**Desarrollo de un modelo 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)

[**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). 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.

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 un dataset balanceado en las diferentes clases de las enfermedades del tomate?
* ¿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

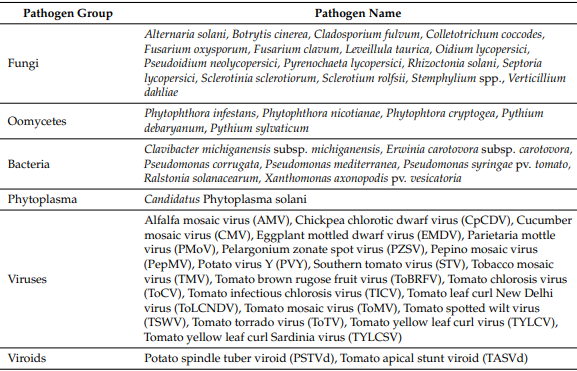
## Objetivo general

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 un dataset balanceado en las diferentes clases de las enfermedades del tomate.
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 de machine learning para la identificación y clasificación de las enfermedades de las hojas del tomate. El proceso de investigación inició con la recolección de 9,000 imágenes de hojas de tomate del dataset de PlantVillage, enfocándose en siete clases de plantas enfermas y una clase de planta sana.

En la fase de preprocesamiento se minimizó el ruido de las imágenes utilizando un filtro de mediana, después se realizó la extracción de características mediante las técnicas GLCM (Gray-Level Co-Occurrence Matrix), Gabor y SURF (Speeded Up Robust Features). Para clasificar las imágenes en categorías normales o enfermas, se emplearon varios algoritmos de clasificación como la máquina de soporte vectorial (SVM), k-Nearest Neighbors (KNN), Naïve Bayes (NB) y árboles de decisión.

Los resultados mostraron que la combinación de SVM y Gabor obtuvo métricas superiores a todas las demás con un 'accuracy' de 0,7339, 'precision' de 0,2525, 'recall' de 0,9492 y 'f1-score' de 0,3989. A pesar de estos resultados, los autores indican que técnicas más avanzadas, como Adaptive Neuro Fuzzy, Neural Networks y algoritmos genéticos, podrían mejorar la clasificación.

No obstante, este trabajo evidencia que los métodos tradicionales de machine learning muestran métricas inferiores en comparación con las técnicas modernas de Deep learning basadas en redes neuronales convolucionales. Esta observación señala 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 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.

La arquitectura CNN desarrollada en el estudio consta de 3 capas convolucionales, 3 capas de max-pooling y 2 capas completamente conectadas. Se evaluaron también modelos CNN preentrenados a través de transferencia de aprendizaje, empleando el mismo dataset, con el objetivo de contrastar la eficacia del modelo propuesto. Las métricas resultantes mostraron una precisión en un rango de 76% a 100% para diversas clases, con una precisión promedio de 91.2%. Este rendimiento supera al de modelos preentrenados como VGG16, con una precisión del 77.2%, InceptionV3, con 63.4%, y MobileNet, con 63.7%

El modelo propuesto tiene un número menor de parámetros entrenables comparado al de los modelos preentrenados, sugiriendo potenciales ventajas en almacenamiento y computación. No obstante, la variabilidad en la exactitud entre clases y la comparación con trabajos más actuales indican que el modelo, aunque eficaz, está superado por enfoques más recientes.

El artículo ***"Tomato plant disease detection using transfer learning with C-GAN synthetic images"*** de Abbas et al. (2021) aborda un enfoque de Deep learning para la detección y clasificación de enfermedades de la planta del tomate. La metodología implementa dos fases principales, la generación de imágenes sintéticas mediante Redes Antagónicas Generativas Condicionales (C-GAN) para extender el dataset y la clasificación de enfermedades utilizando un modelo DenseNet121 previamente entrenado.

El desempeño del método se evaluó con el dataset de PlantVillage, que incluye nueve clases de hojas de tomate con distintas enfermedades además de una clase de hojas sanas. Las imágenes sintéticas generadas con C-GAN se añadieron al dataset existente. Luego, se entrenó el modelo DenseNet121 con el conjunto de imágenes extendido.

Se realizaron dos series de experimentos, el primero centrado en el entrenamiento de C-GAN y la generación de imágenes sintéticas; el segundo, en el entrenamiento del modelo DenseNet121 utilizando el dataset ampliado. Un análisis comparativo con modelos preentrenados como VGG19, ResNet50, Inception-V3, Xception y MobileNet mostró que el modelo propuesto logró el más alto nivel de exactitud, con un 97.11% en la clasificación de las 10 clases del estudio.

No obstante, aunque el modelo propuesto demostró una alta eficacia en la clasificación de enfermedades de la planta de tomate y la aplicación de C-GAN ayudó a evitar el sobreajuste, es importante evaluar la fidelidad de las imágenes sintéticas generadas, ya que podrían no tener el nivel de detalle que presentan las imágenes reales y, por lo tanto, afectar negativamente el rendimiento del modelo en tareas de generalización.

El estudio ***"Early Detection and Classification of Tomato Leaf Disease Using High-Performance Deep Neural Network"*** de Trivedi et al. (2021) aborda la identificación y clasificación de enfermedades en hojas de tomate mediante Convolutional Neural Networks (CNN). Se empleó un conjunto de datos compuesto por 3000 imágenes, distribuidas en nueve clases de hojas afectadas y una clase de hojas sanas. En la fase de preprocesamiento, las imágenes se normalizaron y redimensionaron.

El proceso comprende etapas de preprocesamiento, segmentación y clasificación a través de CNN. La arquitectura CNN especificada consta de ocho capas convolucionales, ocho capas de max pooling, tres capas fully connected, con una tasa de dropout de 0.5 y función de activación Relu. Se realizaron experimentos variando parámetros como la tasa de aprendizaje y el número de épocas.

Los resultados de la métrica de exactitud fueron del 98.49%. Los experimentos adicionales mostraron que la precisión varía levemente con diferentes tasas de aprendizaje y épocas, oscilando en general alrededor del 98.4% al 98.58%. La comparación con modelos estándar como MobileNet, VGG16 e InceptionV3 indicó que el modelo propuesto supera a estos modelos CNN en exactitud.

El estudio resalta la eficacia del modelo propuesto, especialmente en comparación con otros modelos. Sin embargo, omite discusiones sobre limitaciones, aplicabilidad futura o la justificación para la elección de la arquitectura específica empleada.

El artículo titulado ***"Tomato leaf disease classification using supervised learning techniques: contrasting analysis"***, presentado por Rawat et al., (2022), ofrece una contribución al campo de la clasificación de enfermedades en hojas de tomate a través de técnicas de aprendizaje supervisado.

El estudio se articula en torno a cuatro componentes principales: selección del dataset de imágenes de tomates del archivo "Plant Village", el preprocesamiento de estas imágenes para la eliminación de ruido mediante un filtro de mediana, la aplicación de diversos algoritmos de aprendizaje automático como SVM, K-Nearest Neighbour, Naïve Bayes, Decision Tree, Feed Forward Neural Network, Back Propagation Neural Network, Deep Neural Network, Convolutional Neural Network y Multi Kernel SVM, y por último la evaluación de la efectividad de estos algoritmos mediante métricas como Precisión, F1-Score y Recall.

El estudio realiza un análisis comparativo de la eficacia de los diferentes algoritmos. En este análisis, la Red Neuronal Convolucional (CNN) logra una exactitud de 98,5%, superando a otros algoritmos como SVM, que alcanza un 90% de exactitud, y KNN con un 83,6%. La investigación concluye que las CNN ofrecen un rendimiento superior en la clasificación de enfermedades de hojas de tomate en comparación con métodos tradicionales de machine learning.

El estudio titulado ***"AlexNet Convolutional Neural Network for Disease Detection and Classification of Tomato Leaf"*** (Chen et al., 2022) aborda la detección y clasificación de enfermedades en hojas de tomate mediante una red neuronal convolucional (CNN) basada en la arquitectura de AlexNet. Los autores utilizaron un conjunto de datos de 18,345 imágenes de entrenamiento y 4,585 imágenes de prueba, distribuidas en nueve clases correspondientes a diversas enfermedades de la hoja del tomate y una clase de hoja sana.

La modificación realizada a la arquitectura de la red AlexNet involucra tres capas de convolución, tres capas completamente conectadas y una capa de salida. Se utilizó el optimizador Adam con una tasa de aprendizaje de 0.0005, 75 épocas y un tamaño de lote de 128. La función de pérdida fue la entropía cruzada. El modelo mostró una exactitud del 96%, una tasa de precisión del 98%, un valor de recall del 95% y un F1-score del 97%.

Las conclusiones destacan la eficacia del modelo para la clasificación de enfermedades en hojas de tomate. Se señala que el modelo es lo suficientemente ligero como para implementarse en dispositivos móviles, lo que es relevante dadas las limitaciones de capacidad de memoria en estos dispositivos.

El estudio ***"BotanicX-AI: Identification of Tomato Leaf Diseases Using an Explanation-Driven Deep-Learning Model"*** de Bhandari et al., (2023), aborda la clasificación de nueve clases de enfermedades en hojas de tomate y una clase para las hojas sanas utilizando un modelo de aprendizaje profundo basado en EfficientNetB5. El dataset se construye preprocesando las imagenes mediante redimensionamiento y aumentos de datos resultando en 11,000 imágenes de hojas de tomate distribuidas en 10 clases.

La metodología consiste en aplicar transferencia de aprendizaje utilizando un modelo de EfficientNetB5 al que se le agregaron capas adicionales como la normalización por lotes, capas densas y técnicas de regularización como dropout. Para incrementar la interpretabilidad del modelo se integran técnicas como GradCAM y LIME, que ayudan a identificar las regiones de la imagen que más contribuyen a la clasificación.

En cuanto a métricas el modelo propuesto alcanzó una exactitud del 99.07%, superando a otros modelos preentrenados, como MobileNet (94,00%), Xception (95,32%), VGG16 (93,35%), ResNet50 (96,03%) y DenseNet121 (96,30%). Las técnicas de GradCAM y LIME pudieron identificar las regiones más importantes en las imágenes de las hojas para la clasificación. Aunque se observó que GradCAM tenía limitaciones en la identificación de ciertas regiones en algunas clases de enfermedades.

Finalmente, los autores recomiendan la consideración de otras técnicas de interpretabilidad, como HiResCAM y SHAP, para trabajos futuros, argumentando que estas pueden aumentar la confianza en la aplicabilidad del modelo.

Los estudios revisados abordan el problema de la detección y clasificación de enfermedades en hojas de tomate utilizando diversas técnicas que van desde machine learning tradicional hasta deep learning. Los algoritmos de machine learning, como SVM y KNN, muestran métricas de desempeño inferiores a los modelos basados en deep learning. Entre los modelos de deep learning, las Redes Neuronales Convolucionales (CNN) y sus variaciones, como DenseNet121 y EfficientNetB5, muestran métricas superiores en términos de precisión, exactitud y F1-score.

Los métodos que incorporan imágenes sintéticas generadas mediante Redes Antagónicas Generativas Condicionales (C-GAN) y técnicas de transferencia de aprendizaje parecen ofrecer ventajas en términos de desempeño y prevención de sobreajuste. Sin embargo, la calidad de las imágenes sintéticas generadas es una variable a considerar para tareas de generalización.

La aplicabilidad en dispositivos móviles y la interpretabilidad del modelo son aspectos que se resaltan en algunos trabajos, evidenciando la importancia de estos factores en aplicaciones prácticas. Las técnicas de explicabilidad como GradCAM y LIME ofrecen posibilidades para entender las decisiones del modelo, aunque se sugiere que métodos adicionales pueden mejorar la confianza en el modelo.

Los métodos basados en deep learning muestran un desempeño superior y se destacan como el método más efectivo en la clasificación de enfermedades de la planta de tomate. Sin embargo, no han presentado innovaciones en su arquitectura recientemente. Además, no se ha investigado la aplicación de arquitecturas de visión transformer para esta problemática específica, lo cual representa un área no explorada que podría ofrecer soluciones a limitaciones como la interpretabilidad o una mayor exactitud y que por tanto justifica una investigación más exhaustiva.

# MARCO TEÓRICO

**Enfermedades del Tomate** Para abordar este desafío, es fundamental comprender las enfermedades del tomate y sus efectos en la producción. Estas enfermedades pueden afectar negativamente la producción de tomates, reduciendo el rendimiento y la calidad de los cultivos. Además, se propagan de diversas formas, ya sea por contacto directo, a través de vectores como insectos o incluso por condiciones climáticas adversas. El impacto económico a largo plazo de estas enfermedades es significativo y puede influir en la estabilidad económica y los precios de los productos relacionados con el tomate en el mercado global.

**Fitopatología y Epidemiología Vegetal:** La fitopatología se centra en el estudio de las enfermedades de las plantas, y la epidemiología vegetal investiga cómo estas enfermedades se propagan en poblaciones de plantas. Estos campos son fundamentales para comprender la naturaleza y la propagación de las enfermedades que afectan a los cultivos, como los tomates.

**Suministro de Alimentos**: Los tomates y otros cultivos vegetales son fuentes fundamentales de alimentos en la dieta humana. Proporcionan nutrientes esenciales, vitaminas y minerales. Cualquier amenaza a la producción de estos cultivos puede tener un impacto directo en el suministro de alimentos.

**Economía Global:** La producción y comercialización de cultivos como el tomate representan una parte significativa de la economía agrícola global. La seguridad alimentaria está vinculada a la estabilidad económica, ya que la escasez de alimentos o las fluctuaciones en los precios pueden desencadenar crisis económicas y políticas.

**Variabilidad Climática:** El cambio climático está causando fenómenos climáticos extremos, como sequías e inundaciones, que pueden dañar los cultivos. La detección temprana de enfermedades del tomate, que pueden agravarse en condiciones climáticas adversas, es fundamental para la gestión de riesgos climáticos y la seguridad alimentaria.

**Diversidad Genética** de Plantas y Resistencia a Enfermedades Vegetales: La diversidad genética dentro de las plantas desempeña un papel importante en su capacidad para resistir enfermedades. Algunas variedades de tomates pueden ser naturalmente más resistentes que otras, lo que destaca la importancia de la diversidad genética en la seguridad alimentaria.

**Pérdidas por Enfermedades:** Las enfermedades del tomate pueden causar pérdidas significativas en la producción si no se detectan y controlan a tiempo. Estas pérdidas no solo afectan la disponibilidad de alimentos, sino que también pueden aumentar los precios de los productos alimenticios.

**Calidad de los Alimentos**: La seguridad alimentaria no se trata solo de la cantidad de alimentos disponibles, sino también de la calidad. Las enfermedades pueden afectar la calidad de los tomates, lo que a su vez puede afectar la salud de las personas que los consumen.

**Uso de Pesticidas:** La detección temprana y precisa de enfermedades del tomate permite un uso más eficiente y dirigido de pesticidas. Esto es importante para reducir la cantidad de productos químicos utilizados en la agricultura, lo que puede tener beneficios para la salud humana y el medio ambiente.

**Impacto Económico y Medioambiental** La detección temprana y precisa de enfermedades del tomate tiene un impacto significativo en la sostenibilidad agrícola al reducir la necesidad de pesticidas y contribuir a una agricultura más sostenible. Además, promueve la producción de tomates saludables, lo que beneficia a la seguridad alimentaria y la salud pública a nivel global.

**Visión Artificial y Aprendizaje Automático:** La visión artificial y el aprendizaje automático son herramientas esenciales en la detección de enfermedades en cultivos de tomate. La visión artificial permite a las computadoras procesar y entender imágenes, mientras que el aprendizaje automático capacita a las máquinas para aprender patrones y tomar decisiones basadas en datos de entrenamiento.

**Redes Neuronales Convolucionales (CNN):** Dentro de la visión por computadora, las redes neuronales convolucionales (CNN) han sido ampliamente utilizadas para la detección de enfermedades en plantas, incluyendo el tomate. Estas redes utilizan capas convolucionales para detectar patrones y características en las imágenes. Además, se ha aplicado con éxito el transfer learning utilizando modelos CNN pre-entrenados en grandes conjuntos de datos.

**Arquitectura Swin Transformer:** Swin Transformer, que significa "Shifted Window Transformer", introduce una arquitectura innovadora que aborda las limitaciones de los ViTs y las CNN tradicionales. Comienza dividiendo la imagen en parches y utiliza la atención de ventanas deslizantes en lugar de atención global, lo que reduce significativamente la complejidad computacional. Además, implementa una estructura piramidal para procesar imágenes a diferentes resoluciones, permitiendo capturar características a diferentes escalas.

**Transfer Learning y Arquitectura de Red Neuronal:** La transferencia de conocimientos de modelos pre-entrenados en redes neuronales es una técnica poderosa que puede acelerar el entrenamiento y mejorar el rendimiento en la detección de enfermedades.

**Dataset PlantVillage** Para entrenar modelos de detección de enfermedades del tomate, es esencial contar con un conjunto de datos de alta calidad. El dataset PlantVillage proporciona una variedad de clases de enfermedades del tomate y refleja la complejidad de la detección en el campo. La calidad de los datos etiquetados es fundamental para el éxito de la capacitación de modelos.

**Clases del dataset de plantVillage:**

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)

**Resolución de Imagen y Aumento de Datos**: En la detección de enfermedades de tomates mediante visión por computadora, la resolución de imagen y las técnicas de aumento de datos pueden mejorar la precisión al proporcionar imágenes de alta calidad y aumentar la cantidad de datos de entrenamiento.

**Capas de Atención:** En el modelo Swin Transformer, se emplean capas de atención para priorizar ciertas características de las imágenes durante el proceso de análisis y clasificación. Estas capas posibilitan la asignación de pesos diferentes a diversas regiones de la imagen, optimizando así la capacidad del modelo para realizar tareas específicas.

**Interpretabilidad en Modelos de Visión Artificial:** La interpretabilidad en modelos de visión artificial es crucial para evaluar la calidad de las decisiones del modelo en la clasificación de enfermedades de los tomates. Esta interpretabilidad permite comprender cómo se llega a cada decisión y puede guiar acciones correctivas.

**Aprendizaje Supervisado:** es un enfoque en el aprendizaje automático donde un algoritmo se entrena utilizando un conjunto de datos que contiene etiquetas de salida. El objetivo es aprender una función que mapee las entradas a las salidas. Este método se aplica comúnmente en tareas como clasificación y regresión.

**Clasificación de Imágenes:** Esta tarea pertenece al ámbito de la visión por computadora. Consiste en asignar una etiqueta a una imagen de entrada seleccionada de un conjunto de categorías predefinidas. El proceso generalmente implica el uso de modelos de aprendizaje profundo como las redes neuronales convolucionales.

**Matriz de Confusión:** Es una tabla que presenta el rendimiento de un algoritmo de clasificación. Las filas y columnas representan las clases reales y las clases predichas, respectivamente. Los elementos de la matriz son:

* Verdaderos Positivos (TP): Casos positivos correctamente identificados.
* Verdaderos Negativos (TN): Casos negativos correctamente identificados.
* Falsos Positivos (FP): Casos negativos incorrectamente identificados como positivos.
* Falsos Negativos (FN): Casos positivos incorrectamente identificados como negativos.

**Métricas de Evaluación:** Parámetros que cuantifican la calidad de un modelo en tareas específicas. Se definen matemáticamente como:

* 
* 
* 
* 

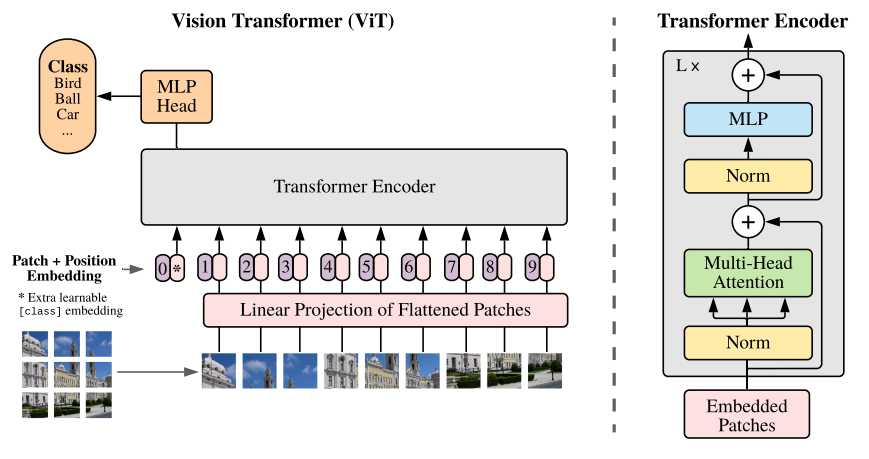
**Overfitting y Underfitting:** Dos problemas comunes en el aprendizaje automático. Overfitting ocurre cuando el modelo captura el ruido en lugar del patrón subyacente del conjunto de entrenamiento, lo que resulta en bajo rendimiento en datos no vistos. Underfitting sucede cuando el modelo no captura suficientemente la tendencia subyacente del conjunto de datos.

**Validación Cruzada:** Este es un método que implica dividir el conjunto de datos en múltiples subconjuntos de entrenamiento y prueba. El modelo se entrena y se evalúa múltiples veces, cambiando los subconjuntos utilizados. El rendimiento del modelo se estima promediando las métricas de rendimiento de cada iteración.

**Hiperparámetros:** Parámetros que no se aprenden del conjunto de entrenamiento, sino que se configuran previamente para el proceso de aprendizaje. Incluyen tasa de aprendizaje, número de épocas, y tamaño del lote en algoritmos como redes neuronales.

**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.

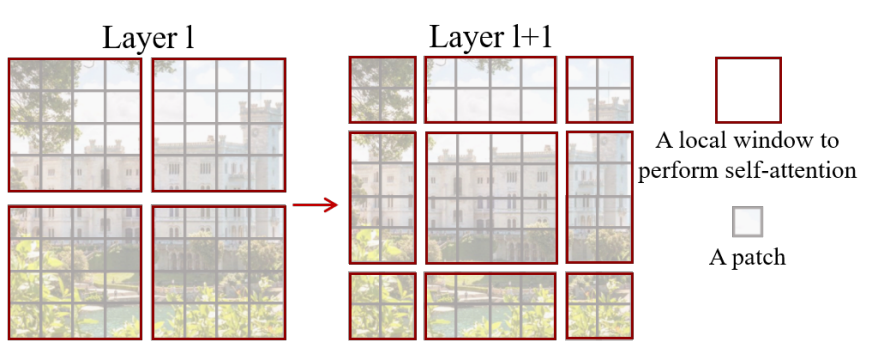
**MLP (Multi-Layer Perceptron):** Refers to a fully connected neural network. In the context of transformers, MLPs are used to transform features after layers of attention.

**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).



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

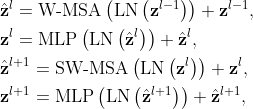
Equations (1) and (2) show the difference in complexity between global care and window-based care. Window-based attention is scalable and computationally more efficient (Liu et al., 2021).

(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.

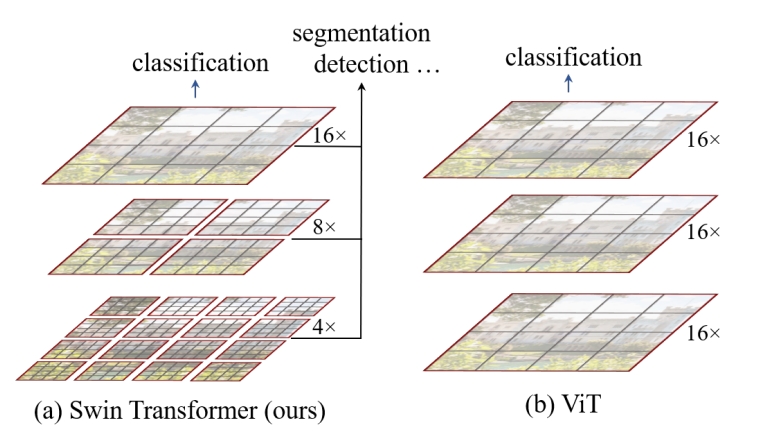
Equations (3) show how features are computed in successive blocks using different window-based attention configurations (Liu et al., 2021).

(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.

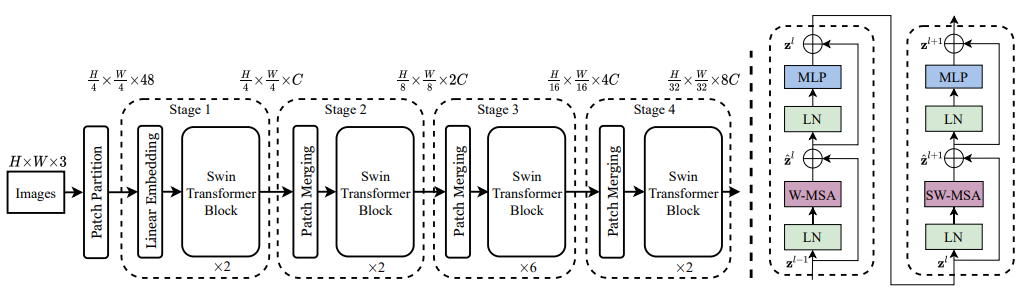
**-GELU:** GELU, or Gaussian Error Linear Unit, is a nonlinear activation function. It is used in neural networks to introduce nonlinearities into the model. In the Swin Transformer block, the GELU is used as an activation function in the 2-layer multilayer perceptron (MLP).

**-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.

Agarwal, Mohit, et al. “ToLeD: Tomato Leaf Disease Detection Using Convolution Neural Network.” *Procedia Computer Science*, vol. 167, no. 293–301, 2020, pp. 293–301, https://doi.org/10.1016/j.procs.2020.03.225.

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). https://doi.org/10.1016/j.compag.2021.106279

Chen, H.-C., Widodo, A. M., Wisnujati, A., Rahaman, M., Lin, J. C.-W., Chen, L., & Weng, C.-E. (2022). AlexNet Convolutional Neural Network for Disease Detection and Classification of Tomato Leaf. Electronics, 11(6), 951. https://doi.org/10.3390/electronics11060951

Bhandari, M., Shahi, T. B., Neupane, A., & Walsh, K. B. (2023). BotanicX-AI: Identification of Tomato Leaf Diseases Using an Explanation-Driven Deep-Learning Model. Journal of Imaging, 9(2), 53. https://doi.org/10.3390/jimaging9020053

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.

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.