Detección y Clasificación de Patógenos Foliares mediante Visión por Computador

Gabriel Sánchez Muñoz
gabrielsm@correo.ugr.es

Germán Rodríguez Vidal
germanrv@correo.ugr.es

Pablo García Bas
pablogb@correo.ugr.es

Miguel Ángel Moreno Castro
miguelangelmc@correo.ugr.es

Abstract

La detección temprana y precisa de enfermedades vegetales es crítica para la agricultura de precisión y la seguridad alimentaria. Los enfoques actuales basados en Deep Learning, a pesar de mostrar resultados prometedores, suelen carecer de información contextual sobre el huésped, lo que limita su precisión en escenarios de múltiples cultivos. Este trabajo propone un enfoque jerárquico de dos etapas para abordar esta limitación. En primer lugar, se realiza una clasificación de la especie arbórea para establecer un contexto biológico y, posteriormente, un detector de objetos localiza y clasifica los patógenos específicos asociados a dicha especie. Los resultados muestran que incorporar la identificación previa del huésped incrementa la precisión media (mAP) en un [X]% respecto a los detectores genéricos de una sola etapa ("end-to-end"), demostrando la superioridad de los modelos condicionados biológicamente.

1. Introducción

Las enfermedades de las plantas representan una amenaza significativa para la seguridad alimentaria mundial y la estabilidad económica agrícola, causando pérdidas estimadas de hasta el 30% en el rendimiento de los cultivos cada año [8]. La identificación temprana y precisa de estas patologías es crucial para aplicar medidas de control eficaces y minimizar el uso de químicos. Tradicionalmente, este diagnóstico ha dependido de la inspección visual manual por parte de expertos, un proceso que resulta laborioso, subjetivo y difícil de escalar a grandes explotaciones [1].

En la última década, la visión por computador y, específicamente, el aprendizaje profundo, han emergido como herramientas poderosas para automatizar esta tarea. El uso de Redes Neuronales Convolucionales (CNNs) ha permitido grandes avances tanto en la clasificación de imágenes (identificar si una hoja está enferma) como en la detección de objetos (localizar la lesión exacta en la hoja) [5]. Sin embargo, la mayoría de las arquitecturas actuales abordan la detección de enfermedades como un problema monolítico donde los modelos se entrenan para detectar cualquier enfermedad en cualquier tipo de hoja simultáneamente, ignorando a menudo la estructura jerárquica natural de la taxonomía biológica.

Debido a que muchas patologías foliares comparten características visuales similares, independientemente de la especie, los detectores agnósticos al huésped sufren de una alta confusión inter-clase. Esto suele derivar en predicciones erróneas donde se asocian enfermedades a plantas incompatibles, reduciendo la precisión y confiabilidad del sistema en aplicaciones del mundo real.

Para resolver esto, proponemos un enfoque jerárquico que imita el diagnóstico experto donde primero se identifica la especie del árbol para simplificar la búsqueda de la enfermedad y luego se aplica un detector especializado para esa especie.

2. Dataset

Para validar nuestro enfoque jerárquico, hemos confeccionado un *dataset* que integra imágenes de tres cultivos: rosa, patata y manzana. Las imágenes provienen de la plataforma Roboflow Universe, que contiene una gran cantidad de *datasets* de código abierto, y han sido seleccionadas por su calidad y variedad de condiciones de iluminación.

- Rose Dataset: Consta de 2,725 imágenes y abarca 4 clases (Black Spot, Downy Mildew, Powdery Mildew y hojas sanas).
- Potato Dataset: Comprende 812 imágenes distribuidas en [X] clases (típicamente Early Blight, Late Blight, Healthy).
- Apple Dataset: Incluye 1582 imágenes que cubren 4 patologías (Alternaria Spot, Brown Spot, Gray Spot,

Rust).

El dataset combinado se dividió aleatoriamente en subconjuntos de Entrenamiento (70%), Validación (15%) y Prueba (15%), preservando la estratificación de clases para garantizar una evaluación equilibrada.

3. Clasificación

Para el diseño del módulo de clasificación, hemos evaluado dos paradigmas principales del estado del arte, las arquitecturas basadas en Vision Transformers (ViT) y las Redes Neuronales Convolucionales (CNNs). Nuestro objetivo fue contrastar la capacidad de modelado global de los Transformers frente a la eficiencia inductiva de las CNNs. Adicionalmente, dada la orientación práctica de este proyecto hacia una futura aplicación móvil, se incluyó en el estudio un modelo diseñado específicamente para entornos de recursos limitados.

3.1. MaxViT

Dentro de la familia de los Transformers, MaxViT (Multi-Axis Vision Transformer) combina tanto mecanismos de atención global como convoluciones locales para capturar características a múltiples escalas [10]. Su característica distintiva es el mecanismo de *Multi-Axis Self-Attention* (Max-SA), que descompone el cálculo de atención en dos operaciones: atención local (*Block Attention*) para capturar texturas finas, y atención global dispersa (*Grid Attention*) para relacionar partes distantes de la imagen.

3.1.1 EfficientNetV2

Esta arquitectura mejora a su predecesora mediante la introducción de bloques Fused-MBConv [9]. Estos bloques reemplazan las convoluciones *depthwise* separables tradicionales por convoluciones estándar 3×3 fusionadas en las primeras capas.

3.2. MobileNet

La base de su eficiencia radica en las convoluciones *depthwise* separables, que factorizan la operación de convolución estándar en dos capas más ligeras (profundidad y punto a punto), reduciendo drásticamente la cantidad de parámetros y operaciones [4]. Aunque su capacidad de representación es menor que los modelos anteriores, su inclusión es crítica para evaluar el compromiso (*trade-off*) entre precisión y latencia en una aplicación real para agricultores.

4. Detección de Objetos

El objetivo fundamental en detección de objetos es localizar y clasificar regiones de interés (*bounding boxes*). Históricamente, métodos dominantes como *Deformable Parts Model* (DPM) abordaban este problema mediante un enfoque de ventana deslizante (*sliding window*) que resultaban inviables computacionalmente al procesar todas las posibles sub-ventanas [2].

La evolución comenzó con **R-CNN**, un modelo de 3 etapas donde una búsqueda selectiva genera regiones potencialmente interesantes, una CNN extrae características y un SVM las clasifica [3]. Sus limitaciones principales son el alto coste de procesar regiones por separado y la falta de aprendizaje en la búsqueda selectiva.

4.1. Faster R-CNN

Faster R-CNN integra la propuesta de regiones dentro de la red neuronal, reemplazando el algoritmo fijo por una *Region Proposal Network* (RPN) entrenable y más rápida [7].

En primer lugar se definen *anchor points* sobre el mapa de características de la imagen, actuando como centros de posibles regiones con diferentes escalas y ratios. Seguidamente, la RPN predice simultáneamente la probabilidad de objeto y las coordenadas del *bounding box*.

A pesar de su alta precisión, el hecho de ser un modelo de dos etapas penaliza su velocidad de inferencia.

4.2. YOLO

Como respuesta a la latencia, *You Only Look Once* (YOLO) redefinió la detección no como una clasificación de regiones, sino como un problema único de regresión, descartando por completo el pipeline disjunto de DPM y R-CNN.

YOLO utiliza una única red neuronal convolucional que procesa la imagen completa de una sola vez. Divide la imagen en una cuadrícula ($S \times S$) donde cada celda predice B bounding boxes y sus respectivas puntuaciones de confianza.

La arquitectura unificada de YOLO permite una inferencia en tiempo real inalcanzable para los métodos de dos etapas [6]. Mientras que Faster R-CNN procesa aproximadamente 0.5 FPS, YOLO alcanza los 45 FPS, ofreciendo la inmediatez necesaria para aplicaciones agrícolas prácticas.

5. Métodos

En esta sección se detalla la propuesta arquitectónica diseñada para la detección de patologías foliares. El sistema se fundamenta en un enfoque jerárquico de dos etapas, compuesto secuencialmente por un módulo de clasificación taxonómica (especie vegetal) y un módulo de detección de objetos (patógeno y localización dentro de la imagen).

5.1. Arquitectura del Pipeline Propuesto

Para evaluar la eficacia del condicionamiento biológico en la detección de enfermedades, se implementaron y com-

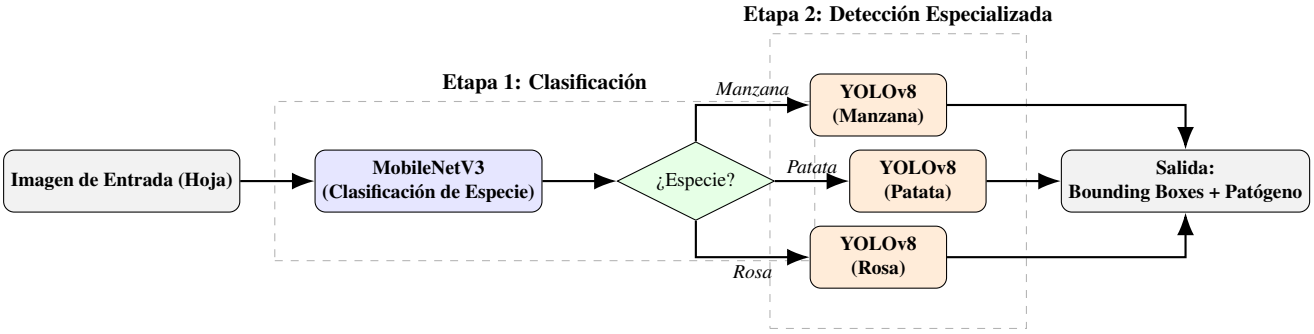

Figure 1. Visión general de la arquitectura jerárquica propuesta. Una imagen de entrada es procesada primero por una red MobileNetV3 para identificar el cultivo huésped (Etapa 1). En base a esta predicción, se selecciona dinámicamente un modelo de detección YOLOv8 especializado (Etapa 2) para localizar y clasificar las patologías foliares correspondientes, evitando la confusión inter-especie.

pararon tres variantes arquitectónicas del pipeline. Todas comparten la etapa inicial de clasificación, pero difieren en la estrategia de detección posterior:

- **V1: Arquitectura de Expertos Especializados (Ramificada).** En este esquema, la predicción del clasificador actúa como un enrutador (*router*). Dependiendo de la especie predicha, la imagen se deriva a uno de los tres detectores YOLO independientes, cada uno entrenado exclusivamente con el subconjunto de datos de dicha especie (Rosa, Patata o Manzana). Esto elimina la confusión inter-clase de especies a nivel de detector.
- **V2: Arquitectura Secuencial Global (Baseline).** El clasificador identifica la especie para aportar contexto, pero la detección se realiza mediante un único modelo YOLO global entrenado con el dataset completo (todas las clases de patógenos y especies simultáneamente). Este enfoque evalúa si la especialización (modelo V1) es realmente necesaria frente a un detector robusto generalista (modelo V2).
- **V3: Arquitectura Global Condicionada.** Una variante híbrida donde se utiliza un único detector global, pero la salida del clasificador se utiliza para filtrar o condicionar las predicciones del detector (e.g., penalizando la confianza de patógenos incompatibles con la especie detectada).

5.2. Selección del Clasificador: MobileNet

De los modelos evaluados en la fase preliminar (MaxViT, EfficientNetV2 y MobileNet), se seleccionó **MobileNet** [4] como el extractor de características para la primera etapa.

Esta decisión se justifica por el *trade-off* entre precisión y coste computacional. Los experimentos mostraron que, dada la baja cardinalidad del problema de clasificación (únicamente 3 clases: rosa, patata, manzana), modelos de alta capacidad como MaxViT no aportaban una ganancia significativa de precisión (saturación de rendimiento), mientras que introducían una latencia considerable. MobileNet, gracias a sus convoluciones separables en profundidad (*depthwise separable convolutions*), ofrece un rendimiento competitivo con una fracción de los parámetros, alineándose con el objetivo de despliegue en dispositivos móviles.

5.3. Selección del Detector: YOLOv8

Para la etapa de localización de patógenos, se realizó una comparativa entre detectores de dos etapas (**Faster R-CNN**) y de una etapa (**YOLO**). Finalmente, se optó por la arquitectura **YOLOv8**, específicamente la variante *small* (YOLOv8s).

La elección se fundamenta en dos factores críticos observados durante la experimentación:

1. **Generalización y Convergencia:** Faster R-CNN mostró una tendencia al sobreajuste (*overfitting*) temprano con nuestro dataset, degradando su rendimiento en el conjunto de test en comparación con YOLO bajo el mismo número de épocas de entrenamiento.
2. **Eficiencia de Inferencia:** Como se discutió en la introducción, la aplicabilidad en campo requiere tiempos de respuesta rápidos. YOLOv8s, al ser un detector *single-stage* libre de la red de propuesta de regiones (RPN), demostró una velocidad de inferencia (FPS) muy superior, esencial para la experiencia de usuario en tiempo real.

Finalmente, la elección de la arquitectura se acotó a dos candidatos (si bien en las pruebas se consideraron hasta 8 versiones distintas de YOLO): **YOLOv8s** y la reciente versión **YOLO11s**. Aunque ambos modelos mostraron un desempeño sobresaliente, la balanza se inclinó a favor de YOLOv8s por dos razones fundamentales:

1. **Consistencia en la precisión:** YOLOv8s demostró una mayor robustez inter-dominio. Superó a YOLO11s en el dataset de rosas con un margen de $+1.5\%$ de mAP y mantuvo un empate técnico en patatas, mientras que la ventaja de YOLO11s en manzanas fue marginal ($< 0.3\%$), no compensando su inestabilidad en otros escenarios.

2. **Eficiencia computacional:** Se observó un patrón constante de mayor eficiencia en la versión 8. YOLOv8s no solo requirió menores tiempos de entrenamiento para converger (e.g., 14.1 min vs 16.8 min en rosas), sino que, críticamente, ofreció una latencia de inferencia menor (2.66 img/s frente a 2.50 img/s de la v11s). Esta ganancia del 6.4% en velocidad es determinante para el despliegue en tiempo real propuesto.

En consecuencia, se selecciona **YOLOv8s** como el detector final, al ofrecer el mejor equilibrio entre estabilidad predictiva y agilidad de procesamiento.

6. Experimentos

Para la elección del mejor metodo para detección y clasificación de patógenos foliares, se ha utilizado el conjunto de datos Global Plant Dataset, un corpus heterogéneo , compuesto por los anteriores datsets mencionados , con ellos se han realizado diversos experimentos. A continuación se describen los experimentos realizados y los resultados obtenidos.

6.1. Metricas de Evaluación

Para la evaluación de los modelos de clasificación, se han utilizado las métricas de precisión, mAP 50 y mAP 50-95. La precisión mide la proporción de predicciones correctas sobre el total de predicciones realizadas. El mAP 50 (mean Average Precision at IoU 0.5) evalúa la precisión promedio considerando un umbral de Intersección sobre Unión (IoU) del 50%, mientras que el mAP 50-95 promedia la precisión en múltiples umbrales de IoU, proporcionando una evaluación más robusta del rendimiento del modelo.

Se diseñaron tres experimentos principales para comparar los enfoques de detección y clasificación:

6.2. Arquitecturas del pipeline implementadas

El primer experimento (Baseline Monolitico) consiste en entrenar un único modelo de detección de objetos (YOLOv8) para identificar todas las clases de patógenos foliares sin considerar la especie del árbol huésped. Este modelo debe distinguir simultáneamente entre todas las enfermedades de todas las plantas, representando el enfoque estándar "end-to-end". Este enfoque sirve como línea base para evaluar la efectividad de los métodos jerárquicos.

El segundo experimento (Pipeline Híbrido Unificado) implementa un enfoque jerárquico en dos etapas. En la primera etapa, se entrenó un modelo de clasificación de imágenes basado en MobileNet para identificar la especie del árbol (rosa, patata o manzana), alcanzando una exactitud del 99.84%. En la segunda etapa, la predicción de la especie actúa como un mecanismo de condicionamiento lógico (masking) sobre un único detector YOLOv8, restringiendo el espacio de búsqueda exclusivamente a las enfermedades biológicamente plausibles para la especie identificada. Este enfoque permite reducir drásticamente las confusiones entre especies sin incrementar significativamente el coste computacional.

El tercer experimento (Pipeline Híbrico Modular) explora una arquitectura jerárquica aún más especializada. Tras la clasificación de la especie vegetal mediante MobileNet, se activan modelos de detección independientes entrenados específicamente para cada especie. Cada detector se enfoca únicamente en distinguir entre los distintos patógenos foliares asociados a una única planta hospedadora. Este diseño maximiza la especialización del modelo y minimiza la ambigüedad inter-especie.

6.3. Resultados

Los resultados obtenidos de los tres enfoques experimentales se resumen en la Tabla 1.

Los datos evidencian que la falta de contexto biológico penaliza severamente al modelo *Baseline*, especialmente en las clases Rosas (0.56 mAP50) y Manzanas (0.6356 mAP50). Esto sugiere que el modelo monolítico sufre de "alucinaciones cruzadas", confundiendo texturas de enfermedades similares entre especies distintas.

La incorporación de la etapa de clasificación en los enfoques jerárquicos genera una mejora sustancial: **Análisis detallado por especie (mAP@50 y mAP@50-95):**

- Rosas Esta clase ilustra el mayor beneficio del enfoque jerárquico. El *Baseline* no solo fallaba en la detección (mAP@50 de 0.5640), sino que su precisión de localización era deficiente (mAP@50-95 de 0.2966). La implementación del pipeline disparó la detección por encima del 0.92 y, crucialmente, el modelo Especializado duplicó la calidad de la localización (0.6509), demostrando que los expertos aprenden mejor la morfología compleja de la hoja de rosa.

- Manzanas Se observa una correlación directa entre la eliminación de la ambigüedad y la precisión geométrica. El mAP@50 saltó del 0.6356 al 0.9378 en el enfoque Unificado, arrastrando consigo una mejora sustancial en el mAP@50-95 (de 0.4981 a 0.7110).

Table 1. Comparativa de rendimiento (mAP@50 y mAP@50-95).

	Baseline		Unificado		Especializado	
Especie	**mAP50**	**50-95**	**mAP50**	**mAP50-95**	**mAP50**	**50-95**
Manzana	0.6356	0.4981	0.9378	0.7110	0.9275	0.7145
Patatas	0.7965	0.6296	0.7965	0.6296	0.8239	0.6610
Rosas	0.5640	0.2966	0.9248	0.5354	0.9574	0.6509
Avg	**0.6654**	**0.4748**	**0.8864**	**0.6253**	**0.9029**	**0.6755**

Esto indica que el modelo condicionado no solo encuentra más enfermedades, sino que delimita sus bordes con mucha mayor exactitud.

- Patatas Aunque la detección general (mAP@50) se mantuvo estable en torno a 0.79 en todos los enfoques —sugiriendo una robustez visual inherente al tizón—, el mAP@50-95 revela una ventaja oculta del *Pipeline Especializado*. Este enfoque alcanzó un 0.6610 frente al 0.6296 del *Baseline*, confirmando que un modelo dedicado exclusivamente a la patata logra ajustar mejor las cajas delimitadoras (bounding boxes) que uno generalista.

7. Conclusiones

Sección que presenta, brevemente y a modo de resumen, las principales conclusiones del trabajo realizado. También suele incluir posibles trabajos futuros. Es decir, cuáles son las líneas más prometedoras para continuar con este trabajo, así como posibles propuestas de mejora. IMPORTANTE: estas son las conclusiones científicas alcanzadas en el proyecto; ¡no tus conclusiones personales sobre el trabajo que has realizado!

References

[1] Jayme Garcia Arnal Barbedo. Digital image processing techniques for detecting, quantifying and classifying plant diseases. *SpringerPlus*, 2(1):660, 2013. 1

[2] Pedro F. Felzenszwalb, Ross B. Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010. 2

[3] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *arXiv preprint arXiv:1311.2524*, 2014. 2

[4] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 2, 3

[5] Sharada P Mohanty, David P Hughes, and Marcel Salathé. Using deep learning for image-based plant disease detection. *Frontiers in Plant Science*, 7:1419, 2016. 1

[6] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection, 2016. 2

[7] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, 2016. 2

[8] Serge Savary, Laetitia Willocquet, Sarah Jane Pethybridge, Paul Esker, Neil McRoberts, and Andy Nelson. The global burden of pathogens and pests on major food crops. *Nature Ecology & Evolution*, 3(3):430–439, 2019. 1

[9] Mingxing Tan and Quoc V. Le. Efficientnetv2: Smaller models and faster training, 2021. 2

[10] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yandong Li. Maxvit: Multi-axis vision transformer. *arXiv preprint arXiv:2204.01697*, 2022. 2