Novasinergia
Revista digital de ciencia, ingeniería y tecnología

Unach
UNIVERSIDAD NACIONAL DE CHIMBORAZO

Artículo de Investigación

Clasificación de enfermedades en hojas de papa utilizando Transformadores de Visión

Potato leaf disease classification using Vision Transformers

Fernando David Valle-Medina[1], Luis Javier Castillo-Heredia[1],
Mirella Azucena Correa-Peralta[1], Jomar Elizabeth Guzmán Seraquive[1]

[1]*Universidad Estatal de Milagro, Milagro, Ecuador, 091050;*

fvallem@unemi.edu.ec; mcorreap@unemi.edu.ec; jseraquive@unemi.edu.ec

*Correspondencia: lcastilloh@unemi.edu.ec

Citación: Valle, F.; Castillo, L.; Correa, M. & Guzmán, J., (2025). Clasificación de enfermedades en hojas de papa utilizando Transformadores de Visión. *Novasinergia. 8*(1). 142-156.

https://doi.org/10.37135/ns.01.15.06

Recibido: 15 mayo 2024
Aceptado: 05 agosto 2024
Publicado: 08 enero 2025

Novasinergia
ISSN: 2631-2654

Resumen: La detección y clasificación de enfermedades en cultivos es crucial para el desarrollo y crecimiento del sector agrícola. El uso de técnicas tradicionales y el bajo nivel técnico aplicado al control de los sembríos generan grandes pérdidas para los agricultores. La visión por computadora aporta soluciones en este campo, no obstante, las investigaciones actuales se centran en el uso de redes neuronales convolucionales (Convolutional Neural Networks, CNNs), las cuales tienen una limitada capacidad para ubicar de forma precisa las características de mayor relevancia en una imagen. Para superar estas limitaciones, nuestro estudio propone un modelo de aprendizaje profundo basado en la arquitectura de Transformadores de Visión (Vision Transformers, ViT) para detectar y clasificar las enfermedades tizón temprano y tizón tardío en las hojas de papa. En esta investigación se evidencia cómo las técnicas de aumento de datos, ajuste fino y aprendizaje por transferencia permiten mejorar el rendimiento del modelo. El conjunto de datos para entrenamiento y prueba fue tomado de la plataforma PlantVillage. El reporte de métricas de evaluación del modelo propuesto alcanza una exactitud de 99.18% y un puntaje F1 de 98.7%. Los resultados demuestran un alto nivel de predicción en las enfermedades foliares de papa y evidencian la eficiencia de los mecanismos de atención. Se concluye que el modelo se consolida como una herramienta innovadora y funcional para los agricultores.

Palabras clave: Agricultura de precisión, Aprendizaje profundo, Clasificación de imágenes, Enfermedades de la papa, Transformadores de visión (ViT).

Abstract: *Disease detection and classification in crops is crucial for the development and growth of the agricultural sector. Traditional techniques and the low technical level applied to crop control generate significant losses for farmers. Computer vision offers solutions in this field; however, current research focuses on using convolutional neural networks (CNNs), which cannot accurately locate the most relevant features in an image. To overcome these limitations, this study proposes a deep learning model based on the Vision Transformers (ViT) architecture to detect and classify early and late blight diseases in potato leaves. This research demonstrates how data augmentation, fine-tuning, and transfer learning techniques can improve the model's performance. The dataset for training and testing was taken from the PlantVillage platform. The report of the proposed model's evaluation metrics reaches an accuracy of 99.18% and an F1 score of 98.7%. The results demonstrate a high level of prediction in potato foliar diseases and evidence of the efficiency of attention mechanisms. It is concluded that the model is an innovative and functional tool for farmers.*

Keywords*: Precision agriculture, Deep learning, Image classification, Potato diseases, Vision transformer (ViT)*

1. Introducción

La Comisión Económica para América Latina y el Caribe plantea la importancia de acelerar los procesos de digitalización como un factor crucial para asegurar la sostenibilidad, la generación de oportunidades y la reducción de brechas entre el sector rural y el urbano en la región (CEPAL, FAO, & IICA, 2021). En Ecuador, el cultivo de papa (Solanum tuberosum) mantiene una posición relevante dentro de la economía y producción en la zona andina. Su importancia cultural, socioeconómica y su aporte como alimento nutritivo son pilares fundamentales en la seguridad y soberanía alimentaria (Ministerio de Agricultura y Ganadería, 2022). En los últimos años el sector agrícola ecuatoriano ha enfrentado varios desafíos relacionados principalmente con la alteración de patrones climáticos, la aparición de nuevas plagas y enfermedades que generan pérdidas significativas en la calidad y producción de la papa (López Calvajar et al., 2017). Entre las enfermedades con mayor impacto y cuyos efectos se hacen visibles en las hojas de papa tenemos: el tizón temprano, causado por el hongo Alternaria solani; el tizón tardío, generado por la bacteria Phytophthora infestans; la rizoctoniasis y diversas virosis, entre otras (Pérez & Forbes, 2016).

En las últimas décadas, el cultivo de papa se ha intensificado en varias zonas del Ecuador; no obstante, el escaso manejo técnico, el limitado asesoramiento al agricultor y el uso excesivo de productos químicos han generado la aparición de nuevas plagas y serios problemas ambientales (Barrera et al., 2002). Las estrategias tradicionales para enfrentar y detectar enfermedades en los sembríos se basan en métodos manuales laboriosos que requieren mucho tiempo para su ejecución (Paucar Buñay, 2016). La identificación automática de enfermedades de las plantas es esencial para una agricultura de precisión, ya que permite monitorear grandes extensiones de cultivos de manera eficiente y precisa. Al adoptar enfoques de aprendizaje automático, se pueden detectar y clasificar enfermedades en las plantas de forma rápida y con una mayor exactitud (Shirahatti et al., 2018).

Los avances en el campo de la Inteligencia Artificial (IA) han impulsado un desarrollo significativo en las técnicas de aprendizaje automático (Ihme at al., 2022). En los últimos años, se han optimizado modelos y algoritmos con capacidades avanzadas que permiten analizar grandes volúmenes de datos, mejorando la precisión y eficiencia en tareas específicas como la predicción y clasificación de información (Pintelas et al., 2020). El aprendizaje supervisado, uno de los paradigmas fundamentales del aprendizaje automático utiliza datos etiquetados para entrenar modelos y predecir resultados futuros (Sen et al., 2020). En el ámbito del procesamiento de imágenes, diferentes técnicas de aprendizaje supervisado se han convertido en herramientas valiosas para la extracción e interpretación de patrones, texturas y objetos (Elngar et al., 2021). Esta capacidad permite generar diversas aplicaciones, tales como la detección de enfermedades en cultivos, el reconocimiento facial y la detección de objetos en la robótica.

El aprendizaje profundo (deep learning, DL) y las redes neuronales convolucionales (Convolutional Neural Networks, CNNs) han transformado notablemente el campo de la visión por computadora (Lecun et al., 2015). Las redes neuronales profundas permiten diagnósticos más rápidos y precisos en el campo de la detección temprana de enfermedades

en cultivos (Mahum et al., 2022). De acuerdo con la literatura sobre la clasificación de enfermedades de la papa se observa una predominancia de la arquitectura de redes neuronales convolucionales. En recientes estudios realizados por Sakkarvarthi et al. (2022) y Lozada-Portilla et al. (2021) se obtienen tasas de exactitud de entre el 88% a 90% en la clasificación exitosa de enfermedades, respectivamente. Pese a que estos valores son relativamente aceptables existe un margen de mejora mediante el uso de metodologías modernas e innovadoras para la visión artificial. La arquitectura de Transformadores (Transformers) ha marcado un hito en el procesamiento del lenguaje natural (PNL) a través de la aplicación de mecanismos de atención expuestos en el trabajo original Attention is all you need (Vaswani et al., 2017). Inspirados por este desempeño se ha desarrollado una adaptación innovadora para tareas de visión por computador, dando origen a los Transformadores de Visión (Vision Transformer, ViT). Esta tecnología concibe cada imagen como una secuencia de palabras equivalente a 16x16 píxeles y utiliza mecanismos de autoatención, evitando así la dependencia convolucional y logrando procesar la información de manera paralela y simultánea (Dosovitskiy et al., 2021).

La arquitectura ViT procesa las imágenes de entrada dividiéndolas en parches, los cuales son convertidos en vectores que mantienen características similares entre sí. Estas nuevas entradas son codificadas posicionalmente conservando su estructura espacial y mediante bloques de atención se define la relación global que existe entre los diferentes parches. Esto permite que el modelo identifique patrones, texturas y colores, facilitando la clasificación en diferentes categorías. Las altas capacidades de aprendizaje que demuestran los modelos ViT se logran mediante el entrenamiento con conjuntos de datos a gran escala (Khan et al., 2023). Para enfrentar el desafío que representa el volumen de información requerida por los modelos para su entrenamiento existen proyectos que proporcionan una extensa base de imágenes diseñadas para investigación. ImageNet, por ejemplo, proporciona una base de datos con más de 14 millones de imágenes (Deng et al., 2009), mientras que JFT-300M de Google contiene aproximadamente 300 millones de imágenes etiquetadas en miles de categorías (Sun et al., 2017).

El aprendizaje por transferencia (transfer learning) es una estrategia que permite aprovechar el conocimiento adquirido por un modelo que ha sido preentrenado. Esta técnica busca aprovechar los patrones y características generales aprendidos, adaptándolos mediante ajuste fino (fine-tuning) a tareas específicas. Este procedimiento implica la modificación de las capas finales del modelo, otorgándole un enfoque particular. En la investigación para el reconocimiento de enfermedades en plantas realizada por Pavel et al. (2021), se compararon 3 arquitecturas preentrenadas de redes neuronales convolucionales: InceptionV3, Inception-ResNetV2, y ResNet34. Esta investigación evaluó su capacidad para clasificar 7600 imágenes distribuidas en 38 categorías. Los resultados revelaron que el modelo ResNet34 logró la precisión más alta con un 97,03%. En el estudio realizado por Shaheed et al. (2023), se propone un nuevo modelo que integra ViT y ResNet50. Los resultados preliminares alcanzaron una exactitud de 97.65%, el modelo demuestra un buen nivel para clasificar enfermedades incluso con muestras distorsionadas. Barman et al. (2024) aplican una solución basada en teléfonos inteligentes utilizando el modelo ViT para clasificar plantas sanas y enfermas. El rendimiento alcanzado en las pruebas fue de 90.99%.

El objetivo principal de este estudio es explorar y evaluar la arquitectura de los Transformadores de Visión (ViT) para detectar y clasificar enfermedades en las hojas de papa. Entre las afectaciones más graves en este cultivo, nos enfocamos en el tizón temprano y el tizón tardío, cuyas afectaciones son devastadoras para la calidad y producción de la papa. A diferencia de la mayoría de las investigaciones que emplean redes neuronales convolucionales, este trabajo propone la hipótesis de que el modelo ViT puede ofrecer un rendimiento superior, demostrando cómo, incluso con datos limitados se pueden lograr resultados significativos en la clasificación de enfermedades foliares.

2. Metodología

La revolución causada por el aprendizaje profundo en aplicaciones de visión por computadora ha influido significativamente en el avance tecnológico agrícola. El enfoque propuesto en este estudio sigue una secuencia de pasos esenciales. Primero, se recopilan imágenes de hojas sanas y de muestras con afectaciones de tizón temprano y tizón tardío. Posteriormente, se aplican técnicas de aumento de datos (data augmentation) para equilibrar el número de imágenes en cada clase y prevenir el sobreajuste. A continuación, se realiza el preprocesamiento de las imágenes para su adaptación a la arquitectura ViT y se ajustan los hiperparámetros del modelo para optimizar su rendimiento en la clasificación. Una vez configurado el modelo se procede al entrenamiento y prueba. Finalmente, se calculan las métricas de evaluación para medir los resultados y cuantificar el rendimiento.

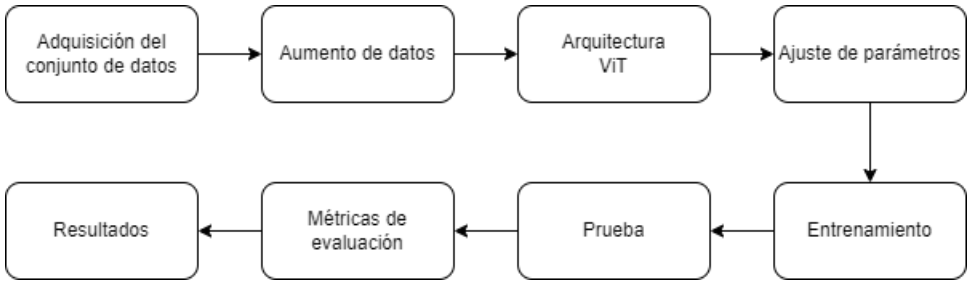

Figura 1: Diagrama de bloques del modelo propuesto.

2.1. Adquisición del conjunto de datos

La base de datos utilizada en nuestro estudio proviene del proyecto PlantVillage (Mohanty et al., 2016), que ofrece una extensa colección de imágenes sobre numerosos cultivos y enfermedades. La perspectiva de acceso abierto y colaborativo de esta plataforma facilita a la comunidad científica realizar investigaciones de aprendizaje profundo aplicadas a la agricultura. El total de 54306 imágenes se encuentra dividido en 38 categorías, que incluyen ejemplares de plantas sanas y 26 tipos de enfermedades en diversas especies. Para nuestro estudio, se seleccionó un subconjunto de datos de hojas de papa que contiene 2152 imágenes y un tamaño de 40 MB. Estas imágenes se organizan en tres clases: saludable, tizón temprano y tizón tardío, cada una con una resolución de 256x256 píxeles en formato RGB. La distribución de imágenes por categoría se detalla en la Tabla 1.

Tabla 1: Cantidad de imágenes clasificadas por categoría.

Clases	Número de imágenes	Imágenes de entrenamiento	Imágenes de prueba
saludable	152	122	30
tizón temprano	1000	800	200
tizón tardío	1000	800	200
Total	2152	1722	430

La Figura 2 ilustra ejemplos de muestras incluidas en el conjunto de datos. En las imágenes se pueden observar las diferencias entre las afectaciones de las hojas de papa. La primera fila evidencia las principales características de hojas saludables; la segunda fila muestra lesiones en los bordes y manchas oscuras circulares causadas por el tizón temprano; la tercera fila presenta ejemplos de marchitamiento y decoloración provocados por el tizón tardío.

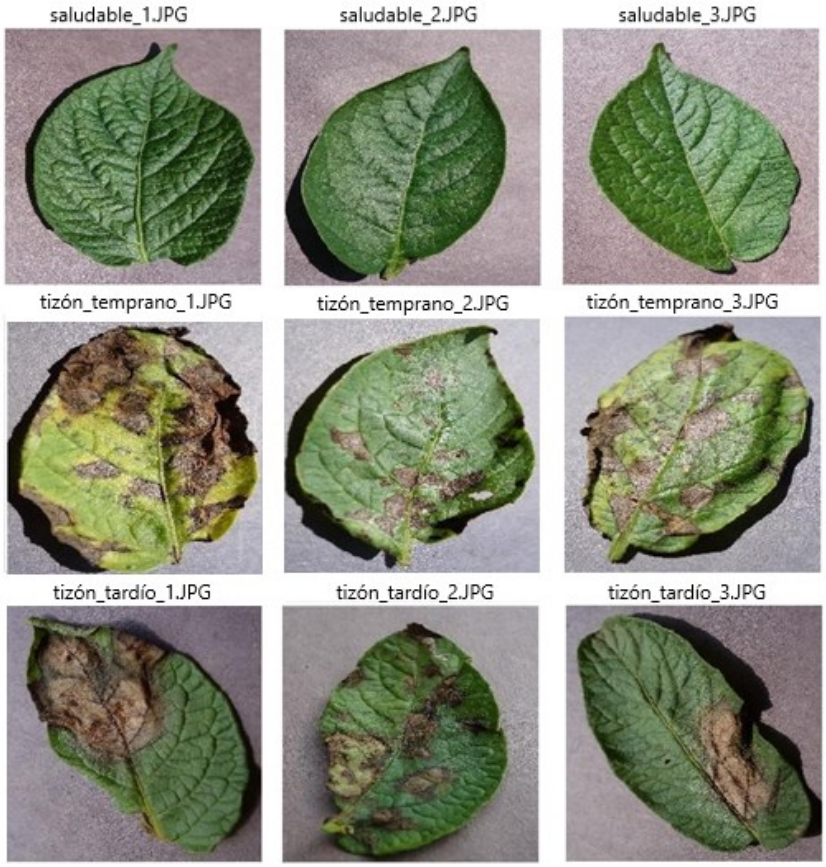

Figura 1: Ejemplos de muestras seleccionadas de cada clase del conjunto de datos

2.2. *Aumento de datos*

En visión artificial las técnicas de aumento de datos son fundamentales para incrementar su diversidad y volumen. En el conjunto de imágenes utilizado, la clase saludable presenta una menor cantidad de muestras, lo cual provoca un desbalance en la distribución total. Para evitar posibles problemas en la generalización del modelo, se utilizaron técnicas de aumento de datos mediante el método ImageDataGenerator de la biblioteca Keras. Las técnicas utilizadas fueron: volteo horizontal, volteo vertical, rotación de imagen (20 grados), adición de ruido (factor=0.1), ajuste de brillo (factor entre 0.8 y 1.2)

y escalamiento (factor entre 0.5 y 1.5). Esta estrategia de enriquecimiento de datos no solo compensa la falta de diversidad, sino que también ayuda a prevenir el sobreajuste, fortaleciendo la capacidad del modelo para generalizar. La configuración de imágenes modificadas se estableció en un 80% para el conjunto de entrenamiento y un 20% para el conjunto de prueba.

Tabla 2: Imágenes aumentadas clasificadas en entrenamiento y prueba

Clases	**Conjunto de datos**	**Conjunto de entrenamiento (80%)**	**Conjunto de prueba (20%)**
saludable	1000	800	200
tizón temprano	1000	800	200
tizón tardío	1000	800	200
Total	3000	2400	600

2.3. *Arquitectura de Transformadores de Visión (ViT)*

En el procesamiento del lenguaje natural (PLN), la arquitectura de Transformadores se ha establecido como el nuevo estándar para manejar secuencias de datos (Vaswani et al., 2017). Recientemente, el enfoque de autoatención está revolucionando el campo de la visión artificial demostrando que la investigación en esta área puede prescindir de las redes neuronales convolucionales. Un transformador puro, aplicado a una secuencia de parches de imágenes puede obtener excelentes resultados en tareas de clasificación. El presente estudio basa la clasificación de enfermedades de la papa en el modelo ViT base propuesto por Dosovitskiy et al. (2021). La Figura 3 muestra la estructura del modelo Transformador de Visión.

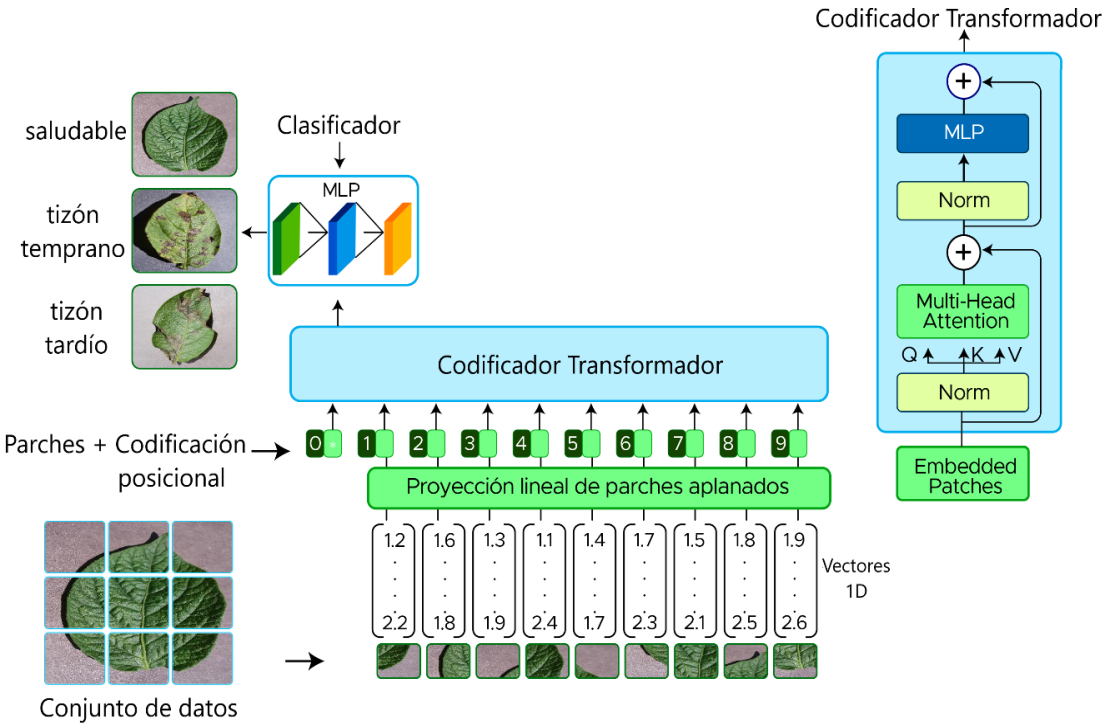

Figura 2: Arquitectura del modelo Transformador de Visión

En la fase inicial, el modelo recibe como entrada imágenes bidimensionales (2D) que son divididas en parches cuadrados de 16x16 píxeles. Consecuentemente, los parches son aplanados y convertidos en vectores unidimensionales (1D). La secuencia de entrada al codificador Transformador, por lo tanto, consiste en vectores numéricos unidimensionales conocidos como *tokens*.

Para una imagen de entrada, $(H \times W)$ es la resolución de la imagen original y C es el número de canales.

$$x \in \mathbb{R}^{H \times W \times C} \tag{1}$$

Para un parche de tamaño P, se crean N parches de la imagen.

$$x_p \in \mathbb{R}^{N \times (P \times P \times C)} \tag{2}$$

Posteriormente, los parches aplanados (vectores 1D) se introducen en la capa de proyección lineal, que opera de manera similar a una red neuronal convolucional. En esta capa, cada valor de color de los píxeles se asigna a neuronas específicas, generando a través de transformaciones lineales un vector de dimensiones reducidas. Este proceso consta de dos operaciones principales: la multiplicación de la entrada por una matriz de pesos, seguida de la adición de un vector de sesgo (bias). Los pesos y sesgos se aprenden y optimizan durante el proceso de entrenamiento. La reducción dimensional del vector resultante permite extraer características esenciales y capturar la información más relevante en la representación del objeto, mientras que simultáneamente se eliminan variaciones irrelevantes en los datos. En la siguiente fase, se añaden incrustaciones de posición (position embeddings) a los parches procesados. Estas incrustaciones, construidas mediante una combinación de funciones sinusoidales y cosenoidales, proporcionan al modelo información crucial sobre la ubicación espacial de cada parche en la imagen original. Este enfoque facilita el aprendizaje del modelo para atender eficientemente a las posiciones relativas dentro de la imagen.

El bloque de Atención Multicabeza (Multi-head Attention) consiste en varias capas de autoatención ejecutándose en paralelo siendo el componente principal para que el modelo entienda la relación entre los parches de una imagen y la imagen completa. La función de atención calcula una puntuación de similitud entre una consulta (Query) y un conjunto de pares clave-valor (Key-Value). Estas puntuaciones se obtienen mediante el producto punto de los vectores de consulta con todos los vectores clave, aplicando una función softmax para normalizar los resultados.

Q, K, V son matrices y d_k es la dimensión de queries and keys.

$$Attention\ (Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3}$$

En modelo mejora su rendimiento al ejecutar h operaciones de autoatención simultáneas, conocidas como bloque Multi-head Attention. Estas operaciones se concatenan y se proyectan linealmente, generando como resultado los valores de salida de ese bloque. De esta manera, se logra que el modelo preste atención a las características principales de diferentes subespacios de representación en diferentes posiciones.

$$Multi - head(Q, K, V) = Concat(head_1, \ldots.. head_h)W^O \tag{4}$$

En (5), W^Q, W^K, W^V son matrices de pesos aprendidos durante el entrenamiento y $W^O \in \mathbb{R}^{hd_v \times d_{model}}$.

El bloque de normalización (Norm) facilita el flujo de información aportando estabilidad y eficiencia al entrenamiento. Este proceso ajusta y reduce dinámicamente el impacto de las variaciones internas del modelo, mejorando su capacidad de generalización (Ba et al., 2016). Por último, el bloque Perceptrón Multicapa (Multi-layer Perceptron, MLP) permite aprender relaciones complejas y no lineales de los datos, potenciando significativamente el rendimiento en la clasificación en cada una de las categorías.

2.4. Ajuste de parámetros

En este estudio se utilizó la biblioteca torchvision de Pytorch, la cual ofrece modelos y pesos preentrendados. De las versiones propuestas por Dosovitskiy et al. (2021), se eligió el modelo base VIT-B/16 para la clasificación de enfermedades en hojas de papa, con un tamaño de parche de 16x16 píxeles. El bloque de clasificación MLP se configuró con tres categorías: saludable, tizón temprano y tizón tardío. Las características de entrada del modelo son: tamaño del lote (batch size) = 32, canales de color (RGB) = 3, alto = 224 píxeles, ancho = 224 píxeles. En la Tabla 3 se propone un resumen que detalla la arquitectura del modelo VIT-B/16 y el ajuste fino realizado.

Tabla 3: Resumen de capas y ajuste de parámetros del modelo VIT-B/16.

Capas	Entrada	Salida	Parámetros
VisionTransformer	[32, 3, 224, 224]	[32, 3]	768
Conv2d	[32, 3, 224, 224]	[32, 768, 14, 14]	(590,592)
Encoder (encoder)	[32, 197, 768]	[32, 197, 768]	151,296
Dropout (dropout)	[32, 197, 768]	[32, 197, 768]	--
EncoderBlock (0 - 12)	[32, 197, 768]	[32, 197, 768]	7,087,872
LayerNorm (ln)	[32, 197, 768]	[32, 197, 768]	1536
Linear (heads)	[32, 768]	[32, 3]	2,307

El modelo cuenta con un total de 85 millones de parámetros preentrenados utilizando el conjunto de datos ImageNet-1K. La mayoría de estos parámetros no se ajustan durante el entrenamiento para preservar la extracción de características aprendidas; sin embargo, 2307 de estos parámetros son entrenables. La arquitectura consta de 12 capas y 768 dimensiones de características. En los bloques codificadores, la información se procesa secuencialmente y al conjunto de 196 parches se le agrega un token de clasificación, lo que resulta en una dimensión final de 197 tokens. Para prevenir el sobreajuste se implementó la técnica de regularización (dropout), mientras que la capa de normalización (LayerNorm) estabiliza el aprendizaje y mejora la convergencia durante el entrenamiento. Además, se ajustó la última capa para que sea entrenable, permitiendo clasificar cada lote de 32 imágenes en una de las 3 clases correspondientes.

2.5. Métricas de evaluación

Para ejecutar y evaluar el modelo se utilizó el siguiente hardware: procesador Intel(R) Core(TM) i7-6500U CPU @ 2.50GHz (4 CPUs), ~2.6GHz con 12 GB de RAM, tarjeta gráfica

NVIDIA GeForce 920M y sistema operativo Windows 10 de 64 bits. Los algoritmos fueron ejecutados en el framework PyTorch 2.2.1+cu121 utilizando Python 3.10.12. El modelo fue entrenado durante 10 épocas (epochs) y utilizando el optimizador Adam con una tasa de aprendizaje (learning rate) de 1×10^{-3}.

El estudio se evaluó en base a las siguientes métricas: exactitud (accuracy), precisión (precision), sensibilidad (recall), puntaje F1 (F1-score). Estas métricas se basan en los siguientes conceptos, Verdaderos Positivos (True Positives, TP), Verdaderos Negativos (True Negatives, TN), Falsos Positivos (False Positives, FP) y Falsos Negativos (False Negatives, FN).

$$Exactitud = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (6)$$

$$Precisión = \frac{TP}{TP + FP} \times 100\% \quad (7)$$

$$Sensibilidad = \frac{TP}{TP + FN} \times 100\% \quad (8)$$

$$Puntaje\ F1 = 2 \times \left(\frac{Precisión \times Sensibilidad}{Precisión + Sensibilidad}\right) \times 100\% \quad (9)$$

3. Resultados

La Figura 4, ilustra las curvas de pérdida y precisión del modelo Transformador de visión (ViT). El eje y representa los valores de pérdida, mientras que el eje x muestra el número de épocas. Durante el proceso de entrenamiento, se observa una progresiva evolución del rendimiento durante las iteraciones de cada época. En la etapa inicial, se registra una pérdida de entrenamiento de 0.3239, en contraste con una pérdida de prueba de 0.1545. El modelo exhibe una rápida convergencia, estabilizándose en la décima época con una reducción significativa de la pérdida de entrenamiento a 0.0178, mientras que la pérdida de prueba alcanza un valor de 0.0357. Los resultados sugieren una generalización eficaz minimizando el sobreajuste. En términos de exactitud, el modelo presenta un valor de 99.92% en el conjunto de entrenamiento y logra un notable 99.18% en el conjunto de prueba, evidenciando una elevada eficacia en la clasificación de enfermedades en hojas de papa.

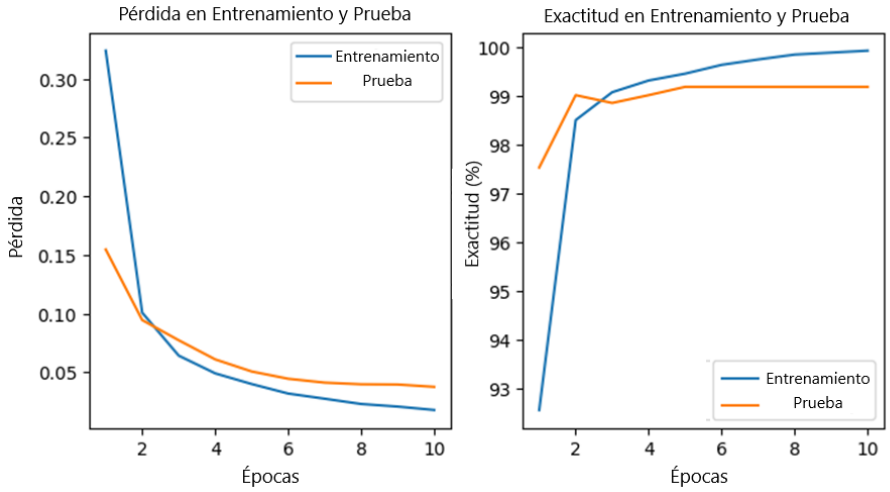

Figura 4: a) Pérdida de entrenamiento y prueba b) Exactitud de entrenamiento y prueba en porcentaje

La Tabla 4 muestra el reporte de clasificación completo con las métricas propuestas en (6), (7), (8) y (9). Los resultados revelan valores equilibrados y eficientes, lo cual se traduce en una clasificación altamente precisa en cada una de las clases.

Tabla 4: Comparación de las métricas de evaluación para cada clase propuesta

Clase	**Precisión**	**Sensibilidad**	**Puntaje F1**	**Exactitud**
saludable	0.975	100	0.987	
tizón temprano	1.00	1.00	1.00	99.18%
tizón tardío	1.00	0.975	0.987	

Para evaluar el rendimiento del modelo en la predicción de 600 muestras de prueba, se utilizó la herramienta matriz de confusión. La Figura 5 ilustra la distribución de predicciones para las categorías: saludable, tizón temprano y tizón tardío, contrastando los valores verdaderos con los predichos. Las celdas diagonales de la matriz representan las clasificaciones correctas para cada clase, mientras que los elementos fuera de la diagonal indican el número de predicciones erróneas. El modelo clasifica acertadamente las 200 instancias de las clases saludable y tizón temprano, con ausencia total de falsos positivos y negativos en estas categorías. La única confusión se observa en la clase tizón tardío, donde cinco instancias fueron erróneamente clasificadas como saludables, lo que, sugiere una ligera dificultad del modelo para distinguir etapas iniciales de estas enfermedades.

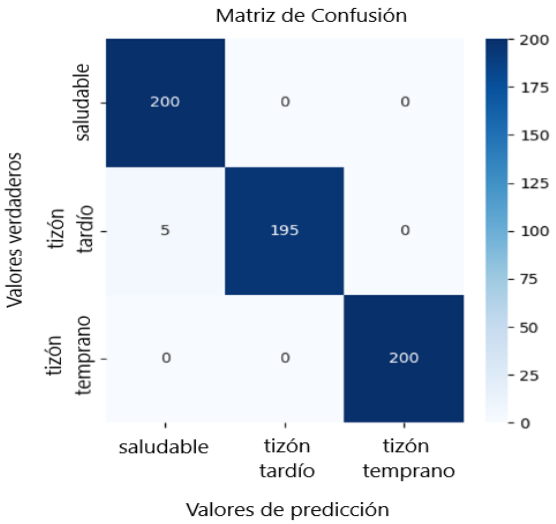

Figura 5: Evaluación del modelo mediante la matriz de confusión para la clasificación de imágenes en las clases: saludable, tizón temprano, tizón tardío.

La Tabla 5 presenta un análisis comparativo del rendimiento de diversas arquitecturas de redes neuronales convolucionales (CNNs) y el modelo Transformador de Visión (ViT) en la tarea de clasificación de enfermedades foliares en papa. Los modelos evaluados incluyen las arquitecturas InceptioV3, ResNet50, VGG16, VGG19 y ViT. Todas las arquitecturas fueron sometidas a condiciones idénticas de preprocesamiento de imágenes y entrenadas mediante técnicas de aprendizaje por transferencia.

Tabla 5: Comparación de rendimiento entre arquitecturas de CNN y ViT para la clasificación de enfermedades foliares en papa

Modelo	Parámetros (M)	Épocas	Exactitud en prueba
InceptionV3	38.5	30	97.17%
ResNet50	36.4	30	96.53%
VGG16	17.9	30	90.28%
VGG19	23.2	30	77.78%
ViT	85.8	10	99.18%

4. Discusión

El presente estudio tuvo como objetivo clasificar enfermedades de las hojas de papa evaluando el modelo de Transformadores de Visión (ViT). Los resultados obtenidos demuestran un rendimiento superior de la arquitectura ViT en comparación con las arquitecturas CNNs. La Figura 4(a) muestra una rápida convergencia del modelo, lo cual es indicativo de una eficiente capacidad de aprendizaje desde las primeras épocas. La reducción significativa de la pérdida en el conjunto de entrenamiento, junto con la disminución de la pérdida en el conjunto de prueba hasta 0.0357, señala una robusta capacidad de generalización del modelo. En la Figura 4(b) se observa una alta exactitud en el conjunto de prueba (99.18%), que se aproxima estrechamente a la exactitud del conjunto de entrenamiento (99.92%). Esta reducida diferencia sugiere que el modelo no solo ha aprendido a clasificar correctamente las imágenes de entrenamiento, sino que también mantiene un desempeño excepcional en datos no vistos previamente. La congruencia en las

curvas de pérdida y la alta exactitud en ambos conjuntos sugieren una regulación adecuada del modelo, evitando el sobreajuste común en arquitecturas complejas de aprendizaje profundo.

El análisis de la matriz de confusión revela que el modelo clasifica correctamente todas las instancias de las clases saludable y tizón temprano, sin cometer errores en estas categorías. En la clase tizón tardío, se observa una confusión mínima donde cinco instancias fueron incorrectamente clasificadas como saludables. Estos resultados demuestran el significativo desempeño obtenido mediante el modelo ViT. El procesamiento de los parches de las imágenes facilita la captura de patrones morfológicos característicos en las muestras, lo que sugiere que el mecanismo de atención empleado presenta mayor efectividad en la detección de enfermedades foliares. El reporte de métricas de evaluación detallado en la Tabla 4, subraya la eficacia de ViT frente a otras arquitecturas. La precisión, sensibilidad y el puntaje F1 son elevados en todas las clases, indicando una clasificación altamente precisa y balanceada.

Para validar el rendimiento del modelo ViT, la Tabla 5 muestra la comparación de la exactitud en la clasificación de imágenes en las tres clases propuestas (saludable, tizón temprano, tizón tardío) en relación con diferentes arquitecturas de CNNs. El modelo InceptionV3 entrenado en 30 épocas y con 38.5 millones de parámetros, alcanzó la exactitud máxima en prueba de 97.17%. Los valores obtenidos señalan que, a pesar de que el modelo ViT requiere una mayor cantidad de datos para su entrenamiento, con técnicas de aprendizaje por transferencia en 10 épocas superó a InceptionV3 en términos de exactitud y capacidad de generalización. El modelo ResNet50 alcanzó una exactitud de 96.53% y el modelo VGG16 un 90.28%. El modelo VGG19 obtuvo la menor exactitud en la validación. Estos resultados indican que el modelo ViT presenta un mejor desempeño en la captura de relaciones globales en las imágenes, lo cual mejora la flexibilidad y adaptabilidad.

Al comparar los resultados de las métricas obtenidas en nuestro estudio con investigaciones similares, se observa que Lozada-Portilla et al. (2021) utilizaron distintas arquitecturas de CNNs para evaluar el rendimiento en la clasificación de hojas de papa infectadas con tizón tardío alcanzando una exactitud máxima del 90%. En la investigación realizada por Shaheed et al. (2023) se evaluaron distintas técnicas de aprendizaje por transferencia y se propuso la integración de un nuevo modelo que combina la arquitectura ResNet50 y ViT, alcanzando una exactitud del 99,12% en un conjunto de datos de hojas de papa especializado. El marco híbrido presentado en la investigación de Arshad et al. (2023) logró una precisión general del 98,66% al combinar características profundas de los modelos VGG19 e InceptionV3 en la predicción de enfermedades en cultivos de papa. Basados en la revisión de estos estudios y respaldados por los resultados obtenidos, se evidencia que la arquitectura propuesta presenta un rendimiento superior en la detección de enfermedades en hojas de papa.

Si bien el modelo presenta un excelente desempeño, es preciso reconocer sus limitaciones. El conjunto de datos de PlantVillage no presenta imágenes de hojas en un escenario o campo agrícola natural. La diversidad entre diferentes condiciones climáticas y variedades de papas son aspectos que aún requieren un mayor enfoque. Para investigaciones futuras, se pretende ampliar el conjunto de datos incluyendo una mayor variabilidad en las características de las imágenes y medir el rendimiento del modelo en estas condiciones.

5. Conclusiones

En este estudio, se utilizó el modelo Transformador de Visión (ViT) para la detección y clasificación de enfermedades en las hojas de papa. Los hallazgos confirman que el preprocesamiento de imágenes mediante parches permite aprender y diferenciar eficazmente entre muestras sanas e infectadas a través de mecanismos de atención. El sobresaliente rendimiento del codificador transformador propuesto reduce la dependencia tradicional de las redes neuronales convolucionales (CNNs) para tareas visuales complejas, abriendo una nueva ruta tecnológica para enfrentar los desafíos agrícolas en la detección automatizada de enfermedades. Las técnicas de aprendizaje por transferencia y ajuste fino demuestran ser una solución altamente efectiva en tareas específicas de visión por computadora orientadas al sector agrícola.

Para futuras investigaciones, se sugiere partir de un conjunto de datos ampliado y balanceado, que incluya muestras de una mayor variedad de enfermedades. Esto permitirá aplicar diferentes ajustes y técnicas a la arquitectura, contribuyendo a su aplicación práctica. Concluimos destacando la relevancia de las técnicas de aprendizaje profundo como herramientas innovadoras y funcionales para los agricultores. Los resultados obtenidos con el modelo ViT consolidan estrategias que abordan y mitigan los desafíos del sector agrícola, proporcionando soluciones efectivas que mejoran la productividad y detección de enfermedades en cultivos de papa.

Contribuciones de los autores

En concordancia con la taxonomía establecida internacionalmente para la asignación de créditos a autores de artículos científicos (https://casrai.org/credit/). Los autores declaran sus contribuciones en la siguiente matriz:

	Valle, F.	Castillo, L.	Correa, M.	Guzmán, J.
Conceptualización	■			
Análisis formal			■	
Investigación	■			
Metodología		■		
Recursos				■
Validación			■	
Redacción – revisión y edición		■		

Conflicto de Interés

Los autores manifiestan que no existe ningún tipo de conflicto de interés, ya sea, financiero, personal o académico, que pueda influir en los resultados y conclusiones de este estudio.

Referencias

Arshad, F., Mateen, M., Hayat, S., Wardah, M., Al-Huda, Z., Gu, Y. H., & Al-antari, M. A. (2023). PLDPNet: End-to-end hybrid deep learning framework for potato leaf disease prediction. *Alexandria Engineering Journal*, *78*, 406-418. https://doi.org/10.1016/J.AEJ.2023.07.076

Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer Normalization. *Statistics – Machine Learning*. https://doi.org/10.48550/arXiv.1607.06450

Barman, U., Sarma, P., Rahman, M., Deka, V., Lahkar, S., Sharma, V., & Saikia, M. J. (2024). ViT-SmartAgri: Vision Transformer and Smartphone-Based Plant Disease Detection for Smart Agriculture. *Agronomy*, *14*(2), 327. https://doi.org/10.3390/AGRONOMY14020327

Barrera, V., Escudero, L., Norton, G. W., & Sherwood, S. (2002). Validación y difusión de modelos de manejo integrado de plagas y enfermedades en el cultivo de la papa: Una experiencia de capacitación participativa en la Provincia del Carchi, Ecuador. *Revista Informativa INIAP*, *16*, 25-28. http://repositorio.iniap.gob.ec/handle/41000/1493

Comisión Económica para América Latina y el Caribe [CEPAL], Organización de las Naciones Unidas para la Agricultura y la Alimentación [FAO], & Instituto Interamericano de Cooperación para la Agricultura [IICA]. (2021). *Perspectivas de la Agricultura y del Desarrollo Rural en las Américas: una mirada hacia América Latina y el Caribe 2021-2022*. IICA. https://hdl.handle.net/11362/47208

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L., (2009). ImageNet: A large-scale hierarchical image database. *IEEE Conference on Computer Vision and Pattern Recognition,* FL, USA. https://doi.org/10.1109/CVPR.2009.5206848

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, G., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *Computer Vision and Pattern Recognition.* https://doi.org/10.48550/arXiv.2010.11929

Elngar, A. A., Arafa, M., Fathy, A., Moustafa, B., Mahmoud, O., Shaban, M., & Fawzy, N. (2021). Image Classification Based On CNN: A Survey. *Journal of Cybersecurity and Information Management (JCIM)*, *6*(1), 18-50. https://doi.org/10.54216/JCIM.060102

Ihme, M., Chung, W. T., & Mishra, A. A. (2022). Combustion machine learning: Principles, progress and prospects. *Progress in Energy and Combustion Science*, *91*. https://doi.org/10.1016/j.pecs.2022.101010

Khan, A., Rauf, Z., Sohail, A., Rehman, A., Asif, H., Asif, A., & Farooq, U. (2023). A survey of the vision transformers and their CNN-transformer based variants. *Artificial Intelligence Review*, *56*, 2917-2970. https://doi.org/10.1007/s10462-023-10595-0

Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521, 436-444. https://doi.org/10.1038/nature14539

López Calvajar, G. A., López Fernández, R. C., & León González, J. L. (2017). Análisis de la influencia de factores climatológicos en la pérdida de superficie sembrada de cultivos transitorios en el Ecuador. Revista Científica Agroecosistemas, 5(3), 176-183. https://aes.ucf.edu.cu/index.php/aes/article/view/155

Lozada-Portilla, W. A., Suarez-Barón, M. J., & Avendaño-Fernández, E. (2021). Aplicación de redes neuronales convolucionales para la detección del tizón tardío Phytophthora infestans en papa Solanum tuberosum. *Revista U.D.C.A Actualidad & Divulgación Científica*, *24*(2), e1917. https://doi.org/10.31910/rudca.v24.n2.2021.1917

Mahum, R., Munir, H., Mughal, Z. U. N., Awais, M., Sher Khan, F., Saqlain, M., Mahamad, S., & Tlili, I. (2022). A novel framework for potato leaf disease detection using an efficient deep learning model. *Human and Ecological Risk Assessment: An International Journal*, *29*(2), 303–326. https://doi.org/10.1080/10807039.2022.2064814

Ministerio de Agricultura y Ganadería. (2022). *Boletín situacional cultivo de papa.* https://sipa.agricultura.gob.ec/index.php/papa/boletines-situacionales-papa-ecuador

Mohanty, S. P., Hughes, D. P., & Salathé, M. (2016). Using deep learning for image-based plant disease detection. *Frontiers in Plant Science*, *7*, 1419. https://doi.org/10.3389/fpls.2016.01419

Paucar Buñay, D. J. (2016). *Factores que influyen en el nivel de conocimientos ancestrales en el manejo del cultivo de papa (Solanum tuberosum) en dos sectores de la provincia de Tungurahua* [Tesis de grado, Universidad Técnica de Ambato]. Repositorio Universidad Técnica de Ambato. https://repositorio.uta.edu.ec:8443/jspui/handle/123456789/24431

Pavel, M. I., Rumi, R. I., Fairooz, F., Jahan, S., & Hossain, M. A. (2021). Deep Residual Learning Approach for Plant Disease Recognition. *International Conference on Mobile Computing and Sustainable Informatics*, 511-521. https://doi.org/10.1007/978-3-030-49795-8_49

Pérez, W., & Forbes, G. (2016). Guía de identificación de plagas que afectan a la papa en la zona andina- Lima (Perú). *Centro Internacional de la Papa (CIP)*. https://doi.org/10.4160/9789290604020

Pintelas, E., Liaskos, M., Livieris, I. E., Kotsiantis, S., & Pintelas, P. (2020). Explainable machine learning framework for image classification problems: Case study on glioma cancer prediction. *Journal of Imaging*, *6*(6), 37. https://doi.org/10.3390/JIMAGING6060037

Sakkarvarthi, G., Sathianesan, G. W., Murugan, V. S., Reddy, A. J., Jayagopal, P., & Elsisi, M. (2022). Detection and Classification of Tomato Crop Disease Using Convolutional Neural Network. *Electronics*, *11*(21), 3618. https://doi.org/10.3390/electronics11213618

Sen, P. C., Hajra, M., & Ghosh, M. (2020). Supervised Classification Algorithms in Machine Learning: A Survey and Review. *Advances in Intelligent Systems and Computing*, *937*, 99-111. https://doi.org/10.1007/978-981-13-7403-6_11

Shaheed, K., Qureshi, I., Abbas, F., Jabbar, S., Abbas, Q., Ahmad, H., & Sajid, M. Z. (2023). EfficientRMT-Net—An Efficient ResNet-50 and Vision Transformers Approach for Classifying Potato Plant Leaf Diseases. *Sensors, 23*(23), 9516. https://doi.org/10.3390/S23239516

Shirahatti, J., Patil, R., & Akulwar, P. (2018). A Survey Paper on Plant Disease Identification Using Machine Learning Approach. *Proceedings of the 3rd International Conference on Communication and Electronics Systems, ICCES*, 1171-1174. https://doi.org/10.1109/CESYS.2018.8723881

Sun, C., Shrivastava, A., Singh, S., & Gupta, A. (2017). Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. *Computer Vision and Pattern Recognition*. https://doi.org/10.48550/arXiv.1707.02968

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Lukasz, K., & Polosukhin, I. (2017). Attention Is All You Need. *Computation and Language.* https://doi.org/10.48550/arXiv.1706.03762