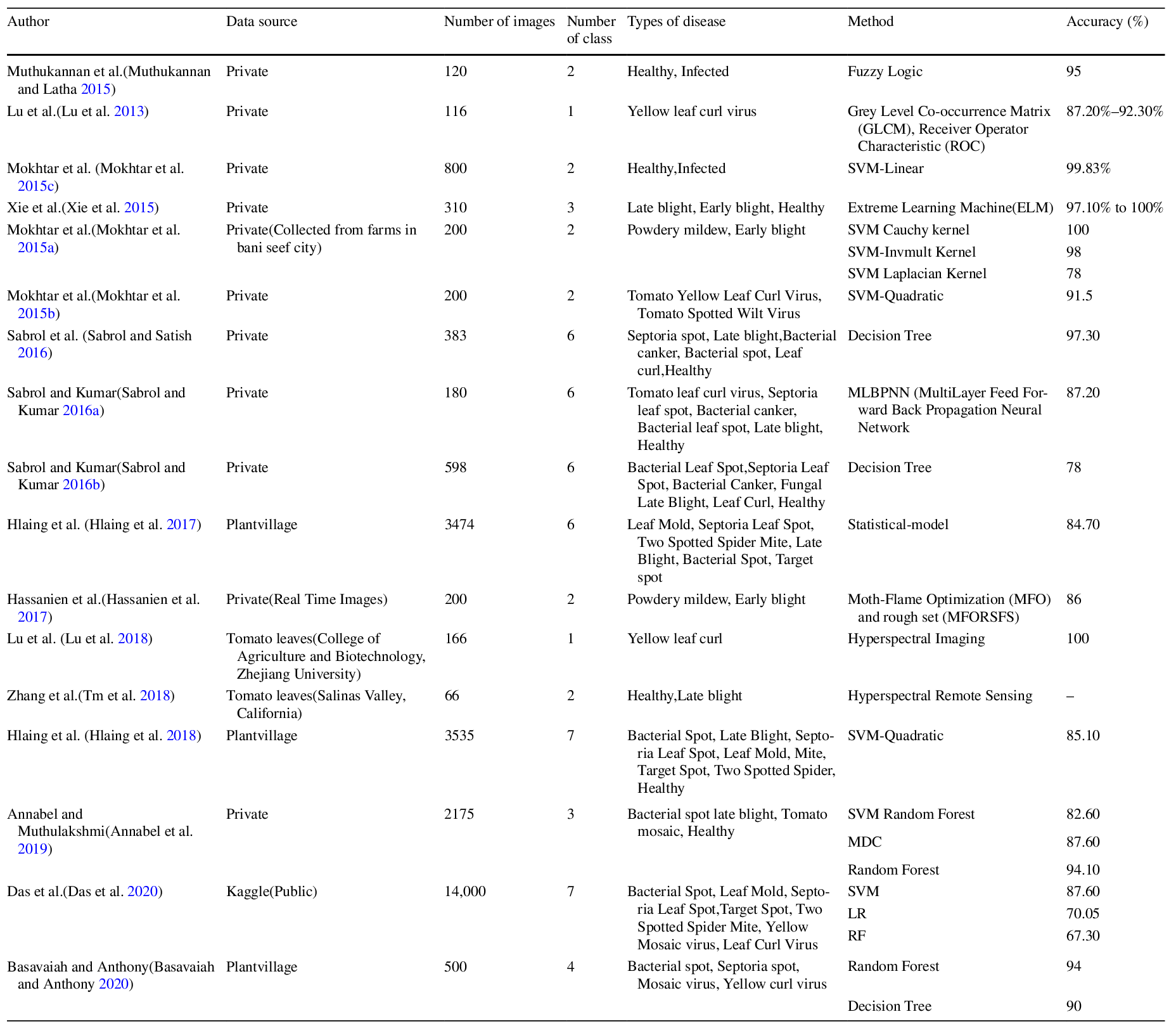
The research process is depicted in Fig. 1. The initial step involved downloading papers relevant to tomato leaf disease detection using AI. After reading, they were categorized based on the techniques used: traditional image processing, ML, or DL. From the search, 79 papers were identified, which was later refined down to 44 after thorough optimization and analysis.

Each article was reviewed in detail, considering:

* The specific plant diseases discussed.
* The AI model used.
* The dataset employed.
* The performance of the ML and DL techniques in the research.

**Results**

**Table 3. Comparison table showing the performance of machine learning algorithm applied to the detection of tomato plant disease through leaf images [12]**

****

**Table 5. Comparison table showing performance of deep learning algorithm applied to the detection of tomato plant disease through leaf images [12]**

****

**Conclusion:**

The agricultural domain has grappled with various challenges in recent times. This study offers a comprehensive review of the latest research on tomato leaf disease identification using artificial intelligence. We examined 44 related studies, focusing on datasets, pre-processing methods, models, and prediction accuracy. Our analysis prioritized data sources, top accuracy rates, and methodologies. The findings indicate that deep learning models surpass traditional methods like image processing and neural networks in identifying tomato diseases from leaf images.

Early detection of these diseases can minimize costs by avoiding unnecessary pesticide use. Emerging techniques like combining deep learning with hyperspectral imaging are particularly promising for early disease detection. As tomato diseases intensify over time, specialized deep learning models can be utilized to track and categorize these diseases throughout their lifecycle. To enhance prediction speed and accuracy, the integration of features from CNN models is beneficial. Looking ahead, the integration of agricultural robots and drones to automatically capture and classify diseased plant images is a promising avenue.

## Title: Computer-aided fusion-based neural network in application to categorize tomato plants. Springer-Verlag London Ltd., part of Springer Nature 2023

**Authors:** Rajyalakshmi Uppada · D. V. A. N. Ravi Kumar

**Methodology**

The methodology employed can be broken down into the following key steps:

1. **Image Pre-processing**: Utilizing Non-Subsampled Contourlet to acquire energy-detail components from the image dataset (Kaggle PlantVillage and Mendeley dataset).
2. **Cluster Extraction**: Modified K-means processing is used to extract colored clusters for tomato-leaf detection.
3. **Classification using SVM**: 38 extracted features are utilized by the Multiclass SVM (M-SVM) classifier for accurate categorization.
4. **Proposed M-CNN model**: A custom deep learning architecture is established for optimized unsupervised categorization. The model is based on the DenseNet-201 architecture and involves pre-processing stages followed by CNN categorization.

**Results**

**Table 6** Distinguished performance of NSC + K-Means + M-CNN classifier on specific tomato-leaves [13]

|  |  |  |  |
| --- | --- | --- | --- |
| **Category of tomato-leaf** | **Number of images** | **Average accuracy** | **M-CNN categorization** |
| Healthy Tomato-Leaf (D0) | 1528 | 99.32 | Healthy |
| Early-Blight (D1) | 1600 | 98.14 | Infected |
| Septoria-Leaf-Spot (D2) | 1701 | 99.17 | Infected |
| Leaf-Mold (D3) | 1523 | 99.39 | Infected |
| Bacterial-Spot (D4) | 1702 | 99.18 | Infected |
| Spider-Mites (D5) | 1610 | 99.32 | Infected |
| Late-Blight (D6) | 1527 | 99.27 | Infected |
| Target-Spot (D7) | 2247 | 98.53 | Infected |
| Mosaic-Virus (D8) | 1555 | 99.7 | Infected |
| Yellow-Leaf-Curl-Virus (D9) | 4286 | 99.55 | Infected |
| Powdery\_Mildew (D10) | 1004 | 99.09 | Infected |

**Table 6** Normalized confusion matrix for M-CNN classifier [13]

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Actual class | Predicted class | | | | | | | | | | |
|  | **D0** | **D1** | **D2** | **D3** | **D4** | **D5** | **D6** | **D7** | **D8** | **D9** | **D10** |
| **D0** | **98.98** | 0 | 0 | 0 | 0.13 | 0 | 0.13 | 0.76 | 0 | 0 | 0.13 |
| **D1** | 0.13 | **95.80** | 0.64 | 0.51 | 0 | 0.25 | 0.25 | 1.65 | 0.76 | 0 | 0.13 |
| **D2** | 0.64 | 0.91 | **94.48** | 0.82 | 0.78 | 0.67 | 0.64 | 0.80 | 0.51 | 0.76 | 0 |
| **D3** | 0.38 | 0.13 | 1.27 | **96.18** | 0.64 | 0.13 | 0.25 | 0.76 | 0.13 | 0.13 | 0.89 |
| **D4** | 0.25 | 0.25 | 1.27 | 0.25 | **96.82** | 0.13 | 0.64 | 0.25 | 0 | 0.13 | 0 |
| **D5** | 0.51 | 0.78 | 0.27 | 0.64 | 0.91 | **98.17** | 0.89 | 0.56 | 0.38 | 0.89 | 0.27 |
| **D6** | 0.16 | 0.13 | 0 | 0.13 | 0 | 0 | **99.95** | 0 | 0.13 | 0 | 0.27 |
| **D7** | 3.56 | 0.25 | 0.89 | 0.25 | 0.51 | 1.27 | 2.54 | **90.33** | 0.13 | 0.25 | 1.27 |
| **D8** | 0.13 | 1.40 | 0 | 0 | 0 | 0 | 0.51 | 0 | **97.96** | 0 | 0 |
| **D9** | 0 | 0 | 0.13 | 0 | 0 | 0 | 0.13 | 0.13 | 0.13 | **99.49** | 0.64 |
| **D10** | 0.64 | 0.91 | 0.82 | 0.78 | 0.67 | 0.64 | 0.80 | 0.51 | 0.76 | 0 | **94.48** |

These are the results of the proposed method:

* Achieved a model accuracy of 99.15% and an average precision of 95.6%.
* The combination of NSC, K-means, and M-CNN classifiers produced average sensitivity and specificity of 98.78% and 97.34% respectively.
* A detailed comparison with state-of-the-art approaches showed the model's superiority in terms of accuracy, precision, and handling more disease classes.

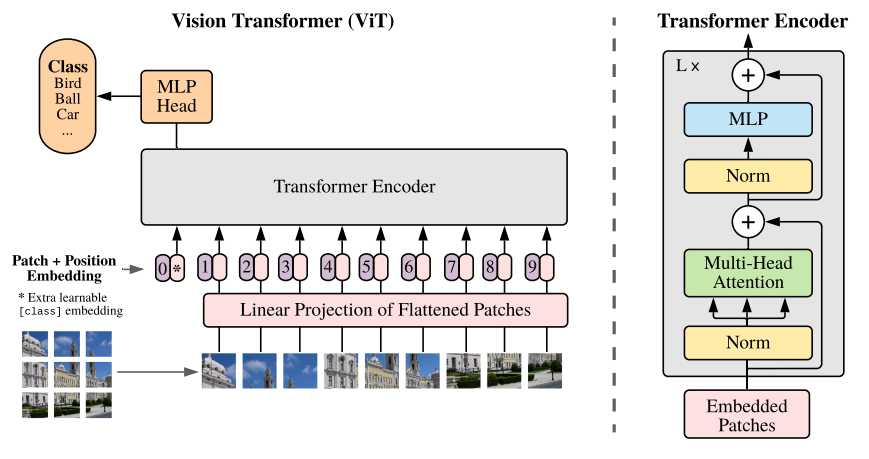
**Conclusion**

The study successfully introduces an efficient and accurate approach to detect and categorize diseased and healthy tomato plants. The proposed method, employing image pre-processing, clustering, SVM, and a customized CNN model, significantly outperformed traditional methods and other state-of-the-art approaches.

# THEORETICAL FRAMEWORK

**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 [14].



**Figure 1:** Model overview of the first visual transformer [14]

**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 [14][15].

**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.

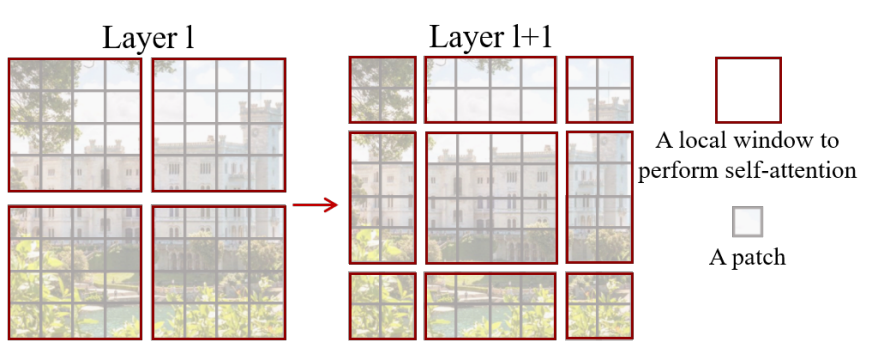
**MLP (Multi-Layer Perceptron):** Refers to a fully connected neural network. In the context of transformers, MLPs are used to transform features after layers of attention.

**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 [16].



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

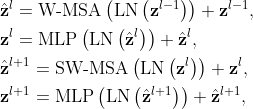
Equations (1) and (2) show the difference in complexity between global care and window-based care. Window-based attention is scalable and computationally more efficient [16].

(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.

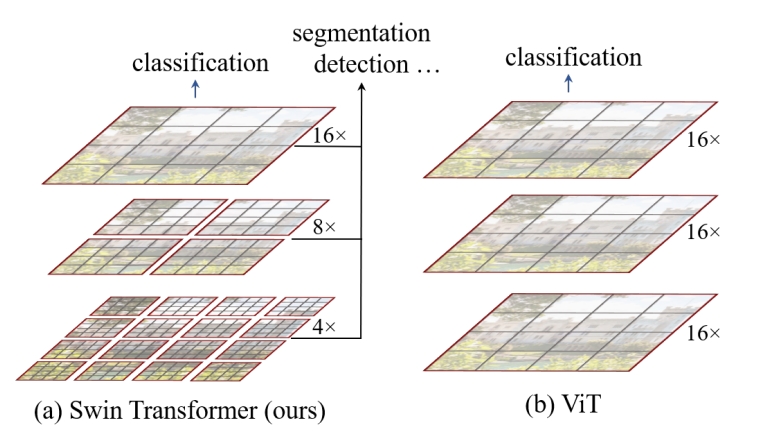
Equations (3) show how features are computed in successive blocks using different window-based attention configurations [16].

(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 [16].



**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 [16].

**- 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 [16].

**-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.

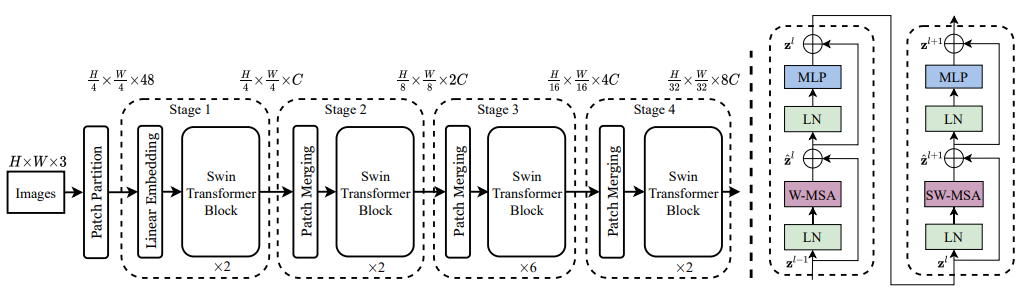
**-GELU:** GELU, or Gaussian Error Linear Unit, is a nonlinear activation function. It is used in neural networks to introduce nonlinearities into the model. In the Swin Transformer block, the GELU is used as an activation function in the 2-layer multilayer perceptron (MLP).

**-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 [16].

**- 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
* 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.
* 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.